Mathematics for IT Engineers

Pál Burai

UNIVERSITY OF DEBRECEN

WWW.UNIDEB.HU

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Contents

1. Complex numbers

1.1 A possible motivation introducing complex numbers

Let us consider the following innocent looking equation:

$$x^2 + 1 = 0.$$

Its right hand side is greater than or equal to 1 for all real x. In other words, the simple quadratic polynomial $x^2 + 1$ has no real root.

In general, a quadratic polynomial has the form

$$p(x) = ax^2 + bx + c,$$

where a, b, c are given real numbers, and a is different from zero. It is a basic fact of hight school mathematics that if the discriminant of p

$$D_p = b^2 - 4ac$$

is less than zero, then there is no real root of p. This was one of the reasons for the extension of the concept of real numbers in the sixteenth century. Gerolamo Cardano, an Italian mathematician was the first who conceived complex numbers in 1545.

Gerolamo Cardano
(1501-1576)

Concerning the polynomial $x^2 + 1$ we can execute the following series of formal calculations:

$$\begin{aligned} x^2 + 1 &= 0 \\ x^2 &= -1 \\ x &= \pm\sqrt{-1}. \end{aligned}$$

This shows the requisiteness of a "number" whose square is negative.

Definition 1.1.1 — The imaginary unit i. The number i is defined by the property

$$i^2 = -1.$$

This number is called the **imaginary unit**.

Using this definition we can perform the following calculations

$$i^2 + 1 = -1 + 1 = 0$$

and

$$(-i)^2 + 1 = i^2 + 1 = 0.$$

This shows that $\pm i$ are the roots of $x^2 + 1$.

1.2 Operations with complex numbers in algebraic form

1.2.1 Algebraic form, real part, imaginary part

Let us start with the definition of algebraic form.

Definition 1.2.1 — Algebraic form of complex numbers. Let a and b be real numbers, then the number

$$z = a + ib$$

is called a **complex number given in algebraic form**.

The set of all complex numbers is denoted by $\mathbb{C}$, that is

$$\mathbb{C} := \{ z \mid z = a + ib,\ a, b \in \mathbb{R} \}.$$

Definition 1.2.2 — Real part and imaginary part of complex numbers. It is conspicuous that a complex number given in algebraic form has two main parts.

$$z = \underbrace{a}_{\mathfrak{Re}(z)} + i \underbrace{b}_{\mathfrak{Im}(z)},$$

where $\mathfrak{Re}(z)$ is the **real part of** z, and $\mathfrak{Im}(z)$ is the **imaginary part of** z.

In other words, $\mathfrak{Re}\colon \mathbb{C} \to \mathbb{R}$ and $\mathfrak{Im}\colon \mathbb{C} \to \mathbb{R}$ are real valued functions with complex domain.

Problem 1.1 Find the real and the imaginary part of the complex numbers

a) $z_1 = 1$ b) $z_2 = -\pi$ c) $z_3 = i$

d) $z_4 = -4i$ e) $z_5 = 1 - 2i$ f) $z_6 = 3i - 4$

Solutions:

a) $\mathfrak{Re}(z_1) = \mathfrak{Re}(1 + 0 \cdot i) = 1$ and $\mathfrak{Im}(z_1) = \mathfrak{Im}(1 + 0 \cdot i) = 0$. The complex number z_1 is a real number as well, so its imaginary part is zero.
This is true in general. The set of real numbers is embedded into the set of complex numbers with zero imaginary part. That is

$$\mathbb{R} \subset \mathbb{C}, \qquad \text{and} \quad \mathfrak{Re}(x) = x, \qquad \mathfrak{Im}(x) = 0,$$

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

b) Taking into account the previous solution, we have $\mathfrak{Re}(z_2) = -\pi$ and $\mathfrak{Im}(z_2) = 0$.
c) A complex number is called **purely imaginary** if its real part is zero. For example $\mathfrak{Re}(z_3) = \mathfrak{Re}(0+1\cdot i) = 0$ and $\mathfrak{Im}(z_3) = \mathfrak{Im}(0+1\cdot i) = 1$, so, i is a purely imaginary number.
d) According to the third example $z_4 = -4i$ is also a purely imaginary number, which gives $\mathfrak{Re}(z_4) = 0$, and $\mathfrak{Im}(z_4) = \mathfrak{Im}(-4i) = -4$.
e) $\mathfrak{Re}(z_5) = \mathfrak{Re}(1-2i) = 1$, and $\mathfrak{Im}(z_5) = \mathfrak{Im}(1-2i) = -2$.
f) $\mathfrak{Re}(z_6) = \mathfrak{Re}(3i-4) = \mathfrak{Re}(-4+3i) = -4$, and $\mathfrak{Im}(z_6) = \mathfrak{Im}(3i-4) = \mathfrak{Im}(-4+3i) = 3$.

1.2.2 Graphical representation of complex numbers on the complex plane

As we saw, a complex number has a real and an imaginary part. These can be considered as the "coordinates" of the complex number. We just rename the axis corresponding to the situation.

More precisely, an arbitrary complex number $z = a + ib$ can be identified as a vector on the plane with coordinates a, b. See figure 1.1.

In this situation the usual x-axis is called the **real axis**, and the usual y-axis is called the **imaginary axis**. There is a one-to-one correspondence between complex numbers in algebraic form and the points of the complex plane.

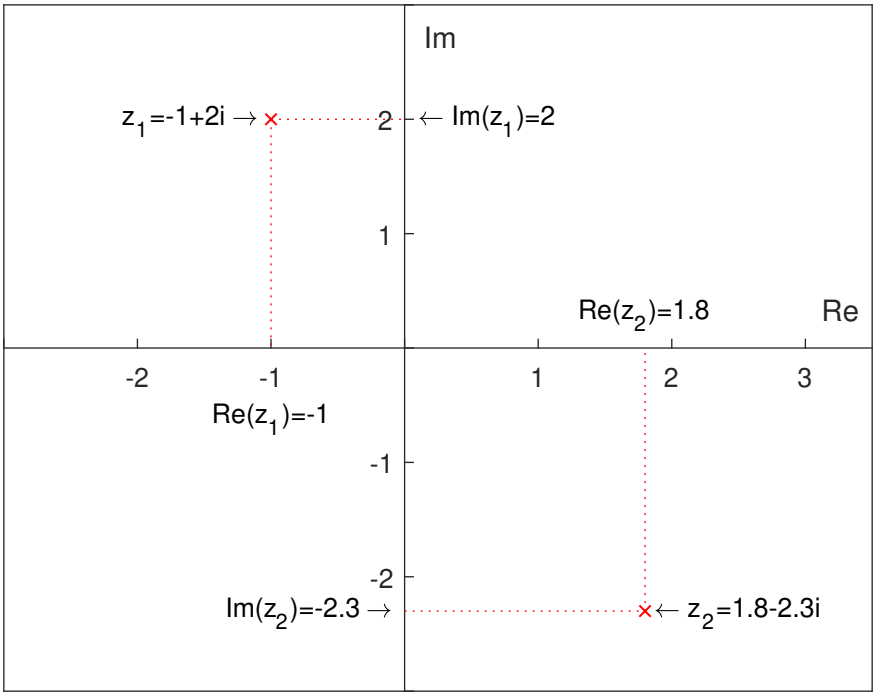
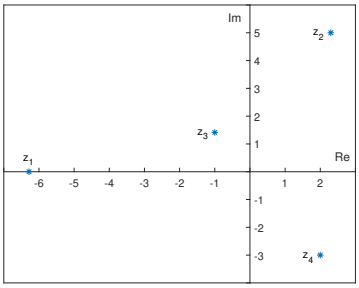
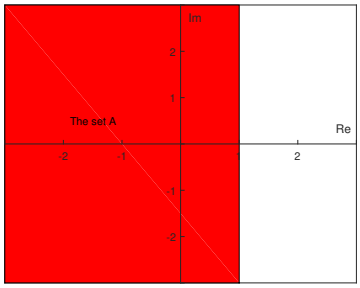


Figure 1.1: Graphical representation of complex numbers

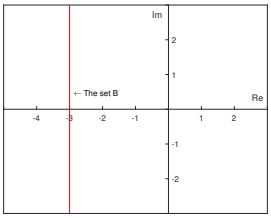
Problem 1.2 Plot the following complex numbers and sets of complex numbers on the complex



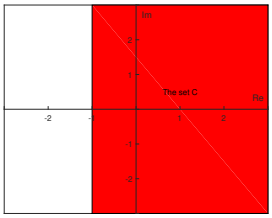
Graphical representation of z_1, z_2, z_3, z_4.



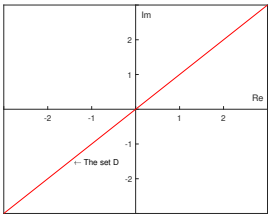
The set A.



The set B.



The set C.



The set D.

plane!

$$z_1 = -2\pi, \quad z_2 = 2.3+5i, \quad z_3 = \sqrt{2}i-1, \quad z_4 = 2-3i,$$

$$A = \{\, z \in \mathbb{C} \mid \mathfrak{Re}(z) \leq 1 \,\}, \quad B = \{\, z \in \mathbb{C} \mid \mathfrak{Re}(z) = -3 \,\},$$

$$C = \{\, z \in \mathbb{C} \mid \mathfrak{Im}(z) \geq -1 \,\}, \qquad D = \{\, z \in \mathbb{C} \mid \mathfrak{Im}(z) = \mathfrak{Re}(z) \,\}.$$

1.2.3 Addition of complex numbers in algebraic form

It is reasonable to define the sum of complex numbers like the sum of two dimensional vectors, that is to say, let's add the corresponding coordinates. This will really be the rule in the case of complex numbers, as the real part of the sum will be the sum of the real parts of the summands and the imaginary part of the sum will be the sum of the imaginary parts of the summands.

Definition 1.2.3 — Addition rule of complex numbers in algebraic form. Let $z = a+ib$ and $w = c+id$ two given complex numbers, then their sum is defined by the following formula

$$z+w = (a+c)+i(b+d).$$

Problem 1.3 Find the sum $z+w$ if

a) $z = i-2$, $w = 2.1+i$ b) $z = -3-5i$, $w = 2.8+5.3i$ c) $z = 2i+3.4$, $w = -1-i$

Solutions:

a) $z+w = (i-2)+(2.1+i) = -2+2.1+I+i = -0.1+2i$,
b) $z+w = (-3-5i)+(2.8+5.3i) = -3+2.8-5i+5.3i = -0.2+0.3i$,
c) $z+w = (2i+3.4)+(-1-i) = 3.4-1+2i-i = 2.4+i$.

Addition has the same properties as the addition of other types of numbers (e.g. real numbers, rational numbers, integers and so on).

Proposition 1.2.1 — Properties of addition of complex numbers.

- Addition of complex numbers

$$+ : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$$

is a **binary operation** on the set of complex numbers. This means that the sum of two complex numbers is a complex number.[1]

- Addition of complex numbers is **associative**, that is to say[2]

$$(z+w)+u = z+(w+u), \text{ for every } \quad z,w,u \in \mathbb{C}.$$

- There exists an **additive unit**, the zero complex number 0 with the property

$$z+0 = 0+z = z, \text{ for every } \quad z \in \mathbb{C}.$$

- All complex numbers z have an **additive inverse** denoted by $-z$ with the property

$$z+(-z) = (-z)+z = 0.$$

- Addition of complex numbers is **commutative**, that is to say

$$z+w = w+z, \text{ for every } \quad z,w \in \mathbb{C}.$$

In short we called the pair $(\mathbb{C},+)$ an **abelian group** or **commutative group**.[3]

1.2.4 Multiplication of complex numbers in algebraic form

Let $z = a+ib$ and $w = c+id$ be given complex numbers. Let us calculate the product formally (multiply every term by every term).

$$zw = (a+ib)(c+id) = ac+aid+ibc+ibid = ac+iad+ibc+\underbrace{i^2}_{=-1}bd = ac-bd+i(ad+bc).$$

Definition 1.2.4 — Multiplication rule of complex numbers in algebraic form. Let $z = a+ib$ and $w = c+id$ two given complex numbers, then their product is defined by the following formula

$$z \cdot w = ac-bd+i(ad+bc).$$

Problem 1.4 Find the product $z \cdot w$ if

a) $z = i-2, w = 2.1+i$ b) $z = -3-5i, w = 2.8+5.3i$ c) $z = 2i+3.4, w = -1-i$

Solutions:

a) $zw = (i-2)(2.1+i) = (-2)\cdot 2.1 - 1\cdot 1 + i((-2)\cdot 1 + 2.1\cdot 1) = -4.2-1+i(-2+2.1) = -5.2+0.1i$,

b) $zw = (-3-5i)(2.8+5.3i) = (-3)\cdot 2.8 - (-5)\cdot 5.3 + i((-3)\cdot 5.3 + (-5)\cdot 2.8) = -8.4+26.5+i(-15.9-14) = 18.1-29.9i$,

c) $zw = (2i+3.4)(-1-i) = 3.4\cdot(-1) - 2\cdot(-1) + i(3.4\cdot(-1)+2\cdot(-1)) = -3.4+2+i(-3.4-2) = -1.4-5.4i$.

[1] This is not always the case. For example subtraction is not a binary operation on the set of natural numbers $1-2=-1$, and -1 is an integer but not a natural number. We say $\mathbb{N}$ **is not closed with respect to substraction**.

[2] Again, not all binary operations are associative. Subtraction on $\mathbb{Z}$ is a binary operation, but it is not associative, e.g. $(3-2)-1 = 0 \neq 2 = 3-(2-1)$.

[3] The concept of group is a central one in algebra. Other important examples are $(\mathbb{Z},+)$, $(\mathbb{R},+)$, and $n \times m$ matrices with addition (this defined later in this book).

Similarly to the case of addition, multiplication also has the same properties as multiplication of real numbers or rational numbers. However, to construct a similar group structure like in the case of addition we should omit zero from $\mathbb{C}$.

Proposition 1.2.2 — Properties of multiplication of complex numbers.

- Multiplication of complex numbers

$$\cdot : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$$

is a **binary operation** on the set of complex numbers. This means that the product of two complex numbers is a complex number.

- Multiplication of complex numbers is **associative**, that is to say

$$(z \cdot w) \cdot u = z \cdot (w \cdot u), \text{ for every } \quad z, w, u \in \mathbb{C}.$$

- There exists a **multiplicative unit**, the complex number 1 with the property

$$1 \cdot z = z \cdot 1 = z, \text{ for every } \quad z \in \mathbb{C}.$$

- All complex numbers z, except for zero, have a **multiplicative inverse** denoted by $\frac{1}{z}$ or z^{-1} with the property

$$z \cdot \frac{1}{z} = 1.$$

- Multiplication of complex numbers is **commutative**, that is to say

$$z \cdot w = w \cdot z, \text{ for every } \quad z, w \in \mathbb{C}.$$

As we see $(\mathbb{C} \setminus \{0\}, \cdot)$ constitutes an abelian group.

Henceforward, if there is no ambiguity we skip the dot for denoting multiplication, so we use the notation zw for the product of z and w instead of $z \cdot w$.

Multiplicative inverse and division

Analogously to real numbers, if z is an arbitrary non-zero complex number, there is a complex number, denoted by z^{-1} or $\frac{1}{z}$, such that

$$z \cdot z^{-1} = 1.$$

Proposition 1.2.3 Let $z = a + ib$ be a given non-zero complex number, then

$$\frac{1}{z} = \frac{a}{a^2 + b^2} - i\frac{b}{a^2 + b^2}.$$

Proof.

$$z \cdot \frac{1}{z} = (a + ib)\left(\frac{a}{a^2 + b^2} - i\frac{b}{a^2 + b^2}\right) = \frac{a^2}{a^2 + b^2} + \frac{b^2}{a^2 + b^2} - i\frac{ab}{a^2 + b^2} + i\frac{ab}{a^2 + b^2} = \frac{a^2 + b^2}{a^2 + b^2} = 1.$$

■

In practice, we use the the following trick to calculate a ratio of two complex numbers given in algebraic form.

■ **Example 1.1**

$$\frac{1+i}{2-3i}=\frac{1+i}{2-3i}\frac{2+3i}{2+3i}=\frac{(1+i)(2+3i)}{(2-3i)(2+3i)}=\frac{-2+5i}{13}=-\frac{2}{13}+\frac{5i}{13}.$$

Here we used the fact that $z\cdot\bar{z}$ is always real.[4] In this way we can avoid division by complex numbers. ■

Problem 1.5 Find the multiplicative inverse of z, where

a) $z=1-i$, b) $z=2i+1$

Solutions:

a) $z=1-i$, so $a=1$, $b=-1$, and

$$\frac{1}{z}=\frac{1}{1^2+(-1)^2}-i\frac{-1}{1^2+(-1)^2}=\frac{1}{2}+\frac{1}{2}i.$$

b) $z=2i+1$, so $a=1$, $b=2$, and

$$\frac{1}{z}=\frac{1}{1^2+2^2}-i\frac{2}{1^2+2^2}=\frac{1}{5}-\frac{1}{5}i.$$

Problem 1.6 Find the ratio of z and w, where

a) $z=1-4i$ $w=8i-2$, b) $z=11i+1$, $w=9+2i$.

Solutions:

a)

$$\frac{1-4i}{-2+8i}=\frac{1-4i}{-2+8i}\cdot\frac{-2-8i}{-2-8i}=\frac{(1-4i)(-2-8i)}{4+64}=\frac{-34}{68}=-\frac{1}{2}.$$

b)

$$\frac{1+11i}{9+2i}=\frac{1+11i}{9+2i}\cdot\frac{9-2i}{9-2i}=\frac{(1+11i)(9-2i)}{81+4}=\frac{31+97i}{85}=\frac{31}{85}+\frac{97}{85}i.$$

Connection between multiplication and addition

In the light of the similarity between addition and multiplication of complex and e.g. real numbers, we can expect an akin connection between multiplication and addition of complex numbers like concerning real numbers. Our requirement is correct. This connection is called **distributivity**.

Proposition 1.2.4 — Distributivity law. For every $z,w,u\in\mathbb{C}$ we have

$$(z+w)u=zu+wu.$$

[4] See the definition and properties of complex conjugate $\bar{z}$ in subsection 1.2.5

If a set is endowed with two binary operations like the set of complex numbers (abelian group with respect to both operations, and the distributivity law connects the two operations), then it is said to be a **field**.[5]

1.2.5 Conjugate of complex numbers

Up till now we saw the similarities between complex and e.g. real numbers. Addition, multiplication, and machinery of calculations do not essentially differ in the mentioned situations.

Here is the firs essential difference between these sets of numbers.

Definition 1.2.5 — Complex conjugate. Let $z \in \mathbb{C}$ be given in algebraic form $z = a + bi$, then the number

$$\bar{z} = a - bi$$

is said to be the **conjugate of** z.

Geometrically conjugation is the reflection with respect to the real-axis on the complex plane, see Figure 1.4. Conjugation changes the sign of the imaginary part and leaves the real part untouched.

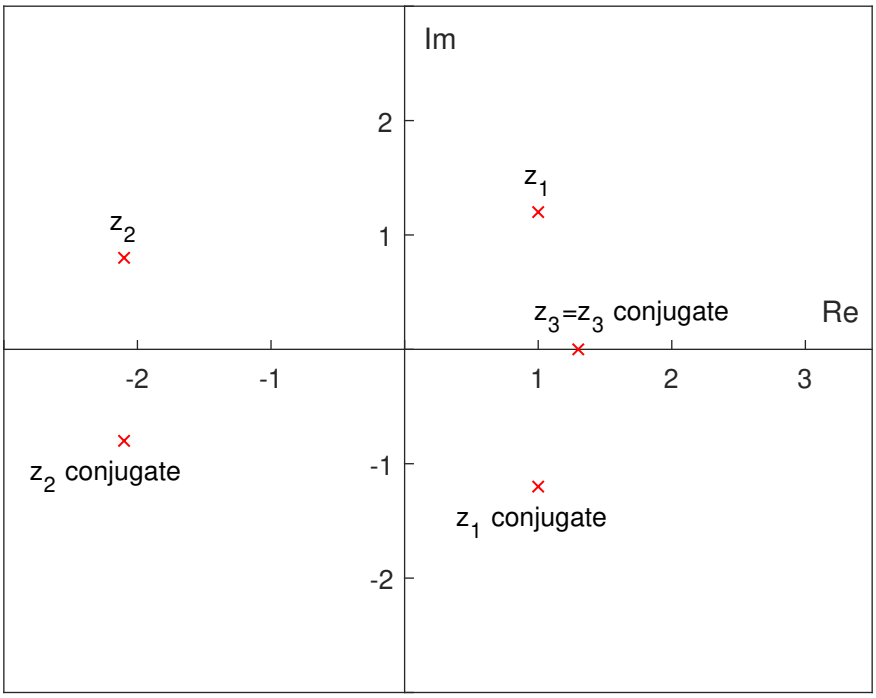


Figure 1.4: Complex numbers and their conjugates

Proposition 1.2.5 — Properties of conjugation of complex numbers.

- The conjugate of

[5] This concept is another very important one in algebra. Other important examples are the field of real numbers, the field of rational numbers, and finite fields e.g. integers modulo p, where p is a prime.

a real number is a real number.[6]

$$\bar{x} = x, \qquad \text{for every } x \in \mathbb{R}.$$

- Conjugation is an idempotent (unary) operation. In other words, double conjugation has no effect.

$$\bar{\bar{z}} = z, \qquad \text{for every } z \in \mathbb{C}.$$

- Conjugate of a sum is the sum of the conjugates of the summands.

$$\overline{z+w} = \bar{z} + \bar{w}, \qquad \text{for every } z, w \in \mathbb{C}.$$

- Conjugate of a product is the product of the conjugates.

$$\overline{z \cdot w} = \bar{z} \cdot \bar{w}, \qquad \text{for every } z, w \in \mathbb{C}.$$

- A complex number multiplied by its conjugate always results a real number. More precisely it results the sum of the squares of the real and the imaginary parts.[7]

$$z\bar{z} = (\mathfrak{Re}z)^2 + (\mathfrak{Im}z)^2, \qquad \text{for every } z \in \mathbb{C}.$$

Problem 1.7 Find the conjugate of z, w, $z-w$, $z+w$, and zw, where

$$z = 1 + \frac{1}{2}i \qquad \text{and} \qquad w = 0.2 - 4i.$$

Solution: Because z and w are given in different forms, at first we have to transform for example z into decimal form: $z = 1 + 0.5i$. Now we can start the calculation.

$$\bar{z} = \overline{1+0.5i} = 1 - 0.5i, \quad \bar{w} = \overline{0.2-4i} = 0.2 + 4i,$$

$$\overline{z-w} = \overline{(1+0.5i)-(0.2-4i)} = \overline{1-0.2+i(0.5+4)} = \overline{0.8+4.5i} = 0.8 - 4.5i,$$

$$\overline{z+w} = \overline{(1+0.5i)+(0.2-4i)} = \overline{1+0.2+i(0.5-4)} = \overline{1.2-3.5i} = 1.2 + 3.5i,$$

$$\overline{zw} = \overline{(1+0.5i)(0.2-4i)} = \overline{0.2+2+i(-4+0.1)} = \overline{2.2-3.9i} = 2.2 + 3.9i.$$

1.3 Polar form and exponential form of complex numbers

As we have seen, a complex number can be considered as a vector on the complex plane.

A vector on the plane is determined uniquely by its coordinates, but this is not the only possibility.

A motivation to look for a new approach can be the laborious calculation of high powers. A simple way of calculating the powers is not the sole advantage of the new approach, easy calculation of ratios and roots are important as well.

Let $z = a + ib$ be a complex number. We can identify it by the vector from the origin to the point (a, b). This vector is also determined uniquely by its length and by its angle (counterclockwise) between the real axis and the vector.

1.3.1 Length and angle of complex numbers

[6] Actually, this is the real axis, which is also the axis of the mentioned reflection which remains fixed during the reflection process. This is why real numbers remain untouched by conjugation.

[7] As a matter of fact, this is the square of the length of the complex number z. See subsection 1.3.1.

Definition 1.3.1 — The length of complex numbers. Let $z = a + ib$ be a given complex number, then the quantity

$$|z| = \sqrt{a^2 + b^2} = \sqrt{(\mathfrak{Re}z)^2 + (\mathfrak{Im}z)^2}$$

is said to be the **length of** z.

This definition comes from Pythagoras' theorem.

Proposition 1.3.1 — Properties of the length of complex numbers. Let $z, w \in \mathbb{C}$ be arbitrary, then

- the length z is always non-negative, and it is zero if and only if $z = 0$;
- the length of the product is the product of the lengths, that is

$$|zw| = |z||w|;$$

- the length of a sum is always less than or equal the sum of the lengths, that is

$$|z + w| \leq |z| + |w|, \qquad \text{triangle-inequality;}$$

- the length of a real number equals to its absolute value;
- taking the square root of the product, a complex number multiplied by its conjugate results the length too, that is

$$|z| = \sqrt{z\bar{z}}.$$

Problem 1.8 Calculate the length of the following complex numbers

a) $z_1 = -1$ b) $z_2 = 3i$ c) $z_3 = 1 - i$ d) $z_4 = -1 - 2i$

Solutions:

a) Because z_1 is a real number its length equals to its absolute value: $|z_1| = 1$,

b) z_2 is purely imaginary, so its length equals to the absolute value of its imaginary part: $|z_2| = 3$,

c) $|z_3| = \sqrt{1^2 + (-1)^2} = \sqrt{1+1} = \sqrt{2}$,

d) $|z_3| = \sqrt{(-1)^2 + (-2)^2} = \sqrt{1+4} = \sqrt{5}$.

It seems to be straightforward (according to the figure 1.5) to use the arctan function for the definition of the angle of complex numbers. However, the situation is a little bit more complicated because of the range of arctan.

Firstly, we cannot assign an angle to the zero complex number.

Important remark: $z = 0$ has no angle, this entails it has no polar form!

Secondly, the angle depends on the position of the complex number. Namely, it is important to know in which quarter it is. The mazy appearance of the definition is the consequence of this second observation.

Definition 1.3.2 — The angle of complex numbers. Let $z = a + ib \neq 0$ be a given complex

number, then the quantity

$$\varphi_z = \begin{cases} \arctan\left(\frac{b}{a}\right), & \text{if } a > 0 \text{ and } b \geq 0 \\ \arctan\left(\frac{b}{a}\right) + 2\pi, & \text{if } a > 0 \text{ and } b < 0 \\ \arctan\left(\frac{b}{a}\right) + \pi, & \text{if } a < 0 \text{ and } b \geq 0 \\ \arctan\left(\frac{b}{a}\right) + \pi, & \text{if } a < 0 \text{ and } b < 0 \\ \frac{\pi}{2} & \text{if } a = 0 \text{ and } b > 0 \\ \frac{3\pi}{2} & \text{if } a = 0 \text{ and } b < 0 \end{cases}$$

is said to be the **angle of** z.

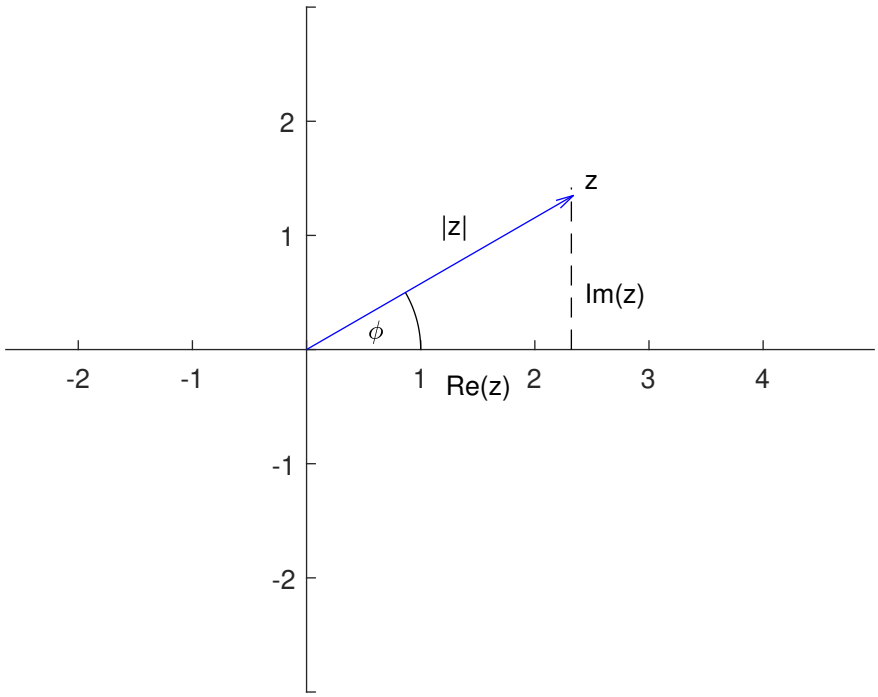


Figure 1.5: The length and angle of z

Important remark: The angle is always greater than, or equal to zero and less than 2π. If it is outside of this range after a certain calculation, then it is necessary to take the angle modulo 2π. We will see examples for this in Subsection 1.3.2.

Problem 1.9 Calculate the angle of the following complex numbers

a) $z_1 = -1$ b) $z_2 = 3i$ c) $z_3 = 1 - i$

d) $z_4 = -1 - \sqrt{3}i$ e) $z_5 = -i$ f) $z_6 = 1 + i$

Solutions:

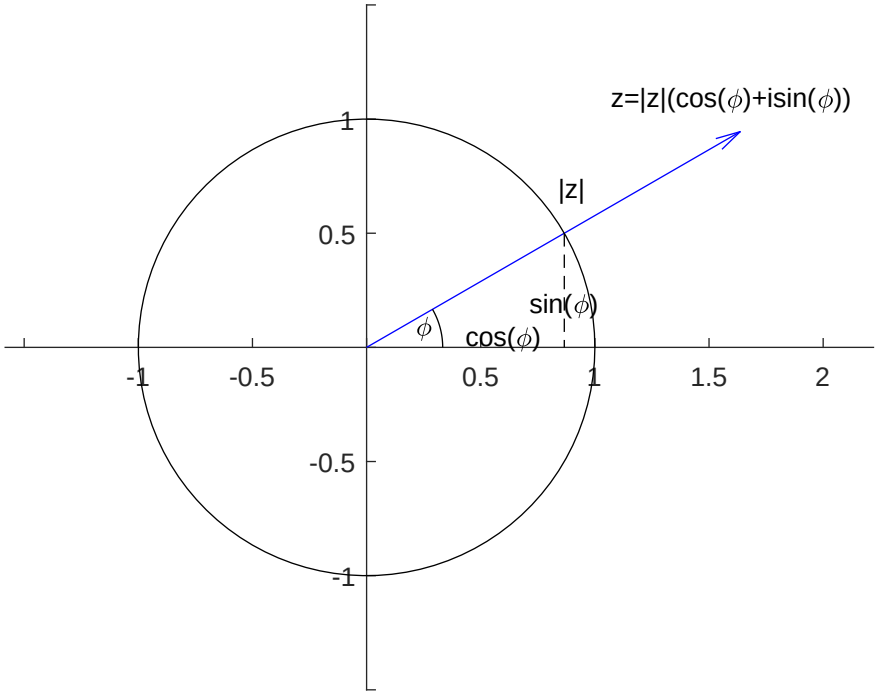


Figure 1.6: Polar form of z

a) $z_1 = -1 + 0 \cdot i$, so $a < 0$, and $b \geq 0$. This is the third case, where the rule is

$$\varphi_z = \arctan \frac{b}{a} + \pi = \arctan 0 + \pi = \pi.$$

b) $z_2 = 0 + 3i$, so $a = 0$, and $b > 0$. This is the fifth case, where the rule is

$$\varphi_z = \frac{\pi}{2}.$$

c) $z_3 = 1 + (-1) \cdot i$, so $a > 0$, and $b < 0$. This is the second case, where the rule is

$$\varphi_z = \frac{b}{a} + 2\pi = \arctan \frac{1}{-1} + 2\pi = \arctan(-1) + 2\pi = \frac{-\pi}{4} + 2\pi = \frac{7\pi}{4}.$$

d) $z_4 = -1 + (-\sqrt{3}) \cdot i$, so $a < 0$, and $b < 0$. This is the fourth case, where the rule is

$$\varphi_z = \arctan\left(\frac{b}{a}\right) + \pi = \arctan\left(\frac{-1}{-\sqrt{3}}\right) + \pi = \arctan\left(\frac{1}{\sqrt{3}}\right) + \pi = \frac{\pi}{3} + \pi = \frac{4\pi}{3}.$$

e) $z_5 = 0 + (-3)\cdot i$, so $a = 0$, and $b < 0$. This is the sixth case, where the rule is

$$\varphi_z = \frac{3\pi}{2}.$$

f) $z_6 = 1 + 1\cdot i$, so $a > 0$, and $b \geq 0$. This is the first case, where the rule is

$$\varphi_z = \arctan\left(\frac{b}{a}\right) = \arctan\left(\frac{1}{1}\right) = \arctan(1) = \frac{\pi}{4}.$$

Definition 1.3.3 — The polar form of complex numbers. Let $z \neq 0$ be a given complex number its length is denoted by $|z|$ and its angle is denoted by φ_z, then it can be written in the form

$$z = |z|(\cos\varphi_z + i\sin\varphi_z)$$

which is said to be the **polar form of** z.

Problem 1.10 Find the polar form of the following complex numbers

a) $z_1 = -5$ b) $z_2 = 11i$ c) $z_3 = 2\sqrt{3} + 2i$ d) $z_4 = -1.2 - 0.4i$

Solutions:

a) z_1 is real, so its length equals to its absolute value.

$$|z_1| = |-5| = 5.$$

Its real part is negative and its imaginary part is zero. This is the third case. So its angle is

$$\varphi_{z_1} = \arctan\left(\frac{0}{-5}\right) + \pi = \arctan(0) + \pi = 0 + \pi = \pi.$$

The polar form of z_1 is

$$z_1 = 5(\cos\pi + i\sin\pi).$$

b) z_2 is purely imaginary, so its length is the absolute value of its imaginary part.

$$|z_2| = |11| = 11.$$

Its real part is zero and its imaginary part is positive. This is the fifth case. So its angle is

$$\varphi_{z_2} = \frac{\pi}{2}.$$

The polar form of z_2 is

$$z_2 = 11\left(\cos\frac{\pi}{2} + i\sin\frac{\pi}{2}\right).$$

c) Both the real and the imaginary parts of z_3 are different from zero, so its length is the square root of the sum of the squares of its real and imaginary parts.

$$|z_3| = \sqrt{(2\sqrt{3})^2 + 2^2} = \sqrt{12+4} = \sqrt{16} = 4.$$

Both the real and imaginary parts are positive. This is the first case. So its angle is

$$\varphi_{z_3} = \arctan\left(\frac{2}{2\sqrt{3}}\right) = \arctan\left(\frac{1}{\sqrt{3}}\right) = \frac{\pi}{6}.$$

The polar form of z_3 is

$$z_3 = 4\left(\cos\frac{\pi}{3} + i\sin\frac{\pi}{3}\right).$$

d) Both the real and the imaginary parts of $z_4 = -1.2 - 0.4i$ are different from zero, so its length is the square root of the sum of the squares of its real and imaginary parts.

$$|z_4| = \sqrt{(-1.2)^2 + (-0.4)^2} = \sqrt{1.44+0.16} = \sqrt{1.6} \approx 1.2649.$$

Here we have only an approximate value of the length.
Both the real and imaginary parts are negative. This is the fourth case. So its angle is

$$\varphi_{z_4} = \arctan\left(\frac{-0.4}{-1.2}\right) = \arctan\left(\frac{1}{3}\right) \approx \underbrace{0.3218}_{\text{in radian}} \approx 18.4349^\circ.$$

Again, we have only an approximate value of the angle.[8]
The (approximate) polar form of z_4 is

$$z_4 \approx 1.2649\left(\cos(18.4349^\circ) + i\sin(18.4349^\circ)\right).$$

1.3.2 Calculation with complex numbers in polar form

As it has been mentioned, algebraic form fits well for addition and subtraction of complex numbers, however, it is clumsy and tiresome to calculate the product of numerous complex numbers or to raise a complex number to a hight power.

