Advanced Engineering Mathematics I
Part I
Linear Algebra and Differential Equations

Authors: Søren Enemark, Steen Markvorsen, and Karsten Schmidt

Translated by: Jesper Kampmann Larsen

Technical University of Denmark

Version: January 11, 2022

eNote 1

Complex Numbers

In this eNote we introduce and investigate the set of numbers $\mathbb{C}$, the complex numbers. Since $\mathbb{C}$ is considered to be an extension of $\mathbb{R}$, the eNote presumes general knowledge of the real numbers, including the elementary real functions such as the trigonometric functions and the natural exponential function. Finally elementary knowledge of vectors in the plane is taken for granted.

Updated 22.09.21. *by David Brander. Version* 29.05.16. *by Karsten Schmidt.*

1.1 Introduction

A simple quadratic equation such as $x^2 = 25$ has two real solutions, viz.

$$x = 5 \text{ and } x = -5 \, ,$$

since $5^2 = 25$ and $(-5)^2 = 25$. Likewise the equation $x^2 = 2$ has two solutions, viz.

$$x = \sqrt{2} \text{ and } x = -\sqrt{2} \, ,$$

since $\sqrt{2}^{\,2} = 2$ and $(-\sqrt{2})^{\,2} = 2$.

In the two examples above the right-hand sides were *positive*. When considering the equation

$$x^2 = k \; , \; k \in \mathbb{R}$$

we must be more careful; here everything depends on the sign of k . If $k \geq 0$, the equation has the solutions

$$x = \sqrt{k} \text{ and } x = -\sqrt{k} \, ,$$

since $\sqrt{k}^{\,2} = k$ and $(-\sqrt{k})^{\,2} = k$. But if $k < 0$ the equation has no solutions, since real numbers with negative squares do not exist.

But now we ask ourselves the question, is it possible to imagine a set of numbers larger than the set of real numbers; a set that includes all the real numbers but in addition *also* includes solutions to an equation like

$$x^2 = -1?$$

The equation should then in analogy to the equations above have two solutions

$$x = \sqrt{-1} \;\text{ and }\; x = -\sqrt{-1}\,.$$

Let us be bold and assume that this is in fact possible. We then choose to call this number $\mathrm{i} = \sqrt{-1}$. The equation $x^2 = -1$ then has two solutions, viz.

$$x = \mathrm{i} \;\text{ and }\; x = -\mathrm{i}$$

since, if we assume that the usual rules of algebra hold,

$$\mathrm{i}^2 = \sqrt{-1}^{\,2} = -1 \;\text{ and }\; (-\mathrm{i})^2 = (-1\cdot\sqrt{-1}\,)^{\,2} = (-1)^2(-\sqrt{-1})^2 = -1\,.$$

As we just mentioned, we make the further demand on the hypothetical number i, that one must be able to use the same algebraic rules that apply to the real numbers. We must e.g. be able to multiply i by a real number b and add this to another real number a. In this way a new kind of number z of the type

$$z = a + \mathrm{i}b\,, \quad (a, b) \in \mathbb{R}^2$$

emerges.

Below we describe how these ambitions about a larger set of numbers can be fulfilled. We look at how the structure of the set of numbers should be and at which rules apply. We call this set of numbers *the complex numbers* and use the symbol $\mathbb{C}$. $\mathbb{R}$ must be a proper subset of $\mathbb{C}$ — that is, $\mathbb{C}$ contains all of $\mathbb{R}$ *together with* the new numbers which fulfill the above ambitions that are impossible in $\mathbb{R}$. As we have already hinted $\mathbb{C}$ must be *two-dimensional* in the sense that a complex number contains *two* real numbers, a and b.

1.2 Complex Numbers Introduced as Pairs of Real Numbers

The common way of writing a complex number z is

$$z = a + \mathrm{i}b\,, \tag{1-1}$$

where a and b are real numbers and i is the new *imaginary* number that satisfies $\mathrm{i}^2 = -1$. This form is very practical in computation with complex numbers. But we have not really clarified the meaning of the expression (1-1). For what is the meaning of a product like $\mathrm{i}b$, and what does the addition $a + \mathrm{i}b$ mean?

A satisfactory way of introducing the complex numbers is as the *set of pairs of real numbers* (a, b). In this section we will show how in this set we can define arithmetic operations (addition, subtraction, multiplication and division) that fulfill the ordinary arithmetic rules for real numbers. This will turn out to give full credit to the form (1-1).

Definition 1.1 The Complex Numbers

The complex numbers $\mathbb{C}$ are defined as the set of ordered pairs of real numbers:

$$\mathbb{C} = \{(a, b) \mid a, b \in \mathbb{R}\} \tag{1-2}$$

equipt with the arithmetic rules described below.

As the symbol for an arbitrary complex number we will use the letter z .

Example 1.2

Here we show five different complex numbers:

$$z_1 = (2, 7)\,,\ z_2 = (7, 2)\,,\ z_3 = (0, 1)\,,\ z_4 = (-5, 0)\,,\ z_5 = (0, 0)\,.$$

First we introduce the arithmetic rule for the addition of complex numbers. Then subtraction as a special form of addition.

|||| Definition 1.3 Addition of Complex Numbers

Let $z_1 = (a, b)$ and $z_2 = (c, d)$ be two complex numbers.

The sum $z_1 + z_2$ is defined as

$$z_1 + z_2 = (a, b) + (c, d) = (a + c, b + d)\,. \tag{1-3}$$

|||| Example 1.4 Addition

For the two complex numbers $z_1 = (2, 7)$ and $z_2 = (4, -3)$ we have:

$$z_1 + z_2 = (2, 7) + (4, -3) = (2 + 4, 7 + (-3)) = (6, 4)\,.$$

The complex number $(0, 0)$ is *neutral* with respect to addition, since for every complex number $z = (a, b)$ we have:

$$z + (0, 0) = (a, b) + (0, 0) = (a + 0, b + 0) = (a, b) = z\,.$$

It is evident that $(0, 0)$ is the only complex number that is neutral with respect to addition.

For every complex number z there exists an *additive inverse* (also calld *opposite number*) denoted $-z$, which, when added to z, gives $(0, 0)$. The complex number $z = (a, b)$ has the additive inverse $-z = (-a, -b)$, since

$$(a, b) + (-a, -b) = (a + (-a), b + (-b)) = (a - a, b - b) = (0, 0)\,.$$

It is clear that $(-a, -b)$ is the only additive inverse for $z = (a, b)$, so the notation $-z$ is well-defined. By use of this, subtraction of complex numbers can be introduced as a special form of addition.

|||| Definition 1.5 Subtraction of Complex Numbers

For the two complex numbers z_1 and z_2 the difference $z_1 - z_2$ is defined as the sum of z_1 and *the additive inverse* for z_2 :

$$z_1 - z_2 = z_1 + (-z_2)\,. \tag{1-4}$$

Let us for two arbitrary complex numbers $z_1 = (a, b)$ and $z_2 = (c, d)$ calculate the difference $z_1 - z_2$ using definition 1.5:

$$z_1 - z_2 = (a, b) + (-c, -d) = (a + (-c), b + (-d)) = (a - c, b - d)\,.$$

This gives the simple formula

$$z_1 - z_2 = (a - c, b - d)\,. \tag{1-5}$$

|||| Example 1.6 Subtraction of Complex Numbers

For the two complex numbers $z_1 = (5, 2)$ and $z_2 = (4, -3)$ we have:

$$z_1 - z_2 = (5 - 4, 2 - (-3)) = (1, 5)\,.$$

While addition and subtraction appear to be simple and natural, multiplication and division of complex numbers appear to be more odd. Later we shall see that all the four arithmetic rules have geometrical equivalents in the so-called *complex plane* that constitutes the graphical representation of the complex numbers. But first we must accept the definitions at their face value. First we give the definition of multiplication. Then follows the definition of division as a special form of multiplication.

|||| Definition 1.7 Multiplication of Complex Numbers

Let $z_1 = (a, b)$ and $z_2 = (c, d)$ be two complex numbers.

The product $z_1\, z_2$ is defined as

$$z_1\, z_2 = z_1 \cdot z_2 = (ac - bd, ad + bc)\,. \tag{1-6}$$

|||| Example 1.8 Multiplication of Complex Numbers

For the two complex numbers $z_1 = (2,3)$ and $z_2 = (1,-4)$ we have:

$$z_1\,z_2 = (2,3)\cdot(1,-4) = (2\cdot 1 - (3\cdot(-4)), 2\cdot(-4) + 3\cdot 1) = (14,-5)\,.$$

The complex number $(1,0)$ is *neutral* with respect to multiplication, since for every complex number $z = (a,b)$ we have that:

$$z\cdot(1,0) = (a,b)\cdot(1,0) = (a\cdot 1 - b\cdot 0\,, a\cdot 0 + b\cdot 1) = (a,b) = z\,.$$

It is clear that $(1,0)$ is the only complex number that is neutral with respect to multiplication.

For every complex number z apart from $(0,0)$ there exists a unique reciprocal number that when multiplied by the given number gives $(1,0)$. It is denoted $\frac{1}{z}$. The complex number (a,b) has the reciprocal number

$$\frac{1}{z} = \left(\frac{a}{a^2+b^2}\,, -\frac{b}{a^2+b^2}\right)\,, \tag{1-7}$$

since

$$(a,b)\cdot\left(\frac{a}{a^2+b^2}, -\frac{b}{a^2+b^2}\right) = \left(\frac{a^2}{a^2+b^2} + \frac{b^2}{a^2+b^2}\,, -\frac{ab}{a^2+b^2} + \frac{ba}{a^2+b^2}\right) = (1,0)\,.$$

|||| Exercise 1.9

Show that every complex number $z \neq (0,0)$ has exactly one reciprocal number.

By the use of reciprocal numbers we can now introduce division as a special form of multiplication.

|||| Definition 1.10 Division of Complex Numbers

Let z_1 and z_2 be arbitrary complex numbers, where $z_2 \neq (0,0)$.

The quotient $\frac{z_1}{z_2}$ is defined as the product of z_1 and *the reciprocal number* $\frac{1}{z_2}$ for z_2 :

$$\frac{z_1}{z_2} = z_1 \cdot \frac{1}{z_2} \ . \tag{1-8}$$

Let us for two arbitrary complex numbers $z_1 = (a,b)$ and $z_2 = (c,d) \neq (0,0)$ compute the quotient $\frac{z_1}{z_2}$ from the Definition 1.10:

$$z_1 \cdot \frac{1}{z_2} = (a,b)\left(\frac{c}{c^2+d^2}\,,\, -\frac{d}{c^2+d^2}\right) = \left(\frac{ac+bd}{c^2+d^2}, \frac{bc-ad}{c^2+d^2}\right) .$$

From this we get the following formula for division:

$$\frac{z_1}{z_2} = \left(\frac{ac+bd}{c^2+d^2}\,,\, \frac{bc-ad}{c^2+d^2}\right) . \tag{1-9}$$

|||| Example 1.11 Division of Complex Numbers

Consider two complex numbers $z_1 = (1,2)$ and $z_2 = (3,4)$.

$$\frac{z_1}{z_2} = \left(\frac{1\cdot 3+2\cdot 4}{3^2+4^2}, \frac{2\cdot 3-1\cdot 4}{3^2+4^2}\right) = \left(\frac{11}{25}, \frac{2}{25}\right) .$$

We end this section by showing that the complex numbers, with the above arithmetic operations, fulfill the computational rules known from the real numbers.

|||| Theorem 1.12 Properties of Complex Numbers

The complex numbers fulfill the following computational rules:

1. Commutative rule for addition: $z_1 + z_2 = z_2 + z_1$
2. Associative rule for addition: $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$
3. The number $(0,0)$ is neutral with respect to addition
4. Every z has an opposite number $-z$ where $z + (-z) = (0,0)$
5. Commutative rule for multiplication: $z_1\, z_2 = z_2\, z_1$
6. Associative rule for multiplication: $(z_1\, z_2)\, z_3 = z_1\, (z_2\, z_3)$
7. The number $(1,0)$ is neutral with respect to multiplication
8. Every $z \neq (0,0)$ has a reciprocal number $\frac{1}{z}$ where $z \cdot \frac{1}{z} = (1,0)$
9. Distributive rule: $z_1\, (z_2 + z_3) = z_1\, z_2 + z_1\, z_3$

|||| Proof

Let us look at property 1, the commutative rule. Given two complex numbers $z_1 = (a,b)$ and $z_2 = (c,d)$. We see that

$$z_1 + z_2 = (a + c, b + d) = (c + a, d + b) = z_2 + z_1\,.$$

To establish the second equality sign we have used that for both the first and the second coordinates the commutative rule for addition of real numbers applies. By this it is seen that the commutative rule also applies to complex numbers.

In the proof of the properties $2, 5, 6$ and 9 we similarly use the fact that the corresponding rules apply to the real numbers. The details are left to the reader. For the properties $3, 4, 7$ and 8 we refer to treatment above in this section.

■

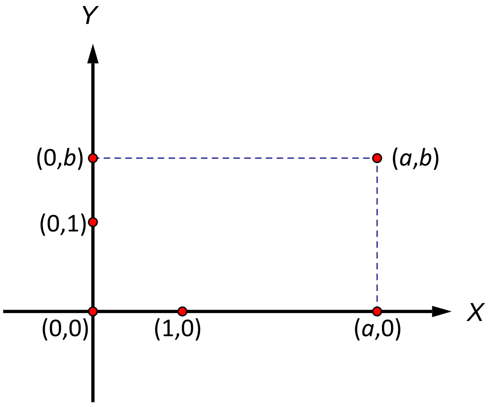


Figure 1.1: Six complex numbers in the (x, y)-plane

1.3 Complex Numbers in Rectangular Form

Since to every ordered pair of real numbers corresponds a unique point in the (x, y)-plane and vice versa, $\mathbb{C}$ can be considered to be the set of points in the (x, y)-plane. Figure 1.1 shows six points in the (x, y)-plane, i.e. six complex numbers.

In the following we will change our manner of writing complex numbers.

First we identify all complex numbers of the type $(a, 0)$, i.e. the numbers that lie on the x-axis, with the corresponding real number a . In particular the number $(0, 0)$ is written as 0 and the number $(1, 0)$ as 1 . Note that this will not be in conflict with the arithmetic rules for complex numbers and the ordinary rules for real numbers, since

$$(a, 0) + (b, 0) = (a + b, 0 + 0) = (a + b, 0)$$

and

$$(a, 0) \cdot (b, 0) = (a \cdot b - 0 \cdot 0\,, a \cdot 0 + 0 \cdot b) = (ab, 0)\,.$$

In this way the x-axis can be seen as an ordinary real number axis and is called the *real axis*. In this way the real numbers can be seen as a subset of the complex numbers. That the y-axis is called the *imaginary axis* is connected to the extraordinary properties of the complex number i which we now introduce and investigate.

|||| Definition 1.13 The Number i

By the complex number i we understand the number $(0,1)$.

A decisive motivation for the introduction of complex numbers was the wish for a set of numbers that contained the solution to the equation

$$x^2 = -1 \,.$$

With the number i we have got such a solution because:

$$\mathrm{i}^2 = \mathrm{i}\cdot\mathrm{i} = (0,1)\cdot(0,1) = (0\cdot 0 - 1\cdot 1\,, 0\cdot 1 + 1\cdot 0) = (-1,0) = -1\,.$$

|||| Theorem 1.14 Complex Numbers in Rectangular Form

Every complex number $z = (a,b)$ can be written in the form

$$z = a + \mathrm{i}\cdot b = a + \mathrm{i}b\,. \tag{1-10}$$

This way of writing the complex number is called *the rectangular form* of z .

|||| Proof

The proof consists of simple manipulations in which we use the new way of writing numbers of this type.

$$(a,b) = (a,0) + (0,b) = (a,0) + (0,1)\cdot(b,0) = a + \mathrm{i}\,b\,.$$

■

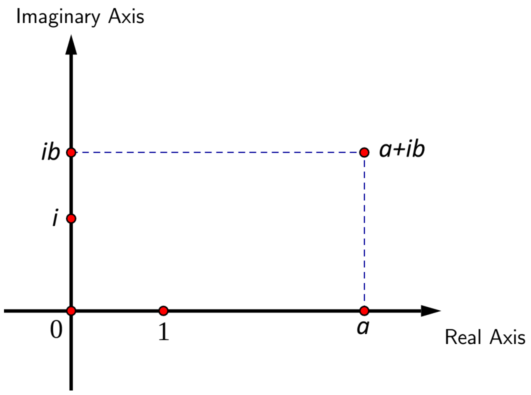


Figure 1.2: Six complex numbers in rectangular form in the complex number plane

Since $0 = (0,0)$ is neutral with respect to addition, and $1 = (1,0)$ is neutral with respect to multiplication, the following identities apply:

$$0 + z = z \text{ and } 1z = z\,.$$

Furthermore it is easily seen that

$$0z = 0\,.$$

Let us now consider all complex numbers of the type $(0, b)$. Since

$$(0,b) = 0 + \mathrm{i}b = \mathrm{i}b\,,$$

i can be understood as the unit of the y-axis, and therefore we refer to i as the *imaginary unit*. From this comes the name the *imaginary axis* for the y-axis.

In Figure 1.2 we see an update of the situation from Figure 1.1, where numbers are given in their rectangular form.

All real numbers are complex but not all complex numbers are real!

|||| Method 1.15 Computation Using the Rectangular form

A decisive advantage arising from the rectangular form of complex numbers is that one does not have to remember the formulas for the arithmetic rules for addition, subtraction, multiplication and division given in the definitions 1.3, 1.5, 1.7 and 1.10. All computations can be carried out by following the usual arithmetic rules for real numbers and treating the number i as one would treat a real variable — with the difference, though, that we replace i^2 by -1 .

In the following example it is shown how multiplication can be carried out through ordinary computation with the rectangular form of the factors.

|||| Example 1.16 Multiplication Using the Rectangular Form

We compute the product of two complex numbers given in rectangular form $z_1 = a + ib$ and $z_2 = c + \mathrm{i}d$:

$$\begin{aligned} z_1 z_2 &= (a + \mathrm{i}b)(c + \mathrm{i}d) = ac + \mathrm{i}ad + \mathrm{i}bc + \mathrm{i}^2 bd = ac + \mathrm{i}ad + \mathrm{i}bc - bd \\ &= (ac - bd) + \mathrm{i}(ad + bc) . \end{aligned}$$

The result corresponds to the definition, see Definition 1.7!

|||| Exercise 1.17

Prove that the following rule for real numbers — the so-called *zero rule* — also applies to complex numbers: "'A product is 0 if and only if at least one of factors is 0 ."'

|||| Remark 1.18 Powers of Complex Numbers

The property 6 in Theorem 1.12 gives us the possibility to introduce integer powers of complex numbers, corresponding to integer powers of real numbers. In the following let n be a natural number.

1. $z^1 = z\,,\ z^2 = z\cdot z\,,\ z^3 = z\cdot z\cdot z$ etc.

2. By definition $z^0 = 1\,.$

3. Finally we put $z^{-n} = \dfrac{1}{z^n}\,.$

It is easily shown that the usual rules for computations with integer powers of real numbers also apply for integer powers of complex numbers:

$$z^n\,z^m = z^{n+m} \quad\text{and}\quad (z^n)^m = z^{n\,m}\,.$$

We end this section by introducing the concepts *real part* and *imaginary part* of complex numbers.

|||| Definition 1.19 Real Part and Imaginary Part

Given a complex number z in rectangular form $z = a + ib$. By the *real part* of z we understand the real number

$$\operatorname{Re}(z) = \operatorname{Re}(a+ib) = a\,, \tag{1-11}$$

and by the *imaginary part* of z we understand the real number

$$\operatorname{Im}(z) = \operatorname{Im}(a+ib) = b\,. \tag{1-12}$$

The expression *rectangular form* refers to the position of the number in the complex number plane, where $\mathrm{Re}(z)$ is the number's perpendicular drop point on the real axis, and $\mathrm{Im}(z)$ its perpendicular drop point on the imaginary axis. In short the real part is the first coordinate of the number while the imaginary part is the second coordinate of the number.

Note that every complex number z can be written in rectangular form like this:

$$z = \mathrm{Re}(z) + \mathrm{i}\,\mathrm{Im}(z)\,.$$

|||| Example 1.20 Real Part and Imaginary Part

Three complex numbers are given by

$$z_1 = 3 - 2\mathrm{i}\,,\; z_2 = \mathrm{i}5\,,\, z_3 = 25 + \mathrm{i}\,.$$

Find the real part and the imaginary part of each number.

$$\begin{aligned}
\mathrm{Re}(z_1) &= 3\,,\; \mathrm{Im}(z_1) = -2\\
\mathrm{Re}(z_2) &= 0\,,\; \mathrm{Im}(z_2) = 5\\
\mathrm{Re}(z_3) &= 25\,,\; \mathrm{Im}(z_3) = 1
\end{aligned}$$

Two complex numbers in rectangular form are *equal* if and only if both their real parts and imaginary parts are equal.

1.4 Conjugation of Complex Numbers

|||| Definition 1.21 Conjugation

Let z be a complex number with the rectangular form $z = a + ib$. By the conjugated number corresponding to z we understand the complex number $\overline{z}$ given by

$$\overline{z} = a - ib\,. \tag{1-13}$$

Conjugating a complex number corresponds to reflecting the number in the real axis as shown in Figure 1.3.

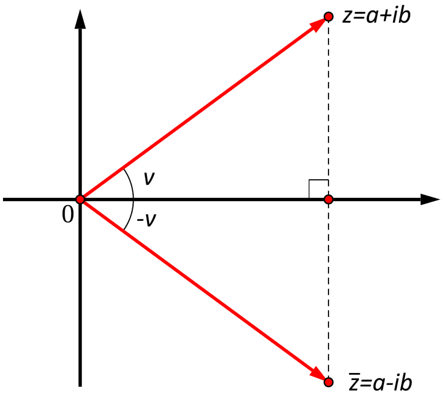


Figure 1.3: Reflection in the real axis

It is obvious that the conjugate number of a conjugate number is the original number:

$$\overline{\overline{z}} = z\,. \tag{1-14}$$

Furthermore the following useful formula for the product of complex number and its conjugate applies:

$$z \cdot \overline{z} = |\,z\,|^2 \tag{1-15}$$

which is shown by simple calculation.

In the following method we show a smart way of finding the rectangular form of a fraction when the denominator is not real: we use the fact that the product of a number $z = a + ib$ and its conjugate $z = a - ib$ is *always* a real number, cf. (1-15).

|||| Method 1.22 Finding the rectangular form of a complex fraction

The way to remember: *Multiply the numerator and the denominator by the conjugate of the denominator*. Here the denominator is written in its rectangular form:

$$\frac{z}{a+ib} = \frac{z(a-ib)}{(a+ib)(a-ib)} = \frac{z(a-ib)}{a^2+b^2}\,.$$

An example:

$$\frac{2-\mathrm{i}}{1+\mathrm{i}} = \frac{(2-\mathrm{i})(1-\mathrm{i})}{(1+\mathrm{i})(1-\mathrm{i})} = \frac{1-3\mathrm{i}}{1^2+1^2} = \frac{1-3\mathrm{i}}{2} = \frac{1}{2} - \frac{3}{2}\,\mathrm{i}\,.$$

In conjugation in connection with the four ordinary arithmetic operations the following rules apply.

|||| Theorem 1.23 Arithmetic Rules for Conjugation

1. $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$
2. $\overline{z_1 - z_2} = \overline{z_1} - \overline{z_2}$
3. $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$
4. $\overline{(z_1/z_2)} = \overline{z_1}/\overline{z_2} \quad , z_2 \neq 0\,.$

|||| Proof

The proof is carried out by simple transformation using the rectangular form of the numbers. As an example we show the first formula. Suppose that $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$. Then:

$$\begin{aligned}\overline{z_1 + z_2} &= \overline{(a_1 + \mathrm{i}b_1) + (a_2 + \mathrm{i}b_2)} = \overline{(a_1 + a_2) + \mathrm{i}(b_1 + b_2)}\\ &= (a_1 + a_2) - \mathrm{i}(b_1 + b_2) = (a_1 - \mathrm{i}b_1) + (a_2 - \mathrm{i}b_2)\\ &= \overline{z_1} + \overline{z_2}\,.\end{aligned}$$

■

Finally we note that all complex numbers on the real axis are identical with their conjugate number and that they are the only complex numbers that fulfill this condition. Therefore we can state a criterion for whether a given number in a set of complex numbers is real:

|||| Theorem 1.24 The Real Criterion

Let A be a subset of $\mathbb{C}$, and let $A_{\mathbb{R}}$ denote the subset of A that consists of real numbers. Then:

$$A_{\mathbb{R}} = \{z \in A \,|\, \overline{z} = z\}\ .$$

|||| Proof

Let z be an arbitrary number in $A \subseteq \mathbb{C}$ with rectangular form $z = a + ib$. Then:

$$\overline{z} = z \Leftrightarrow a - \mathrm{i}b = a + \mathrm{i}b \Leftrightarrow 2\mathrm{i}b = 0 \Leftrightarrow b = 0 \Leftrightarrow z \in A_{\mathbb{R}}\,.$$

■

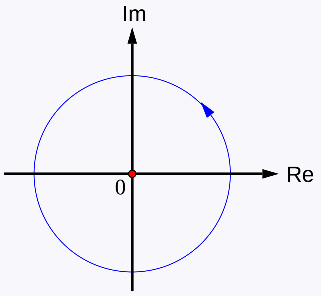
1.5 Polar Coordinates

The obvious way of stating a point (or a position vector) in an ordinary (x, y)-coordinate system is by the point's *rectangular*, i.e. *orthogonal*, coordinates (a, b). In many situations it is, however, useful to be able to determine a point by its *polar coordinates*, consisting of its *distance* to $(0, 0)$ together with its *direction angle* from the x-axis to its position vector. The direction angle is then positive if it is measured counter-clockwise and negative if measured clockwise.

Analogously, we now introduce polar coordinates for complex numbers. Let us first be absolutely clear about the orientation of the complex number plane.

|||| Definition 1.25 Orientation of the Complex Number Plane

The orientation of the complex number plane is determined by a circle with its centre at the origen being traversed *counter-clockwise*.



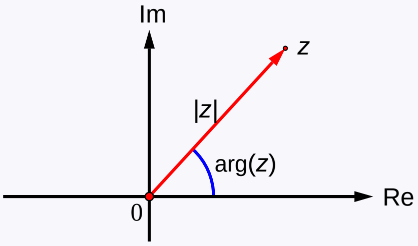
The ingredients in the polar coordinates of complex numbers are (as mentioned above) its distance to $(0, 0)$ called the *absolute value*, and its direction angle called the *argument*. We now introduce these two quantities.

|||| Definition 1.26 Absolute Value and Argument

Given a complex number z .

By the *absolute value* of z we understand the length of the corresponding position vector. The absolute value is written $|\,z\,|$ and is also called the *modulus* or *numerical value*.

Suppose $z \neq 0$. Every angle from the positive part of the real axis to the position vector for z is called an *argument* for z and is denoted $\arg(z)$. The angle is positive or negative relative to the orientation of the complex number plane.



The pair

$$(\,|\,z\,|\,,\arg(z)\,)$$

of the absolute value of z and an argument for z will collectively be called the *polar coordinates* of the number.

Note that the argument for a number z is not unique. If you add 2π to an arbitrary argument for z, you get a new valid direction angle for z and therefore a valid argument. Therefore a complex number has infinitely many arguments corresponding to turning an integer number of times extra clockwise or counter-clockwise in order to reach the same point again.

You can always choose an argument for z that lies in the interval from $-\pi$ to π. Traditionally this argument is given a preferential position. It is called the *principal value* of the argument.

|||| Definition 1.27 Principal Value

Given a complex number z that is not 0. By the *principal argument* $\operatorname{Arg}(z)$ for z we understand the uniquely determined argument for z that satisfies:

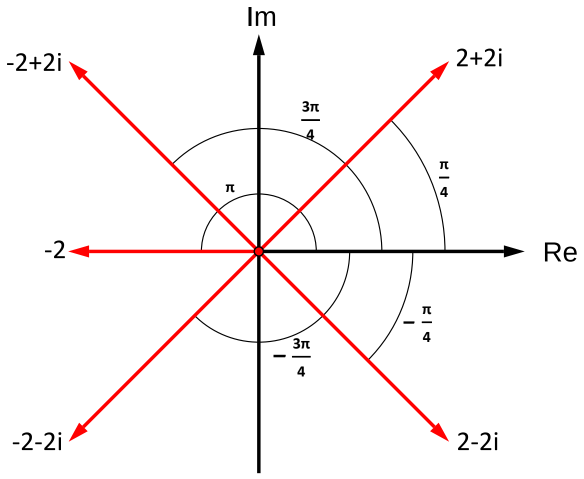
$$\arg(z) \in \,]-\pi, \pi]\,.$$

We denoted the *principal value* with a capital initial $\operatorname{Arg}(z)$ as compared to $\arg(z)$ that denotes an *arbitrary* argument. All arguments for a complex number z are then given by

$$\arg(z) = \operatorname{Arg}(z) + p \cdot 2\pi \ , \ p \in \mathbb{Z}\,. \tag{1-16}$$

Two complex numbers are *equal* if and only if both their absolute values and the principal arguments are equal.

|||| Example 1.28 Principal Arguments



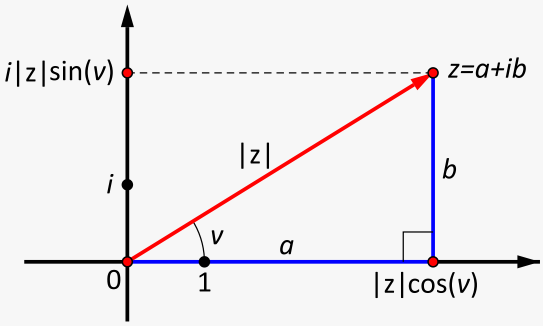
The figure shows five complex numbers, four of which lie on the lines through $(0,0)$ bisecting the four quadrants. We read:

- $2+2\mathrm{i}$ has the principal argument $\frac{\pi}{4}$,
- $2-2\mathrm{i}$ has the principal argument $-\frac{\pi}{4}$,
- $-2+2\mathrm{i}$ has the principal argument $\frac{3\pi}{4}$,
- $-2-2\mathrm{i}$ has the principal argument $-\frac{3\pi}{4}$, and
- -2 has the principal argument π .

Whether it is advantageous to use the rectangular format of the complex numbers or their polar form depends on the situation at hand. In Method (1.29) it is demonstrated how one can shift between the two forms.

|||| Method 1.29 Rectangular and Polar Coordinates

We consider a complex number $z \neq 0$ that has the rectangular form $z = a + ib$ and an argument v :



1. The rectangular form is computed from the polar coordinates like this:

$$a = |z| \cos(v) \quad \text{and} \quad b = |z| \sin(v)\,. \tag{1-17}$$

2. The absolute value is computed from the rectangular form like this :

$$|z| = \sqrt{a^2 + b^2}\,. \tag{1-18}$$

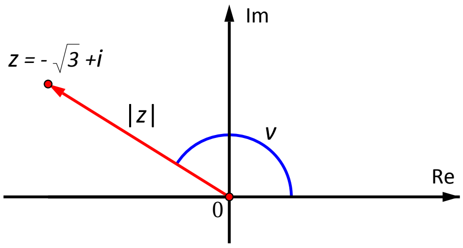
3. An argument is computed from the rectangular form by finding an angle v that satisfies *both* of the following equations:

$$\cos(v) = \frac{a}{|z|} \quad \text{and} \quad \sin(v) = \frac{b}{|z|}\,. \tag{1-19}$$

When z is drawn in the first quadrant it is evident that the computational rules (1-17) and (1-19) are derived from well-known formulas for cosine and sine to acute angles in right-angled triangles and (1-18) from the theorem of Pythagoras. By using the same formulas it can be shown that the introduced methods are valid regardless of the quadrant in which z lies.

|||| Example 1.30 From Rectangular to Polar Form

Find the polar coordinates for the number $z = -\sqrt{3} + \mathrm{i}$.



We use the rules in Method 1.29. Initially we identify the real and the imaginary parts of z as

$$a = -\sqrt{3} \quad \text{and} \quad b = 1.$$

First we determine the absolute value:

$$|\,z\,| = \sqrt{a^2 + b^2} = \sqrt{(-\sqrt{3})^{\,2} + 1^2} = \sqrt{3+1} = 2\,.$$

Then the argument is determined. From the equation

$$\cos(v) = \frac{a}{|\,z\,|} = -\frac{\sqrt{3}}{2}$$

we get two possible principal arguments for z, viz.

$$v = \frac{5\pi}{6} \;\text{ and }\; v = -\frac{5\pi}{6}\,.$$

From the figure it is seen that z lies in the second quadrant, and the correct principal argument must therefore be the first of these possibilities. But this can also be determined without inspection of the figure, since also the equation

$$\sin(v) = \frac{1}{2}$$

must be fulfilled. From this we also get two possible principal arguments for z, viz.

$$v = \frac{\pi}{6} \;\text{ and }\; v = \frac{5\pi}{6}\,.$$

Since only $v = \frac{5\pi}{6}$ satisfies both equations, we see that $\operatorname{Arg}(z) = \frac{5\pi}{6}\,.$

Thus we have found the set of polar coordinates for z :

$$(|z|, \operatorname{Arg}(z)) = \left(2, \frac{5\pi}{6}\right).$$

We end this section with the important product rules for absolute values and arguments.

|||| Theorem 1.31 The Product Rule for Absolute Values

The absolute value of the product of two complex numbers z_1 and z_2 is found by

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|. \tag{1-20}$$

From Theorem 1.31 we get the corollary

|||| Corollary 1.32

The absolute value for the quotient of two complex numbers z_1 and z_2 where $z_2 \neq 0$ is found by

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}. \tag{1-21}$$

The absolute value of the nth power of a complex number z is for every $n \in \mathbb{Z}$ given by

$$|{z_1}^n| = |z_1|^n. \tag{1-22}$$

|||| Exercise 1.33

Write down in words what the formulas (1-20), (1-21) and (1-22) say and prove them.

|||| Theorem 1.34 The Product Rule for Arguments

Given two complex numbers $z_1 \neq 0$ and $z_2 \neq 0$ (which also means $z_1 z_2 \neq 0$). Then if v_1 is an argument for z_1 and v_2 is an argument for z_2, then $v_1 + v_2$ is an argument for the product $z_1 z_2$.

|||| Corollary 1.35

Given two complex numbers $z_1 \neq 0$ and $z_2 \neq 0$.Then:

1. If v_1 is an argument for z_1 and v_2 is an argument for z_2, then $v_1 - v_2$ is an argument for the fraction $\dfrac{z_1}{z_2}$.
2. If v is an argument for z, then $n \cdot v$ is an argument for the power z^n.

|||| Exercise 1.36

Prove Theorem 1.34 and Corollary 1.35.

1.6 Geometric Understanding of the Four Computational Operations

We started by introducing addition, subtraction, multiplication and division of complex numbers as algebraic operations carried out on pairs of real numbers (a, b), see definitions 1.3, 1.5, 1.7 and 1.10. Then we showed that the rectangular form of the complex numbers $a + ib$ leads to a more practical way of computation: One can compute with complex numbers just as with real numbers, as long as the number i is treated as a real parameter and it is understood that $\mathrm{i}^2 = -1$. In this section we shall see that the computational operations can also be viewed as geometrical constructs.

The first exact description of the complex numbers was given by the Norwegian surveyor Caspar Wessel in 1796. Wessel introduced complex numbers as line segments with given lengths and directions, that is what we now call vectors in the plane. Therefore computations with complex numbers were geometric operations carried out on

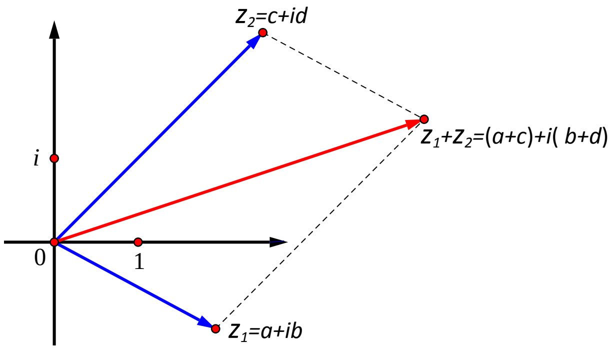


Figure 1.4: Addition by the method of parallelograms

vectors. In the following we recollect the ideas in the definition of Wessel. It is easy to see the equivalence between the algebraic and geometric representations of addition and subtraction — it is more demanding to understand the equivalence when it comes to multiplication and division.

|||| Theorem 1.37 Geometric Addition

Addition of two complex numbers z_1 and z_2 can be obtained geometrically in the following way:

The position vector for $z_1 + z_2$ is the sum of the position vectors for z_1 and z_2 . (See Figure 1.4).

|||| Proof

Suppose that z_1 and z_2 are given in rectangular form as $z_1 = a + ib$ and $z_2 = c + id$. Then the position vector for z_1 has the coordinates (a, b) and the position vector for z_2 has the coordinates (c, d) . The sum of the two position vectors is then $(a + c, b + d)$, being the coordinates of the position vector for the complex number $(a + c) + \mathrm{i}(b + d)$. Since we have that $z_1 + z_2 = (a + \mathrm{i}b) + (c + \mathrm{i}d) = (a + c) + \mathrm{i}(b + d)$, we have proven the theorem.

■

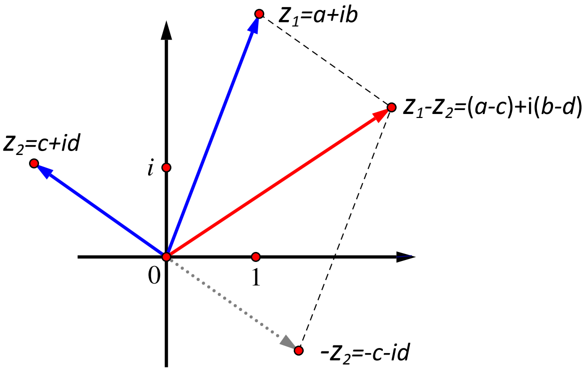


Figure 1.5: Subtraction by the method of parallelograms

Geometric subtraction is given as a special form of geometric addition: The position vector for $z_1 - z_2$ is the sum of the position vectors for z_1 and the opposite vector to the position vector for z_2. This is illustrated in Figure 1.5

While in the investigation of geometrical addition (and subtraction) we have used the rectangular form of complex numbers, in the treatment of geometric multiplication (and division) we shall need their polar coordinates.

|||| Theorem 1.38 Geometrical Multiplication

Given two complex numbers z_1 and z_2 that are both different from 0 (which also means that $z_1 z_2 \neq 0$). Multiplication of z_1 and z_2 can be obtained geometrically in the following way:

1. The absolute value of the product $z_1 z_2$ is found by multiplication of the absolute value of z_1 by the absolute value of z_2.

2. An argument for the product $z_1 z_2$ is found by adding an argument for z_1 and an argument for z_2.

|||| Proof

First part of the theorem appears from Theorem 1.31 while the second part is evident from Theorem 1.34.

■

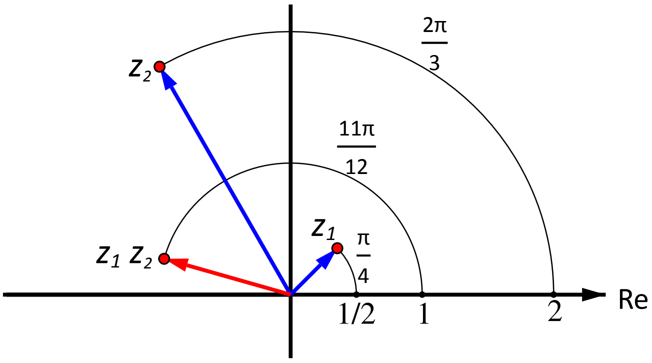


Figure 1.6: Multiplication

|||| Example 1.39 Multiplication by Use of Polar Coordinates

Two complex numbers z_1 and z_2 are given by the polar coordinates $\left(\frac{1}{2}, \frac{\pi}{4}\right)$ and $\left(2, \frac{2\pi}{3}\right)$, respectively. (Figure 1.6,)

We compute the product of z_1 and z_2 by the use of their absolute values and arguments:

$$|\,z_1 z_2\,| = |\,z_1\,|\,|\,z_2\,| = \frac{1}{2} \cdot 2 = 1$$

$$\arg(z_1 z_2) = \arg(z_1) + \arg(z_2) = \frac{\pi}{4} + \frac{2\pi}{3} = \frac{11\pi}{12}\,.$$

Thus the product $z_1 z_2$ is the complex number that has the absolute value 1 and the argument $\frac{11\pi}{12}$.

Note that it is important to observe whether a set of coordinates is given in *rectangular* or in *polar* form.

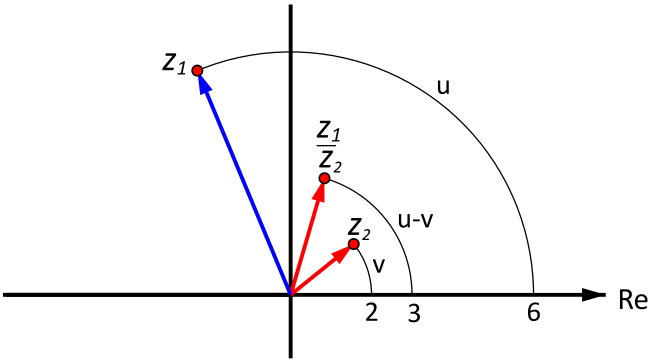


Figure 1.7: Division

|||| Example 1.40 Division by Use of Polar Coordinates

The numbers z_1 and z_2 are given by $|z_1| = 6$ with $\arg(z_1) = u$ and $|z_2| = 2$ with $\arg(z_2) = v$ respectively. Then $\dfrac{z_1}{z_2}$ can be determined as

$$\left|\frac{z_1}{z_2}\right| = \frac{6}{2} = 3 \text{ and } \arg\left(\frac{z_1}{z_2}\right) = u - v\,.$$

1.7 The Complex Exponential Function

The ordinary exponential function $x \mapsto \mathrm{e}^x\,,\ x \in \mathbb{R}$ has, as is well known, the characteristic properties,

1. $\mathrm{e}^0 = 1\,,$
2. $\mathrm{e}^{x_1+x_2} = \mathrm{e}^{x_1} \cdot \mathrm{e}^{x_2}$ for all $x_1, x_2 \in \mathbb{R}$, and
3. $(\mathrm{e}^x)^n = \mathrm{e}^{nx}$ for all $n \in \mathbb{Z}$ and $x \in \mathbb{R}\,.$

In this section we will introduce a particularly useful extension of the real exponential function to a complex exponential function, that turns out to follow the same rules of computation as its real counterpart.

|||| Definition 1.41 Complex Exponential Function

By the complex exponential function $\exp_{\mathbb{C}}$ we understand a function that to each number $z \in \mathbb{C}$ with the rectangular form $z = x + \mathrm{i}y$ attaches the number

$$\exp_{\mathbb{C}}(z) = \exp_{\mathbb{C}}(x + \mathrm{i}y) = \mathrm{e}^x \cdot (\cos(y) + \mathrm{i}\sin(y))\,, \tag{1-23}$$

where e (about $2.7182818\ldots$) is base for the real natural exponential function.

Since we for every *real* number x get

$$\exp_{\mathbb{C}}(x) = \exp_{\mathbb{C}}(x + \mathrm{i} \cdot 0) = \mathrm{e}^x\,(\cos(0) + \mathrm{i}\sin(0)) = \mathrm{e}^x\,,$$

we see that the complex exponential function is everywhere on the real axis identical to the real exponential function. Therefore we do not risk a contradiction when we in the following allow (and often use) the way of writing

$$\exp_{\mathbb{C}}(z) = \mathrm{e}^z \;\text{ for }\; z \in \mathbb{C}\,. \tag{1-24}$$

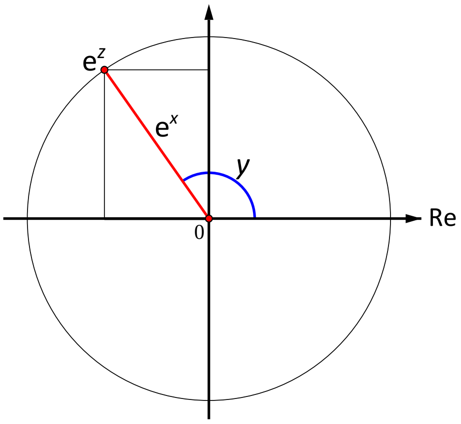


Figure 1.8: Geometric Interpretation of e^z

We now consider the complex number e^z where z is an arbitrary complex number with the rectangular form $z = x + \mathrm{i}y$. Then (by use of Theorem 1.31) we see that

$$|\mathrm{e}^z| = |\mathrm{e}^x\,(\cos(y) + \mathrm{i}\sin(y))| = |\mathrm{e}^x|\,|(\cos(y) + \mathrm{i}\sin(y))| = |\mathrm{e}^x| = \mathrm{e}^x\,. \tag{1-25}$$

Furthermore (by use of Theorem 1.34) we see that

$$\arg\left(\mathrm{e}^z\right) = \arg\left(\mathrm{e}^x\right) + \arg\left(\cos(y) + \mathrm{i}\sin(y)\right) = 0 + y = y\,. \tag{1-26}$$

The polar coordinates for $z = x + \mathrm{i}y$ are then (e^x, y), which is illustrated in Figure 1.8.

For the trigonometric functions $\cos(x)$ and $\sin(x)$ we know that for every integer p $\cos(x + p2\pi) = \cos(x)$ and $\sin(x + p2\pi) = \sin(x)$. If the graph for $\cos(x)$ or $\sin(x)$ is displaced by an arbitrary multiple of 2π, it will be mapped onto itself. Therefore the functions are called *periodic* having a *period* of 2π.

A similar phenomenon is seen for the complex exponential function. It has the *imaginary* period $\mathrm{i}\,2\pi$. This is closely connected to the periodicity of the trigonometric functions as can be seen in the proof of the following theorem.

|||| Theorem 1.42 Periodicity of e^z

For every complex number z and every integer p:

$$\mathrm{e}^{z+\mathrm{i}p2\pi} = \mathrm{e}^z . \tag{1-27}$$

|||| Proof

Suppose that z has the rectangular form $z = x + \mathrm{i}y$ and $p \in \mathbb{Z}$.

Then:

$$\begin{aligned}\mathrm{e}^{z+\mathrm{i}p2\pi} &= \mathrm{e}^{x+\mathrm{i}(y+p2\pi)}\\ &= \mathrm{e}^x\big(\cos(y+p2\pi) + \mathrm{i}\sin(y+p2\pi)\big) = \mathrm{e}^x\big(\cos(y) + \mathrm{i}\sin(y)\big)\\ &= \mathrm{e}^z .\end{aligned}$$

By this the theorem is proved.

■

In the following example the periodicity of the complex exponential function is illustrated.

|||| Example 1.43 Exponential Equation

Determine all solutions to the equation

$$e^z = -\sqrt{3} + \mathrm{i}\,. \tag{1-28}$$

First we write z in rectangular form: $z = x + \mathrm{i}y$. In Example 1.30 we found that the right-hand side in (1-28) has the absolute value $|z| = 2$ and the principal argument $v = \frac{5\pi}{6}$. Since the left-hand and the right-hand sides must have the same absolute value and the same argument, apart from an arbitrary multiple of 2π , we get

$$|\,e^z\,| = |-\sqrt{3} + \mathrm{i}\,| \Leftrightarrow e^x = 2 \;\;\Leftrightarrow\;\; x = \ln(2)$$

$$\arg(e^z) = \arg(-\sqrt{3} + \mathrm{i}) \Leftrightarrow y = v + p2\pi = \frac{5\pi}{6} + p2\pi\ ,\ p \in \mathbb{Z}\,.$$

All solutions for (1-28) are then

$$z = x + \mathrm{i}y = \ln(2) + \mathrm{i}\left(\frac{5\pi}{6} + p2\pi\right)\ ,\ p \in \mathbb{Z}\,.$$

We end this section by stating and proving the rule of computations mentioned in the introduction and known from the real exponential function.

|||| Theorem 1.44 Complex Exponential Function Computation Rules

1. $e^0 = 1$
2. $e^{z_1+z_2} = e^{z_1} \cdot e^{z_2}$ for all $z_1, z_2 \in \mathbb{C}$
3. $(e^z)^n = e^{nz}$ for all $n \in \mathbb{Z}$ og $z \in \mathbb{C}$

|||| Proof

Point 1 in the theorem that $e^0 = 1$, follows from the fact that the complex exponential function is identical with the real exponential function on the real axis, cf. (1-24).

In point 2 we set $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$. From the set of polar coordinates and 1.38 we get:

$$\begin{aligned} \mathrm{e}^{z_1} \cdot \mathrm{e}^{z_2} &= (\mathrm{e}^{x_1}, y_1) \cdot (\mathrm{e}^{x_2}, y_2) = (\mathrm{e}^{x_1} \cdot e^{x_1}, y_1 + y_2) = (\mathrm{e}^{x_1+x_2}, y_1 + y_2) \\ &= \mathrm{e}^{(x_1+x_2)+\mathrm{i}(y_1+y_2)} = \mathrm{e}^{(x_1+\mathrm{i}y_1)+(x_2+\mathrm{i}y_2)} \\ &= \mathrm{e}^{z_1+z_2} \, . \end{aligned}$$

In point 3 we set $z = x + \mathrm{i}y$ and with the use of sets of polar coordinates and the repeated use of Theorem 1.38 we get:

$$\begin{aligned} (\mathrm{e}^z)^n &= ((\mathrm{e}^x)^n, n \cdot y) = (\mathrm{e}^{n \cdot x}, n \cdot y) = \mathrm{e}^{n \cdot x + \mathrm{i} \cdot n \cdot y} = \mathrm{e}^{n(x+\mathrm{i} \cdot y)} \\ &= \mathrm{e}^{n \cdot z} \, . \end{aligned}$$

By this the Theorem is proved.

■

|||| Exercise 1.45

Show that for every $z \in \mathbb{C}$ $\mathrm{e}^z \neq 0$.

1.8 The Exponential Form of Complex Numbers

Let v be an arbitrary real number. If we substitute the pure imaginary number $\mathrm{i}v$ into the complex exponential function we get from the Definition 1.41:

$$\mathrm{e}^{\mathrm{i}v} = \mathrm{e}^{0+\mathrm{i}v} = \mathrm{e}^0 \left(\cos(v) + \mathrm{i}\sin(v)\right),$$

which yields the famous *Euler's formula*.

|||| Theorem 1.46 Euler's Formula

For every $v \in \mathbb{R}$:

$$\mathrm{e}^{\mathrm{i}v} = \cos(v) + \mathrm{i}\sin(v) \, . \tag{1-29}$$

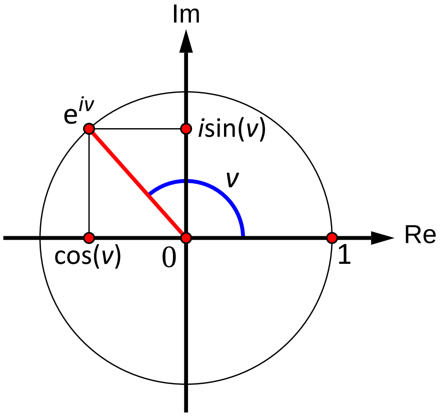


Figure 1.9: The number e^{iv} in the complex number plane

By use of the definition of the complex exponential function, see Definition 1.41, we derived Euler's formula. In return we can now use Euler's formula to write the complex exponential function in the convenient form

$$e^z = e^x\big(\cos(y) + i\sin(y)\big) = e^x e^{iy}\,. \tag{1-30}$$

The two most-used ways of writing complex numbers both in pure and applied mathematics are the *rectangular* form (as is frequently used above) and the *exponential* form. In the exponential form the polar coordinates of the number (absolute value and argument), in connection with the complex exponential function. Since the polar coordinates appear explicitly in this form, it is also called the *polar form*.

|||| Theorem 1.47 The Exponential Form of Complex Numbers

Every complex number $z \neq 0$ can be written in the form

$$z = |z|\, e^{iv}\,, \tag{1-31}$$

where v is an argument for z. This way of writing is called the *exponential form* (or the *polar form*) of the number.

Proof

Let v be an argument for the complex number $z \neq 0$, and put $r = |z|$. We show that re^{iv} has the same absolute value and argument as z, and thus the two numbers are identical:

1. $$\left|re^{iv}\right| = |r|\left|e^{iv}\right| = r\,.$$

2. Since 0 is an argument for r, and v is an argument for e^{iv}, we have that $0 + v = v$ is an argument for the product re^{iv}.

■

Method 1.48 Computations Using the Exponential Form

A decisive advantage of the exponential form of complex numbers is that one does not have to think about the rule of computations for multiplication, division and powers when the polar coordinates are used, see Theorem 1.31, Corollary 1.32 and Theorem 1.34. All computations can be carried out using the ordinary rules of computation on the exponential form of the numbers.

We now give an example of multiplication following Method 1.48; cf. Example 1.39.

Example 1.49 Multiplication in Exponential Form

Two complex numbers are given in exponential form,

$$z_1 = \frac{1}{2}\,e^{\frac{\pi}{4}i} \;\text{ and }\; z_2 = 2\,e^{\frac{3\pi}{2}i}\,.$$

The product of the numbers is found in exponential form as

$$z_1 z_2 = \left(\frac{1}{2}\,e^{\frac{\pi}{4}i}\right)\left(2\,e^{\frac{3\pi}{2}i}\right) = \left(\frac{1}{2}\cdot 2\right)e^{\frac{\pi}{4}i+\frac{3\pi}{2}i} = 1e^{i\left(\frac{\pi}{4}+\frac{3\pi}{2}\right)} = e^{\frac{7\pi}{4}\,i}\,.$$

Exercise 1.50

Show that Method 1.48 is correct.

In the following we will show how so-called *binomial equations* can be solved by the use of the exponential form. A binomial equation is an equation *with two terms* in the form

$$z^n = w\,, \tag{1-32}$$

where $w \in \mathbb{C}$ and $n \in \mathbb{N}$. Binomial equations are described in more detail in eNote 2 about polynomials.

First we show an example of the solution of a binomial equation by use of the exponential form and then we formulate the general method.

Example 1.51 Binomial Equation in Exponential Form

Find all solutions to the binomial equation

$$z^4 = -8 + 8\sqrt{3}\,\mathrm{i}\,. \tag{1-33}$$

The idea is that we write both z and the right-hand side in exponential form.

If z has the exponential form $z = s\mathrm{e}^{\mathrm{i}u}$, then the equation's left-hand side can be computed as

$$z^4 = (s\mathrm{e}^{\mathrm{i}u})^4 = s^4\,(e^{\mathrm{i}u})^4 = s^4\,e^{\mathrm{i}4u}\,. \tag{1-34}$$

The right-hand side is also written in exponential form. The absolute value r of the right-hand side is found by

$$r = |-8 + 8\sqrt{3}\,\mathrm{i}\,| = \sqrt{(-8)^2 + (8\sqrt{3})^2} = 16\,.$$

The argument v of the right-hand side satisfies

$$\cos(v) = \frac{-8}{16} = -\frac{1}{2} \quad\text{and}\quad \sin(v) = \frac{8\sqrt{3}}{16} = \frac{\sqrt{3}}{2}\,.$$

By use of the two equations the principal argument of the right-hand side can be determined to be

$$v = \arg(-8 + 8\sqrt{3}\mathrm{i}) = \frac{2\pi}{3}\,,$$

and so the exponential form of the right-hand side is

$$r\mathrm{e}^{\mathrm{i}v} = 16\mathrm{e}^{\frac{2\pi}{3}\mathrm{i}}\,. \tag{1-35}$$

We now substitute (1-34) and (1-35) into (1-33) in order to replace the right- and left-hand side with the exponential counterparts

$$s^4\, e^{\mathrm{i}4u} = 16\mathrm{e}^{\frac{2\pi}{3}\mathrm{i}}\,.$$

Since the absolute value of the left-hand side must be equal to absolute value of the right-hand side we get

$$s^4 = 16 \;\Leftrightarrow\; s = \sqrt[4]{16} = 2\,.$$

The argument of the left-hand side $4u$ and the argument of the right-hand side $\frac{2\pi}{3}$ must be equal apart from a multiple of 2π . Thus

$$4u = \frac{2\pi}{3} + p2\pi \;\Leftrightarrow\; u = \frac{\pi}{6} + p\frac{\pi}{2}\;,\; p \in \mathbb{Z}\,.$$

These infinitely many arguments correspond, as we have seen earlier, to only *four* half-lines from $(0,0)$ determined by the arguments obtained by putting $p = 0, p = 1, p = 2$ and $p = 3$. For any other value of p the corresponding half-line will be identical to one of the four mentioned above. E.g. the half-line corresponding to $p = 4$ has the argument

$$u = \frac{\pi}{6} + 4\,\frac{\pi}{2} = \frac{\pi}{6} + 2\pi\,,$$

i.e. the same half-line that corresponds to $p = 0$, since the difference in argument is a whole revolution, that is 2π .

Therefore the given equation (1-33) has exactly four solutions that lie on the four mentioned half-lines and that are separated the distance $s = 2$ from 0. Stated in exponential form:

$$z = 2\,\mathrm{e}^{\mathrm{i}(\frac{\pi}{6}+p\frac{\pi}{2})}\;,\; p = 0\,,1\,,2\,,3\,.$$

Or each recomputed to rectangular form by means of Euler's formula (1-29):

$$z_0 = \sqrt{3} + \mathrm{i}\,,\; z_1 = -1 + \mathrm{i}\sqrt{3}\,,\; z_2 = -\sqrt{3} - \mathrm{i}\,,\; z_3 = 1 - \mathrm{i}\sqrt{3}\,.$$

All solutions to a binomial equation lie on a circle with the centre at 0 and radius equal to the absolute value of the right-hand side. The connecting lines between 0 and the solutions divide the circle into equal angles. This is illustrated in Figure 1.10 which shows the solutions to the equation of the fourth degree from Example 1.51.

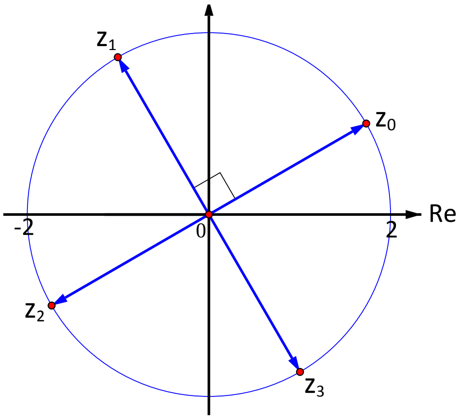


Figure 1.10: The four solutions for $z^4 = -8+8\sqrt{3}i$

The method in Example 1.51 we now generalize in the following theorem. The theorem is proved in eNote 2 about polynomials.

|||| Theorem 1.52 Binomial Equation Solved Using the Exponential Form

Given a complex number w that is different from 0 and that has the exponential form

$$w = |w|\,\mathrm{e}^{\mathrm{i}v}\,.$$

The binomial equation

$$z^n = w\ ,\ n \in \mathbb{N} \tag{1-36}$$

has n solutions that can be found with the formula

$$z = \sqrt[n]{|w|}\ \mathrm{e}^{\mathrm{i}(\frac{v}{n}+p\frac{2\pi}{n})}\ ,\ \text{ where } p = 0\,,1\,,\ldots\,,n-1\,. \tag{1-37}$$

|||| Exercise 1.53 Binomial Equation with a Negative Right-Hand Side

Let r be an arbitrary positive real number. Show by use of Theorem 1.52 that the binomial quadratic equation

$$z^2 = -r$$

has the two solutions

$$z_0 = \mathrm{i}\,\sqrt{r}\ \text{ and }\ z_1 = -\mathrm{i}\,\sqrt{r}\,.$$

1.9 Linear and Quadratic Equations

Let a and b be complex numbers with $a \neq 0$. A *complex linear equation* of the form

$$az = b$$

in analogy with the corresponding real linear equation has exactly one solution

$$z = \frac{b}{a} \,.$$

With a and b in rectangular form, the solution is easily found in rectangular form, as shown in the following example.

|||| Example 1.54 Solution of a Linear Equation

The equation

$$(1-\mathrm{i})\,z = (5+2\mathrm{i})$$

has the solution

$$z = \frac{5+2\mathrm{i}}{1-\mathrm{i}} = \frac{(5+2\mathrm{i})(1+\mathrm{i})}{(1-\mathrm{i})(1+\mathrm{i})} = \frac{3+7\mathrm{i}}{2} = \frac{3}{2} + \frac{7}{2}\,\mathrm{i}\,.$$

Also in the solution of *complex quadratic equations* we use a formula that corresponds to the well-known solution formula for real quadratic equations. This is given in the following theorem that is proved in eNote 2 about polynomials.

|||| Theorem 1.55 Solution Formula for Complex Quadratic Equations

Let a, b and c be arbitrary complex numbers with $a \neq 0$. We define the *discriminant* by $D = b^2 - 4ac$. The quadratic equation

$$az^2 + bz + c = 0 \tag{1-38}$$

has two solutions

$$z_0 = \frac{-b - w_0}{2a} \quad \text{and} \quad z_1 = \frac{-b + w_0}{2a}\,, \tag{1-39}$$

where w_0 is a solution to the binomial quadratic equation $w^2 = D$.

If in particular $D = 0$, we find $z_0 = z_1 = \dfrac{-b}{2a}$.

In this eNote we do not introduce square roots of complex numbers. Therefore the complex solution formula above differs in one detail from the ordinary real solution formula.

Concrete examples of the application of the theorem can be found in Section 30.5.2 in eNote 2 about polynomials.

1.10 Complex Functions of a Real Variable

In this section we use the theory of the so-called epsilon functions for the introduction of differentiability. The material is a bit more advanced than previously and knowledge about epsilon functions from eNote 3 (see Section 3.4) may prove advantageous. Furthermore the reader should be familiar with the rules of differentiation of ordinary real functions.

We will make a special note of functions of the type

$$f : t \mapsto \mathrm{e}^{ct}\,,\; t \in \mathbb{R}\,, \tag{1-40}$$

where c is a given complex number. This type of function has many uses in pure and applied mathematics. A main purpose of this section is to give a closer description of these. They are examples of the so-called *complex functions of a real variable*. Our investigation starts off in a wider sense with this broader class of functions. I.a. we show how concepts such as differentiability and derivatives can be introduced. Then we give a fuller treatment of functions of the type in (1-40).

|||| Definition 1.56 Complex Functions of a Real Variable

By a *complex function of a real variable* we understand a function f that for every $t \in \mathbb{R}$ attaches exactly one complex number that is denoted $f(t)$. A short way of writing a function f of this type is

$$f : \mathbb{R} \mapsto \mathbb{C} \,.$$

The notation $f : \mathbb{R} \mapsto \mathbb{C}$ tells us the function f uses a variable in the real number space, but ends up with a result in the complex number space. Consider e.g. the function $f(t) = \mathrm{e}^{\mathrm{i}t}$. At the real number $t = \frac{\pi}{4}$ we get the complex function value

$$f\left(\frac{\pi}{4}\right) = \mathrm{e}^{\frac{\pi}{4}\mathrm{i}} = \cos\left(\frac{\pi}{4}\right) + \mathrm{i}\sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\mathrm{i} \,.$$

Let us consider a function $f : \mathbb{R} \mapsto \mathbb{C}$. We introduce two real functions g and h by

$$g(t) = \mathrm{Re}\big(f(t)\big) \;\text{ and }\; h(t) = \mathrm{Im}\big(f(t)\big)$$

for all $t \in \mathbb{R}$. By this f can be stated in *rectangular form*:

$$f(t) = g(t) + \mathrm{i} \cdot h(t) \,,\; t \in \mathbb{R} \,. \tag{1-41}$$

When in the following we introduce differentiability of complex functions of one real variable, we shall need a special kind of complex function, viz. the so-called *epsilon functions*. Similar to *real* epsilon functions they are auxiliary functions, whose functional expression is of no interest. The two decisive properties for a real epsilon function $\epsilon : \mathbb{R} \mapsto \mathbb{R}$ are that it satisfies $\epsilon(0) = 0$, and that $\epsilon(t) \to 0$ when $t \to 0$. The complex epsilon function is introduced in a similar way.

|||| Definition 1.57 Epsilon Function

By a *complex epsilon function of a real variable* we understand a function $\epsilon : \mathbb{R} \mapsto \mathbb{C}$, that satisfies:

1. $\epsilon(0) = 0$, and

2. $|\epsilon(t)| \to 0$ for $t \to 0$.

Note that if ϵ is an epsilon function, then it follows directly from the Definition 1.57 that for every $t_0 \in \mathbb{R}$:

$$|\,\epsilon(t-t_0)\,| \to 0 \;\text{ for }\; t \to t_0\,.$$

In the following example a pair of complex epsilon functions of a real variable are shown.

|||| Example 1.58 Epsilon Functions

The function

$$t \mapsto \mathrm{i}\sin(t)\,,\; t \in \mathbb{R}$$

is an epsilon function. This is true because requirement 1 in definition 1.57 is fulfilled by

$$\mathrm{i}\sin(0) = \mathrm{i}\cdot 0 = 0$$

and requirement 2 by

$$|\mathrm{i}\sin(t)| = |\mathrm{i}|\,|\sin(t)| = |\sin(t)| \to 0 \;\text{ for }\; t \to 0\,.$$

Also the function

$$t \mapsto t + \mathrm{i}t\,,\; t \in \mathbb{R}$$

is an epsilon function, since

$$0 + \mathrm{i}\cdot 0 = 0$$

and

$$|t + \mathrm{i}t| = \sqrt{t^2+t^2} = \sqrt{2}\,|t| \to 0 \;\text{ for }\; t \to 0\,.$$

We are now ready to introduce the concept of *differentiability* for complex functions of a real variable.

|||| Definition 1.59 Derivative of a $\mathbb{C}$-valued Function of a Real Variable

A function $f : \mathbb{R} \mapsto \mathbb{C}$ is called *differentiable* at $t_0 \in \mathbb{R}$, if a constant $c \in \mathbb{C}$ and an epsilon function $\epsilon : \mathbb{R} \mapsto \mathbb{C}$ exist such that

$$f(t) = f(t_0) + c(t - t_0) + \epsilon(t - t_0)(t - t_0)\,,\ t \in \mathbb{R}\,. \tag{1-42}$$

If f is differentiable at t_0 then c is called the *derivative* for f at t_0 .

If f is differentiable at every t_0 in an open interval I then f is said to be *differentiable* on I .

Differentiability for a complex function of a real variable is tightly connected to the differentiability of the two real parts of the rectangular form. We now show this.

|||| Theorem 1.60

For a function $f : \mathbb{R} \mapsto \mathbb{C}$ with the rectangular form $f(t) = g(t) + ih(t)$ and a complex number c with the rectangular form $c = a + ib$:

f is differentiable at $t_0 \in \mathbb{R}$ with

$$f'(t_0) = c\,,$$

if and only if g and h are differentiable at t_0 with

$$g'(t_0) = a \quad \text{and} \quad h'(t_0) = b\,.$$

|||| Proof

First suppose that f is differentiable at t_0 and $f'(t_0) = a + ib$, where $a, b \in \mathbb{R}$. Then there exists an epsilon function ϵ such that f for every t can be written in the form

$$f(t) = f(t_0) + (a + ib)(t - t_0) + \epsilon(t - t_0)(t - t_0)\,.$$

We rewrite both the left- and the right-hand side into their rectangular form:

$$\begin{aligned}&g(t) + \mathrm{i}h(t) = \\ &g(t_0) + \mathrm{i}h(t_0) + a(t - t_0) + \mathrm{i}b(t - t_0) + \mathrm{Re}(\epsilon(t - t_0)(t - t_0)) + \mathrm{iIm}(\epsilon(t - t_0)(t - t_0)) = \\ &(g(t_0) + a(t - t_0) + \mathrm{Re}(\epsilon(t - t_0))(t - t_0)) + \mathrm{i}(h(t_0) + b(t - t_0) + \mathrm{Im}(\epsilon(t - t_0))(t - t_0))\,.\end{aligned}$$

From this we get

$$g(t) = g(t_0) + a(t - t_0) + \mathrm{Re}(\epsilon(t - t_0))(t - t_0) \text{ and } h(t) = h(t_0) + b(t - t_0) + \mathrm{Im}(\epsilon(t - t_0))(t - t_0)\,.$$

In order to conclude that g and h are differentiable at t_0 with $g'(t_0) = a$ and $h'(t_0) = b$, it only remains for us to show that $\mathrm{Re}(\epsilon)$ and $\mathrm{Im}(\epsilon)$ are real epsilon functions. This follows from

1. $\epsilon(0) = \mathrm{Re}(\epsilon(0)) + \mathrm{iIm}(\epsilon(0)) = 0$ yields $\mathrm{Re}(\epsilon(0)) = 0$ and $\mathrm{Im}(\epsilon(0)) = 0$, and
2. $|\epsilon(t)| = \sqrt{|\mathrm{Re}(\epsilon(t))|^2 + |\mathrm{Im}(\epsilon(t))|^2} \to 0$ for $t \to 0$ yields that $\mathrm{Re}(\epsilon(t)) \to 0$ for $t \to 0$ and $\mathrm{Im}(\epsilon(t)) \to 0$ for $t \to 0$.

The converse statement in the theorem is similarly proved.

■

|||| Example 1.61 Derivative of a Complex Function

By the expression

$$f(t) = t + \mathrm{i}t^2$$

a function $f : \mathbb{R} \mapsto \mathbb{C}$ is defined. Since the real part of f has the derivative 1 and the imaginary part of f the derivative $2t$ we obtain from Theorem 1.60:

$$f'(t) = 1 + \mathrm{i}2t\,,\; t \in \mathbb{R}\,.$$

|||| Example 1.62 Derivative of a Complex-valued Function

Consider the function $f : \mathbb{R} \mapsto \mathbb{C}$ given by

$$f(t) = \mathrm{e}^{\mathrm{i}t} = \cos(t) + \mathrm{i}\sin(t)\ ,\ t \in \mathbb{R}\,.$$

Since $\cos'(t) = -\sin(t)$ and $\sin'(t) = \cos(t)$, it is seen from Theorem 1.60 that

$$f'(t) = -\sin(t) + \mathrm{i}\cos(t)\,,\ t \in \mathbb{R}\,.$$

In the following theorem we consider the so-called *linear* properties of differentiation. These are well known from real functions.

|||| Theorem 1.63 Computational Rules for Derivatives

Let f_1 and f_2 be differentiable complex functions of a real variable, and let c be an arbitrary complex number. Then:

1. The function $f_1 + f_2$ is differentiable with the derivative

$$(f_1 + f_2)'(t) = {f_1}'(t) + {f_2}'(t)\,. \tag{1-43}$$

2. The function $c \cdot f_1$ is differentiable with the derivative

$$(c \cdot f_1)'(t) = c \cdot {f_1}'(t)\,. \tag{1-44}$$

|||| Proof

Let $f_1(t) = g_1(t) + \mathrm{i}\,h_1(t)$ and $f_2(t) = g_2(t) + \mathrm{i}\,h_2(t)$, where g_1, h_1, g_2 and h_2 are differentiable real functions. Furthermore let $c = a + ib$ be an arbitrary complex number in rectangular form.

First part of the theorem:

$$\begin{aligned}(f_1 + f_2)(t) &= f_1(t) + f_2(t) = g_1(t) + \mathrm{i}\,h_1(t) + g_2(t) + \mathrm{i}\,h_2(t)\\ &= (g_1(t) + g_2(t)) + \mathrm{i}\,(h_1(t) + h_2(t))\ .\end{aligned}$$

We then get from Theorem 1.60 and by the use of computational rules for derivatives for real functions:

$$\begin{aligned}(f_1+f_2)'(t) &= (g_1+g_2)'(t)+\mathrm{i}\,(h_1+h_2)'\,(t)\\ &= \left(g_1'(t)+g_2'(t)\right)+\mathrm{i}\left(h_1'(t)+h_2'(t)\right)\\ &= \left(g_1'(t)+\mathrm{i}\,h_1'(t)\right)+g_2'(t)+\mathrm{i}\,h_2'(t)\\ &= f_1'(t)+f_2'(t)\,.\end{aligned}$$

By this the first part of the theorem is proved.

Second part of the theorem:

$$\begin{aligned}c\cdot f_1(t) &= (a+\mathrm{i}b)\cdot(g_1(t)+\mathrm{i}\,h_1(t))\\ &= (a\,g_1(t)-b\,h_1(t))+\mathrm{i}\,(a\,h_1(t)+b\,g_1(t))\ .\end{aligned}$$

We get from Theorem 1.60 and by the use of computational rules for derivatives for real functions:

$$\begin{aligned}(c\cdot f_1)'(t) &= (a\,g_1-b\,h_1)'\,(t)+\mathrm{i}\,(a\,h_1+b\,g_1)'\,(t)\\ &= \left(a\,g_1'(t)-b\,h_1'(t)\right)+\mathrm{i}\left(a\,h_1'(t)+b\,g_1'(t)\right)\\ &= (a+\mathrm{i}b)\left(g_1'(t)+\mathrm{i}\,h_1'(t)\right)\\ &= c\cdot f_1'(t)\,.\end{aligned}$$

By this the second part of the theorem is proved.

■

|||| Exercise 1.64

Show that if f_1 and f_2 are differentiable complex functions of a real variable, then the function f_1-f_2 is differentiable with the derivative

$$(f_1-f_2)'(t)=f_1'(t)-f_2'(t)\,. \tag{1-45}$$

We now return to functions of the type (1-40). First we give a useful theorem about their conjugation.

|||| Theorem 1.65

For an arbitrary complex number c and every real number t:

$$\overline{\mathrm{e}^{ct}} = \mathrm{e}^{\bar{c}\,t}\,. \tag{1-46}$$

|||| Proof

Let $c = a + ib$ be the rectangular form of c. We then get by the use of Definition 1.41 and the rules of computation for conjugation in Theorem 1.23:

$$\begin{aligned}
\overline{\mathrm{e}^{ct}} &= \overline{\mathrm{e}^{at+ibt}} \\
&= \overline{e^{at}\,(\cos(bt) + \mathrm{i}\sin(bt))} \\
&= \overline{e^{at}}\;\overline{(\cos(bt) + \mathrm{i}\sin(bt))} \\
&= e^{at}\,(\cos(bt) - \mathrm{i}\sin(bt)) \\
&= e^{at}\,(\cos(-bt) + \mathrm{i}\sin(-bt)) \\
&= e^{at-\mathrm{i}bt} \\
&= e^{\bar{c}t}\,.
\end{aligned}$$

Thus the theorem is proved.

■

For ordinary real exponential functions of the type

$$f : x \mapsto \mathrm{e}^{kx}\ ,\ x \in \mathbb{R}\,,$$

where k is a real constant we have the well-known derivative

$$f'(x) = kf(x) = k\mathrm{e}^{kx}\,. \tag{1-47}$$

We end this eNote by showing that the complex exponential function of a real variable satisfies a quite similar rule of differentiation.

|||| Theorem 1.66 Differentiation of e^{ct}

Consider an arbitrary number $c \in \mathbb{C}$. The function $f : \mathbb{R} \mapsto \mathbb{C}$ given by

$$f(t) = \mathrm{e}^{ct}\,,\, t \in \mathbb{R} \tag{1-48}$$

is differentiable and its derivative is determined by

$$f'(t) = cf(t) = c\mathrm{e}^{ct}\,. \tag{1-49}$$

|||| Proof

Let the rectangular form of c be $c = a + ib$. We then get

$$\begin{aligned} e^{ct} &= \mathrm{e}^{at+ibt} \\ &= e^{at}\left(\cos(bt) + \mathrm{i}\sin(bt)\right) \\ &= e^{at}\cos(bt) + \mathrm{i}\left(e^{at}\sin(bt)\right)\,. \end{aligned}$$

Thus we have

$$f(t) = g(t) + \mathrm{i}h(t)\,,\ \text{where}\ \ g(t) = e^{at}\cos(bt)\ \text{ and }\ h(t) = e^{at}\sin(bt)\,.$$

Since g and h are differentiable, f is also differentiable . Furthermore since

$$g'(t) = a\mathrm{e}^{at}\cos(bt) - \mathrm{e}^{at}b\sin(bt)\ \text{ and }\ h'(t) = a\mathrm{e}^{at}\sin(bt) + \mathrm{e}^{at}b\cos(bt)\,,$$

we now get

$$\begin{aligned} f'(t) &= a\mathrm{e}^{at}\cos(bt) - \mathrm{e}^{at}b\sin(bt) + \mathrm{i}\left(a\mathrm{e}^{at}\sin(bt) + \mathrm{e}^{at}b\cos(bt)\right) \\ &= (a + \mathrm{i}b)e^{at}\left(\cos(bt) + \mathrm{i}\sin(bt)\right) \\ &= (a + \mathrm{i}b)e^{at+\mathrm{i}bt} \\ &= c\,\mathrm{e}^{ct}\,. \end{aligned}$$

Thus the theorem is proved.

■

If c in Theorem 1.66 is real, (1-49) naturally only expresses the ordinary differentiation of the real exponential function as in (1-47), as expected.

eNote 2

Polynomials of One Variable

In this eNote complex polynomials of one variable are introduced. An elementary knowledge of complex numbers is a prerequisite, and knowlege of real polynomials of one real variable is recommended.

Updated: 29.05.16. *Karsten Schmidt.* 11.9.2021. *David Brander.*

2.1 Introduction

Polynomials are omnipresent in the technical literature about mathematical models of physical problems. A great advantage of polynomials is the simplicity of computation since only addition, multiplication and powers are needed. Because of this polynomials are especially applicable as approximations to more complicated types of functions.

Knowledge about the *roots* of polynomials is the main road to understanding their properties and efficient usage, and is therefore a major subject in the following. But first we introduce some general properties.

|||| Definition 2.1

By a *polynomial* of degree n we understand a function that can be written in the form

$$P(z) = a_n\, z^n + a_{n-1}\, z^{n-1} + \cdots + a_1\, z + a_0 \tag{2-1}$$

where $a_0,\, a_1,\, \ldots,\, a_n$ are complex constants with $a_n \neq 0$, and z is a complex variable.

a_k is called the *coefficient* of z^k, $k = 0, 1, \ldots, n$, and a_n is the *leading coefficient*.
A *real polynomial* is a polynomial in which all the coefficients are real.
A *real polynomial of a real variable* is a real polynomial in which we assume $z \in \mathbb{R}$.

Polynomials are often denoted by a capital P or similar letter $Q, R, S \ldots$ If the situation requires you to include the variable name, the polynomial is written as $P(z)$ where it is understood that z is an independent complex variable.

|||| Example 2.2 Examples of Polynomials

$P(z) = 2\,z^3 + (1+\mathrm{i})\,z + 5$ is a polynomial of the third degree.
$Q(z) = z^2 + 1$ is a real quadratic polynomial.
$R(z) = 17$ is a polynomial of the $0'$th degree.
$S(z) = 0$ is called the 0-polynomial and is not assigned any degree.
$T(z) = 2\,z^3 + 5\sqrt{z} - 4$ is not a polynomial.

If you multiply a polynomial by a constant, or when you add, subtract, multiply and compose polynomials with each other, you get a new polynomial. This polynomial can be simplified by gathering terms of the same degree and written in the form (2.1).

|||| Example 2.3 Addition and Multiplication of Polynomials

Two polynomials P and Q are given by $P(z) = z^2 - 1$ and $Q(z) = 2\,z^2 - z + 2$. The polynomials $R = P + Q$ and $S = P \cdot Q$ are determined like this:

$$\begin{aligned}
R(z) &= (z^2 - 1) + (2\,z^2 - z + 2) = (z^2 + 2\,z^2) + (-z) + (-1+2) = 3\,z^2 - z + 1\,.\\
S(z) &= (z^2 - 1) \cdot (2\,z^2 - z + 2) = (2\,z^4 - z^3 + 2\,z^2) + (-2\,z^2 + z - 2)\\
&= 2\,z^4 - z^3 + (2\,z^2 - 2\,z^2) + z - 2 = 2\,z^4 - z^3 + z - 2\,.
\end{aligned}$$

2.2 The Roots of Polynomials

Definition 2.4 Root of a Polynomial

By a *root* of a polynomial $P(z)$ we understand a number z_0 such that $P(z_0) = 0$.

Example 2.5 Whether a Given Number Is a Root of a Polynomial

Show that 3 is a root of $P(z) = z^3 - 5z - 12$, and that 1 is not a root.

Since $P(3) = 3^3 - 5 \cdot 3 - 12 = 0$, 3 is a root of P.
Since $P(1) = 1^3 - 5 \cdot 1 - 12 = -16 \neq 0$, 1 is not a root of P.

To develop the theory we shall need the following Lemma.

|||| Lemma 2.6 The Theorem of Descent

A polynomial P of the degree n is given by

$$P(z) = a_n\, z^n + a_{n-1}\, z^{n-1} + \cdots + a_1\, z + a_0\,. \tag{2-2}$$

If z_0 is an arbitrary number, and Q is the polynomial of the n-1$'$ degree given by the coefficients

$$b_{n-1} = a_n \tag{2-3}$$

$$b_k = a_{k+1} + z_0 \cdot b_{k+1} \quad \text{for} \quad k = n{-}2, \ldots, 0 \quad , \tag{2-4}$$

then P can be written in the factorized form

$$P(z) = (z - z_0)Q(z) \tag{2-5}$$

if and only if z_0 is a root of P.

|||| Proof

Let the polynomial P be given as in the theorem, and let α be an arbitrary number. Consider an arbitrary $(n-1)$-degree polynomial

$$Q(z) = b_{n-1}\, z^{n-1} + b_{n-2}\, z^{n-2} + \cdots + b_1\, z + b_0\,.$$

By simple calculation we get

$$(z-\alpha)Q(z) = b_{n-1}\, z^n + (b_{n-2} - \alpha b_{n-1})\, z^{n-1} + \cdots + (b_0 - \alpha b_1)z - \alpha b_0\,.$$

It is seen that the polynomials $(z-\alpha)Q(z)$ and $P(z)$ have the same representation if we in succession write the b_k-coefficients for Q as given in (2-3) and (2-4), and if at the same time the following is valid:

$$-\alpha b_0 = a_0 \Leftrightarrow b_0\, \alpha = -a_0\,.$$

We investigate whether this condition is satisfied by using (2-3) and (2-4) in the opposite order:

$$\begin{aligned} b_0\,\alpha &= (a_1+\alpha b_1)\alpha = b_1\alpha^2 + a_1\alpha \\ &= (a_2+\alpha b_2)\alpha^2 + a_1\alpha = b_2\alpha^3 + a_2\alpha^2 + a_1\alpha \\ &\;\vdots \\ &= b_{n-1}\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_2\alpha^2 + a_1\alpha \\ &= a_n\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_2\alpha^2 + a_1\alpha = -a_0 \\ \Leftrightarrow\; & P(\alpha) = a_n\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_2\alpha^2 + a_1\alpha + a_0 = 0\,. \end{aligned}$$

It is seen that the condition is only satisfied if and only if α is a root of P. By this the proof is complete.

■

|||| Example 2.7 Descent

Given the polynomial $P(z) = 2z^4 - 12z^3 + 19z^2 - 6z + 9$. It is seen that 3 is a root since $P(3) = 0$. Determine a third-degree polynomial Q such that

$$P(z) = (z-3)Q(z)\,.$$

We set $a_4 = 2,\ a_3 = -12,\ a_2 = 19,\ a_1 = -6$ og $a_0 = 9$ and find the coefficients for Q by the use of (2-3) and (2-4):

$$\begin{aligned} b_3 =& a_4 = 2 \\ b_2 =& a_3 + 3b_3 = -12 + 3\cdot 2 = -6 \\ b_1 =& a_2 + 3b_2 = 19 + 3\cdot(-6) = 1 \\ b_0 =& a_1 + 3b_1 = -6 + 3\cdot 1 = -3\,. \end{aligned}$$

We conclude that

$$Q(z) = 2z^3 - 6z^2 + z - 3$$

so

$$P(z) = (z-3)\,(2z^3 - 6z^2 + z - 3)\,.$$

When a polynomial P with the root z_0 is written in the form $P(z) = (z-z_0)Q_1(z)$, where Q_1 is a polynomial, it is possible that z_0 is also a root of Q_1. Then Q_1 can similarly be written as $Q_1(z) = (z - z_0)Q_2(z)$ where Q_2 is a polynomial. And in this way the descent can successively be carried out as $P(z) = (z - z_0)^m R(z)$ where R is a polynomial in which z_0 is not a root. We will now show that this factorization is unique.

|||| Theorem 2.8 The Multiplicity of a Root

If z_0 is a root of the polynomial P, it can in exactly one way be written in factorised form as:

$$P(z) = (z - z_0)^m \, R(z) \tag{2-6}$$

where $R(z)$ is a polynomial for which z_0 is not a root.

The exponent m is called the *algebraic multiplicity* of the root z_0.

|||| Proof

Assume that α is a root of P, and that (contrary to the statement in the theorem) there exist two different factorisations

$$P(z) = (z - \alpha)^r \, R(z) = (z - \alpha)^s \, S(z)$$

where $r > s$, and $R(z)$ and $S(z)$ are polynomials of which α is not a root. We then get

$$(z - \alpha)^r \, R(z) - (z - \alpha)^s \, S(z) = (z - \alpha)^s \big((z - \alpha)^k R(z) - S(z) \big) = 0 \,, \text{ for all } z \in \mathbb{C}$$

where $k = r - s$. This equation is only satisfied if

$$(z - \alpha)^k \, R(z) = S(z) \quad \text{for} \quad \text{all} \quad z \neq \alpha \,.$$

Since both the left-hand and the right-hand sides are continuous functions, they must have the same value at $z = \alpha$. From this we get that

$$S(\alpha) = (z - \alpha)^k R(\alpha) = 0$$

which is contradictory to the assumption that α is not a root of S.

■

|||| Example 2.9

In Example 2.7 we found that

$$P(z) = (z-3)\,(2z^3 - 6z^2 + z - 3)$$

where 3 is a root. But 3 is also a root of the factor $2z^3 - 6z^2 + z - 3$. By using the theorem of descent, Theorem 2.6, on this polynomial we get

$$P(z) = (z-3)\ \ (z-3)(2z^2+1)\ = (z-3)^2\,(2z^2+1)\,.$$

Since 3 is not a root of $2z^2+1$, the root 3 in P has the multiplicity 2.

Now we have started a process of descent! How far can we get along this way? To continue this investigation we will need a fundamental result, viz the *Fundamental Theorem*.

2.2.1 The Fundamental Theorem of Algebra

A decisive reason for the introduction of complex numbers is that every (complex) polynomial has a root in the set of complex numbers. This result was proven by the mathematician Gauss in his ph.d.-dissertation from 1799. The proof of the theorem is demanding, and Gauss strove all his life to refine his proof more. Four versions of the proof by Gauss exist, so there is no doubt that he put a lot of emphasis on this theorem. Here we take the liberty to state Gauss' result without proof:

|||| Theorem 2.10 The Fundamental Theorem of Algebra

Every polynomial of degree $n \geq 1$ has at least one root within the set of complex numbers.

The polynomial $P(z) = z^2 + 1$ has no roots within the set of real numbers. But within the set of complex numbers it has two roots i and $-\mathrm{i}$ because

$$P(\mathrm{i}) = \mathrm{i}^2 + 1 = -1 + 1 = 0 \ \text{ and }\ P(-\mathrm{i}) = (-\mathrm{i})^2 + 1 = \mathrm{i}^2 + 1 = 0\,.$$

The road from the fundamental theorem of algebra until full knowledge of the number of roots is not long. We only have to develop the ideas put forward in the theorem of descent futher.

We consider a polynomial P of degree n with leading coefficient a_n. If $n \geq 1$, P has according to the fundamental theorem of algebra a root α_1 and therefore by the use of method of coefficients, cf. Theorem 2.6 it can be written as

$$P(z) = (z - \alpha_1)Q_1(z) \tag{2-7}$$

where Q_1 is a polynomial of degree n-1 with leading coefficient a_n. If $n \geq 2$, then Q_1 has a root α_2 and can be written as

$$Q_1(z) = (z - \alpha_2)Q_2(z)$$

where Q_2 is a polynomial of degree n-2 also with a leading coefficient a_n. By substitution we now get

$$P(z) = (z - \alpha_1)(z - \alpha_2)Q_2(z)\,.$$

In this way the construction of polynomials of descent Q_k of degree $n - k$ for $k = n - 1, \ldots, 0$ continues until we reach the polynomial Q_n of degree n-$n = 0$, which in accordance with Example 2.2, is equal to its leading coefficient a_n. Hereafter P can be written in its *completely factorized form*:

$$P(z) = a_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n)\,. \tag{2-8}$$

In this expression we should note three things:

- First all the n numbers $\alpha_1, \ldots, \alpha_n$ that are listed in (2-8), are roots of P since substitution into the formula gives the value 0.
- The second thing we notice is that P cannot have other roots than the n given ones. That there cannot be more roots is easily seen as follows: If an arbitrary number $\alpha \neq \alpha_k$, $k = 1, \ldots, n$, is inserted in place of z in (2-8), all factors on the right-hand side of (2-8) will be different from zero. Hence their product will also be different form zero. Therefore $P(\alpha) \neq 0$, and α is not a root of P.
- The last thing we notice in (2-8), is that the roots are not necessarily different. If $z_1, z_2, \ldots, z_p$ are the p *different* roots of P, and m_k is the multiplicity of z_k, $k = 1, \ldots, p$, then the completely factorized form (2-8) can be simplified as follows

$$P(z) = a_n(z - z_1)^{m_1}(z - z_2)^{m_2} \cdots (z - z_p)^{m_p} \tag{2-9}$$

where the following applies:

$$m_1 + m_2 \cdots + m_p = n\,.$$

According to the preceding arguments we can now present the fundamental theorem of algebra in the extended form.

Theorem 2.11 The Fundamental Theorem of Algebra — Version 2

Every polynomial of degree $n \geq 1$ has within the set of complex numbers exactly n roots, when the roots are counted with multiplicity.

Example 2.12 Quadratic Polynomial in Completely Factorized Form

An arbitrary quadratic polynomial $P(z) = az^2 + bz + c$ can be written in the form

$$P(z) = a(z - \alpha)(z - \beta)$$

where α and β are roots of P. If $\alpha \neq \beta$, P has two different roots, both with algebraic multiplicity 1. If $\alpha = \beta$, P has only one root with the algebraic multiplicity 2. The roots are then denoted a *double root*.

Example 2.13 Algebraic Multiplicity

A polynomial P is given in complete factorized form as:

$$P(z) = 7(z-1)^2(z+4)^3(z-5)\,.$$

We see that P has three different roots: $1, -4$ and 5 with the algebraic multiplicities 2, 3 and 1, respectively.

We notice that the sum of the algebraic multiplicities is 6 which equals the degree of P in concordance with the Fundamental Theorem of Algebra — Version 2.

Example 2.14 Algebraic Multiplicity

State the number of roots of $P(z) = z^3$.

P has only one root $z = 0$. The algebraic multiplicity of this root is 3. One says that 0 is a *triple root* in the polynomial.

2.3 Identical Polynomials

Two polynomials P and Q are equal (as functions of z) if $P(z) = Q(z)$ for all z. But what does it *take* for two polynomials to be equal? Is it possible that a fourth-degree and a fifth-degree polynomial take on the same value for all variables as long as you choose the right coefficients? This is not the case as is seen from the following theorem.

|||| Theorem 2.15 The Identity Theorem for Polynomials

Two polynomials are identical if and only if they are of the same degree, and all coefficients for corresponding terms of the same degree from the two polynomials are equal.

|||| Proof

We consider two arbitrary polynomials P og Q. If they are of the same degree, and all the coefficients for terms of the same degree are equal, they must have the same value for all variables and hence they are identical. This proves the first direction of the theorem of identity.

Assume hereafter that P og Q are identical as functions of z, but that not all coefficients for terms of the same degree from the two polynomials are equal. We assume further that P has the degree n and Q the degree m where $n \geq m$. Let a_k be the coefficients for P and let b_k be the coefficients for Q, and consider the difference polynomial

$$\begin{aligned} R(z) =& P(z) - Q(z) \\ =& (a_n - b_n)z^n + (a_{n-1} - b_{n-1})z^{n-1} + \cdots + (a_1 - b_1)z + (a_0 - b_0) \end{aligned} \tag{2-10}$$

where we for the case $n > m$ put $b_k = 0$ for $m < k \leq n$. We note that the 0-degree coefficient $(a_0 - b_0)$ cannot be the only coefficient of $R(z)$ that is different from 0, since this would make $P(0) - Q(0) = (a_0 - b_0) \neq 0$ which contradicts that P and Q are identical as functions. Therefore the degree of R is greater than or equal to 1. On the other hand (2-10) shows that the degree of R at the most is n. Now let z_k, $k = 1, \ldots, n+1$, be $n+1$ different numbers. They are all roots of R since

$$R(z_k) = P(z_k) - Q(z_k) = 0\,,\ k = 1\ldots n+1\,.$$

This contradicts the fundamental theorem of algebra – version 2, Theorem 2.11: R cannot have a number of roots that is higher than its degree. The assumption, that not all coefficients of terms of the same degree from P and Q are equal, must therefore be wrong. From this it also follows that P and Q have the same degree. By this the second part of the identity theorem is proven.

■

|||| Example 2.16 Two Identical Polynomials

The equation

$$3\,z^2 - z + 4 = a\,z^2 + b\,z + c$$

is satisfied for all z exactly when $a = 3$, $b = -1$ og $c = 4$.

|||| Exercise 2.17 To Identical Polynomials

Determine the numbers a, b and c such that

$$(z-2)(a\,z^2 + b\,z + c) = z^3 - 5\,z + 2 \quad \text{for} \quad \text{all} \quad z\,.$$

In the following section we treat methods of finding roots of certain types of polynomials.

2.4 Polynomial Equations

From the fundamental theorem of algebra, Theorem 2.10, we know that every polynomial of degree greater than or equal to 1 has roots. Moreover, in the extended version, Theorem 2.11, it is maintained that for every polynomial the degree is equal to the number of roots if the roots are counted with multiplicity. But the theorem is a theoretical theorem of existence that does not help in *finding* the roots.

In the following methods for finding the roots of simple polynomials are introduced.

But let us keep the level of ambition (safely) low, because in the beginning of the $17'$th century the Norwegian algebraicist Abel showed that one *cannot* establish general methods for finding the roots of arbitrary polynomials of degree larger than four!

For polynomials of higher degree than four a number of smart tricks exist by which one can successfully find a single root. Hereafter one descends to a polynomial of lower degree — and successively decends to a polynomial of fourth degree or lower for which one can find the remaining roots.

Let us at the outset maintain that when you want to find the roots of a polynomial $P(z)$, you should solve the corresponding *polynomial equation* $P(z) = 0$. As a simple illustration we can look at the root of an arbitrary first-degree polynomial:

$$P(z) = az + b\,.$$

To find this we shall solve the equation

$$az + b = 0\,.$$

this is not difficult. It has the solution $z_0 = -\dfrac{b}{a}$ which therefore is a root of $P(z)$.

Finding the *roots* of a polynomial P, is tantamount to finding the *solutions* to the polynomial equation $P(z) = 0$.

|||| Example 2.18 The Root of a Linear Polynomial

Find the root of a linear polynomial P given by

$$P(z) = (1 - \mathrm{i})\,z - (5 + 2\mathrm{i})\,.$$

We shall solve the following equation

$$(1 - \mathrm{i})\,z - (5 + 2\mathrm{i}) = 0 \Leftrightarrow (1 - \mathrm{i})\,z = (5 + 2\mathrm{i})\,.$$

We isolate z on the left-hand side:

$$z = \frac{5 + 2\mathrm{i}}{1 - \mathrm{i}} = \frac{(5 + 2\mathrm{i})(1 + \mathrm{i})}{(1 - \mathrm{i})(1 + \mathrm{i})} = \frac{3 + 7\mathrm{i}}{2} = \frac{3}{2} + \frac{7}{2}\,\mathrm{i}\,.$$

Hence the equation has the solution $z_0 = \dfrac{3}{2} + \dfrac{7}{2}\,\mathrm{i}$ that also is the root of P.

2.4.1 Binomial Equations

A binomial equation is an equation of the degree n in which only the coefficients a_n (the term of highest degree) and a_0 (the constant term) are different from 0. A given binomial equation can only be simplified to the following form:

|||| Definition 2.19 Binomial Equation

A binomial equation has the form $z^n = w$ where $w \in \mathbb{C}$ and $n \in \mathbb{N}$.

For binomial equations an explicit solution formula exists, which we present in the following theorem.

|||| Theorem 2.20 Binomial Equations Solved by Use of the Exponential Form

Let $w \neq 0$ be a complex number with the exponential form

$$w = |w|\,\mathrm{e}^{\mathrm{i}v}\,.$$

The binomial equation

$$z^n = w \tag{2-11}$$

has n different solutions given by the formula

$$z_p = \sqrt[n]{|w|}\ \mathrm{e}^{\mathrm{i}(\frac{v}{n}+p\frac{2\pi}{n})}\ \text{ where } p = 0\,,1\,,\ldots\,,n-1\,. \tag{2-12}$$

|||| Proof

For every $p \in \{0,1,\ldots,n-1\}$ $z_p = \sqrt[n]{|w|}\ \mathrm{e}^{\mathrm{i}(\frac{v}{n}+p\frac{2\pi}{n})}$ is a solution to (2-11), since

$$(z_p)^n = \left(\sqrt[n]{|w|}\mathrm{e}^{\mathrm{i}(\frac{v}{n}+p\frac{2\pi}{n})}\right)^n = |w|\ \mathrm{e}^{\mathrm{i}(v+p\,2\pi)} = |w|\ \mathrm{e}^{\mathrm{i}v}\ = w.$$

It is seen that the n solutions viewed as points in the complex plane all lie on a circle with centre at $z=0$, radius $\sqrt[n]{|w|}$ and a consecutive angular distance of $\frac{2\pi}{n}$. In other words the connecting lines between $z = 0$ and the solutions divide the circle in n angles of the same size.

From this it follows that all n solutions are mutually different. That there are no more solutions is a consequence of the fundamental theorem of algebra – version 2, Theorem 2.11. By this the theorem is proven.

■

In the next examples we will consider some important special cases of binomial equations.

|||| Example 2.21 Binomial Equation of the Second Degree

We consider a complex number in the exponential form $w = |w|\,\mathrm{e}^{\mathrm{i}v}$. It follows from (2-12) that the quadratic equation

$$z^2 = w$$

has two solutions

$$z_0 = \sqrt{|w|}\;\mathrm{e}^{\mathrm{i}\,\frac{v}{2}} \quad\text{and}\quad z_1 = -\sqrt{|w|}\;\mathrm{e}^{\mathrm{i}\,\frac{v}{2}}\,.$$

|||| Example 2.22 Binomial Equation of the Second Degree with a Negative Right-Hand Side

Let r be an arbitrary positive real number. By putting $v = \mathrm{Arg}(-r) = \pi$ in Example 2.21 it is seen that the binomial of the second degree

$$z^2 = -r$$

has two solutions

$$z_0 = \mathrm{i}\,\sqrt{r} \;\text{ og }\; z_1 = -\mathrm{i}\,\sqrt{r}\,.$$

As a concrete example the equation $z^2 = -16$ has the solutions $z = \pm\mathrm{i}\,4$.

Sometimes the method used in Example 2.21 can be hard to carry out. In the following example we show an alternative method.

|||| Example 2.23 Binomial Equation of the Second Degree, Method 2

Solve the equation

$$z^2 = 8 - 6i\,. \tag{2-13}$$

Since we expect the solution to be complex we put $z = x + \mathrm{i}y$ where x and y are real numbers. If we can find x and y, then we have found the solutions for z. Therefore we have $z^2 = (x + \mathrm{i}y)^2 = x^2 - y^2 + 2xy\mathrm{i}$ and we see that (2-13) is equivalent to

$$x^2 - y^2 + 2xy\mathrm{i} = 8 - 6\mathrm{i}\,.$$

Since a complex equation is true exactly when both the real parts and the imaginary parts of the right-hand and the left-hand sides of the equation are identical, (2-13) is equivalent to

$$x^2 - y^2 = 8 \;\text{ and }\; 2xy = -6\,. \tag{2-14}$$

If we put $y = \dfrac{-6}{2x} = -\dfrac{3}{x}$ in $x^2 - y^2 = 8$, and put $x^2 = u$, we get a quadratic equation that can be solved:

$$\begin{gathered}
x^2 - \left(-\frac{3}{x}\right)^2 = 8 \Leftrightarrow x^2 - \frac{9}{x^2} = 8 \Leftrightarrow \\
\left(x^2 - \frac{9}{x^2}\right)x^2 = 8x^2 \Leftrightarrow x^4 - 9 = 8x^2 \Leftrightarrow x^4 - 8x^2 - 9 = 0 \Leftrightarrow \\
u^2 - 8u - 9 = 0 \Leftrightarrow u = 9 \;\text{ or }\; u = -1\,.
\end{gathered}$$

The equation $x^2 = u = 9$ has the solutions $x_1 = 3$ and $x_2 = -3$, while the equation $x^2 = u = -1$ has no solution, since x and y are real numbers. If we put $x_1 = 3$ respective $x_2 = -3$ in (2-14), we get the corresponding y-values $y_1 = -1$ and $y_2 = 1$.

From this we conclude that the given equation (2-13) has the roots

$$z_1 = x_1 + \mathrm{i}y_1 = 3 - \mathrm{i} \;\text{ and }\; z_2 = x_2 + \mathrm{i}y_2 = -3 + \mathrm{i}\,.$$

2.4.2 Quadratic Equations

For the solution of quadratic equations we state below the formula that corresponds to the well-known solution formula for real quadratic equations. There is a single deviation, viz. we do not compute the square-root of the discriminant since we in this theorem do not presuppose knowledge of square-roots of complex numbers.

|||| Theorem 2.24 Solution Formula for Quadratic Equation

For the quadratic equation

$$az^2 + bz + c = 0\ ,\ \ a \neq 0 \tag{2-15}$$

we introduce the *discriminant* D by $D = b^2 - 4ac$. The equations has two solutions

$$z_1 = \frac{-b - w_0}{2a} \ \text{og}\ \ z_2 = \frac{-b + w_0}{2a} \tag{2-16}$$

where w_0 is a solution to the binomial equation of the second degree $w^2 = D$.

If in particular $D = 0$, we have that $z_1 = z_2 = \dfrac{-b}{2a}$.

Proof

Let w_0 be an arbitrary solution to the binomial equation $w^2 = D$. We then have:

$$
\begin{aligned}
az^2 + bz + c &= a\left(z^2 + \frac{b}{a}z + \frac{c}{a}\right) \\
&= a\left(\left(z + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a}\right) \\
&= a\left(\left(z + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a^2}\right) \\
&= a\left(\left(z + \frac{b}{2a}\right)^2 - \frac{D}{4a^2}\right) \\
&= a\left(\left(z + \frac{b}{2a}\right)^2 - \frac{w_0^2}{4a^2}\right) \\
&= a\left(\left(z + \frac{b}{2a}\right) + \frac{w_0}{2a}\right)\left(\left(z + \frac{b}{2a}\right) - \frac{w_0}{2a}\right) \\
&= a\left(z + \frac{b + w_0}{2a}\right)\left(z + \frac{b - w_0}{2a}\right) = 0 \\
\Leftrightarrow z &= \frac{-b - w_0}{2a} \quad \text{or} \quad z = \frac{-b + w_0}{2a} \quad .
\end{aligned}
$$

By this the solution formula (2-16) is derived.

■

Example 2.25 Real Quadratic Equation with a Positive Value of the Discriminant

Solve the following quadratic equation with real coefficients:

$$2z^2 + 5z - 3 = 0\,.$$

We identify the coefficients: $a = 2, b = 5, c = -3$, and find the discriminant as:

$$D = 5^2 - 4 \cdot 2 \cdot (-3) = 49\,.$$

It is seen that $w_0 = 7$ is a solution to the binomial equation of the second degree $w^2 = D =$

49 . Now the solutions can be computed as:

$$z_1 = \frac{-5+7}{2\cdot 2} = \frac{1}{2} \text{ and } z_2 = \frac{-5-7}{2\cdot 2} = -3\,. \tag{2-17}$$

|||| Example 2.26 Real Quadratic Equation with a Negative Discriminant

Solve the following quadratic equation with real coefficients:

$$z^2 - 2z + 5 = 0\,.$$

We identify the coefficients: $a = 1, b = -2, c = 5$, and find the discriminant as:

$$D = (-2)^2 - 4\cdot 1\cdot 5 = -16\,.$$

According to Example 2.22 the solution to the binomial equation of the second degree $w^2 = D = -16$ is given by $w_0 = 4\mathrm{i}$. Now the solutions can be computed as:

$$z_1 = \frac{-(-2)+4\mathrm{i}}{2\cdot 1} = 1+2\mathrm{i} \text{ and } z_2 = \frac{-(-2)-4\mathrm{i}}{2\cdot 1} = 1-2\mathrm{i}\,. \tag{2-18}$$

|||| Example 2.27 A Quadratic Equation with Complex Coefficients

Solve the quadratic equation

$$z^2 - (1+\mathrm{i})z - 2 + 2\mathrm{i} = 0\,. \tag{2-19}$$

First we identify the coefficients: $a = 1, b = -(1+\mathrm{i}), c = -2+2\mathrm{i}$, and we find the discriminant:

$$D = (-(1+\mathrm{i}))^2 - 4\cdot 1\cdot(-2+2\mathrm{i}) = 8-6\mathrm{i}\,.$$

From Example 2.23 we know that the solution to the binomial equation $w^2 = D = 8-6\mathrm{i}$ is $w_0 = 3-\mathrm{i}$. From this we find the solution to (2-19) as

$$z_1 = \frac{-(-(1+\mathrm{i}))+(3-\mathrm{i})}{2\cdot 1} = 2 \text{ and } z_2 = \frac{-(-(1+\mathrm{i}))-(3-\mathrm{i})}{2\cdot 1} = -1+\mathrm{i}\,. \tag{2-20}$$

2.4.3 Equations of the Third and Fourth Degree

From antiquity geometrical methods for the solution of (real) quadratic equations are known. But not until A.D. 800 did algebraic solution formulae became known, through the work (in Arabic) of the Persian mathematician Muhammad ibn Musa al-Khwarismes famous book al-Jabr. In the West the name al-Khwarisme became the well-known word *algorithm*, while the book title became *algebra*.

Three centuries later history repeated itself. Around A.D. 1100 another Persian mathematician (and poet) Omar Khayyám gave exact methods on how to find solutions to real equations of the third and fourth degree by use of advanced geometrical methods. As an example he solved the equation $x^3 + 200x = 20x^2 + 2000$ by intersecting a circle with a hyperbola the equations of which he could derive from the equation of third degree.

Omar Khayyám did not think it possible to draw up algebraic formulae for solutions to equations of degree greater than two. He was proven wrong by the Italian Gerolamo Cardano who in the 16th century published formulae for the solution of Equations of the third and fourth degree.

Khayyáms methods and Cardanos formulae are beyond the scope of this eNote. Here we only give — see the previous Example 2.9 and the following Example 2.28 — a few examples by use of the "method of descent", Theorem 2.6, on how one can find all solutions to equations of degree greater that two if one in advance knows or can guess a sufficient number of the solutions.

|||| Example 2.28 An Equation of the Third Degree with an Initial Guess

Solve the equation of third degree

$$z^3 - 3z^2 + 7z - 5 = 0\,.$$

It is easily guessed that 1 is a solution. By use of the algorithm of descent one easily gets the factorization:

$$z^3 - 3z^2 + 7z - 5 = (z-1)(z^2 - 2z + 5) = 0\,.$$

We know that 1 is a solution, the remaining solutions are found by solving the quadratic equation

$$z^2 - 2z + 5 = 0\,,$$

which, according to Example 2.26, has the solutions $1 + 2\mathrm{i}$ and $1 - 2\mathrm{i}$.

Collectively the equation of the third degree has the solutions $1, 1+2\mathrm{i}$ og $1-2\mathrm{i}$.

2.5 Real Polynomials

The theory that has been unfolded in the previous section applies to all polynomials with complex coefficients. In this section we present two theorems that *only* apply to polynomials with real coefficients — that is the subset called *the real polynomials*. The first theorem shows that non-real roots always appear in pairs.

|||| Theorem 2.29 Roots in Real Polynomials

If the number $a+ib$ is a root of the polynomial that only has real coefficients, then also the conjugate number $a-ib$ is a root of the polynomial.

|||| Proof

Let

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

be a real polynomial. By use of the arithmetic rules for conjugation of the sum and product of complex numbers (see eNote 1 about complex numbers) with the condition that all coefficients are real, we get

$$\begin{aligned} \overline{P(z)} &= \overline{a_n\, z^n + a_{n-1}\, z^{n-1} + \cdots + a_1\, z + a_0} \\ &= a_n\, \overline{z}^n + a_{n-1}\, \overline{z}^{n-1} + \cdots + a_1\, \overline{z} + a_0 \\ &= P(\overline{z}) \ . \end{aligned}$$

If z_0 is a root of P, we get

$$\overline{P(z_0)} = \overline{0} = 0 = P(\overline{z_0})$$

from which it is seen that $\overline{z_0}$ is also a root. Thus the theorem is proven.

■

|||| Example 2.30 Conjugated Roots

Given that the polynomial

$$P(z) = 3z^2 - 12z + 39 \tag{2-21}$$

has the root $2 - 3\mathrm{i}$. Determine all roots of P, and write P in a complete factorized form.

We see that all the three coefficients in P are real. Therefore the conjugate of the given root $2 + 3\mathrm{i}$ is also a root of P. Since P is a quadratic polynomial, there are no more roots.

According to Example 2.12 the complete factorized form for P : is

$$P(z) = 3\,(z - (2 - 3\mathrm{i}))(z - (2 + 3\mathrm{i}))\,.$$

In the complete factorized form of a polynomial it is always possible to multiply the two factors that correspond to a pair of conjugated roots such that the product forms a *real quadratic polynomial* in this way:

$$\begin{aligned}(z - (a + \mathrm{i}b))(z - (a - \mathrm{i}b)) &= ((z - a) + \mathrm{i}b))((z - a) - \mathrm{i}b) \\ &= (z - a)^2 - (\mathrm{i}b)^2 \\ &= z^2 - 2az + (a^2 + b^2)\,.\end{aligned}$$

From Theorem 2.29 we know that complex roots always are present in conjugated pairs. This leads to the following theorem:

|||| Theorem 2.31 Real Factorization

A real polynomial can be written as a product of real polynomials of the first degree and real quadratic polynomials without any real roots.

|||| Example 2.32 Real Factorization

Given that a real polynomial of seventh degree P has the roots $1, \mathrm{i}, 1+2\mathrm{i}$ as well as the double root -2, and that the coefficient to its term of the highest degree is $a_7 = 5$. Write P as a product of real linear and real quadratic polynomials without real roots.

We use the fact that the conjugates of the complex roots are also roots and write P in its complete factorized form:

$$P(z) = 5\,(z-1)(z-\mathrm{i})(z+\mathrm{i})(z-(1+2\mathrm{i}))((z-(1-2\mathrm{i}))(z-2)^2\,.$$

Two pairs of factors correspond to conjugated roots. When we multiply these we obtain the form we wanted:

$$P(z) = 5\,(z-1)(z^2+1)(z^2-2z+5)(z-2)^2\,.$$

By this we end the treatment of polynomials in one variable.

eNote 3

Elementary Functions

In this eNote we will both repeat some of the basic properties for a selection of the (from high school) well-known functions $f(x)$ of one real variable x, and introduce some new functions, which typically occur in a variety of applications. The basic questions concerning any function are usually the following: How, and for which values of x, is the function defined*? Which values for $f(x)$ do we get when we apply the functions to the x-elements in the domain? Is the function* continuous*? What is the* derivative $f'(x)$ *of the function – if it exists? As a new concept, we will introduce a vast* class *of functions, the* ***epsilon functions****, which are denoted by the common symbol $\varepsilon(x)$ and which we will use generally in order to describe continuity and differentiability – also of functions of more variables, which we introduce in the following eNotes.*

(Updated: 22.9.2021 David Brander)

3.1 Domain and Range

In the description of a real function $f(x)$ both the real numbers x where the function is defined and the values that are obtained by applying the function on the domain are stated. The ***Domain*** we denote $D(f)$ and the ***range***, or *image*, we denote $R(f)$.

Note: in higher mathematics, it is usual to define a function by specifying the domain and *codomain*, (the set where the function in principle takes values) rather than the image. For example: $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. The codomain is $\mathbb{R}$, but the range is the set of non-negative numbers $[0, \infty[\subset \mathbb{R}$.

|||| Example 3.1 Some Domains and Ranges

Here are domains and the corresponding ranges for some well-known functions.

$$
\begin{array}{llll}
f_1(x) = \exp(x) & , \; D(f_1) = \mathbb{R} =]-\infty, \infty[& , \; R(f_1) =]0, \infty[\\
f_2(x) = \ln(x) & , \; D(f_2) =]0, \infty[& , \; R(f_2) = \mathbb{R} =]-\infty, \infty[\\
f_3(x) = \sqrt{x} & , \; D(f_3) = \; [0, \infty[& , \; R(f_3) = \; [0, \infty[\\
f_4(x) = x^2 & , \; D(f_4) = \mathbb{R} =]-\infty, \infty[& , \; R(f_4) = \; [0, \infty[\\
f_5(x) = x^7 + 8x^3 + x - 1 & , \; D(f_5) = \mathbb{R} =]-\infty, \infty[& , \; R(f_5) = \mathbb{R} =]-\infty, \infty[\\
f_6(x) = \exp(\ln(x)) & , \; D(f_6) =]0, \infty[& , \; R(f_6) =]0, \infty[\\
f_7(x) = \sin(1/x) & , \; D(f_7) =]-\infty, 0[\cup]0, \infty[& , \; R(f_7) = [-1, 1] \\
f_8(x) = |x|/x & , \; D(f_8) =]-\infty, 0[\cup]0, \infty[& , \; R(f_8) = \{-1\} \cup \{1\}
\end{array}
\tag{3-1}
$$

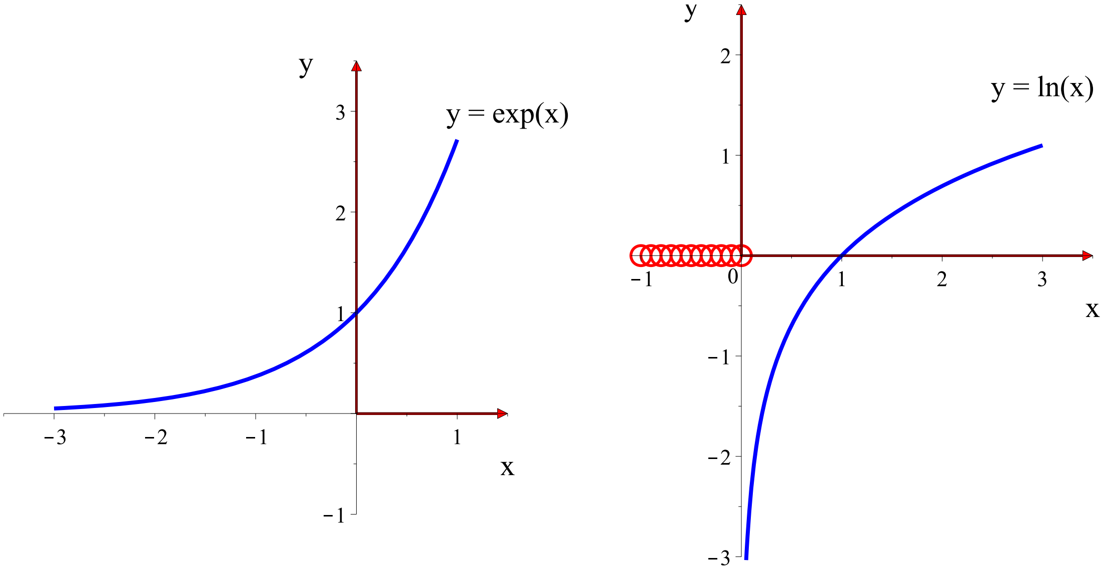


Figure 3.1: The well-known exponential function $e^x = \exp(x)$ and the natural logarithmic function $\ln(x)$. The red circles on the negative x-axis and at 0 indicate that the logarithmic function is not defined on $]-\infty, 0]$.

The function $f_8(x)$ in Example 3.1 is defined using $|x|$, which denotes the absolute value of x, i.e.

$$|x| = \begin{cases} x > 0\,, & \text{for} \quad x > 0 \\ 0\,, & \text{for} \quad x = 0 \\ -x > 0\,, & \text{for} \quad x < 0 \end{cases} \quad . \tag{3-2}$$

From this the domain and range for $f_8(x)$ follow directly.

|||| Example 3.2 Tangent

The function

$$f(x) = \tan(x) = \frac{\sin(x)}{\cos(x)} \tag{3-3}$$

has the domain $D(f) = \mathbb{R} \setminus A$, A denoting those real numbers x for which $\cos(x) = 0$, $\cos(x)$ being the denominator, i.e.

$$D(f) = \mathbb{R} \setminus \{x \mid \cos(x) = 0\} = \mathbb{R} \setminus \{(\pi/2) + p \cdot \pi\,,\ p \text{ being an integer}\} \quad . \tag{3-4}$$

The range $R(f)$ is all real numbers, see Figure 3.2.

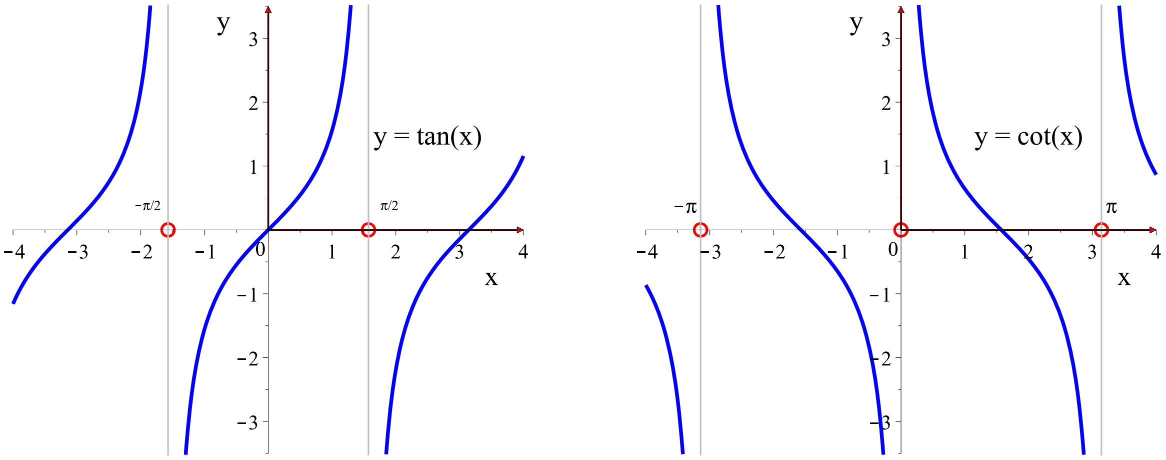


Figure 3.2: The graphs for the functions $\tan(x)$ and $\cot(x)$.

|||| Exercise 3.3

Let $g(x)$ denote the reciprocal function to the function $\tan(x)$:

$$g(x) = \cot(x) = \frac{\cos(x)}{\sin(x)} \tag{3-5}$$

Determine the domain for $g(x)$ and state it in the same way as above for $\tan(x)$, see Figure 3.2.

3.1.1 Extension of the Domain to All of $\mathbb{R}$

A function $f(x)$ that is not defined for all real numbers can easily be *extended* to a function $\widehat{f}(x)$, which has $D(\widehat{f}) = \mathbb{R}$. One way of doing this is by the use of a ***curly bracket*** in the following way:

|||| Definition 3.4

Given a function $f(x)$ with $D(f) \neq \mathbb{R}$. We then define the 0-***extension*** of $f(x)$ by:

$$\widehat{f}(x) = \begin{cases} f(x)\,, & \text{for} \quad x \in D(f) \\ 0\,, & \text{for} \quad x \in \mathbb{R} \setminus D(f) \end{cases} \quad . \tag{3-6}$$

It is evident that depending on the application one can seal and extend the domain for $f(x)$ in many other ways than choosing the constant 0 as the value for the extended function at the points where the original function is not defined.

Naturally, the *Range* $R(\widehat{f})$ for the 0-extended function is the original range for $f(x)$ united with 0, i.e. $R(\widehat{f}) = R(f) \cup \{0\}$.

Hereafter we will assume – unless otherwise stated – that the functions we consider are defined for all $\mathbb{R}$ possibly by extension as above.

3.2 Epsilon Functions

We introduce a special class of functions, which we will use in order to define the important concept of continuity.

|||| Definition 3.5 Epsilon Functions

Every $\varepsilon(x)$ that is defined on an open interval that contains 0 and that assumes the value $\varepsilon(0) = 0$ at $x = 0$ and moreover tends towards 0 when x tends towards 0 is called an ***epsilon function*** of x. Thus epsilon functions are characterized by the properties:

$$\varepsilon(0) = 0 \quad \text{and} \quad \varepsilon(x) \to 0 \quad \text{for} \quad x \to 0 \quad . \tag{3-7}$$

The last condition is equivalent to the fact that the absolute value of $\varepsilon(x)$ can be made as small as possible by choosing the numerical value of x sufficiently small. To be precise the condition means: For every number $a > 0$ there exists a number $b > 0$ such that $|\varepsilon(x)| < a$ for all x satisfying $|x| < b$.

The set of epsilon functions is very large:

|||| Example 3.6 Epsilon Functions

Here are some simple examples of epsilon functions:

$$\begin{aligned} \varepsilon_1(x) &= x \\ \varepsilon_2(x) &= |x| \\ \varepsilon_3(x) &= \ln(1+x) \\ \varepsilon_4(x) &= \sin(x) \quad . \end{aligned} \tag{3-8}$$

The quality 'to be an epsilon function' is rather stable: The product of an epsilon function and an arbitrary other function that only has to be bounded is also an epsilon function. The sum and the product of two epsilon functions are again epsilon functions. The absolute value of an epsilon function is an epsilon function.

Functions that are 0 in other places than $x = 0$ can also be epsilon functions:

If a function $g(x)$ has the properties $g(x_0) = 0$ and $g(x) \to 0$ for $x \to x_0$ then $g(x)$ is an epsilon function of $x - x_0$ i.e. we can write $g(x) = \varepsilon_g(x - x_0)$.

Exercise 3.7

Show that the 0-extension $\widehat{f_8}(x)$ of the function $f_8(x) = |x|/x$ is *not* an epsilon function. Hint: If we choose $k = 10$ then clearly there does *not exist* a value of K such that

$$|f_8(x)| = |\,|x|/x\,| = 1 < \frac{1}{10} \quad , \quad \text{for all } x \text{ with } \quad |x| < \frac{1}{K} \quad . \tag{3-9}$$

Draw the graph for $\widehat{f_8}(x)$. This cannot be drawn without 'lifting the pencil from the paper'!

Exercise 3.8

Show that the 0-extension of the function $f(x) = \sin(1/x)$ is *not* an epsilon function.

3.3 Continuous Functions

We can now formulate the ***concept of continuity*** by use of epsilon functions:

Definition 3.9 Continuity

A function $f(x)$ is continuous at x_0 if there exists an epsilon function $\varepsilon_f(x - x_0)$ such that the following is valid on an open interval that contains x_0:

$$f(x) = f(x_0) + \varepsilon_f(x - x_0) \quad . \tag{3-10}$$

If $f(x)$ is continuous at every x_0 on a given open interval in $D(f)$ we say that $f(x)$ is continuous on the interval.

Note that even though it is clear what the epsilon function precisely is in the definition 3.9, viz. $f(x) - f(x_0)$, then the only property in which we are interested is the following: $\varepsilon_f(x - x_0) \to 0$ for $x \to x_0$ such that $f(x) \to f(x_0)$ for $x \to x_0$, that is precisely as we know the concept of continuity from high school!

|||| Exercise 3.10

According to the above, all epsilon functions are continuous at $x_0 = 0$ (with the value 0 at $x_0 = 0$). Construct an epsilon function that is *not* continuous at any of the points $x_0 = 1/n$ where $n = 1, 2, 3, 4, \cdots$.

Even though the concept of epsilon functions is central to the definition of continuity (and as we shall see below, to the definition of differentiability), epsilon functions need not be continuous for any other values than $x_0 = 0$.

|||| Exercise 3.11

Show that the 0-extension $\widehat{f}(x)$ of the function $f(x) = |x - 7|/(x - 7)$ is *not* continuous on $\mathbb{R}$.

3.4 Differentiable Functions

|||| Definition 3.12 Differentiability

A function $f(x)$ is differentiable at $x_0 \in D(f)$ if both a constant a and an epsilon function $\varepsilon_f(x-x_0)$ exist such that

$$f(x) = f(x_0) + a \cdot (x-x_0) + (x-x_0) \cdot \varepsilon_f(x-x_0) \quad . \tag{3-11}$$

It is the number a that we call $f'(x_0)$ and it is well-defined in the sense that if $f(x)$ *can be* stated at all in the form above (i.e. if $f(x)$ is differentiable at x_0) then there is one and only one value for a that makes this formula true. With this definition of the ***derivative*** $f'(x_0)$ of $f(x)$ at x_0 we then have:

$$f(x) = f(x_0) + f'(x_0) \cdot (x-x_0) + (x-x_0) \cdot \varepsilon_f(x-x_0) \quad . \tag{3-12}$$

If $f(x)$ is differentiable for all x_0 in a given open interval in $D(f)$, we then naturally say that $f(x)$ is differentiable on the interval. We often write the derivative of $f(x)$ at x in the following alternative way:

$$f'(x) = \frac{d}{dx} f(x) \quad . \tag{3-13}$$

|||| Explanation 3.13 The Derivative is Unique

We will show that there is only one value of a that fulfills Equation (3-11). Assume that two different values, a_1 and a_2 both fulfill (3-11) possibly with two different epsilon functions:

$$\begin{aligned} f(x) &= f(x_0) + a_1 \cdot (x-x_0) + (x-x_0) \cdot \varepsilon_1(x-x_0) \\ f(x) &= f(x_0) + a_2 \cdot (x-x_0) + (x-x_0) \cdot \varepsilon_2(x-x_0) \quad . \end{aligned} \tag{3-14}$$

By subtracting (3-14) from the uppermost equation we get:

$$0 = 0 + (a_1 - a_2) \cdot (x-x_0) + (x-x_0) \cdot (\varepsilon_1(x-x_0) - \varepsilon_2(x-x_0)) \quad , \tag{3-15}$$

such that

$$a_2 - a_1 = \varepsilon_1(x-x_0) - \varepsilon_2(x-x_0) \tag{3-16}$$

for all $x \neq x_0$ – and clearly this cannot be true; the right hand side tends towards 0 when x tends towards x_0! Therefore the above assumption, i.e. that $a_1 \neq a_2$, is wrong. The two constants a_1 and a_2 must be equal, and this is what we should realize.

The definition above is quite equivalent to the one we know from high school. If we first subtract $f(x_0)$ from both sides of the equality sign in Equation (3-12) and then divide by $(x - x_0)$ we get

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) + \varepsilon_f(x - x_0) \rightarrow f'(x_0) \quad \text{for} \quad x \rightarrow x_0 \quad , \tag{3-17}$$

i.e. the well-known limit value for the *quotient* between the increment in the function $f(x) - f(x_0)$ and the x-increment $x - x_0$. The reason why we do not apply this known definition of $f'(x_0)$ is simply that for functions of more variables the quotient does not make sense – but more about this in a later eNote.

|||| Theorem 3.14 Differentiable Implies Continuous

If a function $f(x)$ is differentiable at x_0, then $f(x)$ is also continuous at x_0.

|||| Proof

We have that

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0) \cdot (x - x_0) + (x - x_0)\varepsilon_f(x - x_0) \\ &= f(x_0) + \left[f'(x_0) \cdot (x - x_0) + (x - x_0)\varepsilon_f(x - x_0) \right] , \end{aligned} \tag{3-18}$$

and since the function in the square brackets on the right hand side is an epsilon function of $(x - x_0)$ then $f(x)$ is continuous at x_0.

■

But the opposite is not valid – here is an example that shows this:

|||| Example 3.15 Continuous But Not Differentiable

The function $f(x) = |x|$ is continuous but not differentiable at $x_0 = 0$. The function is in itself an epsilon function and therefore $f(x)$ is continuous in 0. But now assume that there exist a constant a and an epsilon function $\varepsilon_f(x - x_0)$ such that

$$f(x) = f(x_0) + a \cdot (x - x_0) + (x - x_0)\varepsilon_f(x - x_0). \tag{3-19}$$

The following will then apply:

$$|x| = 0 + a \cdot x + x \cdot \varepsilon_f(x) \tag{3-20}$$

and hence for all $x \neq 0$:

$$\frac{|x|}{x} = a + \varepsilon_f(x) \quad . \tag{3-21}$$

If so a should both be equal to -1 and to 1 and this is impossible! Therefore the assumption above that there exists a constant a is accordingly wrong; therefore $f(x)$ is not differentiable.

|||| Definition 3.16

The first degree approximating polynomial for $f(x)$ expanded about the point x_0 is defined by:

$$P_{1,x_0}(x) = f(x_0) + f'(x_0) \cdot (x - x_0) \quad . \tag{3-22}$$

Note that $P_{1,x_0}(x)$ really *is* a first degree polynomial in x. The graph for the function $P_{1,x_0}(x)$ is ***the tangent to the graph*** for $f(x)$ at the point $(x_0, f(x_0))$, see Figure 3.3. The equation for the tangent is $y = P_{1,x_0}(x)$, thus $y = f(x_0) + f'(x_0) \cdot (x - x_0)$. The slope of the tangent is clearly $\alpha = f'(x_0)$ and the tangent intersects the y-axis at the point $(0, f(x_0) - x_0 \cdot f'(x_0))$. Later we will find out how we can approximate with polynomials of higher degree n, i.e. polynomials that are then denoted $P_{n,x_0}(x)$.

3.4.1 Differentiation of a Product

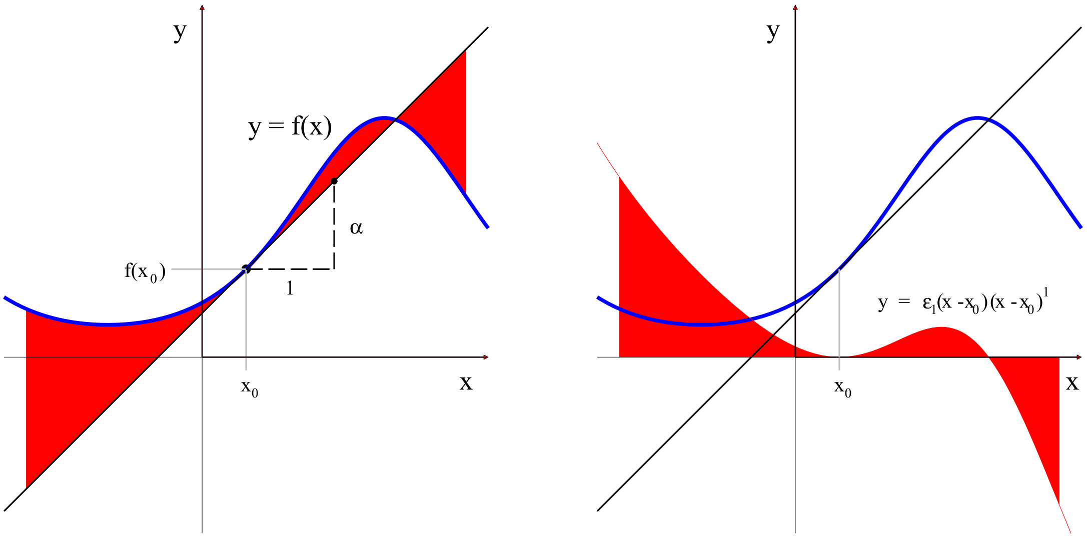


Figure 3.3: Construction of the tangent $y = P_{1,x_0}(x) = f(x_0) + \alpha \cdot (x - x_0)$ with the slope $\alpha = f'(x_0)$ for the function $f(x)$. To the right the difference between $f(x)$ and the 'tangent value' $P_{1,x_0}(x)$.

|||| Theorem 3.17 Differentiation of $f(x) \cdot g(x)$

A product $h(x) = f(x) \cdot g(x)$ of two differentiable functions $f(x)$ and $g(x)$ is differentiable and its derivative is as follows:

$$\frac{d}{dx}\left(f(x) \cdot g(x)\right) = f'(x) \cdot g(x) + f(x) \cdot g'(x) \quad . \tag{3-23}$$

Even though this formula is rather well known from high school we shall give a short sketch of a proof – to illustrate the use of epsilon functions.

|||| Proof

Since $f(x)$ and $g(x)$ are differentiable in x_0, we have:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)\cdot(x-x_0) + (x-x_0)\varepsilon_f(x-x_0) \\ g(x) &= g(x_0) + g'(x_0)\cdot(x-x_0) + (x-x_0)\varepsilon_g(x-x_0) \quad , \end{aligned} \tag{3-24}$$

resulting in the product of the two right hand sides:

$$\begin{aligned} &h(x) = f(x)\cdot g(x) \\ &= f(x_0)\cdot g(x_0) + (f'(x_0)\cdot g(x_0) + f(x_0)\cdot g'(x_0))\cdot(x-x_0) + (x-x_0)\varepsilon_h(x-x_0)\,, \end{aligned} \tag{3-25}$$

where we have used $(x-x_0)\varepsilon_h(x-x_0)$ as short for the remaining part of the product sum. Furthermore any of the addends in the remaining part contains the factor $(x-x_0)^2$ or the product of $(x-x_0)$ with an epsilon function and therefore *can* be written in the stated form. But then the product formula follows directly from the factor in front of $(x-x_0)$ in Equation (3-25):

$$h'(x_0) = f'(x_0)\cdot g(x_0) + f(x_0)\cdot g'(x_0) \quad . \tag{3-26}$$

■

3.4.2 Differentiation of a Quotient

The following differentiation rule is also well known from high school:

|||| Theorem 3.18 Differentiation of $f(x)/g(x)$

A quotient $h(x) = f(x)/g(x)$ involving two differentiable functions $f(x)$ and $g(x)$, is differentiable everywhere that $g(x) \neq 0$, and the derivative is given in this well-known fashion:

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)}{g(x)} - \frac{f(x)\cdot g'(x)}{g^2(x)} = \frac{f'(x)\cdot g(x) - f(x)\cdot g'(x)}{g^2(x)} \quad . \tag{3-27}$$

|||| Exercise 3.19

Use the epsilon function argument in the same way as in the differentiation rule for a product to show Equation 3.18.

3.4.3 Differentiation of Composite Functions

|||| Theorem 3.20 The Chain Rule for Composite Functions

A function $h(x) = f(g(x))$ that is composed of two differentiable functions $f(x)$ and $g(x)$ is in itself differentiable at every x_0 with the derivative

$$h'(x_0) = f'(g(x_0)) \cdot g'(x_0) \tag{3-28}$$

|||| Proof

We exploit that the two functions $f(x)$ and $g(x)$ are differentiable. In particular $g(x)$ is differentiable at x_0:

$$g(x) = g(x_0) + g'(x_0)(x - x_0) + (x - x_0) \cdot \varepsilon_g(x - x_0) \quad , \tag{3-29}$$

and the function $f(u)$ is differentiable at $u_0 = g(x_0)$:

$$f(u) = f(u_0) + f'(u_0)(u - u_0) + (u - u_0) \cdot \varepsilon_f(u - u_0) \quad . \tag{3-30}$$

From this we get, setting $u = g(x)$ and $u_0 = g(x_0)$:

$$\begin{aligned} h(x) &= f(g(x)) \\ &= f(g(x_0)) + f'(g(x_0))(g(x) - g(x_0) + (g(x) - g(x_0) \cdot \varepsilon_f(g(x) - g(x_0) \\ &= h(x_0) + f'(g(x_0))(g'(x_0)(x - x_0) + (x - x_0) \cdot \varepsilon_g(x - x_0)) \\ &\qquad + (g'(x_0)(x - x_0) + (x - x_0) \cdot \varepsilon_g(x - x_0)) \cdot \varepsilon_f(g(x) - g(x_0) \\ &= h(x_0) + f'(g(x_0))g'(x_0) \cdot (x - x_0) + (x - x_0) \cdot \varepsilon_h(x - x_0) \quad , \end{aligned} \tag{3-31}$$

from which we directly read that $h'(x_0) = f'(g(x_0))g'(x_0)$ – because this is exactly the unique coefficient of $(x - x_0)$ in the above expression.

■

|||| Exercise 3.21

Above we have used – at the end of Equation (3-31) – that

$$f'(g(x_0)) \cdot \varepsilon_g(x - x_0) + (g'(x_0) + \cdot\varepsilon_g(x - x_0)) \cdot \varepsilon_f(g(x) - g(x_0)) \tag{3-32}$$

is an epsilon function, which we accordingly can call (and have called) $\varepsilon_h(x - x_0)$. Consider why this is entirely OK.

|||| Exercise 3.22

Find the derivatives of the following functions for every x-value in their respective domains:

$$\begin{aligned} f_1(x) &= (x^2 + 1) \cdot \sin(x) \\ f_2(x) &= \sin(x)/(x^2 + 1) \\ f_3(x) &= \sin(x^2 + 1) \quad . \end{aligned} \tag{3-33}$$

3.5 Inverse Functions

The exponential function $\exp(x)$ and the logarithmic function $\ln(x)$ are ***inverse functions*** to each other – as is well known the following is valid:

$$\begin{aligned} \exp(\ln(x)) = x \quad &\text{for} \quad x \in D(\ln) =]0, \infty[= R(\exp) \\ \ln(\exp(x)) = x \quad &\text{for} \quad x \in D(\exp) =] - \infty, \infty[= R(\ln) \quad . \end{aligned} \tag{3-34}$$

Note that even though $\exp(x)$ is defined for all x, the inverse function $\ln(x)$ is only defined for $x > 0$ – and vice versa (!).

The function $f(x) = x^2$ has an inverse function in its respective intervals of monotony, i.e. where $f(x)$ is either increasing or decreasing: The inverse function on the interval $[0, \infty[$ where $f(x)$ is increasing is the well-known function $g(x) = \sqrt{x}$. Thus the function $f(x)$ maps the interval $A = [0, \infty[$ one-to-one onto the interval $B = [0, \infty[$, and the inverse function $g(x)$ maps the interval B one-to-one onto the interval A such that:

$$\begin{aligned} f(g(x)) &= (\sqrt{x})^2 = x \quad \text{for} \quad x \in B = [0, \infty[\\ g(f(x)) &= \sqrt{x^2} = x \quad \text{for} \quad x \in A = [0, \infty[\quad . \end{aligned} \tag{3-35}$$

The inverse function to $f(x)$ on the interval $]-\infty, 0]$ where $f(x)$ is decreasing is the function $h(x) = -\sqrt{x}$, which is not defined on the *same* interval as $f(x)$. The function $f(x)$ maps the interval $C =]-\infty, 0]$ one-to-one onto the interval $D = [0, \infty[$, and the inverse function $h(x)$ maps the interval D one-to-one onto the interval C such that:

$$\begin{aligned} f(h(x)) &= (-\sqrt{x})^2 = x \quad \text{for} \quad x \in D = [0, \infty[\\ h(f(x)) &= -\sqrt{x^2} = x \quad \text{for} \quad x \in C =]-\infty, 0] \quad . \end{aligned} \tag{3-36}$$

If $f(x)$ is not monotonic on an interval, it means that we can obtain the *same function-value* $f(x)$ for more x-values – in the same way as $x^2 = 1$ both for $x = 1$ and for $x = -1$, and then the function is not one-to-one on the interval. The functions $\cos(x)$ and $\sin(x)$ are only monotonic on certain subintervals of the x-axis, see Figure 3.7. If we wish to define inverse functions to the functions we must choose the interval with care, see Section 3.8 and Figure 3.8.

|||| Definition 3.23 Notation for Inverse Functions

We denote the inverse function for a given function $f(x)$ by $f^{\circ -1}(x)$. The inverse function is generally defined by the following properties on suitably chosen intervals A and B that are part of $D(f)$ and $D(f^{\circ -1})$, respectively

$$\begin{aligned} f^{\circ -1}(f(x)) &= x \quad \text{for} \quad x \in A \subset D(f) \\ f(f^{\circ -1}(x)) &= x \quad \text{for} \quad x \in B \subset D(f^{\circ -1}) \quad . \end{aligned} \tag{3-37}$$

We use here the symbol $f^{\circ -1}(x)$ in order to avoid confusion with $(f(x))^{-1} = 1/f(x)$. However the reader should note that the standard notation is simply f^{-1} for the inverse function. The graph for the inverse function $g(x) = f^{\circ -1}(x)$ to a function $f(x)$ can be obtained by ***mirroring the graph*** for $f(x)$ in the diagonal in the first quadrant in the (x, y)-coordinate system – i.e. the line with the equation $y = x$ – see Figure 3.4.

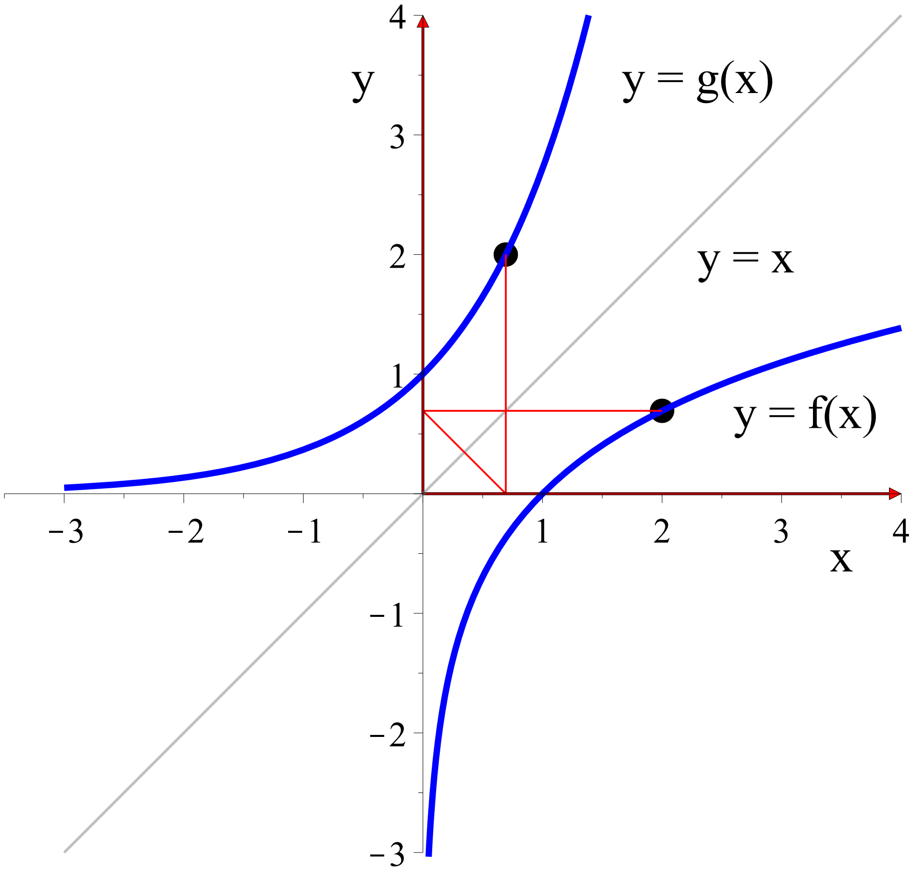


Figure 3.4: The graph for a function $f(x)$ and the graph for the inverse function $g(x)$. It is valid that $g(x) = f^{\circ-1}(x)$ and $f(x) = g^{\circ-1}(x)$, but they each have their own definition intervals.

3.5.1 Differentiation of Inverse Functions

|||| Theorem 3.24 Differentiation of Inverse Functions

If a differentiable function $f(x)$ has the inverse function $f^{\circ-1}(x)$ and if $f'(f^{\circ-1}(x_0)) \neq 0$, then the inverse function$f^{\circ-1}(x)$ is itself differentiable at x_0:

$$(f^{\circ-1})'(x_0) = \frac{1}{f'(f^{\circ-1}(x_0))} \qquad (3\text{-}38)$$

|||| Proof

From the definition of inverse functions we have

$$h(x) = f(f^{\circ -1}(x)) = x \quad , \tag{3-39}$$

so $h'(x_0) = 1$, but we also have from the chain rule in (3-28):

$$h'(x_0) = f'(f^{\circ -1}(x_0)) \cdot (f^{\circ -1})'(x_0) = 1 \quad , \tag{3-40}$$

from which we get the result by dividing by $f'(f^{\circ -1}(x_0))$.

■

3.6 Hyperbolic Functions

|||| Definition 3.25 Hyperbolic Cosine and Hyperbolic Sine

We will define two new functions $\cosh(x)$ and $\sinh(x)$ as the unique solution to the following system of differential equations with initial conditions. The two solutions are denoted ***hyperbolic cosine*** and ***hyperbolic sine***, respectively:

$$\begin{aligned} \cosh'(x) = \sinh(x) \quad &, \quad \cosh(0) = 1 \\ \sinh'(x) = \cosh(x) \quad &, \quad \sinh(0) = 0 \quad . \end{aligned} \tag{3-41}$$

The *names* $\cosh(x)$ and $\sinh(x)$ (often spoken as "cosh" and "sinsh") look like $\cos(x)$ and $\sin(x)$, but the functions are very different, as we shall demonstrate below.

Yet there are also fundamental structural similarities between the two pairs of functions and this is what motivates the names. In the system of differential equations for $\cos(x)$

and $\sin(x)$ only a single minus sign separates this from (3-41):

$$\begin{aligned} \cos'(x) &= -\sin(x) \quad , \quad \cos(0) = 1 \\ \sin'(x) &= \cos(x) \quad , \quad \sin(0) = 0 \quad . \end{aligned} \tag{3-42}$$

In addition (again with the decisive minus sign as the only difference) the following simple analogy to the well-known and often used relation $\cos^2(x) + \sin^2(x) = 1$ applies:

|||| Theorem 3.26 Fundamental Relation of $\cosh(x)$ and $\sinh(x)$

$$\cosh^2(x) - \sinh^2(x) = 1 \quad . \tag{3-43}$$

|||| Proof

Make the derivative with respect to x on both sides of the equation (3-43) and conclude that $\cosh^2(x) - \sinh^2(x)$ is a constant. Finally use the initial conditions.

■

|||| Exercise 3.27

Show directly from the system of differential equations (3-41) that the two ”new” functions are in fact not so new:

$$\begin{aligned} \cosh(x) &= \frac{e^x + e^{-x}}{2} \quad , \quad D(\cosh) = \mathbb{R} \quad , \quad R(\cosh) = [1, \infty[\\ \sinh(x) &= \frac{e^x - e^{-x}}{2} \quad , \quad D(\sinh) = \mathbb{R} \quad , \quad R(\sinh) =]-\infty, \infty[\end{aligned} \tag{3-44}$$

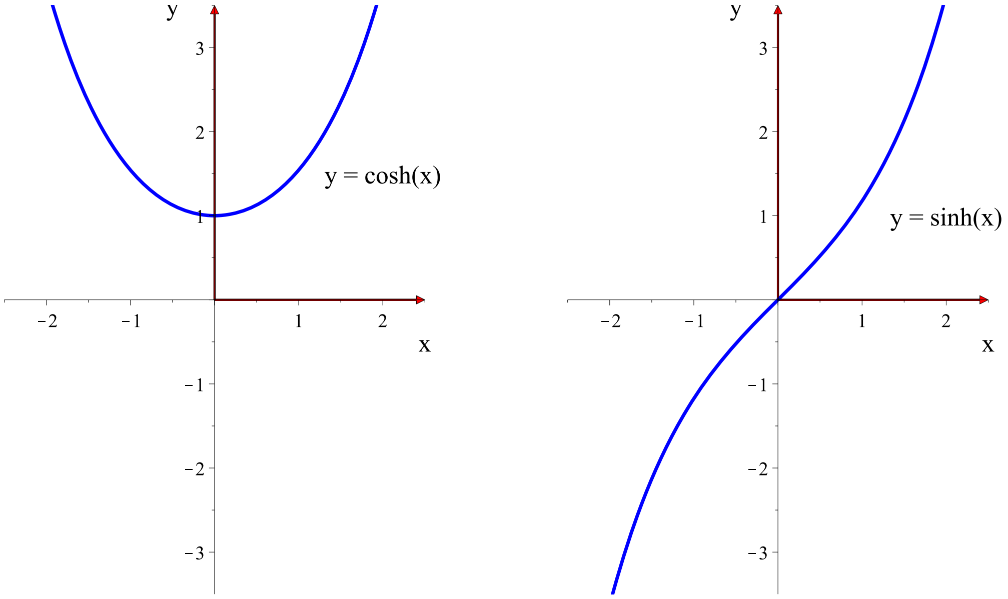


Figure 3.5: Hyperbolic cosine, $\cosh(x)$, and hyperbolic sine, $\sinh(x)$.

|||| Exercise 3.28

Show *directly* from the expressions found in Exercise 3.27, that

$$\cosh^2(x) - \sinh^2(x) = 1 \quad . \tag{3-45}$$

|||| Exercise 3.29

The graph for the function $f(x) = \cosh(x)$ looks a lot like a parabola, viz. the graph for the function $g(x) = 1 + (x^2/2)$ when we plot both functions on a suitably small interval around $x_0 = 0$. Try this! If we instead plot the two graphs in very large x-interval, we learn that the two functions have very different graphical behaviours. Try this, i.e. try to plot both functions on the interval $[-50, 50]$. Comment upon and explain the qualitative differences. Similarly compare the two functions $\sinh(x)$ and $x + (x^3/6)$ in the same way.

It is natural and useful to define hyperbolic analogies to $\tan(x)$ and $\cot(x)$. This is done as follows:

|||| Definition 3.30 Hyperbolic Tangent and Hyperbolic Cotangent

$$\begin{aligned} \tanh(x) &= \frac{\sinh(x)}{\cosh(x)} = \frac{e^{2x}-1}{e^{2x}+1} \;,\; D(\tanh) = \mathbb{R} \;,\; R(\tanh) =]-1,1[\\ \coth(x) &= \frac{\cosh(x)}{\sinh(x)} = \frac{e^{2x}+1}{e^{2x}-1} \;,\; D(\coth) = \mathbb{R} - \{0\} \;, \\ & \qquad R(\coth) =]-\infty,-1[\cup]1,\infty[\quad . \end{aligned} \tag{3-46}$$

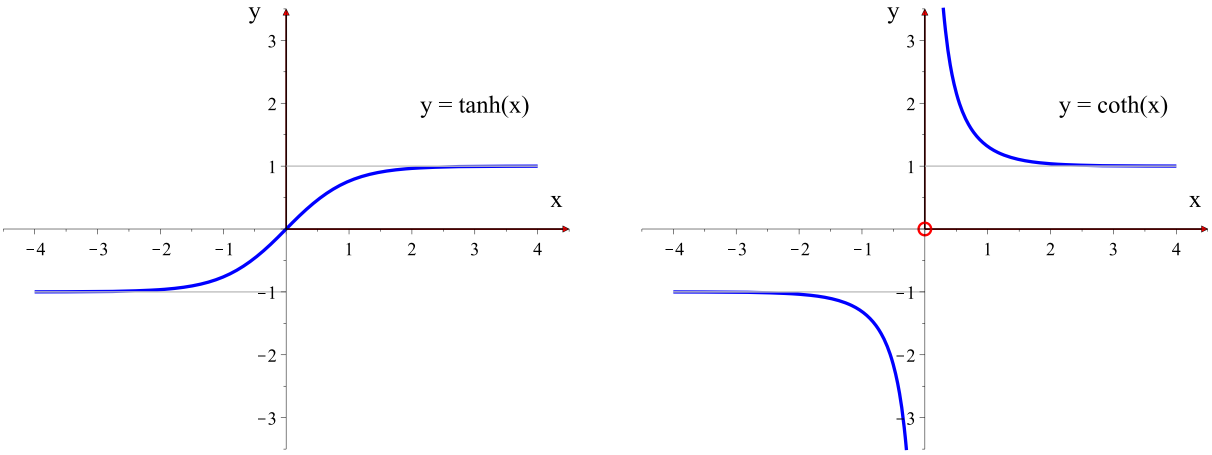


Figure 3.6: Hyperbolic tangent, $\tanh(x)$, and hyperbolic cotangent, $\coth(x)$.

The derivatives of $\cosh(x)$ and of $\sinh(x)$ are already given by the defining system in (3-41).

$$\begin{aligned} \frac{d}{dx}\cosh(x) &= \sinh(x) \\ \frac{d}{dx}\sinh(x) &= \cosh(x) \\ \frac{d}{dx}\tanh(x) &= \frac{1}{\cosh^2(x)} = 1 - \tanh^2(x) \\ \frac{d}{dx}\coth(x) &= \frac{-1}{\sinh^2(x)} = 1 - \coth^2(x) \quad . \end{aligned} \tag{3-47}$$

|||| Exercise 3.31

Show the last two expressions for the derivatives for $\tanh(x)$ and $\coth(x)$ in (3-47) by the use of the differentiation rule in Theorem 3.18.

3.7 The Area Functions

The inverse functions to the hyperbolic functions are called ***area functions*** and are named $\cosh^{\circ -1}(x) = \operatorname{arcosh}(x)$, $\sinh^{\circ -1}(x) = \operatorname{arsinh}(x)$, $\tanh^{\circ -1}(x) = \operatorname{artanh}(x)$, and $\coth^{\circ -1}(x) = \operatorname{arcoth}(x)$, respectively.

Since the functions $\cosh(x)$, $\sinh(x)$, $\tanh(x)$, and $\coth(x)$ all can be expressed in terms of exponential functions it is no surprise that the inverse functions and their derivatives can be expressed by logarithmic functions. We gather the information here:

$$\begin{aligned}
\operatorname{arcosh}(x) &= \ln(x+\sqrt{x^2-1}) \text{ for } \quad x\in[1,\infty[\\
\operatorname{arsinh}(x) &= \ln(x+\sqrt{x^2+1}) \text{ for } \quad x\in\mathbb{R} \\
\operatorname{artanh}(x) &= \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) \quad \text{for} \quad x\in]-1,1[\\
\operatorname{arcoth}(x) &= \frac{1}{2}\ln\left(\frac{x-1}{x+1}\right) \quad \text{for} \quad x\in]-\infty,1[\cup]1,\infty[\quad .
\end{aligned} \tag{3-48}$$

$$\begin{aligned}
\frac{d}{dx}\operatorname{arcosh}(x) &= \frac{1}{\sqrt{x^2-1}} \quad \text{for} \quad x\in]1,\infty[\\
\frac{d}{dx}\operatorname{arsinh}(x) &= \frac{1}{\sqrt{x^2+1}} \quad \text{for} \quad x\in\mathbb{R} \\
\frac{d}{dx}\operatorname{artanh}(x) &= \frac{1}{1-x^2} \quad \text{for} \quad x\in]-1,1[\\
\frac{d}{dx}\operatorname{arcoth}(x) &= \frac{1}{1-x^2} \quad \text{for} \quad x\in]-\infty 1[\cup]1,\infty[\quad .
\end{aligned} \tag{3-49}$$

3.8 The Arc Functions

The inverse functions to the trigonometric functions are a bit more complicated. As mentioned earlier here we must choose for each trigonometric function an interval

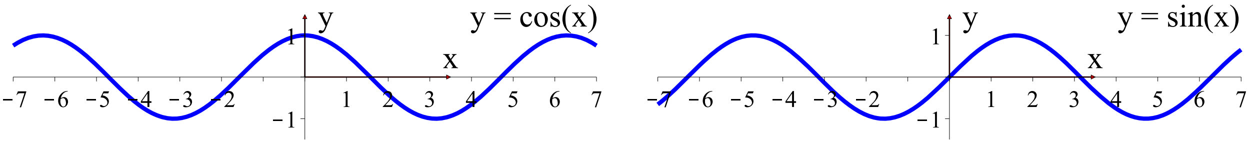


Figure 3.7: Cosine and Sine Functions.

where the function in question is monotonic. In return, once we *have* chosen such an interval, it is clear how the inverse function should be defined and how it should then be differentiated. The inverse functions to $\cos(x)$, $\sin(x)$, $\tan(x)$, and $\cot(x)$ are usually written $\arccos(x)$, $\arcsin(x)$, $\arctan(x)$, and $\operatorname{arccot}(x)$, respectively; their names are arccosine, arcsine, arctangent, and arccotangent. As above we gather the results here:

$$\begin{aligned}
\cos^{\circ -1}(x) &= \arccos(x) \in [0,\pi] \text{ for } \quad x \in [-1,1] \\
\sin^{\circ -1}(x) &= \arcsin(x) \in [-\pi/2,\pi/2] \text{ for } \quad x \in [-1,1] \\
\tan^{\circ -1}(x) &= \arctan(x) \in [-\pi/2,\pi/2] \text{ for } \quad x \in \mathbb{R} \\
\cot^{\circ -1}(x) &= \operatorname{arccot}(x) \in]0,\pi[\text{ for } \quad x \in \mathbb{R} \quad .
\end{aligned} \tag{3-50}$$

$$\begin{aligned}
\frac{d}{dx}\arccos(x) &= \frac{-1}{\sqrt{1-x^2}} \text{ for } \quad x \in]-1,1[\\
\frac{d}{dx}\arcsin(x) &= \frac{1}{\sqrt{1-x^2}} \text{ for } \quad x \in]-1,1[\\
\frac{d}{dx}\arctan(x) &= \frac{1}{1+x^2} \text{ for } \quad x \in \mathbb{R} \\
\frac{d}{dx}\operatorname{arccot}(x) &= \frac{-1}{1+x^2} \text{ for } \quad x \in \mathbb{R} \quad .
\end{aligned} \tag{3-51}$$

Note that the derivatives for $\arccos(x)$ and $\arcsin(x)$ are not defined at $x_0 = 1$ or at $x_0 = -1$. This is partly because, if the function we consider is only defined on a bounded interval then we cannot say that the function is differentiable at the end-points of the interval. Moreover the formulas for $\arccos'(x)$ and $\arcsin'(x)$ show that they are not defined at $x_0 = 1$ or $x_0 = -1$; these values give 0 in the denominators.

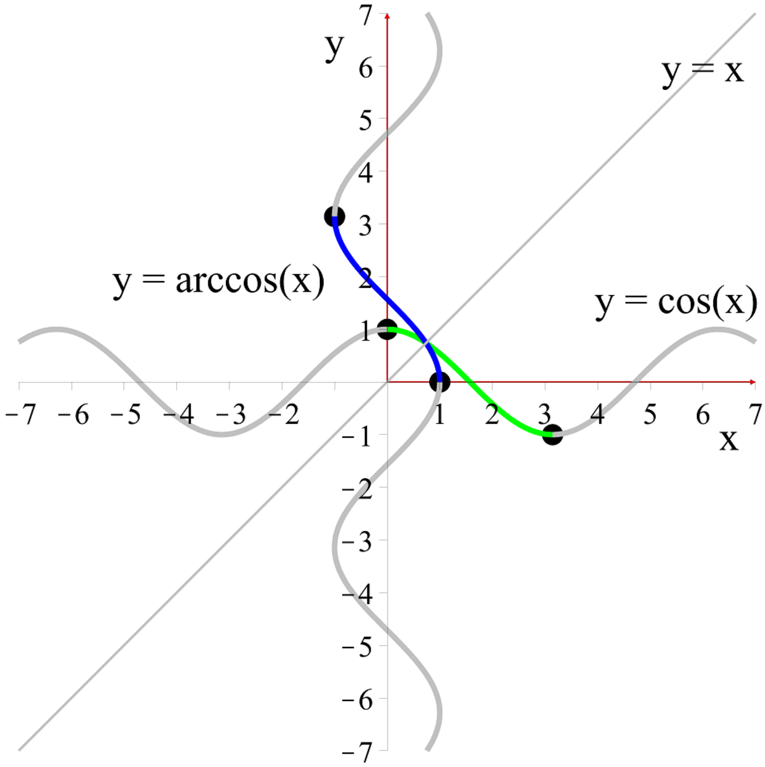


Figure 3.8: The arccosine function is defined here.

|||| Exercise 3.32

Use a suitable modification of $\arctan(x)$ in order to determine a new differentiable (and hence continuous) function $f(x)$ that looks like the 0-extension of $|x|/x$ (which is neither continuous nor differentiable), i.e. we want a function $f(x)$ with the following properties: $1 > f(x) > 0.999$ for $x > 0.001$ while $-0.999 > f(x) > -1$ for $x < -0.001$. See Figure 3.10. Hint: Try to plot $\arctan(1000x)$.

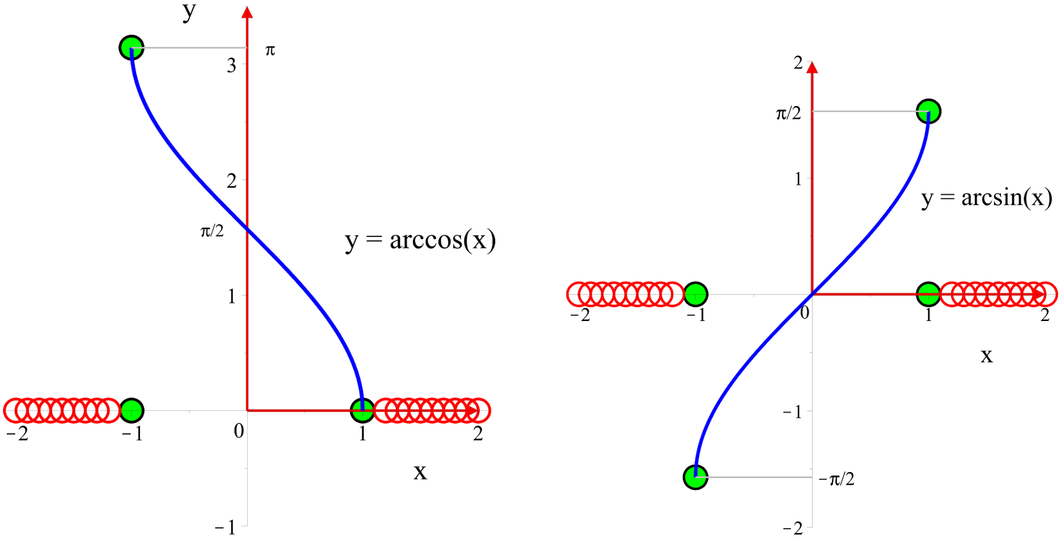


Figure 3.9: Arccosine and arcsine. Again the red circles indicate that the arc-functions are not defined outside the interval $[-1, 1]$. Similarly the green circular disks indicate that the arc-functions *are* defined at the end-points $x = 1$ and $x = -1$.

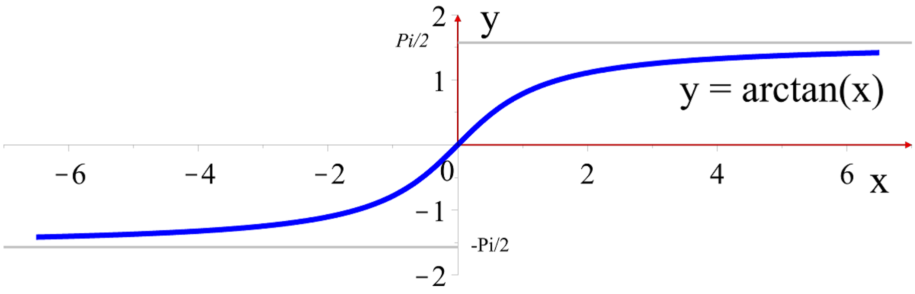


Figure 3.10: The arctangent function.

3.9 Summary

We have treated some of the fundamental properties of some well-known and some not so well-known functions. How are they defined, what are their domains, are they continuous, are they differentiable, and if so what are their derivatives?

- A function $f(x)$ is continuous at x_0 if $f(x) - f(x_0)$ is an epsilon function of $(x - x_0)$, i.e.
$$f(x) = f(x_0) + \varepsilon_f(x - x_0) \quad . \tag{3-52}$$
- A function $f(x)$ is differentiable at x_0 with the derivative $f'(x_0)$ if
$$f(x) = f(x_0) + f'(x_0)(x - x_0) + (x - x_0)\varepsilon_f(x - x_0) \quad .$$
- If a function is differentiable at x_0, then it is also continuous at x_0. The converse does *not* apply.
- The derivative of a product of two functions is
$$\frac{d}{dx}\left(f(x) \cdot g(x)\right) = f'(x) \cdot g(x) + f(x) \cdot g'(x) \quad . \tag{3-53}$$
- The derivative of a quotient of two functions is
$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)}{g(x)} - \frac{f(x) \cdot g'(x)}{g^2(x)} = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)} \quad . \tag{3-54}$$
- The derivative of a composite function is
$$\frac{d}{dx}f(g(x)) = f'(g(x)) \cdot g'(x) \quad . \tag{3-55}$$
- The derivative of the inverse function $f^{\circ -1}(x)$ is
$$\left(f^{\circ -1}\right)'(x) = \frac{1}{f'(f^{\circ -1}(x))} \quad . \tag{3-56}$$

eNote 4

Taylor's Approximation Formulas for Functions of One Variable

In eNotes **??** *and* **??** *it is shown how functions of one and two variables can be approximated by first-degree polynomials at every (development) point and that the graphs for the approximating first-degree polynomial are exactly the tangents and the tangent planes, respectively, for the corresponding graphs of the functions. In this eNote we will show how the functions can be approximated even better by polynomials of higher degree, so if the approximation to a function is sufficiently good then one can use and continue the computations with the approximation polynomial in place of the function itself and hope for a sufficiently small error. But what does it mean that the approximation and the error are* sufficiently good *and* sufficiently small*? And how does this depend on the degree of the approximating polynomial? You will find the answers to these questions in this eNote.*
(Updated: 22.09.2021 David Brander).

4.1 Higher Order Derivatives

First we consider functions $f(x)$ of one variable x on an open interval of the real numbers. We will also assume that the functions can be differentiated an arbitrary number of times, that is, all the derivatives exist for every x in the interval: $f'(x_0)$, $f''(x_0)$, $f'''(x_0)$, $f^{(4)}(x_0)$, $f^{(5)}(x_0)$, etc. where $f^{(4)}(x_0)$ means the 4th derivative of $f(x)$ in x_0. These higher order derivatives we will use in the construction of (the coefficients to) the approximating polynomials.

Definition 4.1

If a function $f(x)$ can be differentiated an abitrary number of times at every point x in a given open interval I we say that the function is ***smooth*** on the interval I.

Example 4.2 Higher-Order Derivatives of Some Elementary Functions

Here are some higher-order derivatives of some well-known smooth functions:

$f(x)$	$f'(x)$	$f''(x)$	$f'''(x)$	$f^{(4)}(x)$	$f^{(5)}(x)$
e^x	e^x	e^x	e^x	e^x	e^x
x^2	$2x$	2	0	0	0
x^3	$3x^2$	$6x$	6	0	0
x^4	$4x^3$	$12x^2$	$24x$	24	0
x^5	$5x^4$	$20x^3$	$60x^2$	$120x$	120
$(x-x_0)^5$	$5\cdot(x-x_0)^4$	$20\cdot(x-x_0)^3$	$60\cdot(x-x_0)^2$	$120\cdot(x-x_0)$	120
$\cos(x)$	$-\sin(x)$	$-\cos(x)$	$\sin(x)$	$\cos(x)$	$-\sin(x)$
$\sin(x)$	$\cos(x)$	$-\sin(x)$	$-\cos(x)$	$\sin(x)$	$\cos(x)$
$\cosh(x)$	$\sinh(x)$	$\cosh(x)$	$\sinh(x)$	$\cosh(x)$	$\sinh(x)$
$\sinh(x)$	$\cosh(x)$	$\sinh(x)$	$\cosh(x)$	$\sinh(x)$	$\cosh(x)$

(4-1)

Note that

1. The nth derivative $f^{(n)}(x)$ of the function $f(x) = (x - x_0)^n$ is

$$f^{(n)}(x) = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 = n! \quad , \tag{4-2}$$

   Where $n!$ (n factorial) is the short way of writing the product of the natural numbers from and including 1 to and including n, cf. Table 4.2 where $n!$ appears as $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$. Note: by definition $0! = 1$, so $n!$ is well-defined for non-negative integers.

2. By repeated differentiation of $\cos(x)$ we get the same set of functions periodically with the period 4: If $f(x) = \cos(x)$ then

$$f^{(p)}(x) = f^{(p+4)}(x) \quad \text{for all } p \geq 1 \quad . \tag{4-3}$$

   The same applies for $f(x) = \sin(x)$.

3. By repeated differentiation of the hyperbolic cosine function $\cosh(x)$ we again get the same ”set” of functions periodically with the period 2: If $f(x) = \cosh(x)$ we get

$$f^{(p)}(x) = f^{(p+2)}(x) \quad \text{for all } p \geq 1 \quad . \tag{4-4}$$

   This applies to the hyperbolic sine function $f(x) = \sinh(x)$, too.

|||| Example 4.3 The Derivatives of a Somewhat Less Elementary Function

A function $f(x)$ can e.g. be given as an integral (that in this case can be expressed by the ordinary elementary functions):

$$f(x) = \int_0^x \mathrm{e}^{-t^2}\, dt \quad . \tag{4-5}$$

But we can easily find the higher order derivatives of the function for every x:

$$f'(x) = \mathrm{e}^{-x^2} \quad , \quad f''(x) = -2 \cdot x \cdot \mathrm{e}^{-x^2} \quad , \quad f'''(x) = -2 \cdot \mathrm{e}^{-x^2} + 4 \cdot x^2 \cdot \mathrm{e}^{-x^2} \quad \text{etc.} \tag{4-6}$$

|||| Example 4.4 The Derivatives of an Unknown Function

We assume that a function $f(x)$ is given as a solution to a differential equation with the initial conditions at x_0:

$$f''(x) + 3f'(x) + 7f(x) = q(x) \quad , \quad \text{where} \quad f(x_0) = 1 \quad , \quad \text{and} \quad f'(x_0) = -3 \tag{4-7}$$

where $q(x)$ is a given smooth function of x. Again we can fairly easily find the higher order derivatives of the function at x_0 by using the initial conditions directly and by *differentiating the differential equation*. We get the following from the initial conditions and from the differential equation itself:

$$f'(x_0) = -3 \quad , \quad f''(x_0) = q(x_0) - 3f'(x_0) - 7f(x_0) = q(x_0) + 2 \quad . \tag{4-8}$$

The third (and the higher-order) derivatives of $f(x)$ we then obtain by differentiating both sides of the differential equation. E.g. by differentiating once we get:

$$f'''(x) + 3f''(x) + 7f'(x) = q'(x) \quad , \tag{4-9}$$

from which we get:

$$\begin{aligned} f'''(x_0) &= q'(x_0) - 3f''(x_0) - 7f'(x_0) \\ &= q'(x_0) - 3 \cdot (q(x_0) + 2) - 7 \cdot (-3) \\ &= q'(x_0) - 3q(x_0) + 15 \quad . \end{aligned} \tag{4-10}$$

4.2 Approximations by Polynomials

The point of the following to find the polynomial of degree n (e.g. the second-degree polynomial) that best approximates a given smooth function $f(x)$ at and around a given x_0 in the domain of the function $D(f)$.

For the case $n = 2$, we try to write $f(x)$ in the following way:

$$f(x) = a_0 + a_1 \cdot (x - x_0) + a_2 \cdot (x - x_0)^2 + R_{2,x_0}(x) \quad , \tag{4-11}$$

where a_0, a_1, and a_2 are suitable constants that are to be chosen so that the ***remainder function*** also known as the Lagrange remainder term $R_{2,x_0}(x)$ is as small as possible at and around x_0. The remainder function we can express by $f(x)$ and the polynomial we are testing:

$$R_{2,x_0}(x) = f(x) - a_0 - a_1 \cdot (x - x_0) - a_2 \cdot (x - x_0)^2 \quad , \tag{4-12}$$

and it is this function that should be as close as possible to 0 when x is close to x_0 such that the difference between the function $f(x)$ and the second-degree polynomial becomes as small as possible – at least in the vicinity of x_0.

The first natural requirement is therefore that:

$$R_{2,x_0}(x_0) = 0 \quad \text{corresponding to} \quad f(x_0) = a_0 \quad , \tag{4-13}$$

by which a_0 is now determined.

The next natural requirement is that the graph of the remainder function has horizontal gradient at x_0 such that the tangent to the remainder function then is identical to the x axis:

$$R'_{2,x_0}(x_0) = 0 \quad \text{such that} \quad f'(x_0) = a_1 \quad , \tag{4-14}$$

by which a_1 is determined.

The next requirement on the remainder function is then:

$$R''_{2,x_0}(x_0) = 0 \quad \text{corresponding to} \quad f''(x_0) = 2 \cdot a_2 \quad , \tag{4-15}$$

by which also $a_2 = \frac{1}{2} f''(x_0)$ then is determined and fixed.

Thus we have found

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!} \cdot (x - x_0) + \frac{f''(x_0)}{2!} \cdot (x - x_0)^2 + R_{2,x_0}(x) \tag{4-16}$$

Where the remainder function $R_{2,x_0}(x)$ satisfies the following requirement that makes it very small in the neighborhood of x_0:

$$R_{2,x_0}(x_0) = R'_{2,x_0}(x_0) = R''_{2,x_0}(x_0) = 0 \quad . \tag{4-17}$$

If similarly we had wished to find an approximating n'th degree polynomial for the same function $f(x)$ we would have found:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!} \cdot (x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!} \cdot (x - x_0)^n + R_{n,x_0}(x) \quad , \tag{4-18}$$

where the remainder function $R_{n,x_0}(x)$ is a smooth function that satisfies all the requirements:

$$R_{n,x_0}(x_0) = R'_{n,x_0}(x_0) = \cdots = R^{(n)}_{n,x_0}(x_0) = 0 \quad . \tag{4-19}$$

At this point it is reasonable to expect, on one hand, that these requirements on the remainder functions can be satisfied; on the other, that the remainder function itself must 'appear like' and be as small as a power of $(x - x_0)$ close to x_0.

This is precisely the content of the following Lemma:

|||| Lemma 4.5 Remainder Functions

The remainder function $R_{n,x_0}(x)$ can be expressed from $f(x)$ in two different ways, and we will use both in what follows:

$$R_{n,x_0}(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot (x - x_0)^{n+1} \quad , \tag{4-20}$$

where $\xi(x)$ lies between x and x_0 in the interval I.

The other way is the following one, that contains an epsilon function:

$$R_{n,x_0}(x) = (x - x_0)^n \cdot \varepsilon_f(x - x_0) \quad , \tag{4-21}$$

where $\varepsilon_f(x - x_0)$ is an epsilon function of $(x - x_0)$.

|||| Proof

We will content ourselves by proving the first statement (4-20) in the simplest case, viz. for $n = 0$, i.e. the following : On the interval between (a fixed) x and x_0 we can always find a value ξ such that the following applies:

$$R_{0,x_0}(x) = f(x) - f(x_0) = \frac{f'(\xi)}{(1)!} \cdot (x - x_0) \quad . \tag{4-22}$$

But this is only a form of the ***mean value theorem***: If a smooth function has values $f(a)$ and $f(b)$, respectively, at the end points of an interval $[a, b]$, then the graph for $f(x)$ has at some position a tangent that is parallel to the line segment connecting the two points $(a, f(a))$ and $(b, f(b))$, see Figure 4.1.

The other statement (4-21) follows from the first (4-20) by observing that $\frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot (x - x_0)^{n+1}$ is an epsilon function of $(x - x_0)$, since $\frac{f^{(n+1)}(\xi)}{(n+1)!}$ is bounded and since $(x - x_0)^{n+1}$ is itself an epsilon function.

■

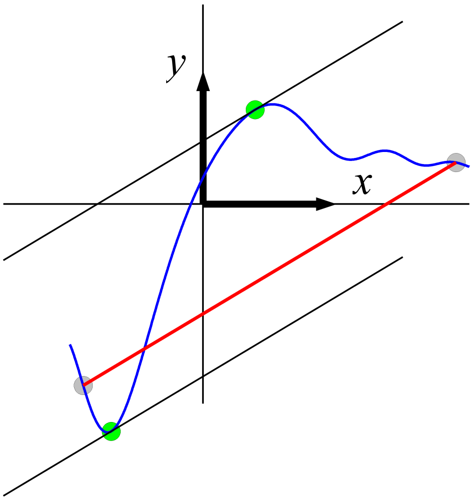


Figure 4.1: Two points on the blue graph curve for a function are connected with a line segment (red). The mean value theorem then says that at least one position exists (in the case shown, exactly two positions, marked in green) on the curve between the two given points where the slope $f'(x)$ for the tangent (black) to the curve is exactly the same as the slope of the straight line segment.

|||| Definition 4.6 Approximating Polynomials

Let $f(x)$ denote a smooth function on an interval I. The polynomial

$$P_{n,x_0}(x) = f(x_0) + \frac{f'(x_0)}{1!} \cdot (x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!} \cdot (x - x_0)^n \tag{4-23}$$

is called the ***approximating polynomial*** of nth degree for the function $f(x)$ with development point x_0.

To sum up:

|||| Theorem 4.7 Taylor's Formulas

Every smooth function $f(x)$ can for every non-negative integer n be divided into an approximating polynomial of degree n and a remainder function like this:

$$f(x) = P_{n,x_0}(x) + R_{n,x_0}(x) \quad , \tag{4-24}$$

where the remainder function can be expressed in the following ways:

$$\begin{aligned} R_{n,x_0}(x) &= \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot (x-x_0)^{n+1} \quad \text{for the } \xi(x) \text{ between } x \text{ and } x_0 \\ &\text{and} \\ R_{n,x_0}(x) &= (x-x_0)^n \cdot \varepsilon_f(x-x_0) \quad . \end{aligned} \tag{4-25}$$

In particular it is Taylor's Limit Formula (where the remainder function is expressed by an epsilon function) that we will make use of in what follows. We mention this version explicitly:

|||| Theorem 4.8 Taylor's Limit Formula

Let $f(x)$ denote a smooth function on an open interval I that contains a given x_0. Then for all x in the interval and for every integer $n \geq 0$ the following applies

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!} \cdot (x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!} \cdot (x-x_0)^n + (x-x_0)^n \cdot \varepsilon_f(x-x_0) \quad ,$$

where $\varepsilon_f(x-x_0)$ denotes an epsilon function of $(x-x_0)$, i.e. $\varepsilon_f(x-x_0) \to 0$ for $x \to x_0$.

|||| Example 4.9 The Approximating Polynomials of a Polynomial

One might be led to believe that every polynomial is its own approximating polynomial because every polynomial must be the best approximation to itself. Here is an example that shows that this is *not* that simple. We look at the third-degree polynomial

$$f(x) = 1 + x + x^2 + x^3 \quad . \tag{4-26}$$

The polynomial $f(x)$ has the following quite different approximating polynomials - dependent on the choice of *development point* x_0 and *degree of development* n:

$$\begin{aligned}
P_{7,x_0=0}(x) &= 1 + x + x^2 + x^3 \\
P_{3,x_0=0}(x) &= 1 + x + x^2 + x^3 \\
P_{2,x_0=0}(x) &= 1 + x + x^2 \\
P_{1,x_0=0}(x) &= 1 + x \\
P_{0,x_0=0}(x) &= 1 \\
P_{7,x_0=1}(x) &= 1 + x + x^2 + x^3 \\
P_{3,x_0=1}(x) &= 1 + x + x^2 + x^3 \\
P_{2,x_0=1}(x) &= 2 - 2 \cdot x + 4 \cdot x^2 \\
P_{1,x_0=1}(x) &= -2 + 6 \cdot x \\
P_{0,x_0=1}(x) &= 4 \\
P_{7,x_0=7}(x) &= 1 + x + x^2 + x^3 \\
P_{3,x_0=7}(x) &= 1 + x + x^2 + x^3 \\
P_{2,x_0=7}(x) &= 344 - 146 \cdot x + 22 \cdot x^2 \\
P_{1,x_0=7}(x) &= -734 + 162 \cdot x \\
P_{0,x_0=7}(x) &= 400 \quad .
\end{aligned} \tag{4-27}$$

|||| Exercise 4.10 Remainder Functions for Polynomials

For the function $f(x) = 1 + x + x^2 + x^3$ we consider the following two splittings into approximating polynomials and corresponding remainder functions:

$$\begin{aligned}
f(x) &= P_{2,x_0=1}(x) + R_{2,x_0=1}(x) \quad \text{and} \\
f(x) &= P_{1,x_0=7}(x) + R_{1,x_0=7}(x) \quad ,
\end{aligned} \tag{4-28}$$

where the two approximating polynomials $P_{2,x_0=1}(x)$ and $P_{1,x_0=7}(x)$ already are stated in example 4.9. Determine the two remainder functions $R_{2,x_0=1}(x)$ and $R_{1,x_0=7}(x)$ expressed in both of the two ways shown in 4-25: For each of the two remainder functions the respective expressions for $\xi(x)$ and for $\varepsilon(x - x_0)$ are stated.

|||| Example 4.11 Taylor's Limit Formula with the Development Point $x_0 = 0$

Here are some often-used functions with their respective approximating polynomials (and corresponding remainder functions expressed by epsilon functions) with the common development point $x_0 = 0$ and arbitrarily high degree:

$$\begin{aligned}
\mathrm{e}^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + x^n \cdot \varepsilon(x) \\
\mathrm{e}^{x^2} &= 1 + x^2 + \frac{x^4}{2!} + \cdots + \frac{x^{2n}}{n!} + x^{2n} \cdot \varepsilon(x) \\
\cos(x) &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \cdot \frac{x^{2n}}{(2n)!} + x^{2n} \cdot \varepsilon(x) \\
\sin(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \cdot \frac{x^{2n+1}}{(2n+1)!} + x^{2n+1} \cdot \varepsilon(x) \\
\ln(1+x) &= x - \frac{x^2}{2} + \cdots + (-1)^{n-1} \cdot \frac{x^n}{n!} + x^n \cdot \varepsilon(x) \\
\ln(1-x) &= -x - \frac{x^2}{2} - \cdots - \frac{x^n}{n!} - x^n \cdot \varepsilon(x) \\
\frac{1}{1+x} &= 1 - x + x^2 - x^3 + \cdots + (-1)^{n-1} \cdot x^{n-1} + x^n \cdot \varepsilon(x) \\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots + x^{n-1} + x^n \cdot \varepsilon(x)
\end{aligned} \tag{4-29}$$

Note that in Taylor's Limit formula we always end with an epsilon function and with the power of x that is precisely the same as the last power used in the preceding approximating polynomial.

4.3 Continuous Extensions

The function $f(x) = \sin(x)/x$ is not defined at $x = 0$. We will investigate whether we can extend the function to having a value at 0, such that the extended function is continuous at 0. I.e. we will find a value a such that the a-extension

$$\widetilde{f} = \begin{cases} \frac{\sin(x)}{x} & \text{for} \quad x \neq 0 \\ a & \text{for} \quad x = 0 \end{cases} \tag{4-30}$$

is continuous at $x = 0$, i.e. such that

$$\frac{\sin(x)}{x} \to a \quad \text{for} \quad x \to 0 \quad . \tag{4-31}$$

A direct application of Taylor's limit formula appears in the determination of limit values for those quotients $f(x)/g(x)$ where both the functions, i.e. the numerator $f(x)$ and the denominator $g(x)$, tend towards 0 for x tending towards 0. What happens to the quotient as x tends towards 0? We illustrate with a number of examples. Note that even though the numerator function and the denominator function both are continuous at 0, the quotient needs not be continuous.

|||| Example 4.12 Limit Values for Function Fractions

$$\frac{\sin(x)}{x} = \frac{x + x^1 \cdot \varepsilon(x)}{x} = 1 + \varepsilon(x) \to 1 \quad \text{for} \quad x \to 0 \quad . \tag{4-32}$$

$$\frac{\sin(x)}{x^2} = \frac{x - \frac{1}{3!}x^3 + x^3 \cdot \varepsilon(x)}{x^2} = \frac{1}{x} - \frac{x}{3!} + x \cdot \varepsilon(x) \quad , \tag{4-33}$$

that has no limit value for $x \to 0$. Therefore a continuous extension does not exist in this case.

$$\frac{\sin(x^2)}{x^2} \to 1 \quad \text{for} \quad x \to 0 \quad \text{because} \quad \frac{\sin(u)}{u} \to 1 \quad \text{for} \quad u \to 0 \quad . \tag{4-34}$$

$$\frac{\sin(x) - x}{x^2} = \frac{x - \frac{1}{3!}x^3 + x^3 \cdot \varepsilon(x) - x}{x^2} = -\frac{x}{3!} + x \cdot \varepsilon(x) \to 0 \quad \text{for} \quad x \to 0 \quad . \tag{4-35}$$

$$\frac{\sin(x) - x}{x^3} = \frac{x - \frac{1}{3!}x^3 + x^3 \cdot \varepsilon(x) - x}{x^3} = -\frac{1}{3!} + \cdot\varepsilon(x) \to -\frac{1}{6} \quad \text{for} \quad x \to 0 \quad . \tag{4-36}$$

By determination of such limit values the approximating polynomials in the numerator and the denominator are developed to such a high degree that limit value "appears" by dividing both the numerator and the denominator by a power of x.

Here is a somewhat more complicated example:

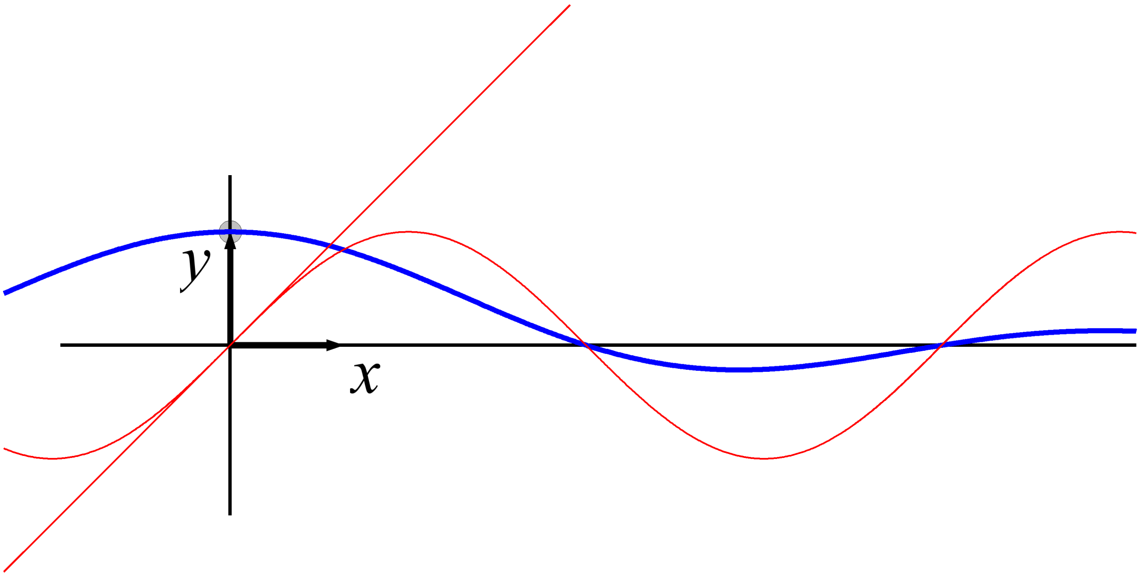


Figure 4.2: The function $f(x) = \sin(x)/x$ (blue) together with the numerator function $\sin(x)$ (red) and the denominator function x (also red). The function $f(x)$ is continuous at $x = 0$ exactly when we use the value $f(1) = 1$.

|||| Example 4.13 The Limit Value for a Fraction Between Functions

$$\begin{aligned}\frac{2\cos(x) - 2 + x^2}{x\cdot\sin(x) - x^2} &= \frac{2\cdot(1 - \frac{1}{2!}\cdot x^2 + \frac{1}{4!}\cdot x^4 + x^4\cdot\varepsilon_1(x)) - 2 + x^2}{x\cdot(x - \frac{1}{3!}\cdot x^3 + \frac{1}{5!}\cdot x^5 + x^5\cdot\varepsilon_2(x)) - x^2} \\ &= \frac{2 - x^2 + \frac{1}{12}\cdot x^4 + 2\cdot x^4\cdot\varepsilon_1(x) - 2 + x^2}{x^2 - \frac{1}{3!}\cdot x^4 + \frac{1}{5!}\cdot x^6 + x^6\cdot\varepsilon_2(x) - x^2} \\ &= \frac{\frac{1}{12}\cdot x^4 + 2\cdot x^4\cdot\varepsilon_1(x)}{-\frac{1}{3!}\cdot x^4 + \frac{1}{5!}\cdot x^6 + x^6\cdot\varepsilon_2(x)} \\ &= \frac{\frac{1}{12} + 2\cdot\varepsilon_1(x)}{-\frac{1}{6} + \frac{1}{5!}\cdot x^2 + x^2\cdot\varepsilon_2(x)} \\ &\to -\frac{1}{2} \quad \text{for} \quad x \to 0 \quad ,\end{aligned} \tag{4-37}$$

since the numerator tends towards $\frac{1}{12}$ for $x \to 0$ and the denominator tends towards $-\frac{1}{6}$ for $x \to 0$.

4.4 Estimation of the Remainder Functions

How large is the error committed by using the approximating polynomial (which it is easy to compute) instead of the function itself (that can be difficult to compute) on a given (typically small) interval around the development point? The remainder function can of course give the answer to this question. We give here a couple of examples that show how the remainder function can be used for such error estimations for given functions.

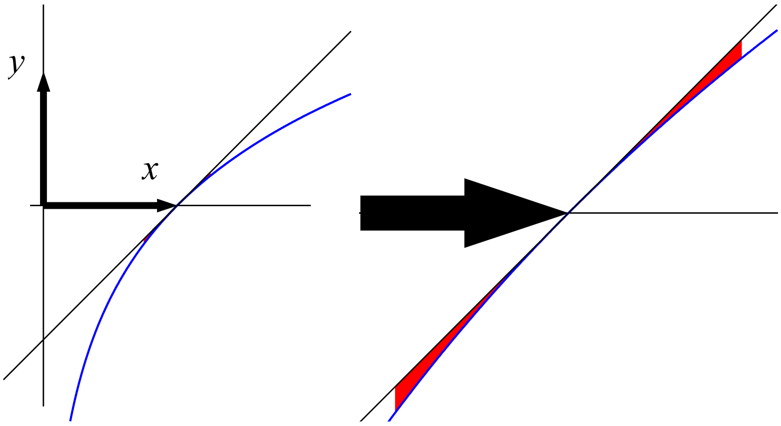


Figure 4.3: The function $f(x) = \ln(x)$ from Example 4.14 (blue), the approximating first-degree polynomial (black) with development point $x_0 = 1$ and the corresponding remainder function (red) illustrated as the difference between $f(x)$ and the approximating polynomial on the interval $\left[\frac{3}{4}, \frac{5}{4}\right]$. To the right is shown the figure around the point $(1, 0)$ close-up.

|||| Example 4.14 Approximation of an Elementary Function

The logarithmic function $\ln(x)$ is defined for positive values of x. We approximate with the approximating first-degree polynomial with the development point at $x_0 = 1$ and will estimate the remainder term on a suitably small interval around $x_0 = 1$, i.e. the starting point is the following:

$$f(x) = \ln(x) \quad , \quad x_0 = 1 \quad , \quad n = 1 \quad , \quad x \in \left[\frac{3}{4}, \frac{5}{4}\right] \quad . \tag{4-38}$$

According to Taylor's formula with the remainder function we have - using the development point $x_0 = 1$ where $f(1) = 0$ and $f'(1) = 1$ and using $f''(x) = -1/x^2$ for all x in the domain:

$$f(x) = \ln(x) = \ln(1) + \frac{f'(1)}{1!}(x-1) + \frac{f''(\xi)}{2!} \cdot (x-1)^2 = x - 1 - \frac{1}{2\cdot\xi^2} \cdot (x-1)^2 \tag{4-39}$$

for a value of ξ between x and 1. Thus we have found:

$$P_{1,x_0=1}(x) = x - 1 \quad , \quad \text{and} \quad R_{1,x_0=1}(x) = -\frac{1}{2\cdot\xi^2} \cdot (x-1)^2 \quad . \tag{4-40}$$

The *absolute value of the remainder function* on the given interval can now be evaluated for all x in the given interval - even if we do not know very much about the position of ξ in the interval apart from the fact that ξ lies between x and 1:

We have

$$|R_{1,x_0=1}(x)| = |-\frac{1}{2\cdot\xi^2} \cdot (x-1)^2| \leq |\frac{1}{2\cdot\xi^2} \cdot \left(\frac{1}{4}\right)^2| \quad . \tag{4-41}$$

Here the minus sign has been removed because we only look at the absolute value and we have also used that $(x-1)^2$ clearly is largest (with the value $(1/4)^2$) for $x = 3/4$ and for $x = 5/4$ in the interval. In addition ξ is *smallest* and thus $(1/\xi)^2$ *largest* on the interval for $\xi = 3/4$. (Note that here we do not use the fact of ξ lying between x and 1 - we simply use the fact of ξ lying in the interval!) I.e.

$$|R_{1,x_0=1}(x)| \leq |\frac{1}{32\cdot\xi^2}| \leq |\frac{1}{32\cdot\left(\frac{3}{4}\right)^2}| = \frac{1}{18} \quad , \tag{4-42}$$

thus we have proved that

$$|\ln(x) - (x-1)| \leq \frac{1}{18} \quad \text{for all} \quad x \in \left[\frac{3}{4}, \frac{5}{4}\right] \quad . \tag{4-43}$$

One may well wonder why the remainder function estimation of such a simple function as $f(x) = \ln(x)$ in Example 4.14 should be so complicated, when it is evident to everybody (!) that the red remainder function in that case assumes its largest numerical (absolute) value at one of the end points of the actual interval, see Figure 4.3 – a statement, moreover, which we can prove by a quite ordinary function investigation.

By differentiation of the remainder function we get:

$$R'_{1,x_0=1}(x) = \frac{d}{dx}\left(\ln(x) - (x-1)\right) = \frac{1}{x} - 1 \quad , \tag{4-44}$$

that is less than 0 precisely for $x > 1$ (such that $R_{1,x_0=1}(x)$ to the right of $x = 1$ is negative and decreasing from the value 0 at $x = 1$) and greater than 0 for $x < 1$ (such that $R_{1,x_0=1}(x)$ to the left of $x = 1$ is negative and increasing towards the value 0 at $x = 1$). But the problem is that we *in principle do not know* what the value of $\ln(x)$ in fact is – neither at $x = 3/4$ nor at $x = 5/4$ unless we use Maple or some other tool as help. The remainder function estimate uses *only* the defined properties of $f(x) = \ln(x)$, i.e. $f'(x) = 1/x$ and $f(1) = 0$ and the estimation gives the values (also at the end points of the interval) with a (numerical) error of at most $1/18$ in this case.

If we actually get the information that $\ln(3/4) = -0.2877$ and $\ln(5/4) = 0.2223$ we then of course get a direct estimate of the largest value of $|R_{1,x_0=1}(x)|$ in the interval $\left[\frac{3}{4}, \frac{5}{4}\right]$:

$$\begin{aligned} |R_{1,x_0=1}(x)| &\leq \max\{|-0.2877 + 0.25|\,,\,|0.2223 - 0.25|\} \\ &= 0.0377 < 1/18 = 0.0556 \quad . \end{aligned} \tag{4-45}$$

With the ordinary function analysis we get a somewhat better estimate of the remainder function – but only because we beforehand can estimate the function value at the end points.

|||| Exercise 4.15 Approximation of a Non-Elementary Function

Given the function from Example 4.3:

$$f(x) = \int_0^x e^{-t^2}\,dt \tag{4-46}$$

An estimate of the magnitude of the difference between $f(x)$ and the approximating first-degree polynomial $P_{1,x_0=0}(x)$ with the development point at $x_0 = 0$ is wished for. The exercise is about determining the largest absolute value that the remainder function $|R_{1,x_0=0}(x)|$

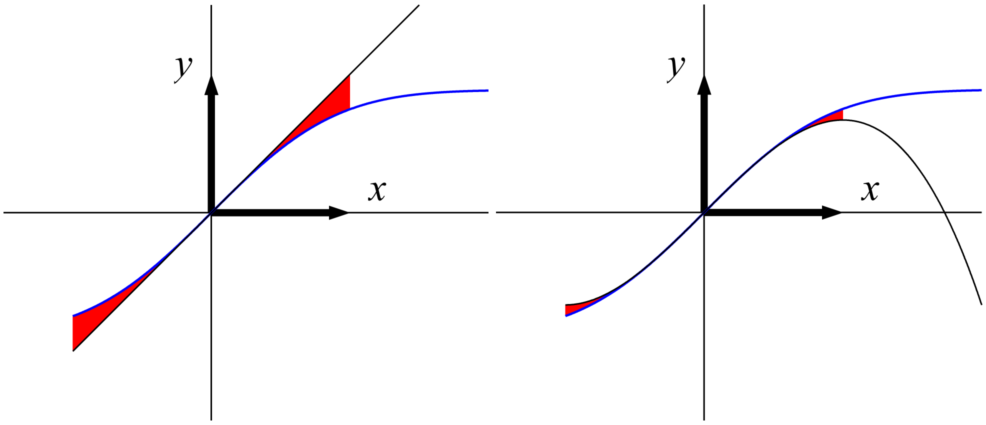


Figure 4.4: The function $f(x)$ from Example 4.15 (blue), the approximating first- and third-degree polynomials (black) with development points $x_0 = 0$ and the corresponding remainder functions (red) on the interval $[-1, 1]$.

assumes on the interval $[-1, 1]$.
Hint: Use higher order derivatives of $f(x)$ evaluated at $x_0 = 0$ found earlier, cf. Example 4.3: $f(0) = 0$, $f'(0) = 1$, $f''(x) = -2 \cdot x \cdot \mathrm{e}^{-x^2}$. See Figure 4.4.

|||| Example 4.16 Approximation of an Unknown (But Elementary) Function

Given the function from example 4.4, i.e. the function satisfies the following differential equation with initial conditions:

$$f''(x) + 3f'(x) + 7f(x) = x^2 \quad , \quad \text{where} \quad f(0) = 1 \quad , \quad \text{and} \quad f'(0) = -3 \quad , \tag{4-47}$$

where we have assumed that the right-hand side of the equation is $q(x) = x^2$ and that the development point is $x_0 = 0$. By this we now get:

$$f'(0) = -3 \quad , \quad f''(0) = 2 \quad , \quad f'''(0) = 15 \quad . \tag{4-48}$$

We have

$$\begin{aligned} f(x) &= f(0) + f'(0) \cdot x + \frac{f''(0)}{2} \cdot x^2 + \frac{f'''(0)}{6} \cdot x^3 + x^3 \cdot \varepsilon(x) \\ &= 1 - 3 \cdot x + x^2 + \frac{5}{2} \cdot x^3 + x^3 \cdot \varepsilon(x) \quad , \end{aligned} \tag{4-49}$$

such that the approximating third-degree polynomial for $f(x)$ with development point $x_0 = 0$ is

$$P_{3,x_0=0}(x) = 1 - 3 \cdot x + x^2 + \frac{5}{2} \cdot x^3 \quad . \tag{4-50}$$

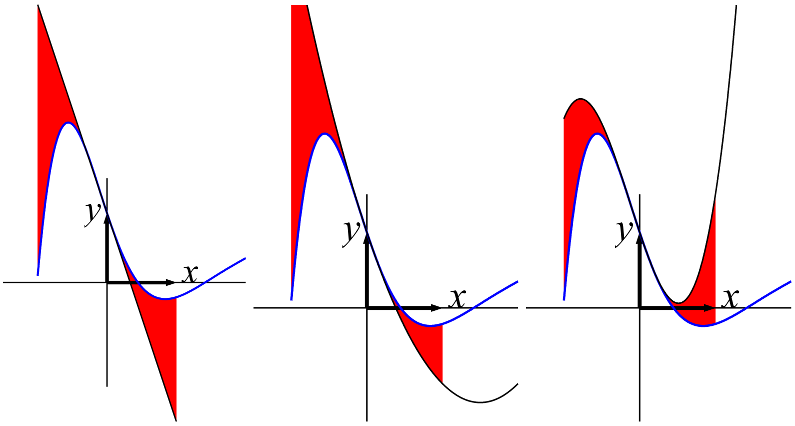


Figure 4.5: The function $f(x)$ from Example 4.16 (blue), the approximating first-, second-, and third-degree polynomials (black) with the development point $x_0 = 0$. The corresponding respective remainder functions (red) are illustrated as the differences between $f(x)$ and the approximating polynomials.