Polar form fits perfectly well for these operations.

Taking an angle modulo 2π

Let us consider the angle $\frac{13\pi}{3}$. This is greater than $2\pi = \frac{6\pi}{3}$. The main point is that it is outside of the interval $[0, 2\pi[$.

If we get an angle like this after some calculations with complex numbers, it is necessary to take it modulo 2π.

Let us see how it works in practice.

$$\frac{13\pi}{3} = \frac{\pi}{3} + \frac{12\pi}{3} = \frac{\pi}{3} + 2\cdot 2\pi, \quad \text{that is} \quad \frac{13\pi}{3} \equiv \frac{\pi}{3} \mod 2\pi.$$

So we use $\frac{\pi}{3}$ instead of $\frac{13\pi}{3}$.

In general, if $\varphi \notin [0, 2\pi[$, then there is a unique integer $k \in \mathbb{Z}$ and an angle $\psi \in [0, 2\pi[$ such that

$$\varphi = \psi + k\cdot 2\pi,$$

[8] For the numerical calculation one can use a calculator or any kind of mathematical software. For example for the conversion of radians into degrees one can use the command `rad2deg` in Matlab.

which is denoted by

$$\varphi \equiv \psi \mod 2\pi,$$

and we say that φ **is congruent to** ψ **modulo** 2π.[9]

If the angle is given in degree instead of radian, one can use $\mod 360^\circ$ instead of $\mod 2\pi$.

Problem 1.11 Find the angle ψ such that $\varphi \equiv \psi \mod 2\pi$ and $\psi \in [0, 2\pi[$, where

a) $\varphi = \frac{23\pi}{5}$, b) $\varphi = \frac{-11\pi}{2}$, c) $\varphi = \frac{-\pi}{4}$, d) $\varphi = 12\pi$, e) $\varphi = \frac{13\pi}{2}$.

Solutions:

a)

$$\varphi = \frac{23\pi}{5} = \frac{3\pi + 4\cdot 5\pi}{5} = \frac{3\pi}{5} + 4\pi = \frac{3\pi}{5} + 2\cdot 2\pi \equiv \frac{3\pi}{5} \mod 2\pi,$$

b)

$$\varphi = \frac{-11\pi}{2} = -\frac{3\pi + 4\cdot 2\pi}{2} = -\frac{3\pi}{2} - 2\cdot 2\pi \equiv -\frac{3\pi}{2} \mod 2\pi \equiv \frac{\pi}{2} \mod 2\pi,$$

c)

$$\varphi = \frac{-\pi}{4} \equiv \frac{3\pi}{2} \mod 2\pi,$$

d)

$$\varphi = 12\pi = 0 + 6\cdot 2\pi \equiv 0 \mod 2\pi,$$

e)

$$\varphi = \frac{13\pi}{2} = \frac{\pi + 6\cdot 2\pi}{2} = \frac{\pi}{2} + 3\cdot 2\pi \equiv \frac{\pi}{2} \mod 2\pi.$$

Multiplication and raising to a power

A reminder from high-school: We will need the following two trigonometric identities. If α and β are angles then

$$\cos\alpha\cos\beta - \sin\alpha\sin\beta = \cos(\alpha+\beta), \quad \text{and} \quad \cos\alpha\sin\beta + \cos\beta\sin\alpha = \sin(\alpha+\beta).$$

[9] The 12 hour clock is a good analogy which can help to understand this concept. This clock uses arithmetic modulo 12.

Theorem 1.3.2 — Multiplication of complex numbers in polar form. Let $z = |z|(\cos\varphi_z + i\sin\varphi_z)$ and $w = |w|(\cos\varphi_w + i\sin\varphi_w)$ be given non-zero complex numbers, then

$$zw = |z||w|(\cos\psi + i\sin\psi), \quad \text{where} \quad (\varphi_z + \varphi_w) \equiv \psi \mod 2\pi.$$

Proof. Let $z = |z|(\cos\varphi_z + i\sin\varphi_z)$ and $w = |w|(\cos\varphi_w + i\sin\varphi_w)$ be the given non-zero complex numbers, then

$$\begin{aligned} zw = |z|(\cos\varphi_z + i\sin\varphi_z)|w|(\cos\varphi_w + i\sin\varphi_w) = \\ = |z||w|(\cos\varphi_z\cos\varphi_w - \sin\varphi_z\sin\varphi_w + i(\cos\varphi_z\sin\varphi_w + \cos\varphi_w\sin\varphi_z)) = \\ = |z||w|(\cos(\varphi_z + \varphi_w) + i\sin(\varphi_z + \varphi_w)). \end{aligned}$$

■

Using the previous theorem, it is easy to derive now the rule of raising to a power if the exponent is a natural number. Indeed, if $n \in \mathbb{N}$ and $z \in \mathbb{C}$, then

$$z^n = \underbrace{z\cdots z}_{n \text{ times}},$$

which allows us the recursive application of the previous definition of multiplication. Let $n \in \mathbb{N}$ be a given natural number, then

$$z^n = \underbrace{z\cdot z\cdots z}_{n \text{ times}} = \underbrace{|z|(\cos\varphi_z + i\sin\varphi_z)\cdots|z|(\cos\varphi_z + i\sin\varphi_z)}_{n \text{ times}} = |z|^n(\cos(n\varphi_z) + i\sin(n\varphi_z)).$$

The only thing we have to take care of is the angle $n\varphi_z$. If $n\varphi_z \notin [0, 2\pi[$, then it is necessary to substract or add 2π as many times to $n\varphi_z$ till the result will be in the required interval $[0, 2\pi[$ (take the angle $\mod 2\pi$). Actually, we have proved the following theorem.

Theorem 1.3.3 — Raising to a power of complex numbers in polar form. [a] Let $z \in \mathbb{C}$, $z = |z|(\cos\varphi_z + i\sin\varphi_z$ and $n \in \mathbb{N}$, then

$$z^n = |z|^n(\cos\psi + i\sin\psi), \quad \text{where} \quad \psi \equiv n\varphi_z \mod 2\pi.$$

[a] This theorem is also known as de Moivre's theorem or de Moivre's formula named after a French mathematician Abraham de Moivre.

■ **Example 1.2** Let $z = 2(cos\frac{\pi}{2} + i\sin\frac{\pi}{2})$ and $n = 10$, then

$$\begin{aligned} z^{10} = 2^{10}\left(\cos\left(10\frac{\pi}{2}\right) + i\sin\left(10\frac{\pi}{2}\right)\right) = 2^{10}(\cos(5\pi) + i\sin(5\pi)) = \\ = 1024(\cos(\pi + 2\cdot 2\pi) + \sin(\pi + 2\cdot 2\pi)) = 1024(\cos(\pi) + \sin(\pi)) \end{aligned}$$

■

Problem 1.12 Find the product zw and the power z^n, where

a) $z = 2\left(\cos\left(\frac{\pi}{11}\right) + i\sin\left(\frac{\pi}{11}\right)\right)$, $w = 5\left(\cos\left(\frac{3\pi}{5}\right) + i\sin\left(\frac{3\pi}{5}\right)\right)$, $n = 10$.

b) $z = 3\left(\cos\left(\frac{17\pi}{10}\right) + i\sin\left(\frac{17\pi}{10}\right)\right)$, $w = -2\left(\cos\left(\frac{\pi}{7}\right) + i\sin\left(\frac{\pi}{7}\right)\right)$, $n = 5$.

Solutions:

a)

$$zw = 2\cdot 5\left(\cos\left(\frac{\pi}{11}+\frac{3\pi}{5}\right)+i\sin\left(\frac{\pi}{11}+\frac{3\pi}{5}\right)\right) =$$

$$= 10\left(\cos\left(\frac{5\pi+33\pi}{55}\right)+i\sin\left(\frac{5\pi+33\pi}{55}\right)\right) = 10\left(\cos\left(\frac{38\pi}{55}\right)+i\sin\left(\frac{38\pi}{55}\right)\right),$$

$$z^{10} = 2^{10}\left(\cos\left(10\frac{\pi}{11}\right)+i\sin\left(10\frac{\pi}{11}\right)\right) = 1024\left(\cos\left(\frac{10\pi}{11}\right)+i\sin\left(\frac{10\pi}{11}\right)\right).$$

b)

$$zw = 3\cdot(-2)\left(\cos\left(\frac{17\pi}{10}+\frac{\pi}{7}\right)+i\sin\left(\frac{17\pi}{10}+\frac{\pi}{7}\right)\right) =$$

$$= -6\left(\cos\left(\frac{119\pi+10\pi}{70}\right)+i\sin\left(\frac{119\pi+10\pi}{70}\right)\right) =$$

$$-6\left(\cos\left(\frac{189\pi}{70}\right)+i\sin\left(\frac{189\pi}{70}\right)\right) =$$

$$= -6\left(\cos\left(\frac{49\pi+70\cdot 2\pi}{70}\right)+i\sin\left(\frac{49\pi+70\cdot 2\pi}{70}\right)\right) =$$

$$-6\left(\cos\left(\frac{49\pi}{70}+2\pi\right)+i\sin\left(\frac{49\pi}{70}+2\pi\right)\right) = -6\left(\cos\left(\frac{49\pi}{70}\right)+i\sin\left(\frac{49\pi}{70}\right)\right)$$

$$z^5 = 3^5\left(\cos\left(5\frac{17\pi}{10}\right)+i\sin\left(5\frac{17\pi}{10}\right)\right) = 243\left(\cos\left(\frac{85\pi}{10}\right)+i\sin\left(\frac{85\pi}{10}\right)\right) =$$

$$= 243\left(\cos\left(\frac{5\pi+40\cdot 2\pi}{10}\right)+i\sin\left(\frac{5\pi+40\cdot 2\pi}{10}\right)\right) =$$

$$= 243\left(\cos\left(\frac{5\pi}{10}+4\cdot 2\pi\right)+i\sin\left(\frac{5\pi}{10}+4\cdot 2\pi\right)\right) = 243\left(\cos\left(\frac{\pi}{2}\right)+i\sin\left(\frac{\pi}{2}\right)\right).$$

Division of complex numbers in polar form

Theorem 1.3.4 Let $z = |z|(\cos\varphi_z + i\sin\varphi_z)$ and $w = |w|(\cos\varphi_w + i\sin\varphi_w)$ be given non-zero complex numbers, then

$$\frac{z}{w} = \frac{|z|}{|w|}(\cos\psi + i\sin\psi), \quad \text{where} \quad (\varphi_z - \varphi_w) \equiv \psi \mod 2\pi.$$

Proof. Using the "reminder from high-school" trigonometric identities from the beginning of the

previous subsection we get

$$\frac{z}{w}=\frac{|z|}{|w|}\frac{\cos\varphi_z+i\sin\varphi_z}{\cos\varphi_w+i\sin\varphi_w}=\frac{|z|}{|w|}\frac{\cos\varphi_z+i\sin\varphi_z}{\cos\varphi_w+i\sin\varphi_w}\cdot\frac{\cos\varphi_w-i\sin\varphi_z}{\cos\varphi_-i\sin\varphi_w}=$$

$$=\frac{|z|}{|w|}\frac{\cos\varphi_z\cos\varphi_w+\sin\varphi_z\sin\varphi_w+i(-\cos\varphi_z\sin\varphi_w+\cos\varphi_w\sin\varphi_z)}{(\cos\varphi_w)^2+(\sin\varphi_w)^2}=$$

$$\frac{|z|}{|w|}(\cos(\varphi_z-\varphi_w)+i\sin(\varphi_z-\varphi_w)).$$

■

Problem 1.13 Find the ratio $\frac{z}{w}$, where

a)

$$z=3(\cos\pi+i\sin\pi),\qquad w=6\left(\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}\right),$$

b)

$$z=12\left(\cos\frac{3\pi}{10}+i\sin\frac{3\pi}{10}\right),\qquad w=3\left(\cos\frac{\pi}{5}+i\sin\frac{\pi}{5}\right).$$

Solutions:

a)

$$\frac{z}{w}=\frac{3}{6}\left(\cos\left(\pi-\frac{\pi}{2}\right)+i\sin\left(\pi-\frac{\pi}{2}\right)\right)=\frac{1}{2}\left(\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}\right),$$

b)

$$\frac{z}{w}=\frac{12}{3}\left(\cos\left(\frac{3\pi}{10}-\frac{\pi}{5}\right)+i\sin\left(\frac{3\pi}{10}-\frac{\pi}{5}\right)\right)=4\left(\cos\frac{\pi}{10}+i\sin\frac{\pi}{10}\right).$$

*n*th roots

Reminder from high school: Let $a\in\mathbb{R}$, $n\in\mathbb{N}$ be given numbers and we are looking for a real number x which is a solution of the equation

$$x^n=a.$$

If a is negative, then n is assumed to be odd. Every real number has a unique odd nth root, and every non-negative real number has a unique even nth root. E.g. there is no real square root of -2.

The situation in the case of complex numbers is totally different. All non-zero complex numbers have n different nth roots. This is not surprising in the light of the raising to a power rule (see. Theorem 1.3.3), because there are exactly n different angles in the interval $[0,2\pi[$ which multiplied by n will be congruent to a given angle φ modulo 2π. Actually these are

$$\frac{\varphi}{n}+k\frac{2\pi}{n},\qquad k=0,1,\ldots,n-1.$$

Indeed,

$$n\left(\frac{\varphi}{n}+k\frac{2\pi}{n}\right)=\varphi+k\cdot 2\pi\equiv\varphi \mod 2\pi, \qquad k=0,1,\ldots,n-1.$$

Using this and Theorem 1.3.3 we immediately get the following statement.

Theorem 1.3.5 — nth roots of complex numbers. Let $z=|z|(\cos\varphi_z+i\sin\varphi_z)$ be a complex number and $n\in\mathbb{N}$ be a natural number greater than or equal to 2. Then z has n different complex nth roots ζ_k, $k=0,1,\ldots,n-1$, which are given by the formula

$$\zeta_k=\sqrt[n]{|z|}\left(\cos\left(\frac{\varphi_z}{n}+k\frac{2\pi}{n}\right)+i\sin\left(\frac{\varphi_z}{n}+k\frac{2\pi}{n}\right)\right), \qquad k=0,1,\ldots,n-1.$$

Geometrically this means that we take a vector with nth root length of the original vector and we also divide the angle of the original vector by n. This results ζ_0. Now we rotate ζ_0 by $\frac{2\pi}{n}$, and we have ζ_1, we rotate ζ_1 by the same angle, and so on. We repeat this process $n-1$ times, at last we get ζ_{n-1}.

■ **Example 1.3** Let $z=16\left(\cos\frac{\pi}{3}+i\sin\frac{\pi}{3}\right)$. Find the fourth roots of z.

Solution: We will have four fourth roots. Let's use the formula from the previous theorem.

$$\zeta_0=\sqrt[4]{16}\left(\cos\left(\frac{\frac{\pi}{3}}{4}+0\cdot\frac{2\pi}{4}\right)+i\sin\left(\frac{\frac{\pi}{3}}{4}+0\cdot\frac{2\pi}{4}\right)\right)=2\left(\cos\left(\frac{\pi}{12}\right)+i\sin\left(\frac{\pi}{12}\right)\right),$$

$$\zeta_1=\sqrt[4]{16}\left(\cos\left(\frac{\frac{\pi}{3}}{4}+1\cdot\frac{2\pi}{4}\right)+i\sin\left(\frac{\frac{\pi}{3}}{4}+1\cdot\frac{2\pi}{4}\right)\right)=2\left(\cos\left(\frac{7\pi}{12}\right)+i\sin\left(\frac{7\pi}{12}\right)\right),$$

$$\zeta_2=\sqrt[4]{16}\left(\cos\left(\frac{\frac{\pi}{3}}{4}+2\cdot\frac{2\pi}{4}\right)+i\sin\left(\frac{\frac{\pi}{3}}{4}+2\cdot\frac{2\pi}{4}\right)\right)=2\left(\cos\left(\frac{13\pi}{12}\right)+i\sin\left(\frac{13\pi}{12}\right)\right),$$

$$\zeta_3=\sqrt[4]{16}\left(\cos\left(\frac{\frac{\pi}{3}}{4}+3\cdot\frac{2\pi}{4}\right)+i\sin\left(\frac{\frac{\pi}{3}}{4}+3\cdot\frac{2\pi}{4}\right)\right)=2\left(\cos\left(\frac{19\pi}{12}\right)+i\sin\left(19\frac{\pi}{12}\right)\right)$$

■

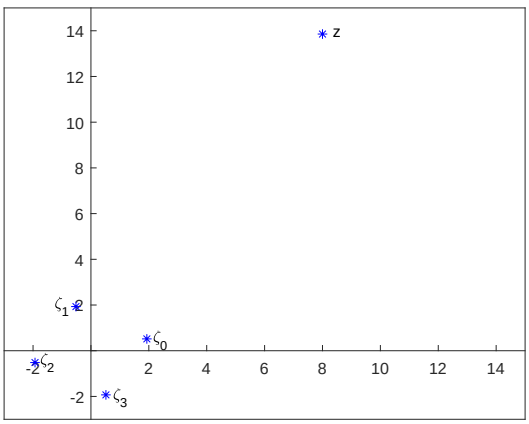


Figure 1.7: Fourth roots of z

Problem 1.14 Find the nth roots of z, where

a) $z = 27(\cos\pi + i\sin\pi), \qquad n = 3,$

b) $z = 32\left(\cos\frac{2\pi}{5} + i\sin\frac{2\pi}{5}\right), \qquad n = 4,$

c) $z = \cos\frac{3\pi}{7} + i\cos\frac{3\pi}{7}, \qquad n = 5.$

Solutions:

a) The length of z is 27, so the length of the third roots will be $\sqrt[3]{27} = 3$ and the rotation angle will be $\frac{2\pi}{3}$.

$$\zeta_0 = 3\left(\cos\left(\frac{\pi}{9} + 0\cdot\frac{2\pi}{3}\right) + i\sin\left(\frac{\pi}{9} + 0\cdot\frac{2\pi}{3}\right)\right) = 3\left(\cos\left(\frac{\pi}{9}\right) + i\sin\left(\frac{\pi}{9}\right)\right),$$

$$\zeta_1 = 3\left(\cos\left(\frac{\pi}{9} + 1\cdot\frac{2\pi}{3}\right) + i\sin\left(\frac{\pi}{9} + 1\cdot\frac{2\pi}{3}\right)\right) = 3\left(\cos\left(\frac{7\pi}{9}\right) + i\sin\left(\frac{7\pi}{9}\right)\right),$$

$$\zeta_2 = 3\left(\cos\left(\frac{13\pi}{9}\right) + i\sin\left(\frac{13\pi}{9}\right)\right) = 3\left(\cos\left(\frac{13\pi}{9}\right) + i\sin\left(\frac{13\pi}{9}\right)\right).$$

b) The length of z is 32, so the length of the fourth roots will be $\sqrt[4]{32} = 2\sqrt[4]{2}$ and the rotation angle will be $\frac{2\pi}{4} = \frac{\pi}{2}$.

$$\zeta_0 = 2\sqrt[4]{2}\left(\cos\left(\frac{2\pi}{20} + 0\cdot\frac{\pi}{2}\right) + i\sin\left(\frac{2\pi}{20} + 0\cdot\frac{\pi}{2}\right)\right) = 2\sqrt[4]{2}\left(\cos\frac{\pi}{10} + i\sin\frac{\pi}{10}\right),$$

$$\zeta_1 = \zeta_0 = 2\sqrt[4]{2}\left(\cos\left(\frac{\pi}{10} + 1\cdot\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{10} + 1\cdot\frac{\pi}{2}\right)\right) = 2\sqrt[4]{2}\left(\cos\frac{6\pi}{10} + i\sin\frac{6\pi}{10}\right),$$

$$\zeta_2 = \zeta_0 = 2\sqrt[4]{2}\left(\cos\left(\frac{\pi}{10} + 2\cdot\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{10} + 2\cdot\frac{\pi}{2}\right)\right) = 2\sqrt[4]{2}\left(\cos\frac{11\pi}{10} + i\sin\frac{11\pi}{10}\right),$$

$$\zeta_3 = \zeta_0 = 2\sqrt[4]{2}\left(\cos\left(\frac{\pi}{10} + 3\cdot\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{10} + 3\cdot\frac{\pi}{2}\right)\right) = 2\sqrt[4]{2}\left(\cos\frac{16\pi}{10} + i\sin\frac{16\pi}{10}\right).$$

c) The length of z is 1, so the length of the third roots will be $\sqrt[5]{1}=1$ and the rotation angle will be $\frac{2\pi}{5}$.

$$\zeta_0=\cos\left(\frac{3\pi}{35}+0\cdot\frac{2\pi}{5}\right)+i\sin\left(\frac{3\pi}{35}+0\cdot\frac{2\pi}{5}\right)=\left(\cos\frac{3\pi}{35}+i\sin\frac{3\pi}{35}\right),$$

$$\zeta_1=\cos\left(\frac{3\pi}{35}+1\cdot\frac{2\pi}{5}\right)+i\sin\left(\frac{3\pi}{35}+1\cdot\frac{2\pi}{5}\right)=\left(\cos\frac{17\pi}{35}+i\sin\frac{17\pi}{35}\right),$$

$$\zeta_2=\cos\left(\frac{3\pi}{35}+2\cdot\frac{2\pi}{5}\right)+i\sin\left(\frac{3\pi}{35}+2\cdot\frac{2\pi}{5}\right)=\left(\cos\frac{31\pi}{35}+i\sin\frac{31\pi}{35}\right),$$

$$\zeta_3=\cos\left(\frac{3\pi}{35}+3\cdot\frac{2\pi}{5}\right)+i\sin\left(\frac{3\pi}{35}+3\cdot\frac{2\pi}{5}\right)=\left(\cos\frac{45\pi}{35}+i\sin\frac{45\pi}{35}\right),$$

$$\zeta_4=\cos\left(\frac{3\pi}{35}+4\cdot\frac{2\pi}{5}\right)+i\sin\left(\frac{3\pi}{35}+4\cdot\frac{2\pi}{5}\right)=\left(\cos\frac{59\pi}{35}+i\sin\frac{59\pi}{35}\right).$$

nth roots of unity

The nth roots of 1 have a special importance, in particular in applications. For example in the calculation of Fast Fourier Transform, which is an important tool in image processing and signal processing.

It is also used in many branches of mathematics like number theory and group theory.

Because of its relevance, this small subsection is devoted to this topic.

Definition 1.3.4 — nth root of unity. Let $n\in\mathbb{N}$ be a natural number. A complex number z is called an **nth root of unity** if it fulfils the following equation:

$$z^n=1.$$

It is easy to check that the nth roots of unity have the form given by the theorem below.

Theorem 1.3.6 Let $n\in\mathbb{N}$ be a natural number. A complex number z is an nth root of unity if and only if it can be written in the following form

$$z=\cos\frac{2k\pi}{n}+i\sin\frac{2k\pi}{n},$$

where k is one of the elements of the set $\{0,1,\ldots,n-1\}$.

The conventional notation in the literature for the nth roots of unity is ε_k instead of ζ_k.

Theorem 1.3.7 Let $n\in\mathbb{N}$ be a natural number, then the set of nth roots constitute an Abelian group with respect to multiplication.

■ **Example 1.4** Let us consider the fourth roots of unity

$$\varepsilon_k=\cos\frac{2k\pi}{n}+i\sin\frac{2k\pi}{n},\qquad k=0,1,2,3.$$

It is easy to find their algebraic form too:

$$\varepsilon_0=1,\qquad \varepsilon_1=i,\qquad \varepsilon_2=-1,\qquad \varepsilon_3=-i.$$

Then

$$\varepsilon_1^0=\varepsilon_0,\quad \varepsilon_1^1=\varepsilon_1,\quad \varepsilon_1^2=\varepsilon_2,\quad \varepsilon_1^3=\varepsilon_3,\quad \text{and}\quad \varepsilon_3^0=\varepsilon_0,\quad \varepsilon_3^1=\varepsilon_3,\quad \varepsilon_3^2=\varepsilon_2,\quad \varepsilon_3^3=\varepsilon_1.$$

So, all the fourth roots of unity can be derived either as the powers of ε_1 or ε_3. They cannot be written as the powers of ε_0 or ε_2.

This is always the case if n and k are coprime, that is to say, their greatest common divisor is one. ∎

Definition 1.3.5 — Primitive nth roots of unity. Let $n\in\mathbb{N}$ be a given natural number. An nth root of unity ε_k is called a **primitive nth root of unity** if $gcd(n,k)=1$[a].

[a] gcd denotes the greatest common divisor of n and k. For example, $gcd(12,9)=3$. So, 12 and 9 are not coprimes.

Problem 1.15

a) Find the sixth roots of unity and the primitive sixth roots of unity.

b) What is the sum of the third roots of unity?

Solutions:

a)

$$\varepsilon_0=1,\qquad \varepsilon_1=\cos\tfrac{2\pi}{6}+i\sin\tfrac{2\pi}{6}=\cos\tfrac{\pi}{3}+i\sin\tfrac{\pi}{3}$$

$$\varepsilon_2=\cos\frac{4\pi}{6}+i\sin\frac{4\pi}{6}=\cos\frac{2\pi}{3}+i\sin\frac{2\pi}{3},\qquad \varepsilon_3=\cos\tfrac{6\pi}{6}+i\sin\tfrac{6\pi}{6}=\cos\pi+i\sin\pi$$

$$\varepsilon_4=\cos\frac{8\pi}{6}+i\sin\frac{8\pi}{6}=\cos\frac{4\pi}{3}+i\sin\frac{4\pi}{3},\quad \varepsilon_5=\cos\tfrac{10\pi}{6}+i\sin\tfrac{10\pi}{6}=\cos\tfrac{5\pi}{3}+i\sin\tfrac{5\pi}{3}.$$

The primitive sixth roots of unity are: ε_1, and ε_5.

b) The third roots of unity are $\varepsilon_0,\varepsilon_1$ and ε_2. All can be written as a power of ε_1. Using the formula for the sum of the first n member of a geometric series[10] we have

$$\varepsilon_0+\varepsilon_1+\varepsilon_2=\varepsilon_1^0+\varepsilon_1^1+\varepsilon_1^2=\frac{\varepsilon_1^3}{\varepsilon_1-1}=\frac{1-1}{\varepsilon_1-1}=0.$$

Actually, a pretty similar calculation shows that the sum of the nth roots of unity is zero for an arbitrary $n\geq 2$.

1.3.3 Euler's formula

Reminder from Calculus: It is known from the theory of infinite series, that the complex exponential function, the complex sine function and the complex cosine function can be defined as infinite series in the following way:

$$e^z=\sum_{n=0}^{\infty}\frac{z^n}{n!},\qquad \sin z=\sum_{n=0}^{\infty}(-1)^n\frac{z^{2n+1}}{(2n+1)!},\qquad \cos z=\sum_{n=0}^{\infty}(-1)^n\frac{z^{2n}}{(2n)!}.$$

It is also a well known fact from calculus that absolutely convergent series can be added term by term. This will be used in the proof of the following nice theorem of Euler.

Leonhard Euler
(1707-1783)

Theorem 1.3.8 — Euler's identity. Let φ be an angle (a real number), then

$$e^{i\varphi} = \cos\varphi + i\sin\varphi.$$

Proof. Here we use the periodicity of the powers of the imaginary unit i

$$\begin{array}{cccccccc} i^0 = 1, & i^1 = i, & i^2 = -1, & i^3 = -i, \\ i^4 = 1, & i^5 = i, & i^6 = -1, & i^7 = -i, \\ i^8 = 1, & i^9 = i, & i^{10} = -1, & i^{11} = -i, \\ \dots & \dots & \dots & \dots\dots & \dots & \dots & \dots \end{array}$$

$$\begin{aligned} e^{i\varphi} = \sum_{n=0}^{\infty} \frac{(i\varphi)^n}{n!} = i^0\frac{\varphi^0}{0!} + i^1\frac{\varphi^1}{1!} + i^2\frac{\varphi^2}{2!} + i^3\frac{\varphi^3}{3!} + i^4\frac{\varphi^4}{4!} + i^5\frac{\varphi^5}{5!} + \cdots = \\ \left((-1)^0\cdot\frac{\varphi^0}{0!} + (-1)^1\cdot\frac{\varphi^2}{2!} + (-1)^2\cdot\frac{\varphi^4}{4!} + (-1)^3\cdot\frac{\varphi^6}{6!} + \cdots\right) + \\ i\left((-1)^0\cdot\frac{\varphi^1}{1!} + (-1)^1\cdot\frac{\varphi^3}{3!} + (-1)^2\cdot\frac{\varphi^5}{5!} + (-1)^3\cdot\frac{\varphi^7}{7!} + \cdots\right) = \\ \sum_{n=0}^{\infty}(-1)^n\frac{\varphi^{2n}}{(2n)!} + i\sum_{n=0}^{\infty}(-1)^n\frac{z^{2n+1}}{(2n+1)!} = \cos\varphi + i\sin\varphi. \end{aligned}$$

■

1.3.4 Calculation with complex numbers in exponential form

With the help of Euler's identity one can write a complex number in a very compact form.

$$z = |z|\underbrace{(\cos\varphi_z + i\sin\varphi_z)}_{e^{i\varphi_z}} = |z|e^{i\varphi_z}.$$

Definition 1.3.6 — Exponential form of complex numbers. Let z be a non-zero complex number with angle φ_z, then the form

$$z = |z|e^{i\varphi_z}$$

is said to be the **exponential form of** z.

All the calculation rules are easy consequences of this definition and the calculation rules of complex number in polar form.

■ **Example 1.5** Let $z = 2e^{i\pi}$ and $w = 3e^{i\frac{\pi}{3}}$, then

$$zw = 2e^{i\pi}\cdot 3e^{i\frac{\pi}{3}} = 6e^{i(\pi+\frac{\pi}{3})} = 6e^{i(\frac{4\pi}{3})}, \qquad \frac{z}{w} = \frac{2}{3}e^{i(\pi-\frac{\pi}{3})} = \frac{2}{3}e^{i(\frac{2\pi}{3})},$$

and

$$z^{10} = 2^{10}e^{i10\pi} = 1024e^{i\cdot 0} = 1024.$$

■

1.4 Exercises

Exercise 1.1 Find the algebraic form of the following expressions!

a) $(2-i)(3-i)$ b) $(3+7i)(2+i)$ c) $(i-2)(i+2)$ d) $(2+2i)(3+i)i$

e) $\frac{3-i}{1+i}$ f) $\frac{-1-i}{1-4i}$ g) $\frac{3+i}{(2+3i)(1+i)}$

■

Exercise 1.2 Find x and y which fulfil the equation

$$(1+2i)x+(1-3i)y=2+i.$$

■

Exercise 1.3 Simplify the following expressions!

a) $(1+i)^3-(1+i)^3$ b) $\frac{i+3}{2i-1}+\frac{5+3i}{3-i}$ c) i^{1023}

■

Exercise 1.4 Find the length, the angle and the polar form of the following complex numbers!

a) $\sqrt{3}-3i$ b) 5 c) -3 d) $-4i$

e) $-1-\sqrt{3}$ f) $6+6i$ g) $9i$ h) $i-1$

■

Exercise 1.5 Give z, w, zw and $\frac{z}{w}$ in polar form!

a) $z=10i \quad , w=1+\sqrt{3}i.$ b) $z=-\sqrt{32}+2i \quad , w=1-i.$

■

Exercise 1.6 Find the angle of $7-24i$ rounding it to four decimal places! ■

Exercise 1.7 We know the length 10 and the angle $\frac{\pi}{4}$ of z. Find the value of $\mathfrak{Re}(z)$ and $\mathfrak{Im}(z)$!
■

Exercise 1.8 Simplify the following expressions!

a)

$$\left(\cos\frac{5\pi}{13}+i\sin\frac{5\pi}{13}\right)\left(\cos\frac{7\pi}{11}+i\sin\frac{7\pi}{11}\right)$$

b)

$$\left(\cos\frac{2\pi}{5}+i\sin\frac{2\pi}{5}\right)^9$$

c)

$$\left(\cos\frac{2\pi}{3}+i\sin\frac{2\pi}{3}\right)\left(\cos\frac{7\pi}{12}+i\sin\frac{7\pi}{12}\right)$$

d)

$$\left(\cos\frac{5\pi}{22}+i\sin\frac{5\pi}{22}\right)^{11}$$

e)

$$\frac{\cos\frac{5\pi}{3}+i\sin\frac{5\pi}{3}}{\cos\frac{6\pi}{5}+i\sin\frac{6\pi}{5}}$$

f)

$$\frac{\cos\frac{2\pi}{9}+i\sin\frac{2\pi}{9}}{\cos\frac{11\pi}{7}+i\sin\frac{11\pi}{7}}$$

g)

$$\left(\cos\frac{5\pi}{22}+i\sin\frac{5\pi}{22}\right)^{-3}$$

h)

$$\left(\cos\frac{\pi}{2}+i\sin\frac{\pi}{2}\right)^{-4}$$

■

Exercise 1.9 Solve the following equations!

a) $z^2-2z+2=0$

b) $z^2-6z+10=0$

c) $4z^2-4z+5=0$

d) $2z^2+3z+2=0$

e) $z^2-z+1=0$

f) $z^2-4z+13=0$

g) $z^2-(1+2i)z+i-1=0$

h) $z^2-(8-3i)z+11-27i=0$

■

Exercise 1.10 Find the fourth roots of $z=2-\sqrt{12}i$! ■

Exercise 1.11 Find the fifth roots of $z=1-i$! ■

Exercise 1.12 Find the sixth roots of $z=i$! ■

Exercise 1.13 Find the sixth roots of unity and the primitive sixth roots of unity! ■

Exercise 1.14 Find the twenty-fourth roots of unity and the primitive twenty-fourth roots of unity! ■

Exercise 1.15 Give the seventh roots of unity in exponential form! ■

2. Linear algebra

2.1 Vectors, Vector spaces

Mathematics is frequently used in other sciences (like physics, engineering and so on) to model natural events. In this events quantities are specified not only by numerical values and unit of measurement but also by their direction. There are several examples for that such as a movement of a car, direction of any kind of force, direction of the wind, and so on.

In applied sciences, like physics and engineering, the quantities which must be specified by magnitude and direction are called *vector quantities* or more simply *vectors*.

Let us consider a shift from P_1 to P_2, where P_1 and P_2 are points in the plane. Then the magnitude (numerical value) is the length of the section joining P_1 with P_2 and the arrow shows the direction of the shift. See the figure below. In the following two vectors are considered to be equal if they have the same magnitude (length) and direction. As a consequence, it can be assumed without losses that the starting point is the origin.

2.1.1 Operation with vectors

The space $\mathbb{R}^2$

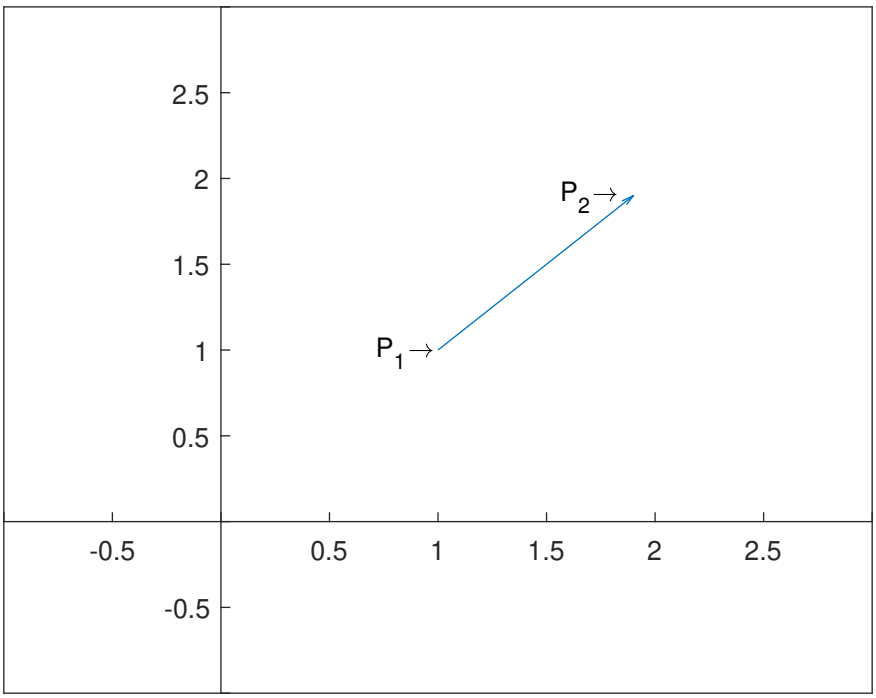
If we fix a point on the plane, which is called the *origin*, then we can characterize the position of all the points on the plane by two numbers. More precisely, we can characterize the position of an arbitrary point x on the plane by an **ordered pair** (x_1, x_2) or (x, y), where the numbers denotes the *coordinates of* x. The vector with the starting point $(0,0)$ (origin) and with the endpoint x is called the **position vector of** x.

Definition 2.1.1 The collection of all ordered pairs of real numbers is denoted by $\mathbb{R}^2$. These ordered pairs are also called **real two-tuple vector**s or **two dimensional real vector**s.

The first and the second numbers of a two dimensional vector $x \in \mathbb{R}^2$ said to be the **first and the second coordinates of** x respectively.

We use the following notation:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{\, x \mid x = (x_1, x_2),\ x_1, x_2 \in \mathbb{R} \,\}.$$



Addition of vectors in R^2

If we have two two- dimensional vectors x and y, their sum geometrically is given by the directed diagonal of the parallelogram with sides x and y. Algebraically this means coordinatewise addition.

Definition 2.1.2 Let $x, y \in \mathbb{R}^2$ be arbitrarily given with coordinates $x = (x_1, x_2)$, and $y = (y_1, y_2)$. Then their sum is defined by the formula

$$x + y = (x_1 + y_1, x_2 + y_2).$$

Problem 2.1 Find the sum $x + y$ where

a) $x = (-10, 1)$, and $y = (1, 1)$,

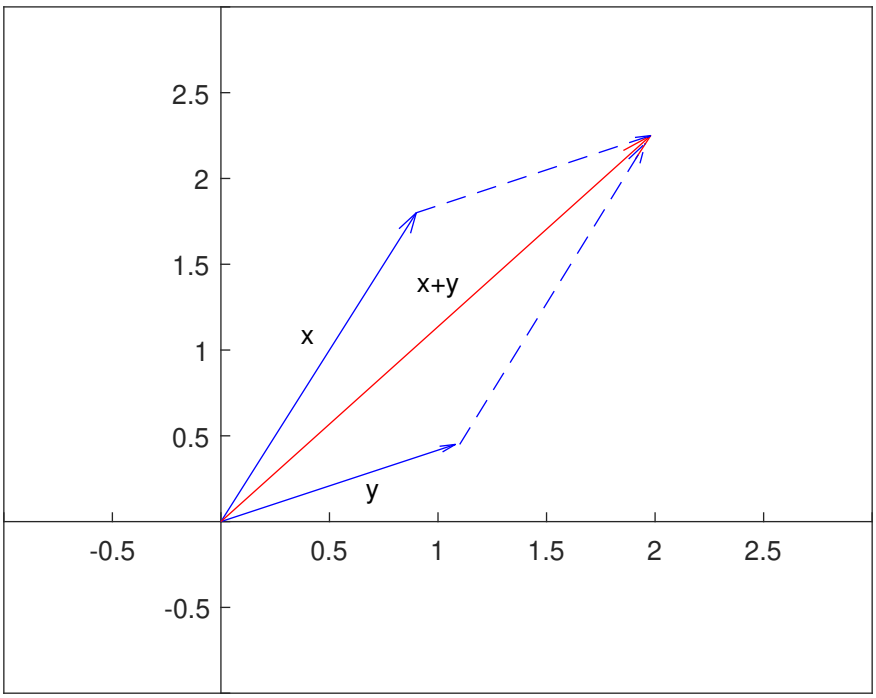
b) $x = (-2, -3.8)$, and $y = (-2.1, 1)$,

c) $x = (-1001, 1001)$, and $y = (1001, -1001)$,

d) $x = (1, 2.342)$, and $y = (-11.2134, 12.3)$.

Solutions:

- $$x + y = (-10, 1) + (1, 1) = (-10 + 1, 1 + 1) = (-9, 2)$$
- $$x + y = (-2, -3.8) + (-2.1, 1) = (-4.1, -2.8)$$
- $$x + y = (-1001, 1001) + (1001, -1001) = (0, 0)$$



Addition of vectors in $\mathbb{R}^2$

- $$x+y=(1,2.342)+(-11.2134,12.3)=(-10.2134,14.642)$$

Because the addition happens coordinatewise, all its properties are inherited from addition of real numbers.

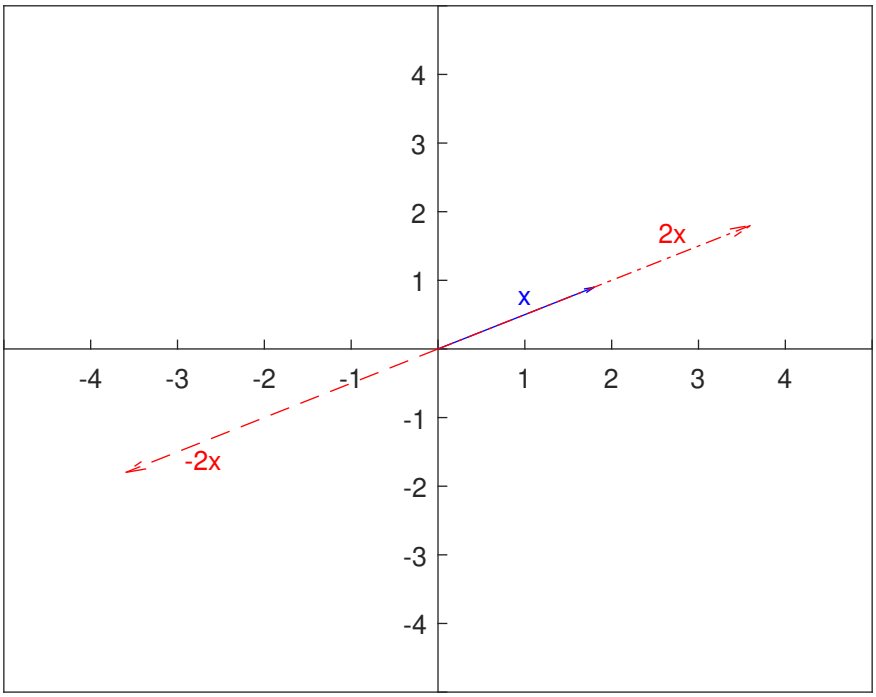
Proposition 2.1.1 — Properties of addition in $\mathbb{R}^2$. The set $\mathbb{R}^2$ with the addition of its elements constitutes an abelian group[1], that is to say

- addition is a binary operation in $\mathbb{R}^2$ (the sum of two two-dimensional vector is a two dimensional vector),
- addition is associative,
$$(x+y)+z=x+(y+z), \qquad x,y,z\in\mathbb{R}^2,$$
- there is an additive unit[2],
$$x+0=0+x=x, \qquad x\in\mathbb{R}^2, \text{ and } 0=(0,0),$$
- there exists an additive inverse of all elements of $\mathbb{R}^2$, that is, for all $x\in\mathbb{R}^2$ there is a vector denoted by $-x\in\mathbb{R}^2$ such that
$$x+(-x)=(-x)+x=0,$$
- addition is commutative
$$x+y=y+x, \qquad x,y\in\mathbb{R}^2.$$

[1] This is the same as in the case of the set of real numbers or the set of complex numbers with their addition.
[2] If there is no ambiguity, we denote by 0 the zero number and the zero vector too.

Multiplication by a scalar in R^2

Multiplication by a scalar also runs coordinatewise. Geometrically the result will be on the same line determined by the original vector, its length depends on the factor, and its direction depends on the sign of the factor.



Multiplication of vectors by scalars in $\mathbb{R}^2$

Definition 2.1.3 Let $x = (x_1, x_2)$ be a vector in $\mathbb{R}^2$ and $\alpha \in \mathbb{R}$ be a scalar, then

$$\alpha x = (\alpha x_1, \alpha x_2).$$

Problem 2.2 Find λx where

a) $\alpha = -2$, and $x = (1,3)$,

b) $\alpha = 2.3$, and $x = (-1, 0.2)$,

c) $\alpha = \pi$, and $x = (\pi, \frac{1}{2})$,

d) $\alpha = 20$, and $x = (0, 0.01)$.

Solutions:

- $$\alpha x = -2(1,3) = (-2,-6)$$
- $$\alpha x = 2.3(-1, 0.2) = (-2.3, 0.46)$$
- $$\alpha x = \pi(\pi, \tfrac{1}{2}) = (\pi^2, \tfrac{\pi}{2})$$

•

$$\alpha x = 20(0, 0.01) = (0, 0.2)$$

Proposition 2.1.2 — Properties of multiplication by a scalar in $\mathbb{R}^2$. Let $x, y \in \mathbb{R}^2$ be vectors and $\alpha, \beta \in \mathbb{R}$ be scalars, then

- $0 \cdot x = 0$, and $1 \cdot x = x$,
- $\alpha(\beta x) = (\alpha\beta)x$,
- $(\alpha+\beta)x = \alpha x + \beta x$, and $\alpha(x+y) = \alpha x + \alpha y$.

Proof. All the properties are inherited again from the properties of operations with real numbers. For example,

$$\begin{aligned}(\alpha+\beta)x = (\alpha+\beta)(x_1,x_2) = ((\alpha+\beta)x_1, (\alpha+\beta)x_2) = (\alpha x_1 + \beta x_1, \alpha x_2 + \beta x_2) = \\ = (\alpha x_1, \alpha x_2) + (\beta x_1, \beta x_2) = \alpha(x_1,x_2) + \beta(x_1,x_2) = \alpha x + \beta x.\end{aligned}$$

■

The space $\mathbb{R}^n$

In a pretty similar way, like in the case of $\mathbb{R}^2$, one can define an ordered n-tuple like a two-tuple.

$$\underbrace{(2,3,-1)}_{\text{Example for a three-tuple.}} \qquad \underbrace{(-1,1,3,4,5,6)}_{\text{Example for a six-tuple.}}$$

■ **Example 2.1** Let us collect the weight of a person on the first days of every months in the previous year in kg.

$$w = (w_1, w_2, \ldots, w_{12}) = (78, 77.4, 81, 82.3, 80.1, 79.3, 79, 80, 80, 80.3, 77.9, 78.9).$$

The result will be a 12-tuple, and the ith member of this vector denotes the weight of the person on the first day of the ith month. For example $w_{11} = 77.9$, this means that the weight of the person was 77.9kg on the first day of November last year. ■

Definition 2.1.4 Let $n \in \mathbb{N}$ be an arbitrary natural number, then $x = (x_1, x_2, \ldots, x_n)$ is called a **real n-tuple** or an n **dimensional real vector**, where $x_i \in \mathbb{R}$, $i = 1, \ldots, n$ denotes the **ith coordinate of** x.

The set of all real n dimensional vectors is denoted by $\mathbb{R}^n$,

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{n\text{-times}} = \{\, x \mid x = (x_1, x_2, \ldots, x_n),\ x_i \in \mathbb{R},\ i = 1, \ldots, n \,\}.$$

One can construct a structure on the set $\mathbb{R}^n$ like in the case of $\mathbb{R}^2$. Definition of addition of n dimensional real vectors or multiplication by a real scalar can be defined componentwise.

Definition 2.1.5 Let $x, y \in \mathbb{R}^n$ be arbitrarily given with coordinates $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, and $\lambda \in \mathbb{R}$ be a scalar. Then

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n), \quad \text{and} \quad \lambda x = (\lambda x_1, \lambda x_2, \ldots \lambda x_n).$$

■ **Example 2.2** Let us assume that a two bakeries produce a certain amount of bread in kg, which are presented by the following 7 dimensional vectors.

$$x = (120, 124, 200, 150, 250, 500, 0), \qquad y = (100, 90, 130, 100, 200, 390, 0),$$

where x represents the produced amount of the first bakery and y represents the produced amount of the second bakery. The ith coordinates correspond to the ith day. For example the first bakery bakes 200kgs bread on Wednesday. What is the daily sum of the produced bread of the two bakeries? This is given by the sum of the production vectors.

$$x+y = (120,124,200,150,250,500,0)+(100,90,130,100,200,390,0) = \\ = (220,214,330,250,450,890,0).$$

Assume that the owner of a new bakery has the same parameters as the bakery with product vector x. What will be the total daily outputs of a week?

$$2x+y = 2(120,124,200,150,250,500,0)+(100,90,130,100,200,390,0) = \\ = (240,248,400,300,500,1000,0)+(100,90,130,100,200,390,0) = \\ = (340,338,530,400,700,1390,0)$$

■

Important remark: It is easy to verify the same properties of vector operations in $\mathbb{R}^n$ as in $\mathbb{R}^2$.

2.1.2 Vector spaces, subspaces

The properties of addition and multiplication by a scalar in $\mathbb{R}^2$ or in $\mathbb{R}^n$ come from the properties of addition and multiplication of real numbers. These properties can also be considered as axioms.[3] This way of thinking leads us to the concept of vector space.

Definition 2.1.6 — Vector space. Let V be a non-empty set. The elements of V are called **vectors**. Assume that a binary map, denoted by $+$ **addition of vectors**, is given on V with the following properties:

- the sum of two elements of V is in V

  $$v+w \in V, \qquad v,w \in V,$$

  that is, $+\colon V \times V \to V$ is a **binary operation on** V;
- addition is **commutative**,

  $$v+w = w+v, \quad v,w \in V;$$

- addition is **associative**,

  $$(u+v)+w = u+(v+w), \qquad u,v,w \in V;$$

- there exists in V a unique vector 0 (called the **zero vector** or the **additive unit**) such that

  $$v+0 = v, \quad v \in V;$$

- to every vector $v \in V$ there corresponds a unique vector $-v$ (called the **additive inverse of** v) such that

  $$v+(-v) = 0.$$

Moreover, assume that there is a binary map, denoted by $\cdot$ **multiplication by a scalar**, is given such that either $\cdot\colon \mathbb{R} \times V \to V$ (multiplication by a real scalar) or $\cdot\colon \mathbb{C} \times V \to V$ (multiplication by a complex scalar)[a] with the following properties:

[3] Collections of assumptions or statements that are taken to be true.

- multiplication by a scalar is **associative**, that is

$$(\alpha\beta)v = \alpha(\beta v), \qquad \alpha,\beta \in \mathbb{R} \text{ or } \mathbb{C},\ v \in V;$$

- for all $v \in V$

$$1v = v;$$

- multiplication by scalars is **distributive** with respect to vector addition,

$$\alpha(v+w) = \alpha v + \alpha w, \qquad \alpha \in \mathbb{R} \text{ or } \mathbb{C},\ v,w \in V;$$

- multiplication by vectors is **distributive** with respect to scalar addition,

$$(\alpha+\beta)v = \alpha v + \beta v, \qquad \alpha,\beta \in \mathbb{R} \text{ or } \mathbb{C},\ v \in V.$$

If V fulfils the above assumptions, then it is called a **vector space over** $\mathbb{R}$ **or** $\mathbb{C}$ depending on the scalar set.

[a] In practice, these are the most frequent cases, but there are other possibilities too, e.g. multiplication by rational numbers, and so on.

Examples for vector spaces

The spaces $\mathbb{R}^2$ and $\mathbb{R}^3$, the real plane and the real three dimensional space, are the best-known examples for vector spaces. These are especially important in applications in physics and in engineering. However, there are other well-known examples, which also have a great importance in practice.

We have dealt with the space $\mathbb{R}^n$. In a quite similar way one can define the space of complex n-tuples $\mathbb{C}^n$. Here the coordinates of the vectors are complex numbers, and the scalar set is usually the set of complex numbers, but it can also be the set of real numbers. In notation

$$C^n = \{\ v \mid v = (v_1,\ldots,v_n),\ v_i \in \mathbb{C}\ \}.$$

■ **Example 2.3** Let $v,w \in \mathbb{C}^3$ $v = (1+i, 1-2i, -1)$, $w = (-10+3i, 8+i, i)$, and $\alpha \in \mathbb{C}$, $\alpha = 3+i$, then

$$v+w = (1+i,1-2i,-1)+(-10+3i,8+i,i) = (-9+4i,9-i,-1+i),$$

and

$$\alpha v = (3+i)(1+i,1-2i,-1) = ((3+i)(1+i),(3+i)(1-2i),(3+i)(-1)) = (2+4i,5-5i,-3-i).$$

■

Besides the previously mentioned "scalar type" vector spaces, there are "function type" vector spaces which also have a great use in applications. The most important vector spaces constituted by functions contain certain polynomials.

Definition 2.1.7 — Real polynomials. Let $n \in \mathbb{N}$, the function $p\colon \mathbb{R} \to \mathbb{R}$ is called a **polynomial of degree** n if it can be written in the form

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_n,\ldots,a_1,a_0 \in \mathbb{R}$ are called the **coefficients of** p. The degree of p is denoted by $\deg p$.

Let $n \in \mathbb{N}$ be given. The set of polynomials with degree at most n is denoted by $\mathscr{P}_n$, that is

$$\mathscr{P}_n = \{p \text{ is a polynomial} \mid \deg p \le n\}.$$

The space of all real polynomials is denoted by $\mathscr{P}$, that is

$$\mathscr{P} = \bigcup_{n\in\mathbb{N}} \mathscr{P}_n.$$

In $\mathscr{P}_n$ and in $\mathscr{P}$ the vectors are polynomials. The usual vector space operations happen pointwise, that is to say, let $p,q \in \mathscr{P}_n$ and $\alpha \in \mathbb{R}$, then

$$(p+q)(x) = p(x)+q(x), \qquad (\alpha p)(x) = \alpha p(x).$$

It is easy to check that with these operations both $\mathscr{P}_n$ and $\mathscr{P}$ are real vector spaces.

It is worthy to note that $\mathscr{P}_n \subset \mathscr{P}$ and that they are vector spaces with respect to the same operations. This trail of thoughts naturally leads us to the concept of subspaces.

Definition 2.1.8 — Subspace. Assume that V is a vector space over the real numbers or over the complex numbers, and let $S \subset V$. If S is a vector space with respect to the same operations over the same scalar set, then it is called a **subspace of** V.

It is relatively easy to check whether a subset of a vector space is really a subspace. According to the following proposition it is not necessary to check all the axioms. This ensures a comfortable and practical method to detect a subspace in a vector space.

Theorem 2.1.3 — Subspace criteria. Let V be a vector space and $S \subset V$ be a non-empty subset of V. Then S is a subspace if and only if for all $v,w \in S$ and for all α scalar we have

$$v - w \in S, \qquad \text{and} \qquad \alpha v \in S.$$

Because of its special importance we prove this theorem.

Proof. Because of the assumption S is closed with respect to addition and multiplication by a scalar.

S is non-empty, so there is $v \in S$. With $0 = v - v \in S$ we have that the additive unit (zero vector) is in S. Using this we get $-v = 0 - v \in S$, so the inverse of v $-v$ is also in S. The properties of the addition are inherited from V. These entails that S is an abelian group with the same addition as V.

The properties of multiplication by a scalar in S are also inherited from the multiplication by a scalar in V.

The reverse statement is trivial. ■

Examples for subspaces

As it has been mentioned $\mathscr{P}_n$ is a subspace of $\mathscr{P}$.

In general, if V is and arbitrary vector space, then V is a subspace of itself, and the set, which contains only the additive unit $\{0\}$ constitutes also a subspace of V. These two are said to be the **trivial subspaces of** V.

In $\mathbb{R}^2$ if we put aside the trivial subspaces, (geometrically) the typical subspaces of $\mathbb{R}^2$ are lines through the origin. In other words, these can be expressed as all the scalar multiplies of a fixed, non-zero vector of $\mathbb{R}^2$.

■ **Example 2.4** The sets

a)

$$S_1 = \{\, x \in \mathbb{R}^2 \mid x = (x_1, x_2) \text{ and } x_2 = 0 \,\}$$

b)

$$S_2 = \{\, x \in \mathbb{R}^2 \mid x = (x_1, x_2) \text{ and } x_1 = 0 \,\}$$

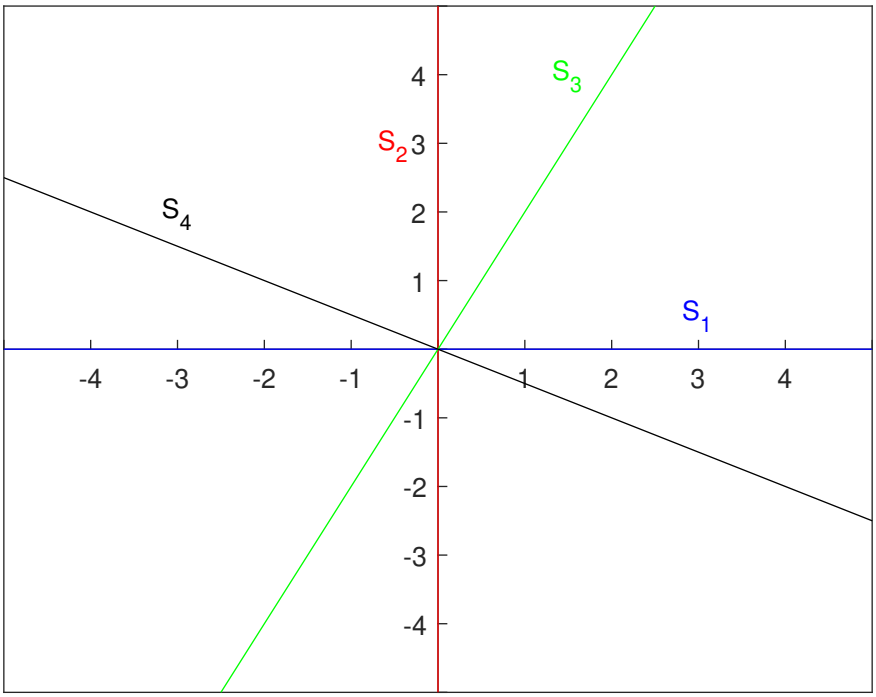
c)

$$S_3 = \{\, x \in \mathbb{R}^2 \mid x = (x_1, x_2) \text{ and } x_1 = 2x_2 \,\}$$

d)

$$S_4 = \{\, x \in \mathbb{R}^2 \mid x = (x_1, x_2) \text{ and } 2x_1 = -x_2 \,\}$$

are subspaces of $\mathbb{R}^2$, and geometrically they represent the x-axis, the y-axis,the line through the origin with slope 2, and the line through the origin with slope $-\frac{1}{2}$ respectively.



Examples for subspaces in $\mathbb{R}^2$

■

Let $n \in \mathbb{N}$ be given, then $\mathbb{R}^n$ can be embedded into $\mathbb{R}^{n+1}$ as a subspace of $\mathbb{R}^{n+1}$ in the following way:

$$S = \{\, x \in \mathbb{R}^{n+1} \mid x = (x_1, \ldots, x_n, x_{n+1}), \text{ and } x_{n+1} = 0 \,\},$$

then S can be identified with $\mathbb{R}^n$.

Further examples for vector spaces and for subspaces

One can define the space of complex polynomials like real polynomials. The set of complex polynomials with degree at most n is a subspace of the vector space of all complex polynomials.

Let us fix an interval $[a,b]$, where $a < b$. The set of all real valued functions defined on $[a,b]$ constitutes a vector space over the real numbers with the pointwise operations. Important subspaces are the space of continuous functions on $[a,b]$, and the space of k-times continuously differentiable functions on $[a,b]$, where k is a given natural number.

2.1.3 Linear combination of vectors

As we have seen, the non-trivial subspaces of $\mathbb{R}^2$ can be written as a scalar multiply of a non-zero vector. If we think about this situation a little bit we perceive that this is a tremendous reduction in the following sense. A line in $\mathbb{R}^2$ contains infinitely many vectors, but all of these vectors can be written as αx, where $\alpha \in \mathbb{R}$ and x is an arbitrarily fixed non-zero vector of the line in question. In other words, all the lines of $\mathbb{R}^2$ trough the origin can be generated by a single vector.

In $\mathbb{R}^n$, if $n > 2$, one can choose two non-zero vectors, x and y, such that there is no α with the property $\alpha x = y$. It is possible to talk about now the expression $\alpha x + \beta y$, where $\alpha, \beta \in \mathbb{R}$. An expression like this is called a linear combination of the vectors x and y. Collecting all such vectors in a set we get a copy of $\mathbb{R}^2$ in $\mathbb{R}^n$, geometrically this is a plane in $\mathbb{R}^n$ generated by the vectors x and y.

It is reasonable to define a linear combination of finitely many vectors in a general vector space V.

Definition 2.1.9 — Linear combination. Let V be a (real or complex) vector space, and $v_1, \dots, v_m \in V$ be given vectors and $\alpha_1, \dots, \alpha_m$ be given (real or complex) scalars, then the expression

$$\alpha_1 v_1 + \cdots \alpha_m v_m$$

is called a **linear combination of the vectors** $v_1, \dots, v_m \in V$ **with coefficients** $\alpha_1, \dots, \alpha_m$.

The result of a linear combination like this will be a vector of V- because of the axioms of vector spaces- so we also say that v **can be combined linearly from the vectors** $v_1, \dots, v_m \in V$ if there are scalars $\alpha_1, \dots, \alpha_m$ such that

$$v = \alpha_1 v_1 + \cdots \alpha_m v_m.$$

■ **Example 2.5** Let us consider $\mathscr{P}_n$, the space of real polynomials with degree at most n. Then all vectors of $\mathscr{P}_n$ can be written as a linear combination of the vectors $1, x, x^2, \dots, x^n$, because if $p \in \mathscr{P}_n$, then there are constants $a_0, a_1, \dots, a_n$ such that

$$p(x) = a_n x^n + \cdots + a_1 x + a_0 \cdot 1 = a_n x^n + \cdots + a_1 x + a_0.$$

So $\mathscr{P}_n$ is the collection of all the linear combinations of the vectors $1, x, x^2, \dots, x^n$. We say that the linear hull of the vectors $1, x, x^2, \dots, x^n$ equals to $\mathscr{P}_n$. ■

Definition 2.1.10 — Linear hull. Let V be a (real or complex) vector space, and $v_1, \dots, v_m \in V$ be given vectors. Then the set

$$\operatorname{span}\{v_1, \dots, v_m\} = \{\, v \in V \mid v = \alpha_1 v_1 + \cdots \alpha_m v_m \,\}$$

is said to be the **linear hull of the vectors** $v_1, \dots, v_m$ or the **linear span of the vectors** $v_1, \dots, v_m$.

If $S \subset V$, then the **linear hull of** S contains all the finite linear combinations of vectors from S. That is

$$\operatorname{span} S = \{\, v \in V \mid v = \alpha_1 v_1 + \cdots \alpha_m v_m, \text{ and } v_i \in S,\ i = 1, \dots, m \,\}.$$

An arbitrary set of vectors not necessarily has a structure. However, their linear hull always has a structure.

Proposition 2.1.4 The linear span of an arbitrary subset of a vector space is a subspace.

Proof. Hint: Immediately follows from the subspace criteria. ■

Problem 2.3 Find the linear hull (the generated subspace) of the following vectors in the corresponding vector space V.

- $(0,1) \in \mathbb{R}^2$
- $(0,0,1),(0,1,0) \in \mathbb{R}^3$
- $(1,1) \in \mathbb{R}^2$
- $1+t, t, 1+t+t^2 \in \mathscr{P}$

Solutions:

- $$\operatorname{span}\{(0,1)\} = \{\, x \in \mathbb{R}^2 \mid \alpha(0,1) = (0,\alpha),\ \alpha \in \mathbb{R} \,\}.$$
  Geometrically this is the y-axis on the plane.
- $$\operatorname{span}\{(0,0,1),(0,1,0)\} = \{\, x \in \mathbb{R}^3 \mid \alpha(0,0,1)+\beta(0,1,0) = (0,\beta,\alpha),\ \alpha,\beta \in \mathbb{R} \,\}.$$
  Geometrically this is the $y-z$-plane in the three dimensional space.
- $$\operatorname{span}\{(1,1)\} = \{\, x \in \mathbb{R}^2 \mid \alpha(1,1) = (\alpha,\alpha),\ \alpha \in \mathbb{R} \,\}.$$
  Geometrically this is the line on the two dimensional plane with the equation $y = x$.
- $$\operatorname{span}\{1+t, t, 1+t+t^2\} =$$
  $$= \{\, p \in \mathscr{P} \mid p(t) = \alpha(1+t)+\beta t+\gamma(1+t+t^2) = \alpha+(\alpha+\beta+\gamma)t+\gamma t^2,\ \alpha,\beta,\gamma \in \mathbb{R} \,\}.$$
  This is the space of polynomials with degree at most two $\mathscr{P}_2$.

2.1.4 Linear dependence, linear independence, basis, dimension

As we saw, the span of finitely many vectors can be very big. For example the space of real polynomials with degree at most n equals to the linear span of finitely many vectors. This is a huge reduction since $\mathscr{P}_n$ contains infinitely many elements.

An important question is, whether it is possible to get such a big reduction in all vector spaces or not. If so, are there finite sets of vectors which have minimal number of elements in the sense that if we cancel one element the resulted span will be contained strictly by the span of the original set? Is there any kind of characterization of sets like these?

These questions lead us to the sequence of closely correlated concepts designated in the title of this subsection.

At first, we define the linear dependence and independence of vectors. If a vector space is spanned by a finite system of vectors, then there are minimal systems. If we omit one vector from a minimal system, then the span of the remaining system will be smaller than the original one. These systems are constituted of linearly independent vectors and they are called bases. These sets are the "skeletons" of the vector space. If we know one "skeleton" we know the whole space.

These finite, minimal systems contain the same number of vectors, and this number will be the dimension of the space.[4]

Let us assume that we have a system of vectors. If we can pick up one which can be written as a linear combination of finitely many vectors from the remaining part of the system, then we can cancel this vector from the system without the change of the span of the system. So, this vector is superfluous from the point of view of the span of the system.

Another approach for the previously discussed situation is, that the zero vector can be combined linearly from the system in a non-trivial way (not only with zero coefficients).

[4] There are infinite dimensional spaces too, when there is no finite system which spans the whole space. An example for infinite dimensional spaces is the space of all real polynomials $\mathscr{P}$.

Definition 2.1.11 — Linear dependence, linear independence. Let V be a vector space, $v_1,\dots,v_m \in V$ be an arbitrary finite system of vectors. If there are $\alpha_1,\dots,\alpha_m$ scalars, at least one is different from zero such that

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0,$$

then the vectors $v_1,\dots,v_m$ are said to be **linearly dependent**.

The vectors $v_1,\dots,v_m$ are said to be **linearly independent** if they are not linearly dependent.

■ **Example 2.6** If a finite system of vectors contains the zero vector, then the system is linearly dependent. Indeed, choose any non-zero coefficient for the zero vector, and zero coefficient for the other vectors. This construction results a non-trivial linear combination of the zero vector. ■

■ **Example 2.7** Let $x,y \in \mathbb{R}^2$ be non-zero vectors. Geometrically their linear dependence means that they are on the same lines. In this case there is a constant $\alpha \neq 0$ such that $x = \alpha y$. Otherwise they are linearly independent. ■

Problem 2.4 Find the linearly dependent and independent pairs x,y where

a) $x=(1,1),\ y=(-1,-1),\ x,y\in\mathbb{R}^2$,
b) $x=(1,1),\ y=(-1,1),\ x,y\in\mathbb{R}^2$,
c) $p_1(t)=t,\ p_2(t)=2t,\ p_1,p_2\in\mathscr{P}$,
d) $p_1(t)=1+t,\ p_2(t)=1-t,\ p_1,p_2\in\mathscr{P}$.

Solutions:

- $$\alpha(1,1)+\beta(-1,-1)=(0,0) \Longleftrightarrow (\alpha-\beta,\alpha-\beta)=(0,0) \Longleftrightarrow \alpha=\beta, \quad \alpha,\beta\in\mathbb{R}.$$
  For example, let $\alpha=\beta=2$, then
  $$2(1,1)+2(-1,-1)=(0,0).$$
  That is, the zero vector can be combined linearly in a non-trivial way. So, the system $(1,1),\ (-1,-1)$ is linearly dependent.
- $$\alpha(1,1)+\beta(-1,1)=(0,0) \Longleftrightarrow (\alpha-\beta,\alpha+\beta)=(0,0) \Longleftrightarrow \alpha-\beta=0, \text{ and } \alpha+\beta=0.$$
  This is a simple system of linear equations. Its solution is $\alpha=\beta=0$. That is, the zero vector can be combined linearly only in the trivial way. So, the system $(1,1),\ (-1,1)$ is linearly independent.
- $$\alpha p_1(t)+\beta p_2(t)=0 \Longleftrightarrow \alpha t+\beta 2t=(\alpha+2\beta)t=0 \Longleftrightarrow \alpha+2\beta=0.$$
  For example with the choice $\alpha=2$ and $\beta=-1$ we have
  $$2t+(-1)2t=0.$$
  That is, the zero polynomial can be combined linearly in a non-trivial way. So, the system $p_1(t)=t,\ p_2(t)=2t$ is linearly dependent.
- $$\alpha p_1(t)+\beta p_2(t)=0 \Longleftrightarrow \alpha(1+t)+\beta(1-t)=\alpha+\beta+(\alpha-\beta)t=0.$$
  That is
  $$\alpha+\beta=0, \quad \text{and} \quad \alpha-\beta=0.$$
  This linear system has only the trivial solution $\alpha=\beta=0$, that is, the system $p_1(t)=1+t$, $p_2(t)=1-t$ is linearly independent.

A linear combination $\alpha_1 v_1 + \cdots + \alpha_m v_m$, where not all the coefficients are zero is called a **non-trivial linear combination**. Otherwise it is called a **trivial linear combination**. This opens the door to give an alternative definition of linear dependence and independence.