> Note that $P_{3,x_0=0}(x)$ satisfies the initial conditions in (4-47) but the polynomial $P_{3,x_0=0}(x)$ is not a solution to the differential equation itself!

4.5 Functional Investigations

A very important property of continuous functions is the following, which means one can control how large and how small values a continuous function can assume on an interval, as long as the interval is sufficiently nice:

|||| Theorem 4.17 Main Theorem for Continuous Functions of One Variable

Let $f(x)$ denote a function that is continuous on all of its domain $D(f) \subset \mathbb{R}$. Let $I = [\,a\,,\,b\,]$ be a bounded, closed, and connected interval in $D(f)$.

Then the range for the function $f(x)$ on the interval I is also a bounded, closed and connected interval $[\,A\,,\,B\,] \subset \mathbb{R}$, thus denoted:

$$R(f_{|_I}) = f(I) = \{f(x)\,|\,x \in I\} = [\,A\,,\,B\,] \quad , \tag{4-51}$$

where the possibility that $A = B$ is allowed and this happens precisely when $f(x)$ is constant on the whole interval I.

|||| Definition 4.18 Global Minimum and Global Maximum

When a function $f(x)$ has the range $R(f_{|_I}) = f(I) = [\,A\,,\,B\,]$ on an interval $I = [a,b]$ we say that

1. A is the ***global minimum value*** for $f(x)$ on I, and if $f(x_0) = A$ for $x_0 \in I$ then x_0 is a ***global minimum point*** for $f(x)$ on I.

2. B is the ***global maximum value*** for $f(x)$ on I, and if $f(x_0) = B$ for $x_0 \in I$ then x_0 is a ***global maximum point*** for $f(x)$ on I.

A well-known and important task is to find the global maximum and minimum values for given functions $f(x)$ on given intervals and to determine the x-values for which these maximum and minimum values are *assumed*, that is, the minimum and maximum points. To solve this task the following is an invaluable help – see Figure 4.6:

|||| Lemma 4.19 Maxima and Minima at Stationary Points

Let x_0 be a global maximum or minimum value for $f(x)$ on I. Assume that x_0 is not an end point for the interval I and that $f(x)$ is differentiable at x_0.
Then x_0 is a ***stationary point*** for $f(x)$,i.e. $f'(x_0) = 0$.

|||| Proof

We outline the argument. Since $f(x)$ is assumed differentiable, we have:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)\cdot(x-x_0) + (x-x_0)\cdot\varepsilon_f(x-x_0) \\ &= f(x_0) + (x-x_0)\cdot(f'(x_0)+\varepsilon_f(x-x_0)) \quad . \end{aligned} \tag{4-52}$$

Now if we assume that $f'(x_0)$ is positive then the parenthesis $(f'(x_0)+\varepsilon_f(x-x_0))$ is also positive for x sufficiently close to x_0 (since $\varepsilon_f(x-x_0) \to 0$ for $x \to x_0$), but then $(x-x_0)\cdot(f'(x_0)+\varepsilon_f(x-x_0))$ is also positive for x sufficiently close to x_0 and then $f(x) > f(x_0)$ for $x > x_0$, and $f(x) < f(x_0)$ for $x < x_0$. Therefore $f(x_0)$ can not be neither a maximum value nor a minimum value for $f(x)$. A similar conclusion appears when the assumption is $f'(x_0) < 0$. If x_0 is a global maximum or minimum value for $f(x)$ on I this assumption must imply that $f'(x_0) = 0$.

■

Hereby we have the following investigation method at our disposal:

|||| Method 4.20 Method of Investigation

Let $f(x)$ be a continuous function and $I = [a,b]$ an interval in the domain $D(f)$.

Maximum and minimum values for the function $f(x)$, $x \in I$, i.e. A and B in the range $[A,B]$ for $f(x)$ restricted to I, are found by finding and comparing the function values at the following points:

1. Interval end points (the boundary points a and b for the interval I).
2. Exception points, i.e. the points in the open interval $]a,b[$ where the function is *not* differentiable.
3. The stationary points, i.e. all the points x_0 in the open interval $]a,b[$ where $f'(x_0) = 0$.

With this method of investigation we not only find the global maximum and minimum values but also the x-values in I for which the global maximum and the global minimum are assumed i.e. maximum and minimum points in the actual interval.

|||| Example 4.21 A Continuous Function Is Investigated

A Continuous function $f(x)$ is defined for all x in the following way:

$$f(x) = \begin{cases} 0.75 & \text{for} \quad x \leq -1.5 \\ 0.5 + (x+1)^2 & \text{for} \quad -1.5 \leq x \leq 0 \\ 1.5 \cdot (1 - x^3) & \text{for} \quad 0 \leq x \leq 1 \\ x - 1 & \text{for} \quad 1 \leq x \leq 2 \\ 1 & \text{for} \quad x > 2 \end{cases} \tag{4-53}$$

See Figure 4.6, where we only consider the function on the interval $I = [-1.5, 2.0]$. There are two exception points where the function is not differentiable: $x_0 = 0$ and $x_0 = 1$. There is one stationary point in $]-1.5, 2.0[$ where $f'(x_0) = 0$ viz. $x_0 = -1$. And finally there are two boundary points (the interval end points $x_0 = -1.5$ and $x_0 = 2$) that need to be investigated.

Therefore we have the following candidates for global maximum and minimum values for f on I:

$x_0 =$	-1.5	-1	0	1	2
$f(x_0) =$	0.75	0.5	1.5	0	1

(4-54)

In conclusion we read from this that the maximum value for $f(x)$ is $B = 1.5$ which is assumed at the maximum point $x_0 = 0$. The minimum value is $A = 0$, assumed at the minimum point $x_0 = 1$. There are no other maximum or minimum points for f on I.

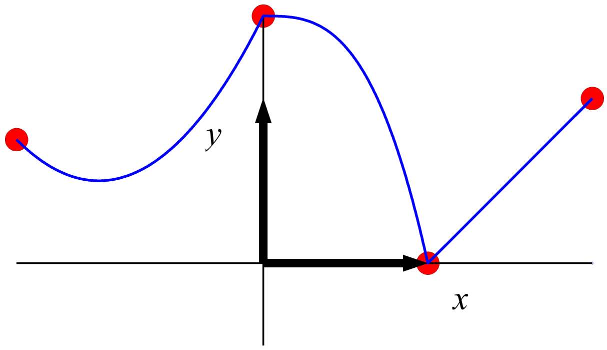


Figure 4.6: The continuous function $f(x)$ from example 4.21 (blue). On the graph we have marked (in red) the 5 points that need to be investigated particularly in order to determine the range for f in the interval $[-1.5, 2]$, cf. Method 4.20.

|||| Definition 4.22 Local Minima and Local Maxima

Let $f(x)$ denote a function on an interval $I = [a, b]$ containing a given $x_0 \in]a, b[$.

1. If $f(x) \geq f(x_0)$ for all x in a (as small as you like) neighborhood of x_0 then $f(x_0)$ is called a ***local minimum value*** for $f(x)$ in I and x_0 is a ***local minimum point*** for $f(x)$ in I. If actually $f(x) > f(x_0)$ for all x in the neighborhood apart from the point x_0 itself then $f(x_0)$ is called a ***proper local minimum value***.

2. If $f(x) \leq f(x_0)$ for all x in a (as small as you like) neighborhood of x_0 then $f(x_0)$ is called a ***local maximum value*** for $f(x)$ in I and x_0 is a ***local maximum point*** for f on I. If actually $f(x) < f(x_0)$ for all x in the neighborhood apart from the point x_0 itself then $f(x_0)$ is called a ***proper local maximum value***.

If the function we want to investigate is smooth at its stationary points then we can qualify the Method 4.20 even better, since the approximating polynomial of degree 2 with development point at the stationary point can help in the decision whether the value of $f(x)$ at the stationary point is a candidate to be a maximum value or a minimum value.

|||| Lemma 4.23 Local Analysis at a Stationary Point

Let $f(x)$ be a smooth function and assume that x_0 is a stationary point for $f(x)$ on an interval $I =]a, b[$. Then the following applies:

1. If $f''(x_0) > 0$ then $f(x_0)$ is a proper local minimum value for $f(x)$.
2. If $f''(x_0) < 0$ then $f(x_0)$ is a proper local maximum value for $f(x)$.
3. If $f''(x_0) = 0$ then this is not sufficient information to decide whether $f(x_0)$ is a local minimum value or a local maximum value or neither.

|||| Exercise 4.24

Prove Lemma 4.23 by using Taylor's limit formula with the approximating second-degree polynomial for $f(x)$ and with the development point x_0. Remember that x_0 is a stationary point, such that $f'(x_0) = 0$.

|||| Example 4.25 Local Maxima and Minima

The continuous function $f(x)$

$$f(x) = \begin{cases} 0.75 & \text{for} \quad x \leq -1.5 \\ 0.5 + (x+1)^2 & \text{for} \quad -1.5 \leq x \leq 0 \\ 1.5 \cdot (1 - x^3) & \text{for} \quad 0 \leq x \leq 1 \\ x - 1 & \text{for} \quad 1 \leq x \leq 2 \\ 1 & \text{for} \quad x \geq 2 \end{cases} \tag{4-55}$$

is shown in Figure 4.6. On the interval $I = [-1.5, 2.0]$ the function has the proper local minimum values 0.5 and 0 in the respective proper local minimum points $x_0 = -1$ and $x_0 = 1$ and the function has a proper local maximum value 1.5 at the proper local maximum point $x_0 = 0$. If we extend the interval to $J = [-7, 7]$ and note that the function values by definition are constant outside the interval I we get the new local maximum values 0.75 and 1 for f on J – not one of them is a *proper* local maximum value. All $x_0 \in]-7, -1.5]$ and all $x_0 \in [2, 7[$ are local maximum points for f on J but not one of them is a *proper* local maximum point. All x_0 in the *open interval* $x_0 \in]-7, -1.5[$ and all x_0 in the *open interval* $x_0 \in]2, 7[$ in addition also local minimum points for $f(x)$ i J but not one of them is a *proper* local minimum point.

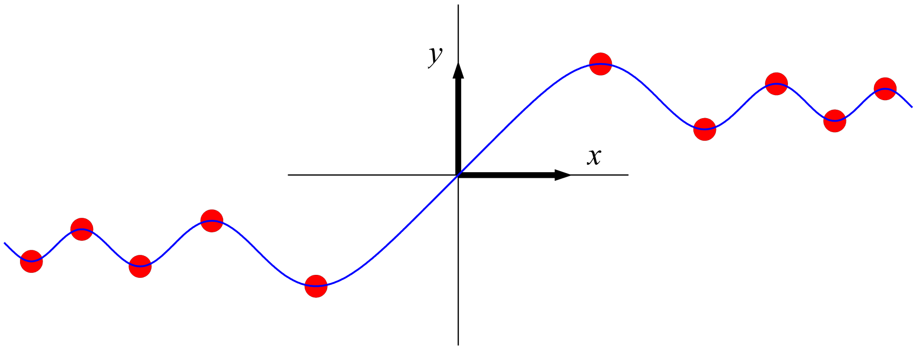


Figure 4.7: Proper local maxima and proper local minima for the function from Example 4.26 are here indicated on the graph for the function. We note: The local maximum and minimum *points* for the function are the x-coordinates of the graph points shown in red, and the local maximum and minimum *values* for the function are the y-coordinates of the graph-points shown in red.

|||| Example 4.26 A Non-Elementary Function

The function $f(x)$

$$f(x) = \int_0^x \cos(t^2)\, dt \tag{4-56}$$

has stationary points at those values of x_0 satisfying:

$$f'(x_0) = \cos(x_0^2) = 0 \quad , \quad \text{dvs.} \quad x_0^2 = \frac{\pi}{2} + p \cdot \pi \quad \text{where } p \text{ is an integer} \quad . \tag{4-57}$$

Since we also have that

$$f''(x) = -2 \cdot x \cdot \sin(x^2) \quad , \tag{4-58}$$

such that at the stated stationary points it applies

$$f''(x_0) = -2 \cdot x_0 \cdot (-1)^p \quad . \tag{4-59}$$

From this it follows – via Lemma 4.23 – that every other stationary point x_0 along the x-axis is a proper local maximum point for $f(x)$ and the other points proper local minimum points. See Figure 4.7. In Figure 4.8 are shown graphs (parabolas) for a pair of the approximating second-degree polynomials for $f(x)$ with the development points at chosen stationary points.

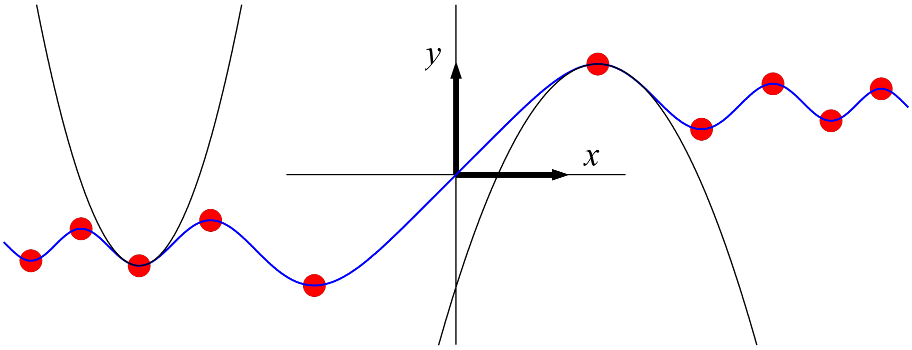


Figure 4.8: The graph for the function in Example 4.26 and two approximating parabolas with development points in two stationary points, which are a proper local minimum point and a proper local maximum point for $f(x)$.

|||| Example 4.27 When the Approximation to Degree 2 is Not Good Enough

As stated in Lemma 4.23 one cannot from $f'(x_0) = f''(x_0) = 0$ decide whether the function has a local maximum or minimum at x_0. This is shown in the three simple functions in Figure 4.9 with all the clarity one could wish for: $f_1(x) = x^4$, $f_2(x) = -x^4$ and $f_3(x) = x^3$. All three functions have a stationary point at $x_0 = 0$ and all have $f''(x_0) = 0$, but $f_1(x)$ has a proper local minimum point at 0, $f_2(x)$ has a proper local maximum point at 0, and $f_3(x)$ has neither a local minimum point nor a local maximum point at 0.

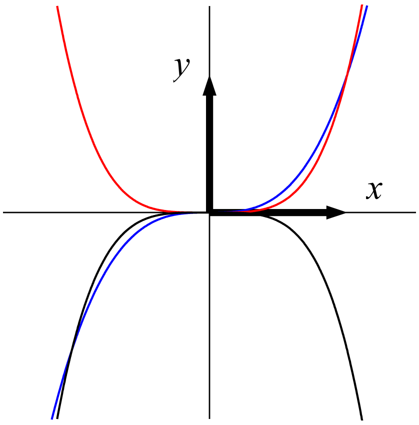


Figure 4.9: Three elementary functions with approximating second-degree polynomials $P_{2,x_0=0}(x) = 0$ for all x. The functions are: $f_1(x) = x^4$ (red), $f_2(x) = -x^4$ (black) and $f_3(x) = x^3$ (blue).

4.6 Summary

In this eNote we have studied how one can approximate smooth functions using polynomials.

- Every smooth function $f(x)$ on an interval I can be split into an approximating n'th-degree polynomial $P_{n,x_0}(x)$ with the development point x_0 and a corresponding remainder function $R_{n,x_0}(x)$ like this:

$$f(x) = P_{n,x_0}(x) + R_{n,x_0}(x) \quad , \tag{4-60}$$

where the polynomial and the remainder function in Taylor's limit formula are written like this:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}\cdot(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}\cdot(x-x_0)^n + (x-x_0)^n\cdot\varepsilon_f(x-x_0) \quad ,$$

with $\varepsilon_f(x-x_0)$ denoting an epsilon function of $(x-x_0)$, i.e. $\varepsilon_f(x-x_0)\to 0$ for $x\to x_0$.

- Taylor's limit formula can be used to find the continuous extension of quotients of functions by finding (if possible) their limit values for $x\to x_0$ where x_0 are the values where the numerator function is 0 such that the quotient at the starting point is not defined at x_0:

$$\frac{\sin(x)}{x} = \frac{x + x^1\cdot\varepsilon(x)}{x} = 1+\varepsilon(x)\to 1 \quad \text{for} \quad x\to 0 \quad . \tag{4-61}$$

- Estimation of the remainder function gives an upper bound for the largest numerical difference between a given function and the approximating polynomial of a suitable degree and with a suitable development point on a given interval of investigation. Such an estimation can also be made for functions that are possibly only "known" via a differential equation or as a non-elementary integral:

$$|\ln(x) - (x-1)| \leq \frac{1}{18} \quad \text{for all} \quad x\in\left[\frac{3}{4},\frac{5}{4}\right] \quad . \tag{4-62}$$

- Taylor's limit formula with approximating second-degree polynomials is used for efficient functional investigation, including determination of range, global and local maxima and minima for given functions.

eNote 5

The Number Spaces $\mathbb{R}^n$ and $\mathbb{C}^n$

This eNote is about the real number space $\mathbb{R}^n$ and the complex number space $\mathbb{C}^n$, which are essential building blocks in Linear Algebra.

Update: 23.09.21 David Brander

5.1 Number Spaces

Remark 5.1 The Common Notion $\mathbb{L}$

Defnitions and rules in this eNote are valid both for the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$. The set of real numbers and the set of complex numbers are examples of *fields*. Fields have common calculation rules concerning elementary arithmetic rules (the same rules as those for $\mathbb{C}$ described in Theorem 1.12 in eNote 1). In the following when we use the symbol $\mathbb{L}$ it means that the notion is valid both for the set of real numbers and for the set of complex numbers.

$\mathbb{R}^n$ is the symbol for the set of all n-tuples that contain n real elements. For example,

$$(1,4,5) \text{ and } (1,5,4)$$

are two different 3-tuples that belong to $\mathbb{R}^3$. Similarly $\mathbb{C}^n$ is the symbol for the set of all n-tuples which contains n complex elements, e.g.

$$(1+2i,0,3i,1,1) \text{ and } (1,2,3,4,5)$$

are two different 5-tuples that belong to $\mathbb{C}^5$. Formally we write $\mathbb{L}^n$ in set notation as:

$$\mathbb{L}^n = \{(a_1, a_2, ..., a_n) \,|\, a_i \in \mathbb{L}\}\,. \tag{5-1}$$

We introduce addition of elements in $\mathbb{L}^n$ and multiplication of elements in $\mathbb{L}^n$ by an element of $\mathbb{L}$ (a scalar) by the following definition:

|||| Definition 5.2

Let $(a_1, a_2, ..., a_n)$ and $(b_1, b_2, ..., b_n)$ be two elements of $\mathbb{L}^n$ and let k be a number in $\mathbb{L}$ (a scalar). *The sum* of the two n-tuples is defined by

$$(a_1, a_2, ..., a_n) + (b_1, b_2, ..., b_n) = (a_1 + b_1, a_2 + b_2, ..., a_n + b_n), \tag{5-2}$$

and *the product* of $(a_1, a_2, ..., a_n)$ by k by

$$k \cdot (a_1, a_2, ..., a_n) = (a_1, a_2, ..., a_n) \cdot k = (k \cdot a_1, k \cdot a_2, ..., k \cdot a_n\,). \tag{5-3}$$

$\mathbb{R}^n$ with the operations (5-2) and (5-3) is called the n-dimensional *real number space*. Similarly, $\mathbb{C}^n$ with the operations (5-2) and (5-3), the n-dimensional *complex number space*.

|||| Example 5.3 Addition

An example of the addition of two 4-tuples in $\mathbb{R}^4$ is

$$(1, 2, 3, 4) + (2, 1, -2, -5) = (3, 3, 1, -1)$$

.

|||| Example 5.4 Multiplication

Denitio An example of multiplication of a 3-tuple in $\mathbb{R}^3$ by a scalar is

$$5 \cdot (2, 4, 5) = (10, 20, 25)\,.$$

An example of multiplication of a 2-tuple in $\mathbb{C}^2$ by a scalar is

$$i \cdot (2 + i, 4) = (-1 + 2i, 4\,i)\,.$$

As a short notation for n-tuples we often use small **bold** letters, we write e.g.

$$\mathbf{a} = (3,2,1) \ \text{ or } \ \mathbf{b} = (b_1, b_2, ..., b_n)\,.$$

For the n-tuple $(0,0,...,0)$, which is called *the zero element* of $\mathbb{L}^n$, we use the notion

$$\mathbf{0} = (0,0,...,0)\,.$$

When more complicated computational exercises in the number spaces are called for, there is a need for the following arithmetic rules.

|||| Theorem 5.5 Arithmetic Rules in $\mathbb{L}^n$

For all values of n, in the number space $\mathbb{L}^n$ the operations introduced in definition 5.2 obey the following eight rules:

1. $\mathbf{a}+\mathbf{b} = \mathbf{b}+\mathbf{a}$ (addition is commutative)
2. $(\mathbf{a}+\mathbf{b})+\mathbf{c} = \mathbf{a}+(\mathbf{b}+\mathbf{c})$ (addition is associative)
3. For all **a**: $\mathbf{a}+\mathbf{0} = \mathbf{a}$ (i.e. **0** is neutral with respect to addition)
4. For all **a** there exists an *opposite element* $-\mathbf{a}$ such that $\mathbf{a}+(-\mathbf{a}) = \mathbf{0}$
5. $k_1(k_2\mathbf{a}) = (k_1k_2)\mathbf{a}$ (multiplication by scalars is associative)
6. $(k_1+k_2)\mathbf{a} = k_1\mathbf{a}+k_2\mathbf{a}$ (distributive rule)
7. $k_1(\mathbf{a}+\mathbf{b}) = k_1\mathbf{a}+k_1\mathbf{b}$ (distributive rule)
8. $1\mathbf{a} = \mathbf{a}$ (the number 1 is neutral in a product with a scalar)

|||| Proof

Concerning rule 4: Given two vectors $\mathbf{a} = (a_1, \ldots a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$. Then

$$\mathbf{a}+\mathbf{b} = (a_1+b_1\,, ..., \,a_n+b_n) = \mathbf{0} \Leftrightarrow b_1 = -a_1\,, ..., \,b_n = -a_n\,.$$

From this we deduce that **a** has an opposite vector $-\mathbf{a}$ given by $-\mathbf{a} = (-a_1\,, ..., -a_n)$. Moreover, this vector is unique.

The other rules are proved by calculating the left and right hand side of the equations and then comparing the two results.

■

From the proof of rule 4 in theorem 5.5 it is evident that for an arbitrary n-tuple $\mathbf{a}$: $-\mathbf{a} = (-1)\mathbf{a}$.

|||| Exercise 5.6

Give a formal proof of rule 2 and rule 5 in Theorem 5.5.

|||| Definition 5.7 Subtraction

Given $\mathbf{a} \in \mathbb{L}^n$ and $\mathbf{b} \in \mathbb{L}^n$. The difference $\mathbf{a} - \mathbf{b}$ is defined as:

$$\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})\,. \tag{5-4}$$

|||| Example 5.8 Subtraction

$(1+2i,1) - (i,2) = (1+2i,1) + (-(i,2)) = (1+2i,1) + (-i,-2) = (1+i,-1)\,.$

|||| Exercise 5.9 The Zero Rule

Show that the following variant of the *zero rule* is valid:

$$k\mathbf{a} = \mathbf{0} \Leftrightarrow k = 0 \text{ or } \mathbf{a} = \mathbf{0}\,. \tag{5-5}$$

|||| Remark 5.10 n-Tuples as Vectors

Often an n-tuple is written as a ***column vector***. We have two equivalent ways of writing, here with an example from $\mathbb{R}^4$:

$$\mathbf{v} = (1,2,3,4) \text{ and } \mathbf{v} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

If in a given context the n-tuple is regarded as a ***row vector*** then a transposition is performed. The transpose of a column vector is a row vector (and vice versa), it has the symbol T:

$$\mathbf{v}^\top = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}.$$

eNote 6

Systems of Linear Equations

(Updated 24.9.2021 David Brander)

6.1 Linear Equations

Remark 6.1 The Common Notion $\mathbb{L}$

Defnitions and rules in this eNote are valid both for the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$. The set of real numbers and the set of complex numbers are examples of *fields*. Fields have common calculation rules concerning elementary arithmetic rules (the same rules as those for $\mathbb{C}$ described in Theorem 1.12 in eNote 1). In the following when we use the symbol $\mathbb{L}$ it means that the notion is valid both for the set of real numbers and for the set of complex numbers.

A ***linear equation*** with n unknowns $x_1, x_2, \ldots x_n$ is an equation of the form

$$a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n = b\,. \tag{6-1}$$

The numbers $a_1, a_2, \ldots, a_n$ are called the *coefficients* and the number b is, in this context, called *the right hand side*. The coefficients and the right hand side are considered known in contrast to the unknowns. The equation is called *homogeneous* if $b = 0$, else *inhomogeneous*.

|||| Definition 6.2 Solution to a Linear Equation

By a *solution* to the equation

$$a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n = b\,. \tag{6-2}$$

we shall understand an n-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{L}^n$ that by substitution into the equation makes the left hand side of the equation equal to the right hand side.

By the ***general solution*** or just the ***solution set*** we understand the set of all solutions to the equation.

|||| Example 6.3 The Equation for a Straight Line in the Plane

An example of a linear equation is the equation for a straight line in the (x, y)-plane:

$$y = 2\,x + 5\,. \tag{6-3}$$

Here y is isolated on the left hand side and the coefficients 2 and 5 have well known geometrical interpretations. But the equation could also be written

$$-2\,x_1 + 1\,x_2 = 5 \tag{6-4}$$

where x and y are substituted by the more general names for unknowns, x_1 and x_2, and the equation is of the form (6-1).

The solution set for the equation (6-3) is of course the coordinate set for all points on the line - by substitution they will satisfy the equation in contrast to all other points!

|||| Example 6.4 Trivial and Inconsistent Equations

The linear equation

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 0 \;\;\Leftrightarrow\;\; 0 = 0 \tag{6-5}$$

where all coefficients and the right hand side are 0, is an example of a ***trivial*** equation. The solution set of the equation consists of all $\mathbf{x} = (x_1, x_2, x_3, x_4) \in \mathbb{L}^4$.

If all the coefficients of the equation are 0 but the right hand side is non-zero, the equation is an ***inconsistent*** equation, that is, an equation without a solution. An example is the equation

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 1 \;\;\Leftrightarrow\;\; 0 = 1\,. \tag{6-6}$$

When you investigate linear equations, you can use the usual ***rule of conversion*** for equations: The set of solutions for the equation is not changed if you add the same number to both sides of the equality sign, and you do not change the solution set if you multiply both sides of the equality sign by a non-zero constant.

All linear equations that are not inconsistent and which contain more than one solution, have infinitely many solutions. The following example shows how the solution set in this case can be written.

|||| Example 6.5 Infinitely Many Solutions in Standard Parameter Form

We consider an inhomogeneous equation with three unknowns:

$$2\,x_1 - x_2 + 4\,x_3 = 5\,. \tag{6-7}$$

By substitution of $x_1 = 1$, $x_2 = 1$ and $x_3 = 1$ into the equation (6-7) we see that $\mathbf{x} = (1,1,1)$ is a solution. But by this we have not found the general solution, because $\mathbf{x} = (\frac{1}{2},0,1)$ is also a solution. How can we describe the complete set of solutions?

First we isolate x_1:

$$x_1 = \tfrac{5}{2} + \tfrac{1}{2}\,x_2 - 2\,x_3\,. \tag{6-8}$$

To every choice of x_2 and x_3 corresponds exactly one x_1. For example, if we set $x_2 = 1$ and $x_3 = 4$, then $x_1 = -5$. This means that the 3-tuple $(-5,1,4)$ is a solution. Therefore we can consider x_2 and x_3 *free parameters* that together determine the value of x_1. Therefore we **rename** x_2 and x_3 to the parameter names s and t, respectively: $s = x_2$ and $t = x_3$. Then x_1 can be expressed as:

$$x_1 = \tfrac{5}{2} + \tfrac{1}{2}\,x_2 - 2\,x_3 = \tfrac{5}{2} + \tfrac{1}{2}\,s - 2\,t\,. \tag{6-9}$$

Now we can write the general solution to (6-7) in the following ***standard parameter form***:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ 0 \\ 0 \end{bmatrix} + s \cdot \begin{bmatrix} \frac{1}{2} \\ 1 \\ 0 \end{bmatrix} + t \cdot \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} \text{ with } s,t \in \mathbb{L}\,. \tag{6-10}$$

Note that the parameter form of the middle equation $x_2 = 0 + s \cdot 1 + t \cdot 0$ only expresses the renaming $x_2 \to s$. Similarly, the last equation only expresses the renaming $x_3 \to t$.

If we consider the equation (6-7) to be an equation for a plane in space, then the equation (6-10) is a *parametric representation* for the same plane. The first column on the right hand side is the *initial point* in the plane, and the two last columns are *directional vectors* for the plane. This is elaborated in the eNote 10 *Geometric Vectors*.

6.2 A System of Linear Equations

A ***system of linear equations*** consisting of m linear equations with n unknowns is written in the form

$$\begin{matrix} a_{11}\cdot x_1 + & a_{12}\cdot x_2 + \ldots + & a_{1n}\cdot x_n = b_1 \\ a_{21}\cdot x_1 + & a_{22}\cdot x_2 + \ldots + & a_{2n}\cdot x_n = b_2 \\ & \vdots & \\ a_{m1}\cdot x_1 + & a_{m2}\cdot x_2 + \ldots + & a_{mn}\cdot x_n = b_m \end{matrix} \tag{6-11}$$

The system has m *rows*, each of which contains an equation. The n unknowns, denoted $x_1, x_2, \ldots x_n$, are present in each of the m equations (unless some of the coefficients are zero, and we choose not to write down the zero terms). The coefficient of x_j in the equation in row number i is denoted a_{ij}. The system is termed *homogeneous* if all the m right hand sides b_i are equal to 0, otherwise *inhomogeneous*.

|||| Definition 6.6 Solution of System of Linear Equations

By a *solution* to the the system of linear equations

$$\begin{matrix} a_{11}\cdot x_1 + & a_{12}\cdot x_2 + \ldots + & a_{1n}\cdot x_n = b_1 \\ a_{21}\cdot x_1 + & a_{22}\cdot x_2 + \ldots + & a_{2n}\cdot x_n = b_2 \\ & \vdots & \\ a_{m1}\cdot x_1 + & a_{m2}\cdot x_2 + \ldots + & a_{mn}\cdot x_n = b_m \end{matrix} \tag{6-12}$$

we understand an n-tuple $\mathbf{x} = (x_1, x_2, \ldots x_n) \in \mathbb{L}^n$ which by substitution into all of the m linear equations satisfies the equations, i.e. makes the left hand side of each equal to the right hand side.

By the ***general solution*** or just the ***solution set*** we understand the set of all solutions to the system. A single solution is often termed a ***particular*** solution.

|||| Example 6.7 A Homogeneous System of Linear Equations

A homogeneous system of linear equations consisting of two equations with four unknowns is given by:

$$\begin{aligned} x_1 + x_2 + 2\,x_3 + x_4 &= 0 \\ 2\,x_1 - x_2 - x_3 + x_4 &= 0 \end{aligned} \tag{6-13}$$

We investigate whether the two 4-tuples $\mathbf{x} = (1,1,2,-6)$ and $\mathbf{y} = (3,0,1,-5)$ are particular solutions to the equations (6-13). Substituting $\mathbf{x}$ into the left hand side of the system we get

$$\begin{aligned} 1+1+2\cdot 2-6 &= 0 \\ 2\cdot 1-1-2-6 &= -7 \end{aligned} \tag{6-14}$$

Because the left hand side is equal to the given right hand side 0 in the first of these equations, $\mathbf{x}$ is only a solution to the first of the two equations. Therefore $\mathbf{x}$ is not a solution to the system.

Substituiting $\mathbf{y}$ we get

$$\begin{aligned} 3+0+2\cdot 1-5 &= 0 \\ 2\cdot 3-0-1-5 &= 0 \end{aligned} \tag{6-15}$$

Since in both equations the left hand side is equal to the right hand side 0, $\mathbf{y}$ is a solution to both of the equations. Therefore $\mathbf{y}$ is a particular solution to the system.

The solution set to a system of linear equations is the ***intersection*** of the solution sets for all the equations comprising the system.

6.3 The Coefficient Matrix and the Augmented Matrix

When we investigate a system of linear equations it is often convenient to use *matrices*. A ***matrix*** is a rectangular array consisting of a number of rows and columns. As an example the matrix $\mathbf{M}$ given by

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 5 \\ 8 & 3 & 2 \end{bmatrix}, \tag{6-16}$$

has two rows and three columns. The six elements are termed the ***elements*** of the matrix. The ***diagonal*** of the matrix consists of the elements with equal row and column numbers. In $\mathbf{M}$ the diagonal consists of the elements 1 and 3.

By the *coefficient matrix* **A** to the system of linear equations (6-11) we understand the matrix whose first row consists of the coefficients in the first equation, whose second row consists of the coefficients in the second equation, etc. In short, the following matrix with m rows and n columns:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{6-17}$$

The *augmented matrix* **T** of the system is constructed by adding a new column to the coefficient matrix consisting of the right hand sides b_i of the system. Thus **T** consists of m rows and $n+1$ columns. If we collect the right hand sides b_i into a column vector **b**, which we denote *the right hand side of the system*, **T** is composed as follows, where the vertical line symbolizes the equality sign of the system:

$$\mathbf{T} = \left[\, \mathbf{A} \,\middle|\, \mathbf{b} \,\right] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right] \tag{6-18}$$

The vertical line in front of the last column in (6-18) has only the didactical function to create a clear representation of the augmented matrix. One can chose to leave out the line if in a given context this does not lead to misunderstandings.

|||| Example 6.8 Coefficient Matrix, Right Hand Side and Augmented Matrix

In the following system of linear equations with 3 equations and 3 unknowns

$$\begin{aligned} -x_2 + x_3 &= 2 \\ 2x_1 + 4x_2 - 2x_3 &= 2 \\ 3x_1 + 4x_2 + x_3 &= 9 \end{aligned} \tag{6-19}$$

we have

$$\mathbf{A} = \begin{bmatrix} 0 & -1 & 1 \\ 2 & 4 & -2 \\ 3 & 4 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 9 \end{bmatrix} \quad \text{and} \quad \mathbf{T} = \left[\begin{array}{ccc|c} 0 & -1 & 1 & 2 \\ 2 & 4 & -2 & 2 \\ 3 & 4 & 1 & 9 \end{array}\right] \tag{6-20}$$

Notice that the 0 that is placed in the top left position in **A** and **T**, denotes that the coefficient of x_1 in the uppermost row of the system is 0.

The clever thing about a coefficient matrix (and an augmented matrix) is that we do not need to write down the unknowns. The unique position of the coefficients in the matrix means that we are sure of which of the unknowns any single particular coefficient belongs to. Thus we have removed redundant symbols!

6.4 Row Reduction of Systems of Linear Equations

Systems or linear equations can be reduced, that is, made simpler using a method called Gaussian elimination. The method has several versions, and the special variant used in these eNotes goes by the name ***Gauss-Jordan elimination*** . The algebraic basis for all variants is that you can reshape a system of linear equations by so-called ***row operations*** without thereby changing the solution set for the system. When a system of equations is reduced as much as possible it is usually easy to read it and to evaluate the solution set.

|||| Theorem 6.9 Row Operations

The solution set of a system of linear equations is not altered if the system is transformed by any of the following three ***row operations***:

ro_1: Let two of the equations swap rows.

ro_2: Multiply one of the equations by a non-zero constant.

ro_3: To a given equation add one of the other equations multiplied by a constant.

Here we introduce a short notation for each of the three row operations:

ro_1: $R_i \leftrightarrow R_j$: The equation in row i is swapped with the equation in row j.

ro_2: $k \cdot R_i$: The equation in row i is multiplied by k.

ro_3: $R_j + k \cdot R_i$: Add the equation in row i, multiplied by k, to the equation in row j.

In the following example we test the three row operations.

|||| Example 6.10 Row Operations

An example of ro$_1$: Consider the system of equations below to the left. We swap two equations in the two rows thus performing $R_1 \leftrightarrow R_2$.

$$\begin{array}{rl} x_1 + 2x_2 = -3 \\ x_1 + x_2 = 0 \end{array} \quad \rightarrow \quad \begin{array}{rl} x_1 + x_2 = 0 \\ x_1 + 2x_2 = -3 \end{array} \tag{6-21}$$

The system to the right has the same solution set as the system on the left.

An example of ro$_2$: Consider the system of equations below to the left. We multiply the equation in the second row by 5, thus performing $5 \cdot R_2$:

$$\begin{array}{rl} x_1 + 2x_2 = -3 \\ x_1 + x_2 = 0 \end{array} \quad \rightarrow \quad \begin{array}{rl} x_1 + 2x_2 = -3 \\ 5\,x_1 + 5\,x_2 = 0 \end{array} \tag{6-22}$$

The system to the right has the same solution set as the system on the left.

An example of ro$_3$: Consider the system of equations below to the left. To the equation in the second row we add the equation in the first row multiplied by 2, thus performing $R_2 + 2 \cdot R_1$:

$$\begin{array}{rl} x_1 + 2x_2 = -3 \\ x_1 + x_2 = 0 \end{array} \quad \rightarrow \quad \begin{array}{rl} x_1 + 2x_2 = -3 \\ 3x_1 + 5x_2 = -6 \end{array} \tag{6-23}$$

The system to the right has the same solution set as the system on the left.

The arrow, →, which is used in the three examples indicates that one or more row operations have taken place.

|||| Proof

The first part of the proof of 6.9 is simple: Since the solution set of a system of equations is equal to the *intersection* F of the solution sets for the various equations comprising the system, F is not altered by the order of the equations being changed. Therefore ro$_1$ is allowed.

Since the solution set of a given equation is not altered when the equation is multiplied by a constant $k \neq 0$, F will not be altered if one of the equations is replaced by the equation multiplied by a constant different from 0. Therefore ro$_2$ is allowed.

Finally consider a system of linear equations A with n unknowns $\mathbf{x} = (x_1, x_2, \ldots x_n)$. We write the left hand side of an equation in A as $L(\mathbf{x})$ and the right hand side as b. Now we perform an arbitrary row operation of the type ro3 in the following way: An arbitrary equation $L_1(\mathbf{x}) = b_1$ is multiplied by an arbitrary number k and is then added to an arbitrary different equation $L_2(\mathbf{x}) = b_2$. This produces a new equation $L_3(\mathbf{x}) = b_3$ where

$$L_3(\mathbf{x}) = L_2(\mathbf{x}) + k\,L_1(\mathbf{x}) \ \text{ and } \ b_3 = b_2 + k\,b_1\,.$$

We now show that the system of equations B that emerges as a result of replacing $L_2(\mathbf{x}) = b_2$ in A by $L_3(\mathbf{x}) = b_3$ has the same solution set as A, and that ro$_3$ thus is allowed. First, assume that $\mathbf{x}_0$ is an arbitrary solution to A . Then it follows from the transformation rules for a linear equation that

$$k\,L_1(\mathbf{x}_0) = k\,b_1$$

and further that

$$L_2(\mathbf{x}_0) + k\,L_1(\mathbf{x}_0) = b_2 + k\,b_1\,.$$

From this it follows that $L_3(\mathbf{x}_0) = b_3$, and that $\mathbf{x}_0$ is a solution to B. Assume vice versa that $\mathbf{x}_1$ is an arbitrary solution to B . Then it follows that

$$-k\,L_1(\mathbf{x}_1) = -k\,b_1$$

and further that

$$L_3(\mathbf{x}_1) - k\,L_1(\mathbf{x}_1) = b_3 - k\,b_1\,.$$

This means that $L_2(\mathbf{x}_1) = b_2$, and that $\mathbf{x}_1$ also is a solution to A. In sum we have shown that ro$_3$ is allowed.

■

From 6.9 follows directly:

|||| Corollary 6.11

The solution set of a system of linear equations is not altered if the system is transformed an arbitrary number of times, in any order, by the three row operations.

We are now ready to use the three row operations for the row reduction of systems of linear equations. In the following example we follow the principles of *Gauss-Jordan elimination*, and a complete description of the method follows in subsection 6.5.

|||| Example 6.12 Gauss-Jordan Elimination

Consider below, to the left, a system of linear equations, consisting of three equations with the three unknowns x_1, x_2 and x_3. On the right the *augmented matrix* for the system is written:

$$\begin{aligned} -x_2+x_3&=2 \\ 2x_1+4x_2-2x_3&=2 \\ 3x_1+4x_2+x_3&=9 \end{aligned} \qquad \mathbf{T}=\left[\begin{array}{ccc|c} 0 & -1 & 1 & 2 \\ 2 & 4 & -2 & 2 \\ 3 & 4 & 1 & 9 \end{array}\right] \tag{6-24}$$

The purpose of reduction is to achieve, by means of row operations, the following situation: x_1 is the only remaining part on left hand side of the upper equation , x_2 is the only one on the left hand side of the middle equation and x_3 is the only one on the left hand side of the lower equation. *If* this is possible then the system of equations is not only reduced but also solved! This is achieved in a series of steps taken in accordance with the Gauss-Jordan algorithm. Simultaneously we look at the effect the row operations have on the augmented matrix.

First we aim to have the topmost equation comprise x_1, and to have the coefficient of this x_1 be 1. This can be achieved in two steps. We swap the two top equations and multiply the equation now in the top row by $\frac{1}{2}$. That is,

$$R_1 \leftrightarrow R_2 \quad \text{and} \quad \frac{1}{2}\cdot R_1\ :$$

$$\begin{aligned} x_1+2x_2-x_3&=1 \\ -x_2+x_3&=2 \\ 3x_1+4x_2+x_3&=9 \end{aligned} \qquad \left[\begin{array}{ccc|c} 1 & 2 & -1 & 1 \\ 0 & -1 & 1 & 2 \\ 3 & 4 & 1 & 9 \end{array}\right] \tag{6-25}$$

Now we remove all other occurrences of x_1. In this example it is only one occurrence, i.e. in row 3. This is achieved as follows: we multiply the equation in row 1 by the number -3 and add the product to the equation in row 3, in short

$$R_3 - 3\cdot R_1\ :$$

$$\begin{aligned} x_1+2x_2-x_3&=1 \\ -x_2+x_3&=2 \\ -2x_2+4x_3&=6 \end{aligned} \qquad \left[\begin{array}{ccc|c} 1 & 2 & -1 & 1 \\ 0 & -1 & 1 & 2 \\ 0 & -2 & 4 & 6 \end{array}\right] \tag{6-26}$$

We have now achieved that x_1 only appears in row 1 . There it must stay! The work on x_1 is finished. This corresponds to the fact that at the top of the first column of the augmented matrix there is 1 and directly below it only 0's. This means that work on the first column is finished !

The next transformations aim at ensuring that the unknown x_2 will be represented only in row 2 and nowhere else. First we make sure that the coefficient of x_2 in row 2 switches coefficient from -1 to 1 by use of the operation

$$(-1)\cdot R_2 : \quad \begin{aligned} x_1+2x_2-x_3&=1\\ x_2-x_3&=-2\\ -2x_2+4x_3&=6 \end{aligned} \qquad \left[\begin{array}{rrr|r} 1 & 2 & -1 & 1\\ 0 & 1 & -1 & -2\\ 0 & -2 & 4 & 6 \end{array}\right] \tag{6-27}$$

We now remove the occurrences of x_2 from row 1 and row 3 with the operations

$$R_1-2\cdot R_2 \quad \text{and} \quad R_3+2\cdot R_2 : \quad \begin{aligned} x_1+x_3&=5\\ x_2-x_3&=-2\\ 2x_3&=2 \end{aligned} \qquad \left[\begin{array}{rrr|r} 1 & 0 & 1 & 5\\ 0 & 1 & -1 & -2\\ 0 & 0 & 2 & 2 \end{array}\right] \tag{6-28}$$

Now the work with x_2 is finished, which corresponds to the fact that in row 2 in the augmented matrix the number in the second column is 1, all the other numbers in the second column being 0. This column must not be altered by subsequent operations.

Finally we wish that the unknown x_3 is represented in row 3 by the coefficient 1 and that x_3 is removed from row 1 and row 2. This can be accomplished in two steps. First

$$\frac{1}{2}\cdot R_3 : \quad \begin{aligned} x_1+x_3&=5\\ x_2-x_3&=-2\\ x_3&=1 \end{aligned} \qquad \left[\begin{array}{rrr|r} 1 & 0 & 1 & 5\\ 0 & 1 & -1 & -2\\ 0 & 0 & 1 & 1 \end{array}\right] \tag{6-29}$$

Then

$$R_1-R_3 \quad \text{and} \quad R_2+R_3 : \quad \begin{aligned} x_1&=4\\ x_2&=-1\\ x_3&=1 \end{aligned} \qquad \left[\begin{array}{rrr|r} 1 & 0 & 0 & 4\\ 0 & 1 & 0 & -1\\ 0 & 0 & 1 & 1 \end{array}\right] \tag{6-30}$$

Now x_3 only appears in row 3. This corresponds to the fact that in column 3 in the third row of the augmented matrix we have 1, each of the other elements in the column being 0. We have now completed a *total reduction* of the system, and from this we can conclude that there exists exactly one solution to the system viz :

$$\mathbf{x}=(x_1,x_2,x_3)=(4,-1,1). \tag{6-31}$$

Let us remember what a solution is: an n-tuple that satisfies all the equations in the system! Let us prove that formula (6-31) actually is a solution to equation (6-24):

$$\begin{aligned} -(-1)+1&=2\\ 2\cdot4+4\cdot(-1)-2\cdot1&=2\\ 3\cdot4+4\cdot(-1)+1&=9 \end{aligned}$$

As expected all three equations are satisfied!

In (6-30) after the row operations the augmented matrix of the system of linear equations has achieved a form of special beauty with three so-called leading 1's in the ***diagonal*** and zeros everywhere else. We say that the transformed matrix is in *reduced row echelon form*. It is not always possible to get the simple representation shown in (6-30). Sometimes the leading 1 in the next row is found more than one column to the right, as one moves down. The somewhat complex definition follows below.

|||| Definition 6.13 Reduced Row Echelon Form

A system of linear equations is denoted to be in ***reduced row echelon form***, if the corresponding augmented matrix fulfills the following four conditions:

1. The first number in a row that is not 0, is a 1. This is called the ***leading*** 1 or the ***pivot*** of the row.

2. In two consecutive rows which both contain a pivot, the upper row's leading 1 is further to the left than the leading 1 in the following row.

3. In a column with a leading 1, all other elements are 0.

4. Any rows with only 0's are placed at the bottom of the matrix.

|||| Example 6.14 Reduced Row Echelon Form

Consider the three matrices

$$\mathbf{A}=\begin{bmatrix}\mathbf{1}&0&0\\0&\mathbf{1}&0\\0&0&\mathbf{1}\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}\mathbf{1}&2&0\\0&0&\mathbf{1}\\0&0&0\end{bmatrix}\quad \text{og}\quad \mathbf{C}=\begin{bmatrix}\mathbf{1}&3&1\\0&0&0\\0&0&0\end{bmatrix}. \qquad (6\text{-}32)$$

The three matrices shown are all in row reduced echelon form. In **A** all the leading 1's are nicely placed in the *diagonal*. **B** has only leading two leading 1's and you have to go two steps to the right to go from the first to the second step. In **C** there is only one leading 1.

|||| Example 6.15

None of the following four matrices is in reduced row echelon form because each violates exactly one of the rules in the definition 6.13 – which, is left to reader to figure out!

$$\mathbf{A}=\begin{bmatrix}1&1&0\\0&1&0\\0&0&1\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}0&0&0\\1&2&0\\0&0&1\end{bmatrix},\quad \mathbf{C}=\begin{bmatrix}1&0&0\\0&2&1\\0&0&0\end{bmatrix}\quad \text{and}\quad \mathbf{D}=\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}. \qquad (6\text{-}33)$$

Note the following important theorem about the relationship between a matrix on the one hand, and the reduced row echelon form of the same matrix produced through the use of row operations, on the other.

|||| Theorem 6.16 Reduced Row Echelon Form

If a given matrix **M** is transformed by two different sequences of row operations into a reduced row echelon form, then the two resulting reduced row echelon forms are identical.

The unique reduced row echelon form a given matrix **M** can be transformed into this way is termed the ***reduced row echelon form***, and given the symbol $\text{rref}(\mathbf{M})$.

|||| Proof

We use the following model for the six matrices that are introduced in the course of the proof:

$$\begin{tikzcd}
\mathbf{A} & \mathbf{M} \arrow[l, "f_1"'] \arrow[r, "f_2"] \arrow[d] & \mathbf{B} \\
\mathbf{A}_1 & \mathbf{M}_1 \arrow[l, "f_1"'] \arrow[r, "f_2"] & \mathbf{B}_1
\end{tikzcd} \tag{6-34}$$

Suppose a matrix $\mathbf{M}$ has been transformed, by two different series of row operations f_1 and f_2, into two different reduced row echelon forms $\mathbf{A}$ and $\mathbf{B}$. Let column number k be the first column of $\mathbf{A}$ and $\mathbf{B}$ where the two matrices differ from one another. We form a new matrix $\mathbf{M}_1$ from $\mathbf{M}$ in the following way. First we remove all the columns in $\mathbf{M}$ whose column numbers are larger than k. Then we remove just the columns in $\mathbf{M}$ whose column numbers are less than k , and have the same column numbers as a column in $\mathbf{A}$ (and thus $\mathbf{B}$) which does not contain a leading 1.

Now we transform $\mathbf{M}_1$ by the series of row operations f_1 and f_2 , and the resulting matrices formed hereby are called $\mathbf{A}_1$ and $\mathbf{B}_1$, respectively. Then $\mathbf{A}_1$ necessarily will be the same matrix that would result if we remove all the columns from $\mathbf{A}$, similar to those we took away from $\mathbf{M}$ to produce $\mathbf{M}_1$. And the same relationship exists between $\mathbf{B}_1$ and $\mathbf{B}$. $\mathbf{A}_1$ and $\mathbf{B}_1$ will therefore have a leading 1 in the diagonal of all columns apart from the last, which is the first column where the two matrices are different from one another. In this last column there are two possibilities: Either one of the matrices has a leading 1 in this column or neither of them has. An example of how the situation in the first case could be is:

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{B}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \tag{6-35}$$

We now interpret $\mathbf{M}_1$ as the augmented matrix for a system of linear equations $\mathcal{L}$. Both $\mathbf{A}_1$ and $\mathbf{B}_1$ will then represent a totally reduced system of equations with the same solution set as $\mathcal{L}$. However, this leads to a contradiction since one of the totally reduced systems is seen to be inconsistent due to one of the equations now being invalid and the other will have just one solution. We can therefore rule out that one of $\mathbf{A}_1$ and $\mathbf{B}_1$ contains a leading 1 in the last column.

We now investigate the other possibility, that neither of $\mathbf{A}_1$ and $\mathbf{B}_1$ contains a leading 1 in the last column. The situation could then be like this:

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B}_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \tag{6-36}$$

Both the totally reduced system of equations as represented by $\mathbf{A}_1$, and that which is represented by $\mathbf{B}_1$, will in this case have exactly one solution. But when the last column is different in the two matrices the solution for $\mathbf{A}_1$'s system of equations will be different from the solution for $\mathbf{B}_1$'s system of equations, whereby we again have ended up in a contradiction.

We conclude that the assumption that **M** might be transformed into two different reduced row echelon forms cannot be true. Hence, to **M** corresponds a unique reduced row echelon form: rref(**M**).

■

From Theorem 6.16 it is relatively easy to obtain the next result about matrices that can transformed into each other through row operations:

|||| Corollary 6.17

If a matrix **M** has been transformed by an arbitrary sequence of row operations into the matrix **N**, then

$$\operatorname{rref}(\mathbf{N}) = \operatorname{rref}(\mathbf{M}). \tag{6-37}$$

|||| Proof

Let s be a sequence of row operations that transforms the matrix **M** to the matrix **N**, and let t be a sequence of row operations that transforms the the matrix **N** to $\operatorname{rref}(\mathbf{N})$. Then the sequence of row operations consisting of s followed by t , transform **M** to $\operatorname{rref}(\mathbf{N})$. But since **M** in accordance with 6.16 has a unique reduced row echelon form, $\operatorname{rref}(\mathbf{M})$ must be equal to $\operatorname{rref}(\mathbf{N})$.

■

If, in the preceding corollary, we interpret **M** and **N** as the augmented matrices for two systems of linear equations, then it follows directly from definition (6.13) that:

Corollary 6.18

If two systems of linear equations can be transformed into one another by the use of row operations, then they are identical in the reduced row echelon form (apart from possible trivial equations).

6.5 Gauss-Jordan Elimination

We are now able to precisely introduce the method of elimination that is applied in these eNotes.

Definition 6.19 Gauss-Jordan Elimination

A system of linear equations is totally reduced by ***Gauss-Jordan elimination*** when the corresponding augmented matrix after the use of the three row operations (see theorem 6.9) is brought into the reduced row echelon form by the following procedure:

> We proceed from left to right : First we treat the first column of the augmented matrix so that it does not conflict with the reduced row echelon form, then the second column is treated so as not to conflict with the reduced row echelon form and so on, as far as and including the last column in the augmented matrix .

This is always possible!

> When you are in the process of reducing systems of linear equations, you are free to deviate from the Gauss-Jordan method if it is convenient in the situation at hand. If you have achieved a reduced row echelon form by using other sequences of row operations, it is the same form that would have been obtained by using the Gauss-Jordan method strictly. This follows from corollary 6.18.

In Example 6.12 it was possible to read the solution from the totally reduced system of linear equations. In the following main example the situation is a bit more complicated owing to the fact that the system has infinitely many solutions.

|||| Example 6.20 Gauss-Jordan Elimination

We want to reduce the following system of four linear equations in five unknowns:

$$\begin{aligned} x_1+3x_2+2x_3+4x_4+5x_5&=9\\ 2x_1+6x_2+4x_3+3x_4+5x_5&=3\\ 3x_1+8x_2+6x_3+7x_4+6x_5&=5\\ 4x_1+14x_2+8x_3+10x_4+22x_5&=32 \end{aligned} \tag{6-38}$$

We write the augmented matrix for the system:

$$\mathbf{T}=\left[\begin{array}{ccccc|c} 1&3&2&4&5&9\\ 2&6&4&3&5&3\\ 3&8&6&7&6&5\\ 4&14&8&10&22&32 \end{array}\right] \tag{6-39}$$

Below we reduce the system using three row operations. This we will do by only looking at the transformations of the augmented matrix!

$$R_2-2\cdot R_1\,,\quad R_3-3\cdot R_1\quad\text{and}\quad R_4-4\cdot R_1:$$

$$\left[\begin{array}{ccccc|c} 1&3&2&4&5&9\\ 0&0&0&-5&-5&-15\\ 0&-1&0&-5&-9&-22\\ 0&2&0&-6&2&-4 \end{array}\right] \tag{6-40}$$

Now we have completed the treatment of the first column, because we have a leading 1 in the first row and only 0's on the other entries in the column.

$$R_2\leftrightarrow R_3\quad\text{and}\quad(-1)\cdot R_2:$$

$$\left[\begin{array}{ccccc|c} 1&3&2&4&5&9\\ 0&1&0&5&9&22\\ 0&0&0&-5&-5&-15\\ 0&2&0&-6&2&-4 \end{array}\right] \tag{6-41}$$

$$R_1-3\cdot R_2\quad\text{and}\quad R_4-2\cdot R_2:$$

$$\left[\begin{array}{ccccc|c} 1&0&2&-11&-22&-57\\ 0&1&0&5&9&22\\ 0&0&0&-5&-5&-15\\ 0&0&0&-16&-16&-48 \end{array}\right] \tag{6-42}$$

The work on the second column is now completed. Now a deviation from the standard situation follows, where leading 1's are established in the diagonal, because it is not possible to produce a leading 1 as the third element in the third row. We are *not* allowed to swap row 1 and row 3, because by doing so the first column would be changed in conflict with the principle that the treatment of the first column is complete. This means that we have also completed the treatment of the third column (the number 2 in the top row cannot be removed). To continue the reduction we move on to the fourth element in row three, where it *is* possible to provide a leading 1.

$$-\tfrac{1}{5}\cdot R_3 :$$

$$\left[\begin{array}{ccccc|c} 1 & 0 & 2 & -11 & -22 & -57 \\ 0 & 1 & 0 & 5 & 9 & 22 \\ 0 & 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & -16 & -16 & -48 \end{array}\right] \tag{6-43}$$

$$R_1 + 11\cdot R_3\,, \quad R_2 - 5\cdot R_3 \quad \text{and} \quad R_4 + 16\cdot R_3 :$$

$$\left[\begin{array}{ccccc|c} 1 & 0 & 2 & 0 & -11 & -24 \\ 0 & 1 & 0 & 0 & 4 & 7 \\ 0 & 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \tag{6-44}$$

Now the Gauss-Jordan elimination has ended and we can write the totally reduced system of equations:

$$\begin{aligned} 1x_1 + 0x_2 + 2x_3 + 0x_4 - 11x_5 &= -24 \\ 0x_1 + 1x_2 + 0x_3 + 0x_4 + 4x_5 &= 7 \\ 0x_1 + 0x_2 + 0x_3 + 1x_4 + 1x_5 &= 3 \\ 0x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 &= 0 \end{aligned} \tag{6-45}$$

First, we note that the original system of equations has actually been reduced (made easier) by the fact that many of the coefficients of the equation system are replaced by 0's. But moreover the system with four equations can now be replaced by a system consisting of only three equations. The last row is indeed a ***trivial*** equation that has the whole $\mathbb{L}^5$ as its solution set. Therefore, the solution set of the system system will not change if the last equation is omitted in the reduced system (since the intersection of the solutions sets of all four equations equals that of the solution sets from the first three equations alone). Quite simply , we can therefore write the totally reduced system of equations as:

$$\begin{aligned} x_1 + 2x_3 - 11x_5 &= -24 \\ x_2 + 4x_5 &= 7 \\ x_4 + x_5 &= 3 \end{aligned} \tag{6-46}$$

But how do we proceed from the reduced system of equations to writing down the solution set in a comprehensible form? We shall return to this example later, see Example 6.30. Before that we need to introduce the concept of *rank*.

6.6 The Concept of Rank

In the example 6.20 a system of linear equations consisting of 4 equations with 5 unknowns has been totally reduced, see equation (6-46). Only three equations are left, because the trivial equation $0x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 = 0$ has been left out since it only expresses the fact that $0 = 0$. That the reduced system of equations contains a trivial equation means that the reduced row echelon form of the the augmented matrix contains a 0-row, as in equation (6-44). This leads to the following definition.

Definition 6.21 Rank

By the ***rank*** ρ of a *matrix* we understand the number of rows that are not 0-***rows***, in the reduced row echelon form of the matrix. The rank thereby corresponds to the number of leading $1'$s in the reduced row echelon form of the matrix.

From the definition 6.21 and corollary 6.18 together with corollary 6.17 we obtain:

Theorem 6.22 Rank and Row Operations

Two matrices that can be transformed into each other by row operations have the same rank.

The rank gives the least possible number of non-trivial equations that a system of equations can be transformed into using row operations. You can never transform a system of linear equations through row operations in such a way that it will contain fewer non-trivial equations than it does when it is totally reduced. This is a consequence of theorem 6.22.

Example 6.23 The Rank of Matrices

A matrix $\mathbf{M}$ with 3 rows and 4 columns is brought into the reduced row echelon form as follows:

$$\mathbf{M} = \begin{bmatrix} 3 & 1 & 7 & -2 \\ -1 & -3 & 3 & 1 \\ 2 & 3 & 0 & -3 \end{bmatrix} \quad \rightarrow \quad \operatorname{rref}(\mathbf{M}) = \begin{bmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6-47}$$

Since $\text{rref}(\mathbf{M})$ does not contain 0-rows, $\rho(\mathbf{M}) = 3$.

A matrix $\mathbf{N}$ with 5 rows and 3 columns is brought into reduced row echelon form like this:

$$\mathbf{N} = \begin{bmatrix} 2 & 2 & 1 \\ -2 & -5 & -4 \\ 3 & 1 & -7 \\ 2 & -1 & -8 \\ 3 & 1 & -7 \end{bmatrix} \quad \rightarrow \quad \text{rref}(\mathbf{N}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6-48}$$

Since $\text{rref}(\mathbf{N})$ contains three rows that are not 0-rows, $\rho(\mathbf{N}) = 3$.

If we interpret $\mathbf{M}$ and $\mathbf{N}$ as augmented matrices for linear systems of equations we see that for both coefficient matrices the rank is 2, this is less than the ranks of the augmented matrices.

We now investigate the relationship between rank and the number of rows and columns. First we notice that from the definition of 6.21 it follows that the rank of a matrix can never be larger than the number of matrix rows.

In Example 6.23 the rank of $\mathbf{M}$ equals the number of rows in $\mathbf{M}$, while the rank of $\mathbf{N}$ is less than the number of rows in $\mathbf{N}$.

Analogously the rank of a matrix cannot be larger than the number of columns. The rank is in fact equal to the number of leading 1's in the reduced row echelon form . And if the echelon form of the matrix contains more leading 1's than there are columns, then there must be at least one column containing more than one leading 1. But this contradicts condition number 3 in the definition 6.13.

In the example 6.23 the rank of $\mathbf{M}$ is less than the number of columns in $\mathbf{M}$, while the rank of $\mathbf{N}$ equals the number of columns in $\mathbf{N}$.

We summarize the above observations in the following theorem:

|||| Theorem 6.24 Rank, Rows and Columns

For a matrix $\mathbf{M}$ with m rows and n columns we have that

$$\rho(\mathbf{M}) \leq \min\{m, n\}\,. \tag{6-49}$$

6.7 From Reduced Row Echelon Form to the Solution Set

Sometimes it is possible to write down the solution set for a system of linear equations immediately when the corresponding augmented matrix is brought into its reduced echelon form. This applies when the system has no solution or when the system has exactly one solution. If the system has infinitely many solutions, work is needed in order to be able to characterize the solution set. This can be achieved by writing the solution in a standard parametric form. The concept of rank proves well suited to give an instructive overview of the classes of solution sets.

6.7.1 When $\rho(\mathbf{A}) < \rho(\mathbf{T})$

The augmented matrix $\mathbf{T}$ for a system of linear equations has the same number of rows as the coefficient matrix $\mathbf{A}$ but one column more, which contains the right hand sides of the equations. There are two possibilities. Either $\rho(\mathbf{T}) = \rho(\mathbf{A})$, or $\rho(\mathbf{T}) = \rho(\mathbf{A}) + 1$, corresponding to the fact that the last column in $\mathrm{rref}(\mathbf{T})$ contains a leading 1. The consequence of the last possibility is investigated in Example 6.25.

|||| Example 6.25 Inconsistent Equation (No Solution)

The augmented matrix for a system of linear equations consisting of three equations in two unknowns is brought into reduced row echelon form

$$\mathrm{rref}(\mathbf{T}) = \left[\begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right] \tag{6-50}$$

The system is thereby reduced to

$$\begin{aligned} x_1 - 2x_2 &= 0 \\ 0x_1 + 0x_2 &= 1 \\ 0x_1 + 0x_2 &= 0 \end{aligned} \tag{6-51}$$

Notice that the equation in the second row is *inconsistent* and thus has no solutions. Because the solution set for the system is the intersection of the solution sets for all the equations, the system has no solutions at all.

Let us look at the reduced row echelon form of the coefficient matrix

$$\mathrm{rref}(\mathbf{A}) = \begin{bmatrix} 1 & -2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{6-52}$$

We note that $\rho(\mathbf{A}) = 1$. This is less than $\rho(\mathbf{T}) = 2$, and this is exactly due to the inconsistency of the equation in the reduced system of equations.

The considerations in example 6.25 can be generalized to the following theorem.

|||| Theorem 6.26 When $\rho(\mathbf{A}) < \rho(\mathbf{T})$

If a system of linear equations with coefficient matrix $\mathbf{A}$ and augmented matrix $\mathbf{T}$ has

$$\rho(\mathbf{A}) < \rho(\mathbf{T})\,, \tag{6-53}$$

then the totally reduced system has an inconsistent equation. Therefore the system has no solutions.

If $\mathrm{rref}(\mathbf{T})$ has a row of the form $\left[\begin{array}{cccc|c} 0 & 0 & \cdots & 0 & 1 \end{array}\right]$, then the system has no solutions.

|||| Exercise 6.27

Determine the reduced rwo echelon form of the augmented matrix for the following system of linear equations, and determine the solution set for the system.

$$\begin{aligned} x_1 + x_2 + 2x_3 + x_4 &= 1 \\ -2x_1 - 2x_2 - 4x_3 - 2x_4 &= 3 \end{aligned} \tag{6-54}$$

6.7.2 When $\rho(\mathbf{A}) = \rho(\mathbf{T}) =$ Number of Unknowns

Let n denote the number of unknowns in a given system of linear equations. Then by the way the coefficient matrices are formed there must be n columns in $\mathbf{A}$.

Further we assume that $\rho(\mathbf{A}) = n$. Then $\mathrm{rref}(\mathbf{A})$ contains exactly n leading 1's. Therefore the leading 1's must be placed in the ***diagonal*** in $\mathrm{rref}(\mathbf{A})$, and all other elements of $\mathrm{rref}(\mathbf{A})$ are zero.

Finally we assume that in the given example $\rho(\mathbf{A}) = \rho(\mathbf{T})$. Then the solution set can be read directly from $\operatorname{rref}(\mathbf{T})$. The first row in $\operatorname{rref}(\mathbf{T})$ will correspond to an equation where the first unknown has the coefficient 1 while all the other unknowns have the coefficient 0. Therefore the value of the first unknown is equal to to the last element in the first row (the right hand side). Similarly with the other rows, row number i corresponds to an equation where unknown number i is the only unknown, and therefore its value is equal to the last element in row number i. Since each unknown there corresponds to exactly one value, and since $\rho(\mathbf{A}) = \rho(\mathbf{T})$ we are certain that there is no inconsistent equation in the given system of equations. Thus the given system of equations has exactly one solution.

|||| Example 6.28 Exactly One Solution

The augmented matrix for a system of linear equations consisting of three equations in two unknowns has been brought onto the reduced row echelon form

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{cc|r} 1 & 0 & -3 \\ 0 & 1 & 5 \\ 0 & 0 & 0 \end{array}\right] \tag{6-55}$$

Consider the reduced row echelon form of the coefficient matrix for the system

$$\operatorname{rref}(\mathbf{A}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{6-56}$$

This has a leading 1 in each column and 0 in all other entries. We further note that $\rho(\mathbf{A}) = \rho(\mathbf{T}) = 2$.

From $\operatorname{rref}(\mathbf{T})$ we can write the totally reduced system of equations as

$$\begin{aligned} 1x_1 + 0x_2 &= -3 \\ 0x_1 + 1x_2 &= 5 \\ 0x_1 + 0x_2 &= 0 \end{aligned} \tag{6-57}$$

which shows that this system of equations has exactly one solution $\mathbf{x} = (x_1, x_2) = (-3, 5)$.

The argument given just before the example proves the following theorem:

|||| **Theorem 6.29 When $\rho(\mathbf{A}) = \rho(\mathbf{T}) =$ Number of Unknowns**

If a linear system with coefficient matrix **A** and augmented matrix **T** has:

$$\rho(\mathbf{A}) = \rho(\mathbf{T}) = \text{number of unknowns}, \tag{6-58}$$

then the system has exactly one solution, and this can be immediately read from $\text{rref}(\mathbf{T})$.

6.7.3 When $\rho(\mathbf{A}) = \rho(\mathbf{T}) <$ the Number of Unknowns

We are now ready to resume the discussion of our main example 6.20, a system of linear equations with 5 unknowns, for which we found the totally reduced system of equations consisting of 3 non-trivial equations. Let us now find the solution set and investigate its properties!

|||| **Example 6.30 Infinitely Many Solutions**

In the example 6.20 the augmented matrix **T** for a system of linear equations with 4 equations in 5 unknowns was reduced to

$$\text{rref}(\mathbf{T}) = \left[\begin{array}{ccccc|c} 1 & 0 & 2 & 0 & -11 & -24 \\ 0 & 1 & 0 & 0 & 4 & 7 \\ 0 & 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \tag{6-59}$$

We see that $\rho(\mathbf{A}) = \rho(\mathbf{T}) = 3$, i.e. less than 5, the number of unknowns.

From $\text{rref}(\mathbf{T})$ we can write the totally reduced system of equations

$$\begin{aligned} x_1 + 2x_3 - 11x_5 &= -24 \\ x_2 + 4x_5 &= 7 \\ x_4 + x_5 &= 3 \end{aligned} \tag{6-60}$$

The system has infinitely many solutions. For every choice of values for x_3 and x_5 we can find exactly one new value for the other unknowns x_1, x_2 and x_4. This can be made more clear by isolating x_1, x_2 and x_4 in the following way

$$\begin{aligned} x_1 &= -24 - 2x_3 + 11x_5 \\ x_2 &= 7 - 4x_5 \\ x_4 &= 3 - x_5 \end{aligned} \tag{6-61}$$

If we, for example, choose $x_3 = 1$ and $x_5 = 2$, we find the solution $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (-4, -1, 1, 1, 2)$. More generally, *any* choice of values for x_3 and x_5 will, in the same way, produce a solution, whilst the other three variables are uniquely determined by the cohice. Therefore we can consider x_3 and x_5 as *free parameters* that determine the value of the three other unknowns, and therefore on the right hand side we rename x_3 and x_5 the parameter names t_1 and t_2, respectively. Then we can write the solution set as:

$$\begin{aligned} x_1 &= -24 - 2t_1 + 11t_2 \\ x_2 &= 7 - 4t_2 \\ x_3 &= t_1 \\ x_4 &= 3 - t_2 \\ x_5 &= t_2 \end{aligned} \tag{6-62}$$

or more clearly in the *standard parameter form*:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -24 \\ 7 \\ 0 \\ 3 \\ 0 \end{bmatrix} + t_1 \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \begin{bmatrix} 11 \\ -4 \\ 0 \\ -1 \\ 1 \end{bmatrix} \quad \text{where} \quad t_1, t_2 \in \mathbb{L}. \tag{6-63}$$

With geometry-inspired wording we term the vector $(-24, 7, 0, 3, 0)$ the *initial point* of the solution set and the two vectors $(-2, 0, 1, 0, 0)$ and $(11, -4, 0, -1, 1)$ its *directional vectors*. Letting $\mathbf{x}_0$, $\mathbf{v}_1$ and $\mathbf{v}_2$ denote the initial point, and the directional vectors, respectively, we can write the parametric representation in this way:

$$\mathbf{x} = \mathbf{x}_0 + t_1\mathbf{v}_1 + t_2\mathbf{v}_2 \quad \text{hvor} \quad t_1, t_2 \in \mathbb{L}. \tag{6-64}$$

Since the solution set has two free parameters corresponding to two directional vectors, we say that it has a *double -infinity* of solutions.

Line 3 and 5 in (6-63) only express that $x_3 = t_1$ and $x_5 = t_2$.

Let us, inspired by example 6.30, formulate a general method for changing the solution set to standard parametic form from the totally reduced system of equations:

|||| Method 6.31 From the Augmented Matrix to the Solution in Standard Parameter Form

We consider a system of linear equations with n unknowns with the coefficient matrix $\mathbf{A}$ and the augmented matrix $\mathbf{T}$. In addition we assume

$$\rho(\mathbf{A}) = \rho(\mathbf{T}) = k < n. \tag{6-65}$$

The solution set of the system is brought into standard parametric form in this way:

1. We find $\mathrm{rref}(\mathbf{T})$ and from this we write the totally reduced system of equations (as is done in (6-60)).

2. In each of the k non-trivial equations in the totally reduced system of equations we isolate the *first* unknowns on the left hand side (as is done in (6-61)).

3. In this way we have isolated k different unknowns on the left hand side of the total system. The other $(n-k)$ unknowns, that are placed on the right hand side are *renamed* the parameter names $t_1, t_2, \ldots, t_{n-k}$.

4. We can now write the solution set in *standard parametic form*:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) = \mathbf{x}_0 + t_1\,\mathbf{v}_1 + t_2\,\mathbf{v}_2 + \cdots + t_{n-k}\,\mathbf{v}_{n-k}\,, \tag{6-66}$$

   where the vector $\mathbf{x}_0$ denotes the *initial point* of the parameter representation, while $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-k}$ are its *directional vectors* (as is done in (6-63)).

Notice that the numbers $t_1, t_2, \ldots, t_{n-k}$ can be chosen freely. Regardless of the choice equation (6-66) will be a valid solution. Therefore they are called ***free parameters***.

If the algorithm of the Gauss-Jordan elimination has been followed perfectly, one arrives at a certain initial point and a certain set of directional vectors for the solution set, see equation (6-66). But the solution set can be written with another choice for the initial point (if the system is inhomogeneous), and with a different choice of directional vectors. However, the *number* of directional vectors will always be $(n-k)$.

Solution sets in which some of the unknowns have definite values are possible. In the following example the free parameter only influences one of the unknowns. The other two are locked:

|||| Example 6.32 Infinitely Many Solutions with a Free Parameter

For a given system of linear equations it is found that

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{ccc|c} 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 5 \end{array}\right] \tag{6-67}$$

We see that $\rho(\mathbf{A}) = \rho(\mathbf{T}) = 2 < n = 3$. Accordingly we have one free parameter. We write the solution set as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 5 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{6-68}$$

where t is a scalar that can be chosen freely.

In general you can prove the following theorem:

|||| Theorem 6.33 When $\rho(\mathrm{A}) = \rho(\mathrm{T}) <$ Number of Unknowns

If a system of linear equations with n unknowns and with the coefficient matrix $\mathbf{A}$ and augment matrix $\mathbf{T}$ has

$$\rho(\mathbf{A}) = \rho(\mathbf{T}) = k < n \tag{6-69}$$

Then the system has infinitely many solutions that can be written in standard parameter form with an initial point and $(n-k)$ directional vectors.

6.8 On the Number of Solutions

Let us consider a system of three linear equations in two unknowns:

$$\begin{aligned} a_1 \cdot x + b_1 \cdot y &= c_1 \\ a_2 \cdot x + b_2 \cdot y &= c_2 \\ a_3 \cdot x + b_3 \cdot y &= c_3 \end{aligned} \tag{6-70}$$

We have previously emphasized that the solution set for a system of equations is the ***intersection*** of the solution sets for each of the equations in the system. Let us now interpret the given system of equations as equations for three straight lines in a coordinate system in the plane. Then the solution set corresponds to a set of points that are *common* to all the three lines. In order to answer the question about "number" of solutions we draw the different situations in Figure 6.1. In situation 2 two of the lines are parallel, and in situation 3 all three lines are parallel. Therefore there are no points that are part of all three lines in the situations 1, 2 and 3. In situation 5 two of the lines are identical (the blue and the red line coincide in the purple line). Hence there is exactly one common point in the situations 4 and 5. In the situation 6 all the three lines coincide (giving the black line). Therefore in this situation there are infinitely many common points.

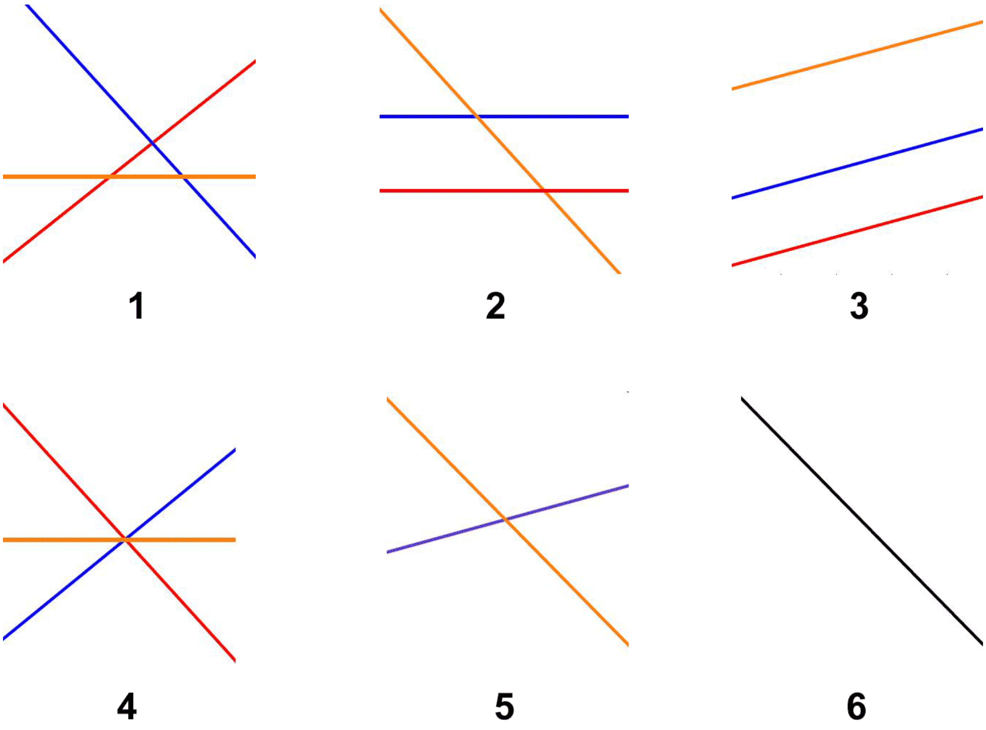


Figure 6.1: The six possible structures of the solutions for three linear equations in two unknowns.

The example with three equations in two unknowns illustrates the following theorem which follows from our study of the solution sets in the previous section, see the theorems 6.26, 6.29 and 6.33:

|||| Theorem 6.34 Remark about the Number of Solutions

A system of linear equations either has no, exactly one, or infinitely many solutions. There are no other possibilities.

6.9 The Linear Structure of the Solution set

In this section we dig a little deeper into the question about the *structure* of the solution set for systems of linear equations. It is particularly important to observe the correspondence between the solution set for an inhomogeneous system of equations and the *corresponding homogeneous system of equations*. We start by investigating the homogenous system.

6.9.1 The Properties of Homogeneous Systems of Equations

A homogenous system of linear equations of m linear equations in n unknowns is written in the form

$$\begin{aligned} a_{11}\cdot x_1 + \ a_{12}\cdot x_2 + \ldots + \ a_{1n}\cdot x_n &= 0 \\ a_{21}\cdot x_1 + \ a_{22}\cdot x_2 + \ldots + \ a_{2n}\cdot x_n &= 0 \\ &\vdots \\ a_{m1}\cdot x_1 + a_{m2}\cdot x_2 + \ldots + a_{mn}\cdot x_n &= 0 \end{aligned} \tag{6-71}$$

In the following theorem we describe an important property of the structure of the solution set for homogeneous systems of linear equations.

|||| Theorem 6.35 Solutions to a Homogeneous System of Linear Equations

Let L_{hom} denote the solution set of a homogeneous system of linear equations. Then there exists at least one solution to the system, namely the zero or *trivial* solution. If

$$\mathbf{x} = (x_1, x_2, \ldots x_n) \quad \text{and} \quad \mathbf{y} = (y_1, y_2, \ldots y_n) \tag{6-72}$$

are two arbitrary solutions, and k is an arbitrary scalar then both the sum

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots x_n + y_n) \tag{6-73}$$

and the product

$$k \cdot \mathbf{x} = (k\cdot x_1, k\cdot x_2, \ldots k\cdot x_n) \tag{6-74}$$

belong to L_{hom}.

|||| Proof

An obvious property of the system (6-71) is that $\rho(\mathbf{A}) = \rho(\mathbf{T})$ (because the right hand side consists of only zeros). Therefor the system has at least one solution - it follows from theorem 6.29. We can also immediately find a solution, viz. the zero vector, $\mathbf{0} \in \mathbb{L}^n$. That this is a solution is evident when we replace all the unknowns in the system with the number 0, then the system consists of m equations of the form $0 = 0$.

Apart from this the theorem comprises two parts that are proved separately:

1. If

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = 0 \quad \text{for every} \quad i = 1, 2, \ldots, m \tag{6-75}$$

and

$$a_{i1}y_1 + a_{i2}y_2 + \cdots + a_{in}y_n = 0 \quad \text{for every} \quad i = 1, 2, \ldots, m \tag{6-76}$$

then by addition of the two equations and a following factorization with respect to the coeficients we get

$$a_{i1}(x_1 + y_1) + a_{i2}(x_2 + y_2) + \cdots + a_{in}(x_n + y_n) = 0 \quad \text{for every} \quad i = 1, 2, \ldots, m \tag{6-77}$$

which shows that $\mathbf{x} + \mathbf{y}$ is a solution.

2. If

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = 0 \quad \text{for every} \quad i = 1, 2, \ldots, m \tag{6-78}$$

and k is an arbitrary scalar, then by multiplying both sides of the equation by k and a following factorization with respect to the coefficients we get

$$a_{i1}(k \cdot x_1) + a_{i2}(k \cdot x_2) + \cdots + a_{in}(k \cdot x_n) = 0 \quad \text{for every} \quad i = 1, 2, \ldots, m \tag{6-79}$$

which shows that $k \cdot \mathbf{x}$ is a solution.

■

|||| Remark 6.36

If you take an arbitrary number of solutions from L_{hom}, multiply these by arbitrary constants and add the products then the so-called ***linear combination*** of solutions also is a solution. This is a consequence of theorem 6.35.

6.9.2 Structural Theorem

We will now consider a decisive relation between an inhomogeneous system of linear equations of the form

$$\begin{array}{c} a_{11}\cdot x_1 + a_{12}\cdot x_2 + \ldots + a_{1n}\cdot x_n = b_1 \\ a_{21}\cdot x_1 + a_{22}\cdot x_2 + \ldots + a_{2n}\cdot x_n = b_2 \\ \vdots \\ a_{m1}\cdot x_1 + a_{m2}\cdot x_2 + \ldots + a_{mn}\cdot x_n = b_m \end{array} \tag{6-80}$$

and *the corresponding homogeneous system of linear equations*, by which we mean the equations (6-80) after all the right hand sides b_i have been replaced by 0. The solution set for the inhomogeneous system of equations is called L_{inhom} and the solution set for the corresponding homogeneous system of equations is called L_{hom}.

|||| Theorem 6.37 Structural Theorem

If you have found just one solution (a so-called ***particular*** solution) $\mathbf{x}_0$ to an inhomogeneous sytem of linear equations, then L_{inhom} can be found as the sum of $\mathbf{x}_0$ and L_{hom}.

In other words

$$L_{inhom} = \left\{ \mathbf{x} = \mathbf{x}_0 + \mathbf{y} \,\middle|\, \mathbf{y} \in L_{hom} \right\}. \tag{6-81}$$

or in short

$$L_{inhom} = \mathbf{x}_0 + L_{hom}. \tag{6-82}$$

|||| Proof

Note that the theorem contains two propositions. One is that the sum of $\mathbf{x}_0$ and an arbitrary vector from L_{hom} belongs to L_{inhom}. The other is that an arbitrary vector from L_{inhom} can be written as the sum of $\mathbf{x}_0$ and a vector from L_{hom}. We prove the two propositions separately:

1. Assume $\mathbf{y} \in L_{hom}$. We want to show that

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{y} = (x_{0_1} + y_1, x_{0_2} + y_2, \ldots, x_{0_n} + y_n) \in L_{inhom}. \tag{6-83}$$

Since

$$a_{i1}x_{0_1} + a_{i2}x_{0_2} + \cdots + a_{in}x_{0_n} = b_i \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-84}$$

and

$$a_{i1}y_1 + a_{i2}y_2 + \cdots + a_{in}y_n = 0 \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-85}$$

then by addition of the two equations and a following factorization with respect to the coeficients we get

$$a_{i1}(x_{0_1} + y_1) + \cdots + a_{in}(x_{0_n} + y_n) = b_i \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-86}$$

which proves the proposition.

2. Assume $\mathbf{x} \in L_{inhom}$. We want to show that a vector $\mathbf{y} \in L_{hom}$ exists that fulfills

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{y}. \tag{6-87}$$

Since both $\mathbf{x}$ and $\mathbf{x}_0$ belong to L_{inhom} we have that

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-88}$$

and

$$a_{i1}x_{0_1} + a_{i2}x_{0_2} + \cdots + a_{in}x_{0_n} = b_i \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-89}$$

When we subtract the lower equation from the upper, we get after factorization

$$a_{i1}(x_1 - x_{0_1}) + \cdots + a_{in}(x_n - x_{0_n}) = 0 \quad \text{for any} \quad i = 1, 2, \ldots, m \tag{6-90}$$

which shows that the vector $\mathbf{y}$ defined by $\mathbf{y} = \mathbf{x} - \mathbf{x}_0$, belongs to L_{hom} and satisfies: $\mathbf{x} = \mathbf{x}_0 + \mathbf{y}$.

■

eNote 7

Matrices and Matrix Algebra

This eNote introduces matrices and arithmetic operations for matrices and deduces the relevant arithmetic rules. Math knowledge comparable to that of a Danish gymnasium (high school) graduate is the only requirement for benefitting from this note, but it is a good idea to acquaint oneself with the number space $\mathbb{R}^n$ that is described in eNote 5 The Number Spaces*.*

(Updated: 24.09.2021 David Brander)

A ***matrix*** is an array of numbers. Here is an example of a matrix called **M**:

$$\mathbf{M} = \begin{bmatrix} 1 & 4 & 3 \\ -1 & 2 & 7 \end{bmatrix} \tag{7-1}$$

A matrix is characterized by the number of *rows* and *columns*, and the matrix **M** above is therefore called a 2×3 matrix. The matrix **M** is said to contain $2 \cdot 3 = 6$ *elements*. In addition to rows and columns a number of further concepts are connected. In order to describe these we write a general matrix, here called **A**, as:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \tag{7-2}$$

The matrix **A** has m rows and n columns, and this can indicated by writing $\mathbf{A}_{m\times n}$ or the *$m \times n$ matrix* **A**. The matrix **A** is also said to be *of the type $m \times n$*.

Two $m \times n$-matrices **A** and **B** are called *equal* if the elements in each matrix are equal, and we then write $\mathbf{A} = \mathbf{B}$.

A matrix with a single column ($n = 1$), is called a *column matrix*. Similarly a matrix with only one row ($m = 1$), a *row matrix*.

A matrix with the same number of row and columns ($m = n$), is called a *square matrix*. Square matrices are investigated in depth in eNote 8 *Square Matrices*.

If all the elements in an $m \times n$-matrix are real numbers, the matrix is called a *real matrix*. The set of these matrices is denoted $\mathbb{R}^{m \times n}$.

7.1 Matrix Sum and the Product of a Matrix by a Scalar

It is possible to add two matrices if they are of the same type. You then add the elements with the same row and column numbers and in this way form a new matrix of the same type. Similarly you can multiply any matrix by a scalar (a number), this is done by multiplying all the elements by the scalar. The matrix in which all elements are equal to 0 is called the *zero matrix* regardless of the type, and is denoted $\mathbf{0}$ or possibly $\mathbf{0}_{m \times n}$. In these notes, all other matrices are called *proper matrices*.

|||| Definition 7.1 Matrix Sum and Multiplication by a Scalar

Given a scalar $k \in \mathbb{R}$ and two real matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{m\times n}$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad \text{og} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} \tag{7-3}$$

The *sum* of the matrices is defined as:

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{bmatrix} \tag{7-4}$$

The sum is only defined when the matrices are of the same type.

The *product* of the matrix $\mathbf{A}$ by the scalar k is written $k\mathbf{A}$ or $\mathbf{A}k$ and is defined as:

$$k\mathbf{A} = \mathbf{A}k = \begin{bmatrix} k\cdot a_{11} & k\cdot a_{12} & \dots & k\cdot a_{1n} \\ k\cdot a_{21} & k\cdot a_{22} & \dots & k\cdot a_{2n} \\ \vdots & \vdots & & \vdots \\ k\cdot a_{m1} & k\cdot a_{m2} & \dots & k\cdot a_{mn} \end{bmatrix} \tag{7-5}$$

The *opposite matrix* $-\mathbf{A}$ (additive inverse) to a matrix $\mathbf{A}$ is defined by the matrix that results when all the elements in $\mathbf{A}$ are multiplied by -1. It is seen that $-\mathbf{A} = (-1)\mathbf{A}$.

|||| Example 7.2 Simple Matrix Operations

Define two matrices $\mathbf{A}$ and$\mathbf{B}$ by:

$$\mathbf{A} = \begin{bmatrix} 4 & -1 \\ 8 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} -4 & 3 \\ 9 & \frac{1}{2} \end{bmatrix} \tag{7-6}$$

The matrices are both of the type 2×2. We wish to determine a third and fourth matrix

$\mathbf{C} = 4\mathbf{A}$ and $\mathbf{D} = 2\mathbf{A} + \mathbf{B}$. This can be done through the use of the definition 7.1.

$$\begin{aligned} \mathbf{C} = 4\mathbf{A} = 4 \cdot \begin{bmatrix} 4 & -1 \\ 8 & 0 \end{bmatrix} = \begin{bmatrix} 4\cdot 4 & 4\cdot(-1) \\ 4\cdot 8 & 4\cdot 0 \end{bmatrix} = \begin{bmatrix} 16 & -4 \\ 32 & 0 \end{bmatrix} \\ \mathbf{D} = 2\mathbf{A} + \mathbf{B} = \begin{bmatrix} 8 & -2 \\ 16 & 0 \end{bmatrix} + \begin{bmatrix} -4 & 3 \\ 9 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 25 & \frac{1}{2} \end{bmatrix} \end{aligned} \tag{7-7}$$

In the following theorem we summarize the arithmetic rules that are valid for sums of matrices and multiplication by a scalar.

|||| Theorem 7.3 Arithmetic Rules for the Matrix Sum and the Product by a Scalar

For arbitrary matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ in $\mathbb{R}^{m\times n}$ and likewise arbitrary real numbers k_1 and k_2 the following arithmetic rules are valid:

1. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ Addition is commutative
2. $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$ Addition is associative
3. $\mathbf{A} + \mathbf{0} = \mathbf{A}$ $\mathbf{0}$ is a neutral matrix for addition in $\mathbb{R}^{m\times n}$
4. $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$ Every matrices in $\mathbb{R}^{m\times n}$ has an opposite matrix
5. $k_1(k_2\mathbf{A}) = (k_1k_2)\mathbf{A}$ Product of a matrix by scalars is associative

$$\left.\begin{array}{ll} 6. & (k_1 + k_2)\mathbf{A} = k_1\mathbf{A} + k_2\mathbf{A} \\ 7. & k_1(\mathbf{A} + \mathbf{B}) = k_1\mathbf{A} + k_1\mathbf{B} \end{array}\right\} \text{The distributive rules are valid}$$

8. $1\mathbf{A} = \mathbf{A}$ The scalar 1 is neutral in the product by a matrix

The arithmetic rules in Theorem 7.3 can be proved by applying the ordinary arithmetic rules for real numbers. The method is demonstrated for two of the rules in the following example.

|||| Example 7.4 Demonstration of Arithmetic Rule

Given the two matrices

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{7-8}$$

plus the constants k_1 and k_2. We now try by way of example to show the distributive rules in

Theorem 7.3. First we have:

$$\begin{aligned} (k_1+k_2)\mathbf{A} &= (k_1+k_2)\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} (k_1+k_2)a_{11} & (k_1+k_2)a_{12} \\ (k_1+k_2)a_{21} & (k_1+k_2)a_{22} \end{bmatrix} \\ k_1\mathbf{A}+k_2\mathbf{A} &= \begin{bmatrix} k_1a_{11} & k_1a_{12} \\ k_1a_{21} & k_1a_{22} \end{bmatrix} + \begin{bmatrix} k_2a_{11} & k_2a_{12} \\ k_2a_{21} & k_2a_{22} \end{bmatrix} = \begin{bmatrix} k_1a_{11}+k_2a_{11} & k_1a_{12}+k_2a_{12} \\ k_1a_{21}+k_2a_{21} & k_1a_{22}+k_2a_{22} \end{bmatrix} \end{aligned} \tag{7-9}$$

If you take a_{11}, a_{12}, a_{21} anda_{22} outside the parentheses in each of the elements in the last expression, it is seen that $(k_1+k_2)\mathbf{A} = k_1\mathbf{A}+k_2\mathbf{A}$ in this case. The operation of taking the a-elements outside the parentheses is exactly equivalent to be using the distributive rule for the real numbers.

The second distributive rule is demonstrated for given matrices and constants:

$$\begin{aligned} k_1(\mathbf{A}+\mathbf{B}) &= k_1\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{bmatrix} = \begin{bmatrix} k_1(a_{11}+b_{11}) & k_1(a_{12}+b_{12}) \\ k_1(a_{21}+b_{21}) & k_1(a_{22}+b_{22}) \end{bmatrix} \\ k_1\mathbf{A}+k_1\mathbf{B} &= \begin{bmatrix} k_1a_{11} & k_1a_{12} \\ k_1a_{21} & k_1a_{22} \end{bmatrix} + \begin{bmatrix} k_1b_{11} & k_1b_{12} \\ k_1b_{21} & k_1b_{22} \end{bmatrix} = \begin{bmatrix} k_1a_{11}+k_1b_{11} & k_1a_{12}+k_1b_{12} \\ k_1a_{21}+k_1b_{21} & k_1a_{22}+k_1b_{22} \end{bmatrix} \end{aligned} \tag{7-10}$$

If k_1 is taken outside of the parenthesis in each of the elements in the matrix in the last expression it is seen that the second distributive rule also is valid in this case: $k_1(\mathbf{A}+\mathbf{B}) = k_1\mathbf{A}+k_1\mathbf{B}$. The distributive rule for real numbers is again used for each element.

Note that the zero matrix in $\mathbb{R}^{m\times n}$ is the only matrix $\mathbb{R}^{m\times n}$ that is neutral with respect to addition, and that $-\mathbf{A}$ is the only solution to the equation $\mathbf{A}+\mathbf{X}=\mathbf{0}$.

|||| Definition 7.5 Difference Between Matrices

The difference $\mathbf{A}-\mathbf{B}$ between two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same type is introduced by:

$$\mathbf{A}-\mathbf{B} = \mathbf{A}+(-\mathbf{B}). \tag{7-11}$$

In other words $\mathbf{B}$ is subtracted from $\mathbf{A}$ by subtracting each element in $\mathbf{B}$ from the corresponding element in $\mathbf{A}$.

|||| Example 7.6 Simple Matrix Operation with Difference

With the matrices given in Example 7.2 we get

$$\mathbf{D} = 2\mathbf{A} - \mathbf{B} = 2\mathbf{A} + (-1)\mathbf{B} = \begin{bmatrix} 8 & -2 \\ 16 & 0 \end{bmatrix} + \begin{bmatrix} 4 & -3 \\ -9 & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 12 & -5 \\ 7 & -\frac{1}{2} \end{bmatrix} \tag{7-12}$$

7.2 Matrix-Vector Products and Matrix-Matrix Products

In this subsection we describe the multiplication of a matrix with a vector and then the multiplication of matrix by another matrix.

A vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ can be written as a column matrix and is then called a *column vector*:

$$\mathbf{v} = (v_1, v_2, \ldots, v_n) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{7-13}$$

Using this concept you can divide a matrix $\mathbf{A}_{m \times n}$ into its column vectors. This is written like this:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_n \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} & \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} & \ldots & \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \end{aligned} \tag{7-14}$$

Accordingly there are n column vectors with m elements each.

Notice that the square brackets around the column vectors can be removed just like that! This can be done in all dealings with matrices, where double square brackets occur. It is always the innermost brackets that are removed. In this way there is no difference between the two expressions. The last expression is always preferred, because it is the easier to read.

We now define the product of a matrix and a vector, in which the matrix has as many columns as the vector has elements:

|||| Definition 7.7 Matrix-Vector Product

Let $\mathbf{A}$ be an arbitrary matrix in $\mathbb{R}^{m\times n}$, and let $\mathbf{v}$ be an arbitrary vector in $\mathbb{R}^n$.

The *matrix-vector product* of $\mathbf{A}$ with $\mathbf{v}$ is defined as:

$$\mathbf{Av} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + \ldots + v_n\mathbf{a}_n \end{bmatrix} \tag{7-15}$$

The result is a column vector with m elements. It is the sum of the products of the k'th column in the matrix and the k'th element in the column vector for all $k = 1, 2, \ldots, n$.

It is necessary that there are as many columns in the matrix as there are rows in the column vector, here n.

Notice the order in the matrix-vector product: first matrix, then vector! It is not a vector-matrix product so to speak. The number of rows and columns will not match in the other configuration unless the matrix is of the type 1 × 1.

|||| Example 7.8 Matrix-Vector Product

The following matrix and vector (a column vector) are given:

$$\mathbf{A} = \mathbf{A}_{2\times 3} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} 3 \\ 4 \\ -1 \end{bmatrix}. \tag{7-16}$$

We now form the matrix-vector product of $\mathbf{A}$ with $\mathbf{v}$ by use of definition 7.7:

$$\mathbf{Av} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} 3 \\ 4 \\ -1 \end{bmatrix} = \left[3\begin{bmatrix} a \\ d \end{bmatrix} + 4\begin{bmatrix} b \\ e \end{bmatrix} + (-1)\begin{bmatrix} c \\ f \end{bmatrix} \right] = \begin{bmatrix} 3a + 4b - c \\ 3d + 4e - f \end{bmatrix} \tag{7-17}$$

If $\mathbf{A}$ is given like this

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & 6 \\ 2 & 1 & 4 \end{bmatrix}, \tag{7-18}$$

you get the product

$$\mathbf{Av} = \begin{bmatrix} 3\cdot(-1) + 4\cdot 2 - 6 \\ 3\cdot 2 + 4\cdot 1 - 4 \end{bmatrix} = \begin{bmatrix} -1 \\ 6 \end{bmatrix} \tag{7-19}$$

It is seen that the result (in both cases) is a column vector with as many rows as there are rows in $\mathbf{A}$.

|||| Exercise 7.9 Matrix-Vector Product

Form the matrix-vector product $\mathbf{A}$ with $\mathbf{x}$ in the equation $\mathbf{Ax} = \mathbf{b}$, when it is given that

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad , \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{7-20}$$

Is this something you have seen before? From where does it come?