Proposition 2.1.5 Let V be a vector space. The finite system of vectors $v_1, \ldots, v_m \in V$ is linearly dependent if and only if the zero vector can be written as a non-trivial linear combination of $v_1, \ldots, v_m$.[5]

The finite system of vectors $v_1, \ldots, v_m \in V$ is linearly independent if and only if the zero vector can be written only as the trivial linear combination of $v_1, \ldots, v_m$.

We can go forward not towards the very important concept of basis.

Definition 2.1.12 — Basis. A vector space V is called **finitely generated** if there are vectors $v_1, \ldots, v_m \in V$ such that

$$\operatorname{span}\{v_1, \ldots, v_m\} = V.$$

In this case $v_1, \ldots, v_m$ is said to be a **generating system of** V.

A linearly independent generating system $v_1, \ldots, v_m \in V$ of a vector space V is called a **basis of** V.

Immediately follows from the definition, that all vectors of V can be written as linear combinations of the basis vectors. The advantage of a basis over a generating system is that all the vectors of V can be written only one way using the vectors of a basis.

Theorem 2.1.6 Every vector of a finitely generated vector space can be written as a unique linear combination of a basis.

Theorem 2.1.7 Every basis of a finitely generated vector space contains the same number of elements. This number is called the **dimension** of the vector space.

2.2 Matrices

Matrices are important objects both in theoretical and applied mathematics. Actually the vectors of $\mathbb{R}^n$ and $\mathbb{C}^n$ can also be considered as matrices.

Let $x \in \mathbb{R}^n$ be a vector $x = (x_1, \ldots, x_n)$, then the same date can also be written as a column instead of the previous row. This is called the **transposition of** x, and we use the notation x^T, where

$$x^T = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

The vector x is a 1 by n matrix type matrix (it has 1 row and n columns), and x^T is an n by 1 matrix (it has n rows and 1 column). The vector x^T is called the **transpose of** x.

A slight extension of this method results a new kind of object, matrices.

Definition 2.2.1 — Matrix. Let $m, n \in \mathbb{N}$. An array of numbers

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1,m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} = (a_{ij})_{i=1,\ldots,n}^{j=1,\ldots,m},$$

[5] Observe that the trivial linear combination is always possible in the case of an arbitrary system.

is called an $n \times m$ real or complex **matrix**, where $a_{ij} \in \mathbb{R}$ or $a_{ij} \in \mathbb{C},\ i = 1,\ldots,n,\ j = 1,\ldots,m$. The set of $n \times m$ real matrices is denoted by $\mathscr{M}_{nm}(\mathbb{R})$, and similarly, the set of $n \times m$ complex matrices is denoted by $\mathscr{M}_{nm}(\mathbb{C})$.

The number a_{ij} is the ij**th entry of the matrix**.

The ith row of A is denoted by A_i and the jth column is denoted by A^j.

■ **Example 2.8** Consider the following table of numbers which contains the number of marks of first year students in some subjects.

Subjects | Marks	5	4	3	2	1
Mathematics for Engineers	3	10	15	4	2
Logic	2	9	19	2	2
Calculus	3	7	20	1	3
Physics	5	5	18	5	1

The data is more transparent given in this way. If there are same data sets from different universities it is easy to compare the results, and it is much easier if we skip the headlines and keep only the numbers.

$$A = \begin{bmatrix} 3 & 10 & 15 & 4 & 2 \\ 2 & 9 & 19 & 2 & 2 \\ 3 & 7 & 20 & 1 & 3 \\ 5 & 5 & 18 & 5 & 1 \end{bmatrix}$$

This is actually a 4 by 5 matrix because it has four rows and five columns.

For example, its third column is the following 4 by 1 column vector

$$A^3 = \begin{bmatrix} 15 \\ 19 \\ 20 \\ 18 \end{bmatrix},$$

and its second row is the following 1 by 5 row vector

$$A_2 = \begin{bmatrix} 2 & 9 & 19 & 2 & 2 \end{bmatrix}$$

■

Special matrices

An n by n matrix is called a **square matrix**. For example

$$\begin{bmatrix} 3 & 2 & -1 \\ 0 & -1 & 9 \\ 0 & 2 & 8 \end{bmatrix}$$

is a 3×3 square matrix.

The entries a_{ii} of a square matrix are called the **diagonal entries of the matrix**. For example the diagonal entries of the previous matrix are $3, -1, 8$.

If a square matrix has non-zero elements only in the diagonal, then it is called a **diagonal matrix**. For example

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

is a 3×3 diagonal matrix.

If a square matrix contains zero elements under(over) the diagonal, then it is said to be **upper(lower) triangular**. For example

$$\begin{bmatrix} 2 & 3 & 6 \\ 0 & -3 & -10 \\ 0 & 0 & 2 \end{bmatrix}$$

is a 3×3 upper triangular matrix, and

$$\begin{bmatrix} 5 & 0 & 0 \\ 2 & -4 & 0 \\ 4 & 5 & 1 \end{bmatrix}$$

is a 3 by 3 lower triangular matrix.

If all the entries of an $n \times m$ matrix are zero, then it is said to be the $n \times m$ **zero matrix**. For example

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is the 4×3 zero matrix. The conventional notation for the $n \times m$ zero matrix is $0_{n\times m}$. However, if there is no ambiguity, we skip the subscript, and use simply 0 denoting all full zero matrices.

2.2.1 Basic operations with matrices

Addition of matrices

If $A = (a_{ij})_{i=1,\dots,n}^{j=1,\dots,m}$ and $B = (b_{ij})_{i=1,\dots,n}^{j=1,\dots,m}$ are of the same type matrices, then their sum is defined in the following way

$$\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} + \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nm} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & \cdots & a_{1,m}+b_{1m} \\ \vdots & & \vdots \\ a_{n1}+b_{n1} & \cdots & a_{nm}+b_{nm} \end{bmatrix}.$$

■ **Example 2.9** ■

Theorem 2.2.1 — Properties of addition of matrices. Let $A, B, C \in \mathscr{M}_{n\times m}$[a] be arbitrary matrices of the same type. Then

- Matrix addition is a **binary operation** on the set of matrices, in other words, the sum of two $n \times m$ real/complex matrices will be an $n \times m$ real/complex matrix.
- Matrix addition is **associative**:

$$(A+B)+C = A+(B+C).$$

- There exists an **additive unit**: the full zero $n \times n$ matrix denoted by 0, with the property:

$$A+0 = 0+A = A.$$

- There exists a unique **additive inverse** of A denoted by $-A$ such that

$$A+(-A) = (-A)+A = 0.$$

- Matrix addition is a **commutative** operation, that is

$$A+B=A+B.$$

For short, $(\mathscr{M}_{n\times m},+)$ is an abelian group.

[a] If the scalar set is not specified, then it is always understood in a way that it equals either to $\mathbb{R}$ or to $\mathbb{C}$, but it is the same for both matrices.

Proof. Hint: All the statements follow from the properties of the addition of (real or complex) numbers, because the addition runs componentwise. ■

■ **Example 2.10** Let

$$A=\begin{bmatrix}0 & -1.02 & 11\\ -\pi & 2.3 & -7\end{bmatrix},\quad B=\begin{bmatrix}\sqrt{3} & 2 & -8\\ 0 & -1.3 & 5\end{bmatrix}$$

then

$$A+B=\begin{bmatrix}0 & -1.02 & 11\\ -\pi & 2.3 & -7\end{bmatrix}+\begin{bmatrix}\sqrt{3} & 2 & -8\\ 0 & -1.3 & 5\end{bmatrix}=\begin{bmatrix}\sqrt{3} & 0.98 & 3\\ -\pi & 1 & -2\end{bmatrix}$$

■

Multiplication by a scalar

Definition 2.2.2 Let α be a scalar and A be a matrix, then their product is defined in the following way:

$$\alpha A=\lambda\begin{bmatrix}a_{11} & \cdots & a_{1m}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nm}\end{bmatrix}=\begin{bmatrix}\alpha a_{11} & \cdots & \alpha a_{1m}\\ \vdots & & \vdots\\ \alpha a_{n1} & \cdots & \alpha a_{nm}\end{bmatrix}.$$

■ **Example 2.11**

$$3\begin{bmatrix}1 & 2 & -1\\ -2 & 4 & 4\end{bmatrix}=\begin{bmatrix}3 & 6 & -3\\ -6 & 12 & 12\end{bmatrix}.$$

■

Theorem 2.2.2 — Properties of multiplication by a scalar. Let A and B be arbitrary matrices of the same type α,β be arbitrary scalars, then

- $$0\cdot A=0 \qquad \text{and} \qquad 1\cdot A=A,$$
- $$(\alpha\beta)A=\alpha(\beta A),$$
- **Distributivity**:

$$(\alpha+\beta)A=\alpha A+\beta A \qquad \text{and1} \qquad \alpha(A+B)=\alpha A+\alpha B.$$

Proof. Hint: All the statements follows from the properties of multiplication of (complex or real) scalars, because the operation runs componentwise. ■

Corollary 2.2.3 The set of $n \times m$ real(complex) matrices constitute an $n \times m$ real(complex) vector space with respect to the introduced addition and multiplication by a scalar.

Transposition of matrices

Transposition is a unary operation. It assigns an $m \times n$ matrix to an $n \times m$ matrix.

Definition 2.2.3 — Transposition. Let

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1,m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} = (a_{ij})_{i=1,\ldots,n}^{j=1,\ldots,m},$$

be an $n \times m$ matrix, then the $m \times n$ matrix

$$A^T = \begin{bmatrix} a_{11} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = (a_{ij})_{j=1,\ldots,m}^{i=1,\ldots,n}$$

is called the **transpose of** A.

■ **Example 2.12** Let

$$A = \begin{bmatrix} 1 & 2 \\ -10 & 4 \\ -1 & 3 \\ 10 & 20 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \qquad C = \begin{bmatrix} -1 & 3 & 8 & -2 & 1.3 & -\pi & 95 \end{bmatrix}$$

be matrices, then their transposes are

$$A^T = \begin{bmatrix} 1 & -10 & -1 & 10 \\ 2 & 4 & 3 & 20 \end{bmatrix}, \qquad B^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}, \qquad C^T = \begin{bmatrix} -1 \\ 3 \\ 8 \\ -2 \\ 1.3 \\ -\pi \\ 95 \end{bmatrix}$$

■

Theorem 2.2.4 — Properties of transposition. Let A and B be arbitrarily given matrices of the same type, and α be a scalar, then

- $$\left(A^T\right)^T = A,$$

  that is, transposition is an **idempotent** operation,
- $$(A+B)^T = A^T + B^T,$$

- $(\alpha A)^T = \alpha A^T$.

Proof. Immediate consequence of the definition. ■

2.2.2 Multiplication of matrices

Multiplication of matrices is a little bit more subtle than addition of matrices or multiplication by a scalar.

Let us consider two matrices. Assume that the first one has the same number of columns as the number of rows of the second, that is to say $A \in \mathscr{M}_{n\times m}$ and $B \in \mathscr{M}_{m\times k}$. In this situation we can define their product $C = AB$, where $C \in \mathscr{M}_{n\times k}$ and the jth element of the ith row of C is resulted by the ith row of A and the jth column of B in the following way

$$c_{ij} = \sum_{t=1}^{m} a_{it} b_{tj}.$$

■ **Example 2.13** Let

$$A = \begin{bmatrix} 1 & 0 & -1 & 2 \\ 3 & 3 & 4 & 5 \end{bmatrix}, \qquad \text{and} \qquad B = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

Then A is a 2×4 matrix and B is a 4×1 matrix, so their product will be a 2 by 1 matrix, that is to say, a 2 dimensional column vector.

$$AB = \begin{bmatrix} 1 & 0 & -1 & 2 \\ 3 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1\cdot 1 + 0\cdot 1 + (-1)\cdot 1 + 2\cdot 1 & 3\cdot 1 + 3\cdot 1 + 4\cdot 1 + 5\cdot 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1+0-1+2 & 3+3+4+5 \end{bmatrix} = \begin{bmatrix} 2 & 15 \end{bmatrix}$$

■

Theorem 2.2.5 — Properties of matrix multiplication.

- Matrix multiplication is not commutative.
- Matrix multiplication is **associative**.
- Transpose of a product is equal to the product of the transposes in the reverse order, that is$(AB)^T = B^T A^T$.
- **Distributivity**: $(A+B)C = AC + BC$ and $A(B+C) = AB + AC$.

■ **Example 2.14** Let

$$A = \begin{bmatrix} -1 & 2 \end{bmatrix}, \qquad \text{and} \qquad B = \begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

then AB will be a 1 by 1 matrix, that is to say, a number. However, BA will be a 2 by 2 matrix.

$$AB = (-1)\cdot 2 + 3\cdot 4 = 5, \quad \text{and} \quad BA = \begin{bmatrix} -3 & 6 \\ -4 & 8 \end{bmatrix}.$$

■

■ **Example 2.15** Let

$$A = \begin{bmatrix} -1 & 2 \\ 3 & -4 \end{bmatrix}, \qquad \text{and} \qquad B = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

then

$$AB = \begin{bmatrix} 2 & 1 \\ 4 & -3 \end{bmatrix}, \quad \text{and} \quad BA = \begin{bmatrix} -3 & -4 \\ -1 & 2 \end{bmatrix}.$$

■

The previous examples verify the non-commutativity of matrix multiplication even in the case of square matrices.

Definition 2.2.4 — Multiplicative inverse of matrices. The $n \times n$ matrix

$$E_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

is said to be the n dimensional **unit matrix**. If there is no ambiguity, we write E instead of E_n. Let $A \in \mathscr{M}_{n\times n}$. We call A **invertible**, if there is such a matrix $B \in \mathscr{M}_{n\times n}$ for which $AB = BA = E$.

The calculation of the inverse practically means the solution of systems of linear equations, which is an independent topic inside linear algebra. So, we come back to numerical calculation of the inverse in that section.

2.3 System of linear equations

Linear systems (system of linear equations) and their solution are the most important parts of linear algebra because lots of problems in mathematics- in particular applied mathematics, and in applied sciences too- lead to a solution of a linear system of equations or frequently to a solution of linear systems.

Let us start, as a warming up, with an example from high school, namely with a solution of a two variables linear system. Its general form is

$$\begin{aligned} ax + by &= c \\ dx + ey &= f, \end{aligned}$$

where a, b, c, d, e, f are given constants, and x, y are the unknowns.

A solution of a system like this geometrically means to find the coordinates of the intersection of the lines represented by the equations in the system on the plane. For example, if our system has the form

$$\begin{aligned} x + y &= 1 \\ -x + y &= 0, \end{aligned}$$

then the solution is represented on Figure 2.1. For the coordinates of the solution we should execute the following steps:

- Add the equations. The result is a single equation with one variable y.
- Solve this equation for y.
- Substitute back the value of y into the first equation. The result is an equation with one variable x.

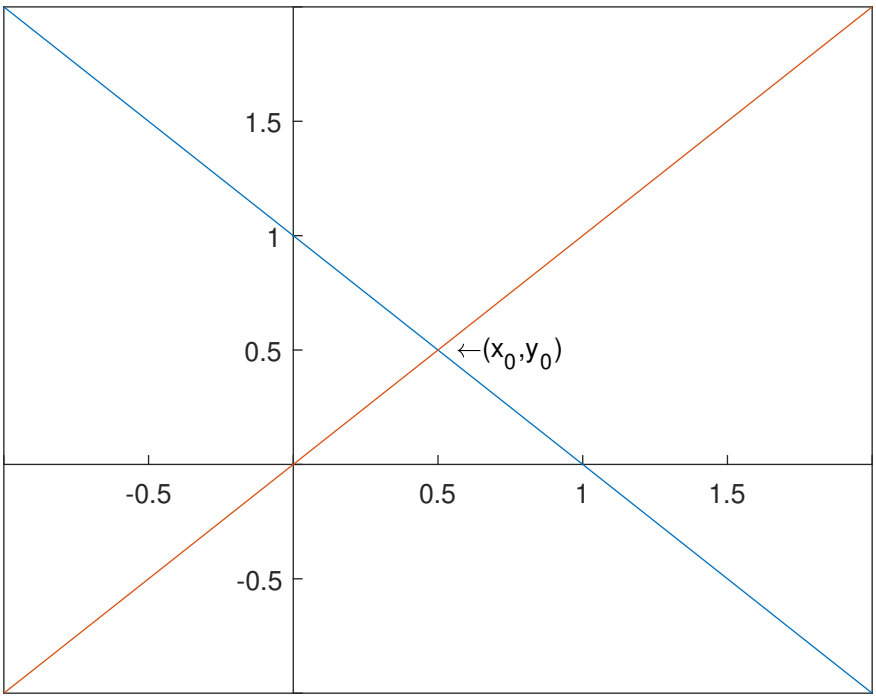


Figure 2.1: Solution of the system is $(x_0, y_0) = (\frac{1}{2}, \frac{1}{2})$

- Solve this equation for x.

$$\begin{array}{lll} x+y=1 & x+y=1 & x+\dfrac{1}{2}=1 \rightarrow x=\dfrac{1}{2} \\ -x+y=0 & 2y=1 & y=\dfrac{1}{2} \end{array}$$

It seems to be evidently advantageous to get rid of the variables and equality signs because in the calculations only the numbers are used. For this we can write the system, and the previous solution process in matrix form separating the right hand side of the system by a vertical line.

$$\left[\begin{array}{cc|c} 1 & 1 & 1 \\ -1 & 1 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 1 & 1 \\ 0 & 2 & 1 \end{array}\right] \Rightarrow 2y=1 \Rightarrow y=\frac{1}{2} \Rightarrow x+\frac{1}{2}=1 \Rightarrow x=\frac{1}{2}.$$

So, the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

Let us see an another example with a little bit more "complicated" coefficients.

$$\begin{aligned} 2x-3y &= 11 \\ 3x+y &= 3. \end{aligned}$$

Using the previous method we have the following, where the first step $-\frac{3}{2}$ times the first row is added to the second row.

$$\left[\begin{array}{cc|c} 2 & -3 & 11 \\ 3 & 1 & 3 \end{array}\right] \sim \left[\begin{array}{cc|c} 2 & -3 & 11 \\ 0 & \frac{11}{2} & -\frac{27}{2} \end{array}\right] \Rightarrow \frac{11}{2}y = -\frac{27}{2} \Rightarrow y = -\frac{27}{11} \Rightarrow 2x + \frac{81}{11} = \frac{121}{11} \Rightarrow x = \frac{20}{11}.$$

So, the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{20}{11} \\ -\frac{27}{11} \end{bmatrix}.$$

One can recognize that the goal of the first step is the reduction of the number of the variables. This strategy is the technique of Gaussian elimination. The method is discussed in detail in the next section.

2.3.1 Classification of linear systems

An arbitrary linear system has the form

$$Ax = b,$$

where $A \in \mathscr{M}_{n\times m}(\mathbb{R})$ or $\mathscr{M}_{n\times m}(\mathbb{C})$ is a given matrix $b \in \mathbb{R}^n$ or $\mathbb{C}^n$ is a given vector and $x \in \mathbb{R}^m$ or $\mathbb{C}^m$ is unknown.

Definition 2.3.1 If the vector b is zero, then the system of linear equations is said to be **homogeneous** otherwise it is called **inhomogeneous**.

If the system is inhomogeneous the matrix

$$[A|b]$$

is said to be the **augmented matrix of the system**.

According to the definition, a homogeneous system has the form

$$Ax = 0.$$

In this case the augmentation with b does not make sense, so A is simply called the **matrix of the homogeneous system**.

The solution set of a homogeneous system has a nice structure.

Theorem 2.3.1 The solution set of a homogeneous system of linear equations is a subspace.

Proof. Let α be a scalar and x, y are solutions of the homogeneous linear system. Then

$$A(\alpha x) = \alpha Ax = \alpha \cdot 0 = 0,$$

so αx is also a solution. Moreover

$$A(x-y) = Ax - Ay = 0 - 0 = 0,$$

so $x-y$ is also a solution. Using the subspace criteria (Theorem 2.1.3) we get the statement. ■

This subspace can be the trivial one, say $\{0\}$, which contains only the additive unit, the zero vector only. This means that an arbitrary homogeneous system always have at least one solution, the zero vector. The important consequence of the previous theorem is the following.

Corollary 2.3.2 The solution set of an arbitrary homogeneous linear system contains either exactly one element (the zero vector) or it contains infinitely many elements (a non-trivial subspace).

These statements imply that a solution of a homogeneous system is practically to find a basis in its solution subspace.

■ **Example 2.16** Let us consider the following homogeneous linear system.

$$\begin{aligned} x_1+x_2+x_3 &= 0\\ x_1+2x_2-x_3 &= 0, \end{aligned}$$

then we can perform the following calculation

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

In the last row there are two unknowns with non-zero coefficients x_2 and x_3. So, we have to choose one as a parameter, and express the variables with the aid of this parameter.

$$x_3 = p, \quad \text{then } x_2-2p=0 \Rightarrow x_2=2p \Rightarrow x_1+x_2+x_3 = x_1+2p+p=0 \Rightarrow x_1=-3p$$

This entails the solution of the linear system:

$$x = p\begin{bmatrix} -3\\ 2\\ 1 \end{bmatrix}, \quad p\in\mathbb{R} \qquad \text{or} \qquad x \in \operatorname{span}\left\{\begin{bmatrix} -3\\ 2\\ 1 \end{bmatrix}\right\}$$

Here the solution of the homogeneous linear system is a one-dimensional subspace of $\mathbb{R}^3$, and $\begin{bmatrix} -3\\ 2\\ 1 \end{bmatrix}$ is a base of the solution subspace. ■

In the next example the solution subspace will be a two-dimensional one.

■ **Example 2.17** Let us find the solution of the following homogeneous system of linear equations.

$$\begin{aligned} 2x_1-2x_2+4x_3+5x_4 &= 0\\ x_1+2x_2-x_3-2x_4 &= 0 \end{aligned}$$

then we have

$$\begin{bmatrix} 2 & -2 & 4 & 5\\ 1 & 2 & -1 & -2 \end{bmatrix} \sim \begin{bmatrix} 2 & -2 & 4 & 5\\ 0 & 3 & -3 & -\frac{9}{2} \end{bmatrix}$$

there are three unknowns with non-zero coefficients in the last row, so we have to choose two parameters.

$$x_4=p_1,\ x_3=p_2 \Rightarrow 3x_2-3p_2-\tfrac{9}{2}p_1=0 \Rightarrow x_2=p_2+\tfrac{3}{2}p_1 \Rightarrow$$

$$2x_1-2(p_2+\tfrac{3}{2}p_1)+4p_2+5p_1=0 \Rightarrow x_1=-p_1-p_2.$$

So, the solution subspace has the form

$$x = p_1\begin{bmatrix} -1\\ \frac{3}{2}\\ 0\\ 1 \end{bmatrix} + p_2\begin{bmatrix} -1\\ 1\\ 1\\ 0 \end{bmatrix}, \quad p_1,p_2\in\mathbb{R}, \qquad \text{or} \qquad x \in \operatorname{span}\left\{\begin{bmatrix} -1\\ \frac{3}{2}\\ 0\\ 1 \end{bmatrix}, \begin{bmatrix} -1\\ 1\\ 1\\ 0 \end{bmatrix}\right\}.$$

This is a two-dimensional subspace of $\mathbb{R}^4$, where

$$\left\{ \begin{bmatrix} -1 \\ \frac{3}{2} \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\}$$

is a base. ∎

The solution set of an inhomogeneous system cannot be a subspace, because the zero vector cannot be a solution if b is different from the zero vector.

$$A0 = 0 \neq b.$$

However, the solution set of the inhomogeneous system also has a nice structure, which depends on the solution subspace of the corresponding homogeneous system $Ax = 0$. Namely, the solution set of the inhomogeneous system is a translation of the solution subspace of the homogeneous system.

Definition 2.3.2 — Affine subspace. Let V be a vector space, S be a subspace of V and $v \in V$ be an arbitrary vector. The set

$$v + S = \{x \in S \mid x = v + s, \text{ where } s \in S\}$$

is said to be and **affine subspace of** V.

If $v \in S$, then $v + S = S$, so subspaces are also affine subspaces as well.

Theorem 2.3.3 The solution set of an inhomogeneous system of linear equations is an affine subspace, where the translation vector is an arbitrary solution of the system (particular solution) and the subspace is the solution set of the corresponding homogeneous system.

Proof. Let us consider the inhomogeneous system

$$Ax = b.$$

Let S be the solution subspace of the corresponding inhomogeneous system $Ax = 0$, and v be a particular solution of the inhomogeneous system, that is to say $Av = b$, then for all $s \in S$ we have

$$A(v + s) = Av + As = b + 0 = b,$$

so all the vectors, which have the form $v + s$ are solutions of the inhomogeneous system.

For the opposite direction, assume that x is an arbitrary solution of the system. Then

$$A(x - v) = Ax - Av = b - b = 0,$$

so $x - v$ is a solution of the corresponding homogeneous system, that is to say, $x - v \in S$ which implies $x \in v + S$, which gives the end of the proof. ∎

∎ **Example 2.18** Let us determine the solution of the following inhomogeneous linear system.

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + 2x_2 - x_3 &= 2, \end{aligned}$$

then we get

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 2 & -1 & 2 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & -2 & 1 \end{array}\right]$$

Similarly to the homogeneous case we have

$$x_3 = p, \quad \text{then } x_2 - 2p = 1 \Rightarrow x_2 = 1 + 2p \Rightarrow x_1 + x_2 + x_3 = x_1 + 1 + 2p + p = 1 \Rightarrow x_1 = -3p$$

This entails the solution of the linear system:

$$x = \begin{bmatrix} -3p \\ 1+2p \\ p \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + p \begin{bmatrix} -3 \\ 2 \\ 1 \end{bmatrix}, \quad p \in \mathbb{R} \qquad \text{or} \qquad x \in \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \operatorname{span}\left\{ \begin{bmatrix} -3 \\ 2 \\ 1 \end{bmatrix} \right\}.$$

■

2.3.2 Gaussian elimination, solution of linear systems

Let us start with an example again. Consider the following linear system.

$$\begin{aligned} x + 2y - 5z &= 1 \\ 2x + 2y + z &= -6 \\ 4x + 2y - 2z &= 0 \end{aligned}$$

Its matrix form is

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 2 & 2 & 1 & -6 \\ 4 & 2 & -2 & 0 \end{array}\right]$$

In the first step we would like to reduce the number of variables from three to two in the second and in the third row using the first row. To get the desired effect subtract two times the first row from the second row, and four times the first row from the third row. As a result we can omit x in the second and in the third rows.

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 2 & 2 & 1 & -6 \\ 4 & 2 & -2 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & -6 & 18 & -4 \end{array}\right]$$

Now, we subtract three times the second row from the third row omitting y from the third row.

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & -6 & 18 & -4 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & 0 & -15 & 20 \end{array}\right]$$

At last, we determine the coordinates of the solution vector in reverse order. We derive at first the value of z using the last row of the last matrix.

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & 0 & -15 & 20 \end{array}\right] \Rightarrow z = -\frac{20}{15} = -\frac{4}{3}.$$

Substituting back into the second row we get y.

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & 0 & -15 & 20 \end{array}\right] \Rightarrow -2y + 11\left(-\frac{4}{3}\right) = -8 \Rightarrow -2y = \frac{44}{3} - \frac{24}{3} = \frac{20}{3} \Rightarrow y = -\frac{10}{3}.$$

The value of x comes from the first row substituting back the values of y and z.

$$\left[\begin{array}{ccc|c} 1 & 2 & -5 & 1 \\ 0 & -2 & 11 & -8 \\ 0 & 0 & -15 & 20 \end{array}\right] \Rightarrow x + 2\left(-\frac{10}{3}\right) - 5\left(-\frac{4}{3}\right) = 1 \Rightarrow x - \frac{20}{3} + \frac{20}{3} = 1 \Rightarrow x = 1.$$

During the Gaussian elimination, our strategy is to transform the matrix into echelon form using some simple row operations.

Definition 2.3.3 — Elementary row operations. Let us consider a matrix. The following operations with the rows of the matrix are said to be **elementary row operations**.

- Interchanging two rows of the matrix.
- Multiplying a row of the matrix by a non-zero scalar.
- Adding two rows of the matrix.

Definition 2.3.4 — Echelon matrix. A matrix is said to be **(row) echelon matrix** or it has **echelon form** if the column index (second index) of the first non-zero element in the ith row is less than the column index of the first non-zero element in the $i+1$th row for all $i = 1, \cdots, n-1$, and if a row does not contain any non-zero element (full zero row), then it is at the bottom of the matrix.

Practically this means that in an echelon matrix if we consider an arbitrary row, then in the rows below zeros are only on the left hand side from the position of the first non-zero element of the row in question. In other words, the first non-zero number in an arbitrary row is strictly to the right of the first non-zero number of the row above it.

■ **Example 2.19** The matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 2 & 91 & -1 & 2 & 1 \\ 0 & 3 & 11 & 6 & 0 & 9 & 0 & 2 & 1 \\ 0 & 0 & 13 & 41 & -2 & 0 & -1 & -2 & 21 \\ 0 & 0 & 0 & 5 & 6 & 91 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & -2 & 22 & 33 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

has echelon form and the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 2 & 91 & -1 & 2 & 1 \\ 0 & 3 & 11 & 6 & 0 & 9 & 0 & 2 & 1 \\ 0 & 1 & 13 & 41 & -2 & 0 & -1 & -2 & 21 \\ 0 & 0 & 0 & 5 & 6 & 91 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & -2 & 22 & 33 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

has not, because the column index of the first non-zero element in the second and also in the third row is two, but according to the definition in the third row this index should be strictly greater than two. ■

Gaussian elimination: This algorithm is a sequence of elementary row operations performed on and arbitrary matrix, which results a (row) echelon matrix. Usually, it is for solving systems of linear equations, however, with some restrictions, it is applicable in other numerical calculations of linear algebra such as determinants, rank of matrices and so on.

■ **Example 2.20** Find the solution of the following homogeneous system of linear equations!

$$\begin{aligned} x_1 - 2x_2 - 4x_3 + x_4 - 3x_5 &= 0 \\ -x_1 + x_2 - 2x_3 - 2x_4 - 2x_5 &= 0 \\ 2x_1 - 5x_2 - 14x_3 + x_4 - 11x_5 &= 0 \end{aligned}$$

Let us write it into matrix form and apply Gaussian elimination.

$$\begin{bmatrix} 1 & -2 & -4 & 1 & -3 \\ -1 & 1 & -2 & -2 & -2 \\ 2 & -5 & -14 & 1 & -11 \end{bmatrix} \sim \begin{bmatrix} 1 & -2 & -4 & 1 & -3 \\ 0 & -1 & -6 & -1 & -5 \\ 0 & -1 & -6 & -1 & -5 \end{bmatrix} \sim \begin{bmatrix} 1 & -2 & -4 & 1 & -3 \\ 0 & -1 & -6 & -1 & -5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We have four unknowns with non-zero coefficients in the last row. So, we have to introduce three parameters: $x_5 = p_1$, $x_4 = p_2$, $x_3 = p_3$. Substituting these back into the equations, we have

$$-x_2 - 6p_3 - p_2 - 5p_1 = 0 \Rightarrow x_2 = -5p_1 - p_2 - 6p_3.$$

From the first equation we get

$$x_1 - 2(-5p_1 - p_2 - 6p_3) - 4p_3 + p_2 - 3p_1 = 0 \Rightarrow x_1 = -7p_1 - 3p_2 - 8p_3.$$

So, the solution set is a three dimensional subspace of $\mathbb{R}^5$, and

$$x = \begin{bmatrix} -7p_1 - 3p_2 - 8p_3 \\ -5p_1 - p_2 - 6p_3 \\ p_3 \\ p_2 \\ p_1 \end{bmatrix} = p_1 \begin{bmatrix} -7 \\ -5 \\ 0 \\ 0 \\ 1 \end{bmatrix} + p_2 \begin{bmatrix} -3 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + p_3 \begin{bmatrix} -8 \\ -6 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad p_1, p_2, p_3 \in \mathbb{R},$$

or

$$x \in \operatorname{span}\left\{ \begin{bmatrix} -7 \\ -5 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -8 \\ -6 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

■

■ **Example 2.21** Find the solution of the following inhomogeneous linear system!

$$\begin{aligned} -x_1 + x_2 + 2x_3 + 10x_4 + x_5 &= -1 \\ 2x_1 + 4x_2 + 8x_3 + 9x_4 - x_5 &= 2 \\ x_1 + x_2 + 2x_3 + 2x_4 + 4x_5 &= 1 \end{aligned}$$

Execute Gaussian elimination on the augmented matrix of the system!

$$\left[\begin{array}{ccccc|c} -1 & 1 & 2 & 10 & 1 & -1 \\ 2 & 4 & 8 & 9 & -1 & 2 \\ 1 & 1 & 2 & 2 & 4 & 1 \end{array}\right] \sim \left[\begin{array}{ccccc|c} -1 & 1 & 2 & 10 & 1 & -1 \\ 0 & 6 & 12 & 29 & 1 & 0 \\ 0 & 2 & 4 & 12 & 5 & 0 \end{array}\right] \sim \left[\begin{array}{ccccc|c} -1 & 1 & 2 & 10 & 1 & -1 \\ 0 & 6 & 12 & 29 & 1 & 0 \\ 0 & 0 & 0 & \frac{7}{3} & \frac{14}{3} & 0 \end{array}\right]$$

In the last row we have two unknowns with non-zero coefficients. We have to introduce one parameter $x_5 = p_1$. From this we get

$$\frac{7}{3}x_4 + \frac{14}{3}p_1 = 0 \Rightarrow x_4 = -p_1.$$

In the second row we have two new unknowns with non-zero coefficients x_2 and x_3. So, we have to introduce a new parameter $x_3 = p_2$. Using these we have

$$6x2 + 12p_2 + 29(-2p_1) + p_1 = 0 \Rightarrow x_2 = \frac{57}{6}p_1 - 2p_2.$$

Substituting everything back into the first equation we can determine x_1.

$$-x_1 + \frac{57}{6}p_1 - 2p_2 + 2p_2 + 10(-2p_1) + p_1 = -1 \Rightarrow x_1 = 1 - \frac{29}{3}p_1.$$

So, the solution set is the following affine subspace of $\mathbb{R}^5$.

$$x = \begin{bmatrix} 1-\frac{29}{3}p_1 \\ \frac{57}{6}p_1 - 2p_2 \\ p_2 \\ -2p_1 \\ p_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + p_1 \begin{bmatrix} -\frac{29}{3} \\ \frac{57}{6} \\ 0 \\ -2 \\ 1 \end{bmatrix} + p_2 \begin{bmatrix} 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad p_1, p_2 \in \mathbb{R},$$

or

$$x \in \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \operatorname{span}\left\{ \begin{bmatrix} -\frac{29}{3} \\ \frac{57}{6} \\ 0 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

■

2.3.3 Calculation of the inverse matrix

As it has been mentioned before, the calculation of the inverse matrix practically means the solution of n inhomogeneous systems of linear equations if the matrix is an n by n one.

However, simple Gaussian elimination is not enough here, we need one of its modifications, namely, the Gauss-Jordan elimination. The only difference here is that the goal of the method is not only the echelon form of the matrix, but the transformation of it into the n by n unit matrix. During this process we write the unit matrix at the beginning beside the original matrix, and at the end we get the inverse of the matrix.

In other words, we combine the elements of the natural base of $\mathbb{R}^n$ from the columns of the matrix linearly.

■ **Example 2.22** Let us calculate the inverse of the following matrix!

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

During the process (Gauss-Jordan elimination), in a nutshell, we execute the following transformation:

$$[A \,|\, E_3] \sim \cdots \sim [E_3 \,|\, A]$$

Let us see the method in practice. At first we divide the first row by 2 to get 1 in the first position of the first row. Subtract now the new first row from the old second row, and subtract four times the new first row form the old third row.

$$\left[\begin{array}{ccc|ccc} 2 & -1 & 4 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ 4 & -1 & 4 & 0 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 1 & -\frac{1}{2} & 2 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & -3 & -\frac{1}{2} & 1 & 0 \\ 0 & 1 & -4 & -2 & 0 & 1 \end{array}\right]$$

Multiply by 2 the second row to get 1 in the second position of the second row. Add the $\frac{1}{2}$ multiply of the second new row to the first old row, and subtract the new row from the third old row.

$$\left[\begin{array}{ccc|ccc} 1 & -\frac{1}{2} & 2 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & -3 & -\frac{1}{2} & 1 & 0 \\ 0 & 1 & -4 & -2 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & -6 & -1 & 2 & 0 \\ 0 & 0 & 2 & -1 & -2 & 1 \end{array}\right]$$

Divide the third row by 2 to get 1 in the third position of the third row. Add six times this new third row to the old second row, and once to the old first row.

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & -6 & -1 & 2 & 0 \\ 0 & 0 & 2 & -1 & -2 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & -4 & -4 & 3 \\ 0 & 0 & 1 & -\frac{1}{2} & -1 & \frac{1}{2} \end{array}\right]$$

So, the inverse of A is

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

■

2.4 Determinants

The determinant is a map, which assigns a number to a square matrix. Geometrically this number is the signed volume of the parallelotope with sides determined by the rows of the matrix. In the case of a two by two matrix this is the signed area of a parallelogram and in the case of a three by three matrix this is the signed volume of a parallelepiped.

At first, we define the determinant of a two by two and a three by three matrix. After some examples, we define the determinant of an arbitrary square matrix in a recursive way using Laplace's expansion theorem.

Definition 2.4.1 — Determinant of a two by two matrix. Let's consider the matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. We assign a number to A in the following manner:

$$\det A = a_{11}a_{22} - a_{12}a_{21}.$$

This number is called the **determinant** of A.

Because its small size, calculation with the determinant of two by two matrices is reasonably transparent. Using this transparency, we verify some useful properties of determinants which will remain also true in the general case.

Proposition 2.4.1 — Properties of the determinant (2×2 case). Let A and B be two 2×2 matrices and α be a scalar, then

- if two rows or columns of A are equal, then the determinant of A is zero;
- if A contains a full zero row, then its determinant is zero;
- if we multiply the elements of a row or a column of A by α , then the determinant will be $\alpha \det A$;
- if we add a scalar multiply of a row/column to an another row/column A, then the value of the determinant does not change;
- the determinant of the unit matrix is one;
- the determinant of A is equal to the determinant of its transpose;
- if we interchange two rows or columns of A, the determinant changes its sign;
- the determinant of a product of two square matrices is equal to the product of their determinants;
- the determinant of an invertible matrix is equal to the reciprocal of the determinant of the inverse matrix;
- the determinant of an upper triangular matrix is equal to the product of the diagonal entries of the matrix.

Proof. Let

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \text{and} \qquad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

We prove statements concerning rows and columns only for rows. The corresponding statements concerning columns run in a pretty similar way.

- Assume that the rows of A are equal, then $a_{11} = a_{21}$ and $a_{12} = a_{22}$, and

$$\det A = a_{11}a_{22} - a_{12}a_{21} = a_{11}a_{12} - a_{12}a_{11} = 0.$$

- Let us assume that the second row of A is a full zero row, then

$$\det A = a_{11}0 - a_{12}0 = 0.$$

- Let us multiply the first row by α, then

$$\det \begin{bmatrix} \alpha a_{11} & \alpha a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \alpha a_{11}a_{22} - \alpha a_{12}a_{21} = \alpha(a_{11}a_{22} - a_{12}a_{21}) = \alpha \det A.$$

- Let us add the first row to the second row, then we get

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

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

- $$\det A = a_{11}a_{22} - a_{12}a_{21} = a_{11}a_{22} - a_{21}a_{12} = \det A^T.$$

- Let us interchange the rows of A.

$$\begin{bmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{bmatrix} = a_{21}a_{12} - a_{22}a_{11} = -(a_{11}a_{22} - a_{12}a_{21}) = -\det A.$$

The remaining part can be proved in a very similar way like the previous statements, however, the corresponding calculations are a little bit lengthier. ■

■ **Example 2.23** Let us calculate the determinants of the following matrices!

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

  We can have the same result if we use the properties of the determinant. Namely, the determinant of the unit matrix is one, and this matrix can be derived from the unit matrix interchanging their rows, which results the minus sign.

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

- $$\det\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix} = 4-6 = -2$$
- $$\det\begin{bmatrix}3 & 4\\ 1 & 2\end{bmatrix} = 6-4 = 2.$$
- $$\det\begin{bmatrix}1+i & i\\ -i & 2\end{bmatrix} = 2(1+i)-i(-i) = 2+2i+i^2 = 2+2i-1 = 1+2i.$$
- $$\det\begin{bmatrix}2+i & 3i\\ -i & 2i\end{bmatrix} = 2i(2+i)-3i(-i) = 4i+2i^2+3i^2 = -5+4i.$$
- $$\det\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} = 0$$

  because their rows are the same.

■

Definition 2.4.2 — Determinant of a three by three matrix. Let us consider the matrix $A = \begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}$. We assign a number to A in the following way:

$$\det A = a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}-a_{12}a_{21}a_{33}-a_{11}a_{23}a_{32}.$$

This number is said to be the **determinant** of A.

■ **Example 2.24** Let

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

then

$$\det A = 1\cdot5\cdot9+2\cdot6\cdot7+3\cdot4\cdot8-3\cdot5\cdot7-2\cdot4\cdot9-1\cdot6\cdot8 = 45+84+96-105-72-48 = 0.$$

■

2.4.1 Laplace expansion theorem, determinant of n by n matrices

Definition 2.4.3 Let A be an $n\times n$ matrix. The ij **minor** of A is denoted by A_{ij} , is the determinant of the $(n-1)\times(n-1)$ matrix that results from A by deleting the ith row and the jth column.

Theorem 2.4.2 — Laplace expansion theorem. If A is an $n \times n$ matrix, then its determinant can be calculated using the expansion below:

$$\det A = \sum_{t=1}^{n} (-1)^{i+t} a_{it} \det A_{it}, \quad \text{(expansion with respect to the } i\text{th row)}$$

or

$$\det A = \sum_{t=1}^{n} (-1)^{t+j} a_{tj} \det A_{tj}, \text{ (expansion with respect to the } j\text{th column).}$$

The advantage of this theorem is that the determinant of an $n \times n$ matrix can be written as a signed sum of $(n-1) \times (n-1)$ matrices. We can continue this process recursively with expressing the $(n-1) \times (n-1)$ determinants as signed sums of $(n-2) \times (n-2)$ determinants till 2×2 determinants.

In practice we combine the power of Laplace's theorem with restricted Gaussian elimination. Namely, we can add a scalar multiply of a row/column to an another row/column but the value of the determinant does not change (see Proposition 2.4.1).

Our strategy is to produce lots of zeros in a row or in a column and applying the theorem to this row or column. This way we can decrease the number of terms of the sum in the expansion. As a result of the repeated application of this strategy we can reduce the calculation of the determinant of an arbitrary matrix into the calculation of the determinant of a two by two matrix.

■ **Example 2.25** Let us expand the following determinant with respect to its second column!

$$\det \begin{bmatrix} 1 & 0 & 2 & -3 \\ 1 & 0 & 1 & 5 \\ 2 & 0 & 4 & 6 \\ 0 & 2 & 12 & -9 \end{bmatrix} = (-1)^{1+2} \cdot 0 \cdot \det \begin{bmatrix} 1 & 1 & 5 \\ 2 & 4 & 6 \\ 0 & 12 & -9 \end{bmatrix} + (-1)^{2+2} \cdot 0 \cdot \det \begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & 6 \\ 0 & 12 & -9 \end{bmatrix} +$$

$$(-1)^{3+2} \cdot 0 \cdot \det \begin{bmatrix} 1 & 2 & -3 \\ 1 & 1 & 5 \\ 0 & 12 & -9 \end{bmatrix} + (-1)^{4+2} \cdot 2 \cdot \det \begin{bmatrix} 1 & 2 & -3 \\ 1 & 1 & 5 \\ 2 & 4 & 6 \end{bmatrix} =$$

$$2 \cdot \det \begin{bmatrix} 1 & 2 & -3 \\ 1 & 1 & 5 \\ 2 & 4 & 6 \end{bmatrix} = 2(6+20-12+6-12-20) = -24$$

■

2.4.2 Calculation of determinants using Gaussian elimination

The simplest way to calculate big determinants is to transform them into upper triangular form, using restricted Gaussian elimination (those elementary row operations which do not change the value of the determinant).

■ **Example 2.26**

$$\det \begin{bmatrix} 1 & -1 & 2 & -3 \\ 11 & -9 & 24 & -30 \\ 2 & 2 & 4 & 6 \\ -9 & 8 & -13 & 30 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 2 & -3 \\ 0 & 2 & 2 & 3 \\ 0 & 4 & 0 & 12 \\ 0 & -1 & 5 & 3 \end{bmatrix} = \det \begin{bmatrix} 1 & -1 & 2 & -3 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & -4 & 6 \\ 0 & 0 & 6 & \frac{9}{2} \end{bmatrix} =$$

$$\det \begin{bmatrix} 1 & -1 & 2 & -3 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & -4 & 6 \\ 0 & 0 & 0 & \frac{27}{2} \end{bmatrix} = 1 \cdot 2 \cdot (-4) \cdot \frac{27}{2} = -108.$$

■

2.5 Euclidean spaces

In the previous part, we saw how a structure can be constructed on a set of vectors (addition and multiplication by a scalar). This is a quite simple configuration with simple calculation rules (properties of the operations), which are very similar to the calculation rules of real or complex numbers.

In applications we need more. In engineering or in physics it is accustomed to measure the length and angel of quantities (e.g. magnitude and direction of forces). Orthogonality also has a special importance in several applications.

Best models, which have all the required additional properties, are Euclidean spaces. The new key concept is the concept of the inner product. All the other additional features of the structure can be derived from the inner product.

There are some differences between real and complex Euclidean spaces. At first we deal with the real case, and the last subsection is devoted to the complex case.

2.5.1 Inner product and length of vectors

The operation in Euclidean spaces[6] is the dot product or inner product. There are several approaches to define it. Here we choose the most elegant and easily memorizable algebraic way, which uses matrix multiplication.

Important remark: Unless stated otherwise, all vectors are assumed to be column vectors!

Definition 2.5.1 — Inner product. Let $x, y \in \mathbb{R}^n$ be given vectors, then the real number

$$x^T y = x_1 y_1 + \cdots + x_n y_n$$

is said to be the **inner product of** x **and** y.[a]

[a] The expression **dot product of** x **and** y is also common in the literature. Here we use the matrix multiplication x transpose times y notation, however the notations $x \cdot y$, and $\langle x, y \rangle$ are also frequently used. The latter one is especially important when the members of the space in question are not vectors.

■ **Example 2.27** Let $x, y \in \mathbb{R}^5$ be given in the following way:

$$x = \begin{bmatrix} 3 \\ 10 \\ -0.2 \\ -1.34 \\ 1 \end{bmatrix}, \qquad y = \begin{bmatrix} -2 \\ 0.2 \\ 3 \\ 1 \\ -5 \end{bmatrix},$$

then their inner product is

$$x^T y = 3 \cdot (-2) + 10 \cdot 0.2 + (-0.2) \cdot 3 + (-1.34) \cdot 1 + 1 \cdot (-5) = -6 + 2 - 0.6 - 1.34 - 5 = -10.94$$

■

Theorem 2.5.1 — Properties of the inner product in $\mathbb{R}^n$. Let $x, y, z \in \mathbb{R}^n$ be vectors, and $\alpha \in \mathbb{R}$ be a scalar, then

[6] Here we deal with finite dimensional, real Euclidean spaces, which can always be identified as $\mathbb{R}^n$ for some n. There are important infinite dimensional Euclidean spaces too.

- the (real) inner product is **symmetric**, that is,

$$x^T y = y^T x.$$

- the (real) inner product is **additive** in its both variables, that is,

$$(x+y)^T z = x^T z + y^T z, \qquad \text{and} \qquad x^T (y+z) = x^T y + x^T z.$$

- the (real) inner product is **homogeneous** in its both variables, that is,

$$(\alpha x)^T y = \alpha (x^T y), \qquad \text{and} \qquad x^T (\alpha y) = \alpha (x^T y).$$

- the (real) inner product is **positive definite**, that is,

$$x^T x \geq 0, \qquad \text{and it is zero if and only if } \; x = 0.$$

The proof immediately follows from the definition of the inner product in $\mathbb{R}^n$.

We can also formulate the statements of the previous theorem that the inner product in $\mathbb{R}^n$ is a **symmetric, bilinear, positive definite map**. Actually, like in the case of the definition of vector spaces, we can say that a vector space V over $\mathbb{R}$ is a real Euclidean space if there is a two variables function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$, which is a symmetric, bilinear, positive definite map. In detail, for all $u, v, w \in V$ and for all $\alpha \in \mathbb{R}$ we have

- symmetry:

$$\langle u, v \rangle = \langle v, u \rangle.$$

- bilinearity:

$$\langle u+v, w \rangle = \langle u, w \rangle + \langle v, w \rangle, \qquad \text{and} \qquad \langle u, v+w \rangle = \langle u, v \rangle + \langle u, w \rangle.$$

- homogeneity:

$$\langle \alpha u, v \rangle = \alpha \langle u, v \rangle, \qquad \text{and} \qquad \langle u, \alpha v \rangle = \alpha \langle u, v \rangle.$$

- positive definitness:

$$\langle u, u \rangle \geq 0, \qquad \text{and it is zero if and only if } \; u = 0.$$

■ **Example 2.28** Let us consider the continuous, real valued functions on the interval $[0, 1]$. This set is a real vector space with respect to the pointwise addition, and scalar multiplication. Then the two variables function

$$(f, g) \mapsto \langle f, g \rangle = \int_0^1 f(t) g(t) dt$$

is a symmetric, bilinear, positive definite map. So, this vector space is also a Euclidean space. ■

Definition 2.5.2 — Length or norm of vectors in $\mathbb{R}^n$. Let $x \in \mathbb{R}^n$ be a vector, then the real number

$$\|x\| = \sqrt{x^T x} = \sqrt{x_1^2 + \cdots + x_n^2}$$

is said to be the **norm of** x or the **length of** x.

■ **Example 2.29** Let

$$x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \qquad \text{and} \qquad y = \begin{bmatrix} 2 \\ -1 \\ 7 \\ 1 \end{bmatrix},$$

then their norms are

$$\|x\| = \sqrt{(-1)^2 + 1^2 + (-1)^2 + 1^2 + (-1)^2} = \sqrt{1+1+1+1+1} = \sqrt{5} \approx 2.2361,$$

and

$$\|y\| = \sqrt{2^2 + (-1)^2 + 7^2 + 1^2} = \sqrt{4+1+49+1} = \sqrt{55} \approx 7.4162.$$

■

Theorem 2.5.2 — Cauchy-Schwarz inequality in $\mathbb{R}^n$. Let $x, y \in \mathbb{R}^n$, then

$$|x^T y| \leq \|x\| \cdot \|y\|.$$

Proof. Let α be an arbitrary scalar and $x, y \in \mathbb{R}^n$ be arbitrary vectors, then by the positive definitness of the inner product we have

$$0 \leq (y+\alpha x)^T (y+\alpha x) = y^T y + y^T(\alpha x) + (\alpha x)^T y + (\alpha x)^T (\alpha x) = \|y\|^2 + 2\alpha(x^T y) + \alpha^2 \|x\|^2.$$

Here we used the properties of the inner product. The quadratic expression for α is greater than zero if and only if its discriminant is non-positive, that is to say

$$(2x^T y)^2 - 4\|x\|^2\|y\|^2 \leq 0,$$

which is equivalent to our statement. ■

Definition 2.5.3 — Orthogonality. Let $x, y \in \mathbb{R}^n$ be given vectors. They are called **orthogonal** or **perpendicular** if their inner product is zero, that is

$$x^T y = 0.$$

■ **Example 2.30** It is clear from the definition that the zero vector is perpendicular to an arbitrary vector.

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$ be given, then $y = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix} \in \mathbb{R}^2$ is always orthogonal to x, because

$$x^T y = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix} = x_1(-x_2) + x_2 x_1 = -x_1 x_2 + x_1 x_2 = 0.$$

■

Reminder from hight school: Pythagoras theorem states that the square of the side opposite the right angle is equal to the sum of the squares of the other two sides of an arbitrary right triangle.

This theorem is also true in $\mathbb{R}^n$.

Theorem 2.5.3 — Pythagoras theorem in $\mathbb{R}^n$. Let $x, y \in \mathbb{R}^n$ be perpendicular vectors, that is to say, $x^T y = 0$, then

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof. Using the properties of the inner product we have

$$\|x+y\|^2 = (x+y)^T(x+y) = x^T x + 2x^T y + y^T y = \|x\|^2 + \|y\|^2.$$

■

Theorem 2.5.4 — Properties of the norm in $\mathbb{R}^n$. Let $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, then

- $\|x\| \geq 0$ and it is zero if and only if $x = 0$, that is, the norm is non-negative,
- $\|\alpha x\| = |\alpha \cdot \|x\|$, that is, the norm is absolute homogeneous,
- $\|x+y\| \leq \|x\| + \|y\|$, that is, the norm fulfils the triangle inequality.[a]

[a] The triangle inequality is also called Minkowski inequality.

Proof. Let $x^T = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$, $y^T = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix} \in \mathbb{R}^n$ be arbitrary vectors, and $\alpha \in \mathbb{R}$ be an arbitrary scalar, then

$$\|x\| = \sqrt{x_1^2 + \dots + x_n^2},$$

which is a square root of a sum of squares, that is a sum of non-negative numbers, so the result must be non-negative too. This sum can be zero if and only if all of its members are zero, so all the x_is are equal to zero, which verifies the first part of the statement.

For the second part let us consider the norm

$$\|\alpha x\| = \sqrt{\alpha^2 x_1^2 + \dots + \alpha^2 x_n^2} = \sqrt{\alpha^2 (x_1^2 + \dots + x_n^2)} = |\alpha| \cdot \sqrt{x_1^2 + \dots + x_n^2} = |\alpha| \cdot \|x\|,$$

which results absolute homogeneity of the norm.

For the last part let us calculate the square of the left hand side of the inequality.

$$\|x+y\|^2 = (x+y)^T(x+y) = \|x\|^2 + 2x^T y + \|y\|^2 \leq \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2,$$

which gives the triangle inequality. Here we used the Cauchy-Schwarz inequality. ■

2.5.2 Gram-Schmidt orthogonalization

Definition 2.5.4 — Orthogonal and orthonormal systems. Let V be a vector space. A system of vectors $\{b_1, \dots, b_k\}$ is called **orthogonal** if b_i perpendicular to b_j if $i \neq j$, that is to say

$$b_i^T b_j = 0, \qquad \text{if} \quad i \neq j.$$

An orthogonal system is said to be **orthonormal** if $\|b_i\| = 1$ for every $i = 1, \dots, k$.

Orthonormal bases are especially important in applications because it is easy to calculate the coefficients of an arbitrary vector with respect to an orthonormal base.

■ **Example 2.31** The base $E_1,\dots,E_n \in \mathbb{R}^n$

$$E_1 = \begin{bmatrix}1\\0\\0\\\vdots\\0\\0\\0\end{bmatrix},\dots, \quad E_i = \begin{bmatrix}0\\\vdots\\0\\1\\0\\\vdots\\0\end{bmatrix} \leftarrow \text{ the } i\text{th coordinate},\dots \quad E_n = \begin{bmatrix}0\\0\\0\\\vdots\\0\\0\\1\end{bmatrix}$$

is said to be the **natural basis of** $\mathbb{R}^n$, which is an orthonormal basis of $\mathbb{R}^n$ as well. ■

■ **Example 2.32** The system

$$a_1 = \begin{bmatrix}1\\1\end{bmatrix}, \qquad a_2 = \begin{bmatrix}-1\\1\end{bmatrix}$$

is an orthogonal basis of $\mathbb{R}^2$. This system is not orthonormal, because the length of both vectors are $\sqrt{2}$. However the system

$$\frac{1}{\sqrt{2}}a_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix} \qquad \frac{1}{\sqrt{2}}a_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}$$

is an orthonormal basis of $\mathbb{R}^2$. ■

If $\{b_1,\dots,b_n\}$ is an arbitrary base in a Euclidean space, one can construct an orthonormal base $\{e_1,\dots,e_n\}$ with the aid of $\{b_1,\dots,b_n\}$ such that

$$\operatorname{span}\{e_1,\dots,e_l\} = \operatorname{span}\{b_1,\dots,b_l\}, \qquad \text{for all} \quad l = 1,\dots,n.$$

This is the **Gram-Schmidt orthogonalization process:**

Let $\{b_1,\dots,b_n\}$ be a base in the Euclidean space E.

1. step: Let

$$e_1 = \frac{b_1}{\|b_1\|}:$$

2. step: For $k = 2,\dots,n$

$$\tilde{e_k} = b_k - (b_k^T e_1)e_1 - \dots - (b_k^T e_{k-1})e_{k-1},$$

and

$$e_k = \frac{\tilde{e_k}}{\|\tilde{e_k}\|}:$$

■ **Example 2.33** Let us apply the Gram-Schmidt process to the following system of vectors!

$$b_1 = \begin{bmatrix}1\\2\\1\end{bmatrix}, \qquad b_2 = \begin{bmatrix}-1\\2\\0\end{bmatrix}, \qquad b_3 = \begin{bmatrix}2\\-1\\1\end{bmatrix}.$$

The norm of b_1 is $\|b_1\| = \sqrt{1^2+2^2+1^2} = \sqrt{6}$, so

$$e_1 = \frac{b_1}{\|b_1\|} = \frac{1}{\sqrt{6}}\begin{bmatrix}1\\2\\1\end{bmatrix}.$$

Using the formula for $\tilde{e_2}$ we have

$$\tilde{e_2} = b_2 - (b_2^T e_1)e_1 = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} - \left(\begin{bmatrix} -1 & 2 & 0 \end{bmatrix} \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right) \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix}.$$

The norm of $\tilde{e_2}$ is

$$\|\tilde{e_2}\| = \frac{1}{2}\sqrt{9+4+1} = \frac{\sqrt{14}}{2}.$$

So,

$$e_2 = \frac{\tilde{e_2}}{\|\tilde{e_2}\|} = \frac{2}{\sqrt{14}} \cdot \frac{1}{2} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{14}} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix}.$$

The formula for $\tilde{e_3}$ is

$$\tilde{e_3} = b_3 - (b_3^T e_1)e_1 - (b_3^T e_2)e_2,$$

which gives

$$\tilde{e_3} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} - \left(\begin{bmatrix} 2 & -1 & 1 \end{bmatrix} \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \right) \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} - \left(\begin{bmatrix} 2 & -1 & 1 \end{bmatrix} \frac{1}{\sqrt{14}} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} \right) \frac{1}{\sqrt{14}} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} =$$

$$= \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} - \frac{1}{6} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \frac{9}{14} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{21} \begin{bmatrix} -2 \\ -1 \\ 4 \end{bmatrix}.$$

The norm of $\tilde{e_3}$ is

$$\|\tilde{e_3}\| = \frac{1}{21}\sqrt{4+1+16} = \frac{\sqrt{21}}{21} = \frac{1}{\sqrt{21}}.$$

Which entails

$$e_3 = \frac{\tilde{e_3}}{\|\tilde{e_3}\|} = \sqrt{21}\frac{1}{21} \begin{bmatrix} -2 \\ -1 \\ 4 \end{bmatrix} = \frac{1}{\sqrt{21}} \begin{bmatrix} -2 \\ -1 \\ 4 \end{bmatrix}.$$

The required orthonormal system is

$$\frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \qquad \frac{1}{\sqrt{14}} \begin{bmatrix} -3 \\ 2 \\ -1 \end{bmatrix}, \qquad \frac{1}{\sqrt{21}} \begin{bmatrix} -2 \\ -1 \\ 4 \end{bmatrix}.$$

■

2.6 Eigenvalues, eigenspaces

Definition 2.6.1 — Eigenvalue, eigenvector. Let A be a square matrix. A scalar λ is called an **eigenvalue of** A if there exists a non-zero vector v such that

$$Av = \lambda v.$$

In this case v is said to be an **eigenvector belonging to** λ.

In the definition, the assumption $v \neq 0$ is crucial, because the defining equation above is fulfilled for an arbitrary scalar if $v = 0$.

■ **Example 2.34** Let us consider the two by two unit matrix

$$E_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

then all the vectors of $\mathbb{R}^2$ are eigenvectors of $\lambda = 1$. Indeed, let $x \in \mathbb{R}^2$, then

$$E_2 x = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \cdot x_1 + 0 \cdot x_2 \\ 0 \cdot x_1 + 1 \cdot x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1 \cdot x.$$

■

■ **Example 2.35** Let A be the following "rotation" matrix.

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

Then we will see later, that it has no real eigenvalue at all. ■

■ **Example 2.36** Let A be the matrix below.

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

Then

$$\begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 3 + 2 \cdot 1 \\ 1 \cdot 2 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 8 \\ 4 = \end{bmatrix} = 4 \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

So, $\lambda = 4$ is an eigenvalue of A and $v = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is eigenvector belonging to 4.

It is worthy to observe the fact that if p is an arbitrary real number then $p \cdot v$ is also an eigenvector of A belonging to 4. ■

Theorem 2.6.1 Eigenvectors belonging to the same eigenvalue constitute a subspace.

Proof. Assume that λ is an eigenvalue of A, v, w are eigenvectors belonging to λ and α is an arbitrary scalar.

Using the subspace criteria (Theorem 2.1.3), we have

$$A(v - w) = Av - Aw = \lambda v - \lambda w = \lambda (v - w),$$

which means that $v - w$ is also an eigenvector of A belonging to λ. Moreover,

$$A(\alpha v) = \alpha A v = \alpha(\lambda v) = \lambda(\alpha v)$$

which means that αv is also an eigenvector of A belonging to λ. These imply the statement. ■

The defining equation of eigenvalues and eigenvectors is

$$Av = \lambda v,$$

which can be written in the form

$$(A-\lambda E)v=0$$

after some rearrangement. Here E denotes the corresponding unit matrix. This is a homogeneous system of linear equations for the unknown vector v. The matrix of the equation is a quadratic one. The zero vector is always a solution of a system like this. This is called the trivial solution of the homogeneous system. A system like this has a non-trivial solution if and only if the determinant of its matrix is zero, that is to say, if

$$\det(A-\lambda E)\neq 0.$$

Expanding this determinant, the result will be a polynomial of λ.

Definition 2.6.2 — Characteristic polynomial. The polynomial p_A

$$p_A(t)=\det(A-tE)$$

is called the **characteristic polynomial of** A.

■ **Example 2.37** Let

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

then its characteristic polynomial p_A is

$$p_A(\lambda)=\det\begin{bmatrix}2-\lambda & 1\\ 2 & 3-\lambda\end{bmatrix}=(2-\lambda)(3-\lambda)-1\cdot 2=\lambda^2-5\lambda+4.$$

■

Because of the definition of the characteristic polynomial we get the following statement.

Theorem 2.6.2 The roots of the characteristic polynomial of a square matrix A are the eigenvalues of A.

■ **Example 2.38** Let A be like in the previous example, then its characteristic polynomial is $p_A(\lambda)=\lambda^2-5\lambda+4$. Its roots are

$$\lambda_{1,2}=\frac{5\pm\sqrt{25-16}}{2}=\frac{5\pm 3}{2}\quad\Rightarrow\quad\lambda_1=4,\ \lambda_2=1.$$

For the eigenspaces we have to solve the corresponding system of linear equations.

$$\lambda_1=4\quad\Rightarrow\quad\begin{bmatrix}2-\lambda_1 & 1\\ 2 & 3-\lambda_1\end{bmatrix}=\begin{bmatrix}2-4 & 1\\ 2 & 3-4\end{bmatrix}=\begin{bmatrix}-2 & 1\\ 2 & -1\end{bmatrix}\sim\begin{bmatrix}-2 & 1\\ 0 & 0\end{bmatrix}$$

Now, let $x_2=p$, then $x_1=\frac{1}{2}p$, so

$$x=p\begin{bmatrix}\frac{1}{2}\\ 1\end{bmatrix},\qquad p\in\mathbb{R}$$

is the eigenspace corresponding to $\lambda_1=4$. This is a one-dimensional subspace of $\mathbb{R}^2$, and $\begin{bmatrix}\frac{1}{2}\\ 1\end{bmatrix}$ constitutes a base in this subspace. A pretty similar calculation can be carried out for the other eigenvalue.

$$\lambda_2=1\quad\Rightarrow\quad\begin{bmatrix}2-\lambda_2 & 1\\ 2 & 3-\lambda_2\end{bmatrix}=\begin{bmatrix}2-1 & 1\\ 2 & 3-1\end{bmatrix}=\begin{bmatrix}1 & 1\\ 2 & 2\end{bmatrix}\sim\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix}$$

Now, let $x_2 = p$, then $x_1 = -p$, so

$$x = p \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \qquad p \in \mathbb{R}$$

is the eigenspace corresponding to $\lambda_2 = 1$. This is a one-dimensional subspace of $\mathbb{R}^2$, and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ constitutes a base in this subspace. ∎

2.7 Exercises

Exercise 2.1 Let

$$a=\begin{bmatrix}1\\1.23\\-2.4\\-8\end{bmatrix},\qquad b=\begin{bmatrix}1\\-2\\7.01\\3.03\end{bmatrix},\qquad u=\begin{bmatrix}1-i & 2+3i\end{bmatrix},\qquad w=\begin{bmatrix}i+3 & 2.1+0.1i\end{bmatrix}.$$

Find the value of the following expressions!

a) $a+b$, b) $a-4b$, c) $2a-3b$,

d) $u-w$, e) $u+5w$, f) $iu+(2-i)w$.

■

Exercise 2.2 Which one is a subspace, which one is not? Apply the subspace criteria!

a)
$$S=\left\{x\in\mathbb{R}^2\,\middle|\,x=\begin{bmatrix}t\\-t\end{bmatrix},\ t\in\mathbb{R}\right\}$$

b)
$$S=\left\{x\in\mathbb{R}^2\,\middle|\,x=\begin{bmatrix}t\\t+1\end{bmatrix},\ t\in\mathbb{R}\right\}$$

c)
$$S=\left\{x\in\mathbb{R}^3\,\middle|\,x=\begin{bmatrix}0\\1\\t\end{bmatrix},\ t\in\mathbb{R}\right\}$$

d)
$$S=\left\{x\in\mathbb{R}^3\,\middle|\,x=\begin{bmatrix}t\\-t\\s\end{bmatrix},\ t,s\in\mathbb{R}\right\}$$

e)
$$S=\left\{\,p\in\mathscr{P}\,\middle|\,p(t)=at^2+bt,\ a,b\in\mathbb{R}\right\}$$

f)
$$S=\left\{p\in\mathscr{P}\middle|p(t)=at^2+bt+1,\ a,b\in\mathbb{R}\right\}$$

■

Exercise 2.3 Find the linearly dependent and independent pairs x,y where

a) $x=(1,2)$, $y=(-2,-4)$, $x,y\in\mathbb{R}^2$, b) $x=(3,4)$, $y=(4,3)$, $x,y\in\mathbb{R}^2$,

c) $x=(1,1)$, $y=(-2,-1)$, $x,y\in\mathbb{R}^2$, d) $x=(1,1)$, $y=(4,4)$, $x,y\in\mathbb{R}^2$.

■

Exercise 2.4 Find the linearly independent systems in $\mathbb{R}^3$!

a)

$$x=\begin{bmatrix}1\\1\\1\end{bmatrix},\quad y=\begin{bmatrix}1\\2\\3\end{bmatrix},\quad z=\begin{bmatrix}-1\\0\\1\end{bmatrix}.$$

b)

$$x=\begin{bmatrix}-1\\0\\2\end{bmatrix},\quad y=\begin{bmatrix}1\\0\\2\end{bmatrix},\quad z=\begin{bmatrix}1\\0\\1\end{bmatrix}.$$

c)

$$x=\begin{bmatrix}0\\1\\10\end{bmatrix},\quad y=\begin{bmatrix}10\\1\\0\end{bmatrix},\quad z=\begin{bmatrix}0\\10\\1\end{bmatrix}.$$

d)

$$x=\begin{bmatrix}-1\\2\\2\end{bmatrix},\quad y=\begin{bmatrix}2\\2\\3\end{bmatrix},\quad z=\begin{bmatrix}3\\3\\3\end{bmatrix}.$$

■

Exercise 2.5 Find the linearly independent systems in $\mathscr{P}$!

a)

$$p_1(t)=t+1,\quad p_2(t)=t^2+1,\quad p_3(t)=2,$$

b)

$$p_1(t)=t+1,\quad p_2(t)=t$$

c)

$$p_1(t)=3t+1,\quad p_2(t)=t^3+1,\quad p_3(t)=2t,$$

d)

$$p_1(t)=t+1,\quad p_2(t)=t-1,\quad p_3(t)=1.$$

■

Exercise 2.6 Let

$$A=\begin{bmatrix}1&2&3&4\\1&2&-1&-2\end{bmatrix},\quad B=\begin{bmatrix}2&-3&0&1\\0&1&0&-4\end{bmatrix},\quad C=\begin{bmatrix}0.1&2.3\\-1.1&2.03\\9.04&10.36\end{bmatrix},\quad D=\begin{bmatrix}2.22&1.1\\0.01&0.2\\1&3\end{bmatrix}.$$

Find the values of the following expressions!

a) $A+B$,

b) $3A-2B$,

c) A^T-B^T,

d) $10C+D$,

e) $100D$,

f) C^T-10D^T.