As we have remarked above a matrix can be viewed as a number of column vectors aligned after one another. This is used in the following definition of a matrix-matrix product as a series of matrix-vector products.

|||| Definition 7.10 Matrix-Matrix Product

Let $\mathbf{A}$ be an arbitrary matrix in $\mathbb{R}^{m \times n}$, and let $\mathbf{B}$ be an arbitrary matrix in $\mathbb{R}^{n \times p}$.

The *matrix-matrix product* or just the ***matrix product*** of $\mathbf{A}$ and $\mathbf{B}$ is defined like this:

$$\mathbf{AB} = \mathbf{A}\begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \ldots & \mathbf{b}_p \end{bmatrix} = \begin{bmatrix} \mathbf{Ab}_1 & \mathbf{Ab}_2 & \ldots & \mathbf{Ab}_p \end{bmatrix} \tag{7-21}$$

The result is matrix of type $m \times p$. The k'th column in the resulting matrix is a matrix-vector product of the first matrix (here $\mathbf{A}$) and the k'th column vector in the last matrix (here $\mathbf{B}$), cf. definition 7.7.

There must be as many columns in the first matrix as there are rows in the last matrix.

|||| Example 7.11 Matrix-Matrix Product

Given two matrices $\mathbf{A}_{2\times2}$ and $\mathbf{B}_{2\times3}$:

$$\mathbf{A} = \begin{bmatrix} 4 & 5 \\ 1 & 2 \end{bmatrix} \quad \text{og} \quad \mathbf{B} = \begin{bmatrix} -8 & 3 & 3 \\ 2 & 9 & -9 \end{bmatrix} \tag{7-22}$$

We wish to form the matrix-matrix product of **A** and **B**. This is done by use of definition 7.10.

$$\begin{aligned}\mathbf{AB} &= \left[\begin{bmatrix}4&5\\1&2\end{bmatrix}\begin{bmatrix}-8\\2\end{bmatrix}\ \begin{bmatrix}4&5\\1&2\end{bmatrix}\begin{bmatrix}3\\9\end{bmatrix}\ \begin{bmatrix}4&5\\1&2\end{bmatrix}\begin{bmatrix}3\\-9\end{bmatrix}\right]\\ &= \begin{bmatrix}4\cdot(-8)+5\cdot 2 & 4\cdot 3+5\cdot 9 & 4\cdot 3+5\cdot(-9)\\ -8+2\cdot 2 & 3+2\cdot 9 & 3+2\cdot(-9)\end{bmatrix} = \begin{bmatrix}-22&57&-33\\-4&21&-15\end{bmatrix}\end{aligned} \tag{7-23}$$

NB: It is *not* possible to form the matrix-matrix product **BA**, because there are not as many columns in **B** as there are rows in **A** ($3 \neq 2$).

|||| Example 7.12 Matrix-Matrix Product two Ways

Given the two matrices $\mathbf{A}_{2\times 2}$ and $\mathbf{B}_{2\times 2}$:

$$\mathbf{A} = \begin{bmatrix}3&2\\-5&1\end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix}4&4\\-1&0\end{bmatrix} \tag{7-24}$$

Because the two matrices are square matrices of the same type both matrix-matrix products **AB** and **BA** can be calculated. We use the definition 7.10.

$$\begin{aligned}\mathbf{AB} &= \left[\begin{bmatrix}3&2\\-5&1\end{bmatrix}\begin{bmatrix}4\\-1\end{bmatrix}\ \begin{bmatrix}3&2\\-5&1\end{bmatrix}\begin{bmatrix}4\\0\end{bmatrix}\right]\\ &= \begin{bmatrix}3\cdot 4+2\cdot(-1) & 3\cdot 4+2\cdot 0\\ -5\cdot 4+1\cdot(-1) & -5\cdot 4+1\cdot 0\end{bmatrix} = \begin{bmatrix}10&12\\-21&-20\end{bmatrix}\\ \mathbf{BA} &= \left[\begin{bmatrix}4&4\\-1&0\end{bmatrix}\begin{bmatrix}3\\-5\end{bmatrix}\ \begin{bmatrix}4&4\\-1&0\end{bmatrix}\begin{bmatrix}2\\1\end{bmatrix}\right]\\ &= \begin{bmatrix}4\cdot 3+4\cdot(-5) & 4\cdot 2+4\\ -1\cdot 3 & -1\cdot 2\end{bmatrix} = \begin{bmatrix}-8&12\\-3&-2\end{bmatrix}\end{aligned} \tag{7-25}$$

We see that $\mathbf{AB} \neq \mathbf{BA}$. The factors are **not** interchangeable!

Here we summarize the arithmetic rules that apply to matrix-matrix products and matrix sums. Because the matrix-vector product is a special case of the matrix-matrix product, the rules also apply for matrix-vector products.

|||| Theorem 7.13 Arithmetic Rules for Matrix Sum and Product

For arbitrary matrices **A**, **B** and **C** and likewise an arbitrary real number k the following arithmetic rules are valid, in so far as the matrix-matrix products can be formed:

$$\begin{array}{rl} (k\mathbf{A})\mathbf{B} = \mathbf{A}(k\mathbf{B}) = k(\mathbf{AB}) & \text{Product with a scalar is associative} \\ \left.\begin{array}{r} \mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \\ (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC} \end{array}\right\} & \text{the distributive rules apply} \\ \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} & \text{Matrix-matrix products are associative} \end{array}$$

Analogous to the demonstration of the arithmetic rules in Theorem 7.3 we demonstrate the last arithmetic rule in Theorem 7.13:

|||| Example 7.14 Are Matrix Products Associative?

The last arithmetic rule in 7.13 is tested on the three matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \;,\quad \mathbf{B} = \begin{bmatrix} -3 & -2 & -1 \\ 0 & 0 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 4 & -5 \\ 2 & 1 \\ 1 & -3 \end{bmatrix} \tag{7-26}$$

First we calculate **AB** and **BC**:

$$\begin{aligned} \mathbf{AB} &= \left[\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} -3 \\ 0 \end{bmatrix} \; \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} -2 \\ 0 \end{bmatrix} \; \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} -1 \\ 7 \end{bmatrix} \right] = \begin{bmatrix} -3 & -2 & 13 \\ -9 & -6 & 25 \end{bmatrix} \\ \mathbf{BC} &= \left[\begin{bmatrix} -3 & -2 & -1 \\ 0 & 0 & 7 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix} \; \begin{bmatrix} -3 & -2 & -1 \\ 0 & 0 & 7 \end{bmatrix}\begin{bmatrix} -5 \\ 1 \\ -3 \end{bmatrix} \right] = \begin{bmatrix} -17 & 16 \\ 7 & -21 \end{bmatrix} \end{aligned} \tag{7-27}$$

Then we calculate $\mathbf{A}(\mathbf{BC})$ and $(\mathbf{AB})\mathbf{C}$:

$$\begin{aligned} \mathbf{A}(\mathbf{BC}) &= \left[\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} -17 \\ 7 \end{bmatrix} \; \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 16 \\ -21 \end{bmatrix} \right] = \begin{bmatrix} -3 & -26 \\ -23 & -36 \end{bmatrix} \\ (\mathbf{AB})\mathbf{C} &= \left[\begin{bmatrix} -3 & -2 & 13 \\ -9 & -6 & 25 \end{bmatrix}\begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix} \; \begin{bmatrix} -3 & -2 & 13 \\ -9 & -6 & 25 \end{bmatrix}\begin{bmatrix} -5 \\ 1 \\ -3 \end{bmatrix} \right] = \begin{bmatrix} -3 & -26 \\ -23 & -36 \end{bmatrix} \end{aligned} \tag{7-28}$$

We see that $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$, and therefore it doesn't matter which of the matrix products **AB** and **BC** we calculate first. This is valid for all matrices (although not proved here).

As is done in example 7.14 we can demonstrate the other arithmetic rules. By writing down carefully the formula for each element of a matrix in the final product, in terms of the elements of the other matrices, one can prove the rules properly.

|||| Exercise 7.15 Demonstration of Arithmetic Rule

Demonstrate the first arithmetic rule in Theorem 7.13 with two real matrices $\mathbf{A}_{2\times 2}$ and $\mathbf{B}_{2\times 2}$ and the constant k.

7.3 Transpose of a Matrix

By interchanging rows and columns in a matrix the ***transpose matrix*** is formed as in this example:

$$\mathbf{A}=\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \quad \text{has the transpose} \quad \mathbf{A}^\top=\begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix} \tag{7-29}$$

$\mathbf{A}^\top$ is '$\mathbf{A}$ *transpose*'. In addition you have that $(\mathbf{A}^\top)^\top = \mathbf{A}$. Here is a useful arithmetic rule for the transpose of a matrix-matrix product.

|||| Theorem 7.16 Transpose of a Matrix

Let there be given two arbitrary matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times p}$. You form the transposed matrices , $\mathbf{A}^\top$ and $\mathbf{B}^\top$ respectively, by interchanging the columns and rows of each matrix.

The transpose of a matrix-matrix product $\mathbf{AB}$ is equal to the matrix-matrix product of $\mathbf{B}^\top$ with $\mathbf{A}^\top$ (that is, in reverse order):

$$(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top \tag{7-30}$$

In the following example Theorem 7.16 is tested.

|||| Example 7.17 Demonstration of Theorem 7.16

Given the two matrices $\mathbf{A} = \begin{bmatrix} 0 & 1 & 6 \\ 7 & -3 & 2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 9 & 1 \\ 1 & 0 \\ -6 & 3 \end{bmatrix}$. Then

$$\begin{aligned} \mathbf{AB} &= \left[\begin{bmatrix} 0 & 1 & 6 \\ 7 & -3 & 2 \end{bmatrix} \begin{bmatrix} 9 \\ 1 \\ -6 \end{bmatrix} \ \begin{bmatrix} 0 & 1 & 6 \\ 7 & -3 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix} \right] \\ &= \begin{bmatrix} 1\cdot 1 - 6\cdot 6 & 6\cdot 3 \\ 7\cdot 9 - 3\cdot 1 - 2\cdot 6 & 7\cdot 1 + 2\cdot 3 \end{bmatrix} = \begin{bmatrix} -35 & 18 \\ 48 & 13 \end{bmatrix} \end{aligned} \tag{7-31}$$

We now try to form the matrix-matrix product $\mathbf{B}^\top \mathbf{A}^\top$ and we find

$$\mathbf{A}^\top = \begin{bmatrix} 0 & 7 \\ 1 & -3 \\ 6 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}^\top = \begin{bmatrix} 9 & 1 & -6 \\ 1 & 0 & 3 \end{bmatrix} \tag{7-32}$$

and then

$$\begin{aligned} \mathbf{B}^\top \mathbf{A}^\top &= \left[\begin{bmatrix} 9 & 1 & -6 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 6 \end{bmatrix} \ \begin{bmatrix} 9 & 1 & -6 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 7 \\ -3 \\ 2 \end{bmatrix} \right] \\ &= \begin{bmatrix} 1\cdot 1 - 6\cdot 6 & 9\cdot 7 - 1\cdot 3 - 6\cdot 2 \\ 3\cdot 6 & 1\cdot 7 + 3\cdot 2 \end{bmatrix} = \begin{bmatrix} -35 & 48 \\ 18 & 13 \end{bmatrix} \end{aligned} \tag{7-33}$$

The two results look identical:

$$\begin{bmatrix} -35 & 18 \\ 48 & 13 \end{bmatrix}^\top = \begin{bmatrix} -35 & 48 \\ 18 & 13 \end{bmatrix} \quad \Leftrightarrow \quad (\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top , \tag{7-34}$$

in agreement with Theorem 7.16

|||| Exercise 7.18 Matrix Product and the Transpose

Given the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \end{bmatrix} \tag{7-35}$$

Calculate if possible the following:

a) $2\mathbf{A} - 3\mathbf{B}$, b) $2\mathbf{A}^\top - 3\mathbf{B}^\top$, c) $2\mathbf{A} - 3\mathbf{B}^\top$, d) $\mathbf{AB}$,

e) $\mathbf{AB}^\top$, f) $\mathbf{BA}^\top$, g) $\mathbf{B}^\top\mathbf{A}$, h) $\mathbf{A}^\top\mathbf{B}$.

7.4 Summary

- Matrices are arrays characterized by the number of *columns* and *rows*, determining the *type* of the matrix. An entry in the matrix is called an *element*.
- The type of a matrix is denoted as: $\mathbf{A}_{m\times n}$. The matrix $\mathbf{A}$ has m rows and n columns.
- Matrices can be multiplied by a scalar by multiplying each element in the matrix by the scalar.
- Matrices can be added if they are of the same type. This is done by adding corresponding elements.
- The matrix-vector product, of $\mathbf{A}_{m\times n}$ with the vector $\mathbf{v}$ with n elements, is defined as:

$$\mathbf{A}_{m\times n}\mathbf{v} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 v_1 + \mathbf{a}_2 v_2 + \dots + \mathbf{a}_n v_n \end{bmatrix}, \tag{7-36}$$

  where $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are the *column vectors* in $\mathbf{A}$.

- The matrix-matrix product (or just the matrix product) is defined as a series of matrix-vector products:

$$\mathbf{A}_{m\times n}\mathbf{B}_{n\times p} = \mathbf{A}\begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_p \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{b}_1 & \mathbf{A}\mathbf{b}_2 & \dots & \mathbf{A}\mathbf{b}_p \end{bmatrix} \tag{7-37}$$

- More arithmetic rules for matrix sums, matrix products and matrix-scalar products are found in Theorem 7.3 and Theorem 7.13.
- The *transpose* $\mathbf{A}^\top$ of a matrix $\mathbf{A}$ is determined by interchanging rows and columns in the matrix.

eNote 8

Square Matrices

In this eNote we explore the basic characteristics of the set of square matrices and introduce the notion of the inverse of certain square matrices. We presume that the reader has a knowledge of basic matrix operations, see e.g. eNote 7, Matrices and Matrix Algebra*.*

(Updated 24.9.2021 David Brander).

Square matrices are simply matrices with *equal number of rows and columns*, that is they are of the type $n \times n$. This note will introduce some of the basic operations that apply to square matrices.

A square $n \times n$ matrix $\mathbf{A}$ looks like this:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{8-1}$$

The elements $a_{11}, a_{22}, \dots, a_{nn}$ are said to be placed on the ***main diagonal*** or just the *diagonal* of $\mathbf{A}$.

A square matrix $\mathbf{D}$, the non-zero elements of which lie exclusively on the main diagonal, is termed a ***diagonal matrix***, and one can write $\mathbf{D} = \mathbf{diag}(a_{11}, a_{22}, \dots, a_{nn})$.

A ***symmetric matrix*** $\mathbf{A}$ is a square matrix that is equal to its own transpose, thus $\mathbf{A} = \mathbf{A}^\top$.

The square matrix with 1's in the main diagonal and zeroes elsewhere, is called the

identity matrix regardless of the number of rows and columns. The identity matrix is here denoted **E**, (more commonly in the literature as **I**). Accordingly

$$\mathbf{E} = \mathbf{E}_{n\times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \tag{8-2}$$

Internationally accepted usage is to denote the identity matrix by **I**.

Theorem 8.1 Identity Matrix

The identity matrix **E** in $\mathbb{R}^{n\times n}$ is the only matrix in $\mathbb{R}^{n\times n}$ that satisfies the following relations:

$$\mathbf{AE} = \mathbf{EA} = \mathbf{A} \tag{8-3}$$

for an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Proof

Suppose another matrix **D**, satisfies the same relations as **E**, that is $\mathbf{AD} = \mathbf{DA} = \mathbf{A}$ for all $n \times n$ matrices **A**. This arbitrary matrix **A** could be the identity matrix, combining the two equations we get : $\mathbf{D} = \mathbf{ED} = \mathbf{DE} = \mathbf{E}$.

Since $\mathbf{E} = \mathbf{D}$ there is no other matrix than the identity matrix **E** that can be a neutral element for the matrix product.

■

The identity matrix can be regarded as the ”1 for matrices": A scalar is not altered by multiplication by 1, likewise a matrix is not altered by the matrix product of the matrix with the identity matrix of the same type.

As is evident from the following it is often crucial for the manipulation of square matrices whether they have full rank or not. Therefore we now introduce a special concept to express this.

Definition 8.2 Invertible and Singular Matrix

A square matrix is called *regular* or *non-singular* (or ***invertible***) if it is of full rank, that is $\rho(\mathbf{A}_{n\times n}) = n$.

A square matrix is called ***singular*** if it not of full rank, that is $\rho(\mathbf{A}_{n\times n}) < n$.

8.1 Inverse Matrix

The reciprocal of a scalar $a \neq 0$ satisfies the following equation: $a \cdot x = 1$, where x is the reciprocal. This can be rewritten as $x = a^{-1}$. This idea will now be generalized to square matrices. Notice that you cannot determine the reciprocal of a scalar a if $a = 0$. A similar exception emerges when we generalize to square matrices.

In order to determine the "reciprocal matrix" to a matrix $\mathbf{A}$, termed the inverse matrix, a *matrix equation* similar to $a \cdot x = 1$ for a scalar:

$$\mathbf{AX} = \mathbf{XA} = \mathbf{E} \tag{8-4}$$

The unknown $\mathbf{X}$ is a matrix. If there is a solution $\mathbf{X}$, it is denoted $\mathbf{A}^{-1}$ and is called the ***inverse matrix*** to $\mathbf{A}$. Hence we wish to find a certain matrix called $\mathbf{A}^{-1}$, for which the matrix product of $\mathbf{A}$ with this matrix yields the identity matrix.

It is not all square matrices that possess an inverse. This is postulated in the following theorem.

Theorem 8.3 Inverse Matrix

A square matrix $\mathbf{A}_{n\times n}$ has an inverse matrix $\mathbf{A}^{-1}$, that satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{E}$, if and only if $\mathbf{A}$ is non-singular.

The inverse matrix is uniquely determined by the solution of the matrix equation $\mathbf{AX} = \mathbf{E}$, where $\mathbf{X}$ is the unknown.

Note: this is why a non-singular square matrix is also called *invertible*.

In the following method it is explained, how the matrix equation described above (8-4) is solved, and thus how the inverse of an invertible matrix is found.

|||| Method 8.4 Determining the Inverse Matrix

You determine the inverse matrix denoted $\mathbf{A}^{-1}$, for the invertible square matrix $\mathbf{A}$ by use of the *matrix equation*

$$\mathbf{A}\mathbf{X} = \mathbf{E}. \tag{8-5}$$

The equation is solved with respect to the unknown $\mathbf{X}$ in the following way:

1. The augmented matrix $\mathbf{T} = \left[\,\mathbf{A} \mid \mathbf{E}\,\right]$ is formed.
2. By ordinary Gauss-Jordan elimination the reduced row echelon form $\mathrm{rref}(\mathbf{T})$ of $\mathbf{T}$ is determined.
3. In the elimination process the identity matrix is finally formed on the left hand side of the vertical line, while the solution (the inverse of $\mathbf{A}$) can be read on the right hand side: $\mathrm{rref}(\mathbf{T}) = \left[\,\mathbf{E} \mid \mathbf{X}\,\right] = \left[\,\mathbf{E} \mid \mathbf{A}^{-1}\,\right]$.

|||| Example 8.5 Inverse Matrix

We wish to find the inverse matrix $\mathbf{A}^{-1}$ to the matrix $\mathbf{A}$, given in this way:

$$\mathbf{A} = \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \tag{8-6}$$

This can be done using method 8.4. First the augmented matrix is formed:

$$\mathbf{T} = \left[\begin{array}{ccc|ccc} -16 & 9 & -10 & 1 & 0 & 0 \\ 9 & -5 & 6 & 0 & 1 & 0 \\ 2 & -1 & 1 & 0 & 0 & 1 \end{array}\right] \tag{8-7}$$

Now we form the leading 1 in the first row: First the row operation $R_1 + R_2$ and then $R_1 + 4 \cdot R_3$. This yields

$$\left[\begin{array}{ccc|ccc} -7 & 4 & -4 & 1 & 1 & 0 \\ 9 & -5 & 6 & 0 & 1 & 0 \\ 2 & -1 & 1 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 1 & 4 \\ 9 & -5 & 6 & 0 & 1 & 0 \\ 2 & -1 & 1 & 0 & 0 & 1 \end{array}\right] \tag{8-8}$$

Then the numbers in the 1st column of the 2nd and 3rd row are eliminated: $R_2 - 9 \cdot R_1$ and

$R_3 - 2 \cdot R_1$. Furthermore the 2nd and 3rd rows are swapped: $R_2 \leftrightarrow R_3$. We then get

$$\left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1 & 4 \\ 0 & -5 & 6 & -9 & -8 & -36 \\ 0 & -1 & 1 & -2 & -2 & -7 \end{array}\right] \rightarrow \left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1 & 4 \\ 0 & -1 & 1 & -2 & -2 & -7 \\ 0 & -5 & 6 & -9 & -8 & -36 \end{array}\right] \tag{8-9}$$

Now we change the sign in row 2: $(-1) \cdot R_2$ and then we eliminate the number in the 2nd column of the 3rd row: $R_3 + 5 \cdot R_2$.

$$\left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1 & 4 \\ 0 & 1 & -1 & 2 & 2 & 7 \\ 0 & -5 & 6 & -9 & -8 & -36 \end{array}\right] \rightarrow \left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1 & 4 \\ 0 & 1 & -1 & 2 & 2 & 7 \\ 0 & 0 & 1 & 1 & 2 & -1 \end{array}\right] \tag{8-10}$$

The last step is then evident: $R_2 + R_3$:

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & 1 & 4 \\ 0 & 1 & 0 & 3 & 4 & 6 \\ 0 & 0 & 1 & 1 & 2 & -1 \end{array}\right] \tag{8-11}$$

We see that $\rho(\mathbf{A}) = \rho(\mathbf{T}) = 3$, thus $\mathbf{A}$ is of full rank, and therefore one can read the inverse to $\mathbf{A}$ on the right hand side of the vertical line:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 1 & 4 \\ 3 & 4 & 6 \\ 1 & 2 & -1 \end{bmatrix} \tag{8-12}$$

Notice that the left hand side of the augmented matrix is the identity matrix. It is so to speak "moved" from the right to the left hand side of the equality signs (the vertical line).

Finally we check whether $\mathbf{A}^{-1}$, as is expected, satisfies $\mathbf{A}\mathbf{A}^{-1} = \mathbf{E}$ and$\mathbf{A}^{-1}\mathbf{A} = \mathbf{E}$:

$$\begin{aligned} \mathbf{A}\mathbf{A}^{-1} &= \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 4 \\ 3 & 4 & 6 \\ 1 & 2 & -1 \end{bmatrix} \\ &= \left[\begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \; \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \; \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \\ -1 \end{bmatrix} \right] \\ &= \begin{bmatrix} -16+27-10 & -16+36-20 & -64+54+10 \\ 9-15+6 & 9-20+12 & 36-30-6 \\ 2-3+1 & 2-4+2 & 8-6-1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{E} \end{aligned} \tag{8-13}$$

It is true! By use of the same procedure it is seen that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{E}$ is also true.

As can be seen in the next example, the inverse can be used in the solution of matrix equations with square matrices. In matrix equations one can *interchange terms* and *multiply by scalars* in order to isolate the unknown just as one would do in ordinary scalar equations. Moreover one can *multiply all terms* by matrices – this can be done either from right or the left on all terms in the equation, yielding different results.

|||| Example 8.6 Matrix Equation

We wish to solve the matrix equation

$$\mathbf{A}\mathbf{X} = \mathbf{B} - \mathbf{C}\mathbf{X} \tag{8-14}$$

where

$$\mathbf{A} = \begin{bmatrix} -4 & 2 & -1 \\ 9 & 5 & -5 \\ 2 & 0 & 7 \end{bmatrix} \;,\quad \mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ 8 & -12 & 5 \\ 5 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} -12 & 7 & -9 \\ 0 & -10 & 11 \\ 0 & -1 & -6 \end{bmatrix} \tag{8-15}$$

First the equation is reduced as far as possible, see e.g. Theorem 7.13, without using the values:

$$\mathbf{A}\mathbf{X} = \mathbf{B} - \mathbf{C}\mathbf{X} \;\Leftrightarrow\; \mathbf{A}\mathbf{X} + \mathbf{C}\mathbf{X} = \mathbf{B} - \mathbf{C}\mathbf{X} + \mathbf{C}\mathbf{X} \;\Leftrightarrow\; (\mathbf{A}+\mathbf{C})\mathbf{X} = \mathbf{B} \tag{8-16}$$

Since $\mathbf{X}$ is the unknown we try to isolate this matrix totally. If $(\mathbf{A}+\mathbf{C})$ is an invertible matrix, one can multiply by the inverse to $(\mathbf{A}+\mathbf{C})$ from the left on both sides of the equality sign. Thus:

$$(\mathbf{A}+\mathbf{C})^{-1}(\mathbf{A}+\mathbf{C})\mathbf{X} = (\mathbf{A}+\mathbf{C})^{-1}\mathbf{B} \;\Leftrightarrow\; \mathbf{E}\mathbf{X} = \mathbf{X} = (\mathbf{A}+\mathbf{C})^{-1}\mathbf{B}\,, \tag{8-17}$$

because $(\mathbf{A}+\mathbf{C})^{-1}(\mathbf{A}+\mathbf{C}) = \mathbf{E}$ according to the definition of inverse matrices. We now form $\mathbf{A}+\mathbf{C}$ and determine whether the matrix is invertible:

$$\mathbf{A}+\mathbf{C} = \begin{bmatrix} -4 & 2 & -1 \\ 9 & 5 & -5 \\ 2 & 0 & 7 \end{bmatrix} + \begin{bmatrix} -12 & 7 & -9 \\ 0 & -10 & 11 \\ 0 & -1 & -6 \end{bmatrix} = \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix} \tag{8-18}$$

The inverse of this matrix is already determined in Example 8.5, and this part of the procedure is therefor skipped. $\mathbf{X}$ is determined as:

$$\begin{aligned} \mathbf{X} = (\mathbf{A}+\mathbf{C})^{-1}\mathbf{B} &= \begin{bmatrix} -16 & 9 & -10 \\ 9 & -5 & 6 \\ 2 & -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 \\ 8 & -12 & 5 \\ 5 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 4 \\ 3 & 4 & 6 \\ 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 8 & -12 & 5 \\ 5 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 28 & -11 & 5 \\ 62 & -45 & 20 \\ 11 & -23 & 10 \end{bmatrix} \end{aligned} \tag{8-19}$$

In the further investigation of the invertibility of the transpose or the inverse of an invertible matrix plus the invertibility of the product of two or more invertible matrices we will need the following corollary, which is stated without proof (see eNote 9, in particular, Theorem 9.20 for one way to prove it).

|||| Lemma 8.7 Inherited Invertibility

1. If $\mathbf{A}$ is an invertible square matrix, both $\mathbf{A}^\top$ and $\mathbf{A}^{-1}$ are invertible.

2. The product $\mathbf{A}\,\mathbf{B}$ of two square matrices is invertible if and only if both $\mathbf{A}$ and $\mathbf{B}$ are invertible.

We can now give arithmetic rules for inverse matrices.

|||| Theorem 8.8 Arithmetic Rules for Inverse Matrices

For the invertible square matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ the following arithmetic rules apply:

1. The inverse of the inverse of a matrix is equal to the matrix itself:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A} \tag{8-20}$$

2. The transpose of an inverse matrix is equal to the inverse of the transpose of the matrix:

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top \tag{8-21}$$

   $\mathbf{A}^\top$ is invertible if and only if $\mathbf{A}$ is invertible.

3. In matrix equations we can multiply all terms by the inverse of a matrix. This can be done either from the right or the left hand side on both sides of the equality sign:

$$\mathbf{A}\mathbf{X} = \mathbf{B} \Leftrightarrow \mathbf{X} = \mathbf{A}^{-1}\mathbf{B} \quad \text{and} \quad \mathbf{X}\mathbf{C} = \mathbf{D} \Leftrightarrow \mathbf{X} = \mathbf{D}\mathbf{C}^{-1} \tag{8-22}$$

4. The inverse of a matrix product of two matrices is equal to the product of the corresponding inverse matrices in reverse order:

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{8-23}$$

All the arithmetic rules in theorem 8.8 are easily proven by checking.

Below one of the rules is tested in an example. The arithmetic rule in equation (8-22) has already been used in example 8.6.

|||| Example 8.9 Checking of Arithmetic Rule for an Inverse Matrix

Two square matrices are given

$$\mathbf{A}=\begin{bmatrix}2&4\\6&10\end{bmatrix}\quad\text{and}\quad\mathbf{B}=\begin{bmatrix}1&1\\2&3\end{bmatrix}\tag{8-24}$$

We wish to test the last arithmetic rule in theorem 8.8, viz. that $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$. First $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ are determined by use of method 8.4.

$$\left[\,\mathbf{A}\mid\mathbf{E}\,\right]=\left[\begin{array}{cc|cc}2&4&1&0\\6&10&0&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&2&\frac{1}{2}&0\\0&-2&-3&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&0&-\frac{5}{2}&1\\0&1&\frac{3}{2}&-\frac{1}{2}\end{array}\right]\tag{8-25}$$

Similarly with $\mathbf{B}$:

$$\left[\,\mathbf{B}\mid\mathbf{E}\,\right]=\left[\begin{array}{cc|cc}1&1&1&0\\2&3&0&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&1&1&0\\0&1&-2&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&0&3&-1\\0&1&-2&1\end{array}\right]\tag{8-26}$$

Since we have obtained the identity matrix on the left hand side of the vertical line in both cases , we get

$$\mathbf{A}^{-1}=\begin{bmatrix}-\frac{5}{2}&1\\\frac{3}{2}&-\frac{1}{2}\end{bmatrix}\quad\text{and}\quad\mathbf{B}^{-1}=\begin{bmatrix}3&-1\\-2&1\end{bmatrix}\tag{8-27}$$

$\mathbf{B}^{-1}\mathbf{A}^{-1}$ is determined:

$$\mathbf{B}^{-1}\mathbf{A}^{-1}=\left[\begin{bmatrix}3&-1\\-2&1\end{bmatrix}\begin{bmatrix}-\frac{5}{2}\\\frac{3}{2}\end{bmatrix}\;\begin{bmatrix}3&-1\\-2&1\end{bmatrix}\begin{bmatrix}1\\-\frac{1}{2}\end{bmatrix}\right]=\begin{bmatrix}-9&\frac{7}{2}\\\frac{13}{2}&-\frac{5}{2}\end{bmatrix}\tag{8-28}$$

On the other side of the equality sign in the arithmetic rule we first calculate $\mathbf{AB}$:

$$\mathbf{AB}=\left[\begin{bmatrix}2&4\\6&10\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix}\;\begin{bmatrix}2&4\\6&10\end{bmatrix}\begin{bmatrix}1\\3\end{bmatrix}\right]=\begin{bmatrix}10&14\\26&36\end{bmatrix}\tag{8-29}$$

Now the inverse of $\mathbf{AB}$ is determined:

$$\left[\,\mathbf{AB}\mid\mathbf{E}\,\right]=\left[\begin{array}{cc|cc}10&14&1&0\\26&36&0&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&\frac{7}{5}&\frac{1}{10}&0\\0&-\frac{2}{5}&-\frac{13}{5}&1\end{array}\right]\rightarrow\left[\begin{array}{cc|cc}1&0&-9&\frac{7}{2}\\0&1&\frac{13}{2}&-\frac{5}{2}\end{array}\right]\tag{8-30}$$

Finally we arrive at

$$(\mathbf{AB})^{-1}=\begin{bmatrix}-9&\frac{7}{2}\\\frac{13}{2}&-\frac{5}{2}\end{bmatrix},\tag{8-31}$$

Comparison of equations (8-28) and (8-31) yields the identity: $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$.

|||| Exercise 8.10 Inverse Matrix

Given the (not square!) matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 3 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \tag{8-32}$$

a) Determine $(\mathbf{BA})^{-1}$.

b) Show that $\mathbf{AB}$ is not invertible and therefore one cannot determine $(\mathbf{AB})^{-1}$.

|||| Exercise 8.11 Inverse Matrix

Given the matrices

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} -1 & 3 \\ 1 & -2 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix} \tag{8-33}$$

a) Calculate $\mathbf{AC}$, $\mathbf{BD}$ and $\mathbf{DC}$.

b) Determine if possible, $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$ and $(\mathbf{AB})^{-1}$.

c) Is it possible to decide whether $(\mathbf{AB})^{-1}$ exists after you have tried to determine $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$? If yes, how?

8.2 Powers of Matrices

We have now seen how the inverse of an invertible matrix is determined and we say that it has the power -1. Similarly we define arbitrary integer ***powers of square matrices***.

|||| Definition 8.12 Powers of a Matrix

For an arbitrary square matrix **A** the following natural powers are defined:

$$\mathbf{A}^0 = \mathbf{E} \quad \text{and} \quad \mathbf{A}^n = \overbrace{\mathbf{A}\mathbf{A}\cdots\mathbf{A}}^{n \text{ times}}, \text{ for } n \in \mathbb{N} \tag{8-34}$$

Furthermore for an arbitrary **invertible** square matrix **B** the negative powers are defined:

$$\mathbf{B}^{-n} = (\mathbf{B}^{-1})^n = \overbrace{\mathbf{B}^{-1}\mathbf{B}^{-1}\cdots\mathbf{B}^{-1}}^{n \text{ times}}, \text{ for } n \in \mathbb{N} \tag{8-35}$$

As a consequence of the definition of powers, some arithmetic rules can be given.

|||| Theorem 8.13 Arithmetic Rules for Powers of Matrices

For an arbitrary square matrix **A** and two arbitrary non-negative integers a and b the following arithmetic rules for powers are valid

$$\mathbf{A}^a\mathbf{A}^b = \mathbf{A}^{a+b} \quad \text{and} \quad (\mathbf{A}^a)^b = \mathbf{A}^{ab} \tag{8-36}$$

If **A** is invertible, these arithmetic rules are also valid for negative integers a and b.

Below is an example of two (simple) matrices that possess some funny characteristics. The characteristics are *not* typical for matrices!

|||| Example 8.14 Two Funny Matrices with Respect to Powers

Given the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \tag{8-37}$$

By use of both the Definition 8.12 and Theorem 8.13 the following calculations are performed. $\mathbf{A}^2$ is determined:

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{E} \tag{8-38}$$

Following the addendum to the fourth arithmetic rule in Theorem 8.8, **A** is invertible, and moreover $\mathbf{A} = \mathbf{A}^{-1}$. This gives

$$\begin{array}{ll} \vdots & \vdots \\ \mathbf{A}^{-3} = (\mathbf{A}\mathbf{A}^2)^{-1} = \mathbf{A}^{-1} = \mathbf{A} & \mathbf{A}^{-2} = (\mathbf{A}^2)^{-1} = \mathbf{E} \\ \mathbf{A}^{-1} = \mathbf{A} & \mathbf{A}^{0} = \mathbf{E} \\ \mathbf{A}^{1} = \mathbf{A} & \mathbf{A}^{2} = \mathbf{E} \\ \vdots & \vdots \end{array} \tag{8-39}$$

Thus all odd powers of **A** give **A**, while even powers give the identity matrix:

$$\mathbf{A}^{2n} = \mathbf{E} \text{ and } \mathbf{A}^{2n+1} = \mathbf{A} \text{ for } n \in \mathbb{Z} \tag{8-40}$$

$\mathbf{B}^2$ is determined:

$$\mathbf{B}^2 = \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{0} \tag{8-41}$$

According to the same arithmetic rule **B** is singular. Then it follows

$$\mathbf{B}^0 = \mathbf{E}\,,\, \mathbf{B}^1 = \mathbf{B}\,,\, \mathbf{B}^2 = \mathbf{0}\,,\, \mathbf{B}^n = \mathbf{0} \text{ for } n \geq 2 \tag{8-42}$$

8.3 Summary

- Square matrices are matrices where the number of rows equals the number of columns.
- The unit matrix $\mathbf{E}$ is a square matrix with the number one in the diagonal and zeros elsewhere:

$$\mathbf{E} = \mathbf{E}_{n\times n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \tag{8-43}$$

- If a square matrix has full rank, it is called regular, otherwise it is called singular.
- A square matrix, the entries of which are all zero except for those on the diagonal, is called a diagonal matrix.
- A square matrix, that is equal to the transpose of itself, is called a symmetric matrix.
- For a *regular* matrix $\mathbf{A}$ there exists a unique inverse, denoted $\mathbf{A}^{-1}$, satisfying:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{E} \tag{8-44}$$

  The inverse can be determined by Method 8.4.
- Rules of computation with square and inverse matrices exist, see Theorem 8.8.
- Powers of suare matrices are defined, see Definition 8.12. In addition some arithmetic rules exist.
- Inverse matrices are e.g. used in connection with *change of basis* and the *eigenvalue problem*. Moreover the *determinant* of a square matrix is defined in eNote 9, *Determinants*.

eNote 9

Determinants

In this eNote we look at ***square matrices****; that is they are of type $n \times n$ for $n \geq 2$, see eNote 8. It is an advantage but not indispensable to have knowledge about the concept of a determinant for (2×2)-matrices in advance. The matrix algebra from eNote 7 is assumed known (sum, product, transpose and inverse of matrices, plus the general solution method for systems of linear equations from eNote 6.*

Updated: 24.9.21 David Brander.

9.1 Intro to Determinants

The ***determinant*** of a real *square* $(n \times n)$-matrix $\mathbf{A}$ is a real number which we denote by $\det(\mathbf{A})$ or sometimes for short by $|\mathbf{A}|$. The determinant of a matrix can, in a way, be considered as a measure of how much the matrix 'weighs' - with sign; we will illustrate this visually and geometrically for (2×2)-matrices and for (3×3)-matrices in eNote 10.

The determinant is a well defined *function* of the total of n^2 numbers, that constitute the elements of an $(n \times n)$-matrix.

In order to define – and then calculate – the value of the determinant of an $(n \times n)$-matrix directly from the n^2 elements in each of the matrices we need two things: First the well-known formula for the determinant of (2×2)-matrices (see the definition 9.1 below) and secondly a method to cut up an arbitrary $(n \times n)$-matrix into (2×2)-matrices and thereby define and calculate arbitrary determinants from the determinants of these (2×2)-matrices.

9.2 Determinants of $(2 \times 2)-$Matrices

|||| Definition 9.1 Determinants of $(2 \times 2)-$Matrices

Let $\mathbf{A}$ be the arbitrary $(2 \times 2)-$matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} . \tag{9-1}$$

Then the determinant of $\mathbf{A}$ is defined by:

$$\det(\mathbf{A}) = a_{11} \cdot a_{22} - a_{21} \cdot a_{12} \,. \tag{9-2}$$

|||| Exercise 9.2 Inverse $(2 \times 2)-$Matrix

Remember that the inverse matrix $\mathbf{A}^{-1}$ of a invertible matrix $\mathbf{A}$ has the characteristic property that $\mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{E}$. Show directly from (9-1) and (9-2), that the inverse matrix $\mathbf{A}^{-1}$ to a $(2 \times 2)-$matrix $\mathbf{A}$ can be expressed in the following way (when $\det(\mathbf{A}) \neq 0$) :

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} . \tag{9-3}$$

|||| Exercise 9.3 Arithmetic Rules for $(2 \times 2)-$Matrices

For allsquare matrices a number of basic arithmetic rules are valid; they are presented in theorem 9.20 below. Check the three first equations in theorem 9.20 for (2×2)-matrices $\mathbf{A}$ and $\mathbf{B}$. Use direct calculation of both sides of the equations using (9-2).

9.3 Submatrices

|||| Definition 9.4 Submatrices

For an $(n \times n)$-matrix $\mathbf{A}$ we define the (i,j) submatrix, $\widehat{\mathbf{A}}_{ij}$, as the $((n-1) \times (n-1))$-submatrix of $\mathbf{A}$ that emerges by deleting the entire row i and the entire column j from the matrix $\mathbf{A}$.

All the total of n^2 submatrices $\widehat{\mathbf{A}}_{ij}$ (where $1 \leq i \leq n$ and $1 \leq j \leq n$) are less than $\mathbf{A}$ and are of the type $(n-1) \times (n-1)$ and therefore have only $(n-1)^2$ elements.

|||| Example 9.5 Submatrices for a $(3 \times 3)-$Matrix

A (3×3)-matrix $\mathbf{A}$ has total of 9 (2×2)-submatricer $\widehat{\mathbf{A}}_{ij}$. For example, if

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 3 & 2 \\ 0 & 5 & 1 \end{bmatrix} , \tag{9-4}$$

then the 9 submatrices belonging to $\mathbf{A}$ are given by:

$$\begin{array}{lll} \widehat{\mathbf{A}}_{11} = \begin{bmatrix} 3 & 2 \\ 5 & 1 \end{bmatrix} , & \widehat{\mathbf{A}}_{12} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} , & \widehat{\mathbf{A}}_{13} = \begin{bmatrix} 1 & 3 \\ 0 & 5 \end{bmatrix} , \\ \widehat{\mathbf{A}}_{21} = \begin{bmatrix} 2 & 1 \\ 5 & 1 \end{bmatrix} , & \widehat{\mathbf{A}}_{22} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} , & \widehat{\mathbf{A}}_{23} = \begin{bmatrix} 0 & 2 \\ 0 & 5 \end{bmatrix} , \\ \widehat{\mathbf{A}}_{31} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} , & \widehat{\mathbf{A}}_{32} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix} , & \widehat{\mathbf{A}}_{33} = \begin{bmatrix} 0 & 2 \\ 1 & 3 \end{bmatrix} . \end{array} \tag{9-5}$$

The corresponding determinants are determinants of $(2 \times 2)-$matrices and each of these can be calculated directly from the definition 9.1 above:

$$\begin{array}{lll} \det(\widehat{\mathbf{A}}_{11}) = -7 \, , & \det(\widehat{\mathbf{A}}_{12}) = 1 \, , & \det(\widehat{\mathbf{A}}_{13}) = 5 \, , \\ \det(\widehat{\mathbf{A}}_{21}) = -3 \, , & \det(\widehat{\mathbf{A}}_{22}) = 0 \, , & \det(\widehat{\mathbf{A}}_{23}) = 0 \, , \\ \det(\widehat{\mathbf{A}}_{31}) = 1 \, , & \det(\widehat{\mathbf{A}}_{32}) = -1 \, , & \det(\widehat{\mathbf{A}}_{33}) = -2 \, . \end{array} \tag{9-6}$$

9.4 Inductive Definition of Determinants

The determinant of a 3×3 matrix can now be defined from the determinants of 3 of the 9 submatrices, and generally: The determinant of an $n \times n$ matrix is defined by the use of the determinants of the n submatrices that belong to a (freely chosen) row r in the following way, which naturally is called ***expansion along the r-th row***:

|||| Definition 9.6 Determinants are Defined by Expansion

For an arbitrary value of the row index r the determinant of a given $(n \times n)$-matrix $\mathbf{A}$ is defined inductively in the following way:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{r+j} a_{rj} \det(\widehat{\mathbf{A}}_{rj}) \quad . \tag{9-7}$$

We here and subsequently use the following short notation for the sum and products of many terms, e.g. n given real numbers $c_1, c_2, \ldots, c_{n-1}, c_n$:

$$\begin{aligned} c_1 + c_2 + \cdots + c_{n-1} + c_n &= \sum_{i=1}^{n} c_i \ , \quad \text{and} \\ c_1 \cdot c_2 \cdot \cdots \cdot c_{n-1} \cdot c_n &= \prod_{i=1}^{n} c_i \quad . \end{aligned} \tag{9-8}$$

|||| Example 9.7 Expansion of a Determinant along the 1. Row

We will use Definition 9.6 directly in order to calculate the determinant of the matrix $\mathbf{A}$ that is given in example 9.5. We choose $r = 1$ and we thus need three determinants of the submatrices, $\det(\widehat{\mathbf{A}}_{11}) = -7$, $\det(\widehat{\mathbf{A}}_{12}) = 1$, and $\det(\widehat{\mathbf{A}}_{13}) = 5$, which we calculated already in example 9.5 above:

$$\begin{aligned} \det(\mathbf{A}) &= \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(\widehat{\mathbf{A}}_{1j}) \\ &= (-1)^{1+1} \cdot 0 \cdot \det(\widehat{\mathbf{A}}_{11}) + (-1)^{1+2} \cdot 2 \cdot \det(\widehat{\mathbf{A}}_{12}) + (-1)^{1+3} \cdot 1 \cdot \det(\widehat{\mathbf{A}}_{13}) \\ &= 0 - 2 + 5 = 3 \quad . \end{aligned} \tag{9-9}$$

Notice that the determinants of the submatrices must be multiplied by the element in $\mathbf{A}$ that is in entry (r, j) and with the sign-factor $(-1)^{r+j}$ before they are added. And notice that the determinants of the submatrices themselves can be expanded by the use of determinants of even smaller matrices, such that finally we only need to determine weighted sums of determinants of $(2 \times 2)-$matrices!

|||| Exercise 9.8 Choice of 'Expansion Row' Arbitrary

Show by direct calculation that we obtain the same value for the determinant by use of one of the other two rows for the expansion of the determinant in example 9.5.

|||| Definition 9.9 Alternative: Expansion along a Column

The determinant of a given $(n \times n)-$matrix $\mathbf{A}$ can alternatively be defined inductively by expansion along an arbitrary chosen *column* :

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+s} a_{i\,s} \det(\widehat{\mathbf{A}}_{i\,s}) \quad . \tag{9-10}$$

Here the expansion is expressed as the expansion along column s.

As is already hinted with the definitions and as shown in the concrete case of the matrix in example 9.5, it doesn't matter which row (or column) defines the expansion:

|||| Theorem 9.10 Choice of Row or Column for the Expansions Immaterial

The two definitions, 9.6 and 9.9, of the determinant of a square matrix give the same value and this they do without regard to the choice of row or column in the corresponding expansions.

|||| Exercise 9.11 Choice of Column for the Expansion is Immaterial

Show by direct calculation that we get the same value for the determinant in 9.5 by using expansion along any of the three columns in **A**.

It is of course wisest to expand along a row (or a column) that contains many 0's.

|||| Exercise 9.12 Determinants of Some Larger Matrices

Use the above instructions and results to find the determinants of each of the following matrices:

$$\begin{bmatrix} 0 & 2 & 7 & 1 \\ 1 & 3 & 0 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 5 & 8 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 2 & 1 & 0 & 5 \\ 1 & 3 & 2 & 0 & 2 \\ 0 & 5 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 5 & 2 & 7 & 1 & 9 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 2 & 1 & 5 & 3 \\ 0 & 1 & 3 & 2 & 2 & 1 \\ 0 & 0 & 5 & 1 & 1 & 4 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 5 & 2 & 7 & 1 & 9 \end{bmatrix}. \tag{9-11}$$

If there are many 0's in a matrix then it is much easier to calculate its determinant! Especially if all the elements in a row (or a column) are 0 except one element then it is clearly wisest to expand along that row (or column). And we are allowed to 'obtain' a lot of 0's by application of the well-known row operations, if you keep record of the constants used for divisions and how often you swap rows. See theorem 9.16 and example 9.17 below.

9.5 Computational Properties of Determinants

We collect some of the most important tools that are often used for the calculation and inspection of the matrix determinants.

It is not difficult to prove the following theorem, e.g. by expansion first along the first column or the first row, after which the pattern shows:

|||| Theorem 9.13 Matrices with 0 above or below the Diagonal

If an $(n \times n)-$matrix has only 0's above or below the diagonal, then the determinant is given by the products of the elements on the diagonal.

As a special case of this theorem we have:

|||| Theorem 9.14 The Determinant of a Diagonal Matrix

Let $\mathbf{\Lambda}$ denote an $(n \times n)$-*diagonal matrix* with the elements in the diagonal $\lambda_1, \lambda_2, ..., \lambda_n$ and 0's outside the diagonal:

$$\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} . \tag{9-12}$$

Then the determinant is

$$\det(\mathbf{\Lambda}) = \lambda_1 \lambda_2 \cdots \lambda_n = \prod_{i=1}^{n} \lambda_i \tag{9-13}$$

|||| Exercise 9.15 Determinant of a Bi-diagonal Matrix

Determine the determinant of the $(n \times n)-$ ***bi-diagonal matrix*** with arbitrarily given values $\mu_1, \ldots, \mu_n$ in the bi-diagonal and 0's elsewhere:

$$\mathbf{M} = \mathbf{bidiag}(\mu_1, \mu_2, \cdots, \mu_n) = \begin{bmatrix} 0 & \cdots & 0 & \mu_1 \\ 0 & \cdots & \mu_2 & 0 \\ \vdots & \cdot^{\cdot^{\cdot}} & \vdots & \vdots \\ \mu_n & \dots & 0 & 0 \end{bmatrix} . \tag{9-14}$$

General matrices (including square matrices), as known from eNote 6, can be reduced to reduced row echelon form by the use of row operations. If you keep an eye on what happens in every step in this reduction then the determinant of the matrix can be read directly from the process. The determinant of a matrix behaves 'nicely' even if you perform row operations on the matrix:

|||| Theorem 9.16 The Influence of Row Operations on the Determinant

The determinant has the following properties:

1. If all the elements in a row in $\mathbf{A}$ are 0 then the determinant is 0, $\det(\mathbf{A}) = 0$.
2. If two rows are swapped in $\mathbf{A}$, $R_i \leftrightarrow R_j$, then the sign of the determinant is shifted.
3. If all the elements in a row in $\mathbf{A}$ are multiplied by a constant k, $k \cdot R_i$, then the determinant is multiplied by k.
4. If two rows in a matrix $\mathbf{A}$ are equal then $\det(\mathbf{A}) = 0$.
5. A row operation of the type $R_j + k \cdot R_i$, $i \neq j$ does not change the determinant.

As indicated above it follows from these properties of the determinant function that the well-known reduction of a given matrix $\mathbf{A}$ to the reduced row echelon form, $\mathrm{rref}(\mathbf{A})$, through row operations as described in eNote 6, in fact *comprises* a totally explicit calculation of the determinant of $\mathbf{A}$. We illustrate with a simple example:

|||| Example 9.17 Inspection of Determinant by Reduction to the Reduced Row Echelon Form

We consider the $(3 \times 3)-$matrix $\mathbf{A}_1 = \mathbf{A}$ from example 9.5:

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 3 & 2 \\ 0 & 5 & 1 \end{bmatrix} \tag{9-15}$$

We reduce $\mathbf{A}_1$ to the reduced row echelon form in the usual way by Gauss–Jordan row operations and all the time we keep an eye on what happens to the determinant by using the rules in 9.16 (and possibly by checking the results by direct calculations):

Operation: Swap row 1 and row 2, $R_1 \leftrightarrow R_2$: The determinant changes sign :

$$\det(\mathbf{A}_2) = -\det(\mathbf{A}_1) \quad : \tag{9-16}$$

$$\mathbf{A}_2 = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 1 \\ 0 & 5 & 1 \end{bmatrix} \tag{9-17}$$

Operation: $\frac{1}{2}R_2$, row 2 is multiplied by $\frac{1}{2}$: The determinant is multiplied by $\frac{1}{2}$:

$$\det(\mathbf{A}_3) = \frac{1}{2}\det(\mathbf{A}_2) = -\frac{1}{2}\det(\mathbf{A}_1) \;\; : \tag{9-18}$$

$$\mathbf{A}_3 = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 1 & 1/2 \\ 0 & 5 & 1 \end{bmatrix} \tag{9-19}$$

Operation: $R_1 - 3R_2$: The determinant is unchanged:

$$\det(\mathbf{A}_4) = \det(\mathbf{A}_3) = \frac{1}{2}\det(\mathbf{A}_2) = -\frac{1}{2}\det(\mathbf{A}_1) \;\; : \tag{9-20}$$

$$\mathbf{A}_4 = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 0 & 5 & 1 \end{bmatrix} \tag{9-21}$$

Operation: $R_3 - 5R_2$: The determinant is unchanged:

$$\det(\mathbf{A}_5) = \det(\mathbf{A}_4) = \det(\mathbf{A}_3) = \frac{1}{2}\det(\mathbf{A}_2) = -\frac{1}{2}\det(\mathbf{A}_1) \;\; : \tag{9-22}$$

$$\mathbf{A}_5 = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 0 & 0 & -3/2 \end{bmatrix} \tag{9-23}$$

Now the determinant is the product of the elements in the diagonal because all the elements below the diagonal are 0, see theorem 9.13. All in all we therefore have:

$$-\frac{3}{2} = \det(\mathbf{A}_5) = \det(\mathbf{A}_4) = \det(\mathbf{A}_3) = \frac{1}{2}\det(\mathbf{A}_2) = -\frac{1}{2}\det(\mathbf{A}_1) \;\; : \tag{9-24}$$

From this we obtain directly – by reading 'backwards':

$$-\frac{1}{2}\det(\mathbf{A}_1) = -\frac{3}{2} \,, \tag{9-25}$$

such that

$$\det(\mathbf{A}_1) = 3 \,. \tag{9-26}$$

In addition we have the following relation between the rank and the determinant of a matrix; the determinant reveals whether the matrix is singular or invertible:

|||| Theorem 9.18 Rank versus Determinant

The rank of a square $(n \times n)$-matrix $\mathbf{A}$ is less than n if and only if the determinant of $\mathbf{A}$ is 0. In other words, $\mathbf{A}$ is singular if and only if $\det(\mathbf{A}) = 0$.

If a matrix contains a variable, a parameter, then the determinant of the matrix is a function of this parameter; in the applications of matrix-algebra it is often crucial to be able to find the zeroes of this function – exactly because the corresponding matrix is singular for those values of the parameter, and hence there might not be a (unique) solution to the corresponding system of linear equations with the matrix as the coefficient matrix.

|||| Exercise 9.19 Determinant of a Matrix with a Variable

Given the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & a & a^2 & a^3 \\ 1 & 0 & a^2 & a^3 \\ 1 & a & a & a^3 \\ 1 & a & a^2 & a \end{bmatrix} \text{, where } a \in \mathbb{R}. \tag{9-27}$$

1. Determine the determinant of $\mathbf{A}$ as a polynomium in a.
2. Determine the roots of this polynomium.
3. Find the rank of $\mathbf{A}$ for $a \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$. What does the rank have to do with the roots of the determinant?
4. Find the rank of $\mathbf{A}$ for all a.

|||| Theorem 9.20 Arithmetic Rules for Determinants

Let $\mathbf{A}$ and $\mathbf{B}$ denote two $(n \times n)-$matrices. Then:

1. $\det(\mathbf{A}) = \det(\mathbf{A}^\top)$
2. $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
3. $\det(\mathbf{A}^{-1}) = (\det(\mathbf{A}))^{-1}$, when $\mathbf{A}$ is invertible, that is $\det(\mathbf{A}) \neq 0$
4. $\det(\mathbf{A}^k) = (\det(\mathbf{A}))^k$, for all $k \geq 1$.
5. $\det(\mathbf{B}^{-1}\mathbf{AB}) = \det(\mathbf{A})$, when $\mathbf{B}$ is invertible, that is $\det(\mathbf{B}) \neq 0$.

|||| Exercise 9.21

Prove the last 3 equations in theorem 9.20 by the use of $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.

|||| Exercise 9.22 The Determinant of a Sum is not the Sum of the Determinants

Show by the most simple example, that the determinant-function det() is *not* additive. That is, find two $(n \times n)-$matrices $\mathbf{A}$ and $\mathbf{B}$ such that

$$\det(\mathbf{A}+\mathbf{B}) \neq \det(\mathbf{A}) + \det(\mathbf{B}) \ . \tag{9-28}$$

|||| Exercise 9.23 Use of Arithmetic Rules for Determinants

Let a denote a real number. The following matrices are given:

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & 4 \\ 1 & a & 2 \\ 2 & 3 & 3 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 2 & -1 & 0 \\ -5 & 3 & -1 \\ 0 & 1 & a \end{bmatrix} . \tag{9-29}$$

1. Find $\det(\mathbf{A})$ and $\det(\mathbf{B})$.
2. Find $\det(\mathbf{A}\,\mathbf{B})$ and $\det\left((\mathbf{A}^\top\mathbf{B})^4\right)$.

3. Determine those values of a for which $\mathbf{A}$ is invertible and find for these values of a the expression for $\det(\mathbf{A}^{-1})$.

9.6 Advanced: Cramer's Solution Method

If $\mathbf{A}$ is a invertible $n \times n$ matrix and $\mathbf{b} = (b_1, ..., b_n)$ is an arbitrary vector in $\mathbb{R}^n$, then there exists (as is known from eNote 6 (invertible coefficient matrix)) exactly one solution $\mathbf{x} = (x_1, ..., x_n)$ to the system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ and we found in that eNote method for finding the solution.

Cramer's method for solving such systems of equations is a *direct* method. Essentially it consists of calculating suitable determinants of matrices constructed from $\mathbf{A}$ and $\mathbf{b}$ and then writing down the solution directly from the calculated determinants.

|||| Theorem 9.24 Cramer's Solution Formula

Let $\mathbf{A}$ be a invertible $n \times n$ matrix and let $\mathbf{b} = (b_1, ..., b_n)$ denote an arbitrary vector in $\mathbb{R}^n$. Then there exists (as is known from eNote 6 (invertible coefficient matrix)) exactly one solution $\mathbf{x} = (x_1, ..., x_n)$ to the system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \ , \tag{9-30}$$

and the elements in the solution are given by:

$$x_j = \frac{1}{\det(\mathbf{A})} \det(\mathbf{A}\dagger_j^{\mathbf{b}}) \ , \tag{9-31}$$

where $\mathbf{A}\dagger_j^{\mathbf{b}}$ denotes the $(n \times n)-$matrix that emerges by replacing column j in $\mathbf{A}$ with $\mathbf{b}$.

|||| Explanation 9.25 What † Means

If $\mathbf{A}$ is the following matrix (from example 9.5)

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 3 & 2 \\ 0 & 5 & 1 \end{bmatrix}, \tag{9-32}$$

and if we let $\mathbf{b} = (b_1, b_2, b_3)$, then

$$\mathbf{A}\dagger_1^{\mathbf{b}} = \begin{bmatrix} b_1 & 2 & 1 \\ b_2 & 3 & 2 \\ b_3 & 5 & 1 \end{bmatrix}, \ \mathbf{A}\dagger_2^{\mathbf{b}} = \begin{bmatrix} 0 & b_1 & 1 \\ 1 & b_2 & 2 \\ 0 & b_3 & 1 \end{bmatrix}, \ \mathbf{A}\dagger_3^{\mathbf{b}} = \begin{bmatrix} 0 & 2 & b_1 \\ 1 & 3 & b_2 \\ 0 & 5 & b_3 \end{bmatrix}. \tag{9-33}$$

|||| Exercise 9.26 Use Cramer's Solution Formula

If in particular we let $\mathbf{A}$ be the same matrix as above and now let $\mathbf{b} = (1, 3, 2)$, then we get by substitution of $\mathbf{b}$ in (9-33) and then computing the relevant determinants:

$$\begin{aligned} \det(\mathbf{A}\dagger_1^{\mathbf{b}}) &= \det\left(\begin{bmatrix} 1 & 2 & 1 \\ 3 & 3 & 2 \\ 2 & 5 & 1 \end{bmatrix}\right) = 4 \\ \det(\mathbf{A}\dagger_2^{\mathbf{b}}) &= \det\left(\begin{bmatrix} 0 & 1 & 1 \\ 1 & 3 & 2 \\ 0 & 2 & 1 \end{bmatrix}\right) = 1 \\ \det(\mathbf{A}\dagger_3^{\mathbf{b}}) &= \det\left(\begin{bmatrix} 0 & 2 & 1 \\ 1 & 3 & 3 \\ 0 & 5 & 2 \end{bmatrix}\right) = 1\,. \end{aligned} \tag{9-34}$$

Since we also know $\det(\mathbf{A}) = 3$ we have now constructed the solution to the system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, through (9-31):

$$\mathbf{x} = (x_1, x_2, x_3) = \left(\frac{1}{3}\cdot 4, \frac{1}{3}\cdot 1, \frac{1}{3}\cdot 1\right) = \left(\frac{4}{3}, \frac{1}{3}, \frac{1}{3}\right). \tag{9-35}$$

1. Check by direct isubstitution, that $\mathbf{x}$ is a solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$.
2. Determine $\mathbf{A}^{-1}$ and use it directly for the solution of the system of equations.

3. Solve the system of equations by reduction of the augmented matrix to the reduced row echelon form as in eNote 2 followed by a reading of the solution.

In order to show what is actually going on in Cramer's solution formula we first define the ***adjoint matrix*** for a matrix $\mathbf{A}$:

|||| Definition 9.27 The Adjoint Matrix

The *(classical) adjoint matrix* $\operatorname{adj}(\mathbf{A})$ (also called the *adjugate matrix*) is defined by the elements that are used in the definition 9.6 of the determinant of $\mathbf{A}$:

$$\operatorname{adj}(\mathbf{A}) = \begin{bmatrix} (-1)^{1+1}\det(\widehat{\mathbf{A}}_{11}) & \cdot & \cdot & (-1)^{1+n}\det(\widehat{\mathbf{A}}_{1n}) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ (-1)^{n+1}\det(\widehat{\mathbf{A}}_{n1}) & \cdot & \cdot & (-1)^{n+n}\det(\widehat{\mathbf{A}}_{nn}) \end{bmatrix}^{\top} \tag{9-36}$$

In other words: The element in entry (j,i) in the adjoint matrix $\operatorname{adj}(\mathbf{A})$ is the sign-modified determinant of the (i,j) submatrix, that is: $(-1)^{i+j}\det(\widehat{\mathbf{A}}_{ij})$. Notice the use of the transpose in (9-36).

|||| Example 9.28 An Adjoint Matrix

In example 9.5 we looked at the following matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 3 & 2 \\ 0 & 5 & 1 \end{bmatrix} . \tag{9-37}$$

The matrix $\mathbf{A}$ has the following adjoint matrix:

$$\operatorname{adj}(\mathbf{A}) = \begin{bmatrix} -7 & 3 & 1 \\ -1 & 0 & 1 \\ 5 & 0 & -2 \end{bmatrix} , \tag{9-38}$$

that is obtained directly from earlier computations of the determinants of the submatrices, remembering that each element is given a sign that depends on the 'entry', and that the expression (9-36) is to be transposed.

$$\begin{aligned} &\det(\widehat{\mathbf{A}}_{11}) = -7\,, \det(\widehat{\mathbf{A}}_{12}) = 1\,, \quad \det(\widehat{\mathbf{A}}_{13}) = 5\,, \\ &\det(\widehat{\mathbf{A}}_{21}) = -3\,, \det(\widehat{\mathbf{A}}_{22}) = 0\,, \quad \det(\widehat{\mathbf{A}}_{23}) = 0\,, \\ &\det(\widehat{\mathbf{A}}_{31}) = 1\,, \quad \det(\widehat{\mathbf{A}}_{32}) = -1\,, \det(\widehat{\mathbf{A}}_{33}) = -2\,. \end{aligned} \tag{9-39}$$

|||| Exercise 9.29 Adjoint Versus Inverse Matrix

Show that all square matrices $\mathbf{A}$ fulfil the following

$$\mathbf{A}\,\mathrm{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{E} \tag{9-40}$$

such that the inverse matrix to $\mathbf{A}$ (which exists precisely if $\det(\mathbf{A}) \neq 0$) can be found in the following way:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\,\mathrm{adj}(\mathbf{A}) \quad . \tag{9-41}$$

Hint: The exercise it not easy. It is recommended to practice on a (2×2)-matrix. The zeroes of the identity matrix in equation (9-40) are obtained by using the property that the determinant of a matrix with two identical columns is 0.

The proof of theorem 9.24 is now rather short:

|||| Proof

By multiplying both sides of equation (9-30) with $\mathbf{A}^{-1}$ we get:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \frac{1}{\det(\mathbf{A})}\,\mathrm{adj}(\mathbf{A})\mathbf{b} \quad , \tag{9-42}$$

and thus – if we denote the elements in $\mathrm{adj}(\mathbf{A})$, α_{ij} :

$$\begin{bmatrix} x_1 \\ \cdot \\ x_n \end{bmatrix} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} \alpha_{11} & \cdot & \alpha_{1n} \\ \cdot & \cdot & \cdot \\ \alpha_{n1} & \cdot & \alpha_{nn} \end{bmatrix} \begin{bmatrix} b_1 \\ \cdot \\ b_n \end{bmatrix} \quad . \tag{9-43}$$

From this we read directly

$$\begin{aligned} x_j &= \frac{1}{\det(\mathbf{A})} \sum_{i=1}^{n} \alpha_{ji} b_i \\ &= \frac{1}{\det(\mathbf{A})} \sum_{i=1}^{n} (-1)^{i+j} b_i \det(\widehat{\mathbf{A}}_{ij}) \\ &= \frac{1}{\det(\mathbf{A})} \det(\mathbf{A}\dagger_j^{\mathbf{b}}) \ , \end{aligned} \tag{9-44}$$

where we in the establishment of the last equality sign have used that

$$\sum_{i=1}^{n} (-1)^{i+j} b_i \det(\widehat{\mathbf{A}}_{ij}) \tag{9-45}$$

is exactly the expansion of $\det(\mathbf{A}\dagger_j^{\mathbf{b}})$ along *column* number j, that is the expansion along the **b**-column in $\det(\mathbf{A}\dagger_j^{\mathbf{b}})$, see the definition in equation (9-10).

■

9.7 Summary

- The determinant of a *square matrix with real elements* is a real number that is calculated from the n^2 elements in the matrix, either by expansion along a row or a column or through inspection of the Gauss-Jordan reduction process to the reduced row echelon form. When expanding along a matrix row or column the intelligent choice for it is a row or column in the matrix with many 0-elements. The expansion along row number r takes place inductively after the following formula that expresses the determinant as a sum of 'smaller' determinants (with suitable signs), see definition 9.6:

  $$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{r+j} a_{rj} \det(\widehat{\mathbf{A}}_{rj}) \quad , \tag{9-46}$$

  where $\widehat{\mathbf{A}}_{rj}$ is the submatrix that emerges by deleting row r and column j from the matrix $\mathbf{A}$, see definition 9.4.

- There exist convenient arithmetic rules for the calculation of determinants of products of matrices, determinants of the inverse matrix, and determinants of the transpose of a matrix. See Theorem 9.20. The most important arithmetic rules are the product-formula

  $$\det(\mathbf{A} \cdot \mathbf{B}) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$$

  and the transpose-formula

  $$\det(\mathbf{A}) = \det(\mathbf{A}^{\top}) \quad .$$

- The determinant of a square matrix that contains a variable, is a function of this variable. The characteristic polynomial is such a very important function: $\mathcal{K}_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda \mathbf{E})$ is n'th degree polynomial in the variable λ, see Definition 9.32.

- Cramer's solution formula gives the direct way (through computations of determinants) to the solution of a inhomogeneous system of linear equations with a invertible coefficient matrix, see Theorem 9.24. If the system of equations is

  $$\mathbf{A}\mathbf{x} = \mathbf{b} \,, \tag{9-47}$$

  the the solutions is given by:

  $$x_j = \frac{1}{\det(\mathbf{A})} \det(\mathbf{A} \dagger_j^{\mathbf{b}}) \quad , \tag{9-48}$$

  where $\mathbf{A} \dagger_j^{\mathbf{b}}$ denotes the matrix that emerges by replacing column j in the matrix $\mathbf{A}$ with $\mathbf{b}$.

9.8 Advanced: Characteristic Polynomial

The material in this subsection naturally belongs to this eNote about determinants due to the involved calculations, but it is only later, when solving the so-called eigenvalue problem, that we will find the characteristic polynomialsreally useful.

Fo a given square matrix $\mathbf{A}$ we define the corresponding ***characteristic matrix*** and the corresponding ***characteristic polynomial*** in the following way:

|||| Definition 9.30 The Characteristic Matrix

Let $\mathbf{A}$ be an $(n \times n)-$matrix. The corresponding *characteristic matrix* is the following real matrix-function of the real variable λ:

$$\mathbf{K_A}(\lambda) = \mathbf{A} - \lambda\,\mathbf{E}\,,\ \text{in which } \lambda \in \mathbb{R}\,, \tag{9-49}$$

where $\mathbf{E} = \mathbf{E}_{n\times n} = \mathbf{diag}(1,1,...,1)$ is the $(n \times n)-$identity matrix.

|||| Example 9.31 A Characteristic Matrix

Given a $(3 \times 3)-$matrix $\mathbf{A}$ by

$$\mathbf{A} = \begin{bmatrix} 3 & -2 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}. \tag{9-50}$$

Then

$$\mathbf{K_A}(\lambda) = \mathbf{A} - \lambda\,\mathbf{E} = \begin{bmatrix} 3-\lambda & -2 & 0 \\ 0 & 1-\lambda & 0 \\ 1 & -1 & 2-\lambda \end{bmatrix}. \tag{9-51}$$

The corresponding ***characteristic polynomial*** for the matrix $\mathbf{A}$ is then the following real polynomial to the variable λ:

Definition 9.32 The Characteristic Polynomial

Given the square matrix $\mathbf{A}$ then the characteristic polynomial for $\mathbf{A}$ is defined like this:

$$\mathcal{K}_{\mathbf{A}}(\lambda) = \det(\mathbf{K}_{\mathbf{A}}(\lambda)) = \det(\mathbf{A} - \lambda\,\mathbf{E})\ , \text{ where } \lambda \in \mathbb{R}\ . \tag{9-52}$$

Example 9.33 A Characteristic Polynomial

With $\mathbf{A}$ as in example 9.31 we get the following characteristic polynomial for $\mathbf{A}$ by expansion of the characteristic matrix along the last column:

$$\begin{aligned}\mathcal{K}_{\mathbf{A}}(\lambda) &= \det\left(\begin{bmatrix} 3-\lambda & -2 & 0 \\ 0 & 1-\lambda & 0 \\ 1 & -1 & 2-\lambda \end{bmatrix}\right) \\ &= (-1)^{3+3}(2-\lambda)\det\left(\begin{bmatrix} 3-\lambda & -2 \\ 0 & 1-\lambda \end{bmatrix}\right) \\ &= (2-\lambda)(3-\lambda)(1-\lambda)\ .\end{aligned} \tag{9-53}$$

The characteristic polynomial for $\mathbf{A}$ thus has the roots 1, 2, and 3.

The characteristic polynomial $\mathcal{K}_{\mathbf{A}}(\lambda)$ – and in particular λ-values for which $\mathcal{K}_{\mathbf{A}}(\lambda) = 0$, i.e. the roots of the polynomial – will play a decisive role in the understanding and application of the operative properties of the $\mathbf{A}$-matrix. This will be described in the eNote about eigenvalues and eigenvectors. The roots of the characteristic polynomial of a matrix are termed the eigenvalues of the matrix.

Exercise 9.34 The Degree of the Characteristic Polynomial

Give the reasons, why the characteristic polynomial $\mathcal{K}_{\mathbf{A}}(\lambda)$ for an $(n \times n)-$matrix $\mathbf{A}$ is a polynomial in λ of the degree n.

|||| Exercise 9.35 Some Characteristic Polynomials and their Roots

Determine the characteristic polynomials for the following matrices and find all real roots in each of the polynomials:

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \,,\; \mathbf{A}_2 = \mathbf{diag}(a_1, a_2, a_3) \,,\; \mathbf{A}_3 = \mathbf{bidiag}(b_1, b_2, b_3) \,. \tag{9-54}$$

|||| Exercise 9.36 Find Matrices with Given Properties

Construct two (4×4)-matrices $\mathbf{A}$ and $\mathbf{B}$ such that one has only real roots in the corresponding characteristic polynomial, and such that the other has no real roots in the corresponding characteristic polynomial.

eNote 10

Geometric Vectors

The purpose of this note is to give an introduction to geometric vectors in the plane and 3-dimensional space, aiming at the introduction of a series of methods that manifest themselves in the general theory of vector spaces. The key concepts are linear independence and linear dependence, plus basis and coordinates. The note assumes knowledge of elementary geometry in the plane and 3-space, of systems of linear equations as described in eNote 6 and of matrix algebra as described in eNote 7.

Updated 25.09.21 David Brander

By a ***geometric vector*** in the plane $\mathbb{R}^2$ or Euclidean 3-space, $\mathbb{R}^3$, we understand a connected pair consisting of a *length* and a *direction*. Euclidean vectors are written as small bold letters, e.g. $\mathbf{v}$. A vector can be represented by an ***arrow*** with a given initial point and a terminal point. If the vector $\mathbf{v}$ is represented by an arrow with the initial point A and the terminal point B, we use the representation $\mathbf{v} = \overrightarrow{AB}$. All arrows with the same length and direction as the arrow from A to B, also represent $\mathbf{v}$.

Example 10.1 Parallel Displacement Using Vectors

Geometric vectors can be applied in ***parallel displacement*** in the plane and 3-space. In Figure 10.1 the line segment CD is constructed from the line segment AB as follows: all points of AB are displaced by the vector $\mathbf{u}$. In the same way the line segment EF emerges from AB by parallel displacement by the vector $\mathbf{v}$. $\overrightarrow{AB} = \overrightarrow{CD} = \overrightarrow{EF}$ but notice that e.g. $\overrightarrow{AB} \neq \overrightarrow{FE}$.

In what follows we assume that a *unit line segment* has been chosen, that is a line segment

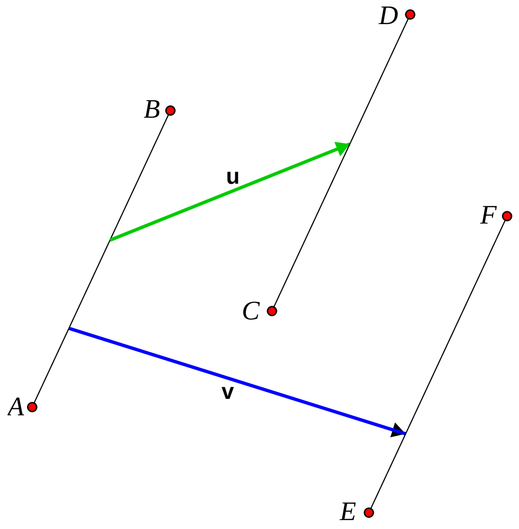


Figure 10.1: Parallel displacement by a vector

that has the length 1. By $|\mathbf{v}|$ we understand the length of the vector $\mathbf{v}$ as the proportionality factor with respect to the unit line segment, that is, a real number. All vectors of the same length as the unit line segment are called ***unit vectors***.

For practical reasons a particular vector that has length 0 and which has no direction is introduced. It is called the ***zero vector*** and is written $\mathbf{0}$. For every point A we put $\overrightarrow{AA} = \mathbf{0}$. Any vector that is not the zero vector is called a *proper* vector.

For every proper vector $\mathbf{v}$ we define the *opposite vector* $-\mathbf{v}$ as the vector that has the same length as $\mathbf{v}$, but the opposite direction. If $\mathbf{v} = \overrightarrow{AB}$, then $\overrightarrow{BA} = -\mathbf{v}$. For the zero vector we put $-\mathbf{0} = \mathbf{0}$.

It is often practical to use a common initial point when different vectors are to be represented by arrows. We choose a fixed point O which we term *the origin*, and consider those representations of the vectors that have O as the initial point. Vectors represented in this way are called ***position vectors***, because every given vector $\mathbf{v}$ has a unique point (position) P that satisifies $\mathbf{v} = \overrightarrow{OP}$. Conversely, every point Q corresponds to a unique vector $\mathbf{u}$ such that $\overrightarrow{OQ} = \mathbf{u}$.

By the *angle between two proper vectors in the plane* we understand the unique angle between their representations radiating from O, in the interval $[0;\pi]$. If a vector $\mathbf{v}$ in the plane is turned the angle $\pi/2$ counter-clockwise, a new vector emerges that is called $\mathbf{v}$'s *hat vector*, it is denoted $\widehat{\mathbf{v}}$.

By the *the angle between two proper vectors in 3-space* we understand the angle between their representations radiating from O in the plane that contains their representations.

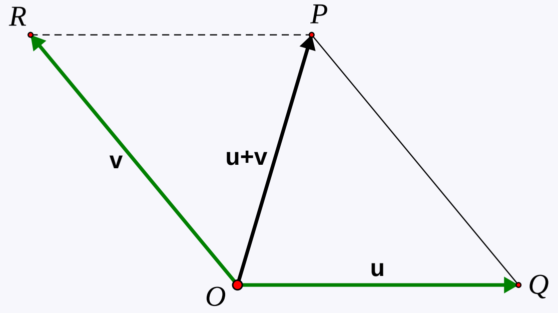
It makes good and useful sense "to add vectors", taking account of the vectors' lengths and directions. Therefore in the following we can introduce some arithmetic operations for geometric vectors. First it concerns two *linear operations*, addition of vectors and multiplication of a vector by a scalar (a real number). Later we will consider three ways of multiplying vectors, viz. the *dot product*, and for vectors in 3-space the *cross product* and the *scalar triple product*.

10.1 Addition and Multiplication by a Scalar

|||| Definition 10.2 Addition

Given two vectors in the plane or 3-space, $\mathbf{u}$ and $\mathbf{v}$. The sum $\mathbf{u}+\mathbf{v}$ is determined in the following way:

- We choose the origin O and mark the position vectors $\mathbf{u} = \overrightarrow{OQ}$ and $\mathbf{v} = \overrightarrow{OR}$.
- By parallel displacement of the line segments OR by $\mathbf{u}$ the line segment QP is constructed.
- $\overrightarrow{OP}$ is then the position vector for the sum of $\mathbf{u}$ and $\mathbf{v}$, in short $\mathbf{u}+\mathbf{v} = \overrightarrow{OP}$.



In physics you talk about the ”parallelogram of forces": If the object O is influenced by the forces $\mathbf{u}$ and $\mathbf{v}$, the *resulting force* can be determined as the vector sum $\mathbf{u}+\mathbf{v}$, the direction of which gives the direction of the resulting force, and the length of which gives the magnitude of the resulting force. If in particular $\mathbf{u}$ and $\mathbf{v}$ are of the same length, but have opposite directions, the resulting force is equal to the $\mathbf{0}$-vector.

We then introduce multiplication of a vector by a scalar:

|||| Definition 10.3 Multiplication by a Scalar

Given a vector $\mathbf{v}$ in the plane or 3-space and a scalar k. If $\mathbf{v}=\mathbf{0}$, we have $k\mathbf{v}=\mathbf{v}k=\mathbf{0}$. Otherwise by the product $k\mathbf{v}$ the following is understood:

- If $k>0$, then $k\mathbf{v}=\mathbf{v}k$ is the vector that has the same direction as $\mathbf{v}$ and which is k times as long as $\mathbf{v}$.
- If $k=0$, then $k\mathbf{v}=\mathbf{0}$.
- If $k<0$, then $k\mathbf{v}=\mathbf{v}k$ is the vector that has the *opposite direction* of $\mathbf{v}$ and which is $-k=|\,k\,|$ as long as $\mathbf{v}$.

|||| Example 10.4 Multiplication by a Scalar

A given vector $\mathbf{v}$ is multiplied by -1 and 2, respectively:

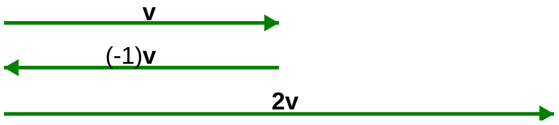


Figure: Multiplication of vector by -1 and 2

It follows immediately from the defintion 10.3 that multiplication of a vector by -1 gives the vector's opposite vector, in short

$$(-1)\mathbf{u} = -\mathbf{u}\,.$$

Thus we use the following way of writing

$$(-5)\mathbf{v} = -(5\mathbf{v}) = -5\mathbf{v}\,.$$

From the definition 10.3 the *zero rule* follows immediately for geometric vectors:

$$k\mathbf{v} = \mathbf{0} \Leftrightarrow k = 0 \text{ or } \mathbf{v} = \mathbf{0}\,.$$

In the following example it is shown that multiplication of an arbitrary vector by an arbitrary scalar can be performed by a genuine compasses and ruler construction.

|||| Example 10.5 Geometrical Multiplication

Given a vector $\mathbf{a}$ and a line segment of length k, we wish to construct the vector $k\mathbf{a}$.

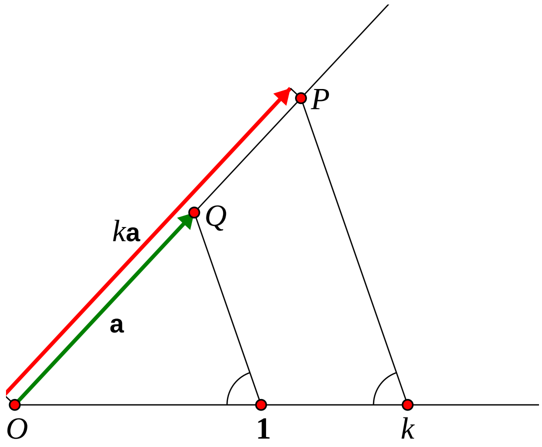


Figure: Multiplication of a vector by an arbitrary scalar

First the position vector $\overrightarrow{OQ}= \mathbf{a}$ is marked. Then with O as the initial point we draw a line which is used as a ruler and which is not parallel to $\mathbf{a}$, and where the numbers 1 and k are marked. The triangle OkP is drawn so it is congruent with the triangle $O1Q$. Since the two triangles are similar it must be true that $k\mathbf{a} = \overrightarrow{OP}$.

|||| Exercise 10.6

Given two parallel vectors **a** and **b** and a ruler line. How can you using a pair of compasses and the ruler line construct a line segment of the length k given that $\mathbf{b} = k\mathbf{a}$.

|||| Exercise 10.7

Given the proper vector **v** and a ruler line. Draw the vector $\frac{1}{|\mathbf{v}|}\mathbf{v}$.

Parametric representations for straight lines in the plane or 3-space are written using proper vectors. Below we first give an example of a line through the origin and then an example of a line not passing through the origin.

|||| Example 10.8 Parametric Representation of a Straight Line

Given a straight line l through the origin, we wish to write the points on the line using a ***parametric representation***:

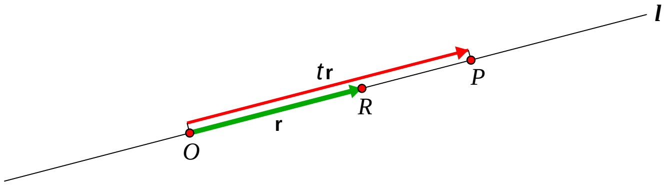


Figure: Parametric representation for a line through the origin

A point R on l different from the origin is chosen. The vector $\mathbf{r} = \overrightarrow{OR}$ is called a ***direction vector*** for l. For every point P on l corresponds exactly one real number t that satisfies $\overrightarrow{OP} = t\mathbf{r}$. Conversely, to every real number t corresponds exactly one point P on l so that $\overrightarrow{OP} = t\mathbf{r}$. As t traverses the real numbers from -∞ to +∞, P will traverse all of l in the direction determined by **r**. Then

$$\{ P \mid \overrightarrow{OP} = t\mathbf{r} \text{ where } t \in \mathbb{R} \}$$

is a parametric representation of l.

|||| Example 10.9 Parametric Representation of a Straight Line

The line m does not go through the origin. We wish to describe the points on m by use of a parametric representation:

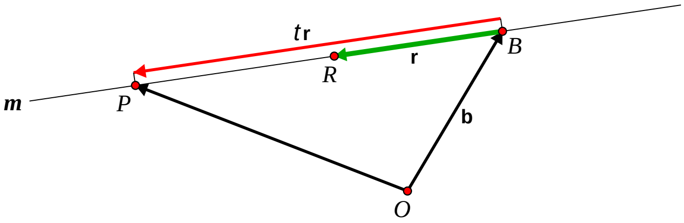


Figure: Parametric representation of a line

First an *initial point* B on m is chosen, and we put $\mathbf{b}\ = \overrightarrow{OB}$. A point $R\in m$ different from B is chosen. The vector $\mathbf{r}\ = \overrightarrow{BR}$ is then a directional vector for m. To every point P on m corresponds exactly one real number t that fulfils $\overrightarrow{OP} = \mathbf{b}+t\mathbf{r}$. Conversely, to every number t exactly one point P on m corresponds so that $\overrightarrow{OP} = \mathbf{b}+t\mathbf{r}$. When t traverses the real numbers from -∞ to +∞, P will traverse all of m in the direction determined by $\mathbf{r}$. Then

$$\{\,P\mid\ \overrightarrow{OP} = \mathbf{b}+t\mathbf{r}\ \text{where } t\in\mathbb{R}\,\}$$

is a parametric representation for m.

Parametric representations can also be used for the description of line segments. This is the subject of the following exercise.

|||| Exercise 10.10

Consider the situation in example 10.9. Draw the oriented line segment with the parametric representation

$$\{\,P\mid\ \overrightarrow{OP} = \mathbf{b}+t\mathbf{r},\ \text{where } t\in[\,-1;\,2\,]\,\}\,.$$

|||| Exercise 10.11

Given two (different) points A and B. Describe with a parametric representation the oriented line segment from A to B.

We will need more advanced arithmetic rules for addition of geometric vectors and multiplication of geometric vectors by scalars than the ones we have given in the examples above. These are sketched in the following theorem and afterwards we will discuss examples of how they can be justified on the basis of already defined arithmetic operations and theorems known from elementary geometry.

|||| Theorem 10.12 Arithmetic Rules

For arbitrary geometric vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ and for arbitrary real numbers k_1 and k_2 the following arithmetic rules are valid:

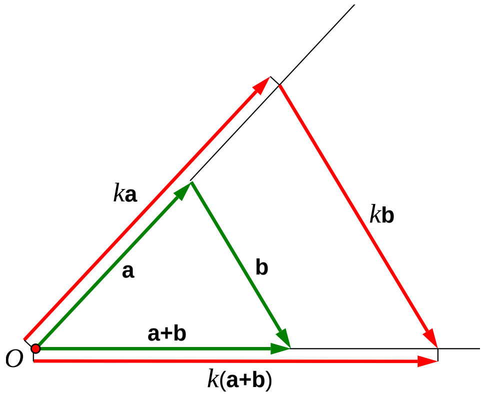
1.	$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$	Addition is commutative
2.	$(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$	Addition is associative
3.	$\mathbf{u}+\mathbf{0}=\mathbf{u}$	The zero vector is neutral for addition
4.	$\mathbf{u}+(-\mathbf{u})=\mathbf{0}$	The sum of a vector and its opposite is $\mathbf{0}$
5.	$k_1(k_2\mathbf{u})=(k_1k_2)\mathbf{u}$	Scalar multiplication is associative
6.	$(k_1+k_2)\mathbf{u}=k_1\mathbf{u}+k_2\mathbf{u}$	} The distributive rules apply
7.	$k_1(\mathbf{u}+\mathbf{v})=k_1\mathbf{u}+k_1\mathbf{v}$	
8.	$1\mathbf{u}=\mathbf{u}$	The scalar 1 is neutral in the product with vectors

The arithmetic rules in Theorem 10.12 can be illustrated and proven using geometric constructions. Let us as an example take the first rule, the commutative rule. Here we just have to look at the figure in the definition 10.2, where $\mathbf{u}+\mathbf{v}$ is constructed. If we construct $\mathbf{v}+\mathbf{u}$, we will displace the line segment OQ with v and consider the emerging line segment RP_2. It must be true that the parallelogram $OQPR$ is identical to the parallelogram OQP_2R and hence $P_2=P$ and $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$.

In the following two exercises the reader is asked to explain two of the other arithmetic rules.

|||| Exercise 10.13

Explain using the diagram the arithmetic rule $k(\mathbf{u}+\mathbf{v}) = k\mathbf{u}+k\mathbf{v}$.



|||| Exercise 10.14

Draw a figure that illustrates the rule $(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$.

For a given vector $\mathbf{u}$ it is obvious that the opposite vector $-\mathbf{u}$ is the only vector that satisfies the equation $\mathbf{u}+\mathbf{x} = \mathbf{0}$. For two arbitrary vectors $\mathbf{u}$ and $\mathbf{v}$ it is also obvious that exactly one vector exists that satisfies the equation $\mathbf{u}+\mathbf{x} = \mathbf{v}$, viz. the vector $\mathbf{x} = \mathbf{v}+(-\mathbf{u})$ which is illustrated in Figure 10.2.

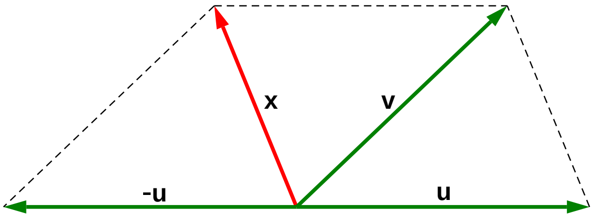


Figure 10.2: Opposite of a vector

Therefore we can introduce *subtraction of vectors* as a variation of addition like this:

Definition 10.15 Subtraction

By the difference of two vectors $\mathbf{v}$ and $\mathbf{u}$ we understand the vector

$$\mathbf{v} - \mathbf{u} = \mathbf{v} + (-\mathbf{u})\,. \tag{10-1}$$

It is not necessary to introduce a formal definition of *division* of a vector by a scalar, we consider this as a rewriting of multiplication by a scalar:

$$\text{Division by a scalar :} \qquad \frac{\mathbf{v}}{k} = \frac{1}{k} \cdot \mathbf{v}\ ;\ k \neq 0$$

10.2 Linear Combinations

A point about the arithmetic rule $(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$ from the theorem 10.12 is that parentheses can be left out in the process of adding a series of vectors, since it has no consequences for the resulting vector in what order the vectors are added. This is the background for *linear combinations* where sets of vectors are multiplied by scalars and thereafter written as a sum.

Definition 10.16 Linear Combination

When the real numbers $k_1, k_2, \ldots, k_n$ are given and in the plane or 3-space the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ then the sum

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \ldots + k_n\mathbf{v}_n$$

is called a ***linear combination*** of the n given vectors.

If all the coefficients $k_1, \cdots, k_n$ are equal to 0, the linear combination is called *improper*, or *trivial*, but if at the least one of the coefficients is different from 0, it is *proper*, or *non-trivial*.

|||| Example 10.17 Construction of a Linear Combination

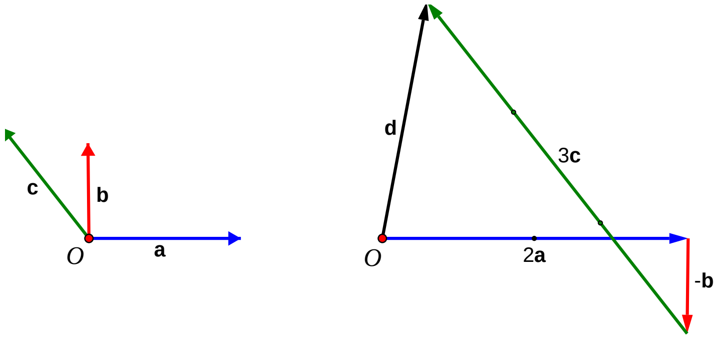


Figure: Construction of a linear combination

In the diagram, to the left the vectors **a**, **b** and **c** are drawn. On the figure to the right we have constructed the linear combination $\mathbf{d} = 2\mathbf{a} - \mathbf{b} + 3\mathbf{c}$.

|||| Exercise 10.18

There are given in the plane the vectors **u**, **v**, **s** and **t**, plus the parallelogram A, see diagram.

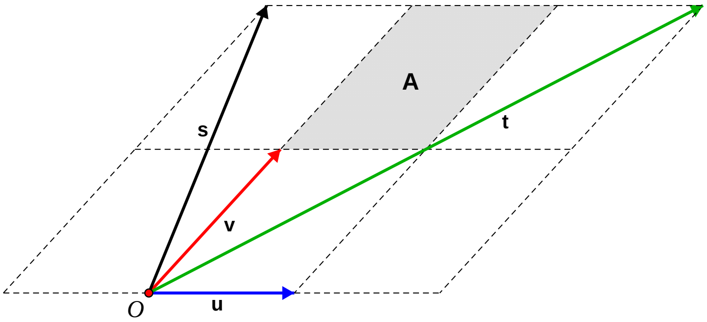


Figure: Linear combinations

1. Write **s** as a linear combination of **u** og **v**.

2. Show that **v** can be expressed by the linear combination $\mathbf{v} = \frac{1}{3}\,\mathbf{s} + \frac{1}{6}\,\mathbf{t}$.

3. Draw the linear combination $\mathbf{s} + 3\mathbf{u} - \mathbf{v}$.

4. Determine real numbers a, b, c and d such that A can be described by the *parametric*

representation

$$A = \{\, P \mid \overrightarrow{OP} = x\mathbf{u} + y\mathbf{v} \text{ with } x \in [\,a;\, b\,] \text{ and } y \in [\,c;\, d\,]\,\}\,.$$

10.3 Linear Dependence and Linear Independence

If two vectors have representations on the same straight line, one says that they are *linearly dependent*. It is evident that two proper vectors are linearly dependent if they are parallel; otherwise they are *linearly independent*. We can formulate it as follows: Two vectors **u** and **v** are linearly dependent if the one can be obtained from the other by multiplication by a scalar different from 0, if e.g. there exists a number $k \neq 0$ such that

$$\mathbf{v} = k\mathbf{u}\,.$$

We wish to generalize this original meaning of the concepts of linear dependence and independence such that the concepts can be used for an arbitrary set of vectors.

|||| Definition 10.19 Linear Dependence and Independence

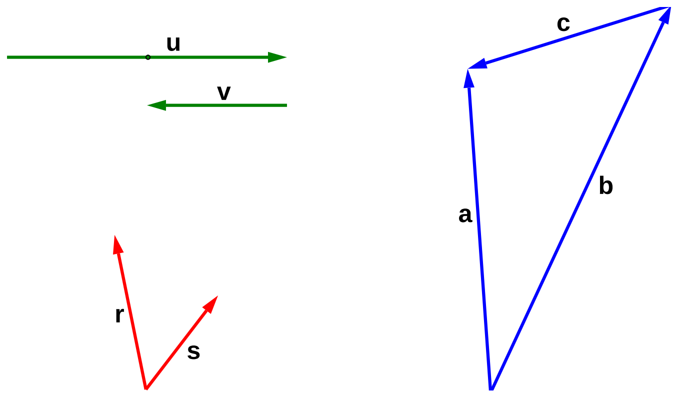
A set of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ where $n \geq 2$,is called ***linearly dependent*** if at least one of the vectors can be written as a linear combination of the others.

If none of the vectors can be written as a linear combination of the others, the set is called ***linearly independent***.

NB: A set that only consists of one vector is called linearly dependent if the vector is the **0**-vector, otherwise linearly independent.

|||| Example 10.20 Linearly Dependent and Linearly Independent Sets of Vectors

In the plane are given three sets of vectors $(\mathbf{u}, \mathbf{v})$, $(\mathbf{r}, \mathbf{s})$ and $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, as shown.



The set $(\mathbf{u}, \mathbf{v})$ is linearly dependent since for this example we have

$$\mathbf{u} = -2\mathbf{v}\,.$$

Also the set $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is linearly dependent, since e.g.

$$\mathbf{b} = \mathbf{a} - \mathbf{c}\,.$$

Only the set $(\mathbf{r}, \mathbf{s})$ is linearly independent.

|||| Exercise 10.21

Explain that three vectors in 3-space are linearly dependent if and only if they have representations lying in the same plane. What are the conditions three vectors must fulfill in order to be linearly independent?

|||| Exercise 10.22

Consider (intuitively) what is the maximum number of vectors a set of vectors in the plane can comprise, if the set is to be linearly independent. The same question in 3-space.

When investigate whether or not a given set of vectors is linearly independent or linearly dependent, the definition 10.19 does not give a practical procedure. It might be easier to use the theorem that follows below. This theorem is based on the fact that a set of vectors is linearly dependent if and only if the **0**-vector can be written as a proper linear combination of the vectors. Assume – as a prerequisite to the theorem – that the set $(\mathbf{a},\mathbf{b},\mathbf{c})$ is linearly dependent because

$$\mathbf{c} = 2\mathbf{a} - 3\mathbf{b}.$$

Then the **0**-vector can be written as the proper linear combination

$$2\mathbf{a} - 3\mathbf{b} - \mathbf{c} = \mathbf{0}\,.$$

Conversely assume that the **0**-vector is a proper linear combination of the vectors $\mathbf{u},\mathbf{v}$ og $\mathbf{w}$ like this:

$$2\mathbf{u} - 2\mathbf{v} + 3\mathbf{w} = \mathbf{0}\,.$$

Then we have (e.g.) that

$$\mathbf{w} = -\frac{2}{3}\mathbf{u} + \frac{2}{3}\mathbf{v}$$

and hence the vectors are linearly dependent.

|||| Theorem 10.23 Linear Independence

Let $k_1, k_2, \ldots, k_n$ be real numbers. That the set of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is linearly independent implies that the equation

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_n\mathbf{v}_n = \mathbf{0} \tag{10-2}$$

is only satisfied when all the coefficients $k_1, k_2, \ldots, k_n$ are equal to 0.

|||| Proof

Assume that the set $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is linearly dependent, and let v_i be a vector that can be written as a linear combination of the other vectors. We reorder (if necessary) the set, such that $i = 1$, following which $\mathbf{v}_1$ can be written in the form

$$\mathbf{v}_1 = k_2\mathbf{v}_2 + \cdots + k_n\mathbf{v}_n \;\Leftrightarrow\; \mathbf{v}_1 - k_2\mathbf{v}_2 - \cdots - k_n\mathbf{v}_n = \mathbf{0}\,. \tag{10-3}$$

The **0**-vector is hereby written in the form (10-2), in which not all the coefficients are 0, because the coefficient to $\mathbf{v}_1$ is 1.

Conversely, assume that the set is written in the form (10-2), and let $k_i \neq 0$. We reorder (if necessary) the set such that $i = 1$ following which we have

$$k_1\mathbf{v}_1 = -k_2\mathbf{v}_2 - \cdots - k_n\mathbf{v}_n \;\Leftrightarrow\; \mathbf{v}_1 = -\frac{k_2}{k_1}\mathbf{v}_2 - \cdots - \frac{k_n}{k_1}\mathbf{v}_n\,. \tag{10-4}$$

From this we see that the set is linearly independent.

■

|||| Example 10.24 Linearly Independent Set

Every set of vectors containing the zero vector is linearly dependent. Consider e.g. the set $(\mathbf{u}, \mathbf{v}, \mathbf{0}, \mathbf{w})$. It is obvious that the zero-vector can be written as the other three vectors:

$$\mathbf{0} = 0\mathbf{u} + 0\mathbf{v} + 0\mathbf{w}\,,$$

where the zerovector is written as a linear combination of the other vectors in the set.

Parametric representations for planes in 3-space is written using two linearly independent vectors. Below we first give an example of a plane through the origin, then an example of a plane that does not contain the origin.

|||| Example 10.25 Parametic Representation for a Plane

Given a plane in 3-space through the origin as shown. We wish to describe the points in the plane by a ***parametric representation***.