■

Exercise 2.7 Calculate the products A^TB, AB^T, CD, CB, A^TD, where

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad B = \begin{bmatrix} -1 \\ i \\ 1-2i \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 5 \\ -1 & i & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 4 & -i \\ 3 & 2 \\ i & i+2 \end{bmatrix}$$

Exercise 2.8 Solve the following systems of linear equations!

a)

$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 4 \\ x_1 - x_2 + x_3 + x_4 &= 8 \\ 3x_1 + x_2 + x_3 - x_4 &= 16 \end{aligned}$$

b)

$$\begin{aligned} x_1 + 3x_2 + x_3 - x_4 &= 7 \\ 2x_1 + 5x_2 - x_3 + 2x_4 &= 22 \\ 3x_1 + 8x_2 + x_3 - x_4 &= 24 \end{aligned}$$

c)

$$\begin{aligned} x_1 - 2x_2 + 3x_3 - x_4 + 2x_5 &= 2 \\ 3x_1 - x_2 + 5x_3 - 3x_4 - x_5 &= 6 \\ 2x_1 + x_2 + 2x_3 - 2x_4 - 3x_5 &= 8 \end{aligned}$$

d)

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1 \\ 2x_1 + 3x_2 + 3x_3 - x_4 &= 3 \end{aligned}$$

e)

$$x_1 + x_2 - x_3 = 2$$

f)

$$\begin{aligned} x_1 + x_2 - x_3 &= 2 \\ 2x_1 + 2x_2 - 2x_3 &= 3 \end{aligned}$$

Exercise 2.9 Find the inverse of the following matrices!

a) $\begin{bmatrix} 1 & -1 \\ 2 & 2 \end{bmatrix}$ b) $\begin{bmatrix} 1 & i \\ 2 & i \end{bmatrix}$ c) $\begin{bmatrix} 1 & 3 \\ 2 & -2 \end{bmatrix}$ d) $\begin{bmatrix} 5 & 0 \\ 9 & 1 \end{bmatrix}$

a)
$$\begin{bmatrix} 1 & 1 & 2 \\ -1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

b)
$$\begin{bmatrix} 2 & 0 & -2 \\ -1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

c)
$$\begin{bmatrix} 6 & 0 & 3 \\ -4 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

■

Exercise 2.10 Find the determinant of the following matrices!

a)
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

b)
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

c)
$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

d)
$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

e)
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

f)
$$\begin{bmatrix} 2 & 2 & 1 \\ 1 & 1 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$$

g)
$$\begin{bmatrix} i & 1 & -i \\ 2 & 4 & 1 \\ 3i & -2i & 1 \end{bmatrix}$$

h)
$$\begin{bmatrix} 1-i & 0 & 2 \\ 0 & -1-i & 3 \\ \pi & 0 & 1 \end{bmatrix}$$

i)
$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & -1 & 0 & 2 \\ 4 & 2 & 0 & -2 \\ 10 & 0 & 0 & -3 \end{bmatrix}$$

■

Exercise 2.11 Find the the norm of x and y, the inner product $x^T y$ if

a)
$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad y = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 3 \end{bmatrix}$$

b)
$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad y = \begin{bmatrix} 10 \\ -10 \\ 20 \end{bmatrix}$$

c)
$$x = \begin{bmatrix} 8 \\ 8 \\ -2 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

■

Exercise 2.12 Apply the Gram-Schmidt orthogonalization technic to the following systems!

a)
$$b_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

b)
$$b_1 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

c)
$$b_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad b_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

a)
$$b_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}, \quad b_3 = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}.$$

b)
$$b_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad b_3 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}.$$

■

Exercise 2.13 Find the eigenvalues and eigenspaces of the following matrices.

a)
$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

b)
$$\begin{bmatrix} -2 & 3 \\ -4 & 5 \end{bmatrix}$$

c)
$$\begin{bmatrix} -2 & -3 \\ 1 & 1 \end{bmatrix}$$

d)
$$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

■

3. Basics of numerical mathematics

3.1 Machine representation of numbers

It is important to emphasise the following crucial fact: *Not all real numbers are representable by a computer, only a finite subset of the reals can be represented.* This entails rounding errors of numerical computations. Without care these can result in serious problems in real life.[1]

3.1.1 Normal form of numbers

A reminder from high school: Let us consider the following number 234.5189. This is the **decimal representation** of this number. In detail

$$234.5189 = 2\cdot 10^2 + 3\cdot 10^1 + 4\cdot 10^0 + 5\cdot 10^{-1} + 1\cdot 10^{-2} + 8\cdot 10^{-3} + 9\cdot 10^{-4},$$

where the digits multiplied by the corresponding powers of the base of the numeral system 10, and the place value of the digits depend on the exponent. If we push this number into the interval $[0,1[$, the we get

$$234.5189 = +10^3 \left(\underbrace{\frac{2}{10} + \frac{3}{10^2} + \frac{4}{10^3} + \frac{5}{10^4} + \frac{1}{10^5} + \frac{8}{10^6} + \frac{9}{10^7}}_{\text{mantissa}} \right).$$

If the base is fixed in advance, it is enough to keep the sign (+), the exponent (3) and the digits.

+	3	2	3	4	5	1	8	9

All real numbers have a finite or an infinite decimal representation. However, one can choose another base number, an arbitrary integer which is greater than 1. If the base is different from 10 one can give a similar representation in a pretty similar way like in the case of 10 using the powers of the new base.

From numerical point of view, the most useful base numbers are 10 (decimal system), 2 (binary system), and 16 (hexadecimal system).

[1] See for example the following webpage: http://www-users.math.umn.edu/ arnold//disasters/

Definition 3.1.1 — Normal form of decimal numbers. Let x be a real number with finite decimal representation, that is to say, there are $d_n, d_{n-1}, \ldots, d_1, d_0, d_{-1}, \ldots, d_{-l}$ digits (numbers from the set $\{0,1,2,3,4,5,6,7,8,9\}$ such that d_n and d_{-l} are different from zero, and

$$x = \pm \sum_{i=-l}^{n} d_i 10^i,$$

then the form

$$\boxed{\pm} \boxed{n+1} \boxed{d_n} \boxed{\ldots} \boxed{d_{-l}} = \pm 10^{n+1} \left(\frac{d_n}{10} + \cdots \frac{d_{-l}}{10^{n+l+1}} \right)$$

is said to be the **normal form of** x. The number $n+1$ is called the **characteristic of** x, and the sequence of the digits is called the **mantissa of** x.

In a computer, only finite number of digits can be representable. The length of the mantissa and the possible range of the characteristic (upper and lower bond) depend on the computer capabilities, but they are fixed. This entails that **not all real numbers can be representable**. A reasonable requirement is the representation of a finite number (this can be quite big) of real numbers given in a certain range. In other words, there is a number which is the biggest representable number, its negative is the smallest representable number. Even between these two, not all the numbers are representable.

If x is an arbitrary real number it is possible to give arbitrarily close to x a number which has a finite decimal representation.

Proposition 3.1.1 Let $x \in \mathbb{R}$, then for every $\varepsilon > 0$ there are digits $d_n, \ldots, d_{-l}$ such that

$$\left| x - \sum_{i=-l}^{n} d_i 10^i \right| \leq \varepsilon.$$

This statement legitimates the (approximate) computer representation of numbers.

Representation in other number systems

It is possible to choose other base numbers for the representation. In numerics the most frequent choices, besides the decimal representation, are the dyadic representation (base number is 2) and hexadecimal representation (base number is 16). In the first case the only two digits (we have) are 0 and 1, in the latter case we have sixteen digits: $0,1,\ldots,9,A,B,C,D,E,F$. However, all positive integers greater than one are feasible as a base number, and number of digits as well.[2]

Definition 3.1.2 — Normal form of numbers. Let a be a positive integer greater than one, and let x be a real number with finite $a$dic representation, that is to say, there are

$$m_n, m_{n-1}, \ldots, m_1, m_0, m_{-1}, \ldots, m_{-l}$$

digits (numbers from the set $\{0,1,\ldots,a-1\}$ such that m_n and m_{-l} are different from zero, and

$$x = \pm \sum_{i=-l}^{n} m_i a^i,$$

[2] A good and well-known example for this phenomenon is the system based on 60. This numeral system is originated with the ancient Sumerians, and it is still used for measuring time, angles, and geographic coordinates.

then the form

$$\boxed{\pm}\boxed{n+1}\boxed{m_n}\boxed{\ldots}\boxed{m_{-l}}_a = \pm a^{n+1}\left(\frac{m_n}{a}+\cdots\frac{m_{-l}}{a^{n+l+1}}\right)$$

is said to be the a**dic normal form of** x. The number $n+1$ is called the **characteristic of** x, and the sequence of the digits is called the **mantissa of** x.

■ **Example 3.1** Let us find the dyadic representation of 13.8125.

$$13 = 8+4+1 = 1101_2.$$

$$0.8125 = \frac{1}{2}+\frac{1}{4}+\frac{1}{16}.$$

So,

$$13.8125_10 = 1101.1101_2.$$

The normal forms in the decimal and in the dyadic systems are:

$$13.8125 = 10^2\cdot 0.138125, \qquad \text{and} \qquad 1101.1101 = 2^4\cdot 0.11011101.$$

■

3.1.2 Floating-point numbers

Assume that we have t positions to store a number. The first position is the representation of the sign of the number, $t-l-1$ positions for the representation of digits from the left of the (decimal, dyadical and so on) point and l positions for the representation of digits from the right of the (decimal, dyadical and so on) point. This is called the **fixed-point representation of numbers**.

If we let the (decimal, dyadical and so on) point move we get the **floating-point representation** of numbers. This is similar to the normal form, but here not all finitely representable numbers can be stored, because the length of the mantissa is given (t).

Definition 3.1.3 — Floating-point numbers. Let t, a be positive integers $a > 1$, ℓ and u be integers (typically $\ell < 0 < u$). Then the numbers

$$\boxed{\pm}\boxed{n}\boxed{m_1}\boxed{\ldots}\boxed{m_t}_a = \pm a^n\left(\frac{m_1}{a}+\cdots+\frac{m_t}{a^t}\right)$$

are called **Floating-point numbers**, where

$$0 < m_1 \le a-1, \qquad 0 \le m_i \le a-1, \quad i = 2,3,\ldots,n, \qquad \ell \le n \le u.$$

The number a is the **base of the representation**, t is the **length of the mantissa**, ℓ, and u are the **lower and upper bounds for the exponent** n respectively.

If there is no ambiguity, we skip the subscript a.

■ **Example 3.2** Let $a = 10$, $t = 4$, $\ell = -3$, and $u = 3$. Give the floating-point representation of 10.32.

$$10.32 = +10^2\left(\frac{1}{10}+\frac{0}{10^2}+\frac{3}{10^3}+\frac{2}{10^4}\right) = \boxed{+}\boxed{2}\boxed{1}\boxed{0}\boxed{3}\boxed{2}_{10}$$

■

■ **Example 3.3** Let $a = 2$, $t = 4$, $\ell = -3$, and $u = 3$. Give the floating-point representation of 3.25. We have only two digits: 0 and 1. At first we transform the number from the decimal system

into the dyadic system, and the dot denotes the dyadic dot instead of the decimal dot. We do not use a different notation for the $a$dic dots in different number systems.

$$3.25_{10} = 11.01_2 = +2^2\left(\frac{1}{2}+\frac{1}{2^2}+\frac{0}{2^3}+\frac{1}{2^4}\right) = \boxed{+}\boxed{2}\boxed{1}\boxed{1}\boxed{0}\boxed{1}_2$$

■

Distribution of Floating-point numbers

If a,t,ℓ and u are given, then we can calculate the largest and the smallest positive representable numbers. All positive representable numbers are within this range. Reflecting this set, with respect to zero, we get the set of negative representable numbers. Because of this phenomenon, it is enough to deal with positive representable numbers.

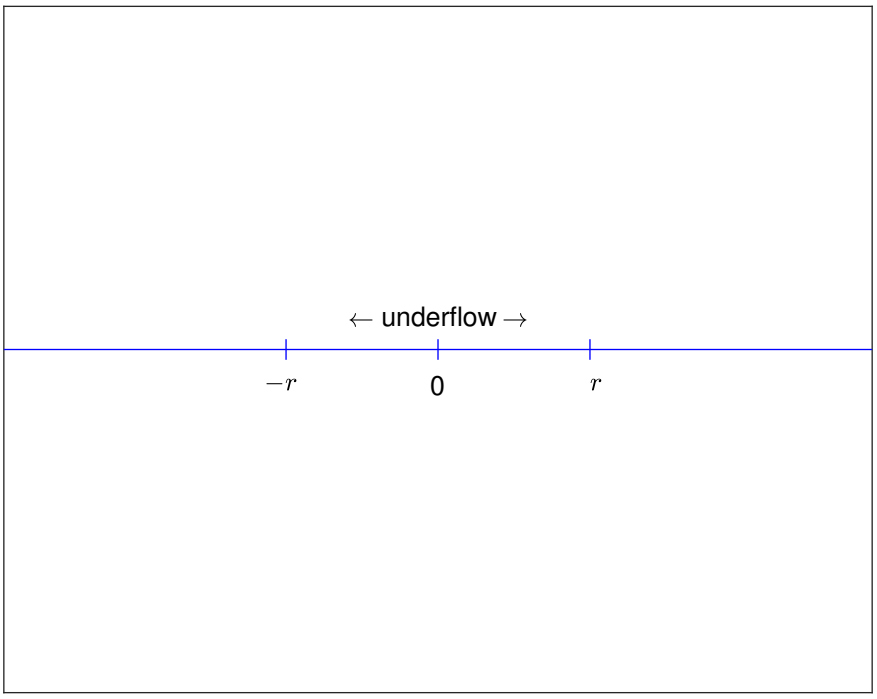
The **smallest positive representable number** is

$$r = +a^{\ell}\left(\frac{1}{a}+\frac{0}{a^2}+\cdots+\frac{0}{a^t}\right) = a^{\ell-1}.$$

The **largest representable number is**

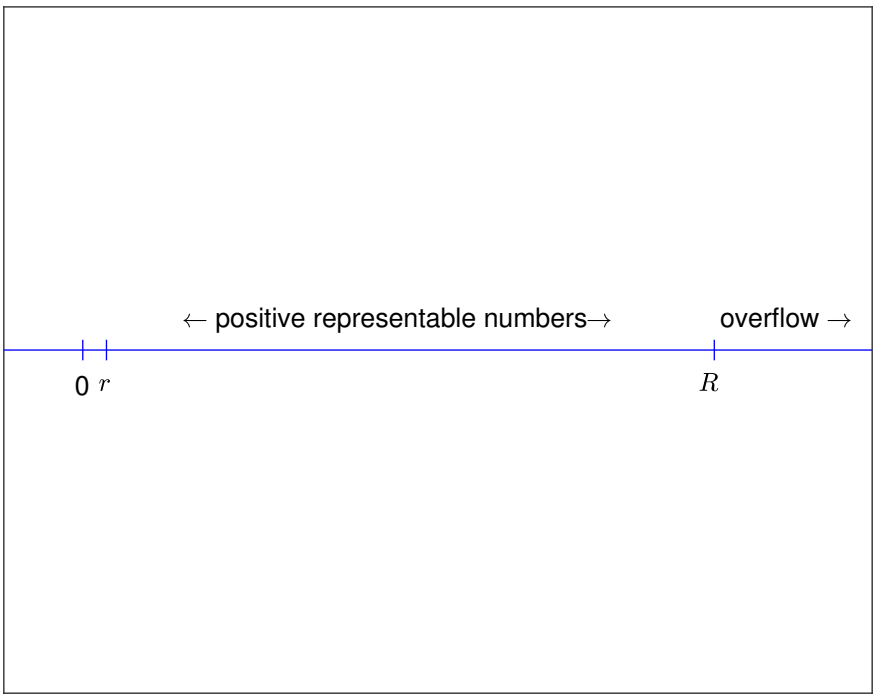
$$\mathscr{R} = +a^{u}\left(\frac{a-1}{a}+\frac{a-1}{a^2}+\cdots+\frac{a-1}{a^t}\right) = a^{u}(1-a^{-t}).$$

All representable positive numbers are between these two numbers, but they are not equally spaced.



The 'hole' around zero

They get dense close to r and their number is getting less and less as $\mathscr{R}$ is approaching.



The area of positive representable numbers

Strictly between $-r$ and r the only representable number is zero, this "hole" can lead to serious drawbacks during numerical calculations.

Let $\mathbb{F}$ denote the set of representable numbers. It is easy to check that for every $x \in \mathbb{F}$

$$a^{-1}\varepsilon_M|x| \leq |x-y| \leq \varepsilon_M|x|,$$

where $y \in \mathbb{F}$ is the next nearest to y, and $\varepsilon_M = a^{1-t}$ is the **machine epsilon**. This latter one is the distance between one and the next nearest floating-point number to one.

In every interval

$$[a^{k-1}, a^k[, \qquad k = \ell, \ldots e(u,t), \quad \text{where the integer } e(u,t) \text{ depends on } u \text{ and } t$$

there are equal number of representable floating-point numbers $(a-1)a^{t-1}$. This causes the inequable distribution (with bigger and bigger gaps) of representable numbers.

■ **Example 3.4** Let $a = 2, t = 4, \ell = -4$, and $u = 3$. Find the smallest and the largest representable floating-point number. What is the value of the machine epsilon? How many positive numbers are representable?

$$a^{\ell-1} = 2^{-4-1} = \frac{1}{32} \quad \text{is the smallest representable positive floating-point number.}$$

$$a^u(1-a^{-t}) = 2^3(1-2-4) = 8\frac{15}{16} = \frac{15}{8} = 7.5 \quad \text{is the largest representable positive floating-point number.}$$

$$\varepsilon_M = a^{1-t} = 2^{1-4} = \frac{1}{8} = 0.125 \quad \text{is the machine epsilon.}$$

We have eight different mantissas. The first digit is always one and we have two choices for the remaining three respectively, which gives 2^3 possibilities.

The number of the exponents are $u - \ell + 1 = 3 - (-4) + 1 = 8$, where $+1$ is for the zero exponent.

All in all, we have $8 \cdot 8 = 64$ different representable, positive numbers. ∎

3.1.3 Rounding and truncation

The set $\mathbb{F}$ is always finite, however, even in the interval $[\nu, \mathcal{R}]$ there are infinitely many real numbers. The representation on an arbitrary real number x in $\mathbb{R} \setminus \mathbb{F}$ is an important practical problem . Furthermore, even if $x, y \in \mathbb{F}$, their sum, product and so on, do not necessary belong to $\mathbb{F}$.

Definition 3.1.4 Let $\nu \leq |x| \leq \mathcal{R}$ be a real number. Then $\tilde{x} \in \mathbb{F}$ is called the **rounded floating-point representation of** x if $\tilde{x}$ is the closest floating-point number to x. $\tilde{x} \in \mathbb{F}$ is called the **truncated floating-point representation of** x if $\tilde{x}$ is the closest floating-point number to x, which is less than x.

If $|x| \leq \nu$ (x is in the 'hole' around zero), then $\tilde{x}$ is equal to zero, even if x is different from zero. This phenomenon is said to be the **underflow**.

If $|x| > \mathcal{R}$, then there is no floating point representation of x, this is called the **overflow**.

If $x \in \mathbb{F}$ then $\tilde{x} = x$ both with rounding and truncation.

∎ **Example 3.5** Let $a = 2$, $t = 4$, $\ell = -3$, and $u = 3$. Find the floating point representation of 2.625 and $\frac{1}{3}$.

$$2.625 = 2 + \frac{1}{2} + \frac{1}{8} = 2^2 \left(\frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} + \frac{1}{2^5} \right).$$

The length of the mantissa is 4, so there is no room for all the digits. We have to choose an approximate value.

With truncation we just omit the last digit, so we have

$$2.625 \approx 2^2 \left(\frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} \right) = \boxed{+} \boxed{2} \boxed{1} \boxed{0} \boxed{1} \boxed{0} = 2.5.$$

The next representable number is 2.75, so 2.625 is exactly the same length both from 2.5 and 2.75. So, both are as good as the rounded approximation for 2.625. ∎

During arithmetic operations with floating-point numbers, the resulted error (caused by rounding or truncation) can be accumulated. This can imply serious mistakes at the end of the calculation. Because of this, one has to design a numerical algorithm with a great care to avoid such failures. For this reason, it is necessary to know more about the spreading of error or error analysis of arithmetic operations. This is far beyond the scope of this work. The interested reader can find detailed description of this in the literature (see e.g. the References at the end of this book).

3.2 Non-linear system of equations

The focus of this part is the following explicit equation

$$F(x) = 0, \tag{3.1}$$

where F is a given real valued function and x is the unknown. If a variable is a solution of (3.1), then we denote it by $\bar{x}$.

■ **Example 3.6** Find a real number x, which sine and square are the same, that is to say, solve the following non-linear equation:

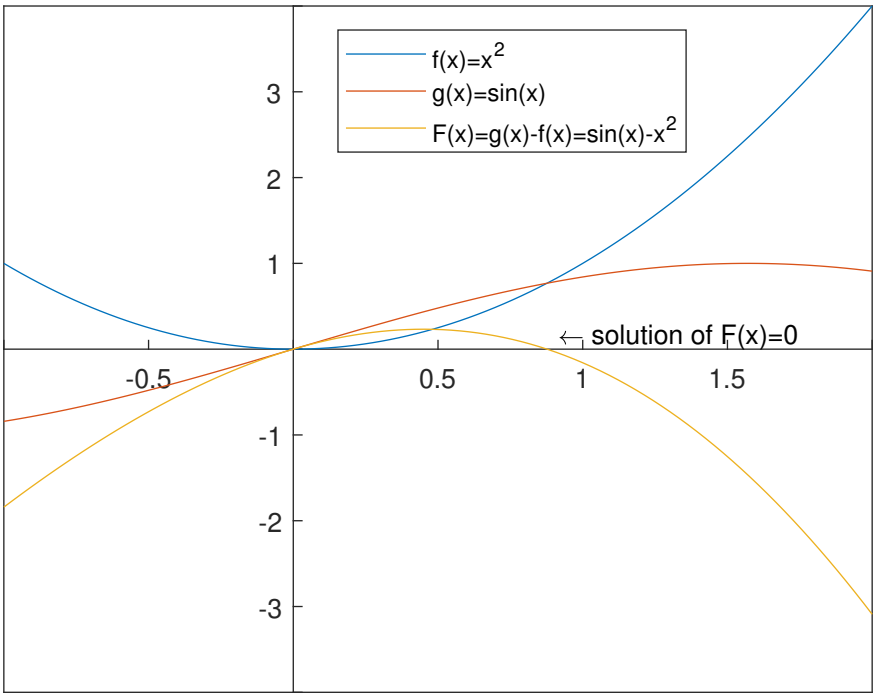
$$\sin x = x^2.$$

Let $F : \mathbb{R} \to \mathbb{R}$, $F(x) = \sin x - x^2$, then we have to solve the following problem

$$\sin x - x^2 = F(x) = 0. \tag{3.2}$$

Here F is a one-variable function, and for the solution of the problem we should solve the non-linear equation above.

This problem has a trivial solution $x = 0$, but it also has an other solution (see the figure below). It is not possible to determine this second solution explicitly. However, numerical methods can be used to find an approximation of it. ■



Graphical solution of (3.2)

■ **Example 3.7** Find the solution of the following non-linear system!

$$\begin{aligned} x_1^2 + x_2^2 &= 1 \\ x_1^3 - x_2 &= 0 \end{aligned}$$

This is equivalent to (3.1), where

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \qquad F(x_1, x_2) = (x_1^2 + x_2^2 - 1, x_1^3 - x_2).$$

■

3.2.1 Main properties of numerical algorithms

Concerning numerical methods, there are some properties which classify their quality and their level of performance. The most important features are:

- rate of convergence;
- local or global convergence;
- wideness of the class where the method is applicable.

The first property characterize the speed of the method.

Definition 3.2.1 — Rate of the convergence. Let $\{x_n\}_{n\in\mathbb{N}}$ be a real sequence, which converges to $\bar{x}$. If there exists a constant $c \in]0,1[$ such that

$$\lim_{n\to\infty} \frac{|x_{n+1}-\bar{x}|}{|x_n-\bar{x}|} = c,$$

then we say that x_n **tends to** $\bar{x}$ **linearly**.

If

$$\lim_{n\to\infty} \frac{|x_{n+1}-\bar{x}|}{|x_n-\bar{x}|} = 1,$$

then we say that x_n **tends to** $\bar{x}$ **sublinearly**.

If

$$\lim_{n\to\infty} \frac{|x_{n+1}-\bar{x}|}{|x_n-\bar{x}|} = 0,$$

then we say that x_n **tends to** $\bar{x}$ **superinearly**.

Moreover, if there is a number $p > 1$ such that

$$\lim_{n\to\infty} \frac{|x_{n+1}-\bar{x}|}{|x_n-\bar{x}|^p} < c,$$

then we say that **the sequence converges with order** p **to** $\bar{x}$.

The case $p = 2$ is called **quadratic convergence**.

One can define the rate of convergence in higher dimension (in $\mathbb{R}^n$) using the norm instead of the absolute value. However, it is also enough to investigate the rate of the convergence of the partial sequences, which is real. In this case, for example linear rate means linear rate of the partial sequences (see the examples below).

It is clear that the worst rate (the slowest one) is sublinear convergence, the linear convergence is better, and superlinear convergence is better than linear convergence. If p is increasing we get a better and better rate. In practice quadratically convergent methods are good.

■ **Example 3.8** Let

$$x_n = \frac{1}{n}, \qquad y_n = \frac{1}{2^n}, \qquad z_n = \frac{1}{n!}, \qquad v_n = \frac{1}{3^{3^n}},$$

then all these sequences tend to zero. Because their elements are positive, we can omit the absolute value. In point of fact, we have to investigate the following ratios:

$$\frac{x_{n+1}}{x_n} = \frac{n}{n+1} \to 1, \qquad \frac{y_{n+1}}{y_n} = \frac{2^n}{2^{n+1}} = \frac{1}{2}, \qquad \frac{z_{n+1}}{z_n} = \frac{n!}{(n+1)!} \to 0, \qquad \frac{v_{n+1}}{v_n^2} = \frac{1}{3^{3^n}} \to 0.$$

So, x_n converges sublinearly, y_n converges linearly, z_n converges superlinearly, and v_n converges quadratically to zero. ■

Definition 3.2.2 An algorithm is called **globally convergent** if starting from any point of the iteration the resulted sequence tends to a solution. Otherwise the method is said to be **locally convergent**.

Local convergence means that the first point of the iteration should be close enough to a solution, if not, it can happen that the method produces a divergent sequence.

The third property, how wide the class of problems is where the method works, cannot be defined in an exact way. For example, some methods are applicable for continuous functions, others can be applied only for differentiable functions and so on.

3.2.2 Non-linear equations

We start with the one-variable case, which is more transparent than the multi-variables case. Therefore it will be easier to understand the main ideas and methods of this topic.

Bisection method

This method is suitable for solving only non-linear equations, there is no generalization for solving multi-variables problems. The reason why we deal with this algorithm is its simplicity, easy implementability, and global convergence.

Let F be a continuous function on a bounded interval $[a,b]$ such that it changes its sign at the end of the interval, that is to say

$$F(a)F(b) < 0 \Leftrightarrow \text{either } F(a) < 0 \text{ and } F(b) > 0 \text{ or } F(a) > 0 \text{ and } F(b) < 0.$$

For example the function on the figure below fulfils the previously mentioned requirements on the interval $[0.5, 2.5]$.

Reminder from calculus: Bolzano's theorem says that if a real function F is continuous on a bounded, closed interval $a,b]$, then it takes all the values between $F(a)$ and $F(b)$.

The consequence of this theorem is that if $F(a)$ is positive and $F(b)$ is negative, then zero is an intermediate value, so there is ξ between a and b, where the function takes zero, that is to say, ξ is a solution of the equation $F(x) = 0$.

Using these facts, we can build an algorithm.

The pseudo code of the algorithm is the following.

Initialization: a, b, F, and

$$x_0 = a, \qquad y_0 = b, \qquad m_0 = \frac{x_0 + y_0}{2}$$

Step 1.: If $F(m_k) = 0$, then m_k is a solution. Otherwise Step 2.

Step 2.: Let

$$m_k = \frac{x_k + y_k}{2}.$$

If $F(x_k)F(m_k) < 0$, then

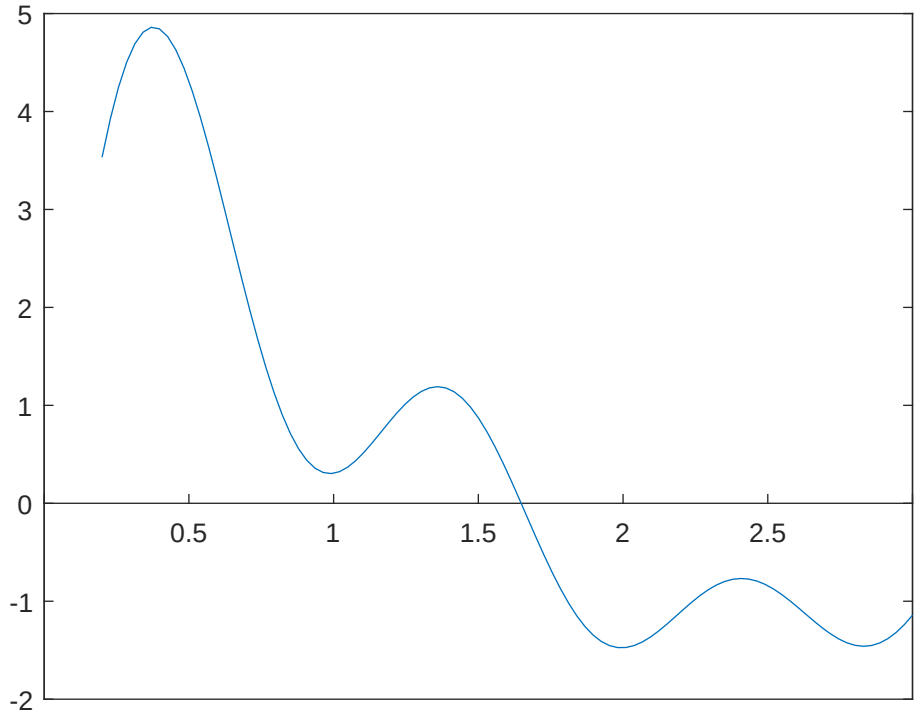
$$x_{k+1} = x_k, \qquad y_{k+1} = m_k.$$

Otherwise,

$$x_{k+1} = m_k, \qquad y_{k+1} = y_k.$$

$k = k+1$. Step 1.

In practice, in the Initialization part we also give a positive ε, and we require $|F(m_k)| \leq \varepsilon$ instead of $F(m_k) = 0$. It is also convenient to give the maximum number of steps in advance.



Bisection method is applicable

Proposition 3.2.1 Besides the required condition of the bisection method, it is globally convergent, and the rate of the convergence is linear.

■ **Example 3.9** Find an approximation of the root in the interval $[0,1]$ of the polynomial $F(x) = x^3 - 2x^2 - x + 1$!

This function is continuous, $F(0) = 1$ and $F(1) = -1$, so we can apply the bisection method. Let $x_0 = 0$, $y_0 = 1$, then $m_0 = \frac{1}{2}$, and

$$F(x_0) = 1, \qquad F(m_0) = \frac{1}{8}, \qquad F(y_0) = -1.$$

$F(y_0)F(m_0) < 0$, so

$$x_1 = m_0 = \frac{1}{2}, \qquad m_1 = \frac{3}{4}, \qquad y_1 = y_0 = 1,$$

and

$$F(x_1) = \frac{1}{8}, \qquad F(m_1) = -\frac{29}{64}, \qquad F(y_1) = -1.$$

$F(x_1)F(m_1) < 0$, so

$$x_2 = x_1 = \frac{1}{2}, \qquad m_2 = \frac{5}{8}, \qquad y_2 = m_1 = \frac{3}{4},$$

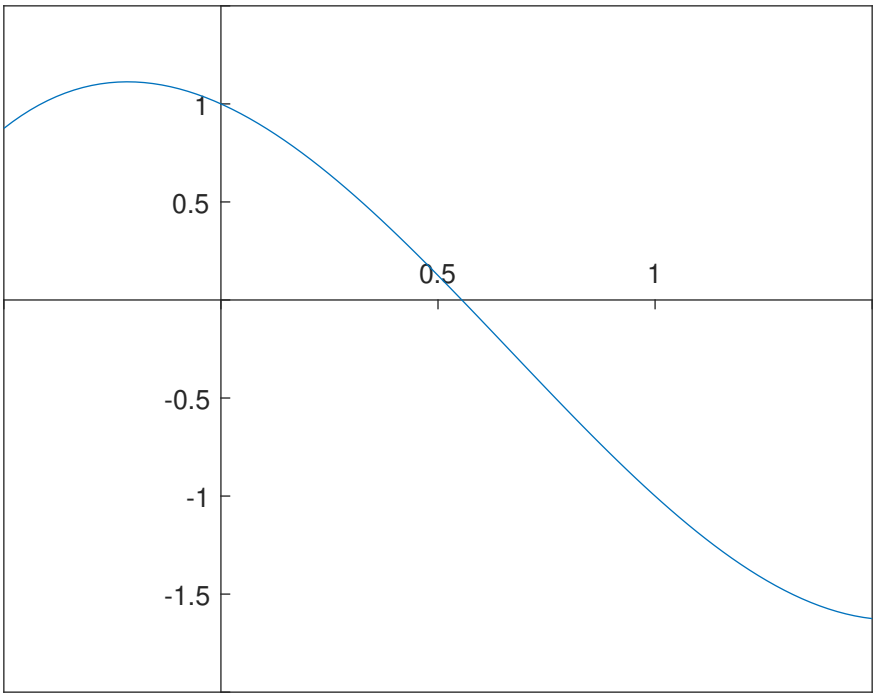
and

$$F(x_2) = \frac{1}{8}, \qquad F(m_2) = -\frac{83}{512}, \qquad F(y_2) = -\frac{29}{64}.$$

Here

$$F(m_2) = -\frac{83}{512} \approx -0.0652, \qquad m_2 = 0.625$$

which is an acceptable approximation of the root in the interval $[0,1]$. ∎



The root of the polynomial $F(x) = x^3 - 2x^2 - x + 1$ in the interval $[0,1]$.

Newton's method in $\mathbb{R}$

To get a better rate than the rate of the bisection method, we need the derivative of the function. The price is the abated competency and local convergence instead of a global one.

For this purpose, let us use the first order Taylor expansion of F around $\bar{x}$ a solution of the problem (3.1). This gives the following linearized version of the problem

$$F(x) = F(\bar{x}) + F'(\xi)(x - \bar{x}), \tag{3.3}$$

where ξ is somewhere in between x and $\bar{x}$.

Because $\bar{x}$ is a solution of (3.1) $F(\bar{x}) = 0$, after some rearrangement, assuming that $F'(\xi) \neq 0$, we have

$$\bar{x} = x - \frac{F(x)}{F'(\xi)}$$

for some ξ, which depends on x. Assuming continuity of the derivative, if x is not too far from $\bar{x}$, then the following expression is a reasonable approximation of the solution:

$$\bar{x} \approx x - \frac{F(x)}{F'(x)}.$$

This formula suggests the iteration scheme below.

Initialization: F, x_0, and $k = 0$

Step 1.: If $F(x_k) = 0$, then x_k is a solution, otherwise Step 2.

Step 2.: Let

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}. \qquad \text{(Newton iteration)}$$

Step 1.

In practice, it is usual to initialize an error $\varepsilon > 0$ and the maximum number of steps n, and in the first step it is better to check $|F(x_k)| \leq \varepsilon$ instead of $F(x_k) = 0$.

Newton's method is usually quadratically convergent, but not globally convergent. It can happen that the iteration is divergent if the starting point is too far from the solutions of the problem.

■ **Example 3.10** Find the approximate value of $\sqrt{2}$ using Newton's method starting from $x_0 = 1$!