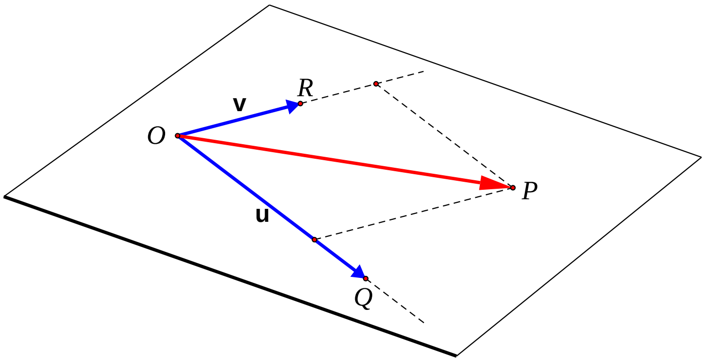


Figure: A plane in 3-space through the origin

In the given plane we choose two points Q and R, both not the origin, and that do not lie on a common line through the origin. The vectors $\mathbf{u} = \overrightarrow{OQ}$ and $\mathbf{v} = \overrightarrow{OR}$ will then be linearly independent, and are called ***direction vectors*** of the plane. For every point P in the plane we have exactly one pair of numbers (s,t) such that $\overrightarrow{OP} = s\mathbf{u} + t\mathbf{v}$. Conversely, for every pair of real numbers (s,t) exists exactly one point P in the plane that satisfies $\overrightarrow{OP} = s\mathbf{u} + t\mathbf{v}$. Then

$$\{P \mid \overrightarrow{OP} = s\mathbf{u} + t\mathbf{v}\,;\ (s,t) \in \mathbb{R}^2\}$$

is a parametric representation of the given plane.

|||| Example 10.26 Parametric Representation for a Plane

A plane in 3-space does not contain the origin. We wish to describe the plane using a parametric representation.

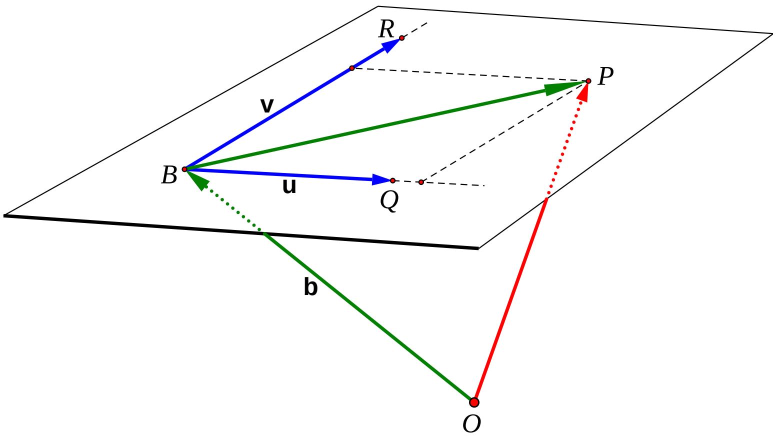


Figure: A plane in 3-space

First we choose an initial point B in the plane, and we put $\mathbf{b} = \overrightarrow{OB}$. Then we choose two linearly independent direction vectors $\mathbf{u} = \overrightarrow{BQ}$ and $\mathbf{v} = \overrightarrow{BR}$ where Q and R belong to the plane. To every point P in the plane corresponds exactly one pair of real numbers (s,t), such that

$$\overrightarrow{OP} = \overrightarrow{OB} + \overrightarrow{BP} = \mathbf{b} + s\mathbf{u} + t\mathbf{v}\,.$$

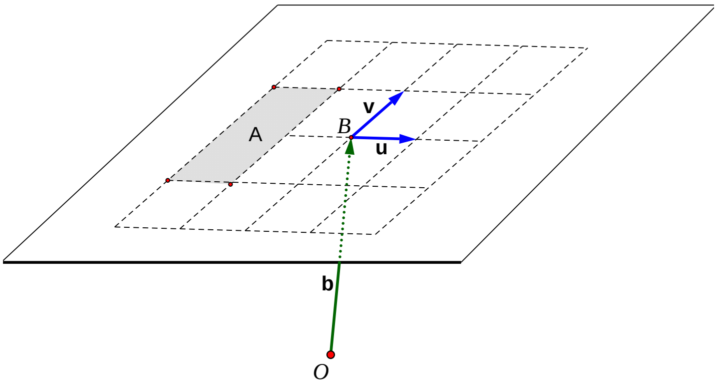
Conversely, to every pair of real numbers (s,t) corresponds exactly one point P in the plane as given by this vector equation. Then

$$\{P \mid \overrightarrow{OP} = \mathbf{b} + s\mathbf{u} + t\mathbf{v}\,;\ (s,t) \in \mathbb{R}^2\}$$

is a parametric representation for the given plane.

|||| Exercise 10.27

Give a parametric representation for the parallelogram A lying in the plane shown:



10.4 The Standard Bases in the Plane and Space

In *analytic geometry* one shows how numbers and equations can describe geometric objects and phenomena including vectors. Here the concept of coordinates is decisive. It is about how we determine the position of the geometric objects in 3-space and relative to one another using numbers and tuples of numbers. To do so we need to choose a number of vectors which we appoint as ***basis vectors***. The basis vectors ***are ordered,*** that is they are given a distinct order, and thus they constitute a ***basis***. When a basis is given all the vectors can be described using coordinates, which we assemble in so called coordinate vectors. How this whole procedure takes place we first explain for the standard bases in the plane and 3-space. Later we show that often it is useful to use other bases than the standard bases and how the coordinates of a vector in different bases are related.

|||| Definition 10.28 Standard Basis in the Plane

By a ***standard basis*** or an ***ordinary basis*** for the geometric vectors in the plane we understand an ordered set of two vectors $(\mathbf{i},\mathbf{j})$ that satisfies:

- $\mathbf{i}$ has the length 1 .
- $\mathbf{j}=\widehat{\mathbf{i}}$ (that is $\mathbf{j}$ is the hat vector of $\mathbf{i}$).

By a ***standard coordinate system in the plane*** we understand a standard basis $(\mathbf{i},\mathbf{j})$ together with a chosen the origin O . The coordinate system is written $(O,\mathbf{i},\mathbf{j})$. By the x-axis and the y-axis we understand oriented number axes through O that are parallel to $\mathbf{i}$ and $\mathbf{j}$,respectively.

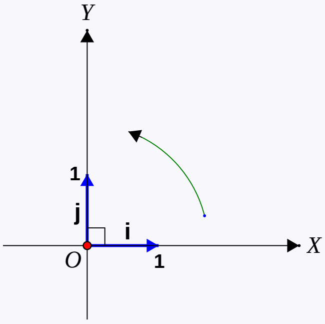


Figure: Standard coordinate system in the plane

|||| Theorem 10.29 Coordinates of a Vector

If $e=(\mathbf{i},\mathbf{j})$ is a standard basis, then any vector $\mathbf{v}$ in the plane can be written in exactly one way as a linear combination of $\mathbf{i}$ and $\mathbf{j}$:

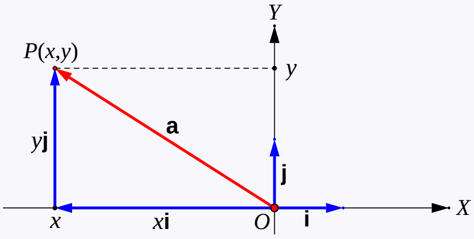
$$\mathbf{v}=x\mathbf{i}+y\mathbf{j}.$$

The coefficients x and y in the linear combination are called $\mathbf{v}$'s *coordinates with respect to the basis e*, or for short $\mathbf{v}$'s e-coordinates, and they are assembled in a ***coordinate vector*** as follows:

$$_{e}\mathbf{v}=\begin{bmatrix}x\\y\end{bmatrix}.$$

|||| Definition 10.30 The Coordinates of a Point

Let P be any point in the plane, and let $(O,\mathbf{i},\mathbf{j})$ be a standard coordinate system in the plane. By the coordinates of P with respect to the coordinate system we understand the coordinates of the position vector $\overrightarrow{OP}$ with respect to the standard basis $(\mathbf{i},\mathbf{j})$.



The introduction of a standard basis and the coordinates of a vector in 3-space is a simple extension of the corresponding coordinates in the plane.

|||| Definition 10.31 Standard Basis in Space

By a ***standard basis*** or an ***ordinary basis*** for the geometric vectors in 3-space we understand an ordered set of three vectors $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ that satisfies:

- $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ all have the length 1.
- $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are pairwise orthogonal.
- When $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are drawn from a chosen point, and we view $\mathbf{i}$ and $\mathbf{j}$ from the endpoint of $\mathbf{k}$, then $\mathbf{i}$ turns into $\mathbf{j}$, when $\mathbf{i}$ is turned by the angle $\frac{\pi}{2}$ counter-clockwise.

By n ***standard coordinate system in 3-space*** we understand a standard basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ together with a chosen the origin O . The coordinate system is written $(O, \mathbf{i}, \mathbf{j}, \mathbf{k})$. By the x-axis, the y-axis and the z-axis we understand oriented number axes through the origin that are parallel to $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$, respectively.

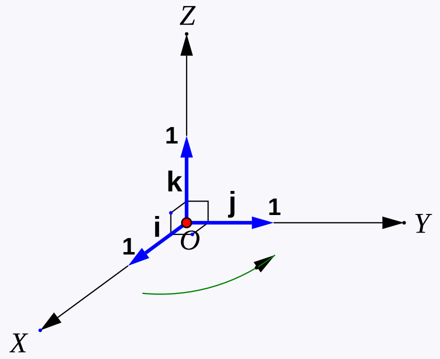


Figure: A standard coordinate system in 3-space.

|||| Theorem 10.32 The Coordinates of a Vector

When $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ is a basis, every vector $\mathbf{v}$ in 3-space can be written in exactly one way as a linear combination of $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$:

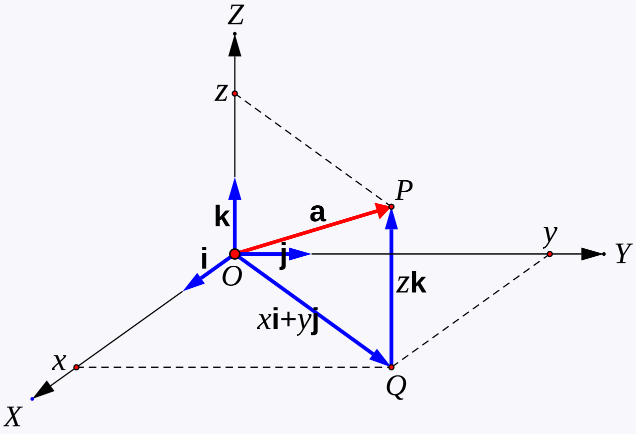
$$\mathbf{v} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

The coefficients x, y and z in the linear combination are called $\mathbf{v}$'s *coordinates with respect to the basis*, or in short $\mathbf{v}$'s e-coordinates, and they are assembled in a ***coordinate vectodr*** as follows:

$$_{\mathrm{e}}\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

|||| Definition 10.33 The Coordinates of a Point

Let P be an arbitrary point in 3-space, and let $(O, \mathbf{i}, \mathbf{j}, \mathbf{k})$ be a standard coordinate system in 3-space. By the coordinates of P with respect to the coordinate system we understand the coordinates of the position vector $\overrightarrow{OP}$ with respect to the standard basis $(\mathbf{i}, \mathbf{j}, \mathbf{k})$.



10.5 Arbitrary Bases for the Plane and Space

If two linearly independent vectors in the plane are given, it is possible to write every other vector as a linear combination of the two given vectors. In Figure 10.3 we consider e.g. the two linearly independent vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ plus two other vectors $\mathbf{u}$ and $\mathbf{v}$: in the plane

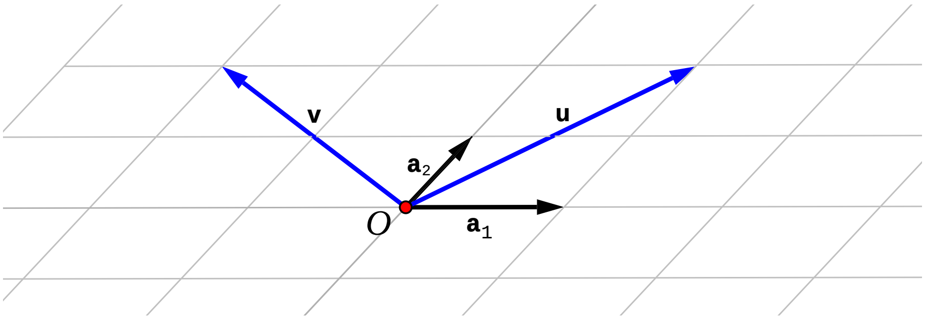


Figure 10.3: Coordinate system in the plane with basis $(\mathbf{a}_1, \mathbf{a}_2)$

We see that $\mathbf{u} = 1\mathbf{a}_1 + 2\mathbf{a}_2$ and $\mathbf{v} = -2\mathbf{a}_1 + 2\mathbf{a}_2$. These linear combinations are unique because $\mathbf{u}$ and $\mathbf{v}$ cannot be written as a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$ using any other coefficients than those written. Similarly, any other vector in the plane can be written as a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$, and our term for this is that the two vectors ***span*** the whole plane.

This makes it possible to generalise the concept of a basis. Instead of a standard basis we can choose to use the set of vectors $(\mathbf{a}_1, \mathbf{a}_2)$ as a basis for the vectors in the plane. If we call the basis a , we say that the coefficients in the linear combinations above are *coordinates* for $\mathbf{u}$ and $\mathbf{v}$, respectively, *with respect to a basis a*, which is written like this:

$$_{a}\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } _{a}\mathbf{v} = \begin{bmatrix} -2 \\ 2 \end{bmatrix} . \tag{10-5}$$

For the set of geometric vectors in 3-space we proceed in a similar way. Given three linearly independent vectors, then every vector in 3-space can be written as a unique-linear combination of the three given vectors. They *span* all of 3-space. Therefore we can choose three vectors as a basis for the vectors in 3-space and express an arbitrary vector in 3-space by coordinates with respect to this basis. A method for determination of the coordinates is shown in Figure 10.4, where we are given an a-basis $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ plus

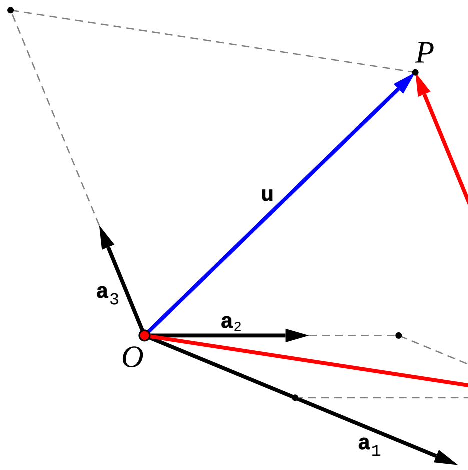


Figure 10.4: Coordinate system with basis $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$

an arbitrary vector $\mathbf{u}$. Through the endpoint P for $\mathbf{u}$ a line parallel to $\mathbf{a}_3$ is drawn, and the point of intersection of this line and the plane that contains $\mathbf{a}_1$ and $\mathbf{a}_2$, is denoted Q. Two numbers k_1 and k_2 exist such that $\overrightarrow{OQ}= k_1\mathbf{a}_1 + k_2\mathbf{a}_2$ because $(\mathbf{a}_1, \mathbf{a}_2)$ constitutes a basis in the plane that contains $\mathbf{a}_1$ and $\mathbf{a}_2$. Furthermore there exists a number k_3 such that $\overrightarrow{QP}= k_3\mathbf{a}_3$ since $\overrightarrow{QP}$ and $\mathbf{a}_3$ are parallel. But then we have

$$\mathbf{u} =\overrightarrow{OQ} + \overrightarrow{QP}= k_1\mathbf{a}_1 + k_2\mathbf{a}_2 + k_3\mathbf{a}_3.$$

$\mathbf{u}$ thereby has the coordinate set (k_1, k_2, k_3) with respect to basis a.

|||| Example 10.34 Coordinates with Respect to an Arbitrary Basis

In 3-space three linearly independent vectors $\mathbf{a}_1, \mathbf{a}_2$ and $\mathbf{a}_3$ are given as shown in the Figure.

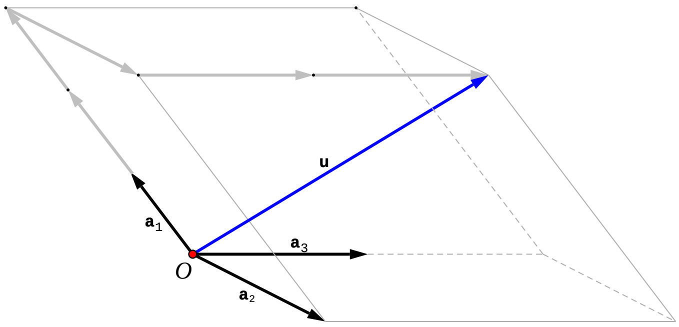


Figure: Coordinate system with basis $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$

Since $\mathbf{u}$ can be written as a linear combination of $\mathbf{a}_1, \mathbf{a}_2$ and $\mathbf{a}_3$ in the following way

$$\mathbf{u} = 3\mathbf{a}_1 + \mathbf{a}_2 + 2\mathbf{a}_3\,, \tag{10-6}$$

then $\mathbf{u}$ has the coordinates $(3, 1, 2)$ with respect to the basis a given by $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ which we write in short as

$$_{\mathrm{a}}\mathbf{u} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}. \tag{10-7}$$

We gather the above considerations about arbitrary bases in the following more formal definition:

|||| Definition 10.35 The Coordinates of a Vector with Respect to a Basis

- By a basis a for the geometric vectors in the plane we will understand an arbitrary ordered set of two linear independent vectors $(\mathbf{a}_1, \mathbf{a}_2)$. Let an arbitrary vector $\mathbf{u}$ be determined by the linear combination $\mathbf{u} = x\mathbf{a}_1 + y\mathbf{a}_2$. The coefficients x and y are called $\mathbf{u}$'s *coordinates with respect to the basis* a, or shorter $\mathbf{u}$'s a-coordinates, and they are assembled in a *coordinate vector* as follows:

$$_{\mathrm{a}}\mathbf{u} = \begin{bmatrix} x \\ y \end{bmatrix}. \tag{10-8}$$

- By a basis b for the geometric vectors in 3-space we understand an arbitrary ordered set of three linear independent vectors $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$. Let an arbitrary vector $\mathbf{v}$ be determined by the linearly combinationen $\mathbf{v} = x\mathbf{b}_1 + y\mathbf{b}_2 + z\mathbf{b}_3$. The coefficients x, y and z are called $\mathbf{v}$'s *coordinates with respect to the basis* b, or shorter $\mathbf{v}$'s b-coordinates, and they are assembled in a *coordinate vector* as follows:

$$_{\mathrm{b}}\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{10-9}$$

The coordinate set of a given vector will change when we change the basis. This crucial point is the subject of the following exercise.

|||| Exercise 10.36

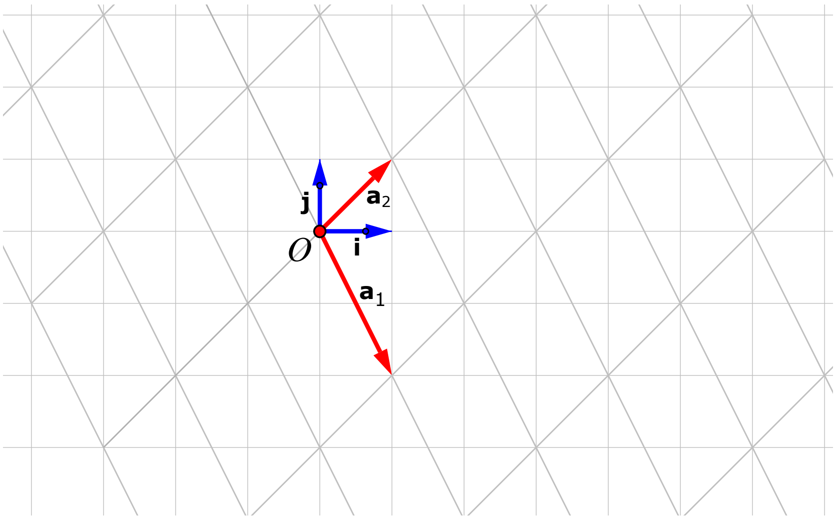


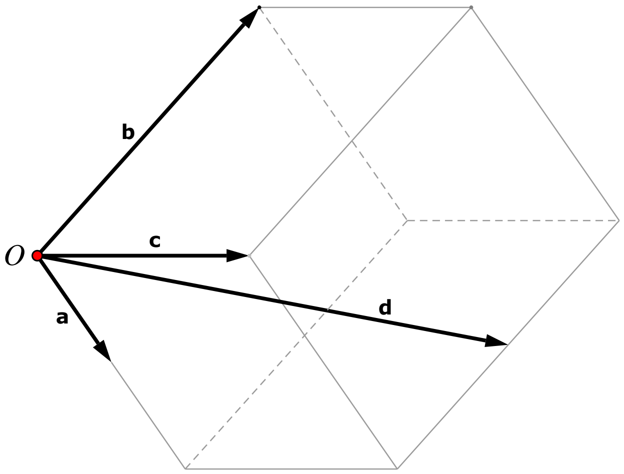
Figure: Change of basis

In the diagram, we are given the standard basis $e = (\mathbf{i}, \mathbf{j})$ in the plane plus another basis $a = (\mathbf{a}_1, \mathbf{a}_2)$.

1. A vector $\mathbf{u}$ has the coordinates $(5, -1)$ with respect to basis e. Determine $\mathbf{u}$'s a-coordinates.

2. A vector $\mathbf{v}$ has the coordinates $(-1, -2)$ with respect to basis a. Determine $\mathbf{v}$'s e-coordinates.

|||| **Exercise 10.37**



1. In the diagram, it is evident that **a**, **b** and **c** are linearly independent. A basis m is therefore given by $(\mathbf{a}, \mathbf{b}, \mathbf{c})$. Determine the coordinate vector ${}_{\mathrm{m}}\mathbf{d}$.

2. It is also evident from the figure that $(\mathbf{a}, \mathbf{b}, \mathbf{d})$ is a basis, let us call it n. Determine the coordinate vector ${}_{\mathrm{n}}\mathbf{c}$.

3. Draw, with the origin as the initial point, the vector **u** that has the m-coordinates

$$ {}_{\mathrm{m}}\mathbf{u} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} . $$

10.6 Vector Calculations Using Coordinates

When you have chosen a basis for geometric vectors in the plane (or in 3-space), then all vectors can be described and determined using their coordinates with respect to the chosen basis. For the two arithmetic operations, addition and multiplication by a scalar, that were introduced previously in this eNote by geometrical construction, we get a particularly practical alternative. Instead of geometrical constructions we can carry out calculations with the coordinates that correspond to the chosen basis.

We illustrate this with an example in the plane with a basis a given by $(\mathbf{a}_1, \mathbf{a}_2)$ plus two vectors $\mathbf{u}$ and $\mathbf{v}$ drawn from O, see Figure 10.5. The exercise is to determine the vector $\mathbf{b} = 2\mathbf{u} - \mathbf{v}$, and we will do this in two different ways.

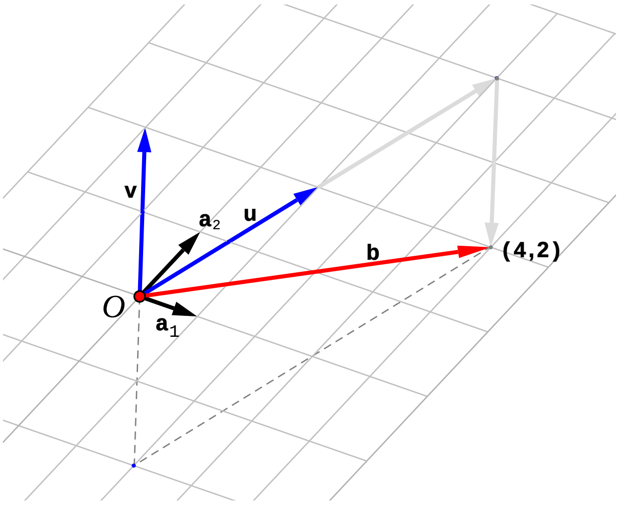


Figure 10.5: Linear combination determined using coordinates

Method 1 (geometric): First we carry through the arithmetic operations as defined in 10.2 and 10.3, cf. the grey construction vectors in Figure 10.5.

Method 2 (algebraic): We read the coordinates for $\mathbf{u}$ and $\mathbf{v}$ and carry out the arithmetic operations directly on the coordinates:

$$ {}_{\mathrm{a}}\mathbf{b} = 2\,{}_{\mathrm{a}}\mathbf{u} - {}_{\mathrm{a}}\,\mathbf{v} = 2\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}. \tag{10-10} $$

Now $\mathbf{b}$ can be drawn directly from its coordinates $(4, 2)$ with respect to basis a.

That it is allowed to use this method is stated in the following theorem.

|||| Theorem 10.38 Basic Rules for Coordinate Calculations

Two vectors $\mathbf{u}$ and $\mathbf{v}$ in the plane or in 3-space plus a real number k are given. Moreover, an arbitrary basis a has been chosen.The two arithmetic operations $\mathbf{u}+\mathbf{v}$ and $k\,\mathbf{u}$ can then be carried out using coordinates as follows:

1. ${}_{\mathrm{a}}(\mathbf{u}+\mathbf{v}) = {}_{\mathrm{a}}\mathbf{u} + {}_{\mathrm{a}}\mathbf{v}$

2. ${}_{\mathrm{a}}(k\mathbf{u}) = k\,{}_{\mathrm{a}}\mathbf{u}$

In other words: the coordinates for a vector sum are obtained by adding the coordinates for the summands. And the coordinates for a vector multiplied by a number are the coordinates of the vector multiplied by that number.

|||| Proof

We carry through the proof for the set of geometric vectors in 3-space. Suppose the coordinates for $\mathbf{u}$ and $\mathbf{v}$ with respect to the chosen basis a are given by

$$ {}_{\mathrm{a}}\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \text{ and } {}_{\mathrm{a}}\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}. \tag{10-11} $$

We then have

$$ \mathbf{u} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3 \text{ og } \mathbf{v} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + v_3\mathbf{a}_3 \tag{10-12} $$

and accordingly, through the application of the commutative, associative and distributive arithmetic rules, see Theorem10.12,

$$ \begin{aligned} \mathbf{u}+\mathbf{v} &= (u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3) + (v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + v_3\mathbf{a}_3) \\ &= (u_1+v_1)\mathbf{a}_1 + (u_2+v_2)\mathbf{a}_2 + (u_3+v_3)\mathbf{a}_3 \end{aligned} \tag{10-13} $$

which yields

$$ {}_{\mathrm{a}}(\mathbf{u}+\mathbf{v}) = \begin{bmatrix} u_1+v_1 \\ u_2+v_2 \\ u_3+v_3 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = {}_{\mathrm{a}}\mathbf{u} + {}_{\mathrm{a}}\mathbf{v} \tag{10-14} $$

so that now the first part of the proof is complete. In the second part of the proof we again use a distributive arithmetic rule, see Theorem 10.12:

$$ k\mathbf{u} = k(u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3) = (k\cdot u_1)\mathbf{a}_1 + (k\cdot u_2)\mathbf{a}_2 + (k\cdot u_3)\mathbf{a}_3 \tag{10-15} $$

which yields

$$ {}_{\mathrm{a}}(k\mathbf{u}) = \begin{bmatrix} k\cdot u_1 \\ k\cdot u_2 \\ k\cdot u_3 \end{bmatrix} = k\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = k\,{}_{\mathrm{a}}\mathbf{u} \tag{10-16} $$

so that now the second part of the proof is complete.

■

Theorem 10.38 makes it possible to perform more complicated arithmetic operations using coordinates, as shown in the following example.

|||| Example 10.39 Coordinate Vectors for a Linear Combination

The three plane vectors **a**, **b** and **c** have the following coordinate vectors with respect to a chosen basis v:

$$ {}_{\mathrm{v}}\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},\ {}_{\mathrm{v}}\mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ and } {}_{\mathrm{v}}\mathbf{c} = \begin{bmatrix} 5 \\ -1 \end{bmatrix}. \tag{10-17}$$

Problem: Determine the coordinate vector $\mathbf{d} = \mathbf{a} - 2\mathbf{b} + 3\mathbf{c}$ with respect to basis v.
Solution:

$$\begin{aligned} {}_{\mathrm{v}}\mathbf{d} &= {}_{\mathrm{v}}(\mathbf{a} - 2\mathbf{b} + 3\mathbf{c}) \\ &= {}_{\mathrm{v}}(\mathbf{a} + (-2)\mathbf{b} + 3\mathbf{c}) \\ &= {}_{\mathrm{v}}\mathbf{a} + {}_{\mathrm{v}}(-2\mathbf{b}) + {}_{\mathrm{v}}(3\mathbf{c}) \\ &= {}_{\mathrm{v}}\mathbf{a} - 2\,{}_{\mathrm{v}}\mathbf{b} + 3\,{}_{\mathrm{v}}\mathbf{c} \\ &= \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 2\begin{bmatrix} 0 \\ 1 \end{bmatrix} + 3\begin{bmatrix} 5 \\ -1 \end{bmatrix} = \begin{bmatrix} 16 \\ -3 \end{bmatrix}. \end{aligned}$$

Here the third equality sign is obtained using the first part of Theorem 10.38 and the fourth equality sign from the second part of that theorem.

|||| Example 10.40 The Parametric Representation of a Plane in Coordinates

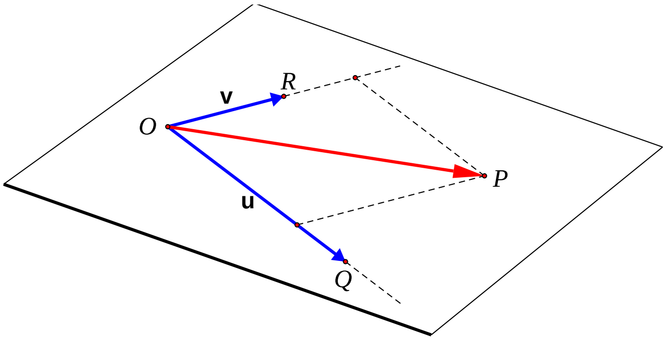


Figure: A plane in 3-space

In accordance with Example 10.25, the plane through the origin shown in the diagram has the parametric representation

$$\{P \mid \overrightarrow{OP} = s\mathbf{u} + t\mathbf{v}\,;\ (s,t) \in \mathbb{R}^2\}. \tag{10-18}$$

Suppose that in 3-space we are given a basis a and that

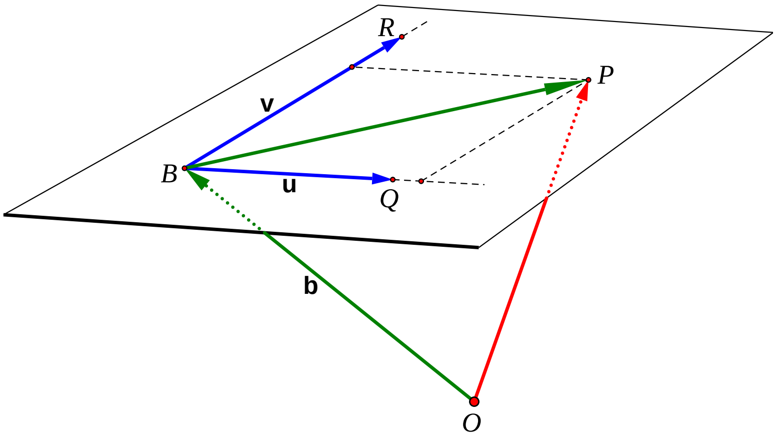
$$_{\text{a}}\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \text{ and } {_{\text{a}}\mathbf{v}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

The parametric representation (10-18) can then be written in coordinate form like this:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = s \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + t \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \tag{10-19}$$

where $_{\text{a}}\overrightarrow{OP} = (x, y, z)$ and $(s,t) \in \mathbb{R}^2$.

|||| Example 10.41 The Parametric Representation of a Plane in Coordinates



In accordance with Example 10.26 the plane through the origin shown in the diagram has the parametric representation

$$\{P \mid \overrightarrow{OP} = \mathbf{b} + s\mathbf{u} + t\mathbf{v}\,;\ (s,t) \in \mathbb{R}^2\}. \tag{10-20}$$

Suppose that in 3-space we are given a basis a and that

$$_{\text{a}}\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, {_{\text{a}}\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \text{ and } {_{\text{a}}\mathbf{v}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

The parametric representation (10-18) can then be written in coordinate form like this:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + s \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + t \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \tag{10-21}$$

where ${}_{\mathrm{a}}\overrightarrow{OP} = (x, y, z)$ and $(s, t) \in \mathbb{R}^2$

10.7 Vector Equations and Matrix Algebra

A large number of vector-related problems are best solved by resorting to vector equations. If we wish to solve these equations using the vector coordinates in a given basis, we get systems of linear equations. The problems can then be solved using matrix methods that follow in eNote 6. This subsection gives examples of this and sums up this approach by introducing the *coordinate matrix* concept in the final Exercise 10.45.

|||| Example 10.42 Whether a Vector is a Linear Combination of Other Vectors

In 3-space are given a basis a and three vectors $\mathbf{u}, \mathbf{v}$ and $\mathbf{p}$ which have the coordinates with respect to the basis a given by:

$${}_{\mathrm{a}}\mathbf{u} = \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix},\ {}_{\mathrm{a}}\mathbf{v} = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \text{ and } {}_{\mathrm{a}}\mathbf{p} = \begin{bmatrix} 0 \\ 7 \\ 1 \end{bmatrix}.$$

Problem: Investigate whether $\mathbf{p}$ is a linear combination of $\mathbf{u}$ and $\mathbf{v}$.

Solution: We will investigate whether we can find coefficients k_1, k_2, such that

$$k_1\mathbf{u} + k_2\mathbf{v} = \mathbf{p}\,.$$

We arrange the corresponding coordinate vector equation

$$k_1 \begin{bmatrix} 2 \\ 1 \\ 5 \end{bmatrix} + k_2 \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 7 \\ 1 \end{bmatrix}$$

which is equivalent to the following system of equations

$$\begin{aligned} 2k_1 + k_2 &= 0 \\ k_1 + 4k_2 &= 7 \\ 5k_1 + 3k_2 &= 1 \end{aligned} \tag{10-22}$$

We consider the augmented matrix $\mathbf{T}$ for the system of equations and give (without details) the reduced row echelon form of the matrix:

$$\mathbf{T} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 7 \\ 5 & 3 & 1 \end{bmatrix} \rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \tag{10-23}$$

We see that the system of equations has exactly one solution, $k_1 = -1$ and $k_2 = 2$, meaning that

$$-1\mathbf{u} + 2\mathbf{v} = \mathbf{p}\,.$$

|||| Example 10.43 Whether a Set of Vectors is Linearly Dependent

In 3-space are given a basis v and three vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ which with respect to this basis have the coordinates

$$_{\mathrm{v}}\mathbf{a} = \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix},\ _{\mathrm{v}}\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix} \text{ and } _{\mathrm{v}}\mathbf{c} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}.$$

Problem: Investigate whether the set of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is linearly dependent.

Solution: Following theorem 10.23 we can investigate whether there exists a proper linear combination

$$k_1\mathbf{a} + k_2\mathbf{b} + k_3\mathbf{c} = \mathbf{0}\,.$$

We look at the corresponding coordinate vector equation

$$k_1 \begin{bmatrix} 5 \\ 1 \\ 3 \end{bmatrix} + k_2 \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix} + k_3 \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

that is equivalent to the following homogeneous system of linear equations

$$\begin{aligned} 5k_1 + k_2 + 2k_3 &= 0 \\ k_1 + 3k_3 &= 0 \\ 3k_1 + 4k_2 + k_3 &= 0 \end{aligned} \tag{10-24}$$

We arrange the augmented matrix $\mathbf{T}$ of the system of equations and give (without details) the reduced row echelon form of the matrix:

$$\mathbf{T} = \begin{bmatrix} 5 & 1 & 2 & 0 \\ 1 & 0 & 3 & 0 \\ 3 & 4 & 1 & 0 \end{bmatrix} \rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{10-25}$$

We see that the system of equations only have the zero solution $k_1 = 0, k_2 = 0$ and $k_3 = 0$. The set of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is therefore linearly independent. Therefore you may choose $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ as a new basis for the set of vectors in 3-space.

In the following example we continue the discussion of the relation between coordinates and change of basis from exercise 10.36

|||| Example 10.44 The New Coordinates after Change of Basis

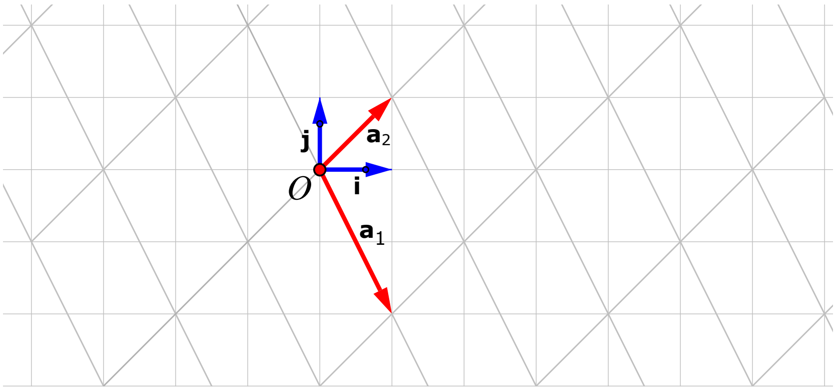


Figure: Change of basis

In the diagram we are given a standard basis e= $(\mathbf{i}, \mathbf{j})$ and another basis a= $(\mathbf{a}_1, \mathbf{a}_2)$. When the basis is changed, the coordinates of any given vector are changed. Here we give a systematic method for expressing the change in coordinates using a matrix-vector product. First we read the e-coordinates of the vectors in basis a:

$$_{\mathrm{e}}\mathbf{a_1} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \text{ and } _{\mathrm{e}}\mathbf{a_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{10-26}$$

1. *Problem*: Suppose a vector $\mathbf{v}$ has the set of coordinates $_{\mathrm{a}}\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$. Determine the e-coordinates of $\mathbf{v}$.

   Solution: We have that $\mathbf{v} = v_1\mathbf{a}_1 + v_2\mathbf{a}_2$ and therefore following Theorem 10.38:

   $$_{\mathrm{e}}\mathbf{v} = v_1\begin{bmatrix} 1 \\ -2 \end{bmatrix} + v_2\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

   If we put $\mathbf{M} = \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}$, we express $\mathbf{v}$'s e-coordinates by the matrix-vector product

   $$_{\mathrm{e}}\mathbf{v} = \mathbf{M} \cdot {_{\mathrm{a}}\mathbf{v}} \tag{10-27}$$

2. *Problem*: Suppose a vector $\mathbf{v}$ has the set of coordinates $_{\mathrm{e}}\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$. Determine the a-coordinates of $\mathbf{v}$.

Solution: We multiply from the left on both sides of 10-27 with the inverse matrix to **M** and get a-coordinates of **v** expressed by the matrix-vector product:

$$ {}_{\mathrm{a}}\mathbf{v} = \mathbf{M}^{-1} \cdot {}_{\mathrm{e}}\mathbf{v} \tag{10-28}$$

|||| Exercise 10.45

By a ***coordinate matrix*** with respect to a given basis *a* for a set of vectors me mean the matrix that is formed by combining the vector's a-coordinate columns to form a matrix.
Describe the matrix **T** in example 10.42 and 10.43 and the matrix **M** in 10.44 as coordinate matrices.

10.8 Theorems about Vectors in a Standard Basis

In this subsection we work with standard coordinate systems, both in the plane and in 3-space. We introduce two different multiplications between vectors, the *dot product* which is defined both in the plane and in 3-space, and the *cross product* that is only defined in 3-space. We look at geometric applications of these types of multiplication and at geometrical interpretations of determinants.

10.8.1 The Dot Product of two Vectors

|||| Definition 10.46 The Dot Product in the Plane

In the plane are given two vectors ${}_e\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and ${}_e\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. By the *dot product* (or the *scalar product*) of **a** and **b** we refer to the number

$$\mathbf{a} \cdot \mathbf{b} = a_1 \cdot b_1 + a_2 \cdot b_2 \,. \tag{10-29}$$

|||| Definition 10.47 The Dot Product in Space

In 3-space are given two vectors ${}_e\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and ${}_e\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. By the *dot product* (or the *scalar product*) of **a** and **b** we understand the number

$$\mathbf{a} \cdot \mathbf{b} = a_1 \cdot b_1 + a_2 \cdot b_2 + a_3 \cdot b_3 \,. \tag{10-30}$$

For the dot product between two vectors the following rules of calculation apply.

|||| Theorem 10.48 Arithmetic Rules for the Dot Product

Given three vectors **a**, **b** and **c** in the plane or in 3-space and the number k. Observe:

1. $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ (commutative rule)
2. $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$ (associative rule)
3. $(k\mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (k\mathbf{b}) = k(\mathbf{a} \cdot \mathbf{b})$
4. $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$
5. $|\mathbf{a} + \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 + 2\mathbf{a} \cdot \mathbf{b}$.

|||| Proof

The Rules 1, 2, 3 follow from a simple coordinate calculation. Rule 4 follows from the Pythagorean Theorem, and Rule 5 is a direct consequence of Rules 1, 2 and 4.

■

In the following three theorems we look at geometric applications of the dot product.

|||| Theorem 10.49 The Length of a Vector

Let **v** be an arbitrary vector in the plane or in 3-space. The length of **v** satisfies

$$|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}}\,. \tag{10-31}$$

|||| Proof

The theorem follows immediately from the arithmetic Rule 4 in 10.48

■

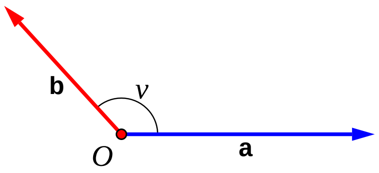


Figure 10.6: Angle between two vectors

|||| Example 10.50 Length of a Vector

Given the vector **v** in 3-space and ${}_e\mathbf{v} = (1, 2, 3)$. We then have

$$|\mathbf{v}| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}\,.$$

The following fact concerns the angle between two vectors, see Figure 10.6.

|||| Theorem 10.51 The Angle between Vectors

In the plane or 3-space we are given two proper vectors **a** and **b**. The angle v between **a** and **b** satisfies

$$\cos(v) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \tag{10-32}$$

|||| Proof

The theorem can be proved using the cosine relation. In carrying out the proof one needs Rule 5 in theorem 10.48. The details are left for the reader.

■

From this theorem it follows directly:

|||| Corollary 10.52 The Size of Angles

Consider the situation in Figure 10.6. We see

1. $\mathbf{a} \cdot \mathbf{b} = 0 \Leftrightarrow \text{angle}(\mathbf{a}, \mathbf{b}) = \frac{\pi}{2}$
2. $\mathbf{a} \cdot \mathbf{b} > 0 \Leftrightarrow \text{angle}(\mathbf{a}, \mathbf{b}) < \frac{\pi}{2}$
3. $\mathbf{a} \cdot \mathbf{b} < 0 \Leftrightarrow \text{angle}(\mathbf{a}, \mathbf{b}) > \frac{\pi}{2}$

The following theorems are dedicated to ***orthogonal projections***. In Figure 10.7 two vectors **a** and **b** in the plane or 3-space are drawn from the origin.

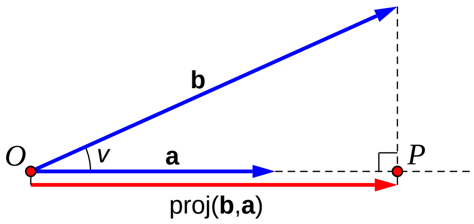


Figure 10.7: Orthogonal projection

Consider P, the foot of the perpendicular from **b**'s endpoint to the line containing **a**. By the orthogonal projection of **b** onto **a** we mean the vector $\overrightarrow{OP}$, denoted $\text{proj}(\mathbf{b}, \mathbf{a})$.

|||| Theorem 10.53 The Length of a Projection

Given two proper vectors **a** and **b** in the plane or 3-space. The length of the orthogonal projection of **b** onto **a** is:

$$|\text{proj}(\mathbf{b}, \mathbf{a})| = \frac{|\mathbf{a} \cdot \mathbf{b}|}{|\mathbf{a}|} \tag{10-33}$$

|||| Proof

Using a known theorem about right angled triangles plus Theorem 10.51 we get

$$|\text{proj}(\mathbf{b},\mathbf{a})| = |\cos(v)|\,|\mathbf{b}| = \frac{|\mathbf{a}\cdot\mathbf{b}|}{|\mathbf{a}|}\,.$$

■

|||| Theorem 10.54 Formula for the Projection Vector

Given two proper vectors **a** and **b** in the plane or 3-space. The orthogonal projection of **b** on **a** is:

$$\text{proj}(\mathbf{b},\mathbf{a}) = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\,\mathbf{a}\,. \tag{10-34}$$

|||| Proof

If **a** and **b** are orthogonal the theorem is true since the projection in that case is the zero vector. Conversely, let $\text{sign}(\mathbf{a}\cdot\mathbf{b})$ denote the sign of $\mathbf{a}\cdot\mathbf{b}$. We have that $\text{sign}(\mathbf{a}\cdot\mathbf{b})$ is positive exactly when **a** and $\text{proj}(\mathbf{b},\mathbf{a})$ have the same direction and negative exactly when they have the opposite direction. Therefore we get

$$\text{proj}(\mathbf{b},\mathbf{a}) = \text{sign}(\mathbf{a}\cdot\mathbf{b})\cdot|\text{proj}(\mathbf{b},\mathbf{a})|\,\frac{\mathbf{a}}{|a|} = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|^2}\,\mathbf{a}\,,$$

where we have used Theorem 10.53, and the fact that $\frac{\mathbf{a}}{|\mathbf{a}|}$ is a unit vector pointing in the direction of **a**.

■

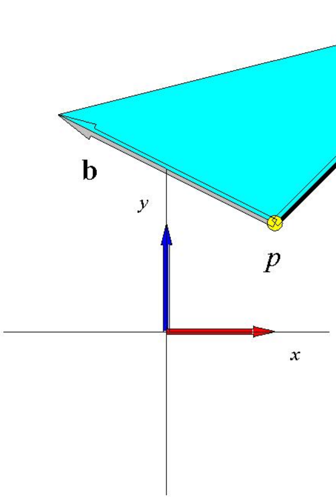


Figure 10.8: A triangle spanned by two vectors in the plane

10.8.2 Geometric Interpretation of the Determinant of a 2×2 Matrix

A triangle $\triangle = \triangle(p, \mathbf{a}, \mathbf{b})$ is determined by two vectors drawn from a common initial point, see the triangle $\triangle = \triangle(p, \mathbf{a}, \mathbf{b})$ in Figure 10.8.

The area of a triangle is known to be half the base times its height. We can choose the length $|\mathbf{a}|$ of $\mathbf{a}$ as the base. And the height in the triangle is

$$|\mathbf{b}|\sin(\theta) = \frac{|\mathbf{b}\cdot\widehat{\mathbf{a}}|}{|\widehat{\mathbf{a}}|} , \tag{10-35}$$

where θ is the angle between the two vectors $\mathbf{a}$ and $\mathbf{b}$, and where $\widehat{\mathbf{a}}$ denotes the ***hat vector*** in the plane to $\mathbf{a}$, that is in coordinates we have $\widehat{\mathbf{a}} = (-a_2, a_1)$. Hence the area is:

$$\begin{aligned}\text{Area}(\triangle(p, \mathbf{a}, \mathbf{b})) &= \frac{1}{2}|\mathbf{b}\cdot\widehat{\mathbf{a}}| \\ &= \frac{1}{2}|a_1b_2 - a_2b_1| \\ &= |\,\frac{1}{2}\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}\,| \\ &= \frac{1}{2}|\ \det\left(\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}\right)\ | \\ &= \frac{1}{2}|\ \det\left(\,[\mathbf{a}\ \mathbf{b}]\,\right)\ | \ .\end{aligned} \tag{10-36}$$

Thus we have proven the theorem:

|||| Theorem 10.55 Area of a Triangle as a Determinant

The area of the triangle $\triangle(p, \mathbf{a}, \mathbf{b})$ is the absolute value of half the determinant of the 2×2 matrix that is obtained by insertion of $\mathbf{a}$ and $\mathbf{b}$ as columns in the matrix.

10.8.3 The Cross Product and the Scalar Triple Product

The *cross product* of two vectors and the *scalar triple product* of three vectors are introduced using determinants:

|||| Definition 10.56 Cross Product

In 3-space two vectors are given ${}_e\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and ${}_e\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$.

By the *cross product* (or the *vector product*) $\mathbf{a} \times \mathbf{b}$ of $\mathbf{a}$ and $\mathbf{b}$ is understood the vector $\mathbf{v}$ given by

$$ {}_e\mathbf{v} = \begin{bmatrix} \det \begin{bmatrix} a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \\ \det \begin{bmatrix} a_3 & b_3 \\ a_1 & b_1 \end{bmatrix} \\ \det \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \end{bmatrix} \tag{10-37} $$

The cross product has a geometric significance. Consider Figure 10.9 and the following theorem:

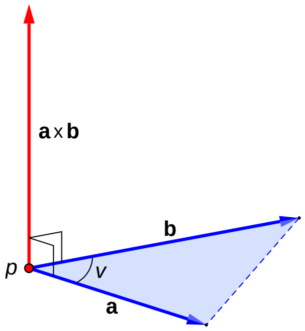


Figure 10.9: Geometry of the cross-product

|||| Theorem 10.57 The Area of a Triangle by the Cross Product

For two linearly independent vectors $\mathbf{a}$ and $\mathbf{b}$ that form the angle v with each other, the cross product $\mathbf{a} \times \mathbf{b}$ satisfies

1. $\mathbf{a} \times \mathbf{b}$ is orthogonal to both $\mathbf{a}$ and $\mathbf{b}$.
2. $|\mathbf{a} \times \mathbf{b}| = 2 \cdot \text{Area}(\triangle(p, \mathbf{a}, \mathbf{b}))$.
3. The vector set $(\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b})$ follows the *right hand rule*: seen from the tip of $\mathbf{a} \times \mathbf{b}$ the direction from $\mathbf{a}$ to $\mathbf{b}$ is counter-clockwise.

|||| Definition 10.58 Scalar Triple Product

The *scalar triple product* $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ of the vectors ${}_{\mathrm{e}}\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$, ${}_{\mathrm{e}}\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ and ${}_{\mathrm{e}}\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ a_3 \end{bmatrix}$

is defined by:

$$
\begin{aligned}
[\mathbf{a}, \mathbf{b}, \mathbf{c}] &= (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} \\
&= (c_1(a_2 b_3 - a_3 b_2) + c_2(a_3 b_1 - a_1 b_3) + c_3(a_1 b_2 - a_2 b_1)) \\
&= \det\left(\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \right) \\
&= \det\left([{}_{\mathrm{e}}\mathbf{a} \; {}_{\mathrm{e}}\mathbf{b} \; {}_{\mathrm{e}}\mathbf{c}] \right) \quad .
\end{aligned}
\tag{10-38}
$$

10.8.4 Geometric Interpretation of the Determinant of a 3×3 Matrix

From elementary Euclidean space geometry we know that the volume of a tetrahedron is one third of the area of the base times the height. Consider the tetrahedron $\boxtimes = \boxtimes(p, \mathbf{a}, \mathbf{b}, \mathbf{c})$ spanned by the vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ drawn from the point p, in Figure 10.10. The area of the base, $\triangle(p, \mathbf{a}, \mathbf{b})$ has been determined in the second part of Theorem 10.57.

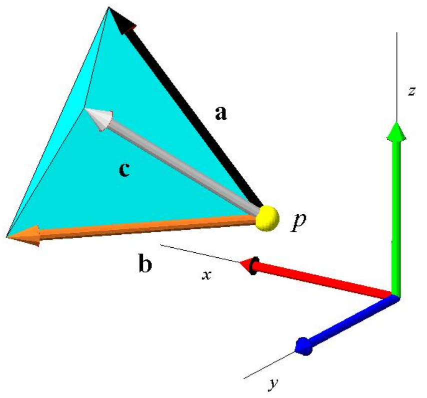


Figure 10.10: A tetrahedron spanned by three vectors in 3-space

The height can then be determined as the scalar product of the third edge vector $\mathbf{c}$ with a unit vector, perpendicular to the base triangle.

But $\mathbf{a} \times \mathbf{b}$ is exactly perpendicular to the base triangle (because the cross product is perpendicular to the edge vectors of the base triangle, see part 2 of Theorem (10.57), so we use this:

$$\begin{aligned}\mathrm{Vol}(\boxtimes(p,\mathbf{a},\mathbf{b},\mathbf{c})) &= |\frac{1}{3}\,\mathrm{Area}(\triangle(p,\mathbf{a},\mathbf{b}))\,\frac{(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}}{|\mathbf{a}\times\mathbf{b}|}| \\ &= \frac{1}{6}|(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}|\end{aligned} \tag{10-39}$$

where we have used part 2 of Theorem 10.57.

By comparing this to the definition of *scalar triple product*, see 10.58, we now get the volume of a tetrahedron written in 'determinant-form':

|||| Theorem 10.59 Volume of a Tetrahedron as a Scalar Triple Product

The volume of the tetrahedron $\boxtimes = \boxtimes(p,\mathbf{a},\mathbf{b},\mathbf{c})$ is:

$$\mathrm{Vol}(\boxtimes(p,\mathbf{a},\mathbf{b},\mathbf{c})) = \frac{1}{6}|\det\left([\mathbf{a}\,\mathbf{b}\,\mathbf{c}]\right)| \quad . \tag{10-40}$$

A tetrahedron has the volume 0, is collapsed, exactly when the determinant in (10-40) is 0, and this occurs exactly when one of the vectors can be written as a linear combination of the two others (why is that?).

|||| Definition 10.60 Regular Tetrahedron

A ***regular tetrahedron*** is a tetrahedron that has a proper volume, that is a volume, that is strictly greater than 0.

|||| Exercise 10.61

Let $\mathbf{A}$ denote a (2×2)-matrix with the column vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{A} = [\,\mathbf{a}\;\mathbf{b}\,] \quad . \tag{10-41}$$

Show that the determinant of $\mathbf{A}$ is 0 if and only if the column vectors $\mathbf{a}$ and $\mathbf{b}$ are linearly dependent in $\mathbb{R}^2$.

|||| Exercise 10.62

Let **A** denote a $(3 \times 3)-$matrix with the column vectors **a**, **b**, and **c**:

$$\mathbf{A} = \left[\, \mathbf{a}\ \mathbf{b}\ \mathbf{c} \,\right] \quad . \tag{10-42}$$

Show, that the determinant of **A** is 0 if and only if the column vectors **a**, **b** and **c** constitute a linearly dependent set of vectors in $\mathbb{R}^3$.

|||| Exercise 10.63

Use the geometric interpretations of the determinant above to show the following Hadamard's inequality for $(2 \times 2)-$matrices and for $(3 \times 3)-$matrices (in fact the inequality is true for all square matrices):

$$(\det(\mathbf{A}))^2 \leq \prod_{j}^{n} \left(\sum_{i}^{n} a_{ij}^2 \right) \quad . \tag{10-43}$$

When is the equality sign valid in (10-43)?

eNote 11

General Vector Spaces

In this eNote a general theory is presented for all mathematical sets where addition and multiplication by a scalar are defined and which satisfy the same arithmetic rules as geometric vectors in the plane and in 3-space. Using the concepts of a basis and coordinates, it is shown how one can simplify and standardize the solution of problems that are common to all these sets, which are called vector spaces. Knowledge of eNote 10 about geometric vectors is an advantage as is knowledge about the solution sets for systems of linear equations, see eNote 6. Finally, elementary matrix algebra and a couple of important results about determinants are required (see eNotes 7 and 8).

Updated: 4.10.21 David Brander

11.1 Generalization of the Concept of a Vector

The vector concept originates in the geometry of the plane and space where it denotes a pair consisting of a length and a direction. Vectors can be represented by a line segment with orientation (an arrow) following which it is possible to define two geometric operations: *addition* of vectors and *multiplication* of vectors *by numbers* (scalar). For the use in more complicated arithmetic operations one proves eight arithmetic rules concerning the two arithmetic operations.

In many other sets of mathematical objects one has a need for defining addition of the objects and multiplication of an object by a scalar. The number spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ and the set of matrices $\mathbb{R}^{m\times n}$ are good examples, see eNote 5 and eNote 6, respectively. The remarkable thing is, that the *arithmetic rules* for addition and multiplication by a scalar, that are possible to prove within each of these sets, are the same as the arithmetic rules for geometric vectors in the plane and in space! Therefore one says: Let us make a *theory* that applies to all the sets where addition and multiplication by a scalar can be defined and where all the eight arithmetic rules known from geometry are valid. By this one carries out a *generalization* of the concept of geometric vectors, and every set that obeys the conditions of the theory is therefore called a *vector space*.

In eNote 10 about geometric vectors it is demonstrated how one can introduce a *basis* for the vectors following which the vectors are determined by their *coordinates* with respect to this basis. The advantage of this is that one can replace the geometric vector calculation by calculation with the coordinates for the vector. It turns out that it is also possible to transfer the concepts of basis and coordinates to many other sets of mathematical objects that have addition and multiplication by a scalar.

In the following, when we investigate vector spaces in the abstract sense, it means that we look at which concepts, theorems and methods follow from the common arithmetic rules, as we ignore the concrete meaning of addition and multiplication by a scalar has in the sets of concrete objects where they are introduced. By this one obtains general methods for *every* set of the kind described above. The application in any particular vector space then calls for *interpretation* in the context of the results obtained. The approach is called *the axiomatic method*. Concerning all this we now give the abstract definition of vector spaces.

|||| Definition 11.1 Vector Spaces

Let $\mathbb{L}$ denote $\mathbb{R}$ or $\mathbb{C}$, and let V be a set of mathematical elements where there is defined two arithmetic operations:

I. *Addition* that from two elements $\mathbf{a}$ and $\mathbf{b}$ in V forms the sum $\mathbf{a}+\mathbf{b}$ that also belongs to V.

II. *Multiplication by a scalar* that from any $\mathbf{a} \in V$ and any scalar $k \in \mathbb{L}$ forms a product $k\mathbf{a}$ or $\mathbf{a}k$ that also belongs to V.

V is called a ***vector space*** and the elements of V ***vectors*** if the following eight arithmetic rules are valid:

1.	$\mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a}$	Addition is commutative
2.	$(\mathbf{a}+\mathbf{b})+\mathbf{c}=\mathbf{a}+(\mathbf{b}+\mathbf{c})$	Addition is associative
3.	$\mathbf{a}+\mathbf{0}=\mathbf{a}$	In V there exists $\mathbf{0}$ that is *neutral* wrt. addition
4.	$\mathbf{a}+(-\mathbf{a})=\mathbf{0}$	For every $\mathbf{a} \in V$ there is an *opposite object* $-\mathbf{a} \in V$
5.	$k_1(k_2\mathbf{a})=(k_1k_2)\mathbf{a}$	Product by a scalar is associative
6.	$(k_1+k_2)\mathbf{a}=k_1\mathbf{a}+k_2\mathbf{a}$	} The distributive rule applies
7.	$k_1(\mathbf{a}+\mathbf{b})=k_1\mathbf{a}+k_1\mathbf{b}$	
8.	$1\mathbf{a}=\mathbf{a}$	The scalar 1 is *neutral* in products with vectors

If $\mathbb{L}$ in the definition 11.1 stands for $\mathbb{R}$ then V is a ***vector space over the real numbers***. This means that the scalar k (only) can be an arbitrary real number. Similarly one talks about V as a ***vector space over the complex numbers*** if $\mathbb{L}$ stands for $\mathbb{C}$, where k can be an arbitrary complex number.

The requirements I and II in the definition 11.1, that the results of addition and of multiplication by a scalar in itself must be an element in V, are called the ***stability requirements***. In other words V must be stable with respect to the two arithmetic operations.

The set of geometric vectors in the plane and the set of geometric vectors in space are naturally the most obvious examples of vector spaces, since the eight arithmetic rules in the definition 11.1 are constructed from the corresponding rules for geometric vectors. But let us check the *stability requirements*. Is the sum of two vectors in the plane itself a vector in the plane? And is a vector in the plane multiplied by a number in itself a vector in the plane? From the definition of the two arithmetic operations (see Definition 10.2 and Definition 10.3), the answer is obviously yes, therefor the set of vectors in the plane is a vector space. Similarly we see that the set of vectors in 3-space is a vector space.

|||| Theorem 11.2 Uniqueness of the 0-Vector and the Opposite Vector

For every vector space V:

1. V only contains one neutral element with respect to addition.
2. Every vector $\mathbf{a} \in V$ has only one opposite element.

|||| Proof

First part:
Let $\mathbf{0}_1$ and $\mathbf{0}_2$ be two elements in V both neutral with respect to addition. Then:

$$\mathbf{0}_1 = \mathbf{0}_1 + \mathbf{0}_2 = \mathbf{0}_2 + \mathbf{0}_1 = \mathbf{0}_2\,,$$

where we have used the fact that addition is commutative. There is only one 0-vector: $\mathbf{0}$.

Second part:
Let $\mathbf{a}_1, \mathbf{a}_2 \in V$ be two opposite elements for $\mathbf{a} \in V$. Then:

$$\mathbf{a}_1 = \mathbf{a}_1 + \mathbf{0} = \mathbf{a}_1 + (\mathbf{a} + \mathbf{a}_2) = (\mathbf{a} + \mathbf{a}_1) + \mathbf{a}_2 = \mathbf{0} + \mathbf{a}_2 = \mathbf{a}_2\,,$$

where we have used the fact that addition is both commutative and associative. Hence there is for $\mathbf{a}$ only one opposite vector $-\mathbf{a}$.

■

|||| Definition 11.3 Subtraction

Let V be a vector space, and let $\mathbf{a}, \mathbf{b} \in V$. By the difference $\mathbf{a} - \mathbf{b}$ we understand the vector

$$\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})\,. \tag{11-1}$$

|||| Exercise 11.4

Prove that $(-1)\mathbf{a} = -\mathbf{a}$.

|||| Exercise 11.5 Zero-Rule

Prove that the following variant of the zero-rule applies to any vector space:

$$k\mathbf{a} = \mathbf{0} \Leftrightarrow k = 0 \;\text{ or }\; \mathbf{a} = \mathbf{0}\,. \tag{11-2}$$

|||| Example 11.6 Matrices as Vectors

For two arbitrary natural numbers m and n, $\mathbb{R}^{m\times n}$ (that is, the set of real $m \times n$-matrices) is a vector space. Similarly $\mathbb{C}^{m\times n}$ (that is, the set of complex $m \times n$-matrices) is a vector space

Consider e.g. $\mathbb{R}^{2\times 3}$. If we add two matrices of this type we get a new matrix of the same type, and if we multiply a 2×3-matrix by a number, we also get a new 2×3-matrix (see Definition 7.1). Thus the stability requirements are satisfied. That $\mathbb{R}^{2\times 3}$ in addition satisfies the eight arithmetic rules, is apparent from Theorem 7.3.

|||| Exercise 11.7

Explain that for every natural number n the number space $\mathbb{L}^n$ is a vector space. Remember to think about the case $n = 1$!

In the following two examples we shall see that the geometrically inspired vector space theory, surprisingly, can be applied to well known sets of *functions*. Mathematic historians have in this connection talked about *the geometrization of mathematical analysis*!

|||| Example 11.8 Polynomials as Vectors

The set of polynomials $P : \mathbb{R} \mapsto \mathbb{R}$ of at most n'th degree is denoted $P_n(\mathbb{R})$. An element P in $P_n(\mathbb{R})$ is given by

$$P(x) = a_0 + a_1 x + \cdots + a_n x^n \tag{11-3}$$

where the coefficients $a_0, a_1, \cdots a_n$ are arbitrary real numbers. Addition of two polynomials in $P_n(\mathbb{R})$ is defined by pairwise addition of coefficients belonging to the same degree of the variable, and multiplication of a polynomial in $P_n(\mathbb{R})$ by a number k is defined as the multiplication of every coefficient with k. As an example of the two arithmetic operations we look at two polynomials from $P_3(\mathbb{R})$:

$$P(x) = 1 - 2x + x^3 = 1 - 2x + 0x^2 + 1x^3$$

and

$$Q(x) = 2 + 2x - 4x^2 = 2 + 2x - 4x^2 + 0x^3.$$

By the sum of P and Q we understand the polynomial $R = P + Q$ given by

$$R(x) = (1+2) + (-2+2)x + (0-4)x^2 + (1+0)x^3 = 3 - 4x^2 + x^3$$

and by the multiplication of P by the scalar $k = 3$ we understand the polynomial $S = 3P$ given by

$$S(x) = (3 \cdot 1) + (3 \cdot (-2))x + (3 \cdot 0)x^2 + (3 \cdot 1)x^3 = 3 - 6x + 3x^3.$$

We will now justify that $P_n(\mathbb{R})$ with the introduced arithmetic operations is a vector space! That $P_n(\mathbb{R})$ satisfies the *stability requirements* follows from the fact that the sum of two polynomials of at most n'th degree in itself is a polynomial of at most n'th degree, and that multiplication of a polynomial of at most n'th degree by a real number again gives a polynomial of at most n'th degree. The conditions 1, 2 and 5 - 8 in the definition 11.1 are satisfied, because the same rules of operation apply to the calculations with coefficients of the polynomials that are used in the definition of the operations. Finally the conditions 3 and 4 are satisfied since the polynomial

$$P(x) = 0 + 0x + \cdots 0x^n = 0$$

constitutes the *zero vector*, and *the opposite vector* to $P(x) \in P_n(\mathbb{R})$ is given by the polynomial

$$-P(x) = -a_0 - a_1 x - \cdots - a_n x^n.$$

In the same way we show that polynomial $P : \mathbb{C} \mapsto \mathbb{C}$ of at most n'th degree, which we denote by $P_n(\mathbb{C})$, is a vector space.

|||| Exercise 11.9

Explain that $P(\mathbb{R})$, that is the set of real polynomials, is a vector space.

|||| Example 11.10 Continuous Functions as Vectors

The set of continuous real functions on a given interval $I \subseteq \mathbb{R}$ is denoted $C^0(I)$. Addition $m = f + g$ of two functions f and g in $C^0(I)$ is defined by

$$m(x) = (f+g)(x) = f(x) + g(x) \text{ for every } x \in I$$

and multiplication $n = k \cdot f$ of the function f by a real number k by

$$n(x) = (k \cdot f)(x) = k \cdot f(x) \text{ for every } x \in I\,.$$

We will now justify that $C^0(I)$, with the introduced operations of calculations, is a vector space. Since $f + g$ and $k \cdot f$ are continuous functions, we see that $C^0(I)$ satisfies the two stability requirements. Moreover: there exists a function that acts as the zero vector, viz. the zero function, that is, the function that has the value 0 for all $x \in I$, and the opposite vector to $f \in C^0(I)$ is the vector $(-1)f$ that is also written $-f$, and which for all $x \in I$ has the value $-f(x)$. Now it is obvious that $C^0(I)$ with the introduced operations of calculation satisfies all eight rules in definition 11.1, and $C^0(I)$ is thus a vector space.

11.2 Linear Combinations and Span

A consequence of arithmetic rules such as $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$ and $(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$ from the definition 11.1 is that one can omit parentheses when one adds a series of vectors: the order of vector addition has no influence on the resulting sum vector. This is the background for *linear combinations* where a set of vectors is multiplied by a scalar and thereafter written as a sum.

|||| Definition 11.11 Linear Combination

When in a vector space V p vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ are given, and arbitrary scalars $k_1, k_2, \ldots, k_p$ are chosen, then the sum

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \ldots + k_p\mathbf{v}_p$$

is called a ***linear combination*** of the p given vectors.

If all the $k_1, \cdots, k_p$ are equal to 0, the linear combination is called *improper*, or *trivial*, but if at least one of the scalars is different from 0, it is called *proper* or *non-trivial*.

In the definition 11.11 only one linear combination is mentioned. In many circumstances it is of interest to consider the total set of possible linear combinations of given vectors. The set is called the *span* of the vectors. Consider e.g. a plane in space, through the origin and containing the position vectors for two non-parallel vectors $\mathbf{u}$ and $\mathbf{v}$. The plane can be considered the span of the two vectors since the position vectors

$$\overrightarrow{OP}= k_1\mathbf{u}+k_2\mathbf{v}$$

"run through" all points P in the plane when k_1 and k_2 take on all conceivable real values, see Figure 11.1.

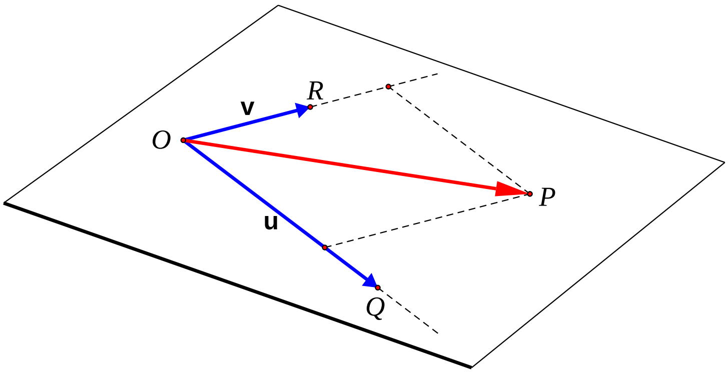


Figure 11.1: $\mathbf{u}$ and $\mathbf{v}$ span a plane in space

|||| Definition 11.12 Span

By the ***span*** of a given set of vectors $\mathbf{v}_1,\ \mathbf{v}_2,\dots,\mathbf{v}_p$ in a vector space V we understand the total set of all possible linear combinations of the vectors. The span of the p vectors is denoted by

$$\text{span}\{\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_p\}\,.$$

|||| Example 11.13 Linear Combination and Span

We consider in the vector space $\mathbb{R}^{2\times 3}$ the three matrices/vectors

$$\mathbf{A}=\begin{bmatrix}1&0&3\\0&2&2\end{bmatrix},\ \mathbf{B}=\begin{bmatrix}2&1&0\\0&3&1\end{bmatrix}\ \text{and}\ \mathbf{C}=\begin{bmatrix}-1&-2&9\\0&0&4\end{bmatrix}. \tag{11-4}$$

An example of a linear combination of the three vectors is

$$2\mathbf{A} + 0\,\mathbf{B} + (-1)\mathbf{C} = \begin{bmatrix} 3 & 2 & -3 \\ 0 & 4 & 0 \end{bmatrix}. \tag{11-5}$$

We can then write

$$\begin{bmatrix} 3 & 2 & -3 \\ 0 & 4 & 0 \end{bmatrix} \in \operatorname{span}\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}. \tag{11-6}$$

11.3 Linear Dependence and Linear Independence

Two geometric vectors $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent if they are parallel, that is if there exists a number k, such that $\mathbf{v} = k\mathbf{u}$. More generally an arbitrary set of vectors are linearly dependent if one of the vectors is a linear combination of the others. We wish to transfer this concept to vector space theory:

|||| Definition 11.14 Linear Dependence and Independence

A set consisting of p vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$ in a vector space V is ***linearly dependent*** if at least one of the vectors can be written as a linear combination of the others: for example

$$\mathbf{v}_1 = k_2\mathbf{v}_2 + k_3\mathbf{v}_3 + \cdots + k_p\mathbf{v}_p\,.$$

If none of the vectors can be written as a linear combination of the others, the set is called ***linearly independent***.

NB: If the set of vectors only consists of a single vector, the set is called linearly dependent if it consists of the 0-vector, and otherwise linearly independent.

|||| Example 11.15 Linear Dependence

Any set of vectors containing the zero vector, is linearly dependent! Consider e.g. the set $\{\mathbf{u}, \mathbf{v}, \mathbf{0}, \mathbf{w}\}$), here the zero vector can trivially be written as a linear combination of the three other vectors in the set:

$$\mathbf{0} = 0\mathbf{u} + 0\mathbf{v} + 0\mathbf{w}\,,$$

where the zero vector is written as a linear combination of the other vectors in the set.

|||| Example 11.16 Linear Dependence

Consider in the vector space $\mathbb{R}^{2\times 3}$ the three matrices/vectors

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & 2 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} -1 & -2 & 9 \\ 0 & 0 & 4 \end{bmatrix}. \tag{11-7}$$

$\mathbf{C}$ can be written as a linear combination of $\mathbf{A}$ and $\mathbf{B}$ since

$$\mathbf{C} = 3\mathbf{A} - 2\mathbf{B}\,.$$

Therefore $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are linearly dependent.

In contrast the set consisting of $\mathbf{A}$ and $\mathbf{B}$ is linearly independent, because these two vectors are not ”parallel’, since a number k obviously does not exist such that $\mathbf{B} = k\mathbf{A}$. Similarly with the sets $\{\mathbf{A},\mathbf{C}\}$ and $\{\mathbf{B},\mathbf{C}\}$.

When you investigate whether a set of vectors is linearly dependent, use of the definition 11.14 provokes the question *which* of the vectors is a linear combination of the others. Where should we begin the investigation? The dilemma can be avoided if bypassing the definition we instead use the following theorem:

|||| Theorem 11.17 Linear Dependence and Independence

A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$ in a vector space V is linearly dependent if and only if the zero vector can be written as proper linear combination of the vectors – that is, if and only if scalars $k_1, k_2, \ldots, k_p$ exist that are not all equal to 0, and that satisfy

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_p\mathbf{v}_p = \mathbf{0}\,. \tag{11-8}$$

. Otherwise the vectors are linearly independent.

.

|||| Proof

Assume first that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$ are linearly dependent, then one can be written as a linear combination of the others, e.g.

$$\mathbf{v}_1 = k_2\mathbf{v}_2 + k_3\mathbf{v}_3 + \cdots + k_p\mathbf{v}_p\,. \tag{11-9}$$

But this is equivalent to

$$\mathbf{v}_1 - k_2\mathbf{v}_2 - k_3\mathbf{v}_3 - \cdots - k_p\mathbf{v}_p = \mathbf{0}\,, \tag{11-10}$$

whereby the zero-vector is written as a linear combination of the vector set in which at least one of the coefficients are not 0, since $\mathbf{v}_1$ has the coefficient 1.

Conversely, assume that the zero-vector is written as a proper linear combination of the set of vectors, where one of the coefficients, for example the $\mathbf{v}_1$ coefficient k_1, is different from 0 (the same argument works for any of other coefficient). Then we have

$$k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_p\mathbf{v}_p = \mathbf{0} \;\Leftrightarrow \mathbf{v}_1 = (-1)\frac{k_2}{k_1}\mathbf{v}_2 + \cdots + (-1)\frac{k_p}{k_1}\mathbf{v}_p\,. \tag{11-11}$$

Thus $\mathbf{v}_1$ is written as a linear combination of the other vectors and the proof is complete.

■

|||| Example 11.18 Linear Dependence

In the number space $\mathbb{R}^4$ the vectors $\mathbf{a} = (1,3,0,2)$, $\mathbf{b} = (-1,9,0,4)$ and $\mathbf{c} = (2,0,0,1)$ are given. Since

$$3\mathbf{a} - \mathbf{b} - 2\mathbf{c} = \mathbf{0}$$

the zero vector is written as a non-trivial linear combination of the three vectors. Thus they are linearly dependent.

11.4 Basis and Dimension of a Vector Space

A compelling argument for the introduction of a basis in a vector space is that all vectors in the vector space then can be written using coordinates. In a later section it is shown how problems of calculation can be simplified and standardized with vectors when we use coordinates. But in this section we will discuss the requirements that a basis should satisfy and investigate the consequences of these requirements.

A basis for a vector space consists of certain number of vectors, usually written in a definite order. A decisive task for the basis vectors is that they should span the vector space, but more precisely we want this task to be done with *as few vectors as possible*. In this case it turns out that all vectors in the vector space can be written *uniquely* as a linear combination of the basis vectors. And it is exactly the *coefficients* in the unique linear combination we will use as coordinates.

Let us start out from some characteristic properties about bases for geometric vectors in the plane.

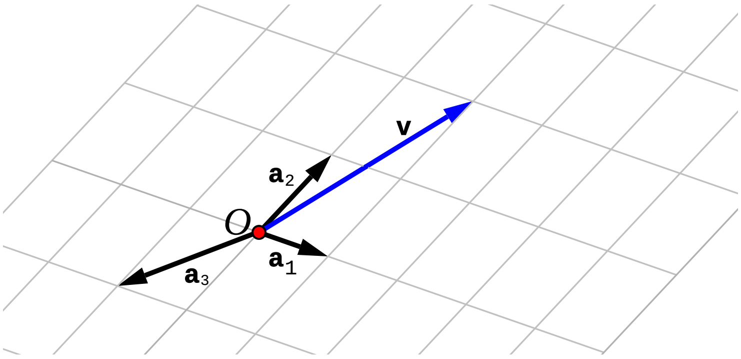


Figure 11.2: Coordinate system in the plane with the basis $(\mathbf{a}_1, \mathbf{a}_2)$

Consider the vector set $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$ in Figure 11.2. There is no doubt that any other vector in the plane can be written as a linear combination of the three vectors. But the linear combination is not unique, for example the vector $\mathbf{v}$ can be written in these two ways:

$$\mathbf{v} = 2\mathbf{a}_1 + 3\mathbf{a}_2 - 1\mathbf{a}_3$$
$$\mathbf{v} = 1\mathbf{a}_1 + 2\mathbf{a}_2 + 0\mathbf{a}_3 \,.$$

The problem is that the a-vectors are not linearly independent, for example $\mathbf{a}_3 = -\mathbf{a}_1 - \mathbf{a}_2$. But if we remove one of the vectors, e.g. $\mathbf{a}_3$, the set is linearly independent, and there is only one way of writing the linear combination

$$\mathbf{v} = 1\mathbf{a}_1 + 2\mathbf{a}_2 \,.$$

We can summarize the characteristic properties of a basis for the geometric vectors in the plane thus:

1. any basis must consist of linearly independent vectors,
2. any basis must contain exactly two vectors (if there are more than two, they are linearly dependent, if there are less than two they do not span the plane), and
3. *every* set consisting of two linear independent vectors is a basis.

These properties can be transferred to other vector spaces. We embark on this now, and we start by the general definition of a basis.

|||| Definition 11.19 Basis

By a ***basis*** for a vector space V we understand a set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ of vectors from V that satisfy:

1. $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ spans V .
2. $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is linearly independent.

When we discuss coordinates later, it will be necessary to consider the basis elements to have a define order, and so we will write them as an *ordered set*, denoted by using parentheses: $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$.

Here we should stop and check that the definition 11.19 does in fact satisfy our requirements of uniqueness of a basis. This is established in the following theorem.

|||| Theorem 11.20 Uniqueness Theorem

If a basis for a vector space V is given, any vector in V can then be written as a *unique* linear combination of the basis vectors.

|||| Proof

We give the idea in the proof by looking at a vector space V that has a basis consisting of three basis vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ and assume that $\mathbf{v}$ is an arbitrary vector in V that in two ways can be written as a linear combination of the basis vectors. We can then write two equations

$$\begin{aligned} \mathbf{v} &= k_1\mathbf{a} + k_2\mathbf{b} + k_3\mathbf{c} \\ \mathbf{v} &= k_4\mathbf{a} + k_5\mathbf{b} + k_6\mathbf{c} \end{aligned} \tag{11-12}$$

By subtracting the lower equation from the upper equation in (11-12) we get the equation

$$\mathbf{0} = (k_1 - k_4)\mathbf{a} + (k_2 - k_5)\mathbf{b} + (k_3 - k_6)\mathbf{c}\,. \tag{11-13}$$

Since $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are linearly independent, the zero vector can only be written as an improper linear combination of these, therefore all coefficients in (11-12) are equal to 0, yielding $k_1 = k_4$, $k_2 = k_5$ and $k_3 = k_6$. But then the two ways $\mathbf{v}$ has been written as linear combinations of the basis vectors, is in reality the same, there is only one way!

This reasoning is immediately extendable to a basis consisting of an arbitrary number of basis vectors.

■

We now return to the fact that every basis for geometric vectors in the plane contains *two* linearly independent basis vectors, and that similarly for geometric vectors in space the basis must consist of *three* linearly independent basis vectors. It turns out that the fixed number of basis vectors is a property of all vector spaces with a basis, and this makes it possible to talk about the *dimension* of a vector space that has a basis. To prove the property we need a lemma.

|||| Lemma 11.21

If a vector space V has a basis consisting of n basis vectors then every set from V that contains more than n vectors will be linearly dependent.

|||| Proof

To get a grasp of the proof's underlying idea, consider a vector space V that has a basis consisting of two vectors $(\mathbf{a},\mathbf{b})$, and investigate three arbitrary vectors $\mathbf{c},\mathbf{d}$ and $\mathbf{e}$ from V. We prove that the three vectors necessarily must be linearly independent.

Since $(\mathbf{a},\mathbf{b})$ is a basis for V, we can write three equations

$$\begin{aligned}\mathbf{c} &= c_1\mathbf{a}+c_2\mathbf{b}\\ \mathbf{d} &= d_1\mathbf{a}+d_2\mathbf{b}\\ \mathbf{e} &= e_1\mathbf{a}+e_2\mathbf{b}\end{aligned} \tag{11-14}$$

Consider further the zero vector written as the following linear combination

$$x_1\mathbf{c}+x_2\mathbf{d}+x_3\mathbf{e}=\mathbf{0}\,, \tag{11-15}$$

which by substitution of the equations (11-14) into (11-15) is equivalent to

$$(x_1c_1+x_2d_1+x_3e_1)\mathbf{a}+(x_1c_2+x_2d_2+x_3e_2)\mathbf{b}=\mathbf{0}\,. \tag{11-16}$$

Since the zero vector only can be obtained as a linear combination of $\mathbf{a}$ and $\mathbf{b}$, if every coefficient is equal to 0, (11-16) is equivalent to the following system of equations

$$\begin{aligned}c_1x_1+d_1x_2+e_1x_3&=0\\ c_2x_1+d_2x_2+e_2x_3&=0\end{aligned} \tag{11-17}$$

This is a homogeneous system of linear equations in which the number of equations is less than the number of unknowns. Therefore the system of equations has infinitely many solutions, which means that (11-16) not only is obtainable with $x_1 = 0, x_2 = 0$ and $x_3 = 0$. Thus we have shown that the ordered set $(\mathbf{c}, \mathbf{d}, \mathbf{e})$ is linearly dependent.

In general: Assume that the basis V consists of n vectors, and that m vectors from V where $m > n$ are given. By following the same procedure as above a homogeneous system of linear equations emerges with n equations in m unknowns that, because $m > n$, similarly has infinitely many solutions. By this it is shown that the m vectors are linearly dependent.

■

Then we are ready to give the following important theorem:

|||| Theorem 11.22 The Number of Basis Vectors

If a vector space V has a basis consisting of n basis vectors, then every basis for V will consist of n basis vectors.

|||| Proof

Assume that V has two bases with different numbers of big(asis vectors. We denote the basis with the least number of basis vectors a and the one with largest number b. According to Lemma 11.21 the b-basis vectors must be linearly dependent, and this contradicts that they form a basis. The assumption that V can have two bases with different numbers of basis vectors, must therefore be untrue.

■

That the number of basis vectors according to theorem 11.22 is a *property* of vector spaces with a basis, motivates the introduction of the concept of dimension:

|||| Definition 11.23 Dimension

By the dimension of a vector space V that has a basis b, we understand the number of vectors in b. If this number is n, one says that V is n-dimesional and write

$$\dim(V) = n\,. \tag{11-18}$$

Remark: There are vector spaces that do not have a finite basis, see Section 11.7.2 below.

|||| Example 11.24 Dimension of Geometric Vector Spaces

Luckily the definition 11.23 confirms an intuitive feeling that the set of geometric vectors in the plane has the dimension two and that the set of geometric vectors in space has the dimension three!

|||| Example 11.25 The Standard Basis for Number Spaces

An arbitrary vector $\mathbf{v} = (a, b, c, d)$ in $\mathbb{R}^4$ or in $\mathbb{C}^4$ (that is in $\mathbb{L}^4$) can in an obvious way be written as a linear combination of four particular vectors in $\mathbb{L}^4$

$$\mathbf{v} = a\,(1, 0, 0, 0) + b\,(0, 1, 0, 0) + c\,(0, 0, 1, 0) + d\,(0, 0, 0, 1)\,. \tag{11-19}$$

We put $\mathbf{e}_1 = (1, 0, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0, 0)$, $\mathbf{e}_3 = (0, 0, 1, 0)$ and $\mathbf{e}_4 = (0, 0, 0, 1)$ and conclude using (11-19) that the ordered set $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ spans $\mathbb{L}^4$.

Since we can see that none of the vectors can be written as a linear combination of the others, the set is linearly independent, and $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ is thereby a basis for $\mathbb{L}^4$. This particular basis is called ***standard basis*** for $\mathbb{L}^4$. Since the number of basis vectors in the standard e-basis is four, $\dim(\mathbb{L}^4) = 4$.

This can immediately be generalized to $\mathbb{L}^n$: For every n the set $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ where

$$\mathbf{e}_1 = (1, 0, 0, \ldots, 0),\ \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \mathbf{e}_n = (0, 0, 0, \ldots, 1)$$

is a basis for $\mathbb{L}^n$. This is called ***standard basis*** for $\mathbb{L}^n$. It is noticed that $\dim(\mathbb{L}^n) = n$.

|||| Example 11.26 Standard Basis for Matrix Spaces

By ***standard basis*** for the vector space $\mathbb{R}^{2\times 3}$ or $\mathbb{C}^{2\times 3}$, we understand the matrix set

$$\left(\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix},\begin{bmatrix}0&1&0\\0&0&0\end{bmatrix},\ldots,\begin{bmatrix}0&0&0\\0&0&1\end{bmatrix}\right) \tag{11-20}$$

Similarly we define a *standard basis* for an arbitrary matrix space $\mathbb{R}^{m\times n}$ and for an arbitrary matrix space $\mathbb{C}^{m\times n}$.

|||| Exercise 11.27

Explain that the matrix set, which in Example 11.26 is referred to as the standard basis for $\mathbb{R}^{2\times 3}$, is in fact a *basis* for this vector space.

|||| Example 11.28 The Monomial Basis for Polynomial Spaces

In the vector space $P_2(\mathbb{R})$ of real polynomials of at most 2nd degree, the ordered set $(1, x, x^2)$ is a basis. This is demonstrated in the following way.

1. Every polynomial $P(x) \in P_2(\mathbb{R})$ can be written in the form

   $$P(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2\,,$$

   that is as a linear combination of the three vectors in the set.

2. The set $\{1, x, x^2\}$ is linearly independent, since the equation

   $$a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2 = 0 \text{ for every } x$$

   according to the ***identity theorem for polynomials*** is only satisfied if all the coefficients a_0, a_1 and a_2 are equal to 0 .

A *monomial* is a polynomial with only one term. Hence, the ordered set $(1, x, x^2)$ is called the ***monomial basis*** for $P_2(\mathbb{R})$, and $\dim(P_2(\mathbb{R})) = 3$.

For every n the ordered set $(1, x, x^2, \ldots, x^n)$ is a basis for $P_n(\mathbb{R})$, and is called the *monomial basis* for $P_n(\mathbb{R})$. Therefore we have that $\dim(P_n(\mathbb{R})) = n + 1$.

Similarly the ordered set $(1, z, z^2, \ldots, z^n)$ is a basis for $P_n(\mathbb{C})$, it is called *monomial basis* for $P_n(\mathbb{C})$. Therefore we have that $\dim(P_n(\mathbb{C})) = n + 1$.

In the set of plane geometric vectors one can choose *any* pair of two linearly independent vectors as basis. Similarly in 3-space *any* set of three linear independent vectors is a basis. We end the section by transferring this to general n-dimensional vector spaces:

|||| Theorem 11.29 Sufficient Conditions for a Basis

In an n-dimensional vector space V, an arbitrary set of n linearly independent vectors from V constitutes a basis for V.

|||| Proof

Since V is assumed to be n-dimensional, it must have a basis b consisting of n basis vectors. Let the a-set $(\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n)$ be an arbitrary linearly independent set of vectors from V. The set is then a basis for V if it spans V. Suppose this is not the case, and let $\mathbf{v}$ be a vector V that does not belong to $\operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n\}$. Then $(\mathbf{v}, \mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n)$ must be linearly independent, but this contradicts theorem 11.21 since there are $n+1$ vectors in the set. Therefore the assumption that the a-set does not span V must be untrue, and it must accordingly be a basis for V.

■

|||| Exercise 11.30

Two geometric vectors $\mathbf{a} = (1, -2, 1)$ and $\mathbf{b} = (2, -2, 0)$ in 3-space are given. Determine a vector $\mathbf{c}$ such that the ordered set $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is a basis for the set of space vectors.

|||| Exercise 11.31

In the 4-dimensional vector space $\mathbb{R}^{2\times 2}$, consider the vectors

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \text{ of } \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{11-21}$$

Explain why the ordered set $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is a linearly independent set, and complement the set with a 2×2 matrix $\mathbf{D}$ such that $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is a basis for $\mathbb{R}^{2\times 2}$.

11.5 Vector Calculations Using Coordinates

Coordinates are closely connected to the concept of a basis. When a basis is chosen for a vector space, any vector in the vector space can be described with the help of its co-ordinates with respect to the chosen basis. By this we get a particularly practical alternative to the calculation operations, addition and multiplication by a scalar, which originally are defined from the 'anatomy' of the specific vector space. Instead of carrying out these particularly defined operations we can implement number calculations with the coordinates that correspond to the chosen basis. In addition it turns out that we can simplify and standardize the solution of typical problems that are common to all vector spaces. But first we give a formal introduction of coordinates with respect to a chosen basis.

|||| Definition 11.32 Coordinates with Respect to a Given Basis

In an n-dimensional vector space V the basis $a = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ and a vector $\mathbf{x}$ are given. We consider the unique linear combination of the basis vectors that according to 11.20 is a way of writing $\mathbf{x}$:

$$\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n\,. \tag{11-22}$$

The coefficients $x_1, x_2, \ldots, x_n$ in (11-22) are denoted $\mathbf{x}$'s *coordinates with respect to the basis* a, or $\mathbf{x}$'s a-coordinates, and they are gathered in a *coordinate vector* as follows:

$$_{\mathrm{a}}\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{11-23}$$

|||| Example 11.33 Coordinates with Respect to a New Basis

In the number space $\mathbb{R}^3$ a basis a is given by $((0,0,1),(1,2,0),(1,-1,1))$. Furthermore the vector $\mathbf{v}=(7,2,6)$ is given. Since

$$2\cdot(0,0,1)+3\cdot(1,2,0)+4\cdot(1,-1,1)=(7,2,6)$$

we see that

$$_{\mathrm{a}}\mathbf{v}=\begin{bmatrix}2\\3\\4\end{bmatrix}.$$

The vector $(7,2,6)$ therefore has the a-coordinates $(2,3,4)$.

In order to be able to manipulate the coordinates of several vectors in various arithmetic operations we will need the following important theorem.

|||| Theorem 11.34 The Coordinate Theorem

In a vector space V two vectors $\mathbf{u}$ and $\mathbf{v}$ plus a real number k are given. In addition an arbitrary basis a is chosen. The two arithmetic operations $\mathbf{u}+\mathbf{v}$ of $k\,\mathbf{u}$ can then be carried out using the a-coordinates like this:

1. $_{\mathrm{a}}(\mathbf{u}+\mathbf{v})={}_{\mathrm{a}}\mathbf{u}+{}_{\mathrm{a}}\mathbf{v}$
2. $_{\mathrm{a}}(k\mathbf{u})=k\,{}_{\mathrm{a}}\mathbf{u}$

In other words: The coordinates for a vector sum are obtained by adding the coordinates for the vectors, and the coordinates for a vector multiplied by a number are the coordinates of the vector multiplied by the number.

|||| Proof

See the proof for the corresponding theorem for geometric vectors in 3-space, Theorem 10.38. The proof for the general case is obtained as a simple extension.

■

||||| Example 11.35 Vector Calculation Using Coordinates

We now carry out a vector calculation using coordinates. The example is not particularly mathematically interesting, but we carry it out in detail in order to demonstrate the technique of Theorem 11.34.

There are given three polynomials in the vector space $P_2(\mathbb{R})$:

$$R(x) = 2 - 3x - x^2\,, S(x) = 1 - x + 3x^2 \text{ and } T(x) = x + 2x^2\,.$$

The task is now to determine the polynomial $P(x) = 2R(x) - S(x) + 3T(x)$. We choose to carry this out using coordinates for the polynomials with respect to the monomial basis for $P_2(\mathbb{R})$.

$$\begin{aligned} {}_{\mathrm{m}}P(x) &= {}_{\mathrm{m}}\big(2R(x) - S(x) + 3T(x)\big) \\ &= {}_{\mathrm{m}}(2R(x)) + {}_{\mathrm{m}}(-S(x)) + {}_{\mathrm{m}}(3T(x)) \\ &= 2\,{}_{\mathrm{m}}R(x) - {}_{\mathrm{m}}S(x) + 3\,{}_{\mathrm{m}}T(x) \\ &= 2\begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ 1 \end{bmatrix}. \end{aligned}$$

We translate the resulting coordinate vector to the wanted polynomial:

$$P(x) = 3 - 2x + x^2\,.$$

11.6 On the Use of Coordinate Matrices

When we embark on problems with vectors and use their coordinates with respect to a given basis it often leads to a system of linear equations which we then solve by matrix calculations. One matrix is of particular importance, viz. the matrix that is formed by gathering the coordinate columns of more vectors in a *coordinate matrix*:

||||| Explanation 11.36 Coordinate Matrix for a Vector Set

If in a n-dimensional vector space V a basis a exists, and a set of m numbered vectors is given, then the ***a*-coordinate matrix** is formed by gathering the a-coordinate columns in the given order to form an $m \times n$ matrix.

By way of example consider a set of three vectors in $\mathbb{R}^2$: $((1,2),(3,4),(5,6))$. The coordinate matrix of the set with respect to the standard e-basis for $\mathbb{R}^2$ is the 2×3-matrix

$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}.$$

We will now show how coordinate matrices emerge in series of examples which we, for the sake of variation, take from different vector spaces. The methods can directly be used on other types of vector spaces, and after each example the method is demonstrated in a concentrated and general form.

It is important for your own understanding of the theory of vector spaces that you practice and realize how coordinate matrices emerge in reality when you start on typical problems.

11.6.1 Whether a Vector is a Linear Combination of Other Vectors

In $\mathbb{R}^4$ we are given four vectors

$$\begin{aligned} \mathbf{a}_1 &= (1,1,1,1) \\ \mathbf{a}_2 &= (1,0,0,1) \\ \mathbf{a}_3 &= (2,3,1,4) \\ \mathbf{b} &= (2,-2,0,1) \end{aligned}$$

Problem: Investigate if $\mathbf{b}$ is a linear combination of $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$.

Solution: We will investigate whether we can find $x_1,\ x_2,\ x_3 \in \mathbb{R}$ such that

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 = \mathbf{b}\,. \tag{11-24}$$

By theorem 11.34 we can rewrite (11-24) as the e-coordinate vector equation

$$x_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 0 \\ 1 \end{bmatrix}$$

which is equivalent to the system of linear equations

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 2 \\ x_1 + 3x_3 &= -2 \\ x_1 + x_3 &= 0 \\ x_1 + x_2 + 4x_3 &= 1 \end{aligned}$$

We form the augmented matrix of the system of equations and give (without further details) its reduced row echelon form

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 0 & 3 & -2 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 4 & 1 \end{bmatrix} \Rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{11-25}$$

From (11-25) it is seen that the rank of the coefficient matrix of the system of equations is 3, while the rank of the augmented matrix is 4. The system of equations has therefore no solutions. This means that (11-24) cannot be solved. We conclude

$$\mathbf{b} \notin \operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}\,.$$

|||| Method 11.37 Linear Combination

You can decide whether a given vector $\mathbf{b}$ is a linear combination of other vectors $\mathbf{a}_1$, $\mathbf{a}_2, \ldots, \mathbf{a}_p$ by solving the system of linear equations which has the augmented matrix that is equal to the coordinate matrix for $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p, \mathbf{b}\,)$ with respect to a given basis.

NB: In general there can be none, one or infinitely many ways a vector can be written as linear combinations of the others.

11.6.2 Whether Vectors are Linearly Dependent

We consider in the vector space $\mathbb{R}^{2\times 3}$ the three matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 2 & 2 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} -1 & -2 & 9 \\ 0 & 0 & 4 \end{bmatrix}. \tag{11-26}$$

Problem: Investigate whether the three matrices are linearly dependent.

Solution: We use theorem 11.17 and try to find three real numbers x_1, x_2 of x_3 that are not all equal to 0, but which satisfy

$$x_1\mathbf{A} + x_2\mathbf{B} + x_3\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{11-27}$$

By theorem 11.34 we can rewrite (11-27) as the e-coordinate vector equation

$$x_1 \begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \\ 2 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 3 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ -2 \\ 9 \\ 0 \\ 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

That is equivalent to the homogeneous system of linear equations with the augmented matrix that here is written together with reduced row echelon form (details are omitted):

$$\mathbf{T} = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & -2 & 0 \\ 3 & 0 & 9 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 2 & 1 & 4 & 0 \end{bmatrix} \Rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{11-28}$$

From (11-28) we see that both the coefficient matrix and the augmented matrix have the rank 2, and since the number of unknowns is larger, viz. 3, we conclude that Equation (11-27) has infinitely many solutions , see Theorem 6.33. Hence the three matrices are linearly dependent. For instance, from $\operatorname{rref}(\mathbf{T})$ one can derive that

$$-3\mathbf{A} + 2\mathbf{B} + \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

|||| **Method 11.38 Linear Dependence or Independence**

One can decide whether the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$ are linearly dependent by solving the linear homogenous system of linear equations with the augmented matrix that is equal to the coordinate matrix for $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p, \mathbf{0})$ with respect to a given basis.

NB: Since the system of equations is homogeneous, there will be either one solution or infinitely many solutions. If the rank of the coordinate matrix is equal to p, there is one solution, and this solution must be the zero solution, and the p vectors are therefore linearly independent. If the rank of the coordinate matrix is less than p, there are infinitely many solutions, including non-zero solutions, and the p vectors are therefore linearly dependent.

11.6.3 Whether a Set of Vectors is a Basis

In an n-dimensional vector space we require n basis vectors, see theorem 11.22. When one has asked whether a given set of vectors can be a basis, one can immediately conclude that this is not the case if the number of vectors in the set is not equal to n. But if there *are* n vectors in the set according to theorem 11.29 we need only investigate whether the set is linear independent, and for this we already have method 11.38. However we can in an interesting way develop the method further by using the determinant of the coordinate matrix of the vector set!

Let us e.g. investigate whether the polynomials

$$P_1(x) = 1 + 2x^2,\ P_2(x) = 2 - x + x^2 \ \text{ of } \ P_3(x) = 2x + x^2$$

form a basis for $P_2(\mathbb{R})$. Since $\dim(P_2(\mathbb{R})) = 3$, the number of polynomials is compatible with being a basis. In order to investigate whether they also are linearly independent, we use their coordinate vectors with respect to the ***monomial basis*** and consider the equation:

$$x_1 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The vectors are linearly independent if and only if the only solution is the trivial solution $x_1 = x_2 = x_3 = 0$. The equation is equivalent to a homogeneous system of linear equations consisting of 3 equations in 3 unknowns. The coefficient matrix and the augmented matrix of the system are:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & -1 & 2 \\ 2 & 1 & 1 \end{bmatrix} \text{ and } \mathbf{T} = \left[\begin{array}{ccc|c} 1 & 2 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 2 & 1 & 1 & 0 \end{array}\right].$$

As for every homogeneous system of linear equations the right hand side of the augmented matrix consists of only 0's, therefore $\rho(\mathbf{A}) = \rho(\mathbf{T})$, and thus solutions *do* exist. There is one solution exactly when $\rho(\mathbf{A})$ is equal to the number of unknowns, that is 3. And this solution must be the zero solution $x_1 = x_2 = x_3 = 0$, since L_{hom} always contains the zero solution.

Here we can use that **A** is a square matrix and thus has a *determinant*. **A** has full rank exactly when it is *invertible*, that is when $\det(\mathbf{A}) \neq 0$.

Since a calculation shows that $\det(\mathbf{A}) = 5$ we conclude that $\big(P_1(x), P_2(x), P_3(x)\big)$ constitutes a basis for $P_2(\mathbb{R})$.

|||| Method 11.39 Proof of a Basis, given n vectors

Given an n-dimensional vector space V. To determine whether a vector set consisting of n vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is a basis for V, we only need to investigate whether the set is linearly independent. A particular option for this investigation occurs because the coordinate matrix of the vector set is a square $n \times n$ matrix:

The set constitutes a basis for V exactly when the determinant of the coordinate matrix of the set with respect to a basis a is non-zero, in short

$$\big(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\big) \text{ is a basis } \Leftrightarrow \det\!\Big(\big[\,{}_{\mathrm{a}}\mathbf{v}_1 \quad {}_{\mathrm{a}}\mathbf{v}_2 \quad \cdots \quad {}_{\mathrm{a}}\mathbf{v}_n\,\big]\Big) \neq 0\,. \tag{11-29}$$

11.6.4 To Find New Coordinates when the Basis is Changed

An important technical problem for the advanced use of linear algebra is to be able to calculate new coordinates for a vector when a new basis is chosen. In this context a particular *change of basis matrix* plays an important role. We now demonstrate how basis matrices emerge.

In a 3-dimensional vector space V a basis a is given. We now choose a new basis b that is determined by the a-coordinates of the basis vectors:

$$ {}_{\mathrm{a}}\mathbf{b}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} , \; {}_{\mathrm{a}}\mathbf{b}_2 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \text{ and } {}_{\mathrm{a}}\mathbf{b}_3 = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix} . $$

Problem 1: Determine the a-coordinates for a vector $\mathbf{v}$ given by the b-coordinates as:

$$ {}_{\mathrm{b}}\mathbf{v} = \begin{bmatrix} 5 \\ -4 \\ -1 \end{bmatrix} . \tag{11-30} $$

Solution: The expression (11-30) corresponds to the vector equation

$$ \mathbf{v} = 5\mathbf{b}_1 - 4\mathbf{b}_2 - 1\mathbf{b}_3 $$

which we below first convert to an a-coordinate vector equation, re-writing the right hand side as a matrix-vector product, before finally computing the result:

$$ \begin{aligned} {}_{\mathrm{a}}\mathbf{v} &= 5 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 4 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - 1 \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 3 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ -4 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix} . \end{aligned} $$

Notice that the 3×3-matrix in the last equation is the coordinate matrix for the b-basis vectors with respect to basis a. It plays an important role, since we apparently can determine the a-coordinates for $\mathbf{v}$ by multiplying b-coordinate vector for $\mathbf{v}$ on the left by this matrix! Therefore the matrix is given the name *change of basis matrix*. The property of this matrix is that it translates b-coordinates to a-coordinates, and it is given the symbol ${}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}$. The coordinate change relation can then be written in this convenient way

$$ {}_{\mathrm{a}}\mathbf{v} = {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\, {}_{\mathrm{b}}\mathbf{v} . \tag{11-31} $$

Problem 2: Determine the b-coordinates for a vector $\mathbf{u}$ that has a-coordinates:

$$ {}_{\mathrm{a}}\mathbf{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} . \tag{11-32} $$

Solution: Since ${}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}$ is the coordinate matrix for a basis, it is *invertible*, and thus has an inverse matrix. We therefore use the coordinate change relation (11-31) as follows:

$$\begin{aligned}{}_{\mathrm{a}}\mathbf{u} &= {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{u} \;\Leftrightarrow\\ {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}{}^{-1}\,{}_{\mathrm{a}}\mathbf{u} &= {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}{}^{-1}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{u} \;\Leftrightarrow\\ {}_{\mathrm{b}}\mathbf{u} &= {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}{}^{-1}\,{}_{\mathrm{a}}\mathbf{u} \;\Leftrightarrow\\ {}_{\mathrm{b}}\mathbf{u} &= \begin{bmatrix}1&1&2\\1&0&3\\1&2&0\end{bmatrix}^{-1}\begin{bmatrix}1\\2\\3\end{bmatrix} = \begin{bmatrix}11\\-4\\-3\end{bmatrix}.\end{aligned}$$

|||| Method 11.40 Coordinate Change when the Basis is Changed

When a basis a is given for a vector space, and when a new basis b is known by the a-coordinates of its basis vectors, the ***change of basis matrix*** ${}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}$ is identical to the a-coordinate matrix for b-basis vectors.

1. If b-coordinates for a vector $\mathbf{v}$ are known, these a-coordinates can be found by the matrix-vector product:
$$ {}_{\mathrm{a}}\mathbf{v} = {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{v}\,. $$

2. Conversely, if the a-coordinates for $\mathbf{v}$ are known, the b-coordinates can be found by the matrix-vector product:
$$ {}_{\mathrm{b}}\mathbf{v} = {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}{}^{-1}\,{}_{\mathrm{a}}\mathbf{v}\,. $$

In short the change of basis matrix that translates a-coordinates to b-coordinates, is the inverse of the change of basis matrix that translate b-coordinates to a-coordinates:

$$ {}_{\mathrm{b}}\mathbf{M}_{\mathrm{a}} = \left({}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\right)^{-1}. $$

11.7 Subspaces

Often you encounter that a subset of a vector space is itself a vector space. In Figure 11.3 are depicted two position vectors $\overrightarrow{OP}$ and $\overrightarrow{OQ}$ that span the plane F:

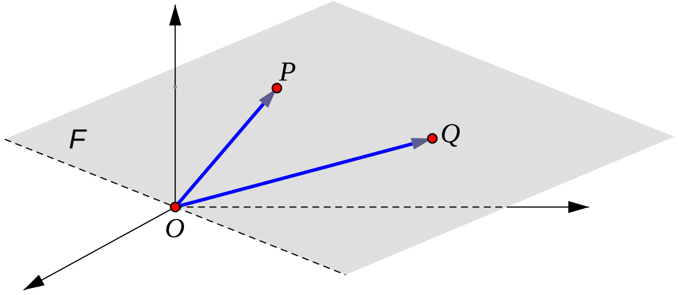


Figure 11.3: A plane through the origin interpreted as a *subspace* in space

Since span$\{\overrightarrow{OP}, \overrightarrow{OQ}\}$ can be considered to be a (2-dimensional) vector space in its own right, it is named a *subspace* of the (3-dimensional) vector space of position vectors in space.

|||| Definition 11.41 Subspace

A subset U of a vector space V is called a ***subspace*** of V if U is itself a vector space.

In any vector space V one can immediately point to two subspaces:
1) V is in itself a subspace of V.
2) The set $\{\,\mathbf{0}\,\}$ is a subspace of V.
These subspaces are called the *trivial* subspaces in V.

When one must check whether a subset is a subspace, one only has to check whether the stability requirements are satisfied:

|||| Theorem 11.42 Sufficient Conditions for a Subspace

A non-empty subset U of a vector space V is a subspace of V if U is *stable* with respect to addition and multiplication by a scalar. This means

1. The sum of two vectors from U belongs to U.
2. The product of a vector in U with a scalar belongs to U.

|||| Proof

Since U satisfies the two stability requirements in 11.1, it only remains to show that U also satisfies the eight arithmetic rules in the definition. But this is evident since all vectors in U are also vectors in V where the rules apply.

■

|||| Example 11.43 Basis for a Subspace

We consider a subset M_1 of $\mathbb{R}^{2\times 2}$, consisting of matrices of the type

$$\begin{bmatrix} a & b \\ b & a \end{bmatrix} \tag{11-33}$$

where a and b are arbitrary real numbers. We try to add two matrices of the type (11-33)

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 4 \end{bmatrix}$$

and we multiply one of type (11-33) by a scalar

$$-3\begin{bmatrix} 2 & -3 \\ -3 & 2 \end{bmatrix} = \begin{bmatrix} -6 & 9 \\ 9 & -6 \end{bmatrix}.$$

in both cases the resulting matrix is of type (11-33) and it is obvious that this would also apply had we used other examples. Therefore M_1 satisfies the stability requirements for a vector space. Thus it follows from theorem 11.42 that M_1 is a subspace of $\mathbb{R}^{2\times 2}$.

Further remark that M_1 is spanned by two linear independent 2×2 matrices since

$$\begin{bmatrix} a & b \\ b & a \end{bmatrix} = a\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Therefore M_1 is a 2-dimensional subspace of $\mathbb{R}^{2\times 2}$, and a possible basis for M_1 is given by

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

|||| Example 11.44 A Subset which is Not a Subspace

The subset M_2 of $\mathbb{R}^{2\times 2}$ consists of all matrices of the type

$$\begin{bmatrix} a & b \\ a\cdot b & 0 \end{bmatrix} \tag{11-34}$$

where a and b are arbitrary real numbers. We try to add two matrices of the type (11-34)

$$\begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 3 \\ 6 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 8 & 0 \end{bmatrix}.$$

Since $8 \neq 3\cdot 5$, this matrix is not of the type (11-34). Therefore M_2 is not stable under linear combinations, and cannot be a subspace.

11.7.1 About Spannings as Subspaces

|||| Theorem 11.45 Spannnings of Subspaces

For arbitrary vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ in vector space V, the set $\operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p\}$ is a subspace of V .

|||| Proof

The stability requirements are satisfied because 1) the sum of two linear combinations of the p vectors in itself is a linear combination of them and 2) a linear combination of the p vectors multiplied by a scalar in itself is a linear combination of them. The rest follows from Theorem 11.42.

■

The solution set L_{hom} for a homogeneous system of linear equations with n unknowns is always a subspace of the number space $\mathbb{R}^n$ and the dimension of the subspace is the same as the number of free parameters in L_{hom} . We show an example of this below.

|||| Example 11.46 L_{hom} is a Subspace

The following homogeneous system of linear equations of 3 equations in 5 unknowns

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

has the solution set (details are omitted):

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = t_1 \begin{bmatrix} -2 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \begin{bmatrix} 11 \\ -4 \\ 0 \\ -1 \\ 1 \end{bmatrix} \quad \text{where } \; t_1, t_2 \in \mathbb{R}\,. \tag{11-35}$$

We see that L_{hom} is a span of two vectors in $\mathbb{R}^5$. Then it is according to theorem 11.45 a subspace of $\mathbb{R}^5$. Since the two vectors evidently are linearly independent, L_{hom} is a 2-dimensional subspace of $\mathbb{R}^5$, with a basis

$$\big(\,(-2,0,1,0,0)\,,(11,-4,0,-1,1)\,\big)\,.$$

In the following example we will establish a method for how one can determine a basis for a subspace that is spanned by a number of given vectors in a subspace.

Consider in $\mathbf{R}^3$ four vectors

$$\mathbf{v}_1 = (1,2,1),\ \mathbf{v}_2 = (3,0,-1),\ \mathbf{v}_3 = (-1,4,3) \text{ and } \mathbf{v}_4 = (8,-2,-4)\,.$$

We wish to find a basis for the subspace

$$U = \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}\ .$$

Let $\mathbf{b} = (b_1, b_2, b_3)$ be an arbitrary vector in U. We thus assume that the following vector equation has a solution:

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + x_3\mathbf{v}_3 + x_4\mathbf{v}_4 = \mathbf{b}\,. \tag{11-36}$$

By substitution of the five vectors into (11-36), it is seen that (11-36) is equivalent to an inhomogeneous system of linear equations with the augmented matrix:

$$\mathbf{T} = \left[\begin{array}{rrrr|c} 1 & 3 & -1 & 8 & b_1 \\ 2 & 0 & 4 & -2 & b_2 \\ 1 & -1 & 3 & -4 & b_3 \end{array}\right] \Rightarrow \operatorname{rref}(\mathbf{T}) = \left[\begin{array}{rrrr|c} 1 & 0 & 2 & -1 & c_1 \\ 0 & 1 & -1 & 3 & c_2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{11-37}$$

Here c_1 is placeholder for the number that b_1 has been transformed into following the row operations leading to the reduced row echelon form $\operatorname{rref}(\mathbf{T})$. Similarly for c_2. Remark that b_3 after the row operations must be transformed into 0, or else (x_1, x_2, x_3, x_4) could not be a solution as we have assumed.

But it is in particular the leading 1's in $\operatorname{rref}(\mathbf{T})$ on which we focus! They show that $\mathbf{v}_1$ and $\mathbf{v}_2$ span all of U, and that $\mathbf{v}_1$ and $\mathbf{v}_2$ are linear independent. We can convince ourselves of both by considering equation (11-36) again.

First: Suppose we had only asked whether $\mathbf{v}_1$ and $\mathbf{v}_2$ span all of U. Then we should have omitted the terms with $\mathbf{v}_3$ and $\mathbf{v}_4$ from (11-36), and then we would have obtained:

$$\operatorname{rref}(\mathbf{T}_2) = \left[\begin{array}{cc|c} 1 & 0 & c_1 \\ 0 & 1 & c_2 \\ 0 & 0 & 0 \end{array}\right]$$

that shows that $c_1\,\mathbf{v}_1 + c_2\,\mathbf{v}_2 = \mathbf{b}$, and that $\mathbf{v}_1$ and $\mathbf{v}_2$ then span all of U.

Secondly: Suppose we had asked whether $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly independent. Then we should have omitted the terms with $\mathbf{v}_3$ and $\mathbf{v}_4$ from (11-36), and put $\mathbf{b} = \mathbf{0}$. And then we would have got:

$$\operatorname{rref}(\mathbf{T}_3) = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

That shows that the zero vector can only be written as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ if both of the coefficients x_1 and x_2 are 0. And thus we show that $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly independent. In total we have shown that $(\mathbf{v}_1, \mathbf{v}_2)$ is a basis for U.

The conclusion is that a basis for U can be singled out by the leading 1's in $\operatorname{rref}(\mathbf{T})$, see (11-37). The right hand side in $\operatorname{rref}(\mathbf{T})$ was meant to serve our argument but its contribution is now unnecessary. Therefore we can summarize the result as the following method:

|||| Method 11.47 About refining a Spanning Set to a Basis

When, in a vector space V, for which a basis a has been chosen, one wishes to find a basis for the subspace

$$U = \text{span}\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$$

everything can be read from

$$\text{rref}\left(\begin{bmatrix} {}_a\mathbf{v}_1 & {}_a\mathbf{v}_2 & \ldots & {}_a\mathbf{v}_p \end{bmatrix}\right). \tag{11-38}$$

If in the i'th column in (11-38) there are no leading 1's, then $\mathbf{v}_i$ is deleted from the set $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$. The set reduced in this way is a basis for U.

Since the number of leading 1's in (11-38) is equal to the number of basis vectors in the chosen basis for U, it follows that

$$\text{Dim}(U) = \rho\left(\begin{bmatrix} {}_a\mathbf{v}_1 & {}_a\mathbf{v}_2 & \ldots & {}_a\mathbf{v}_p \end{bmatrix}\right). \tag{11-39}$$

11.7.2 Infinite-Dimensional Vector Space

Before we end this eNote, that has cultivated the use of bases and coordinates, we must admit that not all vector spaces have a basis. Viz. there exist ***infinite-dimensional vector spaces***.

This we can see through the following example:

|||| Example 11.48 Infinite-Dimensional Vector Spaces

All polynomials in the vector space $P_n(\mathbb{R})$ are continuous functions, therefore $P_n(\mathbb{R})$ is an n+1 dimensional subspace of the vector space $C^0(\mathbb{R})$ of all real continuous functions. Now consider $P(\mathbb{R})$, the set of all real polynomials, that for the same reason is also a subspace of $C^0(\mathbb{R})$. But $P(\mathbb{R})$ must be *infinite-dimensional*, since it has $P_n(\mathbb{R})$, for every n, as a subspace. For the same reason $C^0(\mathbb{R})$ must also be infinite-dimensional.

|||| Exercise 11.49

By $C^1(\mathbb{R})$ is understood the set of all differentiable functions, with $\mathbb{R}$ as their domain, and with continuous derivatives in $\mathbb{R}$.

Explain why $C^1(\mathbb{R})$ is an infinite-dimensional subspace of $C^0(\mathbb{R})$.

eNote 12

Linear Transformations

This eNote investigates an important type of transformation (or map) between vector spaces, viz. linear transformations. It is shown that the kernel and the range for linear transformations are subspaces of the domain and the codomain, respectively. When the domain and the codomain have finite dimensions and a basis has been chosen for each, questions about linear maps can be standardized. In that case a linear transformation can be expressed as a product between a so-called standard matrix for the transformation and the coordinates of the vectors that we want to map. Since standard matrices depend on the chosen bases, we describe how the standard matrices are changed when one of the bases or both are replaced. The prerequisite for the eNote is knowledge about systems of linear equations, see eNote 6, about matrix algebra, see eNote 7 and about vector spaces, see eNote 10.

Updated: 15.11.21 David Brander

12.1 About Maps

A *map* (also known as a *function*) is a rule f that for every element in a set A attaches an element in a set B, and the rule is written $f : A \to B$. A is called the ***domain*** and B the ***codomain***.

CPR-numbering is a map from the set of citizens in Denmark into $\mathbb{R}^{10}$. Note that there is a 10-times infinity of elements in the codomain $\mathbb{R}^{10}$, so luckily we only need a small subset, about five million! The elements in $\mathbb{R}^{10}$ that in a given instant are in use are the ***range*** for the CPR-map.

Elementary functions of the type $f : \mathbb{R} \to \mathbb{R}$. are simple maps. The meaning of the arrow is that f to every real number x attaches another real number $y = f(x)$. Consider e.g. the continuous function:

$$y = f(x) = \frac{1}{2}x^2 - 2\,. \tag{12-1}$$

Here the function has the form of a calculation procedure: Square the number, multiply the result by one half and subtract 2. Elementary functions have a great advantage in that their graph $\{\,(x,y)\,|\,y = f(x)\,\}$ can be drawn to give a particular overview of the map (Figure 12.1).

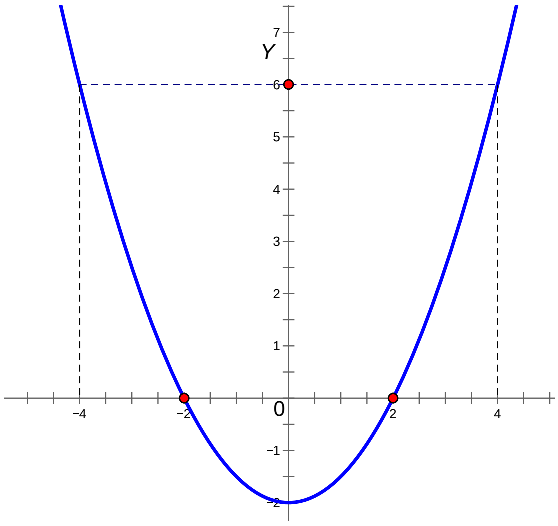


Figure 12.1: Graph of an elementary function

Typical questions in connection with elementary functions reappear in connection with more advanced maps. Therefore let us as an introduction consider some of the most important ones:

1. Determine the zeros of f. This means we must find all x for which $f(x) = 0$. In the example above the answer is $x = -2$ and $x = 2$.

2. Solve for a given b the equation $f(x) = b$. For $b = 6$ there are in the example two solutions: $x = -4$ and $x = 4$.

3. Determine the range for f. We must find all those b for which the equation $f(x) = b$ has a solution. In the example the range is $[\,-2;\infty\,[$.

In this eNote we look at domains, codomains and ranges that are *vector spaces*. A map $f : V \rightarrow W$ attaches to every vector $\mathbf{x}$ in the *domain* V a vector $\mathbf{y} = f(\mathbf{x})$ in the *codomain* W. All the vectors in W that are images of vectors in V together constitute the *range*.

|||| Example 12.1 Mapping from a Vector Space to a Vector Space

A map $g : \mathbb{R}^{2\times 3} \to \mathbb{R}^{2\times 2}$ is given by

$$\mathbf{Y} = g(\mathbf{X}) = \mathbf{X}\mathbf{X}^{\top}\,. \tag{12-2}$$

Then e.g.

$$g\left(\begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 3 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 9 \end{bmatrix}.$$

12.2 Examples of Linear Maps in the Plane

We investigate in the following a map f that has the geometric vectors in the plane as both the domain and codomain. For a given geometric vector $\mathbf{x}$ we will by $\hat{\mathbf{x}}$ understand its *hat vector*, i.e. $\mathbf{x}$ rotated $\pi/2$ counter-clockwise. Consider the map f given by

$$\mathbf{y} = f(\mathbf{x}) = 2\,\hat{\mathbf{x}}\,. \tag{12-3}$$

To every vector in the plane there is attached its hat vector multiplied (extended) by 2. In Figure 12.2 two vectors $\mathbf{u}$ and $\mathbf{v}$ and their images $f(\mathbf{u})$ and $f(\mathbf{v})$ are drawn.

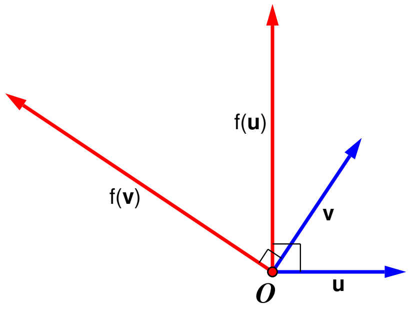


Figure 12.2: Two vectors (blue) and their images (red).

Figure 12.2 gives rise to a couple of interesting questions: How is the vector sum $\mathbf{u}+\mathbf{v}$ mapped? More precisely: How is the image vector $f(\mathbf{u}+\mathbf{v})$ related to the two image vectors $f(\mathbf{u})$ and $f(\mathbf{v})$? And what is the relation between the image vectors $f(k\mathbf{u})$ and $f(\mathbf{u})$, when k is a given real number?

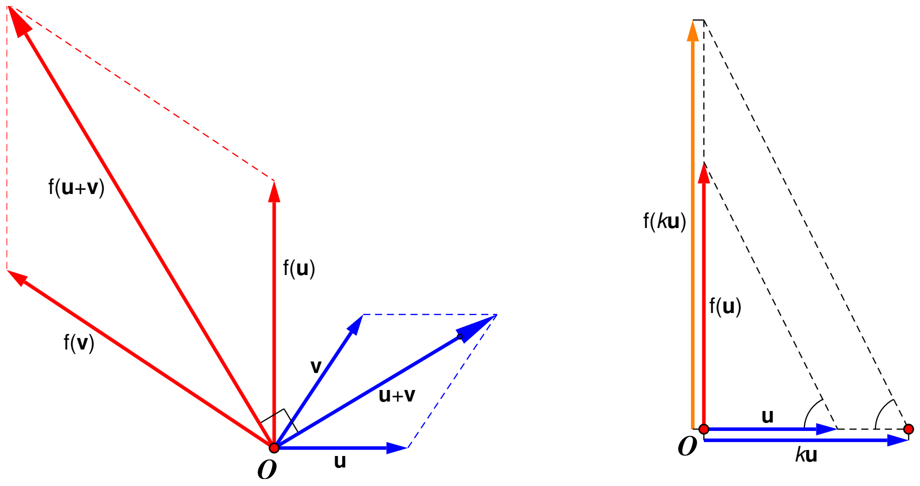


Figure 12.3: Construction of $f(\mathbf{u}+\mathbf{v})$ and $f(k\mathbf{u})$.

As indicated in Figure 12.3, f satisfies two very simple rules:

$$f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}) \text{ and } f(k\mathbf{u}) = k\,f(\mathbf{u})\,. \tag{12-4}$$

Using the well known computational rules for hat vectors

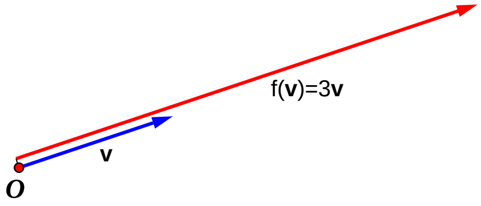
1. $\widehat{\mathbf{u}+\mathbf{v}} = \hat{\mathbf{u}} + \hat{\mathbf{v}}$.
2. $\widehat{k\mathbf{u}} = k\hat{\mathbf{u}}\,$.

we can now confirm the statement (12-4) :

$$\begin{aligned} f(\mathbf{u}+\mathbf{v}) &= 2\widehat{\mathbf{u}+\mathbf{v}} = 2(\hat{\mathbf{u}}+\hat{\mathbf{v}}) = 2\hat{\mathbf{u}} + 2\hat{\mathbf{v}} \\ &= f(\mathbf{u}) + f(\mathbf{v}) \\ f(k\mathbf{u}) &= 2\widehat{k\mathbf{u}} = 2k\hat{\mathbf{u}} = k(2\hat{\mathbf{u}}) \\ &= k\,f(\mathbf{u}) \end{aligned}$$

|||| Exercise 12.2

A map f_1 of plane vectors is given by $f_1(\mathbf{v}) = 3\mathbf{v}$:

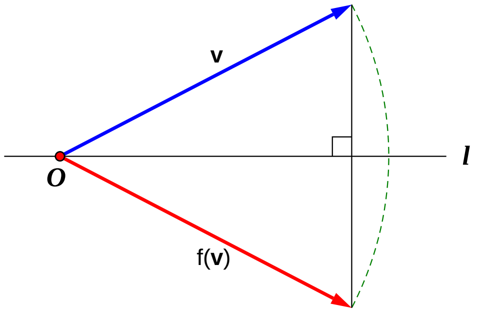


Scaling of vectors

Draw a figure that demonstrates that f_1 satisfies the rules (12-4) .

|||| Exercise 12.3

In the plane a line l through the origin is given. A map f_2 reflects vectors drawn from the origin in l :

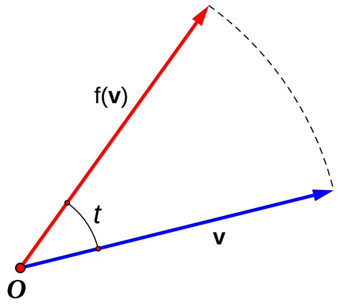


Reflection of a vector

Draw a figure that demonstrates that f_2 satisfies the rules (12-4) .

|||| Exercise 12.4

A map f_3 turns vectors drawn from the origin the angle t about the origin counterclockwise:



Rotation of a vector

Draw a figure that demonstrates that f_3 satisfies the rules (12-4) .

All maps mentioned in this section are *linear*, because they satisfy (12-4) . We now turn to a general treatment of linear mappings between vector spaces.

12.3 Linear Maps

|||| Definition 12.5 Linear Map

Let V and W be two vector spaces and let $\mathbb{L}$ denote either $\mathbb{R}$ or $\mathbb{C}$. A map $f : V \to W$ is called ***linear*** if for all $\mathbf{u}, \mathbf{v} \in V$ and all scalars $k \in \mathbb{L}$ it satisfies the following two *linearity requirements*:

$$\begin{aligned} L_1 : &\quad f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})\,. \\ L_2 : &\quad f(k\mathbf{u}) = k\, f(\mathbf{u})\,. \end{aligned}$$

V is called the *domain* and W the *codomain* for f .

By putting $k = 0$ in the linearity requirement L_2 in the definition 12.5, we see that

$$f(\mathbf{0}) = \mathbf{0}\,. \tag{12-5}$$

In other words for every linear map $f : V \to W$ the zero vector in V is mapped to the zero vector in W.

The image of a linear combination becomes in a very simple way a linear combination of the images of the vectors that are part of the given linear combination:

$$f(k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \ldots + k_p\mathbf{v}_p) = k_1 f(\mathbf{v}_1) + k_2 f(\mathbf{v}_2) + \ldots + k_p f(\mathbf{v}_p)\,. \tag{12-6}$$

This result is obtained by repeated application of L_1 and L_2.

|||| Example 12.6 Linear Map

A map $f : \mathbb{R}^2 \to \mathbb{R}^4$ is given by the rule

$$f(x_1, x_2) = (0, x_1, x_2, x_1 + x_2)\,. \tag{12-7}$$

$\mathbb{R}^2$ and $\mathbb{R}^4$ are vector spaces and we investigate whether f is a linear map. First we test the left hand side and the right hand side of L_1 with the vectors $(1, 2)$ and $(3, 4)$:

$$\begin{aligned} f(\,(1,2) + (3,4)\,) &= f(4,6) = (0,4,6,10)\,. \\ f(1,2) + f(3,4) &= (0,1,2,3) + (0,3,4,7) = (0,4,6,10)\,. \end{aligned}$$

Then L_2 is tested with the vector (2,3) and the scalar 5:

$$\begin{aligned} f(\,5\cdot(2,3)\,) &= f(10,15) = (0,10,15,25)\,. \\ 5\cdot f(2,3) &= 5\cdot(0,2,3,5) = (0,10,15,25)\,. \end{aligned}$$

The investigatíon suggests that f is linear. This is now shown generally. First we test L_1:

$$\begin{aligned} f(\,(x_1,x_2) + (y_1,y_2)\,) &= f(x_1+y_1, x_2+y_2) = (0, x_1+y_1, x_2+y_2, x_1+x_2+y_1+y_2)\,. \\ f(x_1,x_2) + f(y_1,y_2) &= (0, x_1, x_2, x_1+x_2) + (0, y_1, y_2, y_1+y_2) \\ &= (0, x_1+y_1, x_2+y_2, x_1+x_2+y_1+y_2)\,. \end{aligned}$$

Then we test L_2:

$$\begin{aligned} f(\,k\cdot(x_1,x_2)\,) &= f(k\cdot x_1, k\cdot x_2) = (0, k\cdot x_1, k\cdot x_2, k\cdot x_1 + k\cdot x_2)\,. \\ k\cdot f(x_1,x_2) &= k\cdot(0, x_1, x_2, x_1+x_2) = (0, k\cdot x_1, k\cdot x_2, k\cdot x_1 + k\cdot x_2)\,. \end{aligned}$$

It is seen that f satisfies both linearity requirements and therefore is linear.

|||| Example 12.7 A Map that is Not Linear

In the example 12.1 we considered the map $g : \mathbb{R}^{2\times 3} \to \mathbb{R}^{2\times 2}$ given by

$$\mathbf{Y} = g(\mathbf{X}) = \mathbf{X}\,\mathbf{X}^{\top}\,. \tag{12-8}$$

That this map is *not* linear, can be shown by finding an example where either L_1 or L_2 is not valid. Below we give an example of a matrix $\mathbf{X}$ that does not satisfy $g(2\mathbf{X}) = 2\,g(\mathbf{X})$:

$$g\left(2\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix}\right) = g\left(\begin{bmatrix}2&0&0\\0&0&0\end{bmatrix}\right) = \begin{bmatrix}2&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}2&0\\0&0\\0&0\end{bmatrix} = \begin{bmatrix}4&0\\0&0\end{bmatrix}.$$

But

$$2\,g\left(\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix}\right) = 2\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}1&0\\0&0\\0&0\end{bmatrix} = 2\begin{bmatrix}1&0\\0&0\end{bmatrix} = \begin{bmatrix}2&0\\0&0\end{bmatrix}.$$

Therefore g does not satisfy the linearity requirements L_2, hence g is not linear.

|||| Example 12.8 Linear Map

A map $f : P_2(\mathbb{R}) \to \mathbb{R}$ is given by the rule

$$f\big(\,P(x)\,\big) = P'(1)\,. \tag{12-9}$$

For every second degree polynomial the slope of the tangent at $x = 1$ is attached. An arbitrary second degree polynomial P can be written as $P(x) = ax^2 + bx + c$, where a, b and c are real constants. Since $P'(x) = 2ax + b$ we have:

$$f\big(\,P(x)\,\big) = 2a + b\,.$$

If we put $P_1(x) = a_1x^2 + b_1x + c_1$ and $P_2(x) = a_2x^2 + b_2x + c_2$, we get

$$\begin{aligned} f\big(\,P_1(x) + P_2(x)\,\big) &= f\big(\,(a_1+a_2)x^2 + (b_1+b_2)x + (c_1+c_2)\,\big)\\ &= \big(\,2(a_1+a_2) + (b_1+b_2)\,\big)\\ &= (2a_1+b_1) + (2a_2+b_2)\\ &= f\big(\,P_1(x)\,\big) + f\big(\,P_2(x)\,\big)\,. \end{aligned}$$

Furthermore for every real number k and every second degree polynomial $P(x)$:

$$\begin{aligned} f\big(\,k\cdot P(x)\,\big) &= f\big(\,k\cdot ax^2 + k\cdot bx + k\cdot c\,\big)\\ &= (2k\cdot a + k\cdot b) = k\cdot(2a+b)\\ &= k\cdot f\big(\,P(x)\,\big)\,. \end{aligned}$$

It is hereby shown that f satisfies the linearity conditions L_1 and L_2, and that f thus is a linear map.

|||| Exercise 12.9

By $C^\infty(\mathbb{R})$ we understand the vector space consisting of all functions $f : \mathbb{R} \to \mathbb{R}$ that can be differentiated an arbitrary number of times. One example (among infinitely many) is the sine function. Consider the map $\mathrm{D} : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ that to a function $f(x) \in C^\infty(\mathbb{R})$ assigns its derivative:

$$\mathrm{D}(\,f(x)\,) = f'(x)\,.$$

Show that D is a linear map.

12.4 Kernel and Range

The zeros of an elementary function $f : \mathbb{R} \to \mathbb{R}$ are all the real numbers x that satisfy $f(x) = 0$. The corresponding concept for linear maps is called the *kernel*. The range of an elementary function $f : \mathbb{R} \to \mathbb{R}$ are all the real numbers b for each of which a real number x exists such that $f(x) = b$. The corresponding concept for linear maps is also called the *range* or *image*. The kernel is a subspace of the domain and the range is a subspace of the codomain. This is now shown.

|||| Definition 12.10 Kernel and Range

By the ***kernel*** of a linear map $f : V \to W$ we understand the set:

$$\ker(f) = \{\,\mathbf{x} \in V \,|\, f(\mathbf{x}) = \mathbf{0} \in W\,\}\,. \tag{12-10}$$

By the ***range*** or ***image*** of f we understand the set:

$$f(V) = \{\,\mathbf{b} \in W \,|\, \text{At least one } \mathbf{x} \in V \text{ exists with } f(\mathbf{x}) = \mathbf{b}\,\}\,. \tag{12-11}$$

|||| Theorem 12.11 The Kernel and the Range are Subspaces

Let $f : V \rightarrow W$ be a linear map. Then:

1. The kernel of f is a subspace of V.
2. The range $f(V)$ is a subspace of W.

|||| Proof

1) First, the kernel is not empty, as $f(\mathbf{0}) = 0$ by linearity. So we just need to prove that the kernel of f satisfies the stability requirements, see Theorem 11.42. Assume that $\mathbf{x}_1 \in V$ and $\mathbf{x}_2 \in V$, and that k is an arbitrary scalar. Since (using L_1):

$$f(\mathbf{x}_1 + \mathbf{x}_2) = f(\mathbf{x}_1) + f(\mathbf{x}_2) = \mathbf{0} + \mathbf{0} = \mathbf{0}\,,$$

the kernel of f is stable with respect to addition. Moreover (using L_2):

$$f(k\mathbf{x}_1) = k\,f(\mathbf{x}_1) = k\,\mathbf{0} = \mathbf{0}\,,$$

the kernel of f is also stable with respect to multiplication by a scalar. In total we had shown that the kernel of f is a subspace of V.

2) The range $f(V)$ is non-empty, as it contains the zero vector. We now show that it satisfies the stability requirements. Suppose that $\mathbf{b}_1 \in f(V)$ and $\mathbf{b}_2 \in f(V)$, and that k is an arbitrary scalar. There exist, according to the definition, see (12.10), vectors $\mathbf{x}_1 \in V$ and $\mathbf{x}_2 \in V$ that satisfy $f(\mathbf{x}_1) = \mathbf{b}_1$ and $f(\mathbf{x}_2) = \mathbf{b}_2$. We need to show that there exists an $\mathbf{x} \in V$ such that $f(\mathbf{x}) = \mathbf{b}_1 + \mathbf{b}_2$. There is, namely $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, since

$$f(\mathbf{x}_1 + \mathbf{x}_2) = f(\mathbf{x}_1) + f(\mathbf{x}_2) = \mathbf{b}_1 + \mathbf{b}_2\,.$$

Hereby it is shown that $f(V)$ is stable with respect to addition. We will, in a similar way, show that there exists an $\mathbf{x} \in V$ such that $f(\mathbf{x}) = k\mathbf{b}_1$. Here we choose $\mathbf{x} = k\mathbf{x}_1$, then

$$f(\mathbf{x}) = f(k\mathbf{x}_1) = kf(\mathbf{x}_1) = k\mathbf{b}_1\,,$$

from which it appears that $f(V)$ is stable with respect to multiplication by a scalar. In total we have shown that $f(V)$is a subspace of W.

■

But why is it so interesting that the kernel and the range of a linear map are subspaces? The answer is that it becomes simpler to describe them when we know that they possess vector space properties and we thereby in advance know their structure. It is particularly elegant when we can determine the kernel and the range by giving a basis for them. This we will try in the next two examples.

|||| Example 12.12 Determination of Kernel and Range

A linear map $f : \mathbb{R}^3 \rightarrow R^2$ is given by the rule:

$$f(x_1, x_2, x_3) = (x_1 + 2x_2 + x_3\,, -x_1 - 2x_2 - x_3)\,. \tag{12-12}$$

We wish to determine the kernel of f and the range $f(\mathbb{R}^3)$ (note that it is given that f is linear. So we omit the proof of that).

Determination of the kernel:
We shall solve the equation

$$f(\mathbf{x}) = \mathbf{0} \Leftrightarrow \begin{bmatrix} x_1 + 2x_2 + x_3 \\ -x_1 - 2x_2 - x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{12-13}$$

This is a system of linear equations consisting of two equations in three unknowns. The corresponding augmented matrix is

$$\mathbf{T} = \begin{bmatrix} 1 & 2 & 1 & 0 \\ -1 & -2 & -1 & 0 \end{bmatrix} \rightarrow \mathrm{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

We see that the system of equations has the solution set

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = t_1 \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + t_2 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The solution set is spanned by two linearly independent vectors. Therefore we can conclude that the kernel of f is a 2-dimensional subspace of $\mathbb{R}^3$ that is precisely characterized by a basis:

$$\text{Basis for the kernel : } \big(\,(-2,1,0), (-1,0,1)\,\big)\,.$$

There is an entire plane of vectors in the space, that by insertion into the expression for f give the image $\mathbf{0}$. This basis yields all of them.

Determination of the range:
We shall find all those $\mathbf{b} = (b_1, b_2)$ for which the following equation has a solution:

$$f(\mathbf{x}) = \mathbf{b} \Leftrightarrow \begin{bmatrix} x_1 + 2x_2 + x_3 \\ -x_1 - 2x_2 - x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{12-14}$$

Note, that it is not x_1, x_2 and x_3, we are looking for, as we usually do in such a system of equations. Rather it is b_1 and b_2 of the right hand side, which we will determine exactly *in those cases, when solutions exist*! Because when the system has solution of a particular right hand side, then this right-hand side *must* be in the image space that we are looking for.

This is a system of linear equations consisting of two equations in three unknowns. The corresponding augmented matrix is

$$\mathbf{T} = \begin{bmatrix} 1 & 2 & 1 & b_1 \\ -1 & -2 & -1 & b_2 \end{bmatrix} \rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 2 & 1 & b_1 \\ 0 & 0 & 0 & b_1+b_2 \end{bmatrix}$$

If $b_1 + b_2 = 0$, that is if $b_1 = -b_2$, the system of equations has infinitely many solutions. If on the contrary $b_1 + b_2 \neq 0$ there is no solution. All those $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$ that are images of at least one $\mathbf{x} \in R^3$ evidently tcan be written as:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

We conclude that $f(V)$ is a 1-dimensional subspace of $\mathbb{R}^2$ that can be characterized precisely by a basis:

$$\text{Basis for the range : } \big((-1,1)\big).$$

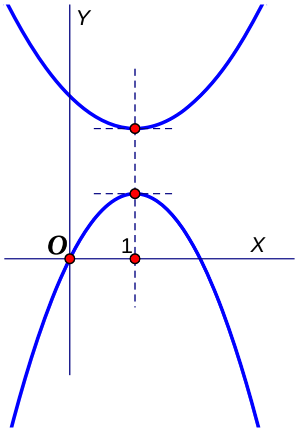


Figure 12.4: Two vectors in the kernel (Exercise 12.13)

Exercise 12.13 Determination of Kernel and Range

In example 12.8 it was shown that the map $f : P_2(\mathbb{R}) \to \mathbb{R}$ given by the rule

$$f\big(\,P(x)\,\big) = P'(1)\,. \tag{12-15}$$

is linear. The kernel of f consists of all second degree polynomials that satisfy $P'(1) = 0$. The graphs for a couple of these are shown in Figure 12.4:

Determine the kernel of f.

In eNote 6 the relation bewteen the solution set for an inhomogeneous system of linear equations and the corresponding homogeneous linear system of equations is presented in Theorem 6.37 (the structural theorem). We now show that a corresponding relation exists for all linear equations.

Theorem 12.14 The Structural Theorem for Linear Equations

Let $f : V \to W$ be a liner map and $\mathbf{y}$ an arbitrary proper vector in W. Furthermore let $\mathbf{x}_0$ be an arbitrary (so-called particular) solution to the inhomogeneous linear equation

$$f(\mathbf{x}) = \mathbf{y}\,. \tag{12-16}$$

Then the general solution L_{inhom} to the linear equation is given by

$$L_{inhom} = \left\{\, \mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1 \;\middle|\; \mathbf{x}_1 \in \ker(f)\,\right\}\,, \tag{12-17}$$

or in short

$$L_{inhom} = \mathbf{x}_0 + \ker(f)\,. \tag{12-18}$$

Proof

The theorem contains two assertions. The one is that the sum of $\mathbf{x}_0$ and an arbitrary vector from the $\ker(f)$ belongs to L_{inhom}. The other is that an arbitrary vector from L_{inhom} can be written as the sum of $\mathbf{x}_0$ and a vector from $\ker(f)$. We prove the two assertions separately:

1. Assume $\mathbf{x}_1 \in \ker(f)$. Then it applies using the linearity condition L_1 :

$$f(\mathbf{x}_1 + \mathbf{x}_0) = f(\mathbf{x}_1) + f(\mathbf{x}_0) = \mathbf{0} + \mathbf{y} = \mathbf{y} \tag{12-19}$$

   by which it is also shown that $\mathbf{x}_1 + \mathbf{x}_0$ is a solution to (12-16).

2. Assume that $\mathbf{x}_2 \in L_{inhom}$. The it applies using the linearity condition L_1 :

$$f(\mathbf{x}_2 - \mathbf{x}_0) = f(\mathbf{x}_2) - f(\mathbf{x}_0) = \mathbf{y} - \mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{x}_2 - \mathbf{x}_0 \in \ker(f)\,. \tag{12-20}$$

   Thus a vector $\mathbf{x}_1 \in \ker(f)$ exists that satisfy

$$\mathbf{x}_2 - \mathbf{x}_0 = \mathbf{x}_1 \Leftrightarrow \mathbf{x}_2 = \mathbf{x}_0 + \mathbf{x}_1 \tag{12-21}$$

   whereby we have stated $\mathbf{x}_2$ in the form wanted. The proof is hereby complete.

■

|||| Exercise 12.15

Consider the map $\mathrm{D} : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ from Exercise 12.9 that to the function $f \in C^\infty(\mathbb{R})$ relates its derivative:

$$\mathrm{D}(\,f(x)\,) = f'(x)\,.$$

State the complete solution to inhomogeneous linear eequation

$$\mathrm{D}(f(x)) = x^2$$

and interprete this in the light of the structural theorem.

12.5 Mapping Matrix

All linear maps from a finite dimensional domain V to a finite dimensional codomain W can be described by a *mapping matrix*. This is the subject of this subsection. The prerequisite is only that a basis for both V and W is chosen, and that we turn from vector calculation to calculation using the coordinates with respect to the chosen bases. The great advantage by this setup is that we can construct general methods of calculation valid for all linear maps between finite dimensional vector spaces. We return to this subject, see section 12.6. First we turn to mapping matrix construction.

Let $\mathbf{A}$ be a real or complex $m \times n-$matrix. We consider a map $f : \mathbb{L}^n \to \mathbb{L}^m$ that has the form of a f matrix-vector product:

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\,\mathbf{x}\,. \tag{12-22}$$

Using the matrix product computation rules from Theorem 7.13, we obtain for every choice of $\mathbf{x}_1\,,\mathbf{x}_2 \in \mathbb{R}^n$ and every scalar k:

$$f(\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{A}\,(\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{A}\mathbf{x}_1 + \mathbf{A}\mathbf{x}_2 = f(\mathbf{x}_1) + f(\mathbf{x}_2)\,,$$

$$f(k\mathbf{x}_1) = \mathbf{A}\,(k\mathbf{x}_1) = k(\mathbf{A}\,\mathbf{x}_1) = kf(\mathbf{x}_1)\,.$$

We see that the map satisfies the linearity requirements L_1 and L_2 . Therefore every map of the form (12-22) is linear.

|||| Example 12.16 Matrix-Vector Product as a Linear Map

The formula:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + 2x_2 \\ 3x_1 + 4x_2 \\ 5x_1 + 6x_2 \end{bmatrix}$$

defines a particular linear map from the vector space $\mathbb{R}^2$ to the vector space $\mathbb{R}^3$.

But also the opposite is true: Every linear map between finite-dimensional vector spaces can be written as a matrix-vector product in the form (12-22) if we replace $\mathbf{x}$ and $\mathbf{y}$ with their coordinates with respect to a chosen basis for the domain and codomain, respectively. This we show in the following.

We consider a linear map $f : V \to W$ where V is an n-dimensional and W is an m-dimensional vector space, see Figure 12.5

For V a basis a is chosen and for W a basis c. This means that a given vector $\mathbf{x} \in V$ can be written as a unique linear combination of the a-basis vectors and that the image $\mathbf{y} = f(\mathbf{x})$ can be written as a unique linear combination of the c-basis vectors:

$$\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_n\mathbf{a}_n \text{ and } \mathbf{y} = y_1\mathbf{c}_1 + y_2\mathbf{c}_2 + \cdots + y_m\mathbf{c}_m\,.$$

This means that $(x_1, x_2, \ldots, x_n)$ is the set of coordinates for $\mathbf{x}$ with respect to the a-basis, and that $(y_1, y_2, \ldots, y_m)$ is the set of coordinates for $\mathbf{y}$ with respect to the c-basis.

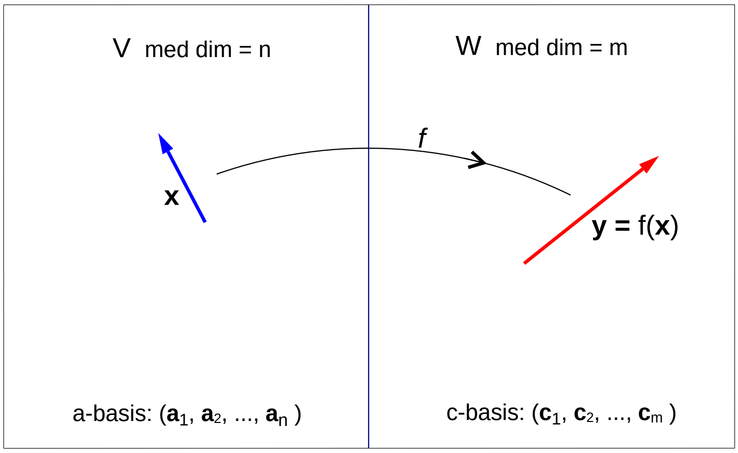


Figure 12.5: Linear map

We now pose the question: How can we describe the relation between the a-coordinate vector for the vector $\mathbf{x} \in V$ and the c-coordinate vector for the image vector $\mathbf{y}$? In other words we are looking for the relation between:

$$_{\mathrm{c}}\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \text{ and } _{\mathrm{a}}\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} .$$

This we develop through the following rewritings where we first, using L_1 and L_2, get $\mathbf{y}$ written as a linear combination of the images of the a-vectos.

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x}) \\ &= f(x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n) \\ &= x_1 f(\mathbf{a}_1) + x_2 f(\mathbf{a}_2) + \cdots + x_n f(\mathbf{a}_n) . \end{aligned}$$

Hereafter we can investigate the coordinate vector for $\mathbf{y}$ with respect to the c-basis, while we first use the cooordinate theorem, see Theorem 11.34, and thereafter the definition on matrix-vector product, see Definition 7.7.

$$\begin{aligned} {}_{\mathrm{c}}\mathbf{y} &= {}_{\mathrm{c}}\big(x_1 f(\mathbf{a}_1) + x_2 f(\mathbf{a}_2) + \cdots + x_n f(\mathbf{a}_n)\big) \\ &= x_1\, {}_{\mathrm{c}}f(\mathbf{a}_1) + x_2\, {}_{\mathrm{c}}f(\mathbf{a}_2) + \cdots + x_n\, {}_{\mathrm{c}}f(\mathbf{a}_n) \\ &= \big[\, {}_{\mathrm{c}}f(\mathbf{a}_1) \quad {}_{\mathrm{c}}f(\mathbf{a}_2) \quad \cdots \quad {}_{\mathrm{c}}f(\mathbf{a}_n) \,\big]\, {}_{\mathrm{a}}\mathbf{x} . \end{aligned}$$

The matrix $\left[\,{}_{\mathrm{c}}f(\mathbf{a}_1) \quad {}_{\mathrm{c}}f(\mathbf{a}_2) \quad \cdots \quad {}_{\mathrm{c}}f(\mathbf{a}_n)\,\right]$ in the last equation is called the *mapping matrix* for f with respect to the a-basis for V and the c-basis for W.

Thus we have achieved this important result: The coordinate vector ${}_{\mathrm{c}}\mathbf{y}$ can be found by multiplying the coordinate vector ${}_{\mathrm{a}}\mathbf{x}$ on the left by the mapping matrix. We now summarize the results in the following.

|||| Definition 12.17 Mapping Matrix

Let $f : V \rightarrow W$ be a linear map from an n-dimensional vector space V to an m-dimensional vector space W. By the ***mapping matrix*** for f with respect to the basis a of V and basis c of W we understand the $m \times n$-matrix:

$$
{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \left[\,{}_{\mathrm{c}}f(\mathbf{a}_1) \quad {}_{\mathrm{c}}f(\mathbf{a}_2) \quad \cdots \quad {}_{\mathrm{c}}f(\mathbf{a}_n)\,\right]. \tag{12-23}
$$

The mapping matrix for f thus consists of the coordinate vectors with respect to the basis c of the images of the n basis vectors in basis a.

The main task for a mapping matrix is of course to determine the images in W of the vectors in V, and this is justified in the following theorem which summarizes the investigations above.

|||| Theorem 12.18 Main Theorem of Mapping Matrices

Let V be an n-dimensional vector space with a chosen basis a and W an m-dimensional vector space with a chosen basis c .

1. For a linear map $f : V \rightarrow W$ it is valid that if $\mathbf{y} = f(\mathbf{x})$ is the image of an arbitrary vector $\mathbf{x} \in V$, then:

$$_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x} \tag{12-24}$$

   where ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$ is the mapping matrix for f with respect to the basis a of V and the basis c of W .

2. Conversely, assume that the images $\mathbf{y} = g(\mathbf{x})$ for a map $g : V \rightarrow W$ can be obtained in the coordinate form as

$$_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{G}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x} \tag{12-25}$$

   where ${}_{\mathrm{c}}\mathbf{G}_{\mathrm{a}} \in \mathbb{L}^{m \times n}$. Then g is linear and ${}_{\mathrm{c}}\mathbf{G}_{\mathrm{a}}$ is the mapping matrix for g with respect to the basis a of V and basis c of W .

Below are three examples of the construction and elementary use of mapping matrices.

|||| Example 12.19 Construction and Use of a Mapping Matrix

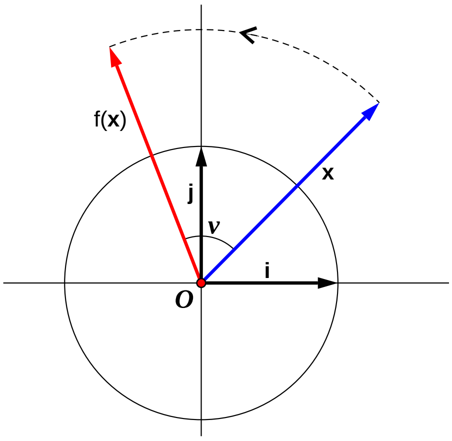


Figure: Linear rotation about the origin

Rotation of plane vectors drawn from the origin is a simple example of a linear map, see Exercise 12.4. Let v be an arbitrary angle, and let f be the linear mapping that rotates an arbitrary vector the angle v about the origin counterclockwise, (see the figure above).

We wish to determine the mapping matrix for f with respect to the standard basis for vectors in the plane. Therefore we need the images of the basis vectors $\mathbf{i}$ and $\mathbf{j}$:

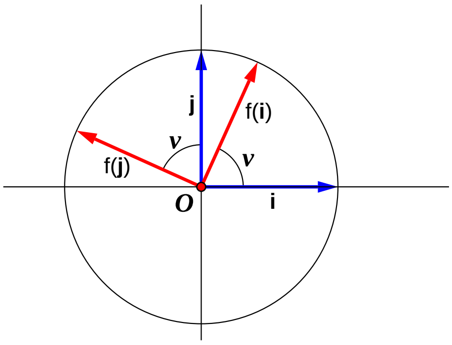


Figure: Determination of mapping matrix

It is seen that $f(\mathbf{i}) = (\cos(v), \sin(v))$ and $f(\mathbf{j}) = (-\sin(v), \cos(v))$. Therefore the mapping matrix we are looking for is

$$ {}_{\mathrm{e}}\mathbf{F}_{\mathrm{e}} = \begin{bmatrix} \cos(v) & -\sin(v) \\ \sin(v) & \cos(v) \end{bmatrix}. $$

The coordinates for the image $\mathbf{y} = f(\mathbf{x})$ of a given vector $\mathbf{x}$ are thus given by the formula:

$$ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos(v) & -\sin(v) \\ \sin(v) & \cos(v) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. $$

|||| Example 12.20 Construction and Use of a Mapping Matrix

In a 3-dimensional vector space V, a basis $a = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ is chosen, and in 2-dimensional vector space W a basis $c = (\mathbf{c}_1, \mathbf{c}_2)$ is chosen. A linear map $f : V \to W$ satisfies:

$$ f(\mathbf{a}_1) = 3\mathbf{c}_1 + \mathbf{c}_2 \,, \; f(\mathbf{a}_2) = 6\mathbf{c}_1 - 2\mathbf{c}_2 \; \text{ and } \; f(\mathbf{a}_3) = -3\mathbf{c}_1 + \mathbf{c}_2 \,. \tag{12-26} $$

We wish to find the image under f of the vector $\mathbf{v} = \mathbf{a}_1 + 2\mathbf{a}_2 + \mathbf{a}_3 \in V$ using the mapping matrix ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$. The mapping matrix is easily constructed since we already from (12-26) know the images of the basis vectors in V :

$$ {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} {}_{\mathrm{c}}f(\mathbf{a}_1) & {}_{\mathrm{c}}f(\mathbf{a}_2) & {}_{\mathrm{c}}f(\mathbf{a}_3) \end{bmatrix} = \begin{bmatrix} 3 & 6 & -3 \\ 1 & -2 & 1 \end{bmatrix} $$

Since $\mathbf{v}$ has the set of coordinates $(1, 2, 1)$ with respect to basis a, we find the coordinate vector for $f(\mathbf{v})$ like this:

$$_{\mathrm{c}}f(\mathbf{v}) = {_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}}\,{_{\mathrm{a}}\mathbf{v}} = \begin{bmatrix} 3 & 6 & -3 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 12 \\ -2 \end{bmatrix}.$$

Hence we have found $f(\mathbf{v}) = 12\mathbf{c}_1 - 2\mathbf{c}_2$.

|||| Example 12.21 Construction and Use of a Mapping Matrix

A linear map $f : \mathbb{R}^4 \to \mathbb{R}^3$ is given by:

$$f(x_1, x_2, x_3, x_4) = \begin{bmatrix} x_1 + 2x_2 + x_4 \\ 2x_1 - x_2 + 2x_3 - x_4 \\ x_1 - 3x_2 + 2x_3 - 2x_4 \end{bmatrix}. \tag{12-27}$$

Let us determine the mapping matrix for f with respect to the standard basis e of $\mathbb{R}^4$ and the standard basis e of $\mathbb{R}^3$. First we find the images of the four basis vectors in $\mathbb{R}^4$ using the rule (12-27):

$$f(1, 0, 0, 0) = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad f(0, 1, 0, 0) = \begin{bmatrix} 2 \\ -1 \\ -3 \end{bmatrix},$$

$$f(0, 0, 1, 0) = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}, \quad f(0, 0, 0, 1) = \begin{bmatrix} 1 \\ -1 \\ -2 \end{bmatrix}.$$

We can now construct the mapping matrix for f:

$$_{\mathrm{e}}\mathbf{F}_{\mathrm{e}} = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & -1 & 2 & -1 \\ 1 & -3 & 2 & -2 \end{bmatrix}. \tag{12-28}$$

We wish to find the image $\mathbf{y} = f(\mathbf{x})$ of the vector $\mathbf{x} = (1, 1, 1, 1)$. At our disposal we have of course the rule (12-27), but we choose to find the image using the mapping matrix:

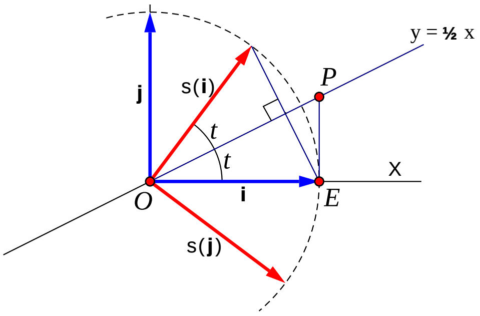
$$_{\mathrm{e}}\mathbf{y} = {_{\mathrm{e}}\mathbf{F}_{\mathrm{e}}}\,{_{\mathrm{e}}\mathbf{x}} = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & -1 & 2 & -1 \\ 1 & -3 & 2 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ -2 \end{bmatrix}.$$

Thus we have found that $y = f(1, 1, 1, 1) = (4, 2, -2)$.

|||| Exercise 12.22

In the plane is given a customary $(O, \mathbf{i}, \mathbf{j})$-coordinate system. Reflection of position vectors about the line $y = \frac{1}{2}x$ is a linear map, let us call it s .

Determine $s(\mathbf{i})$ and $s(\mathbf{j})$, construct the mapping matrix ${}_{\mathrm{e}}\mathbf{S}_{\mathrm{e}}$ for s and determine an expression k for the reflection of an arbitrary position vector $\mathbf{v}$ with the coordinates (v_1, v_2) with respect to the standard basis. The figure below contains some hints for the determination of $s(\mathbf{i})$. Proceed similarly with $s(\mathbf{j})$.



Reflection of the standard basis vectors.

12.6 On the Use of Mapping Matrices

The mapping matrix tool has a wide range of applications. It allows us to translate questions about linear maps between vector spaces to questions about matrices and coordinate vectors that allow immediate calculations. The methods only require that bases in each of the vector spaces be chosen, and that the mapping matrix that belongs to the two bases has been formed. In this way we can reduce problems as diverse as that of finding polynomials with certain properties, finding the result of a geometrical construction and finding the solution of differential equations, to problems that can be solved through the use of matrix algebra.

As a recurrent example in this section we look at a linear map $f : V \rightarrow W$ where V is a 4-dimensional vector space with chosen basis $a = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4)$, and where W is a

3-dimensional vector space with chosen basis $c = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)$. The mapping matrix for f is:

$$ {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 1 & 3 & -1 & 8 \\ 2 & 0 & 4 & -2 \\ 1 & -1 & 3 & -4 \end{bmatrix} . \tag{12-29} $$

12.6.1 Finding the Kernel of *f*

To obtain the kernel of f you must find all the $\mathbf{x} \in V$ that are mapped to $\mathbf{0} \in W$. That is you solve the vector equation

$$ f(\mathbf{x}) = \mathbf{0} . $$

This equation is according to the Theorem 12.18 equivalent to the matrix equation

$$ {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{x} = {}_{\mathrm{c}}\mathbf{0} $$

$$ \Leftrightarrow \begin{bmatrix} 1 & 3 & -1 & 8 \\ 2 & 0 & 4 & -2 \\ 1 & -1 & 3 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} $$

that corresponds to the homogeneous system of linear equations with the augmented matrix:

$$ \mathbf{T} = \left[\begin{array}{cccc|c} 1 & 3 & -1 & 8 & 0 \\ 2 & 0 & 4 & -2 & 0 \\ 1 & -1 & 3 & -4 & 0 \end{array}\right] \to \mathrm{rref}(\mathbf{T}) = \left[\begin{array}{cccc|c} 1 & 0 & 2 & -1 & 0 \\ 0 & 1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] . $$

It is seen that the solution set is spanned by two linear independent vectors: $(-2, 1, 1, 0)$ and $(1, -3, 0, 1)$. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be the two vectors in V that are determined by the a-coordinates like this:

$$ {}_{\mathrm{a}}\mathbf{v}_1 = (-2, 1, 1, 0) \text{ and } {}_{\mathrm{a}}\mathbf{v}_2 = (1, -3, 0, 1) . $$

Since the two coordinate vectors are linearly independent, $(\mathbf{v}_1, \mathbf{v}_2)$ is a basis for the kernel of f, and the kernel of f has the dimension 2.

Point: The number $n = 4$ of unknowns in the solved system of equations is by definition equal to the number of columns in ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$ that again is equal to $\dim(V)$, see definition 12.17. Moreover we notice that the coefficient matrix of the system of equations is equal to ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$. If the rank of the coefficient matrix is k , we know that the solution set, and therefore the kernel, will be spanned by $(n - k)$ linearly independent directional vectors where k is the rank of the coefficient matrix. Therefore we have:

$$ \dim(\ker(f)) = n - \rho\,({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}) = 4 - 2 = 2 . $$

|||| **Method 12.23 Determination of the Kernel**

In a vector space V a basis a is chosen, and in a vector space W a basis c is chosen. The kernel of a linear map $f : V \rightarrow W$, in coordinate form, can be found as the solution set for the homogeneous system of linear equations with the augmented matrix

$$\mathbf{T} = \left[\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} \;\middle|\; {}_{\mathrm{c}}\mathbf{0}\,\right].$$

The kernel is a subspace of V and its dimension is determined by:

$$\dim(\,\ker(f)\,) = \dim(V) - \rho\left({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\right). \tag{12-30}$$

12.6.2 Solving the Vector Equation *f(x) = b*

How can you decide whether a vector $\mathbf{b} \in W$ belongs to the image for a given linear map? The question is whether (at least) one $\mathbf{x} \in V$ exists that is mapped to $\mathbf{b}$. And the question can be extended to how to determine all $\mathbf{x} \in V$ with this property that is mapped in $\mathbf{b}$.

We consider the linear map $f : V \rightarrow W$ that is represented by the mapping matrix (12-29) and choose as our example the vector $\mathbf{b} \in W$ that has c-coordinates $(1, 2, 3)$. We will solve the vector equation

$$f(\mathbf{x}) = \mathbf{b}\,.$$

If we calculate with coordinates the vector equation corresponds to the following matrix equation

$${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x} = {}_{\mathrm{c}}\mathbf{b}$$

that is the matrix equation

$$\begin{bmatrix} 1 & 3 & -1 & 8 \\ 2 & 0 & 4 & -2 \\ 1 & -1 & 3 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

that corresponds to an inhomogeneous system of linear equations with the augmented matrix:

$$\mathbf{T} = \left[\begin{array}{rrrr|r} 1 & 3 & -1 & 8 & 1 \\ 2 & 0 & 4 & -2 & 2 \\ 1 & -1 & 3 & -4 & 3 \end{array}\right]$$

that by Gauss-Jordan elimination is reduced to

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{cccc|c} 1 & 0 & 2 & -1 & 0 \\ 0 & 1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Since the rank of the augmented matrix is larger than the rank of the coefficient matrix, the inhomogeneous system of equations has no solutions. We have thus found a vector in W that has no "original vector" in V.

|||| Method 12.24 Solution of the Vector Equation *f(x) = b*

In a vector space V a basis a is chosen, and in a vector space W a basis c is chosen. For a linear map $f : V \to W$, and a proper vector $\mathbf{b} \in W$, the equation

$$f(\mathbf{x}) = \mathbf{b}$$

can be solved using the inhomogeneous system of linear equations that has the augmented matrix

$$\mathbf{T} = \left[\, {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} \,\middle|\, {}_{\mathrm{c}}\mathbf{b} \,\right]$$

If solutions exist and $\mathbf{x}_0$ is one of these solutions the whole solution set can be written as:

$$\mathbf{x}_0 + \ker(f)\,.$$

An inhomogeneous system of linear equation consisting of m equations in n unknowns, with the coefficient matrix $\mathbf{A}$ and the right-hand side $\mathbf{b}$ can in matrix form be written as

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

The map $f : \mathbb{L}^n \to \mathbb{L}^m$ given by

$$f(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

is linear. The linear equation $f(\mathbf{x}) = \mathbf{b}$ is thus equivalent to the considered system of linear equations. Thus we can see that the structural theorem for systems of linear equations (see eNote 6 Theorem 6.37) is nothing but a particular case of the general structural theorem for linear maps (Theorem 12.14).

12.6.3 Determining the Image Space

Above we have found that the image space for a linear map is a subspace of the codomain, see theorem 12.11. How can this subspace be delimited and characterized?

Again we consider the linear map $f : V \to W$ that is represented by the mapping matrix (12-29). Since the basis $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4)$ for V is chosen we can write all the vectors in V at once:

$$\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 + x_4\mathbf{a}_4 \,,$$

where we imagine that x_1, x_2, x_3 og x_4 run through all conceivable combinations of real values. But then all images in W of vectors in V can be written as:

$$\begin{aligned} f(\mathbf{x}) &= f(x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3 + x_4\mathbf{a}_4) \\ &= x_1 f(\mathbf{a}_1) + x_2 f(\mathbf{a}_2) + x_3 f(\mathbf{a}_3) + x_4 f(\mathbf{a}_4) \,, \end{aligned}$$

where we have used L_1 og L_2, and where we continue to imagine that x_1, x_2, x_3 and x_4 run through all conceivable combinations of real values. But then:

$$f(V) = \text{span}\,\{\, f(\mathbf{a}_1), f(\mathbf{a}_2), f(\mathbf{a}_3), f(\mathbf{a}_4) \,\} \;.$$

The image space is thus spanned by the images of the a-basis vectors! But then we can (according to Method 11.47Method in eNote 11) determine a basis for the image space by finding the leading 1's in the reduced row echelon form of

$$\begin{bmatrix} {}_{c}\mathbf{f}(\mathbf{a}_1) & {}_{c}\mathbf{f}(\mathbf{a}_2) & {}_{c}\mathbf{f}(\mathbf{a}_3) & {}_{c}\mathbf{f}(\mathbf{a}_4) \end{bmatrix} .$$

This is the mapping matrix for f with respect to the chosen bases

$${}_{\text{c}}\mathbf{F}_{\text{a}} = \begin{bmatrix} 1 & 3 & -1 & 8 \\ 2 & 0 & 4 & -2 \\ 1 & -1 & 3 & -4 \end{bmatrix}$$

that by Gauss-Jordan elimination is reduced to

$$\text{rref}({}_{\text{c}}\mathbf{F}_{\text{a}}) = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} .$$

To the two leading 1's in $\text{rref}({}_{\text{c}}\mathbf{F}_{\text{a}})$ correspond the first two columns in ${}_{\text{c}}\mathbf{F}_{\text{a}}$. We thus conclude:

Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be the two vectors in W determined by c-coordinates as:

$${}_{\text{c}}\mathbf{w}_1 = (1, 2, 1) \;\text{ and }\; {}_{\text{c}}\mathbf{w}_2 = (3, 0, -1) \,.$$

Then $(\mathbf{w}_1, \mathbf{w}_2)$ is a basis for the image space $f(V)$.

|||| **Method 12.25 Determination of the Image Space**

In a vector space V a basis a is chosen, and in a vector space W a basis c is chosen. The image space $f(V)$ for a linear mapping $f : V \rightarrow W$ can be found from

$$\operatorname{rref}({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}) \tag{12-31}$$

in the following way: If there is no leading 1 in the i'th column in (12-31) then $f(\mathbf{a}_i)$ is removed from the vector set $\big(f(\mathbf{a}_1), f(\mathbf{a}_2), \ldots, f(\mathbf{a}_n)\big)$. After this thinning the vector set constitutes a basis for $f(V)$.

Since the number of leading 1's in (12-31) is equal to the number of basis vectors in the chosen basis for $f(V)$ it follows that

$$\dim(f(V)) = \rho\,({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}})\,. \tag{12-32}$$

12.7 The Dimension Theorem

In the method of the preceding section 12.23 we found the following expression for the dimension of the kernel of a linear map $f : V \rightarrow W$:

$$\dim(\ker(f)) = \dim(V) - \rho\,({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}})\,. \tag{12-33}$$

And in method 12.25 a corresponding expression for the image space $f(V)$:

$$\dim(f(V)) = \rho\,({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}})\,. \tag{12-34}$$

By combining (12-33) and (12-34) a remarkably simple relationship between the domain, the kernel and the image space for a linear map is achieved:

|||| **Theorem 12.26 The Dimension Theorem (or Rank-Nullity Theorem)**

Let V and W be two finite dimensional vector spaces. For a linear map $f : V \rightarrow W$ we have:

$$\dim(V) = \dim(\ker(f)) + \dim(f(V))\,.$$

Here are some direct consequences of Theorem 12.26:

The image space for a linear map can never have a higher dimension than the domain.

If the kernel only consists of the **0**-vector, the image space *keeps* the dimension of the domain.

If the kernel has the dimension $p > 0$, then p dimensions *disappear* through the map.

Exercise 12.27

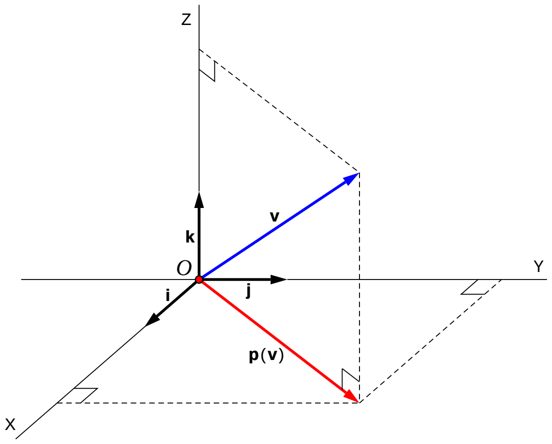
A linear map $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ has, with respect to the standard basis for $\mathbb{R}^3$, the mapping matrix

$$_{\mathrm{e}}\mathbf{F}_{\mathrm{e}} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 0 \\ 3 & 6 & 0 \end{bmatrix}.$$

It is stated that the kernel of f has the dimension 1. Find by mental calculation, a basis for $f(V)$.

|||| Exercise 12.28

In 3-space a standard $(O, \mathbf{i}, \mathbf{j}, \mathbf{k})$-coordinate system is given. The map p projects position vectors down into (x, y)-plane in space:



Projection down into the (X, Y)-plane

Show that p is linear and construct the mapping matrix ${}_{\mathrm{e}}\mathbf{P}_{\mathrm{e}}$ for p. Determine a basis for the kernel and the image space of the projection. Check that the Dimension Theorem is fulfilled.

12.8 Change in the Mapping Matrix when the Basis is Changed

In eNote 11 it is shown how the coordinates of a vector change when the basis for the vector space is changed, see method 11.40. We begin this section by repeating the most important points and showing two examples.

Assume that in V an a-basis $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ is given, and that a new b-basis $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)$ is chosen in V. If a vector $\mathbf{x}$ has the b-coordinate vector ${}_{\mathrm{b}}\mathbf{x}$, then its a-coordinate vector can be calculated as the matrix vector-product

$$ {}_{\mathrm{a}}\mathbf{v} = {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\, {}_{\mathrm{b}}\mathbf{v} \tag{12-35} $$

where the *change of basis matrix* ${}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}$ is given by

$$ {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \left[\, {}_{\mathrm{a}}\mathbf{b}_1 \quad {}_{\mathrm{a}}\mathbf{b}_2 \quad \ldots \quad {}_{\mathrm{a}}\mathbf{b}_n \,\right]. \tag{12-36} $$

We now show two examples of the use of (12-36). In the first example the "new" coordinates are given following which the "old" are calculated. In the second example it is vice versa: the "old" are known, and the "new" are determined.

|||| Example 12.29 From New Coordinates to Old

In a 3-dimensional vector space V a basis $a = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ is given, following which a new basis b is chosen consisting of the vectors

$$\mathbf{b}_1 = \mathbf{a}_1 - \mathbf{a}_3,\ \mathbf{b}_2 = 2\mathbf{a}_1 - 2\mathbf{a}_2 + \mathbf{a}_3 \ \text{ and } \ \mathbf{b}_3 = -3\mathbf{a}_1 + 3\mathbf{a}_2 - \mathbf{a}_3\,.$$

Problem: Determine the coordinate vector ${}_{\mathrm{a}}\mathbf{x}$ for $\mathbf{x} = \mathbf{b}_1 + 2\mathbf{b}_2 + 3\mathbf{b}_3\,.$

Solution: First we see that

$${}_{\mathrm{b}}\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \ \text{ and } \ {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \begin{bmatrix} 1 & 2 & -3 \\ 0 & -2 & 3 \\ -1 & 1 & -1 \end{bmatrix}. \tag{12-37}$$

Then we get

$${}_{\mathrm{a}}\mathbf{x} = {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\, {}_{\mathrm{b}}\mathbf{x} = \begin{bmatrix} 1 & 2 & -3 \\ 0 & -2 & 3 \\ -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -4 \\ 5 \\ -2 \end{bmatrix}. \tag{12-38}$$

|||| Example 12.30 From Old Coordinates to New

In a 2-dimensional vector space W a basis $c = (\mathbf{c}_1, \mathbf{c}_2)$ is given, following which a new basis d is chosen consisting of the vectors

$$\mathbf{d}_1 = 2\mathbf{c}_1 + \mathbf{c}_2 \ \text{ and } \ \mathbf{d}_2 = \mathbf{c}_1 + \mathbf{c}_2\,.$$

Problem: Determine the coordinate vector ${}_{\mathrm{d}}\mathbf{y}$ for $\mathbf{y} = 10\mathbf{c}_1 + 6\mathbf{c}_2\,.$

Solution: First we see that

$${}_{\mathrm{c}}\mathbf{y} = \begin{bmatrix} 10 \\ 6 \end{bmatrix} \ \text{ and } \ {}_{\mathrm{c}}\mathbf{M}_{\mathrm{d}} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}} = ({}_{\mathrm{c}}\mathbf{M}_{\mathrm{d}})^{-1} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}. \tag{12-39}$$

Then we get

$${}_{\mathrm{d}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\, {}_{\mathrm{c}}\mathbf{y} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 10 \\ 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}. \tag{12-40}$$

We now continue to consider how a mapping matrix is changed when the basis for the domain or the codomain is changed.

For two vector spaces V and W with finite dimension the mapping matrix for a linear map $f : V \rightarrow W$ can only be constructed when a basis for V and a basis for W are chosen. By using the mapping matrix symbol ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$ we show the foundation to be the pair of given bases a of V and c of W.

Often one wishes to change the basis of V or the basis of W. In the *first* case the coordinates for those vectors $\mathbf{x} \in V$ will change while the coordinates for their images $\mathbf{y} = f(\mathbf{x})$ are unchanged; in the *second* case it is the other way round with the $\mathbf{x}$ coordinates remaining unchanged while the image coordinates change. If the bases of both $V$and W are changed then the coordinates for both $\mathbf{x}$ and $\mathbf{y} = f(\mathbf{x})$ change.

In this section we construct methods for finding the new mapping matrix for f, when we change the basis for either the domain, the codomain or both. First we show how a vector's coordinates change when the basis for the domain is changed (as in detail in Method 11.40 in eNote 11.)

12.8.1 Change of Basis in the Domain

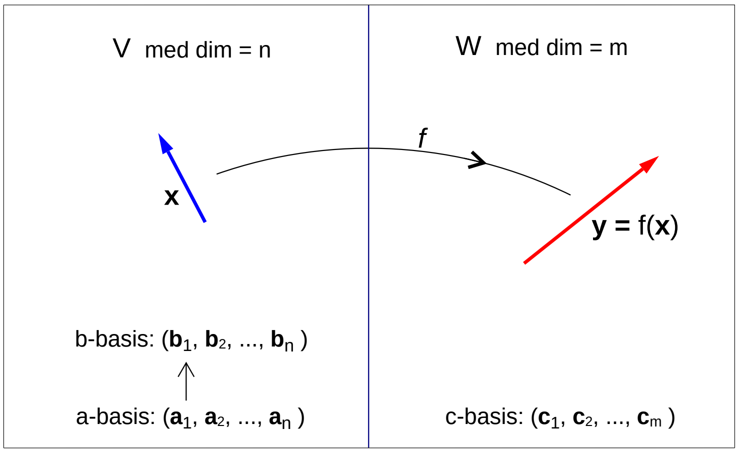

V med dim = n
W med dim = m
f
x
y = f(x)
b-basis: (b1, b2, ..., bn)
a-basis: (a1, a2, ..., an)
c-basis: (c1, c2, ..., cm)

Figure 12.6: Linear map

In Figure 12.6 a linear map $f : V \to W$ is given that, with respect to basis a of V and basis c of W, has the mapping matrix ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$. We change the basis for V from basis a to basis b. The mapping matrix for f now has the symbol ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{b}}$. Let us find it. The equation

$$\mathbf{y} = f(\mathbf{x})$$

is translated into coordinates and rewritten as:

$${}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\left({}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{x}\right) = \left({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\right){}_{\mathrm{b}}\mathbf{x}\,.$$

From this we deduce that the mapping matrix for f with respect to the basis b of V and basis c of W is formed by a matrix product:

$${}_{\mathrm{c}}\mathbf{F}_{\mathrm{b}} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,. \tag{12-41}$$

|||| Example 12.31 Change of a Mapping Matrix

We consider the 3-dimensional vector space V that is treated in Example 12.29 and the 2-dimensional vector space W that is treated in Example 12.30. A linear map $f : V \to W$ is given by the mapping matrix:

$${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix}.$$

Problem: Determine $\mathbf{y} = f(\mathbf{x})$ where $\mathbf{x} = \mathbf{b}_1 + 2\mathbf{b}_2 + 3\mathbf{b}_3$.

Solution: We try two different ways. 1) We use a-coordinates for $\mathbf{x}$ as found in (12-37):

$${}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix}\begin{bmatrix} -4 \\ 5 \\ -2 \end{bmatrix} = \begin{bmatrix} 10 \\ 6 \end{bmatrix}.$$

2) We change the mapping matrix for f :

$${}_{\mathrm{c}}\mathbf{F}_{\mathrm{b}} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix}\begin{bmatrix} 1 & 2 & -3 \\ 0 & -2 & 3 \\ -1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}.$$

Then we can directly use the given b-coordinates for $\mathbf{x}$:

$${}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{x} = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 6 \end{bmatrix}.$$

In either case we get $\mathbf{y} = 10\mathbf{c}_1 + 6\mathbf{c}_2$.

12.8.2 Change of Basis in the Codomain

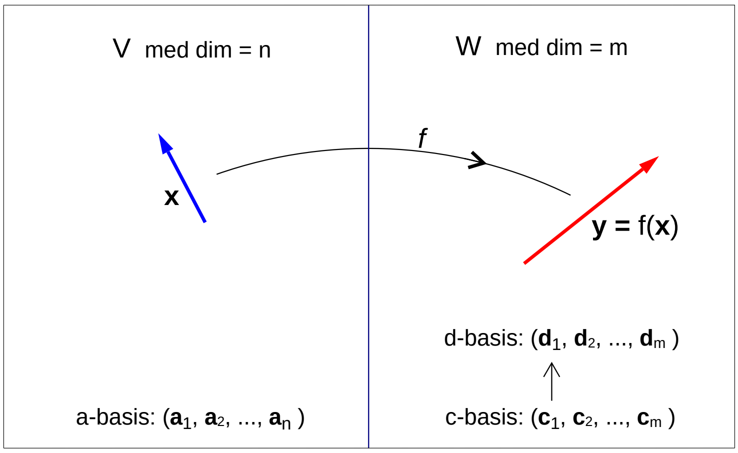

V med dim = n
W med dim = m
f
x
y = f(x)
d-basis: (d1, d2, ..., dm)
a-basis: (a1, a2, ..., an)
c-basis: (c1, c2, ..., cm)

Figure 12.7: Linear map

In Figure 12.7 a linear map $f : V \to W$ is given that, with respect to the basis a of V and basis c of W has a mapping matrix ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$. We change the basis for W from basis c to basis d . The mapping matrix for f now has the symbol ${}_{\mathrm{d}}\mathbf{F}_{\mathrm{a}}$. Let us find it. The equation

$$\mathbf{y} = f(\mathbf{x})$$

is translated into the matrix equation

$${}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x}$$

that is equivalent to

$${}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\left({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x}\right)$$

from which we get that

$${}_{\mathrm{d}}\mathbf{y} = \left({}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\right){}_{\mathrm{a}}\mathbf{x}\,.$$

From this we deduce that the mapping matrix for f with respect to the a-basis for V and the d-basis for W is formed by a matrix product:

$${}_{\mathrm{d}}\mathbf{F}_{\mathrm{a}} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,. \tag{12-42}$$

|||| Example 12.32 Change of Mapping Matrix

We consider the 3-dimensional vector space V that is treated in Example 12.29 and the 2-dimensional vector space W that is treated in Example 12.30. A linear map $f : V \to W$ is given by the mapping matrix:

$$ {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix}. $$

Problem: Given the vector $\mathbf{x} = -4\mathbf{a}_1 + 5\mathbf{a}_2 - 2\mathbf{a}_3$. Determine the image $\mathbf{y} = f(\mathbf{x})$ as a linear combination of $\mathbf{d}_1$ and $\mathbf{d}_2$.

Solution: We try two different ways.
1) We use the given mapping matrix:

$$ {}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{x} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix} \begin{bmatrix} -4 \\ 5 \\ -2 \end{bmatrix} = \begin{bmatrix} 10 \\ 6 \end{bmatrix}. $$

And translate the result to d-coordinates using (12-40):

$$ {}_{\mathrm{d}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\, {}_{\mathrm{c}}\mathbf{y} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 10 \\ 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} $$

2) We change the mapping matrix for f using (12-39):

$$ {}_{\mathrm{d}}\mathbf{F}_{\mathrm{a}} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 2 \\ 3 & 4 & 3 \end{bmatrix}. $$

Then we can directly read the d-coordinates:

$$ {}_{\mathrm{d}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{x} = \begin{bmatrix} 3 & 4 & 2 \\ 3 & 4 & 3 \end{bmatrix} \begin{bmatrix} -4 \\ 5 \\ -2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}. $$

In both cases we get $\mathbf{y} = 4\mathbf{d}_1 + 2\mathbf{d}_2$.

12.8.3 Change of Basis in both the Domain and Codomain

In Figure 12.8 a linear map $f : V \to W$ is given that, with respect to the basis a for V and basis c for W, has the mapping matrix ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$. We change the basis for V from basis a to basis b, and for W from basis c to basis d. The mapping matrix for f now has the

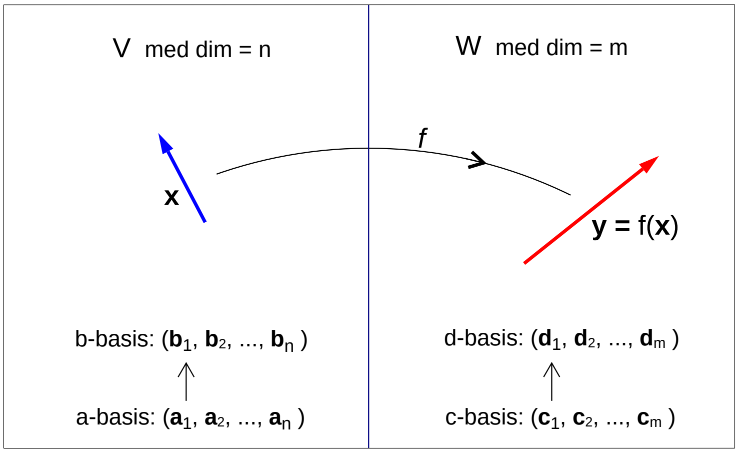

V med dim = n
W med dim = m
f
x
y = f(x)
b-basis: (b1, b2, ..., bn)
a-basis: (a1, a2, ..., an)
d-basis: (d1, d2, ..., dm)
c-basis: (c1, c2, ..., cm)

Figure 12.8: Linear map

symbol ${}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}}$. Let us find it. The equation

$$y = f(\mathbf{x})$$

corresponds in coordinates to

$${}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{x}$$

that is equivalent to

$${}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\left({}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\left({}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{x}\right)\right)$$

from which we obtain

$${}_{\mathrm{d}}\mathbf{y} = \left({}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\right){}_{\mathrm{b}}\mathbf{x}\,.$$

From here we deduce that the mapping matrix for f with respect to b-basis of V and d-basis of W is formed by a matrix product:

$${}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,. \tag{12-43}$$

|||| Example 12.33 Change of Mapping Matrix

We consider the 3-dimensional vector space V that is treated in example 12.29, and the 2-dimensional vector space W that is treated in example 12.30. A linear map $f : V \rightarrow W$ is given by the mapping matrix:

$${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix}.$$

Problem: Given the vector $\mathbf{x} = \mathbf{b}_1 + 2\mathbf{b}_2 + 3\mathbf{b}_3$. Determine $\mathbf{y} = f(\mathbf{x})$ as a linear combination of $\mathbf{d}_1$ and $\mathbf{d}_2$.

Solution: We change the mapping matrix using (12-39) and (12-37):

$$
{}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}} = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\,{}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,{}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 9 & 12 & 7 \\ 6 & 8 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & -3 \\ 0 & -2 & 3 \\ -1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} .
$$

Then we can directly use the given b-coordinates and directly read the d-coordinates:

$$
{}_{\mathrm{d}}\mathbf{y} = {}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}}\,{}_{\mathrm{b}}\mathbf{x} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} .
$$

Conclusion: $\mathbf{y} = 4\mathbf{d}_1 + 2\mathbf{d}_2$.

The change of basis in this example turns out to be rather practical. With the new mapping matrix ${}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}}$ it is much easier to calculate the image vector: You just add the first and the third coordinates of the given vector and keep the second coordinate!

12.8.4 Summary Concerning Change of Basis

We gather the results concerning change of basis in the subsections above in the following method:

|||| Method 12.34 Change of Mapping Matrix when Changing the Basis

For the vector space V are given a basis $a = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ and a new basis $b = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)$. For the vector space W are given a basis $c = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m)$ and a new basis $d = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_m)$.

If f is a linear map $f : V \to W$ that, with respect to basis a of V and basis c of W, has the mapping matrix ${}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}$, then:

1. The mapping matrix for f with respect to basis b of $V$and basis c of W is

$$ {}_{\mathrm{c}}\mathbf{F}_{\mathrm{b}} = {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,. \tag{12-44}$$

2. The mapping matrix for f with respect to basis a of V and basis d of W is

$$ {}_{\mathrm{d}}\mathbf{F}_{\mathrm{a}} = \left({}_{\mathrm{c}}\mathbf{M}_{\mathrm{d}}\right)^{-1} {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}} \; = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\, {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\,. \tag{12-45}$$

3. The mapping matrix for f with respect to basis b of V and basis d of W is

$$ {}_{\mathrm{d}}\mathbf{F}_{\mathrm{b}} = \left({}_{\mathrm{c}}\mathbf{M}_{\mathrm{d}}\right)^{-1} {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} \; = {}_{\mathrm{d}}\mathbf{M}_{\mathrm{c}}\, {}_{\mathrm{c}}\mathbf{F}_{\mathrm{a}}\, {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}}\,. \tag{12-46}$$

In the three formulas we have used the change of basis matrices:

$$ {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \begin{bmatrix} {}_{\mathrm{a}}\mathbf{b}_1 & {}_{\mathrm{a}}\mathbf{b}_2 & \cdots & {}_{\mathrm{a}}\mathbf{b}_n \end{bmatrix} \text{ and } {}_{\mathrm{c}}\mathbf{M}_{\mathrm{d}} = \begin{bmatrix} {}_{\mathrm{c}}\mathbf{d}_1 & {}_{\mathrm{c}}\mathbf{d}_2 & \cdots & {}_{\mathrm{c}}\mathbf{d}_m \end{bmatrix}. $$

eNote 13

Eigenvalues and Eigenvectors

This note introduces the concepts of eigenvalues and eigenvectors for linear maps in arbitrary general vector spaces and then delves deeply into eigenvalues and eigenvectors of square matrices. Therefore the note is based on knowledge about general vector spaces, see eNote 11, on knowledge about algebra with matrices, see eNote 7 and eNote 8, and on knowledge about linear maps see eNote 12.

Update: 7.10.21 David Brander.

13.1 The Eigenvalue Problem for Linear Maps

13.1.1 Introduction

In this eNote we consider linear maps of the type

$$f : V \to V\,, \tag{13-1}$$

that is, linear maps where the *domain* and the *codomain* are the same vector space. This gives rise to a special phenomenon, that a vector can be equal to its image vector:

$$f(\mathbf{v}) = \mathbf{v}\,. \tag{13-2}$$

Vectors of this type are called *fixed points* of the map f. More generally we are looking for ***eigenvectors***, that is vectors that are proportional to their image vectors. In this connection one talks about the *eigenvalue problem*: to find a scalar λ and a proper (i.e. non-zero) vector $\mathbf{v}$ satisfying the vector equation:

$$f(\mathbf{v}) = \lambda\mathbf{v}\,. \tag{13-3}$$

If λ is a scalar and $\mathbf{v}$ a proper vector satisfying 13-3 the proportionality factor λ is called an ***eigenvalue*** of f and $\mathbf{v}$ an *eigenvector* corresponding to λ. Let us, for example, take a linear map $f : G_3 \to G_3$, that is, a linear map of the set of space vectors into itself, mapping three given vectors as shown in Figure 13.1.

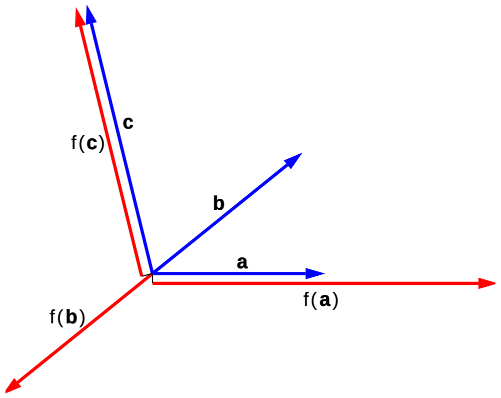


Figure 13.1: Three eigenvectors in space and their image vectors.

As hinted in Figure 13.1 $f(\mathbf{a}) = 2\mathbf{a}$. Therefore 2 is an eigenvalue of f with corresponding eigenvector $\mathbf{a}$. Furthermore $f(\mathbf{b}) = -\mathbf{b}$, so -1 is also an eigenvalue of f with corresponding eigenvector $\mathbf{b}$. And since finally $f(\mathbf{c}) = \mathbf{c}$, 1 is an eigenvalue of f with corresponding eigenvector $\mathbf{c}$. More specifically $\mathbf{c}$ is a fixed point for f.

To solve eigenvalue problems for linear maps is one of the most critical problems in engineering applications of linear algebra. This is closely connected to the fact that a linear map whose mapping matrix with respect to a given basis is a *diagonal matrix* is particularly simple to comprehend and work with. And here the nice rule, that if one chooses a basis consisting of eigenvectors for the map, then the mapping matrix automatically becomes a diagonal matrix.

In the following example we illustrate these points using linear maps in the plane.

|||| Example 13.1 Eigenvalues and Eigenvectors in the Plane

The vector space of vectors in the plane has the symbol $G_2(\mathbb{R})$. We consider a linear map

$$f : G_2(\mathbb{R}) \to G_2(\mathbb{R}) \tag{13-4}$$

of the set of plane vectors into itself, that with respect to a given basis $(\mathbf{a}_1, \mathbf{a}_2)$ has the following diagonal matrix as its mapping matrix:

$$_{\mathrm{a}}\mathbf{F}_{\mathrm{a}} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}. \tag{13-5}$$

Since

$$_{\mathrm{a}}f(\mathbf{a}_1) = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$_{\mathrm{a}}f(\mathbf{a}_2) = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} = 3 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

we have that $f(\mathbf{a}_1) = 2\mathbf{a}_1$ and $f(\mathbf{a}_2) = 3\mathbf{a}_2$. Both basis vectors are thus eigenvectors for f , because $\mathbf{a}_1$ corresponds to the eigenvalue 2 and $\mathbf{a}_2$ corresponds to the eigenvalue 3 . The eigenvalues are the diagonal elements in $_{\mathrm{a}}\mathbf{F}_{\mathrm{a}}$.

We now consider an arbitrary vector $\mathbf{x} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2$ and find its image vector:

$$_{\mathrm{a}}f(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 3x_2 \end{bmatrix}.$$

By the map the x_1-coordinate is multiplied by the eigenvalue 2, while the x_2-coordinate is multiplied by the eigenvalue 3. Geometrically this means that through the map all of the plane "is stretched" first by the factor 2 in the direction $\mathbf{a}_1$ and then by the factor 3 in the direction $\mathbf{a}_2$, see the effect on an arbitrarily chosen vector $\mathbf{x}$ in the figure A:

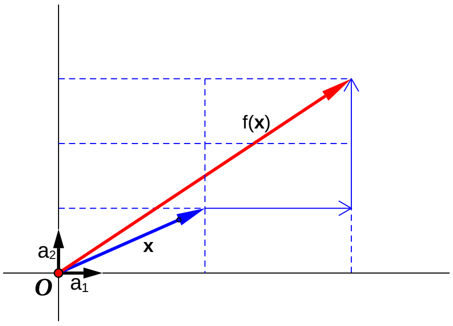


Figure A: The vector $\mathbf{x}$ is stretched horizontally by a factor 2 and vertically by a factor 3.

In Figure B we have chosen the standard basis $(\mathbf{i},\mathbf{j})$ and illustrate how the linear map g that has the mapping matrix

$$_{\mathrm{e}}\mathbf{G}_{\mathrm{e}} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix},$$

maps the "blue house" into the "red house" by stretching all position vectors in the blue house by the factor 2 in the horizontal direction and by the factor 3 in the vertical direction.

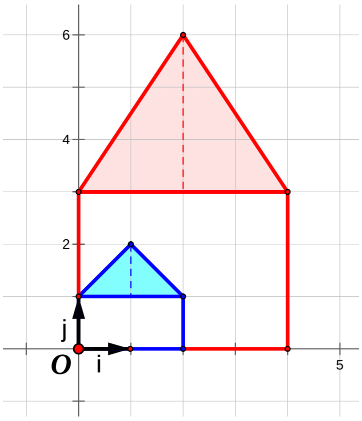


Figure B: The blue house is stretched in the horizontal direction by the factor 2 and vertically by the factor 3.

We now investigate another map h, the mapping matrix of which, with respect to the standard basis, is not a diagonal matrix:

$$_{\mathrm{e}}\mathbf{H}_{\mathrm{e}} = \begin{bmatrix} 7/3 & 2/3 \\ 1/3 & 8/3 \end{bmatrix}.$$

Here it is not possible to decide directly whether the map is composed of two stretchings in two given directions. And the mapping of the blue house by h as shown in the figure below does not give a clue directly:

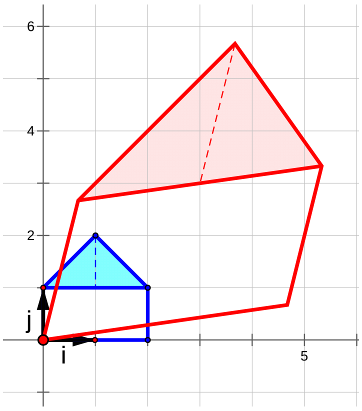


Figure C: House

But it is actually also possible in the case of h to choose a basis consisting of two linearly independent eigenvectors for h. Let $\mathbf{b}_1$ be given by the e-coordinates $(2,-1)$ and $\mathbf{b}_2$ by the e-coordinates $(1,1)$. Then we find that

$$_{\mathrm{e}}h(\mathbf{b}_1) = \begin{bmatrix} 7/3 & 2/3 \\ 1/3 & 8/3 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \end{bmatrix} = 2\cdot\begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

and

$$_{\mathrm{e}}h(\mathbf{b}_2) = \begin{bmatrix} 7/3 & 2/3 \\ 1/3 & 8/3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3\cdot\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

In other words, $h(\mathbf{b}_1) = 2\mathbf{b}_1$ and $h(\mathbf{b}_2) = 3\mathbf{b}_2$. We see that $\mathbf{b}_1$ and $\mathbf{b}_2$ are eigenvectors for h, and when we choose $(\mathbf{b}_1, \mathbf{b}_2)$ as basis, the mapping matrix for h with respect to this basis takes the form:

$$_{\mathrm{b}}\mathbf{G}_{\mathrm{b}} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}.$$

Surprisingly it thus shows that the mapping matrix for h also can be written in the form (13-5). The map h is also composed of two stretchings with the factors 2 and 3. Only the stretching *directions* are now determined by the eigenvectors $\mathbf{b}_1$ and $\mathbf{b}_2$. This is more evident if we map a new blue house whose principal lines are parallel to the b-basis vectors:

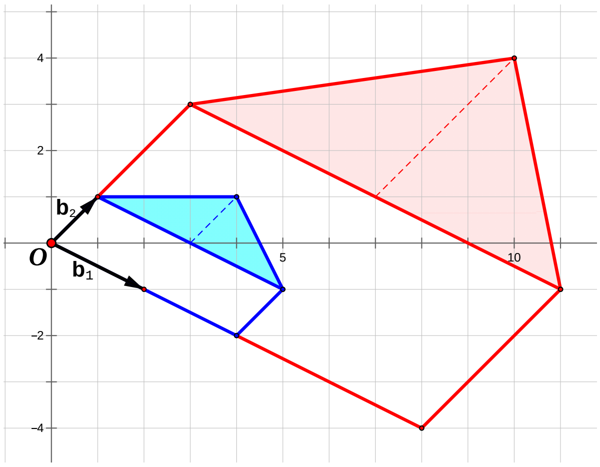


Figure D: The blue house is stretched by the factor 2 and the factor 3, respectively, in the directions of the eigenvectors

Thus we have illustrated: If you can find two linearly independent eigenvectors for a linear map in the plane it is possible:

1. to write its mapping matrix in diagonal form by choosing the eigenvectors as basis
2. to describe the map as stretchings in the directions of the eigenvectors with the corresponding eigenvectors as stretching factors.