Let

$$F(x) = x^2 - 2.$$

Then $\sqrt{2}$ is a solution of $F(x) = 0$, which is "close" to 1. Because $F'(x) = 2x$, we have the following iteration.

$$\begin{aligned} x_0 &= 1 \\ x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} &= 1 - \frac{-1}{2} = \frac{3}{2} \\ x_2 = x_1 - \frac{F(x_1)}{F'(x_1)} &= \frac{3}{2} - \frac{\frac{1}{4}}{3} = \frac{17}{12} \\ x_3 = x_2 - \frac{F(x_2)}{F'(x_2)} &= \frac{17}{12} - \frac{\frac{1}{144}}{\frac{17}{6}} = \frac{577}{408} \approx 1.414215 \end{aligned}$$

Here the first five digits are correct after the decimal dot. ■

Theorem 3.2.2 — Properties of Newton's method. Let $F\colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function, $\bar{x} \in \mathbb{R}$ is a solution of (3.1), where the derivative is non-zero $F'(\bar{x}) \neq 0$. Then there are $\varepsilon > 0$ and $C > 0$ constants such that

- $\bar{x}$ is the unique solution of problem (3.1) in the interval $]\bar{x} - \varepsilon, \bar{x} + \varepsilon[$;
- $|(F'(x))^{-1}| \leq C$ for all $x \in |x - \bar{x}| \leq \varepsilon$;
- for every $x_0 \in]\bar{x} - \varepsilon, \bar{x} + \varepsilon[$ the Newton iteration either stops after finitely many steps on $\bar{x}$ or results a superlinearly convergent sequence with limit $\bar{x}$;
- if F' is a Lipschitz continuous function on the interval $]\bar{x} - \varepsilon, \bar{x} + \varepsilon[$ with constant L, that is

$$|F'(x) - F'(y)| \leq L|x - y|, \qquad x, y \in]\bar{x} - \varepsilon, \bar{x} + \varepsilon[,$$

then the rate of the convergence of Newton iteration is quadratic, that is

$$|x_{n+1} - x_n| \leq \frac{CL}{2}|x_n - \bar{x}|^2, \qquad n \in \mathbb{N}.$$

The proof is beyond the scoop of this note.

The most important part of the theorem is the fourth part, which says that Newton's method is locally convergent, and the rate is quadratic. Practically this means two new correct digits at every steps of the iteration.

Fixed-point iteration in $\mathbb{R}$

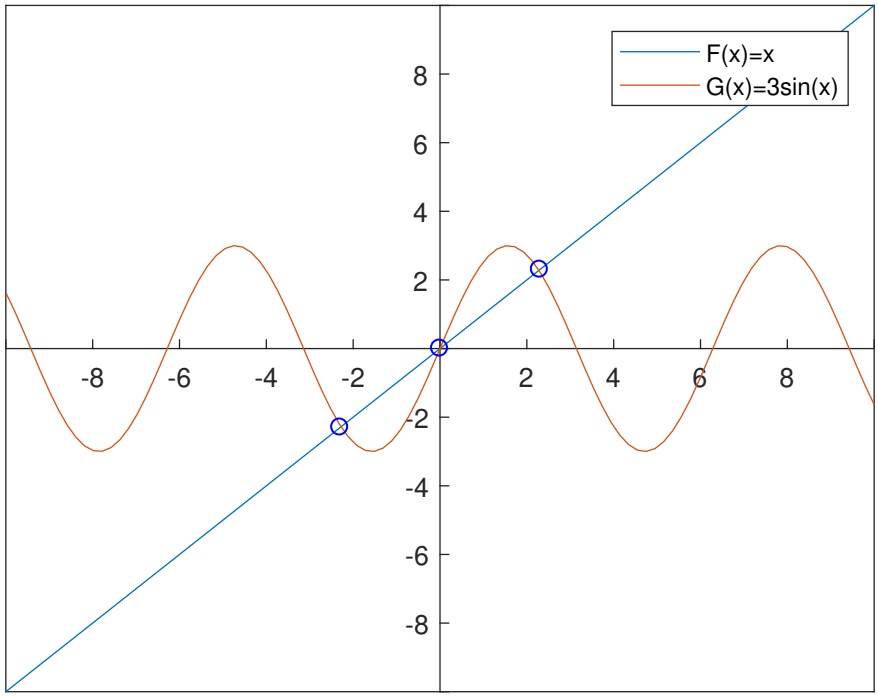
This iteration is essentially applicable for problems of the form

$$T(x) = x. \tag{3.4}$$

Definition 3.2.3 — Fixed-point. Let $T\colon [a,b] \to \mathbb{R}$ be a function, where $a < b$ are real numbers. Then $\bar{x} \in [a,b]$ is said to be a **fixed-point of** T if it is a solution of the equation (3.4).

■ **Example 3.11** Let us define the functions $F, G\colon \mathbb{R} \to \mathbb{R}$ as $F(x) = x$, and $G(x) = x + 1$. Then an arbitrary real number is a fixed-point of F, and there is no fixed point of G. ■

The previous example suggests a geometric interpretation of fixed-points. If $\bar{x}$ is a fixed-point of a function G, then its graph intersects the graph of the identity function $F(x) = x$ at $\bar{x}$.



The function $G(x) = 3\sin(x)$ has three fixed-points denoted by blue circles.

As the original problem (3.1) has a different form, it is needed to transform the fixed-point problem (3.4) into a non-linear equation.

Let us assume that $\bar{x}$ is a solution of the fixed-point problem (3.4), then $\bar{x}$ is also a solution of the non-linear equation

$$F(x) = 0,$$

where

$$F(x) = x - T(x).$$

For some practical purposes, it is more convenient to use the transformation of the problem (3.1) into a fixed-point problem using the following transformation:

$$T(x) = x - \omega F(x),$$

where $\omega \neq 0$ is a given constant.[3]

It is true that $\bar{x}$ is a solution of (3.4) if and only if it is a solution of (3.1). Indeed, let $\bar{x}$ be a solution of the non-linear equation (3.1), then

$$T(\bar{x}) = \bar{x} - \omega F(\bar{x}) = \bar{x} - \omega \cdot 0 = \bar{x},$$

that is, $\bar{x}$ is a fixed-point of T.

Let us assume now, that $\bar{x}$ is a fixed point of T, then

$$T(\bar{x}) = \bar{x} - \omega F(\bar{x}) \quad \Rightarrow \quad \bar{x} = \bar{x} - \omega F(\bar{x}) \quad \Rightarrow \quad F(\bar{x}) = 0.$$

So, $\bar{x}$ is a solution of (3.1).

The method is applicable if T is a contraction with a factor between zero and one defined on a compact[4] interval of the real line.

Definition 3.2.4 — Contraction. Let $T\colon [a,b] \to [a,b]$, then T is said to be a **q-contraction** if there is a non-negative real number q such that

$$|T(x) - T(y)| \leq q|x-y|, \qquad x,y \in [a,b].$$

[Contraction]

■ **Example 3.12** Let $F(x) = x^2 - 2$, then with $\omega = \frac{1}{4}$, the map

$$T(x) = x - \omega F(x) = x - \frac{1}{4}(x^2 - 2)$$

will be a contraction on the interval $[1,2]$.

To prove this, let us apply the Lagrange's mean value theorem.

$$|T(x) - T(y)| \leq \max_{\xi \in [1,2]} |T'(\xi)| \cdot |x-y| \leq \frac{1}{2}|x-y|, \qquad x,y \in [1,2].$$

Because

$$|T'(\xi)| = \left|1 - \frac{1}{2}\xi\right| \leq \frac{1}{2}, \qquad \xi \in [1,2].$$

■

In the definition, the assumption for the range of T is essential, namely, T is a self-map of the compact interval $[a,b]$.

Theorem 3.2.3 — Banach fixed-point theorem on $\mathbb{R}$. Let T be a self-map of the compact interval $[a,b]$. If T is a q-contraction with $q < 1$, then T has a unique fixed point, which is the limit of the sequence $\{x_n\}_{n\in\mathbb{N}}$, where $x_0 \in [a,b]$ arbitrary, and

$$x_{n+1} = T(x_n), \qquad n \in \mathbb{N}.$$

[3] This relaxed form of the transformation shows that Newton's iteration is also a very special fixed-point iteration with a sequence of parameters $\omega_n = \frac{1}{F'(x_n)}$.

[4] On the real line this is equivalent to boundedness and closedness, so, all intervals of the form $[a,b]$ with $-\infty < a \leq b < \infty$ are compact.

In the proof, we use the following facts from elementary Calculus.

Reminder from Calculus: A sequence $\{x_n\}_{n\in\mathbb{N}}$ is called a Cauchy sequence if for all $\varepsilon > 0$ there is a natural number N such that

$$|x_n - x_m| \le \varepsilon \qquad \text{if} \qquad n, m > N.$$

It is also well-known that real Cauchy sequences are convergent.

Proof. We prove that the resulted sequence of the iteration

$$x_{n+1} = T(x_n), \qquad n \in \mathbb{N},$$

is a Cauchy sequence.

Firstly, let m be an arbitrary positive integer, then

$$|x_{m+1} - x_m| = |T(x_m) - T(x_{m-1})| \le q|x_m - x_{m-1}| =$$

$$= |T(x_{m-1}) - T(x_{m-2})| \le q^2|x_{m-1} - x_{m-2}| \le \cdots \le q^m|x_1 - x_0|.$$

Secondly, let $m \le n$ be arbitrary positive integers, then using the triangle inequality and the previous estimate we have

$$|x_{n+1} - x_{m+1}| = |x_{n+1} - x_n + x_n - x_{n-1} + x_{n-1} - \cdots - x_{m+1} + x_{m+1} - x_m| \le$$

$$\le |x_{n+1} - x_n| + |x_n - x_{n-1}| + \cdots + |x_{m+1} - x_m| \le q^{n-1}|x_1 - x_0| + \cdots + q^m|x_1 - x_0| =$$

$$= (q^{n-1} + \cdots + q^m)|x_1 - x_0| = \frac{q^{m+1} - q^n}{1-q}|x_1 - x_0| = q^{m+1}\frac{1 - q^{n-m-1}}{1-q}|x_1 - x_0| \le$$

$$\le q^{m+1}\frac{1}{1-q}|x_1 - x_0| \longrightarrow 0, \qquad \text{as} \quad m \longrightarrow \infty.$$

This means that the resulted sequence is a Cauchy sequence, which is convergent. Let us denote the limit by $\bar{x}$. Then for every $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that

$$|x_n - \bar{x}| \le \varepsilon, \qquad n \ge N.$$

Using this, we have the following estimate for $n - 1 \ge N$.

$$|T(\bar{x}) - \bar{x}| = |T(\bar{x}) - x_n + x_n - \bar{x}| \le |T(\bar{x}) - x_n| + |x_n - \bar{x}| = |T(\bar{x}) - T(x_{n-1})| \le$$

$$\le q|\bar{x} - x_{n-1}| + \varepsilon \le q\varepsilon + \varepsilon = (1+q)\varepsilon.$$

Because $\varepsilon > 0$ was arbitrary, we get

$$|T(\bar{x}) - \bar{x}| = 0,$$

which means that the limit of the sequence $\bar{x}$ is a fixed point of T.

For the uniqueness assume that $\tilde{x}$ is also a fixed-point of T. Then

$$|\tilde{x} - \bar{x}| = |T(\tilde{x}) - T(\bar{x})| \le q|\tilde{x} - \bar{x}|,$$

this implies that $q \ge 1$, which is a contradiction. ■

Using the previous theorem, we can build up a very simple globally convergent algorithm.

Fixed-point iteration algorithm:

Initialization: T, x_0, ε, $k = 0$

Step 1.: If $|T(x_k) - x_k| \leq \varepsilon$, then we accept x_k as a reasonable approximate solution, otherwise Step 2.

Step 2.: Let

$$x_{k+1} = T(x_k).$$

$k = k+1$, Step 1.

For some practical reasons, it is accustomed to give the maximum number of steps n.

Proposition 3.2.4 The rate of the convergence of the fixed-point iteration is linear.

Proof. Let $\bar{x}$ be the limit of the iteration, then we get

$$|x_{n+1} - \bar{x}| = |T(x_n) - T\bar{x}| \leq q|x_n - \bar{x}|.$$

Because of the assumption $q < 1$, we get the statement. ■

■ **Example 3.13** Let us find an approximate value of $\sqrt{2}$ with fixed-point iteration.

We can use the map

$$T(x) = x - \frac{1}{4}(x^2 - 2) = -\frac{1}{4}x^2 + x + \frac{1}{2},$$

which is a contraction with $q = \frac{1}{2}$, as it was shown in the previous example. Then our iteration results the following:

$$\begin{aligned} x_0 &= 1 \\ x_1 = T(x_0) &= \frac{5}{4} \\ x_2 = T(x_1) &= \frac{87}{64} \\ x_3 = T(x_2) &= \frac{1505}{1077} \\ x_4 = T(x_3) &= \frac{1009}{716} \approx 1.4092 \end{aligned}$$

The fourth iteration gives the first exact digit, which shows that this method is quite slow in general. However, its global convergence and easy implementability makes it important not only from a theoretical, but also from a practical point of view. ■

3.2.3 Newton's method

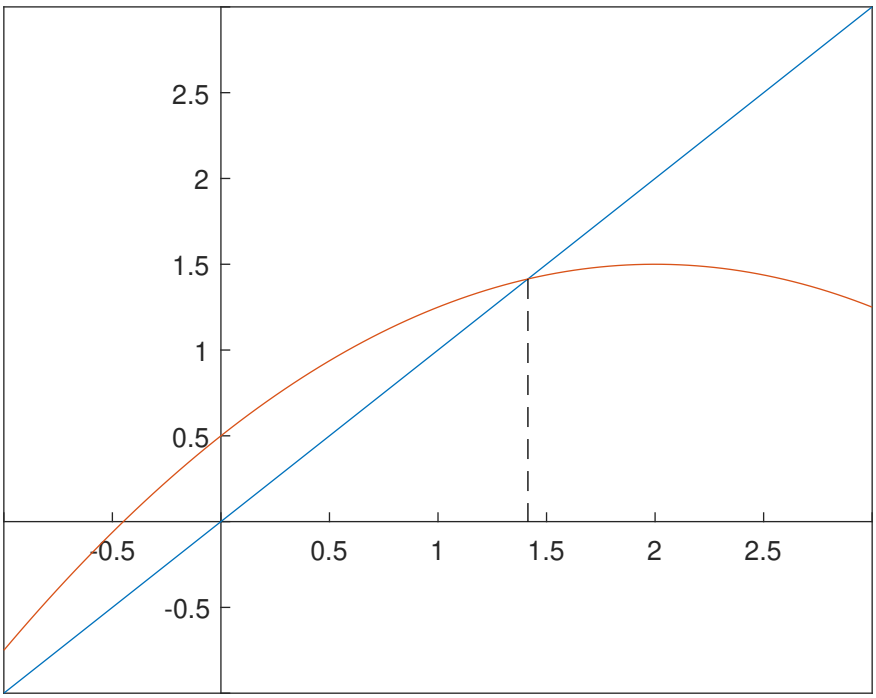
We assume in this section that $F\colon \mathbb{R}^n \to \mathbb{R}^n$, and we shall continue to find the solution of the equation $F(x) = 0$, which is practically a system of non-linear equations. Here we use the norm instead of the absolute value. The essential difference between the multivariable and the one variable case is the more complicated structure of sets.

In the one variable case, the ground of actions was an interval (at least in most of the cases). However, in the multivariable case, more complicated sets can be important in applications. For the sake of simplicity we will use multidimensional intervals (product of ordinary intervals) or balls.

The linearized version of the problem again has the form

$$F(x) = F(\bar{x}) + F'(\xi)(x - \bar{x}),$$

where $F(x)$, $F(\bar{x})$, x, $\bar{x}$ are vectors in $\mathbb{R}^n$, $F'(\xi)$ is an n by n matrix, which contains the partial derivatives of the coordinate functions of F, and ξ is on the section determined by x and $\bar{x}$. Using



Graphical solution of the fixed-point equation $T(x) = x - \frac{1}{4}(x^2 - 2) = x$.

that $\bar{x}$ is a solution of (3.1), if the derivative matrix is invertible at ξ, then after rearrangement we have

$$\bar{x} = x - (F'(\xi))^{-1} F(\bar{x}).$$

In a very similar way to the one variable case, we get the following iteration:

Initialization: F, x_0, and $k = 0$

Step 1.: If $F(x_k) = 0$, then x_k is a solution, otherwise Step 2.

Step 2.: Let

$$x_{k+1} = x_k - (F'(x_k))^{-1} F(x_k). \qquad \text{(Newton iteration)}$$

Step 1.

It is also reasonable, just like in the one-dimensional case, to give the maximum number of steps n, and a tolerance $\varepsilon > 0$. With the latter one, we require $\|F(x_k)\| \leq \varepsilon$ instead of $F(x_k) = 0$.

In practice, there is an other very important modification of the iteration scheme in the multivariable case. Calculation of the inverse matrix of the derivative $(F'(x_k))^{-1}$ is very costly. So, instead of the formula in Step 2., we use the following one in practice:

$$F'(x_k)(x_k - x_{k+1}) = F(x_k). \qquad \text{(modified Newton iteration)}$$

This is a system of linear equations for the unknown vector x_{k+1}. Using the modified formula, we have the following modified Newton algorithm:

Initialization: F, x_0, $F(x_0)$, n, ε, and $k = 0$

Step 1.: If $\|F(x_k)\| \leq \varepsilon$ or $n \leq k$, then stop with solution x_k, otherwise Step 2.

Step 2.: Let x_{k+1} be the unique solution ($F'(x_k)$ is invertible) of the following system of linear equations.

$$F'(x_k)(x_k - x_{k+1}) = F(x_k). \qquad \text{(modified Newton iteration)} \tag{3.5}$$

Step 1.

Theorem 3.2.5 — Properties of Newton's method. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function, $\bar{x} \in \mathbb{R}$ be a solution of (3.1), where the derivative matrix $F'(\bar{x})$ is invertible. Then there are $\varepsilon > 0$ and $C > 0$ constants such that

- $\bar{x}$ is the unique solution of problem (3.1) in the ball $B_\varepsilon(\bar{x}) = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| < \varepsilon\}$;
- $\|(F'(x))^{-1}\| \le C$ for all $x \in B_\varepsilon(\bar{x})$;
- for every $x_0 \in B_\varepsilon(\bar{x})$ the Newton iteration either stops after finitely many steps on $\bar{x}$ or results a superlinearly convergent sequence with limit $\bar{x}$;
- if F' is a Lipschitz continuous function on the ball $B_\varepsilon(\bar{x})$ with constant L, that is

$$\|F'(x) - F'(y)\| \le L\|x - y\|, \qquad x, y \in B_\varepsilon(\bar{x}),$$

then the rate of the convergence of Newton iteration is quadratic, that is

$$\|x_{n+1} - x_n\| \le \frac{CL}{2}\|x_n - \bar{x}\|^2, \qquad n \in \mathbb{N}.$$

The proof is beyond the scoop of this note.

Here, just like in the one-variable case, the most important part is the fourth one, which states quadratic convergence of Newton's method if the derivative is nice.

■ **Example 3.14** Let

$$F(x_1, x_2) = \begin{bmatrix} x_1^2 - 2x_1x_2 \\ x_2^2 - 3 + x_1 \end{bmatrix}.$$

Then

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{3} \end{bmatrix}$$

is a solution of the equation $F(x) = 0$. The derivative of F is

$$F'(x_1, x_2) = \begin{bmatrix} 2x_1 - 2x_2 & -2x_1 \\ 1 & 2x_2 \end{bmatrix}.$$

Let the starting point of the iteration be

$$x^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Find x^1!

The function F and its derivative at x^0 are

$$F(x^0) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \qquad F'(x^0) = \begin{bmatrix} 2 & -2 \\ 1 & 0 \end{bmatrix},$$

and the corresponding linear system is

$$\begin{bmatrix} 2 & -2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 - x_1^1 \\ 0 - x_2^1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

which is a system of linear equations, where the unknown is

$$x^1 = \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix}.$$

The solution of the linear system is

$$x^1 = \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ 3 \end{bmatrix}.$$

■

3.2.4 Fixed point iteration

Let $M \subset \mathbb{R}^n$ be a bounded, closed set[5], and $T : M \to M$ be a map. Similarly to the one-variable case, we examine the conditions under which the non-linear equation

$$Tx = x, \qquad x \in M, \tag{3.6}$$

may be solved by successive approximation

$$x_{n+1} = Tx_n, \quad x_0 \in M \text{ is arbitrary} \quad n = 0, 1, 2, \ldots \tag{3.7}$$

Definition 3.2.5 — Contraction. Let $TM \to M$, where $M \subset \mathbb{R}^n$, then T is said to be a **q-contraction** if there is a non-negative real number q such that

$$\|T(x) - T(y)\| \leq q\|x - y\|, \qquad x, y \in M.$$

Theorem 3.2.6 — Banach fixed-point theorem in $\mathbb{R}^n$. If T is a q-contraction with a factor $0 \geq q < 1$ on a compact subset M of $\mathbb{R}^n$, then it has a unique solution of (3.6), which is the limit of the iteration (3.7).

Proof. Formally it is exactly the same as in the one-dimensional case, but here we use the norm of vectors in $\mathbb{R}^n$ instead of the absolute value of real numbers. ■

3.3 Interpolation

Interpolation is a type of approximation of a model function. The basis of the interpolation is the observed (given) data, which are usually pairs (x_i, f_i), $i = 0, \ldots, n$, where $x_i \neq x_j$ if $i \neq j$.

The first coordinates of the pairs are the **nodes** and the second coordinates are the **observed values**.

The goal of the interpolation process is to find a function φ, which interpolates f_i at x_i, that is to say,

$$\varphi(x_i) = f_i, \qquad i = 0, \ldots, n.$$

In this section our interpolation function will be a polynomial or a piecewise polynomial function.

[5] According to the Heine-Borel theorem, this is equivalent to the compactness of M in this case.

3.3.1 Lagrange interpolation

Here the interpolation function will be a polynomial with degree at most n. So, we are looking for φ in the following form:

$$\varphi(x) = a_n x^n + \cdots a_1 x + a_0,$$

where $a_n, \ldots, a_1, a_0$ are real numbers. They depend on the observed data $(x_i, f_i),\ i = 0, 1, \ldots, n$.

When we are looking for φ, the form above the corresponding problem is called **Lagrange interpolation problem**.

Elementary Lagrange polynomials and the Lagrange polynomial

The Lagrange interpolation problem always have a unique, at most degree n solution.

Theorem 3.3.1 Let $x_0, \ldots, x_n$ be given distinct nodes, and $f_0, \ldots, f_n$ be the corresponding observed values. Then there is a unique polynomial with degree at most n $\mathscr{L}_n \in \mathscr{P}_n$ such that

$$\mathscr{L}_n(x_i) = f_i, \qquad i = 0, 1, \ldots, n.$$

Proof. This proof is a constructive one, that is to say, we give the form of the polynomial, which fulfils the requirement of the theorem.

At first, we construct the **elementary Lagrange polynomials**. The form of the ith elementary Lagrange polynomial depending on the given nodes is

$$\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \qquad i = 0, 1, \ldots, n.$$

The importance of ℓ_i is that it takes 1 at x_i and it takes 0 at $x_j,\ j \neq i$. Indeed,

$$\ell_i(x_i) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x_i - x_j}{x_i - x_j} = 1,$$

and

$$\ell_i(x_k) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x_k - x_j}{x_i - x_j} = 0,$$

because there is a factor $x_j - x_j = 0$ in the nominator of the product above, when $k = j$.

The construction of the Lagrange interpolation polynomial is very easy now. It is just a certain linear combination of the elementary Lagrange polynomials with coefficients of the observed data. In detail, let

$$\mathscr{L}_n(x) = \sum_{i=0}^{n} f_i \ell_i(x).$$

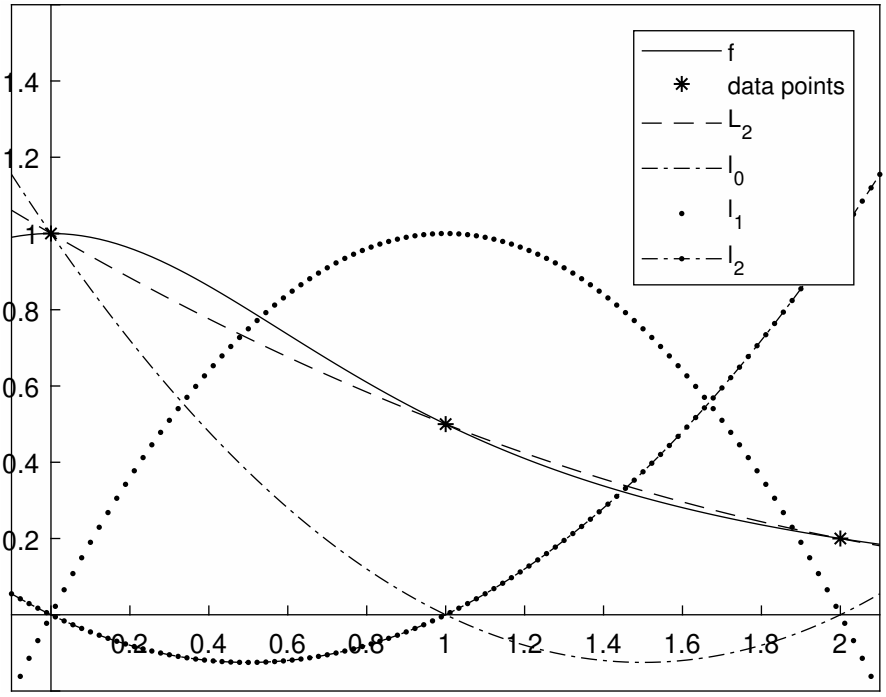
Then the degree of $\mathscr{L}_n$ is at most n because it is a linear combination of polynomials of degree n. Moreover, because of the definition of the elementary Lagrange polynomials, we have

$$\mathscr{L}_n(x_i) = f_i, \qquad i = 0, 1, \ldots, n.$$

So, $\mathscr{L}_n$ really interpolates the data.

The only remaining part is its uniqueness. If L_n is also a polynomial with degree at most n, which interpolates the same data like $\mathscr{L}_n$, then their difference polynomial $\mathscr{L}_n - L_n$ vanishes at $n+1$ points $(x_0, x_1, \ldots, x_n)$, so this difference polynomial has at least $n+1$ different roots, and its degree is at most n, which means that it must be the zero polynomial. This completes the proof. ■

The above construction of $\mathscr{L}_n$ is quite complicated, because of the tricky structure of the elementary Lagrangian polynomials. Moreover, if we would like to augment a new data point, it is necessary to reconstruct the whole interpolation polynomial executing the tiring calculation from the very beginning.



Lagrange interpolation of the function $\frac{1}{1+x^2}$ on the interval $[0,2]$ with three equidistant nodes

■ **Example 3.15** Let $x_0 = 0,\ x_1 = 1,\ x_2 = 2$, and interpolate the function $f(x) = \frac{1}{x^2+1}$. So, $f_0 = 1,\ f_1 = \frac{1}{2},\ f_2 = \frac{1}{5}$. See the figure above.

The elementary Lagrange polynomials are:

$$\ell_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-2)}{(0-1)(0-2)} = \frac{1}{2}x^2 - \frac{3}{2}x + 1.$$

$$\ell_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-0)(x-2)}{(1-0)(1-2)} = -x^2 + 2x.$$

$$\ell_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-0)(x-1)}{(2-0)(2-1)} = \frac{1}{2}x^2 - \frac{1}{2}x.$$

So, we get

$$\mathscr{L}_2 = 1\cdot\ell_0(x) + \frac{1}{2}\ell_1(x) + \frac{1}{5}\ell_2(x) = \frac{1}{10}x^2 - \frac{3}{5}x + 1.$$

■

It is possible to avoid these uncomfortable circumstances using Newton's recursion, which will be the topic of a later subsection.

Behaviour of Lagrange interpolation

Let us start with the estimation of the interpolation error. For this, we need the **nodal polynomial** ω_{n+1}, which vanishes at the nodes $x_0, x_1, \ldots, x_n$.

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i).$$

Theorem 3.3.2 — Error estimate. Using the earlier notations, let us denote by $E_n(x)$ the error of the interpolation at x, that is to say the difference between the interpolated function and the Lagrange interpolation polynomial.

$$E_n(x) = f(x) - \mathscr{L}_n(x).$$

If f is $n+1$ times continuously differentiable on the smallest interval which contains all the nodes, then there is an element of this interval ξ such that

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x).$$

We do not prove this theorem.

Runge's example for the bad behaviour of Lagrange interpolation

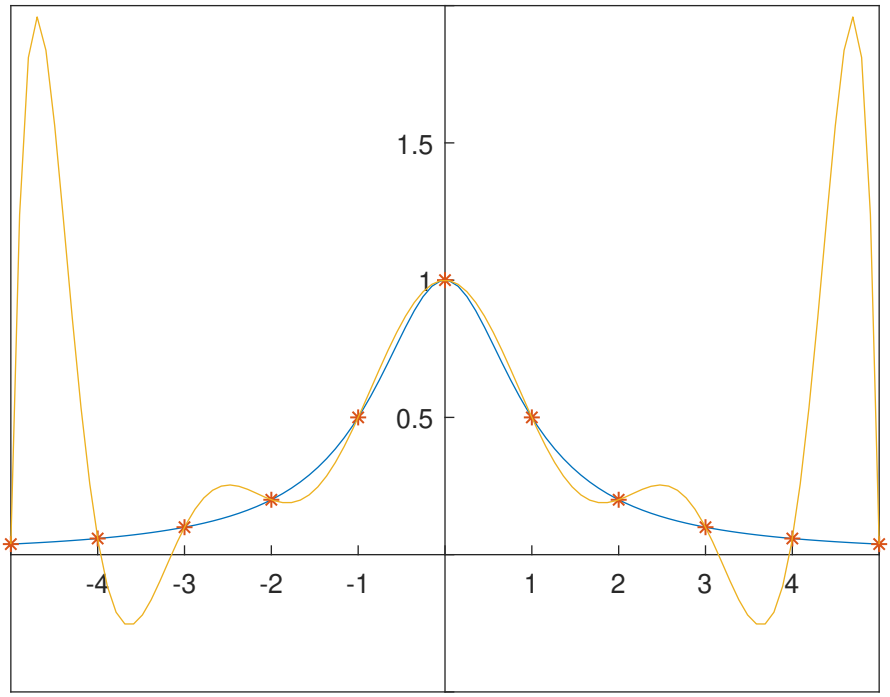


Figure 3.1: Lagrange interpolation of $\frac{1}{1+x^2}$ on the interval $[-5,5]$ using equidistant nodes.

This theorem suggests the false thing that the Lagrange interpolation polynomial tends to nicely to the interpolant if the number of nodes is increased.

Unfortunately the interpolation error can diverge as n tends to infinity. This phenomenon is particularly evident in the neighborhood of the endpoints of the interval, as shown in Figure 3.1. However, it is possible to choose the nodes in such a way which guarantee uniform convergence on the whole interpolation interval. For further information see [**QSS2007**] and [**BP1978**].

Newton's recursion

From a practical aspect, the previous construction of Lagrangian interpolation polynomial using elementary Lagrange polynomials is not the best choice. It is too complicated and its computational cost is too expensive. The goal of this subsection is to get rid of these inconveniences, and give a recursive method, which is more transparent, and its computational cost is cheaper.

If we fix a set of nodes $x_0,\dots,x_n$ and observed data $f_0,\dots,f_n$, then the corresponding Lagrange polynomial $\mathscr{L}_n$ can be written as the sum of the Lagrange polynomial $\mathscr{L}_{n-1}$ and a polynomial with degree at most n. Let us denote it q_n, that is to say

$$\mathscr{L}_n = \mathscr{L}_{n-1} + q_n \quad \Rightarrow \quad q_n(x) = \mathscr{L}_n(x) - \mathscr{L}_{n-1}(x).$$

Using their definitions, we have

$$q_n(x_i) = \mathscr{L}_n(x_i) - \mathscr{L}_{n-1}(x_i) = f_i - f_i = 0, \qquad i = 0,1,\dots,n-1.$$

So, q_n is a scalar multiply of the nodal polynomial ω_n, that is,

$$q_n(x) = a_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) = a_n\omega_n(x)$$

for some constant a_n. This entails the formula

$$a_n = \frac{q_n(x_n)}{\omega_(x_n)} = \frac{f_n - \mathscr{L}_{n-1}(x_n}{\omega_n(x_n)} =: f[x_0,\dots,x_n]$$

Definition 3.3.1 — Divided differences. The coefficient

$$f[x_0,\dots,x_n]$$

in the formula above is said to be **the nth Newton divided difference**.

More precisely it is the nth Newton divided difference of the function f with respect to the points $x_0,\dots,x_n$. This is too lengthy, so, in practice, we use the shorter name without the specification of the function and the base points if there is no ambiguity.

The advantage of these differences is clear from the definition of q_n. Indeed,

$$\mathscr{L}_n = \mathscr{L}_{n-1} + q_n = \mathscr{L}_{n-1} + f[x_0,\dots,x_n]\omega_n(x).$$

This implies the following expression for $\mathscr{L}_n$.

$$\mathscr{L}_n(x) = \sum_{k=0}^{n} f[x_0,\dots,x_k]\omega_k(x) = \sum_{k=0}^{n} a_k\omega_k(x) = a_0 + a_1(x-x_0) + \cdots + a_n(x-x_0)\cdots(x-x_{n-1}).$$

So, if we know all the coefficients (the divided differences), we have explicit formula for the Lagrange interpolation polynomial.

Luckily, these coefficients can be determined in an easy, recursive way.

Theorem 3.3.3

$$f[x_0,\dots,x_n] = \frac{f[x_1,\dots,x_n] - f[x_0,\dots,x_{n-1}]}{x_n - x_0}.$$

Proof. **Hint:** Prove at first the formula:

$$\mathscr{L}_n(x) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x)}{(x-x_i)\omega_n'(x_i)} f_i.$$

Using this and the expression $\mathscr{L}_n(x) = \sum_{k=0}^{n} f[x_0,\dots,x_k]\omega_k(x)$, prove the following formula:

$$f[x_0,\dots,x_n] = \sum_{i=0}^{n} \frac{f_i}{\omega'_{n+1}(x_i)}.$$

This last expression, after some algebraic manipulation, gives the statement. ■

According to the recursion given by the theorem above, the required coefficients are contained by the diagonal boxed entries of the following tableau:

$$\begin{array}{llllll}
x_0 & \boxed{f_0 = f[x_0]} & & & & \\
x_1 & f_1 = f[x_1] & \boxed{f[x_0,x_1]} & & & \\
x_2 & f_2 = f[x_2] & f[x_1,x_2] & \boxed{f[x_0,x_1,x_2]} & & \\
\vdots & \vdots & \vdots & \vdots & & \\
x_n & f_n = f[x_n] & f[x_{n-1},x_n] & f[x_{n-2},x_{n-1},x_n] & \dots & \boxed{f[x_0,\dots,x_n]}
\end{array}$$

The corresponding interpolation polynomial will be

$$\mathscr{N}_n(x) = f[x_0] + f[x_0,x_1](x-x_0) + \dots + f[x_0,\dots,x_n](x-x_0)\cdots(x-x_n).$$

■ **Example 3.16** Let us find the Lagrange interpolation polynomial, which interpolates the data below.

x_i	0	1	2	3	4
f_i	1	2	-1	2	3

Using the Newton recursion, we have the following divided difference tableau:

$$\begin{array}{lllll}
0 & \boxed{1} & & & \\
1 & 2 & \boxed{1} & & \\
2 & -1 & -3 & \boxed{-2} & \\
3 & 2 & 3 & 3 & \boxed{-\frac{5}{3}} \\
4 & 3 & 1 & 1 & -\frac{2}{3} \quad \boxed{-\frac{7}{12}}
\end{array}$$

The corresponding Newton polynomial will be (see the figure below)

$$\mathscr{N}_4(x) = 1 + x - 2x(x-1) + \frac{5}{3}x(x-1)(x-2) - \frac{7}{12}x(x-1)(x-2)(x-3).$$

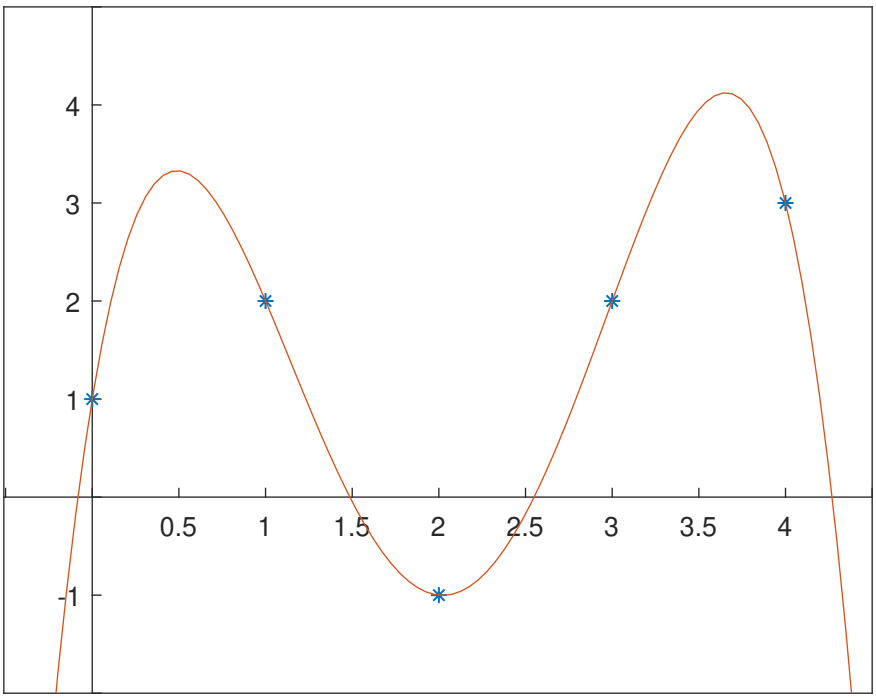
■

3.4 Least square approximation

Like in the case of interpolation, here we have a given data $t_1,\dots,t_n$, which usually (not always) denote (not necessarily different) time instants and $f_1,\dots,f_n$ are the observed values of the model. Here we are looking for a simple function "close" to the given data.