13.1.2 Eigenvalues and their Corresponding Eigenvectors

The *eigenvalue problem* for a linear map is briefly about answering the question: do any proper vectors, each with its image vector proportional to the vector itself, exist. The short answer to this is that this cannot be answered in general, it depends on the particular map. In the following we try to pinpoint what can actually be said generally about the eigenvalue problem.

Definition 13.2 Eigenvalue and Eigenvector

Let $f : V \to V$ be a linear map of the vector space V into itself. If a proper vector $\mathbf{v} \in V$ and a scalar λ exist such that

$$f(\mathbf{v}) = \lambda\mathbf{v}\,, \tag{13-6}$$

then the proportionality factor λ is called an ***eigenvalue*** of f, while $\mathbf{v}$ is called an ***eigenvector*** corresponding to λ.

If, in Definition 13.2, it were not required to find a *proper* vector that satisfies $f(\mathbf{v}) = \lambda\mathbf{v}$, then every scalar λ would be an eigenvalue, since for any scalar λ $f(\mathbf{0}) = \lambda\,\mathbf{0}$ is valid. On the other hand, for a given eigenvalue, it is a matter of convention whether or not to say that the zero vector is also a corresponding eigenvector. Most commonly, the zero vector is not considered to be an eigenvector.

The number 0 can be an eigenvalue. This is so if a proper vector $\mathbf{v}$ exists such that $f(\mathbf{v}) = \mathbf{0}$, since we then have $f(\mathbf{v}) = 0\mathbf{v}$.

If a linear map f has one eigenvector $\mathbf{v}$, then it has infinitely many eigenvectors. This is a simple consequence of the following theorem.

Theorem 13.3 Eigenspace

If λ is an eigenvalue of a linear map $f : V \to V$, denote by E_λ the set: $E_\lambda := \{\mathbf{v} \in V \mid f(\mathbf{v}) = \lambda\mathbf{v})\}$. Then E_λ is a vector subspace of V.

|||| Proof

Let $f : V \to V$ be a linear map of the vector space V into itself, and assume that λ is an eigenvalue of f. Obviously E_λ is not empty, since it contains the zero vector. We shall show that the it satisfies the two stability requirements for subspaces, see Theorem 11.42. Let k be an arbitrary scalar, and let $\mathbf{u}$ and $\mathbf{v}$ be two arbitrary elements of E_λ. Then the following is valid :

$$f(\mathbf{u}+\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}) = \lambda\mathbf{u} + \lambda\mathbf{v} = \lambda(\mathbf{u}+\mathbf{v})\,.$$

Thus the vector sum $\mathbf{u}+\mathbf{v} \in E_\lambda$ and thus we have shown that E_λ satisfies the stability requirement with respect to addition. Furthermore the following is valid:

$$f(k\mathbf{u}) = kf(\mathbf{u}) = k(\lambda\mathbf{u}) = \lambda(k\mathbf{u})\,.$$

Thus we have shown stabilit with respect to multiplication by a scalar. Together we have shown that E_λ is a subspace of the domain.

■

Theorem 13.3 yields the following definition:

|||| Definition 13.4 Eigenvector Space

Let $f : V \to V$ be a linear map of the vector space V to itself, and let λ be an eigenvalue of f.

By the *eigenvector space* (or in short the *eigenspace*) E_λ corresponding to λ we understand the subspace:

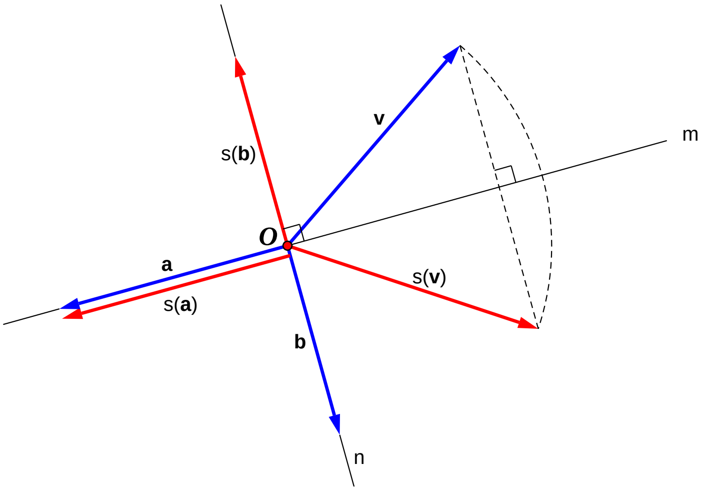
$$E_\lambda = \{\mathbf{v} \in V \mid f(\mathbf{v}) = \lambda\mathbf{v}\}\;.$$

If E_λ is finite-dimensional, $\dim(E_\lambda)$ is called the *geometric multiplicity* of λ, denoted $\mathrm{gm}(\lambda)$.

In the following example we consider a linear map that has two eigenvalues, both with the geometric multiplicity 1.

|||| Example 13.5 Eigenspace for Reflection

In the plane a straight line through the origin is drawn. By s we denote the linear map that maps a vector $\mathbf{v}$, drawn from the origin, in its reflection $s(\mathbf{v})$ in m:



The eigenvalue problem for the reflection in m.

Let $\mathbf{a}$ be an arbitrary proper vector that lies on m. Since

$$s(\mathbf{a}) = \mathbf{a} = 1 \cdot \mathbf{a}$$

1 is an eigenvalue of s. The eigenspace E_1 is the set of vectors that lie on m.

We now draw a straight line n through the origin, perpendicular to m. Let $\mathbf{b}$ be an arbitrary proper vector lying on n. Since

$$s(\mathbf{b}) = -\mathbf{b} = (-1) \cdot \mathbf{b}\,,$$

-1 is an eigenvalue of s. The eigenspace E_{-1} is the set of vectors that lie on n.

That not all linear maps have eigenvalues and thus eigenvectors is evident from the following example.

|||| Example 13.6

Let us investigate the eigenvalue problem for the linear map $f : G_2 \to G_2$ that to every proper vector $\mathbf{v}$ in the plane assigns its hat vector:

$$f(\mathbf{v}) = \widehat{\mathbf{v}}\,.$$

Since a proper vector $\mathbf{v}$ never can be proportional (parallel) to its hat vector, then for any scalar λ we have

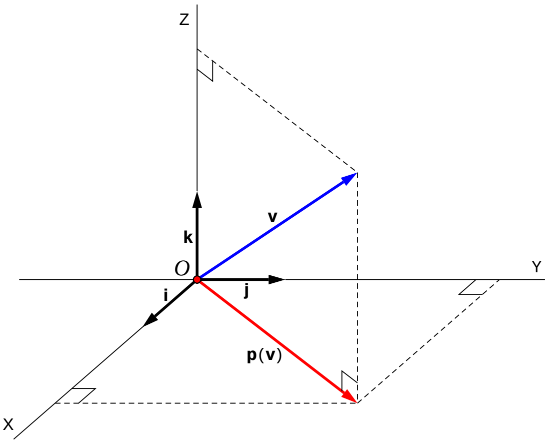
$$\widehat{\mathbf{v}} \neq \lambda \mathbf{v}\,.$$

Therefore eigenvalues and eigenvectors for f do not exist.

From the following exercise we see that the dimension of an eigenspace can be greater than 1.

|||| Exercise 13.7

In space an ordinary $(O, \mathbf{i}, \mathbf{j}, \mathbf{k})$-coordinate system is given. All vectors are drawn from the origin. The map p projects vectors down onto the (X, Y)-plane in space:



Eigenvalue problem for the projection down onto the (X, Y)-plane.

It is shown in Exercise 12.28 that p is linear. Determine all eigenvalues and the eigenspaces that correspond to the eigenvalues, solely by mental calculation (ponder).

|||| Example 13.8 The Eigenvalue Problem for Differentiation

We consider the linear map $f : C^\infty(\mathbb{R}) \to C^\infty$ given by

$$f(x(t)) = x'(t)\,.$$

Let λ be an arbitrary scalar. Since

$$f(\mathrm{e}^{\lambda t}) = \lambda\,\mathrm{e}^{\lambda t}\,,$$

λ is an eigenvalue of f and $\mathrm{e}^{\lambda t}$ is an eigenvector that corresponds to λ .

Since all solutions to the differential equation

$$x'(t) = \lambda x(t)$$

is given by $k \cdot \mathrm{e}^{\lambda t}$ where k is an arbitrary real number, the eigenspace corresponding to λ is determined by

$$E_\lambda = \left\{\, k \cdot \mathrm{e}^{\lambda t} \;\middle|\; k \in \mathbb{R} \,\right\}\,.$$

13.1.3 Theoretical Points

The following corollary gives an important result for linear maps of a vector space into itself. It is valid even if the vector space considered is of infinite dimension.

|||| Corollary 13.9

Let $f : V \to V$ be a linear map of a vector space V into itself, and assume

1. that f has a series of eigenvalues with corresponding eigenspaces,
2. that some of the eigenspaces are chosen, and within each of the chosen eigenspaces some linearly independent vectors are chosen,
3. and that all the so chosen vectors are consolidated in a single set of vectors v .

Then v is a linearly independent set of vectors.

|||| Proof

Let $f : V \to V$ be a linear map, and let v be a set of vectors that are put together according to points 1. to 3. in Corollary 13.9. We shall prove that v is linearly independent. The flow of the proof is that we assume the opposite, that is, v is linearly dependent, and show that this leads to a contradiction.

First we delete vectors from v to get a basis for span$\{v\}$. There must be at least one vector in v that does not correspond to the basis. We choose one of these, let us call it $\mathbf{x}$. Now we write $\mathbf{x}$ as a linear combination of the basis vectors, in doing so we leave out the trivial terms, i.e. those with the coefficient 0:

$$\mathbf{x} = k_1\mathbf{v}_1 + \cdots + k_m\mathbf{v}_m \tag{13-7}$$

We term the eigenvalue that corresponds to $\mathbf{x}$ λ, and the eigenvalues corresponding to $\mathbf{v}_i$ λ_i. From (13-7) we can obtain an expression for $\lambda\mathbf{x}$ in two different ways, partly by multiplying (13-7) by λ, partly by finding the image by f of the right and left hand side in (13-7):

$$\begin{aligned}\lambda\mathbf{x} &= \lambda k_1\mathbf{v}_1 + \cdots + \lambda k_m\mathbf{v}_m\\ \lambda\mathbf{x} &= \lambda_1 k_1\mathbf{v}_1 + \cdots + \lambda_m k_m\mathbf{v}_m\end{aligned}$$

Subtracting the lower from the upper equation yields:

$$\mathbf{0} = k_1(\lambda-\lambda_1)\mathbf{v}_1 + \cdots + k_m(\lambda-\lambda_m)\mathbf{v}_m\,. \tag{13-8}$$

If all the coefficients to the vectors on the right hand side of (13-8) are equal to zero, then $\lambda = \lambda_i$ for all $i = 1, 2, \ldots, m$. But then $\mathbf{x}$ and all the basis vectors $\mathbf{v}_i$ that are chosen form the same eigenspace, and therefore they should collectively be linearly independent, this is how they are chosen. This contradicts that $\mathbf{x}$ is a linear combination of the basis vectors.

Therefore at least one of the coefficients in (13-8) must be different from 0. But then the zero vector is written as a proper linear combination of the basis vectors. This contradicts the requirement that a basis is linearly independent.

Conclusion: the assumption that v is a linearly independent set of vectors, necessarily leads to a contradiction. Therefore v is linearly independent.

■

|||| Example 13.10 The Linear Independence of Eigenvectors

A linear map $f : V \to V$ has three eigenvalues λ_1, λ_2 and λ_3 that have the geometric multiplicities 2, 1 and 3, respectively. The set of vectors $(\mathbf{a}_1, \mathbf{a}_2)$ is a basis for E_{λ_1}, $(\mathbf{b})$ is a basis for

E_{λ_2} , and $(\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)$ is a basis for E_{λ_3} . Then it follows from corollary 13.9 that any selection of the six basis vectors is a linearly independent set of vectors.

Corollary 13.9 is useful because it leads directly to the following important results:

|||| Theorem 13.11 General Properties

Let V be a vector space with $\dim(V) = n$, and let $f : V \to V$ be a linear map of V into itself. Then:

1. Proper eigenvectors that correspond to different eigenvalues for f , are linearly independent.
2. f can at the most have n different eigenvalues.
3. If f has n different eigenvalues, then a basis for V exists consisting of eigenvectors for f .
4. The sum of the geometric multiplicities of eigenvalues for f can at the most be n .
5. If and only if the sum of the geometric multiplicities of the eigenvalues for f is equal to n, a basis for V exists consisting of eigenvectors for f .

|||| Exercise 13.12

The first point in 13.11 is a simple special case of Corollary 13.9 and therefore follows directly from the corollary. The second point can be proved like this:

Assume that a linear map has k different eigenvalues. We choose a proper vector from each of the k eigenspaces. The set of the k chosen vectors is then (in accordance with the corollary 13.9) linearly independent, and k must therefore be less than or equal to the dimension of the vector space (see Corollary 11.21).

Similarly, show how the last three points in Theorem 13.11 follow from Corollary 13.9.

Motivated by Theorem 13.11 we introduce the concept eigenbasis:

|||| Definition 13.13 Eigenvector basis

Let $f : V \to V$ be a linear map of a finite-dimensional vector space V into itself.

By an *eigenvector basis*, or in short *eigenbasis*, for V with respect to f we understand a basis consisting of eigenvectors for f .

Now we can present this subsection's main result:

|||| Theorem 13.14 Main Theorem

Let $f : V \to V$ be a linear map of an n-dimensional vector space V into itself, and let $v = (\mathbf{v}_1, \ldots \mathbf{v}_n)$ be a basis for V . Then:

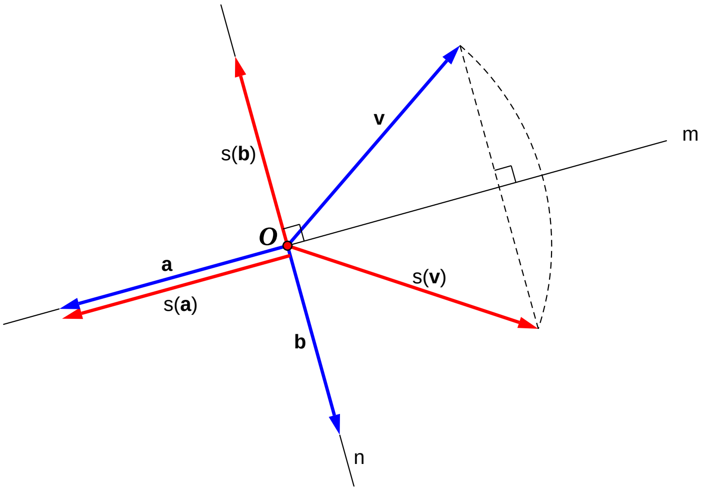
1. The mapping matrix ${}_{\mathrm{v}}\mathbf{F}_{\mathrm{v}}$ for f with respect to v is a diagonal matrix if and only if v is an eigenbasis for V with respect to f .

2. Assume that v is an eigenbasis for V with respect to f . Let $\mathbf{\Lambda}$ denote the diagonal matrix that is the mapping matrix for f with respect to v . The order of the diagonal elements in $\mathbf{\Lambda}$ is then determined from the basis like this: The basis vector $\mathbf{v}_i$ corresponds to the eigenvalue λ_i that is in the i'th column in $\mathbf{\Lambda}$.

The proof of this theorem can be found in eNote 14 (see Theorem 14.7).

|||| Example 13.15 Diagonal Matrix for Reflection

Let us again consider the situation in example 13.5, where we considered the map s that reflects vectors drawn from the origin in the line m:



Reflection about m.

We found that $\mathbf{a}$ is an eigenvector that corresponds to the eigenvalue 1, and that $\mathbf{b}$ is an eigenvector that corresponds to the eigenvalue -1. Since the plane has the dimension 2 it follows from Theorem 13.14 that if we choose the basis $(\mathbf{a},\mathbf{b})$, then f has the following mapping matrix with respect to this basis:

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

|||| Example 13.16 Linear Maps without Eigenvalues

In the example 13.6 we found that the map, which maps a vector in the plane onto its hat vector, has no eigenvalues. Therefore there is no eigenbasis for the map, and therefore it cannot be described by a diagonal matrix for this map.

|||| Example 13.17 Diagonalisation of a Complex Map

Let $f : \mathbb{C}^2 \to \mathbb{C}^2$ be a linear map that satisfies

$$f(z_1, z_2) = (-z_2, z_1).$$

Since:

$$f(i,1) = (-1,i) = i\,(i,1) \quad \text{and} \quad f(-i,1) = (-1,-i) = (-i)(-i,1)\,,$$

it is seen that i is an eigenvalue of f with a corresponding eigenvector $(i,1)$, and that $-i$ is an eigenvalue of f with a corresponding eigenvector $(-i,1)$.

Since $(i,1)$ and $(-i,1)$ are linearly independent, $\big((i,1),(-i,1)\big)$ is an eigenbasis for $\mathbb{C}^2$ with respect to f . The mapping matrix for f with respect to this basis is in accordance with Theorem 13.14

$$\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}.$$

|||| Exercise 13.18

Consider once more the situation in Example 13.7. Choose two different eigenbases (bases consisting of eigenvectors for p), and determine in each of the two cases the diagonal matrix that will become the mapping matrix for p with respect to the chosen basis.

13.2 The Eigenvalue Problem for Square Matrices

When a linear map $f : V \to V$ maps an n-dimensional vector space V into the vector space itself the mapping matrix for f with respect to the arbitrarily chosen basis a becomes a *square* matrix. The eigenvalue problem $f(\mathbf{v}) = \lambda\mathbf{v}$ is the equivalent of the matrix equation:

$${}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}} \cdot {}_{\mathrm{a}}\mathbf{v} = \lambda \cdot {}_{\mathrm{a}}\mathbf{v}\,. \tag{13-9}$$

Thus we can formulate an eigenvalue problem for square matrices generally, that is without necessarily having to think about a square matrix as a mapping matrix. We will standardize the method, when eigenvalues and eigenvectors for square matrices are to be determined. At the same time, due to (13-9), we get methods for finding eigenvalues and eigenvectors for all linear maps of a vector space into itself, that can be described by mapping matrices.

First we define what is to be understood by the eigenvalue problem for a square matrix.

|||| Definition 13.19 The Eigenvalue Problem for Matrices

Solving the eigenvalue problem for a square real $n \times n$-matrix $\mathbf{A}$ means to find a scalar λ and a proper vector $\mathbf{v} = (v_1, \ldots, v_n)$ satisfying the equation:

$$\mathbf{A}\,\mathbf{v} = \lambda \mathbf{v}\,. \tag{13-10}$$

If this equation is satisfied for a pair af λ and $\mathbf{v} \neq \mathbf{0}$, λ is termed an *eigenvalue* of $\mathbf{A}$ and $\mathbf{v}$ an *eigenvector* of $\mathbf{A}$ corresponding to λ.

|||| Example 13.20 The Eigenvalue Problem for a Square Matrix

We wish to investigate whether $\mathbf{v}_1 = (2,3)$, $\mathbf{v}_2 = (4,4)$ and $\mathbf{v}_3 = (2,-1)$ are eigenvectors for $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix} \tag{13-11}$$

For this we write the eigenvalue problem, as stated in Definition 13.19.

$$\begin{aligned} \mathbf{A}\mathbf{v}_1 &= \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 1 \cdot \mathbf{v}_1 \\ \mathbf{A}\mathbf{v}_2 &= \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix}\begin{bmatrix} 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 8 \\ 8 \end{bmatrix} = 2 \cdot \mathbf{v}_2 \\ \mathbf{A}\mathbf{v}_3 &= \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 10 \\ 7 \end{bmatrix} \neq \lambda \cdot \mathbf{v}_3\,. \end{aligned} \tag{13-12}$$

From this we see that $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors for $\mathbf{A}$. $\mathbf{v}_1$ corresponding to the eigenvalue 1, and $\mathbf{v}_2$ corresponding to the eigenvalue 2.

Furthermore we see that $\mathbf{v}_3$ is not an eigenvector for $\mathbf{A}$.

|||| Example 13.21 The Eigenvalue Problem for a Square Matrix

Given the matrix

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

Since

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

0 is an eigenvalue of $\mathbf{A}$ and $(1,1)$ an eigenvector for $\mathbf{A}$ corresponding to the eigenvalue 0.

|||| Example 13.22 The Eigenvalue Problem for a Square Matrix

Given the matrix

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

Since

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

i is a complex eigenvalue of $\mathbf{A}$ and $(-i, 1)$ is a complex eigenvector for $\mathbf{A}$ corresponding to the eigenvalue i.

For the use in the following investigations we make some important comments to Definition 13.19 .

First we note that even if the square matrix $\mathbf{A}$ in Definition 13.19 is real, one is often interested not only in the real solutions to (13-10), but more generally complex solutions. In other words we seek a scalar $\lambda \in \mathbb{C}$ and a vector $\mathbf{v} \in \mathbb{C}^n$, satisfying (13-10).

Therefore it can be convenient to regard the left-hand side of (13-10) as a map $f : \mathbb{C}^n \to \mathbb{C}^n$ given by:

$$f(\mathbf{v}) = \mathbf{A}\,\mathbf{v}\,.$$

This map is linear, viz. let $\mathbf{u} \in \mathbb{C}^n$, $\mathbf{v} \in \mathbb{C}^n$ and $k \in \mathbb{C}$., then according to the usual arithmetic rules for matrices

$$\begin{aligned} &1. \quad f(\mathbf{u}+\mathbf{v}) = \mathbf{A}\,(\mathbf{u}+\mathbf{v}) = \mathbf{A}\,\mathbf{u} + \mathbf{A}\,\mathbf{v} \\ &2. \quad f(k\,\mathbf{u}) = \mathbf{A}(k\,\mathbf{u}) = k(\mathbf{A}\,\mathbf{u}) \end{aligned}$$

By this the linearity is established. Since the eigenvalue problem $f(\mathbf{v}) = \lambda\mathbf{v}$ in this case is *identical* to the eigenvalue problem $\mathbf{Av} = \lambda\mathbf{v}$, we can conclude that results obtained in subsection 9.1 for the eigenvalue problem in general, can be transferred directly to the eigenvalue problem for matrices. Thus let us immediately characterize the set of eigenvectors that correspond to a given eigenvalue of a square, real matrix, compare with Theorem 13.3.

|||| Theorem 13.23 Subspaces of Eigenvectors

Let λ be a real or complex eigenvalue of a real $n \times n$-matrix $\mathbf{A}$. Then the set of complex eigenvectors for $\mathbf{A}$ corresponding to λ, is a subspace in $\mathbb{C}^n$.

If one is only interested in real solutions to the eigenvalue problem for real square matrices, one can alternatively see the left hand side of (13-10) as a real map $f : \mathbb{R}^n \to \mathbb{R}^n$ given by:

$$f(\mathbf{v}) = \mathbf{A}\,\mathbf{v}\,.$$

Of course, this map is linear, too. We get the following version of Theorem 13.23:

|||| Theorem 13.24 Subspaces of Eigenvectors

Let λ be a real eigenvalue of a real $n \times n$-matrix $\mathbf{A}$. Then the set of real eigenvectors for $\mathbf{A}$ corresponding to λ, is a subspace in $\mathbb{R}^n$.

In the light of Theorem 13.23 and Theorem 13.24 we now introduce the concept eigenvector space, compare with Definition 13.4.

|||| Definition 13.25 The Eigenvector Space

Let $\mathbf{A}$ be a square, real matrix, and let λ be an eigenvalue of $\mathbf{A}$.

The subspace of all the eigenvectors that correspond to λ is termed the ***eigenvector space*** (or in short the ***eigenspace***) corresponding to λ and is termed E_λ .

Now we have sketched the structural framework for the eigenvalue problem for square matrices, and we continue in the following two subsections by investigating in an elementary way, how one can begin to find eigenvalues and eigenvectors for square matrices.

13.2.1 To Find the Eigenvalues for a Square Matrix

We wish to determine the eigenvalues that correspond to a real $n \times n$ matrix $\mathbf{A}$. The starting point is the equation

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}\,, \tag{13-13}$$

First we put $\lambda\mathbf{v}$ onto the left hand of the equality sign, and then $\mathbf{v}$ "is placed outside a pair of brackets". This is possible because $\mathbf{v} = \mathbf{E}\,\mathbf{v}$ where $\mathbf{E}$ is the identity matrix:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \;\Leftrightarrow\; \mathbf{A}\mathbf{v} - \lambda(\mathbf{E}\mathbf{v}) = \mathbf{A}\mathbf{v} - (\lambda\mathbf{E})\mathbf{v} = \mathbf{0} \;\Leftrightarrow\; (\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}\,. \tag{13-14}$$

The last equation in (13-14) corresponds to a homogeneous system of linear equations consisting of n equations in the n unknowns v_1, ..., v_n, that are the elements in $\mathbf{v} = (v_1, ..., v_n)$. However, it is not possible to solve the system of equations directly, precisely because we do not know λ. We have to continue the work with the coefficient matrix of the system of equations. We give this matrix a special symbol:

$$\mathbf{K}_{\mathbf{A}}(\lambda) = (\mathbf{A} - \lambda\mathbf{E})$$

and is called ***the characteristic matrix*** of $\mathbf{A}$.

Since it is a homogeneous system of linear equations that we have to solve we have two possibilities for the structure of the solution. Either the characteristic matrix is *invertible*, and the the only solution is $\mathbf{v} = \mathbf{0}$. Or the matrix is *singular*, and then infinitely many solutions $\mathbf{v}$ exist. But since Definition 13.19 requires that $\mathbf{v}$ must be a proper vector, that is a vector different from the zero vector, the characteristic matrix must be singular. To investigate whether this is true, we take the determinant of the square matrix. This is zero exactly when the matrix is singular:

$$\det(\mathbf{A} - \lambda\mathbf{E}) = 0\,. \tag{13-15}$$

Note that the left hand side in (13-15) is a polynomial in the variable λ. The polynomial is given a special symbol:

$$K_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{E}) = \det(\mathbf{K}_{\mathbf{A}}(\lambda))$$

and is termed ***the characteristic polynomial*** of $\mathbf{A}$.

The equation that results when the characteristic polynomial is set equal to zero

$$K_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{E}) = \det(\mathbf{K}_{\mathbf{A}}(\lambda)) = 0$$

is termed the ***characteristic equation*** of $\mathbf{A}$.

By the use of the method for calculating the determinant we see that the characteristic polynomial is always an n'th degree polynomial. See also the following examples. The main point is that the roots in the characteristic polynomial (solutions to the character equation) are the eigenvalues of the matrix, because the eigenvalues precisely satisfy that the characteristic matrix is singular.

It is also common to define the characteristic matrix as $\lambda\mathbf{E} - \mathbf{A}$, since the homogeneous equation for this matrix has the same solutions, and the zeros of the corresponding characteristic polynomial $\det(\lambda\mathbf{E} - \mathbf{A}) = 0$ are also the same. But note that $\det(\lambda\mathbf{E} - \mathbf{A}) = (-1)^n \det(\mathbf{A} - \lambda\mathbf{E})$.

|||| Example 13.26 Eigenvalues for 2×2 Matrices

Given two matrices **A** and **B**:

$$\mathbf{A} = \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} -1 & 4 \\ -2 & 3 \end{bmatrix}. \tag{13-16}$$

We wish to determine the eigenvalues for **A** and **B**.

First we consider **A**. Its characteristic matrix reads:

$$\mathbf{K_A}(\lambda) = \mathbf{A} - \lambda\mathbf{E} = \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 4-\lambda & -2 \\ 3 & -1-\lambda \end{bmatrix}. \tag{13-17}$$

Now we determine the characteristic polynomial:

$$\begin{aligned} K_{\mathbf{A}}(\lambda) = \det(\mathbf{K_A}(\lambda)) &= \det\left(\begin{bmatrix} 4-\lambda & -2 \\ 3 & -1-\lambda \end{bmatrix}\right) \\ &= (4-\lambda)(-1-\lambda) - (-2)\cdot 3 = \lambda^2 - 3\lambda + 2\,. \end{aligned} \tag{13-18}$$

The polynomial as expected has the degree 2. The characteristic equation can be written and the solutions determined:

$$K_{\mathbf{A}}(\lambda) = 0 \;\Leftrightarrow\; \lambda^2 - 3\lambda + 2 = 0 \;\Leftrightarrow\; \lambda = 1 \;\text{ or }\; \lambda = 2\,. \tag{13-19}$$

Thus **A** has two eigenvalues: $\lambda_1 = 1$ and $\lambda_2 = 2$.

The same technique is used for the determination of possible eigenvalues of **B**.

$$\begin{aligned} \mathbf{K_B}(\lambda) = \mathbf{B} - \lambda\mathbf{E} &= \begin{bmatrix} -1 & 4 \\ -2 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} -1-\lambda & 4 \\ -2 & 3-\lambda \end{bmatrix} \\ K_{\mathbf{B}}(\lambda) = \det(\mathbf{K_B}(\lambda)) &= \det\left(\begin{bmatrix} -1-\lambda & 4 \\ -2 & 3-\lambda \end{bmatrix}\right) \\ &= (-1-\lambda)(3-\lambda) - 4\cdot(-2) = \lambda^2 - 2\lambda + 5\,. \end{aligned} \tag{13-20}$$

In this case there are no real solutions to $K_{\mathbf{B}}(\lambda) = 0$, because the discriminant $d = (-2)^2 - 4 \cdot 1 \cdot 5 = -16 < 0$, and therefore **B** has no real eigenvalues. But it has two complex eigenvalues. We use the complex "toolbox": The discriminant can be rewritten as $d = (4i)^2$, which gives the two complex solutions

$$\lambda = \frac{2 \pm 4i}{2} \;\Leftrightarrow\; \lambda = 1 + 2i \;\text{ and }\; \bar{\lambda} = 1 - 2i \tag{13-21}$$

Thus **B** has two complex eigenvalues: $\lambda_1 = 1 + 2i$ and $\lambda_2 = 1 - 2i$.

In the following theorem the conclusions of this subsection are summarized.

|||| Theorem 13.27 The Characteristic Polynomial

For the square real $n \times n$-matrix $\mathbf{A}$ consider

1. *The characteristic matrix* $\mathbf{K_A}(\lambda) = \mathbf{A} - \lambda\mathbf{E}$.
2. *The characteristic polynomial* $K_{\mathbf{A}}(\lambda) = \det(\mathbf{K_A}(\lambda)) = \det(\mathbf{A} - \lambda\mathbf{E})$.
3. *The characteristic equation* $K_{\mathbf{A}}(\lambda) = 0$.

Then:

1. The characteristic polynomial is an n'th degree polynomial with the variable λ, and similarly the characteristic equation is an n'th degree equation with the unknown λ.

2. The roots of the characteristic polynomial (the solutions to the characteristic equation) are all the eigenvalues of $\mathbf{A}$.

13.2.2 To Find the Eigenvectors of a Square Matrix

After the eigenvalues of a real $n \times n$ matrix $\mathbf{A}$ are determined, it is possible to determine the corresponding eigenvectors. The procedure starts with the equation

$$(\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}, \tag{13-22}$$

that was achieved in (13-14). Since the eigenvalues are now known, the homogeneous system of linear equations corresponding to (13-22) can be solved with respect to the n unknowns $v_1, ..., v_n$ that are the elements in $\mathbf{v} = (v_1, ..., v_n)$. We just have to substitute the eigenvalues one after one. As mentioned above, the characteristic matrix is singular when the substituted λ is an eigenvalue. Therefore infinitely many solutions to the system of equations exist. Finding these corresponds to finding all eigenvectors $\mathbf{v}$ that correspond to λ.

In the following method we summarize the problem of determining eigenvalues and the corresponding eigenvectors of a square matrix.

|||| Method 13.28 Determination of Eigenvectors

All (real or complex) eigenvalues λ for the square matrix $\mathbf{A}$ are found as the solutions to the ***characteristic equation*** of $\mathbf{A}$:

$$K_{\mathbf{A}}(\lambda) = 0 \;\Leftrightarrow\; \det(\mathbf{A} - \lambda\mathbf{E}) = 0\,. \tag{13-23}$$

Then the eigenvectors $\mathbf{v}$ corresponding to each of the eigenvalues λ can be determined. They are the solutions to the following system of linear equations

$$(\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}\,, \tag{13-24}$$

when the eigenvalue λ is substituted. $\mathbf{E}$ is the identity matrix.

Method 13.28 is unfolded in the following three examples that also show a way of characterizing the set of eigenvectors corresponding to a given eigenvalue, in the light of Theorem 13.23 and Theorem 13.24.

|||| Example 13.29 Eigenvectors Belonging to Given Eigenvalues

Given the square matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{13-25}$$

We wish to determine eigenvalues and eigenvectors to $\mathbf{A}$ and use method 13.28. First the characteristic matrix is found:

$$\mathbf{K}_{\mathbf{A}}(\lambda) = \mathbf{A} - \lambda\mathbf{E} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \tag{13-26}$$

Then the characteristic polynomial is formed:

$$\begin{aligned} K_{\mathbf{A}}(\lambda) &= \det(\mathbf{A} - \lambda\mathbf{E}) \\ = \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\right) &= (2-\lambda)(2-\lambda) - 1\cdot 1 = \lambda^2 - 4\lambda + 3\,. \end{aligned} \tag{13-27}$$

The characteristic equation that is $\lambda^2 - 4\lambda + 3 = 0$, has the solutions $\lambda_1 = 1$ and $\lambda_2 = 3$, which are all the real eigenvalues of $\mathbf{A}$.

In order to determine the eigenvectors corresponding to λ_1, it is substituted into $(\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}$, and then we solve the system of linear equations with the augmented matrix:

$$\mathbf{T} = [\,\mathbf{A} - \lambda_1\mathbf{E}\,|\,\mathbf{0}\,] = \left[\begin{array}{cc|c} 2-1 & 1 & 0 \\ 1 & 2-1 & 0 \end{array}\right]. \tag{13-28}$$

By Gauss-Jordan elimination we get

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] \tag{13-29}$$

Thus there are infinitely many solutions $\mathbf{v} = (v_1, v_2)$, since there is only one non-trivial equation: $v_1 + v_2 = 0$. If we are only looking for one proper eigenvector corresponding to the eigenvalue λ_1, we can put v_2 equal to 1, and we get the eigenvector $\mathbf{v}_1 = (-1, 1)$. All real eigenvectors corresponding to λ_1 can then be written as

$$\mathbf{v} = t \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} \, , \, t \in \mathbb{R} \, . \tag{13-30}$$

This is a one-dimensional subspace in $\mathbb{R}^2$, viz. the eigenspace that corresponds to 1 that can also be written like this:

$$E_1 = \operatorname{span}\{(-1, 1)\} \, . \tag{13-31}$$

Now λ_2 is substituted in $(\mathbf{A} - \lambda \mathbf{E})\mathbf{v} = \mathbf{0}$, and we then solve the corresponding system of linear equations that has the augmented matrix

$$\mathbf{T} = [\mathbf{A} - \lambda_2 \mathbf{E} \,|\, \mathbf{0}\,] = \left[\begin{array}{cc|c} 2-3 & 1 & 0 \\ 1 & 2-3 & 0 \end{array}\right] . \tag{13-32}$$

By Gauss-Jordan elimination we get

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right] . \tag{13-33}$$

From this we see that $\mathbf{v}_2 = (1, 1)$ is an eigenvector corresponding to the eigenvalue λ_2. All real eigenvectors corresponding to λ_2 can be written as

$$\mathbf{v} = t \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \, , \, t \in \mathbb{R} \, . \tag{13-34}$$

This is a one-dimensional subspace in $\mathbb{R}^2$ that can also be written as:

$$E_3 = \operatorname{span}\{(1, 1)\} \, . \tag{13-35}$$

We will now check our understanding: When $\mathbf{v}_1 = (-1, 1)$ is mapped by $\mathbf{A}$, will the image vector only be a scaling (change of length) of $\mathbf{v}_1$?

$$\mathbf{A}\mathbf{v}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \mathbf{v}_1 \, . \tag{13-36}$$

It is true! It is also obvious that the eigenvalue is 1 .

Now we check $\mathbf{v}_2$:

$$\mathbf{A}\mathbf{v}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \cdot \mathbf{v}_2 \, . \tag{13-37}$$

$\mathbf{v}_2$ is also as expected an eigenvector and the eigenvalue is 3.

|||| Example 13.30 Complex Eigenvalues and Eigenvectors

In Example 13.26 a matrix **B** is given

$$\mathbf{B} = \begin{bmatrix} -1 & 4 \\ -2 & 3 \end{bmatrix} \tag{13-38}$$

that has no real eigenvalues. But we found two complex eigenvalues, $\lambda_1 = 1 + 2i$ and $\lambda_2 = 1 - 2i$.

We substitute λ_1 in $(\mathbf{B} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}$ and then we solve the corresponding system of linear equations that has the augmented matrix

$$\mathbf{T} = [\,\mathbf{B} - \lambda_1\mathbf{E}\,|\,\mathbf{0}\,] = \left[\begin{array}{cc|c} -1-(1+2i) & 4 & 0 \\ -2 & 3-(1+2i) & 0 \end{array}\right] \tag{13-39}$$

By Gauss-Jordan elimination we get

$$\operatorname{rref}(\mathbf{T}) = \left[\begin{array}{cc|c} 1 & -1+i & 0 \\ 0 & 0 & 0 \end{array}\right] \tag{13-40}$$

This corresponds to one non-trivial equation $v_1 + (-1+i)v_2 = 0$, and if we put $v_2 = s$, we see that all the complex eigenvectors corresponding to λ_1 are given by

$$\mathbf{v} = s \cdot \begin{bmatrix} 1-i \\ 1 \end{bmatrix},\ s \in \mathbb{C}\,. \tag{13-41}$$

This is a one-dimensional subspace in $\mathbb{C}^2$, viz. the eigenspace corresponding to the eigenvalue $1 + 2i$ which we also can state like this:

$$E_{1+2i} = \operatorname{span}\{(1-i, 1)\}\,. \tag{13-42}$$

Similarly all complex solutions corresponding to λ_2 are given by

$$\mathbf{v} = s \cdot \begin{bmatrix} 1+i \\ 1 \end{bmatrix},\ s \in \mathbb{C}\,. \tag{13-43}$$

This is a one-dimensional subspace in $\mathbb{C}^2$ which we also can state like this:

$$E_{1-2i} = \operatorname{span}\{(1+i, 1)\}\,. \tag{13-44}$$

In the following example we find eigenvalues and corresponding eigenspaces for a 3 × 3-matrix. It turns out that in this case to one of the eigenvalues corresponds a two-dimensional eigenspace.

|||| Example 13.31 Eigenvalue with Multiplicity 2

Given the matrix **A**:

$$\mathbf{A}=\begin{bmatrix}6 & 3 & 12\\ 4 & -5 & 4\\ -4 & -1 & -10\end{bmatrix} \tag{13-45}$$

First we wish to determine the eigenvalues of **A** and use Method 13.28.

$$\det\left(\begin{bmatrix}6-\lambda & 3 & 12\\ 4 & -5-\lambda & 4\\ -4 & -1 & -10-\lambda\end{bmatrix}\right)=-\lambda^3-9\lambda^2+108=-(\lambda-3)(\lambda+6)^2=0 \tag{13-46}$$

From the last factorization it is seen that **A** has two different eigenvalues. The eigenvalue $\lambda_1=-6$ is a double root in the characteristic equation, while the eigenvalue $\lambda_2=3$ is a single root.

Now we determine the eigenspace corresponding to $\lambda_1=-6$, see Theorem 13.23:

$$\begin{gathered}\left[\begin{array}{ccc|c}6-(-6) & 3 & 12 & 0\\ 4 & -5-(-6) & 4 & 0\\ -4 & -1 & -10-(-6) & 0\end{array}\right]\rightarrow\\ \left[\begin{array}{ccc|c}12 & 3 & 12 & 0\\ 4 & 1 & 4 & 0\\ -4 & -1 & -4 & 0\end{array}\right]\rightarrow\left[\begin{array}{ccc|c}4 & 1 & 4 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{array}\right]\end{gathered} \tag{13-47}$$

Here is only one nontrivial equation: $4x_1+x_2+4x_3=0$. If we put x_1 and x_3 equal to the two free parameters s and t all real eigenvectors corresponding to the eigenvalue -6 are given by:

$$\mathbf{x}=\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=s\cdot\begin{bmatrix}1\\ -4\\ 0\end{bmatrix}+t\cdot\begin{bmatrix}0\\ -4\\ 1\end{bmatrix}\ ,\ s,t\in\mathbb{R}\,. \tag{13-48}$$

This is a two-dimensional subspace in $\mathbb{R}^3$ which can also be stated like this:

$$E_{-6}=\operatorname{span}\{(1,-4,0),(0,-4,1)\}\,. \tag{13-49}$$

It is thus possible to find two linearly independent eigenvectors corresponding to λ_1. What about the number of linearly independent eigenvectors for $\lambda_2=3$?

$$\begin{gathered}\left[\begin{array}{ccc|c}6-3 & 3 & 12 & 0\\ 4 & -5-3 & 4 & 0\\ -4 & -1 & -10-3 & 0\end{array}\right]\rightarrow\\ \left[\begin{array}{ccc|c}3 & 3 & 12 & 0\\ 4 & -8 & 4 & 0\\ -4 & -1 & -13 & 0\end{array}\right]\rightarrow\left[\begin{array}{ccc|c}1 & 1 & 4 & 0\\ 0 & -3 & -3 & 0\\ 0 & 3 & 3 & 0\end{array}\right]\rightarrow\left[\begin{array}{ccc|c}1 & 1 & 4 & 0\\ 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 0\end{array}\right]\end{gathered} \tag{13-50}$$

Here are two non-trivial equations: $x_1 + x_2 + 4x_3 = 0$ and $x_2 + x_3 = 0$. If we put $x_3 = s$ equal to the free parameter, then all real eigenvectors corresponding to the eigenvalue 3 are given by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = s \cdot \begin{bmatrix} -3 \\ -1 \\ 1 \end{bmatrix}, \, s \in \mathbb{R}\,. \tag{13-51}$$

This is a one-dimensional subspace in $\mathbb{R}^3$ that can also be stated like this:

$$E_3 = \operatorname{span}\{(-3,-1,1)\}\,. \tag{13-52}$$

Thus it is only possible to find one linearly independent eigenvector corresponding to λ_2.

|||| Exercise 13.32

Given the square matrix

$$\mathbf{A} = \begin{bmatrix} 5 & -4 & 4 \\ 0 & -1 & 6 \\ 0 & 1 & 4 \end{bmatrix}. \tag{13-53}$$

1. Determine all eigenvalues of $\mathbf{A}$.
2. Determine for each of the eigenvalues the corresponding eigenspace.
3. State at least 3 eigenvectors (not necessarily linearly independent) corresponding to each eigenvalue.

13.2.3 Algebraic and Geometric Multiplicity

As is evident from Example 13.31 it is important to pay attention to whether an eigenvalue is a single root or a multiple root of the characteristic equation of a square real matrix and to the dimension of the corresponding eigenspace. In this subsection we investigate the relation between the two phenomena. This gives rise to the following definitions.

|||| Definition 13.33 Algebraic and Geometric Multiplicity

Let $\mathbf{A}$ be a square, real matrix, and let λ be an eigenvalue of $\mathbf{A}$.

1. λ is said to have ***the algebraic multiplicity*** n when λ is a n-double root in the characteristic equation of the square matrix $\mathbf{A}$. This is is termed $\operatorname{am}(\lambda) = n$.

2. λ is said to have ***the geometric multiplicity*** m when the dimension of the eigenvector space corresponding to λ is m. This is termed $\operatorname{gm}(\lambda) = m$. In other words: $\dim(E_\lambda) = \operatorname{gm}(\lambda)$.

We do not always have $\operatorname{am}(\lambda) = \operatorname{gm}(\lambda)$. This is dealt with in Theorem 13.34.

The following theorem has some important properties concerning algebraic and geometric multiplicity of eigenvalues of square matrices, cf. Theorem 13.11.

|||| Theorem 13.34 Properties of Multiplicities

Given a real $n \times n$-matrix $\mathbf{A}$.

1. $\mathbf{A}$ has at the most n different real eigenvalues, and also the sum of algebraic multiplicities of the real eigenvalues is at the most n .

2. $\mathbf{A}$ has at the most n different complex eigenvalues, but the sum of the algebraic multiplicities of the complex eigenvalues is equal to n .

3. If λ is a real or complex eigenvalue of $\mathbf{A}$, then:

$$1 \leq \operatorname{gm}(\lambda) \leq \operatorname{am}(\lambda) \leq n \tag{13-54}$$

That is, the geometric multiplicity of an eigenvalue will at the least be equal to 1, it will be less than or equal to the algebraic multiplicity of the eigenvalue, which in turn will be less than or equal to the number of rows and columns in $\mathbf{A}$.

|||| Exercise 13.35

Check that all three points in Theorem 13.34 are valid for the eigenvalues and eigenvectors in example 13.31.

Let us comment upon 13.34:

Points 1 and 2 follow directly from the theory of polynomials. The characteristic polynomial for a real $n \times n$-matrix $\mathbf{A}$ is an n'th degree polynomial, and it has at the most n different roots, counting both real and complex ones. Furthermore the sum of the multiplicities of the real roots is at the most n, whereas the sum of the multiplicities to the complex roots is equal to n.

We have previously shown that for every linear map of an n-dimensional vector space into itself the sum of the geometric multiplicities of the eigenvalues for f can at the most be n, see Theorem 13.11. Note that this can be deduced directly from the statements about multiplicities in Theorem 13.34.

As something new and interesting it is postulated in point 3 that the geometric multiplicity of a single eigenvalue can be less than the algebraic multiplicity. This is demonstrated in the following summarizing Example 13.36. Furthermore the geometric multiplicity of a single eigenvalue cannot be greater than the algebraic one. The proof of point 3 in Theorem 13.34 is left out.

|||| Example 13.36 Geometric Multiplicity Less than Algebraic Multiplicity

Given the matrix

$$\mathbf{A} = \begin{bmatrix} -9 & 10 & 0 \\ -3 & 1 & 5 \\ 1 & -4 & 6 \end{bmatrix} \tag{13-55}$$

The eigenvalues of $\mathbf{A}$ are determined:

$$\det\left(\begin{bmatrix} -9-\lambda & 10 & 0 \\ -3 & 1-\lambda & 5 \\ 1 & -4 & 6-\lambda \end{bmatrix}\right) = -\lambda^3 - 2\lambda^2 + 7\lambda - 4 = -(\lambda+4)(\lambda-1)^2 = 0\,. \tag{13-56}$$

From the factorization in front of the last equality sign we get that $\mathbf{A}$ has two different eigenvalues: $\lambda_1 = -4$ and $\lambda_2 = 1$. Moreover $\operatorname{am}(-4) = 1$ and $\operatorname{am}(1) = 2$, as can be seen from the factorization.

The eigenspace corresponding to $\lambda_1 = -4$ is determined by solving $(\mathbf{A} - \lambda_1 \mathbf{E})\mathbf{v} = \mathbf{0}$:

$$\begin{array}{c}\left[\begin{array}{ccc|c} -9-(-4) & 10 & 0 & 0 \\ -3 & 1-(-4) & 5 & 0 \\ 1 & -4 & 6-(-4) & 0 \end{array}\right] \rightarrow \\ \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & -1 & 5 & 0 \\ 0 & -2 & 10 & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -5 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]\end{array} \tag{13-57}$$

There are two non-trivial equations: $v_1 - 2v_2 = 0$ and $v_2 - 5v_3 = 0$. If we put v_3 equal to the free parameter we see that all real eigenvectors corresponding to λ_1 can be stated as

$$E_{-4} = \left\{ s \cdot (10, 5, 1) \;\middle|\; s \in \mathbb{R} \right\} = \operatorname{span}\{(10, 5, 1)\}\,. \tag{13-58}$$

We have that $\operatorname{gm}(-4) = \dim(E_{-4}) = 1$, and that an eigenvector to λ_1 is $\mathbf{v}_1 = (10, 5, 1)$. It is seen that $\operatorname{gm}(-4) = \operatorname{am}(-4)$.

Similarly for $\lambda_2 = 1$:

$$\begin{array}{c}\left[\begin{array}{ccc|c} -9-1 & 10 & 0 & 0 \\ -3 & 1-1 & 5 & 0 \\ 1 & -4 & 6-1 & 0 \end{array}\right] \rightarrow \\ \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & -3 & 5 & 0 \\ 0 & -3 & 5 & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 3 & -5 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]\end{array} \tag{13-59}$$

Again we have two non-trivial equations: $v_1 - v_2 = 0$ and $3v_2 - 5v_3 = 0$. If we put $v_3 = 3s$ we see that all to λ_2 corresponding real eigenvectors can be stated as

$$E_1 = \left\{ s \cdot (5, 5, 3) \;\middle|\; s \in \mathbb{R} \right\} = \operatorname{span}\{(5, 5, 3)\}\,. \tag{13-60}$$

This gives the following results: $\operatorname{gm}(1) = \dim(E_1) = 1$ and that an eigenvector to $\lambda_2 = \lambda_3$ is $\mathbf{v}_2 = (5, 5, 3)$. Furthermore it is seen that $\operatorname{gm}(1) < \operatorname{am}(1)$.

13.2.4 More About the Complex Problem

We will use the matrix

$$\mathbf{B} = \begin{bmatrix} -1 & 4 \\ -2 & 3 \end{bmatrix}. \tag{13-61}$$

From Example 13.30 in order to make more precise some special phenomena for square, real matrices when their eigenvalue problems are studied in a complex framework.

We found that $\mathbf{B}$ has the eigenvalues, $\lambda_1 = 1+2i$ and $\lambda_2 = 1-2i$. We see that the eigenvalues are conjugate numbers. Another remarkable thing in Example 13.30 is that where

$$\mathbf{v} = \begin{bmatrix} 1-i \\ 1 \end{bmatrix}$$

is an eigenvector corresponding to $\lambda_1 = 1+2i$, then the conjugate vector

$$\overline{\mathbf{v}} = \begin{bmatrix} 1+i \\ 1 \end{bmatrix}$$

is an eigenvector for $\lambda_2 = 1-2i$. Both are examples of general rules:

|||| Theorem 13.37 Conjugate Eigenvalues and Eigenvectors

For a square, real matrix $\mathbf{A}$ we have:

1. If λ is a complex eigenvalue of $\mathbf{A}$ in rectangular form $\lambda = a+ib$, then $\overline{\lambda} = a-ib$ is also an eigenvalue of $\mathbf{A}$.

2. If $\mathbf{v}$ is an eigenvector for $\mathbf{A}$ coresponding to the complex eigenvalue λ, then the conjugate vector $\overline{\mathbf{v}}$ is an eigenvector for $\mathbf{A}$ corresponding to the conjugate eigenvalue $\overline{\lambda}$.

|||| Proof

The first part of Theorem 13.37 follows from the theory of polynomials. The characteristic polynomial of a square, real matrix is a polynomial with real coefficients. The roots of such a polynomial come in conjugate pairs.

■

By the *trace* of a square matrix we understand the sum of the diagonal elements. The trace of $\mathbf{B}$ is thus $-1+3 = 2$. Now notice that the sum of the eigenvalues of $\mathbf{B}$ is $(1-i)+(1+i) = 2$, that is equal to the trace of $\mathbf{B}$. This is also a general phenomenon, which we state without proof:

|||| Theorem 13.38 The Trace

For a square, real matrix **A**, the trace **A**, i.e. the sum of the diagonal elements in **A**, is equal to the sum of all (real and/or complex) eigenvalues of **A** , where every eigenvalue is counted in the sum the number of times corresponding to the algebraic multiplicity of the eigenvalue.

|||| Exercise 13.39

In Example 13.31 we found that the characteristic polynomial for the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & 3 & 12 \\ 4 & -5 & 4 \\ -4 & -1 & -10 \end{bmatrix}$$

has the double root -6 and the single root 3 . Prove that Theorem 13.38 is valid in this case.

eNote 14

Similarity and Diagonalization

In this eNote it is explained how certain square matrices can be diagonalized by the use of eigenvectors. Therefore it is presumed that you know how to determine eigenvalues and eigenvectors for a square matrix and furthermore that you know about algebraic and geometric multiplicity.

Updated: 4.11.21 David Brander

If we consider a linear map $f : V \to V$ of an n-dimensional vector space V to itself, then the mapping matrix for f with respect to an arbitrary basis for f is a square, $n \times n$ matrix. If two bases a and b for V are given, then the relation between the corresponding mapping matrices ${}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}}$ and ${}_{\mathrm{b}}\mathbf{F}_{\mathrm{b}}$ are given by

$$ {}_{\mathrm{b}}\mathbf{F}_{\mathrm{b}} = ({}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}})^{-1} \cdot {}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}} \cdot {}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} \tag{14-1} $$

where ${}_{\mathrm{a}}\mathbf{M}_{\mathrm{b}} = \begin{bmatrix} {}_{\mathrm{a}}\mathbf{b}_1 & {}_{\mathrm{a}}\mathbf{b}_2 & \cdots & {}_{\mathrm{a}}\mathbf{b}_n \end{bmatrix}$ is the change of basis matrix that shifts from b to a coordinates.

It is of special interest if a basis v consisting of eigenvectors for f can be found. Viz. let a be an arbitrary basis for V and ${}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}}$ the corresponding mapping matrix for f. Furthermore let v be an eigenvector basis for V with respect to f. From Theorem 13.14 in eNote 13 it appears that the mapping matrix for f with respect to the v-basis is a diagonal matrix $\mathbf{\Lambda}$ in which the diagonal elements are the eigenvalues of f. If $\mathbf{V}$ denotes the change of basis matrix that shifts from v-coordinates to the a-coordinate vectors, according to (14-1)$\mathbf{\Lambda}$ will appear as

$$ \mathbf{\Lambda} = \mathbf{V}^{-1} \cdot {}_{\mathrm{a}}\mathbf{F}_{\mathrm{a}} \cdot \mathbf{V}\,. \tag{14-2} $$

Naturally formula 14-1 and formula 14-2 inspire questions that take their starting point in square matrices: Which conditions should be satisfied in order for two given square matrices to be interpreted as mapping matrices for the same linear map with respect to two different bases? And which conditions should a square matrix satisfy in order to be a mapping matrix for a linear map that in another basis has a diagonal matrix as a mapping matrix? First we study these questions in a pure matrix algebra context and return in the last subsection to the mapping viewpoint. For this purpose we now introduce the concept similar matrices.

14.1 Similar Matrices

|||| Definition 14.1 Similar Matrices

Given the $n \times n$-matrices $\mathbf{A}$ and $\mathbf{B}$. One says that $\mathbf{A}$ is *similar to* $\mathbf{B}$ if an invertible matrix $\mathbf{M}$ can be found such that

$$\mathbf{B} = \mathbf{M}^{-1}\,\mathbf{A}\,\mathbf{M}\,. \tag{14-3}$$

|||| Example 14.2 Similar Matrices

Given the matrices $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 3 & -4 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 8 & 21 \\ -3 & -10 \end{bmatrix}$.

The matrix $\mathbf{M} = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}$ is invertible and has the inverse matrix $\mathbf{M}^{-1} = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}$.

Consider the following calculation:

$$\begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 3 & -4 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 8 & 21 \\ -3 & -10 \end{bmatrix}.$$

This shows that $\mathbf{A}$ is similar to $\mathbf{B}$.

If $\mathbf{A}$ is similar to $\mathbf{B}$, then $\mathbf{B}$ is also similar to $\mathbf{A}$. If we put $\mathbf{N} = \mathbf{M}^{-1}$ then $\mathbf{N}$ is invertible and

$$\mathbf{B} = \mathbf{M}^{-1}\,\mathbf{A}\,\mathbf{M} \;\Leftrightarrow\; \mathbf{M}\,\mathbf{B}\,\mathbf{M}^{-1} = \mathbf{A} \;\Leftrightarrow\; \mathbf{A} = \mathbf{N}^{-1}\,\mathbf{B}\,\mathbf{N}\,,$$

Therefore one uses the phrase: $\mathbf{A}$ and $\mathbf{B}$ are *similar matrices* .

|||| Theorem 14.3 Similarity Is Transitive

Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be $n \times n$-matrices. If $\mathbf{A}$ is similar to $\mathbf{B}$ and $\mathbf{B}$ is similar to $\mathbf{C}$ then $\mathbf{A}$ is similar to $\mathbf{C}$.

|||| Exercise 14.4

Prove Theorem 14.3.

Regarding the eigenvalues of similar matrices the following theorem applies.

|||| Theorem 14.5 Similarity and Eigenvalues

If $\mathbf{A}$ is similar to $\mathbf{B}$ then the two matrices have identical eigenvalues with the same corresponding algebraic and geometric multiplicities.

|||| Proof

Let $\mathbf{M}$ be an invertible matrix that satisfies $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$ and let, as usual, $\mathbf{E}$ denote the identity matrix of the same size as the three given matrices. Then:

$$\begin{aligned}\det(\mathbf{B} - \lambda\mathbf{E}) &= \det(\mathbf{M}^{-1}\mathbf{A}\mathbf{M} - \lambda\mathbf{M}^{-1}\mathbf{E}\mathbf{M}) \\ &= \det(\mathbf{M}^{-1}(\mathbf{A} - \lambda\mathbf{E})\mathbf{M}) \qquad (14\text{-}4) \\ &= \det(\mathbf{A} - \lambda\mathbf{E})\,.\end{aligned}$$

Thus it is shown that the two matrices have the same characteristic polynomial and thus the same eigenvalues with the same corresponding algebraic multiplicities. Moreover, that they have the same eigenvalues appears from Theorem 14.13 which is given below: When $\mathbf{A}$ and

B can represent the same linear map f with respect to different bases they have identical eigenvalues, viz. the eigenvalues of f.

But the eigenvalues also do have the same geometric multiplicities. This follows from the fact that the eigenspaces for **A** and **B** with respect to any of the eigenvalues can be interpreted as two different coordinate representations of the same eigenspace, viz. the eigenspace for f with respect to the said eigenvalue.

■

Note that Theorem 14.5 says that two similar matrices have the same eigenvalues, but not vice versa: that two matrices, which have the same eigenvalues, are similar. There is a difference and only the first statement is true.

Two similar matrices **A** and **B** have the same eigenvalues, but an eigenvector for the one is not generally and eigenvector for the other. But if **v** is an eigenvector for **A** corresponding to the eigenvalue λ then $\mathbf{M}^{-1}\mathbf{v}$ is an eigenvector for **B** corresponding to the eigenvalue λ, where **M** is the invertible matrix that satisfies $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$. Viz.:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad \Leftrightarrow \quad \mathbf{M}^{-1}\mathbf{A}\mathbf{v} = \mathbf{M}^{-1}\lambda\mathbf{v} \quad \Leftrightarrow \quad \mathbf{B}(\mathbf{M}^{-1}\mathbf{v}) = \lambda(\mathbf{M}^{-1}\mathbf{v})\,. \qquad (14\text{-}5)$$

|||| Exercise 14.6

Explain that two square $n \times n$-matrices are similar, if they have identical eigenvalues with the same corresponding geometric multiplicities and that the sum of the geometric multiplicities is n.

14.2 Matrix Diagonalization

Consider a matrix **A** and an invertible matrix **V** given by

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix} \text{ and } \mathbf{V} = \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix}. \qquad (14\text{-}6)$$

Since

$$\mathbf{V}^{-1}\,\mathbf{A}\,\mathbf{V} = \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix},$$

$\mathbf{A}$ possesses a special property: it is similar to a diagonal matrix, viz. the diagonal matrix

$$\mathbf{\Lambda} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

In this context one says that $\mathbf{A}$ has been ***diagonallzed by similarity transformation***.

Now we will ask the question whether or not an arbitrary square matrix $\mathbf{A}$ can be diagonalized by a similarity transformation. Therefore we form the equation

$$\mathbf{V}^{-1}\,\mathbf{A}\,\mathbf{V} = \mathbf{\Lambda},$$

where $\mathbf{V}$ is an invertible matrix and $\mathbf{\Lambda}$ is a diagonal matrix. Below we prove that the equation has exactly one solution if the columns of $\mathbf{V}$ are linearly independent eigenvectors for $\mathbf{A}$, and the diagonal elements in $\mathbf{\Lambda}$ are the eigenvalues of $\mathbf{A}$ written such that the i-th column of $\mathbf{V}$ is an eigenvector corresponding to the eigenvalue for the i-th column in $\mathbf{\Lambda}$.

We note that this is in agreement with the example-matrices in (14-6) above:

$$\begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -3 \\ 3 \end{bmatrix} = 3\begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{14-7}$$

and

$$\begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 2 \end{bmatrix} = 2\begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{14-8}$$

We see from (14-7) that the first column of $\mathbf{V}$ as expected is an eigenvector for $\mathbf{A}$ corresponding to the first diagonal element in $\mathbf{\Lambda}$, and we see in (14-8) that the second column of $\mathbf{V}$ is an eigenvector corresponding to the second diagonal element in $\mathbf{\Lambda}$.

|||| Theorem 14.7 Diagonalization by Similarity Transformation

If a square $n \times n$-matrix $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ corresponding to the n (not necessarily different) eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, it can be diagonalized by the similarity transformation

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{\Lambda} \quad \Leftrightarrow \quad \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}, \tag{14-9}$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{14-10}$$

If $\mathbf{A}$ does not have n linearly independent eigenvectors, it cannot be diagonalized by a similarity transformation.

|||| Proof

Suppose that $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and $\mathbf{v}_i$ corresponds to the eigenvalue λ_i, for $i = 1 \ldots n$. Then the following equations are valid:

$$\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \quad , \quad \mathbf{A}\mathbf{v}_2 = \lambda_2 \mathbf{v}_2 \quad , \quad \ldots \quad , \quad \mathbf{A}\mathbf{v}_n = \lambda_n \mathbf{v}_n \tag{14-11}$$

The n equations can be gathered in a system of equations:

$$\begin{aligned} &\begin{bmatrix} \mathbf{A}\mathbf{v}_1 & \mathbf{A}\mathbf{v}_2 & \cdots & \mathbf{A}\mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1\lambda_1 & \mathbf{v}_2\lambda_2 & \cdots & \mathbf{v}_n\lambda_n \end{bmatrix} \\ \Leftrightarrow\ &\mathbf{A}\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \\ \Leftrightarrow\ &\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda} \end{aligned} \tag{14-12}$$

Now all the eigenvectors are inserted (vertically one after the other) in the matrix $\mathbf{V}$ in the same order as that of the eigenvalues in the diagonal of the matrix $\mathbf{\Lambda}$ that outside the diagonal contains only zeroes. Since the eigenvectors are linearly independent the matrix $\mathbf{V}$ is invertible. Therefore the inverse $\mathbf{V}^{-1}$ exists, and we multiply by this from the left on both sides of the equality sign:

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda} \ \Leftrightarrow\ \mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{\Lambda}\,. \tag{14-13}$$

Thus the first part of the theorem is proved. Suppose on the contrary that $\mathbf{A}$ can be diagonalized by a similarity transformation. Then an invertible matrix $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix}$ and a diagonal matrix $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots \lambda_n)$ exist such that

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathbf{\Lambda}\,. \tag{14-14}$$

If we now repeat the transformations in the first part of the proof only now in the opposite order, it is seen that (14-14) is the equivalent of the following n equations:

$$\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \quad , \quad \mathbf{A}\mathbf{v}_2 = \lambda_2 \mathbf{v}_2 \quad , \quad \ldots \quad , \quad \mathbf{A}\mathbf{v}_n = \lambda_n \mathbf{v}_n \tag{14-15}$$

from which it appears that $\mathbf{v}_i$ for $i = 1 \ldots n$ is an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_i.

Therefore diagonalization by similarity transformation can only be obtained by the method described in the first part of the theorem.

■

The following theorem can be of great help when one investigates whether matrices can be diagonalized by similarity in different contexts . The main result is already given in Theorem 14.7, but here we refine the conditions by drawing upon previously proven theorems about the eigenvalue problem for linear maps and matrices.

|||| Theorem 14.8 Matrix Diagonalizability

For a given $n \times n$-matrix $\mathbf{A}$ we have:

$\mathbf{A}$ can be diagonalized by a similarity transformation

1. if n different eigenvalues for $\mathbf{A}$ exist.
2. if the sum of the geometric multiplicities of the eigenvalues is n .

$\mathbf{A}$ *cannot* be diagonalized by similarity transformation

3. if the sum of the geometric multiplicities of the eigenvalues is less than n .
4. if an eigenvalue λ with $\text{gm}(\lambda) < \text{am}(\lambda)$ exists.

|||| Proof

Ad. 1. If a proper eigenvector from each of the n eigenspaces is chosen, it follows from Corollary 13.9 that the collected set of n eigenvectors is linearly independent. Therefore, according to Theorem 14.7, $\mathbf{A}$ can be diagonalized by similarity transformation.

Ad. 2: If a basis from each of the eigenspaces is chosen, then the collected set of the chosen n eigenvectors according to Corollary 13.9 is linearly independent. Therefore, according to Theorem 14.7 $\mathbf{A}$ can be diagonalized by similarity transformation .

Ad. 3: If the sum of the geometric multiplicities is less than n, n linearly independent eigenvectors for $\mathbf{A}$ do not exist. Therefore, according to Theorem 14.7 $\mathbf{A}$ cannot be diagonalized by similarity transformation.

Ad. 4: Since according to Theorem 13.34 point 1, the sum of the algebraic multiplicity is less than or equal to n, and since according to the same theorem point 2 for every eigenvalue λ $\text{gm}(\lambda) \leq \text{am}(\lambda)$, the sum of the geometric multiplicities cannot become n, if one of the geometric multiplicities is less than its algebraic one. Therefore, according to what has just been proved, $\mathbf{A}$ cannot be diagonalized by similarity transformation.

■

A typical special case is that of a square $n \times n$-matrix with n different eigenvalues. Theorem 14.8 point 1 guarantees that all matrices of this type can be diagonalized by similarity transformation.

In the following examples we will see how to investigate in practice whether diagonalization by similarity transformation is possible and, if so, carry it through.

|||| Example 14.9

The square matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 3 & 2 \\ 2 & 10 & 4 \\ 2 & 6 & 8 \end{bmatrix} \tag{14-16}$$

has the eigenvalues $\lambda_1 = 4$ and $\lambda_2 = 15$. The vectors $\mathbf{v}_1 = (-2,0,1)$ and $\mathbf{v}_2 = (-3,1,0)$ are linearly independent vectors correponding to λ_1, and the vector $\mathbf{v}_3 = (1,2,2)$ is a proper eigenvector corresponding to λ_2. The collected set of the three eigenvectors is linearly independent according to Corollary 13.9. Therefore, according to Theorem 14.7, it is possible to diagonalize $\mathbf{A}$, because $n = 3$ linearly independent eigenvectors exist. Therefore we can write $\mathbf{\Lambda} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}$, where

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 15 \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} -2 & -3 & 1 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix}. \tag{14-17}$$

14.3 Complex Diagonalization

What we so far have said about similar matrices is generally valid for square, *complex* matrices. Therefore the basic equation for diagonalization by similarity transformation:

$$\mathbf{V}^{-1}\,\mathbf{A}\,\mathbf{V} = \mathbf{\Lambda},$$

will be understood in the broadest sense, where the matrices $\mathbf{A}$, $\mathbf{V}$ and $\mathbf{\Lambda}$ are complex $n \times n$-matrices. Until now we have limited ourselves to real examples, that is examples where it has been possible to satisfy the basic equation (14.3) with real matrices. We will in the following look upon a special situation that is typical in technical applications of diagonalization: For a given *real* $n \times n$ matrix $\mathbf{A}$ we seek an invertible matrix $\mathbf{M}$ and a diagonal matrix $\mathbf{\Lambda}$ satisfying the basic equation in a broad context where $\mathbf{M}$ and $\mathbf{\Lambda}$ possibly are complex (not real) $n \times n$ matrices.

The following example shows a real 3×3 matrix that cannot be diagonalized (with only non-complex entries in the diagonal) because its characteristic polynomial only has one real root. On the other hand it can be diagonalized in a complex sense.

|||| Example 14.10 Complex Diagonalization of a Real Matrix

The square matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 5 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \tag{14-18}$$

has the eigenvalues $\lambda_1 = 2$, $\lambda_2 = -i$ and $\lambda_3 = i$. $\mathbf{v}_1 = (1, 0, 0)$ is a proper eigenvector corresponding to λ_1, $\mathbf{v}_2 = (-2 + i, i, 1)$ is a proper eigenvector corresponding to λ_2, and $\mathbf{v}_3 = (-2 - i, -i, 1)$ is a proper eigenvector belonging to λ_3. The collected set of the three said eigenvectors is linearly independent according to Corallary 13.9. Therefore, according to Theorem 14.7, it is possible to diagonalize $\mathbf{A}$. Therefore we can write $\mathbf{\Lambda} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}$, where

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -i & 0 \\ 0 & 0 & i \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} 1 & -2+i & -2-i \\ 0 & i & -i \\ 0 & 1 & 1 \end{bmatrix}. \tag{14-19}$$

The next example shows a real, square matrix that cannot be diagonalized either in a real or in a complex way.

|||| Example 14.11 Non-Diagonalizable Square Matrix

Given the square matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & -2 \\ 1 & 4 & 1 \\ 0 & 0 & 3 \end{bmatrix}, \tag{14-20}$$

and $\mathbf{A}$ has the eigenvalues $\lambda_1 = 3$ and $\lambda_3 = 5$. The eigenvalue 3 has the algebraic multiplicity 2, but only one linearly independent eigenvector can be chosen, e.g. $\mathbf{v}_1 = (1, -1, 0)$. Thus the eigenvalue has the geometric multiplicity 1. Therefore, according to Theorem 14.7, it is not possible to diagonalize $\mathbf{A}$ by similarity transformation.

|||| Exercise 14.12

For the matrix

$$\mathbf{A}=\begin{bmatrix}2 & 9\\ 1 & -6\end{bmatrix} \tag{14-21}$$

the following should be determined:

1. All eigenvalues and their algebraic multiplicities.
2. All corresponding linearly independent eigenvectors and thus the geometric multiplicities of the eigenvectors.
3. If possible, $\mathbf{A}$ is to be diagonalized: Determine a diagonal matrix $\boldsymbol{\Lambda}$ and an invertible matrix $\mathbf{V}$ for which $\mathbf{V}^{-1}\mathbf{A}\mathbf{V}=\boldsymbol{\Lambda}$. What are the requirements for the diagonalization to be carried through? Which numbers and vectors are used in $\boldsymbol{\Lambda}$ and $\mathbf{V}$?

14.4 Diagonalization of Linear Maps

In the introduction to this eNote we asked the question: What conditions should be satisfied so that two given square matrices can be interpreted as mapping matrices for the same linear map with respect to two different bases? The answer is simple:

|||| Theorem 14.13 Similar Matrices as Mapping Matrices

An n-dimensional vector space V is given. Two $n\times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are mapping matrices for the same linear map $f: V\to V$ with respect to two different bases for V if and only if $\mathbf{A}$ and $\mathbf{B}$ are similar.

|||| Exercise 14.14

Prove Theorem 14.13

In the introduction we also asked the question: Which conditions should a square matrix satisfy in order to be a mapping matrix for a linear map that in another basis has a diagonal matrix as a mapping matrix? The answer appears from Theorem 14.7 combined with Theorem 14.13: the matrix must have n linearly independent eigenvectors.

We end the eNote by an example on diagonalization of a linear map, that is finding a suitable basis in which the mapping matrix is diagonal.

|||| Example 14.15 Diagonalization of a Linear Map

A linear map $f : P_1(\mathbb{R}) \to P_1(\mathbb{R})$ is given by the following mapping matrix with respect to the standard monomial basis m:

$$ {}_{\mathrm{m}}\mathbf{F}_{\mathrm{m}} = \begin{bmatrix} -17 & -21 \\ 14 & 18 \end{bmatrix} \tag{14-22}$$

This means that $f(1) = -17 + 14x$ and $f(x) = -21 + 18x$. We wish to investigate whether a (real) eigenbasis for f can be found and if so, how the mapping matrix looks with respect to this basis, and what the basis vectors are.

The eigenvalues of ${}_{\mathrm{m}}\mathbf{F}_{\mathrm{m}}$ are determined:

$$\det\left(\begin{bmatrix} -17-\lambda & -21 \\ 14 & 18-\lambda \end{bmatrix}\right) = \lambda^2 - \lambda - 12 = (\lambda+3)(\lambda-4) = 0\,. \tag{14-23}$$

It is already now possible to confirm that a real eigenbasis for f exists since $2 = \dim(P_2(\mathbb{R}))$, viz. $\lambda_1 = -3$ and $\lambda_2 = 4$ each with the algebraic multiplicity 1. Eigenvectors corresponding to λ_1 are determined:

$$\left[\begin{array}{cc|c} -17+3 & -21 & 0 \\ 14 & 18+3 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & \frac{3}{2} & 0 \\ 0 & 0 & 0 \end{array}\right]. \tag{14-24}$$

This yields an eigenvector ${}_{\mathrm{m}}\mathbf{v}_1 = (-3, 2)$, if the free parameter is put equal to 2. Similarly we get the other eigenvector:

$$\left[\begin{array}{cc|c} -17-4 & -21 & 0 \\ 14 & 18-4 & 0 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]. \tag{14-25}$$

This yields an eigenvector ${}_{\mathrm{m}}\mathbf{v}_2 = (-1, 1)$, if the free parameter is put equal to 1.

Thus a real eigenbasis v for f, given by the basis vectors ${}_{\mathrm{m}}\mathbf{v}_1$ and ${}_{\mathrm{m}}\mathbf{v}_2$, exists. We then get

$$ {}_{\mathrm{m}}\mathbf{M}_{\mathrm{v}} = \begin{bmatrix} -3 & -1 \\ 2 & 1 \end{bmatrix} \quad \text{and} \quad {}_{\mathrm{v}}\mathbf{F}_{\mathrm{v}} = \begin{bmatrix} -3 & 0 \\ 0 & 4 \end{bmatrix} \tag{14-26}$$

The basis consists of the vectors $\mathbf{v}_1 = -3 + 2x$ and $\mathbf{v}_2 = -1 + x$ and the map is "simple" with respect to this basis.

One can check with the map of $\mathbf{v}_1$:

$$\begin{aligned} f(\mathbf{v}_1) &= f(-3+2x) = -3\cdot f(1) + 2\cdot f(x) \\ &= -3\cdot(-17+14x) + 2\cdot(-21+18x) \\ &= 9-6x = -3(-3+2x) = -3\mathbf{v}_1 \end{aligned} \tag{14-27}$$

It is true!

eNote 15

Symmetric Matrices

In this eNote we will consider one of the most used results from linear algebra – the so-called spectral theorem *for symmetric matrices. In short it says that* all *symmetric matrices can be* diagonalized *by a similarity transformation – that is, by change of basis with a suitable substitution matrix.*
The introduction of these concepts and the corresponding method were given in eNotes 10, 13 and 14, which therefore is a necessary basis for the present eNote.
Precisely in that eNote it became clear that not all matrices can be diagonalized. Diagonalization requires a sufficiently large number of eigenvalues (the algebraic multiplicities add up to be as large as possible) and that the corresponding eigenvector spaces actually span all of the vector space (the geometric multiplicities add up to be as large as possible). It is these properties we will consider here, but now for symmetric matrices, which turn out to satisfy the conditions and actually more: the eigenvectors we use in the resulting substitution matrix can be chosen pairwise orthogonal, such that the new basis is the result of a rotation of the old standard basis in $\mathbb{R}^n$.
In order to be able to discuss and apply the spectral theorem most effectively we must first introduce a natural scalar product for vectors in $\mathbb{R}^n$ *in such a way that we will be able to measure angles and lengths in all dimensions. We do this by generalizing the well-known standard scalar product from* $\mathbb{R}^2$ *and* $\mathbb{R}^3$*. As indicated above we will in particular use bases consisting of pairwise orthogonal vectors in order to formulate the spectral theorem, understand it and what use we can make of this important theorem.*

Updated: 20.11.21 David Brander

15.1 Scalar Product

In the vector space $\mathbb{R}^n$ we introduce an inner product, i.e. a scalar product that is a natural generalization of the well-known scalar product from plane geometry and space geometry, see eNote 10.

|||| Definition 15.1 Scalar Product

Let **a** and **b** be two given vectors in $\mathbb{R}^n$ with the coordinates $(a_1, ..., a_n)$ and $(b_1, ..., b_n)$, respectively, with respect to the standard basis shalle in $\mathbb{R}^n$:

$$ {}_{\mathrm{e}}\mathbf{a} = (a_1, ..., a_n) \quad , \quad \text{and} \quad {}_{\mathrm{e}}\mathbf{b} = (b_1, ..., b_n) \quad . \tag{15-1}$$

Then we define the ***scalar product***, the ***inner product***, (also called the *dot product*) of the two vectors in the following way:

$$ \mathbf{a}\boldsymbol{\cdot}\mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots a_n b_n = \sum_{i=1}^{n} a_i b_i \quad . \tag{15-2}$$

When $\mathbb{R}^n$ is equipped with this scalar product $(\mathbb{R}^n, \boldsymbol{\cdot})$ is thereby an example of a so-called ***Euclidian vector space*** , or a ***vector space with inner product***.

The scalar product can be expressed as a matrix product:

$$ \mathbf{a} \cdot \mathbf{b} = {}_{\mathrm{e}}\mathbf{a}^{\top} \cdot {}_{\mathrm{e}}\mathbf{b} = \left[\begin{array}{cccc} a_1 & \cdot & \cdot & \cdot & a_n \end{array}\right] \cdot \left[\begin{array}{c} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{array}\right] \tag{15-3}$$

For the scalar product introduced above the following arithmetic rules apply:

|||| Theorem 15.2 Arithmetic Rules for the Scalar Product

If $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are vectors in $(\mathbb{R}^n, \cdot)$ and k is an arbitrary real number then:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \tag{15-4}$$

$$\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} \tag{15-5}$$

$$\mathbf{a} \cdot (k\mathbf{b}) = (k\mathbf{a}) \cdot \mathbf{b} = k(\mathbf{a} \cdot \mathbf{b}) \tag{15-6}$$

A main point about the introduction of a scalar product is that we can now talk about the *lengths of the vectors* in $(\mathbb{R}^n, \cdot)$:

|||| Definition 15.3 The Length of a Vector

Let $\mathbf{a}$ be a vector in $(\mathbb{R}^n, \cdot)$ with the coordinates $(a_1, ..., a_n)$ with respect to the standard e-basis in $\mathbb{R}^n$. Then the ***length of* a** is defined by

$$|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{\sum_{i=1}^{n} a_i^2} \quad . \tag{15-7}$$

The length of $\mathbf{a}$ is also called the ***norm of* a** with respect to the scalar product in $(\mathbb{R}^n, \cdot)$. A vector $\mathbf{a}$ is called a ***proper vector*** if $|\mathbf{a}| > 0$.

It follows from Definition 15.1 that

$$\begin{aligned} &\mathbf{a}\cdot\mathbf{a} \geq 0 \;\text{ for all }\; \mathbf{a} \in (\mathbb{R}^n, \cdot) \;\text{ and} \\ &\mathbf{a}\cdot\mathbf{a} = 0 \Leftrightarrow \mathbf{a} = \mathbf{0}\,. \end{aligned} \tag{15-8}$$

From this we immediately see that

$$\begin{aligned} &|\mathbf{a}| \geq 0, \;\text{ for all }\; \mathbf{a} \in (\mathbb{R}^n, \cdot) \;\text{ and} \\ &|\mathbf{a}| = 0 \Leftrightarrow \mathbf{a} = \mathbf{0}\,. \end{aligned} \tag{15-9}$$

Thus a *proper* vector is a vector that is not the **0**-vector.

Finally it follows from Definition 15.1 and Definition 15.3 that for $\mathbf{a} \in (\mathbb{R}^n, \cdot)$ and an arbitrary real number k we have that

$$|k\mathbf{a}| = |k| \; |\mathbf{a}| \;. \tag{15-10}$$

We can now prove the following important theorem:

|||| Theorem 15.4 Cauchy-Schwarz Inequality

For arbitrary vectors **a** and **b** in $(\mathbb{R}^n, \cdot)$

$$|\mathbf{a}\cdot\mathbf{b}| \leq |\,\mathbf{a}\,| \; |\,\mathbf{b}\,| \;. \tag{15-11}$$

Equality holds if and only if **a** and **b** are linearly dependent.

|||| Proof

If $\mathbf{b} = \mathbf{0}$, both sides of (15-11) are equal to 0 and the inequality is thereby satisfied. We now assume that **b** is a proper vector.

We put $k = \mathbf{b}\cdot\mathbf{b}$ and $\mathbf{e} = \dfrac{1}{\sqrt{k}}\,\mathbf{b}$. It then follows from (15-6) that

$$\mathbf{e}\cdot\mathbf{e} = \Big(\frac{1}{\sqrt{k}}\mathbf{b}\Big)\cdot\Big(\frac{1}{\sqrt{k}}\mathbf{b}\Big) = \frac{1}{k}\,(\mathbf{b}\cdot\mathbf{b}) = 1$$

and thereby that $|\mathbf{e}| = 1$.

By substituting $\mathbf{b} = \sqrt{k}\,\mathbf{e}$ in the left hand side and the right hand side of (15-11) we get using

(15-6) and (15-10):

$$|\mathbf{a}\cdot\mathbf{b}| = |\,\mathbf{a}\cdot(\sqrt{k}\,\mathbf{e})\,| = \sqrt{k}\,|\mathbf{a}\cdot\mathbf{e}|$$

and

$$|\mathbf{a}|\ |\mathbf{b}| = |\mathbf{a}|\ |\,\sqrt{k}\,\mathbf{e}\,| = \sqrt{k}\,|\mathbf{a}|\ |\mathbf{e}|\ .$$

Therefore we only have to show that for arbitrary $\mathbf{a}$ and $\mathbf{e}$, where $|\mathbf{e}| = 1$,

$$|\mathbf{a}\cdot\mathbf{e}| \leq |\,\mathbf{a}\,| \qquad (15\text{-}12)$$

where equality holds if and only if $\mathbf{a}$ and $\mathbf{e}$ are linearly dependent.

For an arbitrary $t \in \mathbb{R}$ it follows from (15-8), (15-5) and (15-6) that:

$$0 \leq (\mathbf{a}-t\mathbf{e})\cdot(\mathbf{a}-t\mathbf{e}) = \mathbf{a}\cdot\mathbf{a} + t^2(\mathbf{e}\cdot\mathbf{e}) - 2t(\mathbf{a}\cdot\mathbf{e}) = \mathbf{a}\cdot\mathbf{a} + t^2 - 2t(\mathbf{a}\cdot\mathbf{e})\,.$$

If in particular we choose $t = \mathbf{a}\cdot\mathbf{e}$, we get

$$0 \leq \mathbf{a}\cdot\mathbf{a} - (\mathbf{a}\cdot\mathbf{e})^2 \;\Leftrightarrow\; |\mathbf{a}\cdot\mathbf{e}| \leq \sqrt{\mathbf{a}\cdot\mathbf{a}} = |\mathbf{a}|\ .$$

Since it follows from (15-8) that $(\mathbf{a}-t\mathbf{e})\cdot(\mathbf{a}-t\mathbf{e}) = 0$ if and only if $(\mathbf{a}-t\mathbf{e}) = \mathbf{0}$, we see that $|\mathbf{a}\cdot\mathbf{e}| = |\,\mathbf{a}\,|$ if and only if $\mathbf{a}$ and $\mathbf{e}$ are linearly dependent. The proof is hereby complete.

■

From the Cauchy-Schwarz inequality follows the triangle inequality that is a generalization of the well-known theorem from elementary plane geometry, that a side in a triangle is always less than or equal to the sum of the other sides:

|||| Corollary 15.5 The Triangle Inequality

For arbitrary vectors $\mathbf{a}$ and $\mathbf{b}$ in $(\mathbb{R}^n, \cdot)$

$$|\mathbf{a}+\mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}|\ . \qquad (15\text{-}13)$$

|||| Exercise 15.6

Prove Corollary 15.5.

Note that from the Cauchy-Schwarz inequality it follows that:

$$-1 \leq \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| \cdot |\mathbf{b}|} \leq 1 \quad . \tag{15-14}$$

Therefore the *angle between two vectors* in $(\mathbb{R}^n, \cdot)$ can be introduced as follows:

|||| Definition 15.7 The Angle Between Vectors

Let $\mathbf{a}$ and $\mathbf{b}$ be two given proper vectors in $(\mathbb{R}^n, \cdot)$ with the coordinates $(a_1, ..., a_n)$ and $(b_1, ..., b_n)$ with respect to the standard basis in $(\mathbb{R}^n, \cdot)$. Then the ***angle between* a *and* b** is defined as the value θ in interval $[0, \pi]$ that satisfies

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| \cdot |\mathbf{b}|} \quad . \tag{15-15}$$

If $\mathbf{a} \cdot \mathbf{b} = 0$ we say that the two proper vectors are ***orthogonal*** or ***perpendicular*** with respect to each other. This occurs exactly when $\cos(\theta) = 0$, that is, when $\theta = \pi/2$.

15.2 Symmetric Matrices and the Scalar Product

We know the symmetry concept from square matrices:

|||| Definition 15.8

A square matrix $\mathbf{A}$ is ***symmetric*** if it is equal to its own transpose

$$\mathbf{A} = \mathbf{A}^\top \quad , \tag{15-16}$$

that is if $a_{ij} = a_{j\,i}$ for all elements in the matrix.

What is the relation between symmetric matrices and the scalar product? This we consider here:

|||| Theorem 15.9

Let $\mathbf{v}$ and $\mathbf{w}$ denote two vectors in the vector space $(\mathbb{R}^n, \cdot)$ with scalar product introduced above. If $\mathbf{A}$ is an arbitrary $(n \times n)-$matrix then

$$(\mathbf{A}\,\mathbf{v})\,\cdot \mathbf{w} = \mathbf{v} \cdot \left(\mathbf{A}^\top\,\mathbf{w}\right) \quad . \tag{15-17}$$

|||| Proof

We use the fact that the scalar product can be expressed as a matrix product:

$$\begin{aligned}(\mathbf{A}\,\mathbf{v})\cdot\mathbf{w} &= (\mathbf{A}\,\mathbf{v})^\top\cdot\mathbf{w}\\ &= \left(\mathbf{v}^\top\mathbf{A}^\top\right)\cdot\mathbf{w}\\ &= \mathbf{v}^\top\cdot\left(\mathbf{A}^\top\mathbf{w}\right)\\ &= \mathbf{v}\cdot\left(\mathbf{A}^\top\mathbf{w}\right) \quad .\end{aligned} \tag{15-18}$$

■

This we can now use to characterize symmetric matrices:

|||| Theorem 15.10

A matrix $\mathbf{A}$ is a symmetric $(n \times n)-$matrix if and only if

$$(\mathbf{A}\,\mathbf{v})\,\cdot \mathbf{w} = \mathbf{v} \cdot (\mathbf{A}\,\mathbf{w}) \tag{15-19}$$

for all vectors $\mathbf{v}$ and $\mathbf{w}$ in $(\mathbb{R}^n, \cdot)$.

|||| Proof

If $\mathbf{A}$ is symmetric then we have that $\mathbf{A} = \mathbf{A}^\top$ and therefore Equation (15-19) follows directly from Equation (15-17). Conversely, if we assume that (15-19) applies for all $\mathbf{v}$ and $\mathbf{w}$, we will prove that $\mathbf{A}$ is symmetric. But this follows easily just by *choosing* suitable vectors, e.g. $\mathbf{v} = \mathbf{e}_2 = (0, 1, 0, ..., 0)$ and $\mathbf{w} = \mathbf{e}_3 = (0, 0, 1, ..., 0)$ and substitute these into (15-19) as seen below. Note that $\mathbf{A}\,\mathbf{e}_i$ is the i^{th} column vector in $\mathbf{A}$.

$$\begin{aligned}(\mathbf{A}\,\mathbf{e}_2)\cdot\mathbf{e}_3 &= a_{23}\\ &= \mathbf{e}_2\cdot(\mathbf{A}\,\mathbf{e}_3)\\ &= (\mathbf{A}\,\mathbf{e}_3)\cdot\mathbf{e}_2\\ &= a_{32}\quad ,\end{aligned} \tag{15-20}$$

such that $a_{23} = a_{32}$. Quite similarly for all other choices of indices i and j we get that $a_{ij} = a_{j\,i}$ – and this is what we had set out to prove.

■

A basis a in $(\mathbb{R}^n, \cdot)$ consists (as is known from eNote 11) of n linearly independent vectors $(\mathbf{a}_1, ..., \mathbf{a}_n)$. If in addition the vectors are pairwise orthogonal and have length 1 with respect to the scalar product, then $(\mathbf{a}_1, ..., \mathbf{a}_n)$ is an ***orthonormal basis for*** $(\mathbb{R}^n, \cdot)$:

|||| Definition 15.11

A basis $a = (\mathbf{a}_1, ..., \mathbf{a}_n)$ is an *orthonormal basis* if

$$\mathbf{a}_i \cdot \mathbf{a}_j = \begin{cases} 1 & \text{for } i = j \quad , \\ 0 & \text{for } i \neq j \quad . \end{cases} \tag{15-21}$$

|||| Exercise 15.12

Show that if n vectors $(\mathbf{a}_1, ..., \mathbf{a}_n)$ in $(\mathbb{R}^n, \cdot)$ satisfy Equation (15-21) then $a = (\mathbf{a}_1, ..., \mathbf{a}_n)$ is automatically a *basis* for $(\mathbb{R}^n, \cdot)$, i.e. the vectors are linearly independent and span all of $(\mathbb{R}^n, \cdot)$.

|||| Exercise 15.13

Show that the following 3 vectors $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ constitute an orthonormal basis for $(\mathbb{R}^3, \cdot)$ for any given value of $\theta \in \mathbb{R}$:

$$\begin{aligned} \mathbf{a}_1 &= (\cos(\theta), 0, -\sin(\theta)) \\ \mathbf{a}_2 &= (0, 1, 0) \\ \mathbf{a}_3 &= (\sin(\theta), 0, \cos(\theta)) \quad . \end{aligned} \tag{15-22}$$

If we put the vectors from an orthonormal basis into a matrix as columns we get an ***orthogonal matrix***:

|||| Definition 15.14

An $(n \times n)-$matrix $\mathbf{A}$ is said to be *orthogonal* if the column vectors in $\mathbf{A}$ constitute an orthonormal basis for $(\mathbb{R}^n, \cdot)$, that is if the column vectors are pairwise orthogonal and all have length 1 – as is also expressed in Equation (15-21).

Note that *orthogonal matrices* alternatively (and maybe also more descriptively) could be called *orthonormal*, since the columns in the matrix are not only pairwise orthogonal but also normalized such that they all have length 1. We will follow international tradition and call the matrices orthogonal.

It is easy to check whether a given matrix is orthogonal:

|||| Theorem 15.15

An $(n \times n)-$matrix $\mathbf{Q}$ is orthogonal if and only if

$$\mathbf{Q}^\top \cdot \mathbf{Q} = \mathbf{E}_{n \times n} \quad , \tag{15-23}$$

which is equivalent to

$$\mathbf{Q}^\top = \mathbf{Q}^{-1} \quad . \tag{15-24}$$

|||| Proof

See eNote 7 about the computation of the matrix product and then compare with the condition for orthogonality of the column vectors in $\mathbf{Q}$ (Equation (15-21)).

■

We can now explain the geometric significance of an orthogonal matrix: as a linear map it *preserves lengths of, and angles between, vectors*. That is the content of the following theorem, which follows immediately from Theorems 15.9 and Theorem 15.15:

|||| Theorem 15.16

An $n \times n$ matrix $\mathbf{A}$ is orthogonal if and only if the linear mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ given by $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ preserves the scalar product, i.e.:

$$(\mathbf{A}\mathbf{x})\cdot(\mathbf{A}\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}, \qquad \text{for any } \mathbf{x},\, \mathbf{y} \in \mathbb{R}^n.$$

Orthogonal matrices are *regular* and have determinant ±1:

|||| Exercise 15.17

Show that for a matrix $\mathbf{A}$ to be orthogonal, it is necessary that

$$|\det(\mathbf{A})| = 1 \quad . \tag{15-25}$$

Show that this condition is not sufficient, thus matrices exist that satisfy this determinant-condition but that are not orthogonal.

|||| Definition 15.18

An orthogonal matrix $\mathbf{Q}$ is called ***special orthogonal*** or ***positive orthogonal*** if $\det(\mathbf{Q}) = 1$ and it is called ***negative orthogonal*** if $\det(\mathbf{Q}) = -1$.

In the literature, orthogonal matrices with determinant 1 are called special orthogonal, and those with determinant -1 are usually not given a name.

|||| Exercise 15.19

Given the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & -a & 0 & a \\ a & 0 & a & 0 \\ 0 & -a & 0 & -a \\ -a & 0 & a & 0 \end{bmatrix} \quad , \quad \text{with } a \in \mathbb{R} \quad . \tag{15-26}$$

Determine the values of a for which $\mathbf{A}$ is orthogonal and state in every case whether $\mathbf{A}$ is positive orthogonal or negative orthogonal.

15.3 Gram–Schmidt Orthonormalization

Here we describe a procedure for determining an orthonormal basis for a subspace of the vector space $(\mathbb{R}^n, \cdot)$. Let U be a $p-$dimensional subspace of $(\mathbb{R}^n, \cdot)$; we assume that U is spanned by p given linearly independent vectors $(\mathbf{u}_1, \cdots, \mathbf{u}_p)$, constituting a basis u for U. Gram–Schmidt orthonormalization aims at constructing a new basis v $= (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$ for the subspace of U from the given basis u such that the new vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$ are *pairwise orthogonal and have length* 1.

|||| Method 15.20 Gram–Schmidt Orthonormalization

Orthonormalization of p linearly independent vectors $\mathbf{u}_1, \cdots, \mathbf{u}_p$ in $(\mathbb{R}^n, \cdot)$:

1. Start by normalizing $\mathbf{u}_1$ and call the result $\mathbf{v}_1$, i.e.:

$$\mathbf{v}_1 = \frac{\mathbf{u}_1}{|\mathbf{u}_1|} \quad . \tag{15-27}$$

2. The next vector $\mathbf{v}_2$ in the basis v is now chosen in span$\{\mathbf{u}_1, \mathbf{u}_2\}$ but such that at the same time $\mathbf{v}_2$ is orthogonal to $\mathbf{v}_1$, i.e. $\mathbf{v}_2 \cdot \mathbf{v}_1 = 0$; finally this vector is normalized. First we construct an *auxiliary vector* $\mathbf{w}_2$.

$$\begin{aligned} \mathbf{w}_2 &= \mathbf{u}_2 - (\mathbf{u}_2 \cdot \mathbf{v}_1)\, \mathbf{v}_1 \\ \mathbf{v}_2 &= \frac{\mathbf{w}_2}{|\mathbf{w}_2|} \quad . \end{aligned} \tag{15-28}$$

Note that $\mathbf{w}_2$ (and therefore also $\mathbf{v}_2$) then being orthogonal to $\mathbf{v}_1$:

$$\begin{aligned} \mathbf{w}_2 \cdot \mathbf{v}_1 &= (\mathbf{u}_2 - (\mathbf{u}_2 \cdot \mathbf{v}_1)\, \mathbf{v}_1) \cdot \mathbf{v}_1 \\ &= \mathbf{u}_2 \cdot \mathbf{v}_1 - (\mathbf{u}_2 \cdot \mathbf{v}_1)\, \mathbf{v}_1 \cdot \mathbf{v}_1 \\ &= \mathbf{u}_2 \cdot \mathbf{v}_1 - (\mathbf{u}_2 \cdot \mathbf{v}_1)\, |\mathbf{v}_1|^2 \\ &= \mathbf{u}_2 \cdot \mathbf{v}_1 - (\mathbf{u}_2 \cdot \mathbf{v}_1) \\ &= 0 \quad . \end{aligned} \tag{15-29}$$

3. We continue in this way

$$\begin{aligned} \mathbf{w}_i &= \mathbf{u}_i - (\mathbf{u}_i \cdot \mathbf{v}_1)\, \mathbf{v}_1 - (\mathbf{u}_i \cdot \mathbf{v}_2)\, \mathbf{v}_2 - \cdots - (\mathbf{u}_i \cdot \mathbf{v}_{i-1})\, \mathbf{v}_{i-1} \\ \mathbf{v}_i &= \frac{\mathbf{w}_i}{|\mathbf{w}_i|} \quad . \end{aligned} \tag{15-30}$$

4. Until the last vector $\mathbf{u}_p$ is used:

$$\begin{aligned} \mathbf{w}_p &= \mathbf{u}_p - \left(\mathbf{u}_p \cdot \mathbf{v}_1\right) \mathbf{v}_1 - \left(\mathbf{u}_p \cdot \mathbf{v}_2\right) \mathbf{v}_2 - \cdots - \left(\mathbf{u}_p \cdot \mathbf{v}_{p-1}\right) \mathbf{v}_{p-1} \\ \mathbf{v}_p &= \frac{\mathbf{w}_p}{|\mathbf{w}_p|} \quad . \end{aligned} \tag{15-31}$$

The constructed v-vectors span the same subspace U as the given linearly independent u-vectors, $U = \text{span}\{\mathbf{u}_1, \cdots, \mathbf{u}_p\} = \text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ and $\text{v} = (\mathbf{v}_1, \cdots, \mathbf{v}_p)$ constituting an orthonormal basis for U.

|||| Example 15.21

In $(\mathbb{R}^4, \cdot)$ we will by the use of the Gram–Schmidt orthonormalization method find an orthonormal basis $\text{v} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ for the $3-$dimensional subspace U that is spanned by the three given linearly independent (!) vectors having the following coordinates with respect to the standard e-basis in $\mathbb{R}^4$:

$$\mathbf{u}_1 = (2,2,4,1) \quad , \quad \mathbf{u}_2 = (0,0,-5,-5) \quad , \quad \mathbf{u}_3 = (5,3,3,-3) \quad .$$

We construct the new basis vectors with respect to the standard e-basis in $\mathbb{R}^4$ by working through the orthonormalization procedure. There are 3 'steps' since there are in this example 3 linearly independent vectors in U :

1.
$$\mathbf{v}_1 = \frac{\mathbf{u}_1}{|\mathbf{u}_1|} = \frac{1}{5}(2,2,4,1) \quad . \tag{15-32}$$

2.
$$\begin{aligned} \mathbf{w}_2 &= \mathbf{u}_2 - (\mathbf{u}_2 \cdot \mathbf{v}_1)\, \mathbf{v}_1 = \mathbf{u}_2 + 5\mathbf{v}_1 = (2,2,-1,-4) \\ \mathbf{v}_2 &= \frac{\mathbf{w}_2}{|\mathbf{w}_2|} = \frac{1}{5}(2,2,-1,-4) \quad . \end{aligned} \tag{15-33}$$

3.
$$\begin{aligned} \mathbf{w}_3 &= \mathbf{u}_3 - (\mathbf{u}_3 \cdot \mathbf{v}_1)\, \mathbf{v}_1 - (\mathbf{u}_3 \cdot \mathbf{v}_2)\, \mathbf{v}_2 = \mathbf{u}_3 - 5\mathbf{v}_1 - 5\mathbf{v}_2 = (1,-1,0,0) \\ \mathbf{v}_3 &= \frac{\mathbf{w}_3}{|\mathbf{w}_3|} = \frac{1}{\sqrt{2}}(1,-1,0,0) \quad . \end{aligned} \tag{15-34}$$

Thus we have constructed an orthonormal basis for the subspace U consisting of those vectors that with respect to the standard basis have the coordinates:

$$\mathbf{v}_1 = \frac{1}{5} \cdot (2,2,4,1) \quad , \quad \mathbf{v}_2 = \frac{1}{5} \cdot (2,2,-1,-4) \quad , \quad \mathbf{v}_3 = \frac{1}{\sqrt{2}} \cdot (1,-1,0,0) \quad .$$

We can check that this is really an orthonormal basis by posing the vectors as columns in a matrix, which then is of the type (4×3). Like this:

$$\mathbf{V} = \begin{bmatrix} 2/5 & 2/5 & 1/\sqrt{2} \\ 2/5 & 2/5 & -1/\sqrt{2} \\ 4/5 & -1/5 & 0 \\ 1/5 & -4/5 & 0 \end{bmatrix} \tag{15-35}$$

The matrix $\mathbf{V}$ cannot be an orthogonal matrix (because of the type), but nevertheless $\mathbf{V}$ can satisfy the following equation, which shows that the three new basis vectors indeed are pairvise orthogonal and all have length 1 !

$$\mathbf{V}^{\top} \cdot \mathbf{V} = \begin{bmatrix} 2/5 & 2/5 & 4/5 & 1/5 \\ 2/5 & 2/5 & -1/5 & -4/5 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2/5 & 2/5 & 1/\sqrt{2} \\ 2/5 & 2/5 & -1/\sqrt{2} \\ 4/5 & -1/5 & 0 \\ 1/5 & -4/5 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad . \tag{15-36}$$

|||| Exercise 15.22

In $(\mathbb{R}^4, \cdot)$ the following vectors are given with respect to the standard basis e:

$$\mathbf{u}_1 = (1,1,1,1) \quad , \quad \mathbf{u}_2 = (3,1,1,3) \quad , \quad \mathbf{u}_3 = (2,0,-2,4) \quad , \quad \mathbf{u}_4 = (1,1,-1,3) \quad .$$

We let U denote the subspace in $(\mathbb{R}^4, \cdot)$ that is spanned by the four given vectors, that is

$$U = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\} \quad . \tag{15-37}$$

1. Show that u $= (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ is a basis for U and find coordinates for $\mathbf{u}_4$ with respect to this basis.
2. State an orthonormal basis for U.

|||| Example 15.23

In $(\mathbb{R}^3, \cdot)$ a given first unit vector $\mathbf{v}_1$ is required for the new orthonormal basis v$= (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ and the task is to find the two other vectors in the basis. Let us assume that the given vector is $\mathbf{v}_1 = (3,0,4)/5$. We see immediately that e.g. $\mathbf{v}_2 = (0,1,0)$ is a unit vector that is orthogonal to $\mathbf{v}_1$. A last vector for the orthonormal basis can then be found directly using the *cross product*: $\quad \mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2 = \frac{1}{5} \cdot (-4,0,3)$.