In the case of interpolation, the complexity of the interpolation function was determined by the number of the data. Namely, $n+1$ nodes imply interpolation at most degree n. It was seen that this number cannot be less besides the requirements $\mathscr{L}_n(x_i) = f_i$.

Here we do not require exact interpolation of the data at the nodes, therefore the reward is a simpler approximation function.



$$\mathcal{N}_4(x) = 1 + x - 2x(x-1) + \tfrac{5}{3}x(x-1)(x-2) - \tfrac{7}{12}x(x-1)(x-2)(x-3)$$

Actually, it is necessary to give a set of functions, which highly depends on the data itself. For example, if we have a periodic data, it is convenient to give a set of periodic functions. The approximation will be a certain linear combination of the elements of the given set. The coefficients are given as a solution of a linear system (Gaussian normal equation), which can be constructed with the aid of the data and the set of the given functions.

To avoid sensitivity of the model, with respect to translations along the y axis, it is always reasonable to take the constant one function into the set of the given functions.

3.4.1 Linear case

First we deal with the simplest case, when our set of functions contains only two maps, besides the constant one function the identity function. So, the approximation function will have the form

$$F(t) = x_1 + x_2 t.$$

This case deserves a special interest, because it has a great use and significant importance in very different parts of mathematics, applied mathematics and other sciences as well.

The main question now is, how to determine the exactness of the approximation, in other words, what the approximation is "close" enough to the model means.

We will measure the sum of the squares of the discrepancies between the model and the approximation, and we will minimize this sum at last. This is where the name of the process comes from. In detail, let us find the following minimum:

$$\min_{x_1,x_2} \mathcal{J}(x_1,x_2) = \sum_{i=1}^{n} (F(t_i) - f_i)^2 = \sum_{i=1}^{n} (x_1 + x_2 t_i - f_i)^2. \tag{3.8}$$

$\mathcal{J}$ is a convex, differentiable function. It is known, that a function like this has a global minimum at points where its derivative is zero.

Let us calculate the partial derivatives and make them equal to zero. At last we get a linear system, which is called the **Gaussian normal equation**, whose solution gives the required parameters x_1 and x_2.

$$\frac{\mathcal{J}(x_1,x_2)}{\partial x_1} = 2\sum_{i=1}^{n}(x_1 + x_2 t_i - f_i) = 0, \qquad \text{and} \qquad \frac{\mathcal{J}(x_1,x_2)}{\partial x_2} = 2\sum_{i=1}^{n}(x_1 + x_2 t - f_i)t_i = 0.$$

The resulted linear system is:

$$\begin{bmatrix} n & \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} t_i & \sum_{i=1}^{n} t_i^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} f_i \\ \sum_{i=1}^{n} t_i f_i \end{bmatrix}$$

Introducing the notations

$$A = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \qquad f = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

we have the following form for the Gaussian normal equation:

$$A^T A x = A^T f, \qquad \text{where} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{3.9}$$

■ **Example 3.17** Let us find the line, which is the closest one to the following data in the least square sense!

t_i	0	0.1	0.1	0.3	0.4	0.4	0.6	0.7	0.8	0.9	1
f_i	1	1.1	1.2	1.1	1.8	1.85	2.6	1.7	2	2	1.95

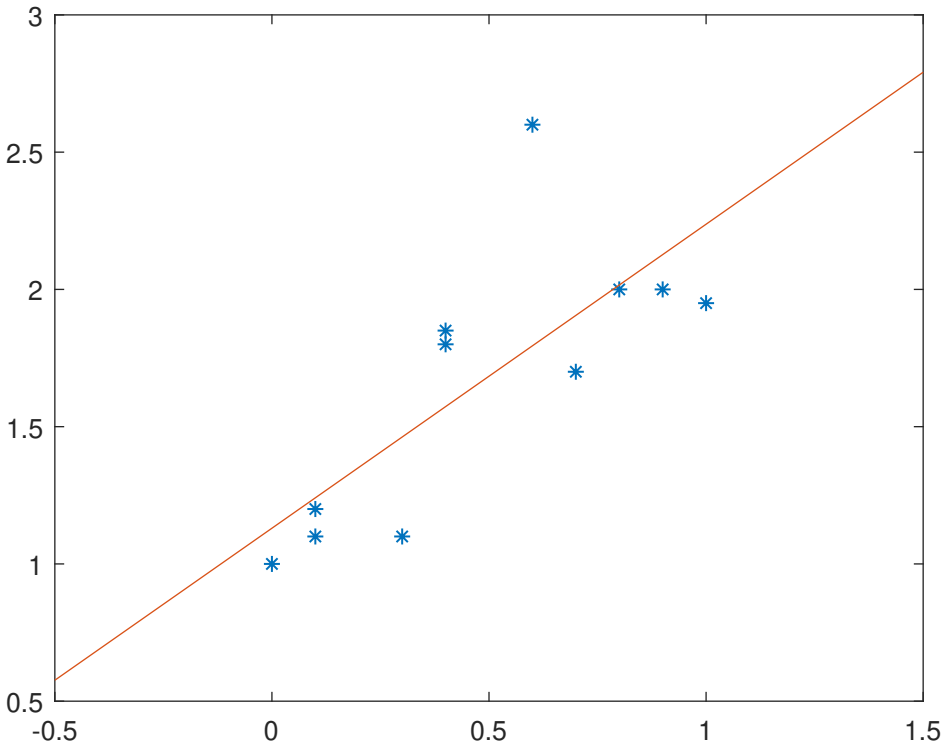
Then

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0.1 \\ 1 & 0.1 \\ 1 & 0.3 \\ 1 & 0.4 \\ 1 & 0.4 \\ 1 & 0.6 \\ 1 & 0.7 \\ 1 & 0.8 \\ 1 & 0.9 \\ 1 & 1 \end{bmatrix}, \qquad \text{and} \qquad f = \begin{bmatrix} 1 \\ 1.1 \\ 1.2 \\ 1.1 \\ 1.8 \\ 1.85 \\ 2.6 \\ 1.7 \\ 2 \\ 2 \\ 1.95 \end{bmatrix}.$$

The corresponding Gaussian normal equation and its solution are

$$A^T A x = \begin{bmatrix} 11 & 5.3 \\ 5.3 & 3.75 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A^T f = \begin{bmatrix} 18.3 \\ 10.12 \end{bmatrix}, \qquad \text{and} \qquad x = \begin{bmatrix} 1.1301 \\ 1.1074 \end{bmatrix}.$$

■



Solution of Example 3.17

Solvability of the Gaussian normal equation

We require a unique solution of the Gaussian normal equation, otherwise there is unwelcome doubtfulness in our model. An inhomogeneous linear system, like the normal equation, has a unique solution if and only if the matrix A^TA is non-singular, in other words, the square matrix A^TA is invertible.

If not all the t_is are equal, then the Gaussian normal equation always have a unique solution in the linear case!

It is very important to emphasize the fact that this is only true in the linear case. We will see in the next subsection that the situation is a little bit more subtle if the model function is non-linear.

3.4.2 General case

Like in the linear case, let us assume that we have the given data $t_1,\dots,t_n$ and $f_1,\dots,f_n$. We are looking for the model function now in the form

$$F(t) = x_1\varphi_1(t) + \cdots + x_m\varphi_m(t) = \sum_{i=1}^{m} x_i\varphi_i(t),$$

where φ_i, $i = 1,\dots,m$ are given functions, and x_i, $i = 1,\dots,m$ are the unknown parameters.

Let us observe that this case contains the linear one. Indeed, with the choices $\varphi_1(t) = 1$ and $\varphi_2(t) = t$, we have $F(t) = x_1\varphi_1(t) + x_2\varphi_2(t) = x_1 \cdot 1 + x_2 \cdot t = x_1 + x_2t$, which is really the linear model.

The unknown parameter vector $x^T = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix}$ arises as the solution of the Gaussian normal equation of the problem. This system can be derived in a pretty similar way, like in the linear case.

Definition 3.4.1 — Gaussian normal equation. Let us introduce the notations below

$$A = \begin{bmatrix} \varphi_1(t_1) & \dots & \varphi_m(t_1) \\ \vdots & \dots & \vdots \\ \varphi_1(t_n) & \dots & \varphi_m(t_n) \end{bmatrix}, \qquad f = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

then the following system of linear equations is said to be the **Gaussian normal equation of the problem**.

$$A^T Ax = A^T f, \qquad \text{where} \qquad x^T = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \quad \text{is the unknown parameter vector.} \tag{3.10}$$

■ **Example 3.18** Let us consider the following data.

t_i	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
f_i	-1	-0.4	0	0.1	0.8	1.02	0.2	-0.31	-0.45	-1	-1.2

This data has a "periodic nature" at the first view, it is reasonable to choose a different model than the linear one. As it is seen on Figure 3.2, the line, which is the best one in the lest square sense, is rather far from the data.

Let

$$\varphi_1(t) = 1, \qquad \varphi_2(t) = \cos(\pi t), \qquad \varphi_3(t) = \cos(2\pi t).$$

Then the Gaussian normal equation of the problem is

$$A^T Ax = \begin{bmatrix} 11 & 0 & 1 \\ 0 & 6 & 0 \\ 1 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} A^T f = \begin{bmatrix} -2.24 \\ 1.5611 \\ -5.2358 \end{bmatrix}.$$

The solution of the system is

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -0.1262 \\ 0.2602 \\ -0.8516 \end{bmatrix}.$$

So, the model function is

$$F(t) = x_1 + x_2\cos(\pi t) + x_3\cos(2\pi t) = -0.12621 + 0.2602\cos(\pi t) - 0.8516\cos(2\pi t).$$

It is clear (see Figure 3.2) that this non-linear model fits much better than the linear one. ■

Solvability of the Gaussian normal equation in the non-linear case

Like in the linear case, a unique solution of the Gaussian normal equation is required. Otherwise the resulted model function F is not unique and it has no use.

Uniqueness is ensured by the non-singularity of the matrix $A^T A$. If it is not the case, then we either have too many functions in the model or we have not enough data.

In the first case, we have to reduce the number of φ_is, in the second case, we have to perform more observations for the sake of the enlargement of the data set.

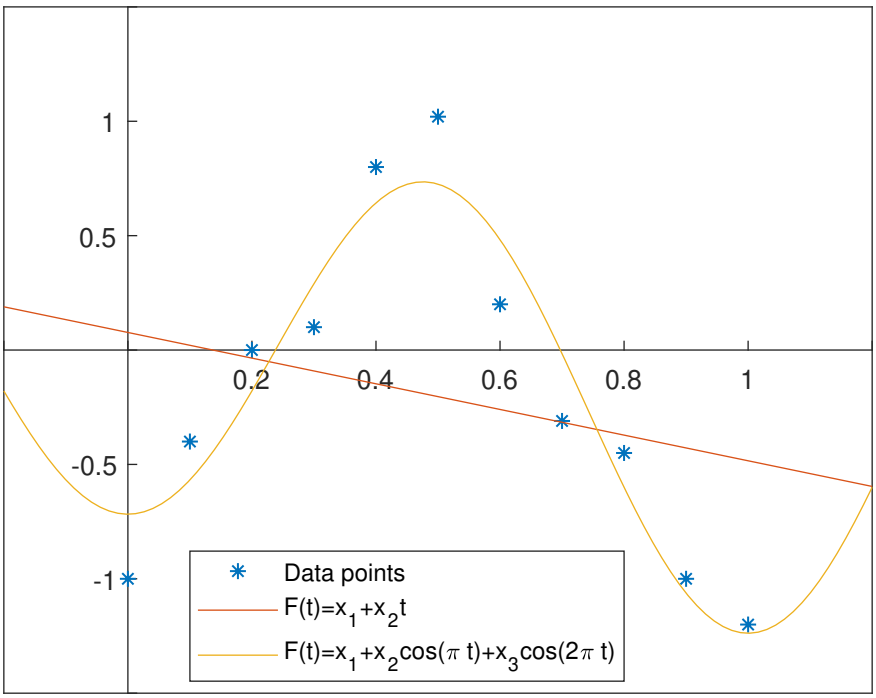


Figure 3.2: Solution of Example 3.18

3.5 Numerical integration

Let $f\colon [a,b] \to \mathbb{R}$ be integrable, where $[a,b]$ is a proper, finite interval. We intend to find the value of the integral[6]

$$I(f) = \int_a^b f(x)dx.$$

Even if the Newton-Leibniz formula is applicable, it is not easy to find the exact value of this integral.

We are looking for formulae, which provide an approximation of $I(f)$.

We will consider only one-dimensional integrals over bounded intervals. The interested reader can find more information about the approximation of multi-dimensional and indefinit integrals in [**Kre1988**] and in [**QSS2007**].

The basic idea of numerical integration is to divide the interval $[a,b]$ into pieces (taking a set of nodes), and substituting the function f with an approximation on the subintervals. This approximation function should be easily integrable, for example a polynomial (like the Lagrange interpolation polynomial) seems to be a good choice from this point of view. After summation of the approximate integrals over the whole interval, we get an approximation of $I(f)$ as well.

Several technically and theoretically difficult questions are popping up during this process. So, here we deal only with the most elementary three quadrature rules (approximation formulae for $I(f)$).

[6]The value of $I(f)$ depends on the interval $[a,b]$. However, if there is no ambiguity, we use this short notation instead of the more correct one $I(f,a,b)$.

To reach our designated goal, we use the integral of the Lagrangian interpolation polynomial of degree zero, one or two of the function in question. Depending on the degree of the interpolation polynomial, we use the notation $I_0(f)$, $I_1(f)$ and $I_2(f)$ respectively for the resulted approximate value of the integral.

3.5.1 The Midpoint formula

This quadrature rule is obtained by approximating f with the constant function equal to the value attained by f at the midpoint of the interval (see Figure 3.3), that is to say,

$$I_0(f) = (b-a) f\left(\frac{a+b}{2}\right).$$

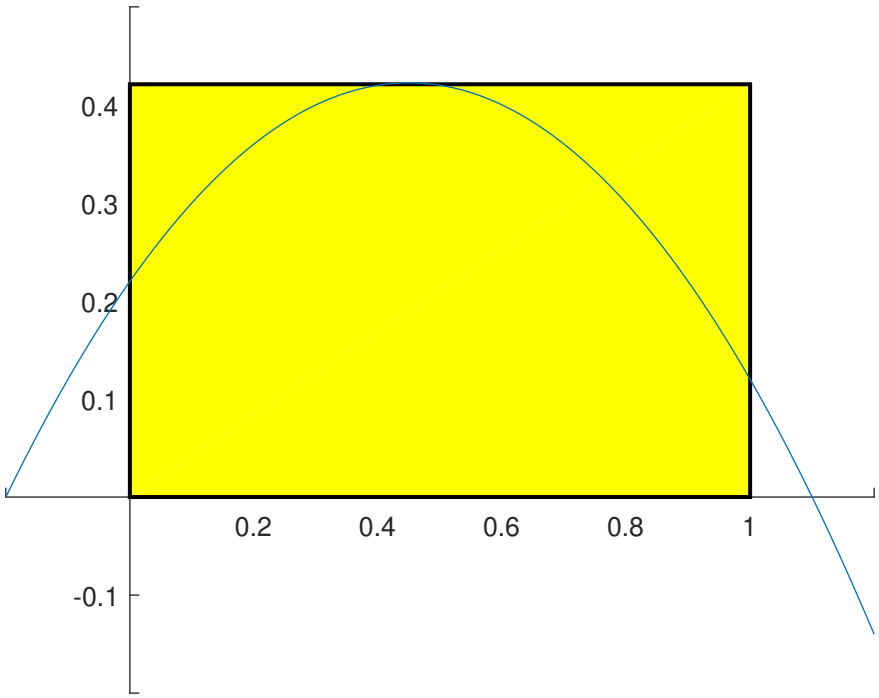


Figure 3.3: Numerical integration over the interval $[0,1]$ using midpoint rule

If f is two times continuously differentiable, then we have the following error estimate.

$$|I_0(f) - I(f)| \leq \frac{(b-a)^3}{24} \max_{x\in[a,b]} |f''(x)|.$$

From the estimation above, it follows that the Midpoint formula is exact for polynomials with degree at most one. Indeed, in this case the second derivative of the function is zero, so we have zero error on the right hand side of the previous inequality.

Actually, one can derive this by direct calculation in the following way. Let $f(x) = \alpha x + \beta$ be a polynomial with degree one, and $[a,b]$ a bounded interval. Then the exact integral is

$$I(f) = \int_a^b f(x)dx = \int_a^b (\alpha x+\beta)dx = \left[\alpha\frac{x^2}{2} + \beta x\right]_a^b = \frac{\alpha}{2}(b^2-a^2) + \beta(b-a) = (b-a)\left(\alpha\frac{b+a}{2} + \beta\right).$$

The approximation of the integral with the Midpoint formula is

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right) = (b-a)\left(\alpha\frac{b+a}{2}+\beta\right).$$

So,

$$I(f) = I_0(f)$$

if f is a polynomial with degree at most one.

■ **Example 3.19** Let $f(x) = \sin(x)$, and $[a,b] = [0, \frac{\pi}{2}]$. Then

$$I(f) = \int_0^{\frac{\pi}{2}} \sin(x)dx = [-\cos(x)]_0^{\frac{\pi}{2}} = 1,$$

$$I_0(f) = \left(\frac{\pi}{2}-0\right)\sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}\pi}{4} \approx 1.1107.$$

The estimated error is

$$\frac{\left(\frac{\pi}{2}\right)^3}{24}\max_{x\in[0,\frac{\pi}{2}]}|-\sin(x)| = \frac{\pi^3}{192} \approx 0.1615,$$

which is greater than 0.1107 (the difference of the exact and the approximated value of the integral).

■

3.5.2 The Trapezoidal formula

This formula is obtained by replacing f with its Lagrangian interpolation polynomial of degree one over the interval.

The corresponding quadrature rule is

$$I_1(f) = \frac{b-a}{2}(f(a)+f(b)).$$

It is worthy to mention that $I_1(f)$ is the area of the trapezoidal determined by the points $(a,0)$, $(a,f(a))$, $(b,f(b))$, $(b,0)$.

If f is two times continuously differentiable, then we have the following error estimate.

$$|I_1(f) - I(f)| \leq \frac{(b-a)^3}{12}\max_{x\in[a,b]}|f''(x)|.$$

As a consequence of the error estimate, we have that the Trapezoidal formula, like the midpoint formula, is exact for polynomials with degree at most one. The proof is also similar to the case of midpoint formula, so, we omit it.

■ **Example 3.20** Let $f(x) = \sin(x)$, and $[a,b] = [0, \frac{\pi}{2}]$. Then

$$I(f) = \int_0^{\frac{\pi}{2}} \sin(x)dx = [-\cos(x)]_0^{\frac{\pi}{2}} = 1,$$

$$I_1(f) = \frac{\frac{\pi}{2}-0}{2}\left(\sin\tfrac{\pi}{4}+\sin 0\right) = \frac{\pi}{4} \approx 0.7854.$$

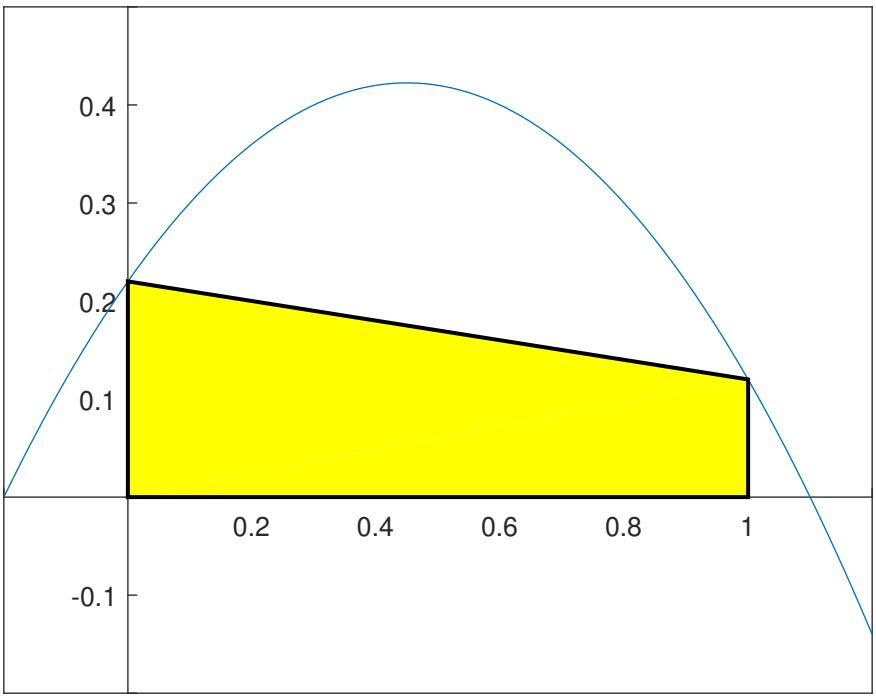


Figure 3.4: Numerical integration over the interval $[0,1]$ using trapezoidal rule

The estimated error is

$$\frac{\left(\frac{\pi}{2}\right)^3}{12} \max_{x\in[0,\frac{\pi}{2}]} |-\sin(x)| = \frac{\pi^3}{96} \approx 0.323,$$

which is greater than $1-0.7854 = 0.2146$ (the difference of the exact and the approximated value of the integral). ■

3.5.3 The Simpson formula

This formula can be derived by replacing f over $[a,b]$ with its interpolation polynomial of degree 2. The rule is the following:

$$I_2(f) = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right).$$

If f is four times continuously differentiable, then we have the following error estimate.

$$|I_2(f)-I(f)| \le \frac{(b-a)^5}{2880} \max_{x\in[a,b]} |f^{(iv)}(x)|.$$

This estimate shows that the Simpson formula is exact for polynomials with degree at most 3.

■ **Example 3.21** Let $f(x) = \sin(x)$, and $[a,b] = [0,\frac{\pi}{2}]$. Then

$$I(f) = \int_0^{\frac{\pi}{2}} \sin(x)dx = [-\cos(x)]_0^{\frac{\pi}{2}} = 1,$$

$$I_2(f) = \frac{\frac{\pi}{2}-0}{6}\left(\sin\frac{\pi}{2}+4\sin\frac{\pi}{4}\sin 0\right) = \frac{\pi}{12}(1+2\sqrt{2}) \approx 1.0023$$

The estimated error is

$$\frac{\left(\frac{\pi}{2}\right)^5}{2880}\max_{x\in[0,\frac{\pi}{2}]}|-\cos(x)| = \frac{\pi^5}{2880} \approx 0.0033,$$

which is greater than $1.0023-1=0.0023$ (the difference of the exact and the approximated value of the integral). ■

3.6 Basic optimization algorithms

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a given function. We are looking for a vector $\bar{x} \in \mathbb{R}^n$, which is a solution of the following minimization problem:

$$\min_{x\in\mathbb{R}^n} f(x), \tag{3.11}$$

that is to say,

$$f(\bar{x}) \le f(x), \qquad \text{for every} \qquad x \in \mathbb{R}^n.$$

This problem is called a global minimization problem. In general it is hard to solve it. Here we always assume that f is differentiable in a certain order.

Even in the smooth case, without further knowledge about the **objective function** f, we have chance to find only a **local solution** instead of a global one.

It is known, that if f has a local minima at $\bar{x}$ then its derivative is zero there. So, looking for a minima of a differentiable function can be transferred to the solution of the non-linear system[7] of equations:

$$f'(x) = 0.$$

As a consequence of this simple observation, we have that all the numerical methods for solving such systems can be applied for numerical solution of optimization problems.

Bisection method can be applied only for one-dimensional optimization problems. Newton's method, and Banach iteration can also be applied for multi-dimensional optimization problems.

■ **Example 3.22** Let us find the minimum of the function

$$f(x) = e^x + x^2.$$

Its derivative is

$$f'(x) = e^x + 2x.$$

As it is seen in Figure 3.5 the global minimum[8] of this function is around -0.3. So we have to solve the following non-linear equation.

$$e^x + 2x = 0.$$

[7] The derivative $f'(x)$ is a vector of $\mathbb{R}^n$ if f is defined on $\mathbb{R}^n$.

[8] This function is strictly convex, so, it has at most one global minimum point. It is also known that if it has a stationary point, then this is a global minimum place of the function as well.

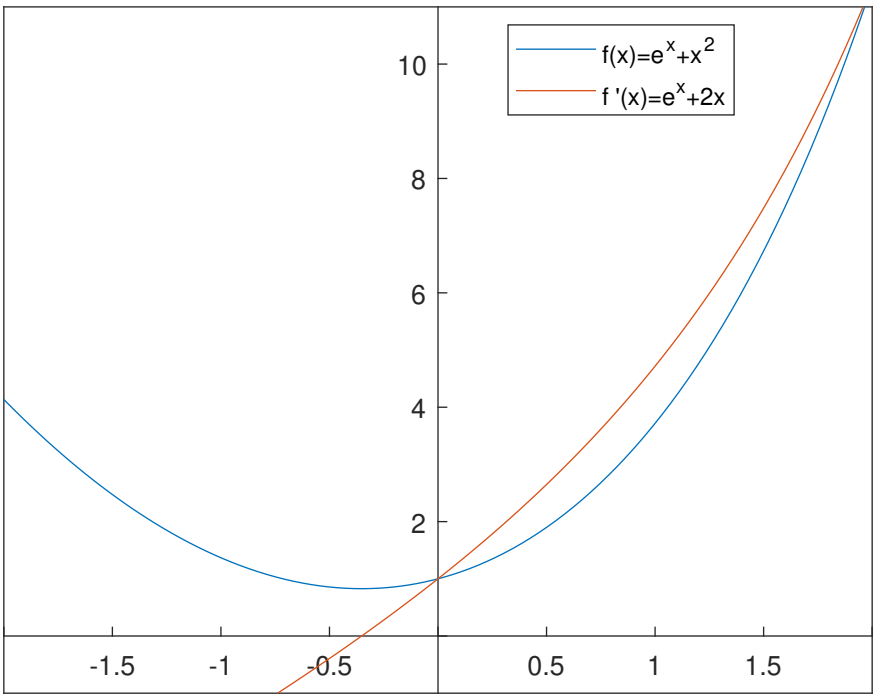


Figure 3.5: The function $f(x) = e^x + x^2$ and its derivative

Because f is a one-variable one, we can apply all the learned methods, namely, Bisection method, Newton's method, and Banach iteration.

Let us solve this problem using Newton's method.

According to the figure $x_0 = 0$ is an acceptable starting point. Our function is $F(x) = f'(x) = e^x + 2x$. Its derivative is $F'(x) = e^x + 2$. The first three iterations, and the corresponding value of F and f are

k	x_k	$f(x_k)$	$f'(x_k)$
0	0	1	1
1	-0.3333	0.8276	0.0499
2	-0.3517	0.8272	0.00011998

Practically the global minimum place is $\bar{x} = -0.3517$. ∎

Here we deal only with one additional method, with the basics of gradient method or steepest descent method. The interested can find more material in the excellent book about numerical optimization [**NW2006**].

3.6.1 Steepest descent method

This method is a typical line search type method. This means the following. We pick up a point $x_0 \in \mathbb{R}^n$, choose a direction $d_0 \in \mathbb{R}^n$, and try to minimize f along the half-line $x_0 + \alpha d_0$. So we are looking for an α_0, which is the solution of the minimization problem:

$$\min_{\alpha \geq 0} f(x_0 + \alpha d_0).$$

The next iteration will be

$$x_1 = x_0 + \alpha_0 d_0.$$

In general, the $k+1$th iterate will be:

$$x_{k+1} = x_k + \alpha_k d_k.$$

We continue this process still we have an iterate where the derivative of f is zero.

Two questions are popping up immediately. How can we choose the direction d_k, and how can we choose the step length α_k in the kth iteration?[9]

The choice of the direction

There are several possibilities in the literature for the choice of the direction. Probably the best known direction is the steepest descent direction. This vector is resulted by the quite reasonable strategy, find that direction at a point from where one can go down in the fastest way. In other words, choose that direction, which has the biggest negative slope.

Definition 3.6.1 — Steepest descent direction. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The solution of the problem (if there exists)

$$\min_{\|d\|=1} f'(x)^T d \tag{3.12}$$

is called a **steepest descent direction of** f **at** x.

This definition makes sense in a pretty large class of functions as the next theorem says.

Theorem 3.6.1 Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. If $f'(x) \neq 0$, then the problem (3.12) has a unique solution, which is given by the formula

$$d = -\frac{f'(x)}{\|f'(x)\|}.$$

Proof. Let d be an arbitrary vector with norm one. From the Cauchy-Schwarz inequality we have

$$-f'(x)^T d \leq |-f'(x)^T d| \leq \|-f'(x)\| \cdot \|d\| = \|f'(x)\| \quad \Rightarrow \quad f'(x)^T d \geq -\|f'(x)\|.$$

Equality if and only if, when $-f'(x)$ and d are linearly dependent, that is to say, there is a scalar α different from zero for which $d = -\alpha f'(x)$. Because $\|d\| = 1$, we have the statement. ■

The choice of the step length

For the step length, the best choice is the solution of the problem

$$\min_{\alpha \geq 0} f(x + \alpha d).$$

However, its computational cost is too much.

It is reasonable to choose a numerically cheaper algorithm, which matches to the following two requirements:

- it gives sufficiently large decrease in the objective function during the resulted iteration step,
- it is long enough to ensure sufficient progress toward a local minimum.

The following rule complies with both the above prerequisites.

[9] The starting point x_0 can be chosen arbitrarily.

Definition 3.6.2 — Goldstein condition. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, $x \in \mathbb{R}^n$ is a point where the derivative of f is different from zero, and d be the steepest descent direction at x. A step size α fulfils the **Goldstein conditions** if

$$f(x) + (1-c)\alpha f'(x)^t d \leq f(x+\alpha d) \leq f(x) + c\alpha f'(x)^T d, \tag{3.13}$$

where $c \in]0, 0.5[$ is fixed in advance.

3.7 Exercises

Exercise 3.1 Give the floating point representation of the following numbers if $a = 2$, $t = 4$, $\ell = -2$, and $u = 3$.

a) 0.25 b) 0.125 c) -2.5

d) $\frac{1}{3}$ e) $\frac{1}{7}$ f) $\frac{1}{8}$

■

Exercise 3.2 Give the smallest, and the largest representable numbers, the machine epsilon, and the number of the positive representable floating-point numbers besides the given data below.

a) $a = 2, t = 4, \ell = -2$, and $u = 3$ b) $a = 2, t = 5, \ell = -2$, and $u = 2$

c) $a = 2, t = 4, \ell = -4$, and $u = 4$ d) $a = 2, t = 3, \ell = -2$, and $u = 2$

■

Exercise 3.3 Find the approximate solution of the non-linear equations below, using Bisection method, Fixed-point iteration, and Newton's method in the given interval.

a) $\sin(x) = 0.5x, \quad x \in [1,2]$

b) $e^x = 2x, \quad x \in [0,1]$

c) $x^3 - x^2 + 1 = 0, \quad x \in [-1,0]$

■

Exercise 3.4 Find the Lagrangian polynomial, which interpolates the given data.

a) $(-3,-6),(-2,-17),(-1,-8),(1,-2),(2,19)$

b) $(-3,-31),(-2,-8),(1,1),(2,22)$

c) $(-2,13),(-1,-4),(1,2)$

d) $(-2,-5),(-1,3),(0,1),(2,15)$

e) $(-1,4),(1,2),(2,10),(2,15)$

f) $(-2,38),(-1,5),(1,-1),(2,-10),(3,-7)$

■

Exercise 3.5 Find the line, which fits the best in least square sense to the following data.

a)

t_i	1	2	3	4	5
f_i	1	0	2	-1	1

b)

t_i	0	0.2	0.4	0.5
f_i	1	1.1	1.2	1.3

■

Exercise 3.6 Find the function in the given form, which fits the best in least square sense to the following data.

a)

t_i	-1	-0.5	0	0.5	1
f_i	1	0	-1.5	0.3	1

, $F(t) = x_1 + x_2 \sin(\pi t)$

b)

t_i	0.1	0.5	1.2	1.5	1
f_i	-0.6	1.5	2.5	2.9	1

, $F(t) = x_1 + x_2 \log(t)$

■

Exercise 3.7 Find the exact value and the numerical approximation of the following integrals using Midpoint formula, Trapezoidal formula, and Simpson formula. Give an estimation for the error for all the three cases.

a) $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} x \sin(x^2) dx$

b) $\int_0^1 x^5 - x^4 + 3x^2 + 1 dx$

c) $\int_{-1}^1 \sqrt{1-x^2} dx$

■

Bibliography

[BP1978] de Boor, Carl and Pinkus, Allan. *Proof of the conjectures of Bernstein and Erdős concerning the optimal nodes for polynomial interpolation*. J. Approx. Theory 24/4 (1978), 289–303.

[Kre1988] Kress, Rainer. *Numerical analysis*. Springer-Verlag, Berlin, 1998.

[NW2006] Nocedal, Jorge and Wright, Stephen J. *Numerical optimization*. Springer-Verlag, Berlin, 2006.

[QSS2007] Quarteroni, Alfio and Sacco, Riccardo and Saleri, Fausto. *Numerical mathematics*. Springer-Verlag, Berlin, 2007.

[UU2011] Ulbrich, Michael and Ulbrich, Stefan. *Nichtlineare Optimierung*. Birkhäuser, 2011.