15.4 The Orthogonal Complement to a Subspace

Let U be a subspace in $(\mathbb{R}^n, \cdot)$ that is spanned by p given linearly independent vectors, $U = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p\}$. The set of those vectors in $(\mathbb{R}^n, \cdot)$ that are all orthogonal to all vectors in U is itself a subspace of $(\mathbb{R}^n, \cdot)$, and it has the dimension $n - p$:

|||| Definition 15.24

The ***orthogonal complement*** to a subspace U of $(\mathbb{R}^n, \cdot)$ is denoted $U^\perp$ and consists of all vectors in $(\mathbb{R}^n, \cdot)$ that are orthogonal to all vectors in U:

$$U^\perp = \{ \mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x} \cdot \mathbf{u} = 0 \ , \text{ for all } \ \mathbf{u} \in U \} \quad . \tag{15-38}$$

|||| Theorem 15.25

The orthogonal complement $U^\perp$ to a given $p-$dimensional subspace U of $(\mathbb{R}^n, \cdot)$ is itself a subspace in $(\mathbb{R}^n, \cdot)$ and it has dimension $\dim(U^\perp) = n - p$.

|||| Proof

It is easy to check all subspace-properties for $U^\perp$; it is clear that if $\mathbf{a}$ and $\mathbf{b}$ is orthogonal to all vectors in U and k is a real number, then $\mathbf{a} + k\mathbf{b}$ are also orthogonal to all vectors in U. Since the only vector that is orthogonal to itself is $\mathbf{0}$ this is also the only vector in the intersection: $U \cap U^\perp = \{\mathbf{0}\}$. If we let v$= (\mathbf{v}_1, \cdots, \mathbf{v}_p)$ denote an *orthonormal basis* for U and w$= (\mathbf{w}_1, \cdots, \mathbf{w}_r)$ an orthonormal basis for $U^\perp$, then $(\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{w}_1, \cdots, \mathbf{w}_r)$ is an orthonormal basis for the subspace $S = \text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{w}_1, \cdots, \mathbf{w}_r\}$ in $(\mathbb{R}^n, \cdot)$. If we now assume that S is not all of $(\mathbb{R}^n, \cdot)$, then the basis for S can be extended with at least one vector such that the extended system is linearly independent in $(\mathbb{R}^n, \cdot)$; by this we get - through the last step in the Gram–Schmidt method - a new vector that is orthogonal to all vectors in U but which is not an element in $U^\perp$; and thus we get a contradiction, since $U^\perp$ are defined to be *all* those vectors in $(\mathbb{R}^n, \cdot)$ that are orthogonal to every vector in U. Therefore the assumption that S is not all of $(\mathbb{R}^n, \cdot)$ is wrong. I.e. $S = \mathbb{R}^n$ and therefore $r + p = n$, such that $\dim(U^\perp) = r = n - p$; and this is what we had to prove.

■

|||| Example 15.26

The orthogonal complement to $U = \text{span}\{\mathbf{a}, \mathbf{b}\}$ in $\mathbb{R}^3$ (for linearly independent vectors – and therefore proper vectors – $\mathbf{a}$ and $\mathbf{b}$) is $U^\perp = \text{span}\{\mathbf{a} \times \mathbf{b}\}$.

Exercise 15.27

Determine the orthogonal complement to the subspace $U = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in $(\mathbb{R}^4, \cdot)$, when the spanning vectors are given by their respective coordinates with respect to the standard basis e in $\mathbb{R}^4$ as:

$$\mathbf{u}_1 = (1,1,1,1) \quad , \quad \mathbf{u}_2 = (3,1,1,3) \quad , \quad \mathbf{u}_3 = (2,0,-2,4) \quad . \tag{15-39}$$

15.5 The Spectral Theorem for Symmetric Matrices

We will now start to formulate the spectral theorem and start with the following non-trivial observation about symmetric matrices:

Theorem 15.28

Let $\mathbf{A}$ denote a symmetric $(n \times n)-$matrix. Then the characteristic polynomial $\mathcal{K}_{\mathbf{A}}(\lambda)$ for $\mathbf{A}$ has exactly n real roots (counted with multiplicity):

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \quad . \tag{15-40}$$

I.e. $\mathbf{A}$ has n real eigenvalues (counted with multiplicity).

If e.g. $\{7,3,3,2,2,2,1\}$ are the roots of $\mathcal{K}_{\mathbf{A}}(\lambda)$ for a $(7 \times 7)-$matrix $\mathbf{A}$, then these roots must be represented *with their respective multiplicity* in the eigenvalue-list:

$$\lambda_1 = 7 \geq \lambda_2 = 3 \geq \lambda_3 = 3 \geq \lambda_4 = 2 \geq \lambda_5 = 2 \geq \lambda_6 = 2 \geq \lambda_7 = 1 \quad .$$

Since Theorem 15.28 expresses a decisive property about symmetric matrices, we will here give a proof of the theorem:

|||| Proof

From the fundamental theorem of algebra we know that $\mathcal{K}_{\mathbf{A}}(\lambda)$ has exactly n complex roots - but we do not know whether the roots are real; this is what we will prove. So we let $\alpha + i\,\beta$ be a complex root of $\mathcal{K}_{\mathbf{A}}(\lambda)$ and we will then show that $\beta = 0$. Note that α and β naturally both are real numbers.

Therefore we have

$$\det\left(\mathbf{A} - (\alpha + i\,\beta)\mathbf{E}\right) = 0 \quad , \tag{15-41}$$

and thus also that

$$\det\left(\mathbf{A} - (\alpha + i\,\beta)\mathbf{E}\right) \cdot \det\left(\mathbf{A} - (\alpha - i\,\beta)\mathbf{E}\right) = 0 \tag{15-42}$$

such that

$$\begin{aligned} \det\left((\mathbf{A} - (\alpha + i\,\beta)\mathbf{E}) \cdot (\mathbf{A} - (\alpha - i\,\beta)\mathbf{E})\right) &= 0 \\ \det\left((\mathbf{A} - \alpha\,\mathbf{E})^2 + \beta^2\,\mathbf{E}\right) &= 0 \quad . \end{aligned} \tag{15-43}$$

The last equation yields that the rank of the real matrix $\left((\mathbf{A} - \alpha\,\mathbf{E})^2 + \beta^2\,\mathbf{E}\right)$ is less than n; this now means (see eNote 6) that proper real solutions $\mathbf{x}$ to the corresponding system of equations must exist.

$$\left((\mathbf{A} - \alpha\,\mathbf{E})^2 + \beta^2\,\mathbf{E}\right)\mathbf{x} = \mathbf{0} \quad . \tag{15-44}$$

Let us choose such a proper real solution $\mathbf{v}$ to (15-44) with $|\mathbf{v}| > 0$. Using the assumption that $\mathbf{A}$ (and therefore $\mathbf{A} - \alpha\mathbf{E}$ also) is assumed to be symmetric, we have:

$$\begin{aligned} 0 &= \left(\left((\mathbf{A} - \alpha\,\mathbf{E})^2 + \beta^2\,\mathbf{E}\right)\mathbf{v}\right) \cdot \mathbf{v} \\ &= \left((\mathbf{A} - \alpha\,\mathbf{E})^2\,\mathbf{v}\right) \cdot \mathbf{v} + \beta^2\,(\mathbf{v} \cdot \mathbf{v}) \\ &= ((\mathbf{A} - \alpha\,\mathbf{E})\,\mathbf{v}) \cdot ((\mathbf{A} - \alpha\,\mathbf{E})\,\mathbf{v}) + \beta^2|\mathbf{v}|^2 \\ &= |\,(\mathbf{A} - \alpha\,\mathbf{E})\,\mathbf{v}|^2 + \beta^2|\mathbf{v}|^2 \quad . \end{aligned} \tag{15-45}$$

Since $|\mathbf{v}| > 0$ we are bound to conclude that $\beta = 0$, because all terms in the last expression are non-negative. And this is what we had to prove.

■

|||| Exercise 15.29

Where was it exactly that we *actually used* the symmetry of $\mathbf{A}$ in the above proof?

> To every eigenvalue λ_i for a given matrix $\mathbf{A}$ is associated an eigenvector space E_{λ_i}, which is subspace of $(\mathbb{R}^n, \cdot)$. If two or more eigenvalues for a given matrix are equal, i.e. if we have a multiple root (e.g. k times) $\lambda_i = \lambda_{i+1} = \cdots \lambda_{thati+k-1}$ of the characteristic polynomial, then the corresponding eigenvector spaces are of course also equal: $E_{\lambda_i} = E_{\lambda_{i+1}} = \cdots E_{\lambda_{i+k-1}}$. We will see below in Theorem 15.31 that for symmetric matrices the dimension of the common eigenvector space E_{λ_i} is exactly equal to the algebraic multiplicity k of the eigenvalue λ_i.

If two eigenvalues λ_i and λ_j for a *symmetric* matrix are *different*, then the two corresponding eigenvector spaces are *orthogonal*, $E_{\lambda_i} \perp E_{\lambda_j}$ in the following sense:

|||| Theorem 15.30

Let $\mathbf{A}$ be a symmetric matrix and let λ_1 and λ_2 be two different eigenvalues for $\mathbf{A}$ and let $\mathbf{v}_1$ and $\mathbf{v}_2$ denote two corresponding eigenvectors. Then $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, i.e. they are orthogonal.

|||| Proof

Since $\mathbf{A}$ is symmetric we have from (15-19):

$$\begin{aligned} 0 &= (\mathbf{A}\mathbf{v}_1) \cdot \mathbf{v}_2 - \mathbf{v}_1 \cdot (\mathbf{A}\mathbf{v}_2) \\ &= \lambda_1 \mathbf{v}_1 \cdot \mathbf{v}_2 - \mathbf{v}_1 \cdot (\lambda_2 \mathbf{v}_2) \\ &= \lambda_1 \mathbf{v}_1 \cdot \mathbf{v}_2 - \lambda_2 \mathbf{v}_1 \cdot \mathbf{v}_2 \\ &= (\lambda_1 - \lambda_2)\, \mathbf{v}_1 \cdot \mathbf{v}_2 \quad , \end{aligned} \tag{15-46}$$

and since $\lambda_1 \neq \lambda_2$ we therefore get the following conclusion: $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, and this is what we had to prove.

■

We can now formulate one of the most widely applied results for symmetric matrices, the ***spectral theorem for symmetric matrices*** that, with good reason, is also called the theorem about ***diagonalization of symmetric matrices***:

|||| Theorem 15.31

Let $\mathbf{A}$ denote a *symmetric* $(n \times n)-$*matrix*. Then a special orthogonal matrix $\mathbf{Q}$ exists such that

$$\mathbf{\Lambda} = \mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{Q}^{\top}\mathbf{AQ} \quad \text{is a diagonal matrix} \quad . \tag{15-47}$$

I.e. that a real symmetric matrix can be diagonalized by application of a positive orthogonal substitution, see eNote 14.

The diagonal matrix can be constructed very simply from the n real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of $\mathbf{A}$ as:

$$\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, ..., \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdot & 0 \\ 0 & \lambda_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \lambda_n \end{bmatrix} \quad , \tag{15-48}$$

Remember: A symmetric matrix has exactly n real eigenvalues when we count these with multiplicity.

The special orthogonal matrix $\mathbf{Q}$ is next constructed as columns of the matrix by using the eigenvectors from the corresponding eigenvector-spaces $E_{\lambda_1}, E_{\lambda_2}, \cdots, E_{\lambda_n}$ in the corresponding order:

$$\mathbf{Q} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \mathbf{v}_n] \quad , \tag{15-49}$$

where $\mathbf{v}_1 \in E_{\lambda_1}, \mathbf{v}_2 \in E_{\lambda_2}, \cdots, \mathbf{v}_n \in E_{\lambda_n}$, and the choice of eigenvectors in the respective eigenvector spaces is made so that

1. Any eigenvectors corresponding to the same eigenvalue are chosen orthogonal (use Gram–Schmidt orthogonalization in every common eigenvector space)

2. The chosen eigenvectors are normalized to have length 1.

3. The resulting matrix $\mathbf{Q}$ has determinant 1 (if not then multiply one of the chosen eigenvectors by -1 to flip the sign of the determinant)

That this is so follows from the results and remarks – we go through a series of enlightening examples below.

15.6 Examples of Diagonalization

Here are some typical examples that show how one diagonalizes some small symmetric matrices, i.e. symmetric matrices of type (2×2) or type (3×3):

|||| Example 15.32 Diagonalization by Orthogonal Substitution

A symmetric $(3 \times 3)-$matrix $\mathbf{A}$ is given as:

$$\mathbf{A} = \begin{bmatrix} 2 & -2 & 1 \\ -2 & 5 & -2 \\ 1 & -2 & 2 \end{bmatrix} \quad . \tag{15-50}$$

We will determine a special orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix:

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{\Lambda} \quad . \tag{15-51}$$

First we determine the eigenvalues for $\mathbf{A}$: The characteristic polynomial for $\mathbf{A}$ is

$$\mathcal{K}_{\mathbf{A}}(\lambda) = \det\left(\begin{bmatrix} 2-\lambda & -2 & 1 \\ -2 & 5-\lambda & -2 \\ 1 & -2 & 2-\lambda \end{bmatrix}\right) = (\lambda-1)^2 \cdot (7-\lambda) \quad , \tag{15-52}$$

so $\mathbf{A}$ has the eigenvalues $\lambda_1 = 7$, $\lambda_2 = 1$, and $\lambda_3 = 1$. Because of this we already know through Theorem 15.31 that it is possible to construct a positive orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{diag}(7,1,1) = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad . \tag{15-53}$$

The rest of the problem now consists in finding the eigenvectors for $\mathbf{A}$ that can be used as columns in the orthogonal matrix $\mathbf{Q}$.

Eigenvectors for $\mathbf{A}$ corresponding to the eigenvalue $\lambda_1 = 7$ are found by solving the homogeneous system of equations that has the coefficient matrix

$$\mathbf{K}_{\mathbf{A}}(7) = \mathbf{A} - 7\mathbf{E} = \begin{bmatrix} -5 & -2 & 1 \\ -2 & -2 & -2 \\ 1 & -2 & -5 \end{bmatrix} \quad , \tag{15-54}$$

which by suitable row operations is seen to have

$$\mathrm{rref}(\mathbf{K}_{\mathbf{A}}(7)) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \quad . \tag{15-55}$$

The eigenvector solutions to the corresponding homogeneous system of equations are seen to be

$$\mathbf{u} = t \cdot (1, -2, 1) \quad , \quad t \in \mathbb{R} \quad , \tag{15-56}$$

such that $E_7 = \mathrm{span}\{(1, -2, 1)\}$. The normalized eigenvector $\mathbf{v}_1 = (1/\sqrt{6}) \cdot (1, -2, 1)$ is therefore an orthonormal basis for E_7 (and it can also be used as the first column vector in the wanted $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} 1/\sqrt{6} & * & * \\ -2/\sqrt{6} & * & * \\ 1/\sqrt{6} & * & * \end{bmatrix} \quad . \tag{15-57}$$

We know from Theorem 15.31 that the two last columns are found by similarly determining all eigenvectors E_1 belonging to the eigenvalue $\lambda_2 = \lambda_3 = 1$ and then choosing two orthonormal eigenvectors from E_1.

The reduction matrix corresponding to the eigenvalue 1 is

$$\mathbf{K_A}(1) = \mathbf{A} - \mathbf{E} = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} , \tag{15-58}$$

which again by suitable row operations is seen to have

$$\mathrm{rref}(\mathbf{K_A}(1)) = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad . \tag{15-59}$$

The eigenvector solutions to the corresponding homogeneous system of equations are seen to be

$$\mathbf{u} = t_1 \cdot (2, 1, 0) + t_2 \cdot (-1, 0, 1) \quad , \quad t_1 \in \mathbb{R} \quad , \quad t_2 \in \mathbb{R} \quad , \tag{15-60}$$

such that $E_1 = \mathrm{span}\{(-1, 0, 1), (2, 1, 0)\}$.

We find an orthonormal basis for E_1 using the Gram–Schmidt orthonormalization of $\mathrm{span}\{(-1, 0, 1), (2, 1, 0)\}$ like this: Since we have already defined $\mathbf{v}_1$ we put $\mathbf{v}_2$ to be

$$\mathbf{v}_2 = \frac{(-1, 0, 1)}{|(-1, 0, 1)|} = (1/\sqrt{2}) \cdot (-1, 0, 1) \quad , \tag{15-61}$$

and then as in the Gram–Schmidt process:

$$\mathbf{w}_3 = (2, 1, 0) - ((2, 1, 0) \bullet \mathbf{v}_2) \cdot \mathbf{v}_2 = (1, 1, 1) \quad . \tag{15-62}$$

By normalization we finally get $\mathbf{v}_3 = (1/\sqrt{3}) \cdot (1, 1, 1)$ and then we finally have all the ingredients to the wanted orthogonal matrix $\mathbf{Q}$:

$$\mathbf{Q} = [\mathbf{v}_1\, \mathbf{v}_2\, \mathbf{v}_3] = \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ -2/\sqrt{6} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix} \quad . \tag{15-63}$$

Finally we investigate whether the chosen eigenvectors give a positive orthogonal matrix. Since

$$\det\left(\begin{bmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right) = -6 < 0 \quad , \tag{15-64}$$

$\mathbf{Q}$ has negative determinant. A special orthogonal matrix is found by multiplying one of the columns of $\mathbf{Q}$ by -1, e.g. the last one. Note that a vector $\mathbf{v}$ is an eigenvector for $\mathbf{A}$ if and only if $-\mathbf{v}$ is also an eigenvector for $\mathbf{A}$. Therefore we have that

$$\mathbf{Q} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & (-\mathbf{v}_3) \end{pmatrix} = \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \\ -2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix} \tag{15-65}$$

is a *positive* orthogonal matrix that diagonalizes $\mathbf{A}$.

This is checked by a direct computation:

$$\begin{aligned} &\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} \\ &= \begin{bmatrix} 1/\sqrt{6} & -2/\sqrt{6} & 1/\sqrt{6} \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{3} & -1/\sqrt{3} & -1/\sqrt{3} \end{bmatrix} \cdot \begin{bmatrix} 2 & -2 & 1 \\ -2 & 5 & -2 \\ 1 & -2 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \\ -2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix} \\ &= \begin{bmatrix} 7 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \end{aligned} \tag{15-66}$$

which we wanted to show.

We should finally remark here that, since we are in three dimensions, instead of using Gram–Schmidt orthonormalization for the determination of $\mathbf{v}_3$ we could have used the cross product $\mathbf{v}_1 \times \mathbf{v}_2$ (see 15.23):

$$\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2 = (1/\sqrt{3}) \cdot (-1, -1, -1) \quad . \tag{15-67}$$

|||| Example 15.33 Diagonalization by Orthogonal Substitution

A symmetric $(2 \times 2)-$matrix $\mathbf{A}$ is given as:

$$\mathbf{A} = \begin{bmatrix} 11 & -12 \\ -12 & 4 \end{bmatrix} \quad . \tag{15-68}$$

We will determine a special orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix:

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{\Lambda} \quad . \tag{15-69}$$

First we determine the eigenvalues for $\mathbf{A}$: The characteristic polynomial for $\mathbf{A}$ is

$$\mathcal{K}_{\mathbf{A}}(\lambda) = \det\left(\begin{bmatrix} 11-\lambda & -12 \\ -12 & 4-\lambda \end{bmatrix}\right) = (\lambda - 20)\cdot(\lambda+5) \quad , \tag{15-70}$$

so $\mathbf{A}$ has the eigenvalues $\lambda_1 = 20$ and $\lambda_2 = -5$. Therefore we now have:

$$\mathbf{\Lambda} = \begin{bmatrix} 20 & 0 \\ 0 & -5 \end{bmatrix} \quad . \tag{15-71}$$

The eigenvectors for $\mathbf{A}$ corresponding to the eigenvalue $\lambda_1 = 20$ are found by solving the homogeneous system of equations having the coefficient matrix

$$\mathbf{K}_{\mathbf{A}}(20) = \mathbf{A} - 20\mathbf{E} = \begin{bmatrix} -9 & -12 \\ -12 & -16 \end{bmatrix} \quad , \tag{15-72}$$

which, through suitable row operations, is shown to have the equivalent reduced matrix:

$$\operatorname{rref}(\mathbf{K}_{\mathbf{A}}(20)) = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix} \quad . \tag{15-73}$$

The eigenvector solutions to the corresponding homogeneous system of equations are found to be

$$\mathbf{u} = t\cdot(4,-3) \quad , \quad t \in \mathbb{R} \quad , \tag{15-74}$$

such that $E_{20} = \operatorname{span}\{(4,-3)\}$. The normalized eigenvector $\mathbf{v}_1 = (1/5)\cdot(4,-3)$ is therefore an orthonormal basis for E_{20} (and it can therefore be used as the first column vector in the wanted $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} 4/5 & * \\ -3/5 & * \end{bmatrix} \quad . \tag{15-75}$$

The last column in $\mathbf{Q}$ is an eigenvector corresponding to the second eigenvalue $\lambda_2 = -5$ and can therefore be found from the general solution E_{-5} to the homogeneous system of equations having the coefficient matrix

$$\mathbf{K}_{\mathbf{A}}(-5) = \mathbf{A} - 5\cdot\mathbf{E} = \begin{bmatrix} 16 & -12 \\ -12 & 9 \end{bmatrix} \quad , \tag{15-76}$$

but since we know that the wanted eigenvector is orthogonal to the eigenvector $\mathbf{v}_1$ we can just use a vector perpendicular to the first eigenvector, $\mathbf{v}_2 = (1/5)\cdot(3,4)$, evidently a unit vector, that is orthogonal to $\mathbf{v}_1$. It is easy to check that $\mathbf{v}_2$ is an eigenvector for $\mathbf{A}$ corresponding to the eigenvalue -5:

$$\mathbf{K}_{\mathbf{A}}(-5)\cdot\mathbf{v}_2 = \begin{bmatrix} 16 & -12 \\ -12 & 9 \end{bmatrix}\cdot\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad . \tag{15-77}$$

Therefore we substitute $\mathbf{v}_2$ as the second column in $\mathbf{Q}$ and get

$$\mathbf{Q} = \begin{bmatrix} 4/5 & 3/5 \\ -3/5 & 4/5 \end{bmatrix} \quad . \tag{15-78}$$

This matrix has the determinant $\det(\mathbf{Q}) = 1 > 0$, so $\mathbf{Q}$ is a positive orthogonal substitution matrix satisfying that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix:

$$\begin{aligned} \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} &= \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} \\ &= \begin{bmatrix} 4/5 & -3/5 \\ 3/5 & 4/5 \end{bmatrix} \cdot \begin{bmatrix} 11 & -12 \\ -12 & 4 \end{bmatrix} \cdot \begin{bmatrix} 4/5 & 3/5 \\ -3/5 & 4/5 \end{bmatrix} \\ &= \begin{bmatrix} 20 & 0 \\ 0 & -5 \end{bmatrix} \\ &= \mathbf{diag}(20, -5) = \mathbf{\Lambda} \quad . \end{aligned} \tag{15-79}$$

|||| Example 15.34 Diagonalization by Orthogonal Substitution

A symmetric $(3 \times 3)-$matrix $\mathbf{A}$ is given like this:

$$\mathbf{A} = \begin{bmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{bmatrix} \quad . \tag{15-80}$$

We will determine a positive orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix:

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} = \mathbf{\Lambda} \quad . \tag{15-81}$$

First we determine the eigenvalues for $\mathbf{A}$: The characteristic polynomial for $\mathbf{A}$ is

$$\mathcal{K}_{\mathbf{A}}(\lambda) = \det\left(\begin{bmatrix} 7-\lambda & -2 & 1 \\ -2 & 6-\lambda & -2 \\ 1 & -2 & 5-\lambda \end{bmatrix}\right) = -(\lambda-3)\cdot(\lambda-6)\cdot(\lambda-9) \quad , \tag{15-82}$$

from which we read the three different eigenvalues $\lambda_1 = 9$, $\lambda_2 = 6$, and $\lambda_3 = 3$ and then the diagonal matrix we are on the road to describe as $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$:

$$\mathbf{\Lambda} = \mathbf{diag}(9, 6, 3) = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{bmatrix} \tag{15-83}$$

The eigenvectors for $\mathbf{A}$ corresponding to the eigenvalue $\lambda_3 = 3$ are found by solving the homogeneous system of equations having the coefficient matrix

$$\mathbf{K}_{\mathbf{A}}(3) = \mathbf{A} - 3 \cdot \mathbf{E} = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 3 & -2 \\ 0 & -2 & 2 \end{bmatrix} \quad , \tag{15-84}$$

which through suitable row operations is seen to have

$$\mathrm{rref}(\mathbf{K_A}(3)) = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad . \tag{15-85}$$

The eigenvector solutions to the corresponding homogeneous system of equations are found to be

$$\mathbf{u}_3 = t \cdot (1,2,2) \quad , \quad t \in \mathbb{R} \quad , \tag{15-86}$$

such that $E_3 = \mathrm{span}\{(1,2,2)\}$. The normalized eigenvector $\mathbf{v}_1 = (1/3) \cdot (1,2,2)$ is therefore an orthonormal basis for E_3 so it can be used as the *third column vector* in the wanted $\mathbf{Q}$; note that we have just found the eigenvector space to the *third eigenvalue* on the list of eigenvalues for $\mathbf{A}$:

$$\mathbf{Q} = \begin{bmatrix} * & * & 1/3 \\ * & * & 2/3 \\ * & * & 2/3 \end{bmatrix} \quad . \tag{15-87}$$

We know from Theorem 15.31 that the two last columns are found by similarly determining the eigenvector space E_6 corresponding to eigenvalue $\lambda_2 = 6$, and the eigenvector space E_9 corresponding to the eigenvalue $\lambda_1 = 9$.

For $\lambda_2 = 6$ we have:

$$\mathbf{K_A}(6) = \mathbf{A} - 6 \cdot \mathbf{E} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 0 & -2 \\ 0 & -2 & -1 \end{bmatrix} \quad , \tag{15-88}$$

which by suitable row operations is found to have the following equivalent reduced matrix:

$$\mathrm{rref}(\mathbf{K_A}(6)) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad . \tag{15-89}$$

The eigenvector solutions to the corresponding homogeneous system of equations are found to be

$$\mathbf{u}_2 = t \cdot (-2,-1,2) \quad , \quad t \in \mathbb{R} \quad , \tag{15-90}$$

so that $E_6 = \mathrm{span}\{(-2,-1,2)\}$. The normalized eigenvector $\mathbf{v}_2 = (1/3) \cdot (-2,-1,2)$ is therefore an orthonormal basis for E_6 (and it can therefore be used as the *second column vector* in the wanted $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} * & -2/3 & 1/3 \\ * & -1/3 & 2/3 \\ * & 2/3 & 2/3 \end{bmatrix} \quad . \tag{15-91}$$

Instead of determining the eigenvector space E_9 for the last eigenvalue $\lambda_1 = 9$ in the same way we use the fact that this eigenvector space is spanned by a vector $\mathbf{v}_1$ that is orthogonal to both $\mathbf{v}_3$ and $\mathbf{v}_2$, so we can use $\mathbf{v}_1 = \mathbf{v}_2 \times \mathbf{v}_3 = (1/3) \cdot (-2, 2, -1)$, and then we finally get

$$\mathbf{Q} = \begin{bmatrix} -2/3 & -2/3 & 1/3 \\ 2/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & 2/3 \end{bmatrix} \quad . \tag{15-92}$$

This matrix is positive orthogonal since $\det(\mathbf{Q}) = 1 > 0$, and therefore we have determined a positive orthogonal matrix $\mathbf{Q}$ that diagonalizes $\mathbf{A}$ to the diagonal matrix $\mathbf{\Lambda}$. This is easily proved by direct computation:

$$\begin{aligned} &\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{\top}\mathbf{A}\mathbf{Q} \\ &= \begin{bmatrix} -2/3 & 2/3 & -1/3 \\ -2/3 & -1/3 & 2/3 \\ 1/3 & 2/3 & 2/3 \end{bmatrix} \cdot \begin{bmatrix} 7 & -2 & 0 \\ -2 & 6 & -2 \\ 0 & -2 & 5 \end{bmatrix} \cdot \begin{bmatrix} -2/3 & -2/3 & 1/3 \\ 2/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & 2/3 \end{bmatrix} \\ &= \begin{bmatrix} 9 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{bmatrix} \\ &= \mathbf{diag}(9, 6, 3) = \mathbf{\Lambda} \quad . \end{aligned} \tag{15-93}$$

15.7 Controlled Construction of Symmetric Matrices

In the light of the above examples it is clear that if only we can construct all orthogonal (2×2)- and (3×3)-matrices $\mathbf{Q}$ (or for that matter $(n \times n)$-matrices), then we can *produce* all symmetric (2×2)- and (3×3)-matrices $\mathbf{A}$ as $\mathbf{A} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{\top}$. We only have to *choose* the wanted eigenvalues in the diagonal for $\mathbf{\Lambda}$.

Every special orthogonal 2×2-matrix has the following form, which shows that it is a rotation given by a rotation angle φ :

$$\mathbf{Q} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \quad , \tag{15-94}$$

where φ is an angle in the interval $[-\pi, \pi]$. Note that the column vectors are orthogonal and both have length 1. Furthermore the determinant $\det(\mathbf{Q}) = 1$, so $\mathbf{Q}$ is special orthogonal.

|||| Exercise 15.35

Prove the statement that *every* special orthogonal matrix can be stated in the form (15-94) for a suitable choice of the rotation angle φ.

If $\varphi > 0$ then $\mathbf{Q}$ rotates vectors in the positive direction, i.e. counter-clockwise; if $\varphi < 0$ then $\mathbf{Q}$ rotates vectors in the negative direction, i.e. clockwise.

|||| Definition 15.36 Rotation Matrices

Every special orthogonal (2×2)-matrix is also called a ***rotation matrix***.

Since every positive orthogonal 3×3-matrix similarly can be stated as a product of rotations about the three coordinate axes – see below – we will extend the naming as follows:

|||| Definition 15.37 Rotation Matrices

Every special orthogonal (3×3)-matrix is called a ***rotation matrix***.

A rotation about a coordinate axis, i.e. a rotation by a given angle about one of the coordinate axes, is produced with one of the following special orthogonal matrices:

$$\mathbf{R}_x(u) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(u) & -\sin(u) \\ 0 & \sin(u) & \cos(u) \end{bmatrix}$$

$$\mathbf{R}_y(v) = \begin{bmatrix} \cos(v) & 0 & \sin(v) \\ 0 & 1 & 0 \\ -\sin(v) & 0 & \cos(v) \end{bmatrix} \tag{15-95}$$

$$\mathbf{R}_z(w) = \begin{bmatrix} \cos(w) & -\sin(w) & 0 \\ \sin(w) & \cos(w) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad ,$$

where the rotation angles are u, v, and w, respectively.

|||| Exercise 15.38

Show by direct calculation that the three axis-rotation matrices and every product of axis-rotation matrices really *are* special orthogonal matrices, i.e. they satisfy $\mathbf{R}^{-1} = \mathbf{R}^{\top}$ and $\det(\mathbf{R}) = 1$.

|||| Exercise 15.39

Find the image vectors of every one of the given vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ by use of the given mapping matrices $\mathbf{Q}_i$:

$$\begin{aligned}
&\mathbf{Q}_1 = \mathbf{R}_x(\pi/4) \quad , \quad \mathbf{a} = (1,0,0),\ \mathbf{b} = (0,1,0),\ \mathbf{c} = (0,0,1) \\
&\mathbf{Q}_2 = \mathbf{R}_y(\pi/4) \quad , \quad \mathbf{a} = (1,1,1),\ \mathbf{b} = (0,1,0),\ \mathbf{c} = (0,0,1) \\
&\mathbf{Q}_3 = \mathbf{R}_z(\pi/4) \quad , \quad \mathbf{a} = (1,1,0),\ \mathbf{b} = (0,1,0),\ \mathbf{c} = (0,0,1) \\
&\mathbf{Q}_4 = \mathbf{R}_y(\pi/4)\cdot\mathbf{R}_x(\pi/4) \quad , \quad \mathbf{a} = (1,0,0),\ \mathbf{b} = (0,1,0),\ \mathbf{c} = (0,0,1) \\
&\mathbf{Q}_5 = \mathbf{R}_x(\pi/4)\cdot\mathbf{R}_y(\pi/4) \quad , \quad \mathbf{a} = (1,0,0),\ \mathbf{b} = (0,1,0),\ \mathbf{c} = (0,0,1) \quad .
\end{aligned} \tag{15-96}$$

The combination of rotations about the coordinate axes by given rotation angles u, v, and w about the $x-$axis, $y-$axis, and $z-$axis is found by computing the matrix product of the three corresponding rotation matrices.

Here is the complete general expression for the matrix product for all values of u, v and w:

$$\mathbf{R}(u,v,w) \;=\; \mathbf{R}_z(w)\cdot\mathbf{R}_y(v)\cdot\mathbf{R}_x(u)$$

$$= \begin{bmatrix} \cos(w)\cos(v) & -\sin(w)\cos(u)-\cos(w)\sin(v)\sin(u) & \sin(w)\sin(u)-\cos(w)\sin(v)\cos(u) \\ \sin(w)\cos(v) & \cos(w)\cos(u)-\sin(w)\sin(v)\sin(u) & -\cos(w)\sin(u)-\sin(w)\sin(v)\cos(u) \\ \sin(v) & \cos(v)\sin(u) & \cos(v)\cos(u) \end{bmatrix} .$$

As one might suspect, it is possible to prove the following theorem:

|||| Theorem 15.40 Axis Rotation Angles for a Given Rotation Matrix

Every rotation matrix $\mathbf{R}$ (i.e. every special orthogonal matrix $\mathbf{Q}$) can be written as the product of 3 axis-rotation matrices:

$$\mathbf{R} = \mathbf{R}(u,v,w) = \mathbf{R}_z(w) \cdot \mathbf{R}_y(v) \cdot \mathbf{R}_x(u) \quad . \tag{15-97}$$

In other words: the effect of every rotation matrix can be realized by three consecutive rotations about the coordinate axes – with the rotation angles u, v, and w, respectively, as given in the above matrix product.

When a given special orthogonal matrix $\mathbf{R}$ is given (with its matrix elements r_{ij}), it is not difficult to find these axis rotation angles. As is evident from the above matrix product we have e.g. that $\sin(v) = r_{31}$ such that $v = \arcsin(r_{31})$ or $v = \pi - \arcsin(r_{31})$, and $\cos(w)\cos(v) = r_{11}$ such that $w = \arccos(r_{11}/\cos(v))$ or $v = -\arccos(r_{31}/\cos(v))$, if only $\cos(v) \neq 0$ i.e. if only $v \neq \pm\pi/2$.

|||| Exercise 15.41

Show that if $v = \pi/2$ or $v = -\pi/2$ then there exist many values of u and w giving the *same* $\mathbf{R}(u,v,w)$. I.e. not all angle values are uniquely determined in the interval $]-\pi,\pi]$ for every given rotation matrix $\mathbf{R}$.

|||| Exercise 15.42

Show that if $\mathbf{R}$ is a rotation matrix (a positive orthogonal matrix) then $\mathbf{R}^\top$ is also a rotation matrix, and vice versa: if $\mathbf{R}^\top$ is a rotation matrix then $\mathbf{R}$ is also a rotation matrix.

|||| Exercise 15.43

Show that if $\mathbf{R}_1$ and $\mathbf{R}_2$ are rotation matrices then $\mathbf{R}_1 \cdot \mathbf{R}_2$ and $\mathbf{R}_2 \cdot \mathbf{R}_1$ are also rotation matrices. Give examples that show that $\mathbf{R}_1 \cdot \mathbf{R}_2$ is not necesarily the same rotation matrix as $\mathbf{R}_2 \cdot \mathbf{R}_1$.

15.8 Structure of Rotation Matrices

As mentioned above (Exercise 15.35), every 2×2 special orthogonal matrix has the form:

$$\mathbf{Q} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

This is a rotation of the plane anticlockwise by the angle ϕ. The angle ϕ is related to the eigenvalues of $\mathbf{Q}$:

|||| Exercise 15.44

Show that the eigenvalues of the matrix $\mathbf{Q}$ above are:

$$\lambda_1 = e^{i\phi}, \qquad \lambda_2 = e^{-i\phi}.$$

How about the 3×3 case? We already remarked that any 3×3 special orthogonal matrix can be written as a composition of rotations about the three coordinate matrices: $\mathbf{Q} = \mathbf{R}_z(w) \cdot \mathbf{R}_y(v) \cdot \mathbf{R}_x(u)$. But is $\mathbf{Q}$ itself a rotation about some axis (i.e. some line through the origin)? We can prove this is so, by examining the eigenvalues and eigenvectors of $\mathbf{Q}$.

|||| Theorem 15.45

The eigenvalues of any orthogonal matrix all have absolute value 1.

Proof. If λ is an eigenvalue of an orthogonal matrix $\mathbf{Q}$, there is, by definition, a non-zero complex eigenvector $\mathbf{v}$ in $\mathbb{C}^n \setminus \{\mathbf{0}\}$. Writing $\mathbf{v}$ as a column matrix, we then have:

$$\begin{aligned} \lambda\bar{\lambda}\mathbf{v}^T \cdot \bar{\mathbf{v}} &= (\lambda\mathbf{v})^T \cdot \overline{(\lambda\mathbf{v})} \\ &= (\mathbf{Q}\cdot\mathbf{v})^T \cdot \overline{(\mathbf{Q}\cdot\mathbf{v})} && (\mathbf{Q}\cdot\mathbf{v} = \lambda\mathbf{v}) \\ &= \mathbf{v}^T \cdot \mathbf{Q}^T \cdot \mathbf{Q} \cdot \bar{\mathbf{v}} && (\bar{\mathbf{Q}} = \mathbf{Q}) \\ &= \mathbf{v}^T \cdot \mathbf{E} \cdot \bar{\mathbf{v}} && (\mathbf{Q}^T \cdot \mathbf{Q} = \mathbf{E}) \\ &= \mathbf{v}^T \cdot \bar{\mathbf{v}}. \end{aligned}$$

Since $\mathbf{v} \neq \mathbf{0}$, it follows that $\mathbf{v}^T \cdot \bar{\mathbf{v}}$ is a non-zero (real) number:

$$\begin{aligned} \mathbf{v}^T \cdot \bar{\mathbf{v}} &= v_1\bar{v}_1 + v_2\bar{v}_2 \dots v_n\bar{v}_n \\ &= |v_1|^2 + |v_2|^2 + \dots |v_n|^2 > 0. \end{aligned}$$

Dividing $\lambda\bar{\lambda}\mathbf{v}^T \cdot \bar{\mathbf{v}} = \mathbf{v}^T \cdot \bar{\mathbf{v}}$ by this number we get:

$$|\lambda|^2 = \lambda\bar{\lambda} = 1.$$

□

We can now apply this to the eigenvalues of a 3 × 3 special orthogonal matrix:

|||| Theorem 15.46

Let $\mathbf{Q}$ be a 3×3 special orthogonal matrix, i.e. $\mathbf{Q}^T\mathbf{Q} = \mathbf{E}$, and $\det \mathbf{Q} = 1$. Then the eigenvalues are:

$$\lambda_1 = 1, \quad \lambda_2 = e^{i\phi}, \quad \lambda_3 = e^{-i\phi},$$

for some $\phi \in]-\pi, \pi]$.

Proof. $\mathbf{Q}$ is a real matrix, so all eigenvalues are either real or come in complex conjugate pairs. There are 3 of them, because $\mathbf{Q}$ is a 3×3 matrix, so the characteristic polynomial has degree 3. Hence there is at least one real eigenvalue:

$$\lambda_1 \in \mathbb{R}.$$

Now there are two possibilities:
Case 1: All roots are real: then, since all eigenvalues have absolute value 1 (by Theorem 15.45), and

$$1 = \det \mathbf{Q} = \lambda_1\lambda_2\lambda_3$$

either one or all three of the eigenvalues are equal to 1.

Case 2: λ_1 is real and the other two are complex conjugate, $\lambda_3 = \bar{\lambda}_2$, so:

$$1 = \det \mathbf{Q} = \lambda_1\lambda_2\bar{\lambda}_2 = \lambda_1|\lambda_2|^2 = \lambda_1,$$

where we used that $|\lambda_2| = 1$. Any complex number λ with absolute value 1 is of the form $e^{i\phi}$, where $\phi = \operatorname{Arg}(\lambda)$, so this gives the claimed form of λ_1, λ_2 and λ_3.

Note that the case $\lambda_2 = \lambda_3 = 1$ or -1 (in Case 1) correspond respectively to $\phi = 0$ and $\phi = \pi$ in the wording of the theorem. □

We can also say something about the eigenvectors corresponding to the eigenvalues.

|||| Theorem 15.47

Let $\mathbf{Q}$ be a special orthogonal matrix, and denote the eigenvalues as in Theorem 15.46. If the eigenvalues are not all real, i.e. $\mathrm{Im}(\lambda_2) \neq 0$, then the eigenvectors corresponding to λ_2 and λ_3 are necessarily of the form:

$$\mathbf{v}_2 = \mathbf{x} + i\mathbf{y}, \qquad \mathbf{v}_3 = \bar{\mathbf{v}}_2 = \mathbf{x} - i\mathbf{y},$$

where $\mathbf{x}$ and $\mathbf{y}$ are respectively the real and imaginary parts of $\mathbf{v}_2$, and

$$\mathbf{x} \cdot \mathbf{y} = 0 \qquad \text{and} \quad |\mathbf{x}| = |\mathbf{y}|.$$

If $\mathbf{v}_1$ is an eigenvector for $\lambda_1 = 1$, then:

$$\mathbf{v}_1 \cdot \mathbf{x} = \mathbf{v}_1 \cdot \mathbf{y} = 0$$

Proof. We have $\mathbf{Q}\mathbf{v}_2 = \lambda_2 \mathbf{v}_2$, and $\mathbf{Q}\bar{\mathbf{v}}_2 = \bar{\lambda}_2 \bar{\mathbf{v}}_2$. So clearly a third eigenvector, corresponding to $\bar{\lambda}_2$, is $\mathbf{v}_3 = \bar{\mathbf{v}}_2$. Using $\mathbf{Q}^T\mathbf{Q} = \mathbf{E}$, we have

$$\mathbf{v}_2^T \mathbf{v}_2 = \mathbf{v}_2^T \mathbf{Q}^T \mathbf{Q} \mathbf{v}_2 = (\mathbf{Q}\mathbf{v}_2)^T (\mathbf{Q}\mathbf{v}_2) = \lambda_2^2 \mathbf{v}_2^T \mathbf{v}_2.$$

If $\mathbf{v}_2^T \mathbf{v}_2 \neq 0$, then we can divide by this number to get $\lambda_2^2 = 1$. But $\lambda_2 = a + bi$, with $b \neq 0$, so this would mean: $1 = \lambda_2^2 = a^2 - b^2 + 2iab$. The imaginary part is: $ab = 0$, which implies that $a = 0$ and hence $\lambda_2^2 = -b^2$, which cannot be equal to 1. Hence:

$$\mathbf{v}_2^T \mathbf{v}_2 = 0.$$

Writing $\mathbf{v}_2 = \mathbf{x} + i\mathbf{y}$, this is:

$$\begin{aligned} 0 &= (\mathbf{x}^T + i\mathbf{y}^T)(\mathbf{x} + i\mathbf{y}) \\ &= \mathbf{x}^T\mathbf{x} - \mathbf{y}^T\mathbf{y} + i(\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{x}). \end{aligned}$$

The real part of this equation is:

$$\mathbf{x}^T\mathbf{x} - \mathbf{y}^T\mathbf{y} = 0, \qquad \text{i.e.,} \quad \mathbf{x} \cdot \mathbf{x} = |\mathbf{x}|^2 = \mathbf{y} \cdot \mathbf{y} = |\mathbf{y}|^2,$$

and the imaginary part is:

$$0 = \mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{x} = \mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{x} = 2\mathbf{x} \cdot \mathbf{y}.$$

Lastly, if $\mathbf{v}_1$ is an eigenvector for $\lambda_1 = 1$, then, by the same argument as above,

$$\mathbf{v}_1^T\mathbf{v}_2 = 1 \cdot \lambda_2 \cdot \mathbf{v}_1^T\mathbf{v}_2,$$

which must be zero, since $\lambda_2 \neq 1$. This is:

$$0 = \mathbf{v}_1^T(\mathbf{x} + i\mathbf{y}) = \mathbf{v}_1 \cdot \mathbf{x} + i\, \mathbf{v}_1 \cdot \mathbf{y}.$$

Since $\mathbf{v}_1$ is real, the real and imaginary parts of this give $\mathbf{v}_1 \cdot \mathbf{x} = \mathbf{v}_1 \cdot \mathbf{y} = 0$.

□

Now we can give a precise description of the geometric effect of a 3 × 3 rotation matrix:

|||| Theorem 15.48

Let $\mathbf{Q}$ be a 3×3 special orthogonal matrix, and $\lambda_1 = 1$, $\lambda_2 = e^{i\phi}$, $\lambda_3 = e^{-i\phi}$ be its eigenvalues, with corresponding eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3 = \bar{\mathbf{v}}_2$. Then:

1. The map $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by $f(\mathbf{x}) = \mathbf{Q}\mathbf{x}$ is a rotation by angle ϕ around the line spanned by $\mathbf{v}_1$.
2. If λ_2 is not real then an orthonormal basis for $\mathbb{R}^3$ is given by:

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{|\mathbf{v}_1|}, \quad \mathbf{u}_2 = \frac{\operatorname{Im} \mathbf{v}_2}{|\operatorname{Im} \mathbf{v}_2|}, \quad \mathbf{u}_3 = \frac{\operatorname{Re} \mathbf{v}_2}{|\operatorname{Re} \mathbf{v}_2|},$$

   where $\mathbf{v}_2$ is an eigenvector for $\lambda_2 = e^{i\phi}$. The mapping matrix for f with respect to this basis is:

$$_u f_u = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}.$$

3. If λ_2 is real then $\mathbf{Q}$ is either the identity map ($\lambda_2 = \lambda_3 = 1$) or a rotation by angle π ($\lambda_2 = \lambda_3 = -1$).

Proof. Statement 1 follows from statements 2 and 3, since these represent rotations by angle ϕ around the $\mathbf{v}_1$ axis.

For statement 2, by Theorem 15.47, if λ_2 is not real, then $u = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ as defined above are an orthonormal basis for $\mathbb{R}^3$, since they are mutually orthogonal and of length 1.

To find the mapping matrix, we have $f(\mathbf{u}_1) = \mathbf{u}_1 = 1 \cdot \mathbf{u}_1 + 0 \cdot \mathbf{u}_2 + 0 \cdot \mathbf{u}_3$, which gives the first column. For $\mathbf{u}_2$ and $\mathbf{u}_3$, according to Theorem 15.47, the real and imaginary parts of $\mathbf{v}_2$ have the same length, so we can rescale $\mathbf{v}_2$ by dividing by this number to get

$$\mathbf{w} = \mathbf{u}_3 + i\mathbf{u}_2, \qquad \mathbf{u}_2 = \frac{\operatorname{Im} \mathbf{v}_2}{|\operatorname{Im} \mathbf{v}_2|}, \quad \mathbf{u}_3 = \frac{\operatorname{Re} \mathbf{v}_2}{|\operatorname{Re} \mathbf{v}_2|},$$

where $\mathbf{w}$ is an eigenvector for f with eigenvalue $e^{i\phi}$. That is:

$$\begin{aligned} e^{i\phi}\mathbf{w} = (\cos\phi + i\sin\phi)(\mathbf{u}_3 + i\mathbf{u}_2) &= f(\mathbf{u}_3 + i\mathbf{u}_2) \\ (\cos\phi\mathbf{u}_3 - \sin\phi\mathbf{u}_2) + i(\sin\phi\mathbf{u}_3 + \cos\phi\mathbf{u}_2) &= f(\mathbf{u}_3) + if(\mathbf{u}_2). \end{aligned}$$

The imaginary and real parts of this equation give:

$$\begin{aligned} f(\mathbf{u}_2) &= \cos\phi\mathbf{u}_2 + \sin\phi\mathbf{u}_3 \\ f(\mathbf{u}_3) &= -\sin\phi\mathbf{u}_2 + \cos\phi\mathbf{u}_3, \end{aligned}$$

and this gives us the second and third columns of the mapping matrix.

This mapping matrix is precisely the matrix of a rotation by angle ϕ around the $\mathbf{v}_1$ axis (compare ${}_u f_u$ with the matrix $\mathbf{R}_x(u)$ discussed earlier).

For statement 3, the special case that λ_2 is real, if $\lambda_2 = \lambda_3 = 1$, then $\phi = 0$ and $\mathbf{Q}$ is the identity matrix, which can be regarded as a rotation by angle 0 around any axis.

Finally, for the case $\lambda_2 = \lambda_3 = -1$, briefly: let $E_1 = \operatorname{span}\{\mathbf{v}_1\}$. Choose any orthonormal basis for the orthogonal complement $E_1^{\perp}$. Using this, one can show that the restricition of f to $E_1^{\perp}$ is a 2×2 rotation matrix with a repeated eigenvalue -1. This means it is minus the identity matrix on $E_1^{\perp}$, i.e. a rotation by angle π, from which the claim follows. □

|||| Example 15.49

The axis of rotation for a 3×3 rotation matrix is sometimes called the *Euler axis*. Let's find the Euler axis, and the rotation angle for the special orthogonal matrix:

$$\mathbf{Q} = \frac{1}{3}\begin{bmatrix} -2 & -2 & 1 \\ 2 & -1 & 2 \\ -1 & 2 & 2 \end{bmatrix},$$

which was used for a change of basis in Example 15.34.

The eigenvalues are:

$$\lambda_1 = 1, \qquad \lambda_2 = -\frac{2}{3} + i\frac{\sqrt{5}}{3}, \qquad \lambda_3 = -\frac{2}{3} - i\frac{\sqrt{5}}{3},$$

with corresponding eigenvectors:

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} -i\sqrt{5} \\ -2 \\ 1 \end{bmatrix}, \qquad \mathbf{v}_3 = \overline{\mathbf{v}_2}.$$

So the axis of rotation is the line spanned by $\mathbf{v}_1 = (0, 1, 2)$, and the angle of rotation is:

$$\phi = \mathrm{Arg}(\lambda_2) = -\arctan\left(\frac{\sqrt{5}}{2}\right) + \pi$$

We can set:

$$\mathbf{u}_1 = \mathbf{v}_1 = \frac{1}{\sqrt{5}}\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad \mathbf{u}_2 = \mathrm{Im}\mathbf{v}_2 = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \mathrm{Re}\mathbf{v}_2 = \frac{1}{\sqrt{5}}\begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix},$$

and, setting $U = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$, the matrix of f in this basis is:

$$_{u}f_{u} = U^T\mathbf{Q}U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{2}{3} & -\frac{\sqrt{5}}{3} \\ 0 & \frac{\sqrt{5}}{3} & -\frac{2}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}.$$

|||| Exercise 15.50

Find the axis and angle of rotation for the rotation matrix: $\mathbf{Q} = \frac{1}{2}\begin{bmatrix} 0 & -\sqrt{2} & \sqrt{2} \\ \sqrt{2} & 1 & 1 \\ -\sqrt{2} & 1 & 1 \end{bmatrix}$.

Conversely, we can construct a matrix that rotates by any desired angle around any desired axis:

|||| Example 15.51

Problem: Construct the matrix for the linear map $f : \mathbb{R}^3 \to \mathbb{R}^3$ that rotates 3-space around the axis spanned by the vector $\mathbf{a} = (1, 1, 0)$ anti-clockwise by the angle $\pi/2$.

Solution: Choose any orthonormal basis $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ where $\mathbf{u}_1$ points in the direction of $\mathbf{a}$. For example:

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

We have chosen them such that $\det([\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]) = 1$. This means that the orientation of space is preserved by this change of basis, so we know that the rotation from the following construction will be anti-clockwise around the axis.

The matrix with respect to the u-basis that rotates anti-clockwise around the $\mathbf{u}_1$-axis by the angle $\pi/2$ is:

$$ {}_u f_u = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi/2) & -\sin(\pi/2) \\ 0 & \sin(\pi/2) & \cos(\pi/2) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}. $$

The change of basis matrix from u to the standard e-basis is:

$$ {}_e M_u = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix}, $$

so the matrix of f with respect to the standard basis is:

$$ {}_e f_e = {}_e M_u \, {}_u f_u \, {}_e M_u^T = \frac{1}{2} \begin{bmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ -\sqrt{2} & \sqrt{2} & 0 \end{bmatrix}. $$

Note: for the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, it would have made no difference what choice we make as long as they are orthogonal to $\mathbf{a}$, and orthonormal. If we rotate them in the plane orthogonal to $\mathbf{a}$, this rotation will cancel in the formula ${}_e M_u \, {}_u f_u \, {}_e M_u^T$.

|||| Exercise 15.52

Find an orthogonal matrix $\mathbf{Q}$ that, in the standard e-basis for $\mathbb{R}^3$, represents a rotation about the axis spanned by $\mathbf{a} = (1, 1, 1)$ by an angle $\pi/2$.

15.9 Reduction of Quadratic Polynomials

A ***quadratic form*** in $(\mathbb{R}^n, \cdot)$ is a quadratic polynomial in n variables – but without linear and constant terms.

|||| Definition 15.53

Let $\mathbf{A}$ be a symmetric $(n \times n)$-matrix and let $(x_1, x_2, \cdots, x_n)$ denote the coordinates for an arbitrary vector $\mathbf{x}$ in $(\mathbb{R}, \cdot)$ with respect to the standard basis e in $\mathbb{R}^n$.

A ***quadratic form*** in $(\mathbb{R}, \cdot)$ is a function of the n variables $(x_1, x_2, \cdots, x_n)$ in the following form:

$$\begin{aligned} P_{\mathbf{A}}(\mathbf{x}) = P_{\mathbf{A}}(x_1, x_2, \cdots, x_n) &= \left[\begin{array}{ccccc} x_1 & x_2 & \cdot & \cdot & x_n \end{array}\right] \cdot \mathbf{A} \cdot \left[\begin{array}{c} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{array}\right] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot x_i \cdot x_j \quad , \end{aligned} \tag{15-98}$$

a_{ij} being the individual elements in $\mathbf{A}$.

|||| Example 15.54 Quadratic Form as Part of a Quadratic Polynomial

Let $f(x, y)$ be the following quadratic polynomial in the two variables x and y.

$$f(x, y) = 11 \cdot x^2 + 4 \cdot y^2 - 24 \cdot x \cdot y - 20 \cdot x + 40 \cdot y - 60 \quad . \tag{15-99}$$

Then we can separate the polynomial in two parts:

$$f(x, y) = P_{\mathbf{A}}(x, y) + (-20 \cdot x + 40 \cdot y - 60) \quad , \tag{15-100}$$

where $P_{\mathbf{A}}(x, y)$ is the quadratic form

$$P_{\mathbf{A}}(x, y) = 11 \cdot x^2 + 4 \cdot y^2 - 24 \cdot x \cdot y \tag{15-101}$$

that is represented by the matrix

$$\mathbf{A} = \left[\begin{array}{cc} 11 & -12 \\ -12 & 4 \end{array}\right] \tag{15-102}$$

We will now see how the spectral theorem can be used for the description of every quadratic form by use of the eigenvalues for the matrix that represents the quadratic form.

|||| Theorem 15.55 Reduction of Quadratic Forms

Let $\mathbf{A}$ be a symmetric matrix and let $P_{\mathbf{A}}(x_1, \cdots, x_n)$ denote the corresponding quadratic form in $(\mathbb{R}^n, \cdot)$ with respect to standard coordinates. By a change of basis to new coordinates $\widetilde{x}_1, \cdots, \widetilde{x}_n$ given by the positive orthogonal change of basis matrix $\mathbf{Q}$ that diagonalizes $\mathbf{A}$ we get the reduced expression for the quadratic form:

$$P_{\mathbf{A}}(x_1, \cdots, x_n) = \widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}_1, \cdots, \widetilde{x}_n) = \lambda_1 \cdot \widetilde{x}_1^2 + \cdots + \lambda_n \cdot \widetilde{x}_n^2 \quad , \tag{15-103}$$

where $\lambda_1, \cdots, \lambda_n$ are the n real eigenvalues for the symmetric matrix $\mathbf{A}$.

The *reduction* in the theorem means that the new expression does not contain any product terms of the type $x_i \cdot x_j$ for $i \neq j$.

|||| Proof

Since $\mathbf{A}$ is symmetric it *can* according to the spectral theorem be diagonalized by an orthogonal substitution matrix $\mathbf{Q}$. The gathering of column vectors $(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ in $\mathbf{Q}$ constitutes a new basis v in $(\mathbb{R}^n, \cdot)$.

Let $\mathbf{x}$ be an arbitrary vector in $\mathbb{R}^n$. Then we have the following set of coordinates for $\mathbf{x}$, partly with respect to the standard e-basis and partly with respect to the new basis v

$$\begin{aligned} {}_{\mathrm{e}}\mathbf{x} &= (x_1, \cdots, x_n) \quad , \\ {}_{\mathrm{v}}\mathbf{x} &= (\widetilde{x}_1, \cdots, \widetilde{x}_n) \quad . \end{aligned} \tag{15-104}$$

Then

$$\begin{aligned}
P_{\mathbf{A}}(\mathbf{x}) &= P_{\mathbf{A}}(x_1, \cdots, x_n) \\
&= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \cdot \mathbf{A} \cdot \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \cdot \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{-1} \cdot \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \\
&= \left(\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \cdot \mathbf{Q} \right) \cdot \mathbf{\Lambda} \cdot \left(\mathbf{Q}^{\top} \cdot \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \right) \\
&= \begin{bmatrix} \widetilde{x}_1 & \cdots & \widetilde{x}_n \end{bmatrix} \cdot \mathbf{\Lambda} \cdot \begin{bmatrix} \widetilde{x}_1 \\ \cdot \\ \cdot \\ \widetilde{x}_n \end{bmatrix} \\
&= \begin{bmatrix} \widetilde{x}_1 & \cdots & \widetilde{x}_n \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \cdot \begin{bmatrix} \widetilde{x}_1 \\ \cdot \\ \cdot \\ \widetilde{x}_n \end{bmatrix} \\
&= \widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}_1, \cdots, \widetilde{x}_n) \quad = \quad \lambda_1 \cdot \widetilde{x}_1^2 + \cdots + \lambda_n \cdot \widetilde{x}_n^2 \quad .
\end{aligned} \tag{15-105}$$

■

Note that the matrix that represents the quadratic form in Example 15.54, Equation (15-102), is not much different from the Hessian Matrix $\mathbf{H}f(x,y)$ for $f(x,y)$, which is also a constant matrix, because $f(x,y)$ is a second degree polynomial. See eNote **??**. In fact we observe that:

$$\mathbf{A} = \frac{1}{2} \cdot \mathbf{H}f(x,y) \quad , \tag{15-106}$$

and this is no coincidence.

|||| Lemma 15.56

Let $f(x_1, x_2, \cdots, x_n)$ denote an arbitrary quadratic polynomial without linear and constant terms. Then $f(x_1, x_2, \cdots, x_n)$ can be expressed as a quadratic form in exactly one way – i.e. there exists exactly one symmetric matrix $\mathbf{A}$ such that:

$$f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n) = P_{\mathbf{A}}(x_1, x_2, \cdots, x_n) \quad . \tag{15-107}$$

The sought matrix is:

$$\mathbf{A} = \frac{1}{2} \cdot \mathbf{H} f(\mathbf{x}) \quad , \tag{15-108}$$

where $\mathbf{H}f(\mathbf{x})$ is (the constant) Hessian matrix for the function $f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n)$.

|||| Proof

We limit ourselves to the case $n = 2$ and refer the analysis to functions of two variables in eNote **??**: If $f(x, y)$ is a polynomial in two variables without linear (and constant) terms, i.e. a quadratic form in $(\mathbf{R}^2, \cdot)$, then the wanted $\mathbf{A}$-matrix is exactly the (constant) Hesse-matrix for $f(x, y)$.

■

This applies generally, if we extend the definition of Hessian matrices to functions of more variables as follows: Let $f(x_1, x_2, \cdots, x_n)$ be an arbitrary smooth function of n variables in the obvious meaning for functions of more variables (than two). Then the corresponding Hessian matrices are the following symmetric $(n \times n)$-matrices which contain all the second-order partial derivatives for the function $f(\mathbf{x})$ evaluated at an arbitrary point $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{H}f(x_1, x_2, \cdots, x_n) = \begin{bmatrix} f''_{x_1\,x_1}(\mathbf{x}) & f''_{x_1\,x_2}(\mathbf{x}) & \cdots & f''_{x_1\,x_n}(\mathbf{x}) \\ f''_{x_2\,x_1}(\mathbf{x}) & f''_{x_2\,x_2}(\mathbf{x}) & \cdots & f''_{x_2\,x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f''_{x_n\,x_1}(\mathbf{x}) & f''_{x_n\,x_2}(\mathbf{x}) & \ldots & f''_{x_n\,x_n}(\mathbf{x}) \end{bmatrix} \quad . \tag{15-109}$$

In particular if $f(x, y, z)$ is a smooth function of three variables (as in Example 15.57

below) we get at every point $(x,y,z) \in \mathbb{R}^3$:

$$\mathbf{H}f(x,y,z) = \begin{bmatrix} f''_{xx}(x,y,z) & f''_{xy}(x,y,z) & f''_{xz}(x,y,z) \\ f''_{xy}(x,y,z) & f''_{yy}(x,y,z) & f''_{yz}(x,y,z) \\ f''_{xz}(x,y,z) & f''_{yz}(x,y,z) & f''_{zz}(x,y,z) \end{bmatrix} \quad , \tag{15-110}$$

where we explicitly have used the symmetry of the Hessian matrix, e.g. $f''_{zx}(x,y,z) = f''_{xz}(x,y,z)$.

|||| Example 15.57 Quadratic Form with a Representing Matrix

Let $f(x,y,z)$ denote the following function of three variables:

$$f(x,y,z) = x^2 + 3 \cdot y^2 + z^2 - 8 \cdot x \cdot y + 4 \cdot y \cdot z \quad . \tag{15-111}$$

Then $f(x,y,z)$ is a quadratic form $P_{\mathbf{A}}(x,y,z)$ with

$$\mathbf{A} = \frac{1}{2} \cdot \mathbf{H}f(x,y,z) = \frac{1}{2} \cdot \begin{bmatrix} f''_{xx}(x,y,z) & f''_{xy}(x,y,z) & f''_{xz}(x,y,z) \\ f''_{xy}(x,y,z) & f''_{yy}(x,y,z) & f''_{yz}(x,y,z) \\ f''_{xz}(x,y,z) & f''_{yz}(x,y,z) & f''_{zz}(x,y,z) \end{bmatrix} = \begin{bmatrix} 1 & -4 & 0 \\ -4 & 3 & 2 \\ 0 & 2 & 1 \end{bmatrix} \quad . \tag{15-112}$$

We can prove 15-108 by direct computation:

$$\begin{aligned} P_{\mathbf{A}}(x,y,z) &= \begin{bmatrix} x & y & z \end{bmatrix} \cdot \begin{bmatrix} 1 & -4 & 0 \\ -4 & 3 & 2 \\ 0 & 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \\ &= \begin{bmatrix} x & y & z \end{bmatrix} \cdot \begin{bmatrix} x - 4 \cdot y \\ 3 \cdot y - 4 \cdot x + 2 \cdot z \\ z + 2 \cdot y \end{bmatrix} \\ &= x \cdot (x - 4 \cdot y) + y \cdot (3 \cdot y - 4 \cdot x + 2 \cdot z) + z \cdot (z + 2 \cdot y) \\ &= x^2 + 3 \cdot y^2 + z^2 - 8 \cdot x \cdot y + 4 \cdot y \cdot z \\ &= f(x,y,z) \quad . \end{aligned} \tag{15-113}$$

As is shown in Section **??** in eNote **??** the signs of the eigenvalues for the Hessian matrix play a decisive role when we analyse and inspect a smooth function $f(x,y)$ at and about a stationary point. And since it is again the very same Hessian matrix that appears in the present context we will here tie a pair of definitions to this sign-discussion – now for the general $(n \times n)$ Hessian matrices, and thus also for general quadratic forms represented by symmetric matrices $\mathbf{A}$:

|||| Definition 15.58 Definite and Indefinite Symmetric Matrices

We let $\mathbf{A}$ denote a symmetric matrix. Let $\mathbf{A}$ have the n real eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. the we say that

1. $\mathbf{A}$ is ***positive definite*** if all eigenvalues λ_i are positive.
2. $\mathbf{A}$ is ***positive semi-definite*** if all eigenvalues λ_i are non-negative (every eigenvalue is greater than or equal to 0).
3. $\mathbf{A}$ is ***negative definite*** if all eigenvalues λ_i are negative.
4. $\mathbf{A}$ is ***negative semi-definite*** if all eigenvalues λ_i are non-positive (every eigenvalue is less than of equal to 0).
5. $\mathbf{A}$ is ***indefinite*** if $\mathbf{A}$ is neither positive semi-definite nor negative semi-definite.

We now formulate an intuitively reasonable result that relates this ”definiteness” to the values which the quadratic polynomial $P_{\mathbf{A}}(\mathbf{x})$ assumes for different $\mathbf{x} \in \mathbb{R}^n$.

|||| Theorem 15.59 The Meaning of Positive Definiteness

If $\mathbf{A}$ is a symmetric positive definite matrix then the quadratic form $P_{\mathbf{A}}(\mathbf{x})$ is positive for all $\mathbf{x} \in \mathbb{R}^n - \mathbf{0}$.

|||| Proof

We refer to Theorem 15.55 and from that we can use the reduced expression for the quadratic form:

$$P_{\mathbf{A}}(x_1, \cdots, x_n) = \widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}_1, \cdots, \widetilde{x}_n) = \lambda_1 \cdot \widetilde{x}_1^2 + \cdots + \lambda_n \cdot \widetilde{x}_n^2 \quad , \tag{15-114}$$

from which it is clear to see that since $\mathbf{A}$ is positive definite we get $\lambda_i > 0$ for all $i = 1, \cdots, n$ and then $P_{\mathbf{A}}(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$, which corresponds to the fact that none of the sets of coordinates for $\mathbf{x}$ can be $(0, \cdots, 0)$.

■

Similar theorems can be formulated for negative definite and indefinite matrices, and they are obviously useful in investigations of functions, in particular in investigations of the functional values around stationary points, as shown in eNote **??**.

15.10 Reduction of Quadratic Polynomials

By reducing the quadratic form part of a quadratic polynomial we naturally get an equivalently simpler quadratic polynomial – now without product terms. We give a couple of examples.

|||| Example 15.60 Reduction of a Quadratic Polynomial, Two Variables

We consider the following quadratic polynomial in two variables:

$$f(x,y) = 11\cdot x^2 + 4\cdot y^2 - 24\cdot x\cdot y - 20\cdot x + 40\cdot y - 60 \qquad (15\text{-}115)$$

The part of the polynomial that can be described by a quadratic form is now

$$P_{\mathbf{A}}(x,y) = 11\cdot x^2 + 4\cdot y^2 - 24\cdot x\cdot y \quad , \qquad (15\text{-}116)$$

where

$$\mathbf{A} = \begin{bmatrix} 11 & -12 \\ -12 & 4 \end{bmatrix} \quad . \qquad (15\text{-}117)$$

Exactly this matrix is diagonalized by a positive orthogonal substitution **Q** in Example 15.32: The eigenvalues for **A** are $\lambda_1 = 20$ and $\lambda_2 = -5$ and

$$\mathbf{Q} = \begin{bmatrix} 4/5 & 3/5 \\ -3/5 & 4/5 \end{bmatrix} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \quad , \quad \text{where} \quad \varphi = -\arcsin(3/5) \quad . \qquad (15\text{-}118)$$

The change of coordinates $\widetilde{x}, \widetilde{y}$ consequently is a rotation of the standard coordinate system by an angle of $-\arcsin(3/5)$.

We use the reduction theorem 15.55 and get that the quadratic form $P_{\mathbf{A}}(x,y)$ in the new coordinates has the following reduced expression:

$$P_{\mathbf{A}}(x,y) = \widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x},\widetilde{y}) = 20\cdot\widetilde{x}^2 - 5\cdot\widetilde{y}^2 \quad . \qquad (15\text{-}119)$$

By introducing the reduced expression for the quadratic form in the polynomial $f(x,y)$ we get:

$$f(x,y) = 20\cdot\widetilde{x}^2 - 5\cdot\widetilde{y}^2 + (-20\cdot x + 40\cdot y - 60) \quad , \qquad (15\text{-}120)$$

where all that remains is to express the last parenthesis by using the new coordinates. This is done using the substitution matrix $\mathbf{Q}$. We have the linear relation between the coordinates (x, y) and $(\widetilde{x}, \widetilde{y})$:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{Q} \cdot \begin{bmatrix} \widetilde{x} \\ \widetilde{y} \end{bmatrix} = \begin{bmatrix} 4/5 & 3/5 \\ -3/5 & 4/5 \end{bmatrix} \cdot \begin{bmatrix} \widetilde{x} \\ \widetilde{y} \end{bmatrix} \tag{15-121}$$

so that:

$$\begin{aligned} x &= \frac{1}{5} \cdot (4 \cdot \widetilde{x} + 3 \cdot \widetilde{y}) \\ y &= \frac{1}{5} \cdot (-3 \cdot \widetilde{x} + 4 \cdot \widetilde{y}) \quad . \end{aligned} \tag{15-122}$$

We substitute these rewritings of x and y in (15-120) and get:

$$\begin{aligned} f(x,y) &= 20 \cdot \widetilde{x}^2 - 5 \cdot \widetilde{y}^2 + (-4 \cdot (4 \cdot \widetilde{x} + 3 \cdot \widetilde{y}) + 8 \cdot (-3 \cdot \widetilde{x} + 4 \cdot \widetilde{y}) - 60) \\ &= 20 \cdot \widetilde{x}^2 - 5 \cdot \widetilde{y}^2 - 40 \cdot \widetilde{x} + 20 \cdot \widetilde{y} - 60 \quad . \end{aligned} \tag{15-123}$$

Thus we have reduced the expression for $f(x, y)$ to the following expression in new coordinates $\widetilde{x}$ and $\widetilde{y}$, that appears by a suitable rotation of the standard coordinate system:

$$\begin{aligned} f(x,y) &= 11 \cdot x^2 + 4 \cdot y^2 - 24 \cdot x \cdot y - 20 \cdot x + 40 \cdot y - 60 \\ &= 20 \cdot \widetilde{x}^2 - 5 \cdot \widetilde{y}^2 - 40 \cdot \widetilde{x} + 20 \cdot \widetilde{y} - 60 \\ &= \widetilde{f}(\widetilde{x}, \widetilde{y}) \quad . \end{aligned} \tag{15-124}$$

Note again that the *reduction* in Example 15.60 results in the reduced quadratic polynomial $\widetilde{f}(\widetilde{x}, \widetilde{y})$ not containing any product terms of the form $\widetilde{x} \cdot \widetilde{y}$. This reduction technique and the output of the large work becomes somewhat more clear when we consider quadratic polynomials in three variables.

|||| Example 15.61 Reduction of a Quadratic Polynomial, Three Variables

In Example 15.34 we have diagonalized the matrix $\mathbf{A}$ that represents the quadratic form in the following quadratic polynomial in three variables:

$$f(x,y,z) = 7 \cdot x^2 + 6 \cdot y^2 + 5 \cdot z^2 - 4 \cdot x \cdot y - 4 \cdot y \cdot z - 2 \cdot x + 20 \cdot y - 10 \cdot z - 18 \quad . \tag{15-125}$$

This polynomial is reduced to the following quadratic polynomial in the new variables obtained using the same directives as in Example 15.60:

$$\begin{aligned} f(x,y) &= \widetilde{f}(\widetilde{x}, \widetilde{y}, \widetilde{z}) \\ &= \quad 9 \cdot \widetilde{x}^2 + 6 \cdot \widetilde{y}^2 + 3 \cdot \widetilde{z}^2 + 18 \cdot \widetilde{x} - 12 \cdot \widetilde{y} + 6 \cdot \widetilde{z} - 18 \end{aligned} \tag{15-126}$$

with the positive orthogonal substitution

$$\mathbf{Q} = \begin{bmatrix} -2/3 & -2/3 & 1/3 \\ 2/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & 2/3 \end{bmatrix} \quad . \tag{15-127}$$

The substitution matrix $\mathbf{Q}$ can be factorized to a product of axis-rotation matrices like this:

$$\mathbf{Q} = \mathbf{R}_z(w) \cdot \mathbf{R}_y(v) \cdot \mathbf{R}_x(u) \quad , \tag{15-128}$$

where the rotation angles are respectively:

$$u = \frac{\pi}{4} \quad , \quad v = -\arcsin\left(\frac{1}{3}\right) \quad , \quad \text{and} \quad w = 3 \cdot \frac{\pi}{4} \quad , \tag{15-129}$$

By rotation of the coordinate system and by using the new coordinates $\widetilde{x}$, $\widetilde{y}$, and $\widetilde{z}$ we obtain a reduction of the polynomial $f(x, y, z)$ to the end that the polynomial $\widetilde{f}(\widetilde{x}, \widetilde{y}, \widetilde{z})$ does not contain product terms while $f(x, y, z)$ contains two product terms, with $x \cdot y$ and $y \cdot z$, respectively.

15.11 Summary

The main result in this eNote is that symmetric $(n \times n)$-matrices are precisely those matrices that can be diagonalized by a special orthogonal change of basis matrix $\mathbf{Q}$. We have used this theorem for the reduction of quadratic polynomials in n variables – though particularly for $n = 2$ and $n = 3$.

- A symmetric $(n \times n)$-matrix $\mathbf{A}$ has precisely n real eigenvalues $\lambda_1, \cdots, \lambda_n$.
- In the vector space $\mathbb{R}^n$ a scalar product is introduced by extending the standard scalar product of $\mathbb{R}^2$ and $\mathbb{R}^3$, and we refer to this scalar product when we write $(\mathbb{R}^n, \cdot)$. If $\mathbf{a} = (a_1, \cdots, a_n)$ and $\mathbf{b} = (b_1, \cdots, b_n)$ with respect to the standard basis e in $\mathbb{R}^n$, then
$$\mathbf{a} \cdot \mathbf{b} = \sum_i^n a_i \cdot b_i \quad . \tag{15-130}$$
- The length, the norm, of a vector $\mathbf{a}$ is given by
$$|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{\sum_{i=1}^n a_i^2} \quad . \tag{15-131}$$
- The Cauchy-Schwarz inequality is valid for all vectors $\mathbf{a}$ and $\mathbf{b}$ in $(\mathbb{R}^n, \cdot)$
$$|\mathbf{a} \cdot \mathbf{b}| \leq |\,\mathbf{a}\,|\;|\,\mathbf{b}\,|\;, \tag{15-132}$$
and the equality sign applies if and only if $\mathbf{a}$ and $\mathbf{b}$ are linearly dependent.
- The angle $\theta \in [0, \pi]$ between two proper vectors $\mathbf{a}$ and $\mathbf{b}$ in $(\mathbb{R}^n, \cdot)$ is determined by
$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| \cdot |\mathbf{b}|} \quad . \tag{15-133}$$
- Two proper vectors $\mathbf{a}$ and $\mathbf{b}$ in $(\mathbb{R}^n, \cdot)$ are orthogonal if $\mathbf{a} \cdot \mathbf{b} = 0$.
- A matrix $\mathbf{Q}$ is orthogonal if the column vectors are pairwise orthogonal and each has length 1 with respect to the scalar product introduced. This corresponds exactly to
$$\mathbf{Q}^\top \cdot \mathbf{Q} = \mathbf{E} \tag{15-134}$$
or equivalently:
$$\mathbf{Q}^{-1} = \mathbf{Q}^\top \quad . \tag{15-135}$$

- The spectral theorem: If $\mathbf{A}$ is symmetric, then a special orthogonal change of basis matrix $\mathbf{Q}$ exists such that

  $$\mathbf{A} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^{\top} \quad , \tag{15-136}$$

  where $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \cdots, \lambda_n)$.

- Every special orthogonal matrix $\mathbf{Q}$ is change of basis matrix that rotates the coordinate-system. It can for $n = 3$ be factorized in three axis-rotation matrices:

  $$\mathbf{Q} = \mathbf{R}_z(w) \cdot \mathbf{R}_y(v) \cdot \mathbf{R}_x(u) \quad , \tag{15-137}$$

  for suitable choices of rotation angles u, v, and w.

- For $n = 3$: By rotation of the coordinate-system, i.e. by use of a special orthogonal change of basis matrix $\mathbf{Q}$, the quadratic form $P_{\mathbf{A}}(x, y, z)$ (which is a quadratic polynomial without linear terms and without constant terms) can be expressed by a quadratic form $\widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}, \widetilde{y}, \widetilde{z})$ in the new coordinates $\widetilde{x}$, $\widetilde{y}$, and $\widetilde{z}$ such that

  $$P_{\mathbf{A}}(x, y, z) = \widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}, \widetilde{y}, \widetilde{z}) \quad \text{for all } (x, y, z), \tag{15-138}$$

  and such that the reduced quadratic form $\widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}, \widetilde{y}, \widetilde{z})$ does not contain any product term of the type $\widetilde{x} \cdot \widetilde{y}$, $\widetilde{x} \cdot \widetilde{z}$, or $\widetilde{y} \cdot \widetilde{z}$:

  $$\widetilde{P}_{\mathbf{\Lambda}}(\widetilde{x}, \widetilde{y}, \widetilde{z}) = \lambda_1 \cdot \widetilde{x}^2 + \lambda_2 \cdot \widetilde{y}^2 + \lambda_3 \cdot \widetilde{z}^2 \quad , \tag{15-139}$$

  where λ_1, λ_2, and λ_3 are the three real eigenvalues for $\mathbf{A}$.

eNote 16

First Order Linear Differential Equations

In this eNote we first give a short introduction to differential equations in general and then the main subject is a special type of differential equation the so-called first order differential equations. The eNote is based on knowledge of special functions, differential and integral calculus and linear maps.

Version 19.10.18 *by Karsten Schmidt/Jesper Kampmann Larsen. Updated 2.11.21 David Brander.*

16.1 What Is a Differential Equation?

A *differential equation* is an equation, in which one or more unknown *functions* appear together with one or more of their derivatives. In this eNote we only consider differential equations that contain one unknown function. Differential equations naturally occur in the modelling of physical, mechanical, economic, chemical and manifold other problems, and this is why it is an important subject.

One says that a differential equation has the *order* n if it contains the nth derivative of the unknown function, but no derivatives of order higher than n. The unknown function is in this eNote denoted by x or $x(t)$, if the name of the independent variable t is important in the context.

An example of a differential equation is

$$x'''(t) - 2x'(t) + x(t) = t\,, \quad t \in \mathbb{R}. \tag{16-1}$$

The equation has *order* 3, since the highest number of times the unknown function x is differentiated in the equation is 3. A *solution* to the equation is a function x_0 which, inserted into the equation, makes it true. If we, for example, want to investigate whether the function

$$x_0(t) = \mathrm{e}^t + t + 2\,, \quad t \in \mathbb{R}$$

is a solution to (16-1), we test this by insertion of x_0 in place of x in the equation. Since

$$\begin{aligned} x_0'''(t) - 2x_0'(t) + x_0(t) &= (\mathrm{e}^t + t + 2)''' - 2(\mathrm{e}^t + t + 2)' + ((\mathrm{e}^t + t + 2)) \\ &= \mathrm{e}^t - 2(\mathrm{e}^t + 1) + \mathrm{e}^t + t + 2 \\ &= t, \end{aligned}$$

x_0 is a solution.

This eNote is about an important type of first order differential equation, the so-called *linear* differential equations. In order to be able to investigate these precisely we first express them in a standard way.

16.2 Introduction to First Order Linear Differential Equations

|||| Definition 16.1

By a first order *linear* differential equation we understand a differential equation that can be brought into the standard form

$$x'(t) + p(t)x(t) = q(t)\,, \quad t \in I \tag{16-2}$$

where I is an open interval in $\mathbb{R}$, and p and q are (known) continuous functions defined on I.

The equation is called *homogeneous* if $q(t) = 0$ for all t. Otherwise it is called *inhomogeneous*.

|||| Example 16.2 Standard Form

The first order differential equation

$$x'(t) + 2x(t) = 30 + 8t\,, \quad t \in \mathbb{R}\,. \tag{16-3}$$

is immediately seen to be in standard form (16-2) with $p(t) = 2$ and $q(t) = 30 + 8t\,, \quad t \in \mathbb{R}\,.$ Therefore it is linear.

|||| Example 16.3 Standard form

Let I be an open interval in $\mathbb{R}$. Consider the first order linear differential equation

$$t \cdot x'(t) + 2x(t) - 8t^2 = -10\,, \quad t \in \mathbb{R}\,. \tag{16-4}$$

In order to bring this into standard form we first have to add $8t^2$ to both sides of the equation, since on the left-hand side only terms containing the unknown function $x(t)$ must appear. Then we divide both sides by t, since the coefficient of $x'(t)$ must be 1 in the standard form. To avoid division by 0, we must assume that t is either greater or less than zero: let us choose the first:

$$x'(t) + \frac{2}{t}\,x(t) = 8t - \frac{10}{t}, \quad t > 0\,. \tag{16-5}$$

Now the differential equation is in the standard form with $p(t) = \dfrac{2}{t}$ and $q(t) = 8t - \dfrac{10}{t}, \quad t > 0\,.$

|||| Exercise 16.4 Standard Form

Explain why the first order differential equation

$$x'(t) + \frac{1}{2}\,t^2 = 0\,, \quad t \in \mathbb{R}$$

is not homogeneous.

Exercise 16.5 More Solutions

Given the differential equation

$$x'(t) + 2x(t) = 30 + 8t\,, \quad t \in \mathbb{R}.$$

Show that, for any of $c = 1$, $c = 2$ or $c = 3$, the function

$$x_0(t) = 13 + 4t + ce^{-2t}\,, \quad t \in \mathbb{R}$$

is a solution.

From Exercise 16.5 it appears that a differential equation can have more than one solution. We will in what follows investigate in more detail the question about the number of solutions. In order to understand precisely what is meant by a first order differential equation being linear and what this means for its solution set, we will need the following lemma.

Lemma 16.6

Let p be a continuous function defined on an open interval I in $\mathbb{R}$. Then the map $f : C^1(I) \to C^0(I)$ given by

$$f(x(t)) = x'(t) + p(t)x(t) \tag{16-6}$$

is linear.

Proof

We will show that f satisfies the two linearity requirements L_1 and L_2. Let $x_1, x_2 \in C^1(I)$ (i.e. the two functions are arbitrary differentiable functions with continuous derivatives on I), and let $k \in \mathbb{R}$. That f satisfies L_1 appears from

$$\begin{aligned} f(x_1(t) + x_2(t)) &= (x_1(t) + x_2(t))' + p(t)(x_1(t) + x_2(t)) \\ &= x_1'(t) + x_2'(t) + p(t)(x_1(t) + p(t)x_2(t) \\ &= (x_1'(t) + p(t)x_1(t)) + (x_2'(t) + p(t)x_2(t)) \\ &= f(x_1(t)) + f(x_2(t))\,. \end{aligned}$$

That f satisfies L_2 appears from

$$\begin{aligned} f(kx_1(t)) &= (kx_1'(t)) + p(t)(kx_1(t)) = k(x_1'(t) + p(t)x_1(t)) \\ &= kf(x_1(t))\,. \end{aligned}$$

By this the proof is completed.

■

From Lemma 16.6 we can deduce important properties for the solution set for first order linear differential equations. First we introduce convenient notations for the solution sets that we will treat.

L_{inhom} denotes all solutions for a given inhomogeneous differential equation. L_{inhom} is briefly known as the *solution set* or the *general solution*.

L_{hom} denotes the solution set for a homogeneous differential equation corresponding to an inhomogeneous equation (where the right-hand side $q(t)$ is replaced by 0).

|||| Theorem 16.7 Three Properties

For a first order linear differential equation $x'(t) + p(t)x(t) = q(t)\,,\ t \in I$:

1. If the equation is homogeneous (i.e. $q(t)$ is the 0-function), then the solution set is a vector subspace of $C^1(I)$.

2. *Structure Theorem*: If the equation is inhomogeneous the general solution L_{inhom} can be written in the form

$$L_{inhom} = x_0(t) + L_{hom} \tag{16-7}$$

   where $x_0(t)$ is a *particular solution* to the inhomogeneous differential equation, and L_{hom} is the solutions set to the corresponding homogeneous differential equation.

3. *Superposition principle*: If $x_1(t)$ is a solution when the right-hand side of the differential equation is replaced by the function $q_1(t)$, and $x_2(t)$ is a solution when the right-hand side is replaced by the function $q_2(t)$, then $x_1(t) + x_2(t)$ is a solution when the right-hand side is replaced by the function $q_1(t) + q_2(t)$.

|||| Proof

We consider the map between vector spaces, $f : C^1(I) \to C^0(I)$ given by

$$f(x(t)) = x'(t) + p(t)x(t)\,. \tag{16-8}$$

This is, according to Lemma 16.6, linear. Therefore we have:

1. L_{hom} is equal to $\ker(f)$. Since the kernel for every linear map is a subspace of the domain, L_{hom} is a subspace of $C^1(I)$.
2. Since the equation $f(x(t)) = x'(t) + p(t)x(t) = q(t)$ is linear, the structure theorem follows directly from the general structure theorem for linear equations (see eNote 12, Theorem 12.14).
3. The superposition principle follows from the fact that f satisfies the linearity requirement L_1. Assume that $f(x_1(t)) = q_1(t)$ and $f(x_2(t)) = q_2(t)$. Then

$$f(x_1(t) + x_2(t)) = f(x_1(t)) + f(x_2(t)) = q_1(t) + q_2(t)\,.$$

By this the proof is completed.

■

When we call a first order differential equation of the form (16-2) linear, it is – as shown above – closely related to the fact that its left-hand side represents a linear map, and that its solution set therefore has the unique properties of Theorem 16.7. In the following example we juggle with the properties in order to decide whether a given differential equation is not linear.

|||| Example 16.8 First Order Differential Equation That Is Nonlinear

We consider a first order differential equation

$$x'(t) - (x(t))^2 = q(t)\,,\ t \in \mathbb{R}\,. \tag{16-9}$$

where we in the usual way have isolated the terms that contain the unknown function on the left-hand side. The left-hand side represents the map $f : C^1(\mathbb{R}) \to C^0(\mathbb{R})$ given by

$$f(x(t)) = x'(t) - (x(t))^2\,. \tag{16-10}$$

Here we will show that one, in different ways, can demonstrate that the differential equation is not linear.

1. We can show directly that f does not satisfy the linearity conditions. To show this we can test L_2, e.g. with $k = 2$. We compute the two sides in L_2 :

$$\begin{aligned} f(2x(t)) &= (2x(t))' - (2x(t))^2 = 2x'(t) - 4(x(t))^2 \\ 2f(x(t)) &= 2(x'(t) - (x(t))^2) = 2x'(t) - 2(x(t))^2 . \end{aligned}$$

   By subtraction of the two equation we get:

$$f(2x(t)) - 2f(x(t)) = -2(x(t))^2$$

   where the right-hand side is only the 0-function when $x(t)$ is the 0-function. Since L_2 applies for all $x(t) \in C^1(\mathbb{R})$, L_2 is not satisfied. Therefore the equation is nonlinear.

2. The solution set to the corresponding homogeneous equation is not a subspace. E.g. it does not satisfy the stability requirement with repsect to multiplication by a scalar which we can show as follows:

   The function $x_0(t) = -\frac{1}{t}$ is a solution to the homogeneous eqution because

$$x_0'(t) - (x_0(t))^2 = \frac{1}{t^2} - \frac{1}{t^2} = 0 .$$

   But $2 \cdot x_0(t) = -\frac{2}{t}$ is not, because

$$(2 \cdot x_0(t))' - (2 \cdot x_0(t))^2 = \frac{2}{t^2} - \frac{4}{t^2} = -\frac{2}{t^2} \neq 0 .$$

   Therefore the differential equation is not linear.

3. The solution set does not satisfy the superposition principle. E.g. we see that

$$f\left(-\frac{1}{t}\right) = 0 \quad \text{and} \quad f\left(\frac{1}{t}\right) = -\frac{2}{t^2}, \quad \text{while}$$
$$f\left(-\frac{1}{t} + \frac{1}{t}\right) = 0 \neq 0 - \frac{2}{t^2} .$$

   Therefore the differential equation is not linear.

It follows from the structure theorem that homogeneous equations play a special role for linear differential equations. Therefore we treat them separately in the next section.

16.3 Homogeneous First Order Linear Differential Equations

We now establish a solution formula for homogeneous first order linear equations.

|||| Theorem 16.9 Solution of the Homogeneous Equation

Let $p(t)$ be a continuous function defined on an open real interval I, and let $P(t)$ be an arbitrary antiderivative for $p(t)$, i.e., a function satisfying $P'(t) = p(t)$.

The general solution for the homogeneous first order linear differential equation

$$x'(t) + p(t)x(t) = 0\,,\ t \in I\,. \tag{16-11}$$

is then given by

$$x(t) = c\,\mathrm{e}^{-P(t)}\,,\ t \in I \tag{16-12}$$

where c is an arbitrary real number.

|||| Proof

The theorem follows from the fact that the derivative of a function $g(t)$ is zero on an interval if and only if that function is constant. We apply this to the function $g(t) = x(t)e^{P(t)}$. Using the chain rule and the product rule for differentiation we have:

$$\begin{aligned}\left(x(t)e^{P(t))}\right)' &= e^{P(t)}x'(t) + p(t)e^{P(t)}x(t) \\ &= e^{P(t)}\left(x'(t) + p(t)x(t)\right).\end{aligned}$$

Since $e^{P(t)} \neq 0$, the above expression is zero if and only if the equation (16-11) holds. That is, the differential equation (16-11) is equivalent to the equation:

$$\left(x(t)e^{P(t))}\right)' = 0.$$

As mentioned, this is equivalent to the statement:

$$x(t)e^{P(t)} = c, \tag{16-13}$$

where c is some real constant, i.e. that $x(t) = ce^{-P(t)}$. This shows that not only is $ce^{-P(t)}$ a solution, for any constant c, but that *any* solution to (16-11) must be of this form, since it must satisfy Equation (16-13) for some c.

■

We already know that the solution set is a subspace of $C^1(I)$. From the formula (16-12) we now know that the subspace is 1-dimensional, and that the function $\mathrm{e}^{-P(t)}$ is a basis for the solution set.

Remark 16.10

Theorem 16.9 is also valid if p is a continuous *complex*-valued function, with the slight modification that the arbitrary constant c is now a complex constant. The proof is exactly the same, because the product rule is the same for complex-valued functions of t, and, writing $P(t) = u(t) + iv(t)$, one finds that the derivative of $e^{P(t)}$ is still $P'(t)e^{P(t)}$. Finally, by separating the function into real and imaginary parts, one again finds that the derivative of a complex-valued function is zero if and only if the function is equal to a complex constant.

Exercise 16.11

In Theorem 16.9 an arbitrary antiderivative $P(t)$ for $p(t)$ is used. Explain why it is immaterial to the solution set which antiderivative you use when you apply the theorem.

Example 16.12 Solution of a Homogeneous Equation

A homogeneous first order linear differential equation is given by

$$x'(t) + \cos(t)x(t) = 0, \quad t \in \mathbb{R}\,. \tag{16-14}$$

We see that that the coefficient function $p(t) = \cos(t)$. An antiderivative for $p(t)$ is $P(t) = \sin(t)$. Then the general solution can be written as

$$x(t) = c\mathrm{e}^{-P(t)} = c\mathrm{e}^{-\sin(t)}, \quad t \in \mathbb{R} \tag{16-15}$$

where c is an arbitrary real number.

16.4 Inhomogeneous Equations Solved by the Guess Method

Now that we know how to find the general solution for homogeneous first order linear differential equations, it is about time to look at the inhomogeneous ones. If you already know or can guess a particular solution to the inhomogeneous equation, it is obvious to use the structure theorem, see Theorem 16.7. This is demonstrated in the following examples.

|||| Example 16.13 Solution Using a Guess and the Structure Theorem

An inhomogeneous first order linear differential equation is given by

$$x'(t) + tx(t) = t, \quad t \in \mathbb{R}. \tag{16-16}$$

It is easily seen that $x_0(t) = 1$ is a particular solution. Then we solve the corresponding homogeneous differential equation

$$x'(t) + tx(t) = 0, \quad t \in \mathbb{R}. \tag{16-17}$$

Using symbols from Theorem 16.9 we have $p(t) = t$ that has the antiderivative

$$P(t) = \frac{1}{2}t^2.$$

The general solution therefore consists of the following functions where c is an arbitrary real number:

$$x(t) = ce^{-\frac{1}{2}t^2}, \quad t \in \mathbb{R}. \tag{16-18}$$

In short:

$$L_{hom} = \left\{ ce^{-\frac{1}{2}t^2}, \ t \in \mathbb{R} \;\middle|\; c \in \mathbb{R} \right\}. \tag{16-19}$$

Now we can establish the general solution to the inhomogeneous differential equation using the structure theorem as:

$$L_{inhom} = x_0(t) + L_{hom} = \left\{ 1 + ce^{-\frac{1}{2}t^2}, \ t \in \mathbb{R} \;\middle|\; c \in \mathbb{R} \right\}.$$

|||| Example 16.14 Solution Using a Guess and the Structure Theorem

An inhomogeneous first order linear differential equation is given by

$$x'(t) + 2x(t) = 30 + 8t, \quad t \in \mathbb{R}\,. \tag{16-20}$$

First let us try to guess a particular solution. Since the right-hand side is first degree polynomial, one can – with the given left-hand side, where you only differentiate and multiply by 2 – assume that a first degree polynomial could be a solution. Therefore we try to insert an arbitrary first degree polynomial $x_0(t) = b + at$ in the left-hand side of the differential equation:

$$x_0'(t) + 2x_0(t) = (b + at)' + 2(b + at) = a + 2b + 2at\,.$$

We compare the resulting expression with the given right-hand side:

$$a + 2b + 2at = 30 + 8t$$

that is satisfied for all $t \in \mathbb{R}$ eactly when

$$a + 2b = 30 \text{ and } 2a = 8 \Leftrightarrow a = 4 \text{ and } b = 13\,.$$

Thus we have found a particular solution

$$x_0(t) = 13 + 4t\,,\; t \in \mathbb{R}\,.$$

Then we solve the corresponding homogeneous differential equation

$$x'(t) + 2x(t) = 0, \quad t \in \mathbb{R}\,. \tag{16-21}$$

Using symbols from Theorem 16.9 we have $p(t) = 2$ that has the antiderivative $P(t) = 2t$. Therefore the general solution consists of the following functions where c is an arbitrary real number:

$$x(t) = ce^{-2t}, \quad t \in \mathbb{R}\,. \tag{16-22}$$

In short:

$$L_{hom} = \left\{\, ce^{-2t}\,,\; t \in \mathbb{R} \;\middle|\; c \in \mathbb{R} \,\right\}. \tag{16-23}$$

Now its is possible to establish the general solution to the inhomogeneous differential equation using the structure theorem:

$$L_{inhom} = x_0(t) + L_{hom} = \left\{\, 13 + 4t + ce^{-2t}\,,\; t \in \mathbb{R} \;\middle|\; c \in \mathbb{R} \,\right\}.$$

|||| Example 16.15 Solution Using a Guess and the Structure Theorem

An inhomogeneous first order linear differential equation is given by

$$x'(t) + x(t) = 1 + \sin(2t), \quad t \geq 0 \,. \tag{16-24}$$

First let us try to guess a particular solution. Since the right-hand side consists of constant plus a sine function with the angular frequency 2, it is obvious to guess a solution the type

$$x(t) = k + a\cos(2t) + b\sin(2t)\,.$$

By insertion of this in the differential equation we get:

$$\begin{aligned} &-2a\sin(2t) + 2b\cos(2t) + k + a\cos(2t) + b\sin(2t) = 1 + \sin(2t) \\ &\Leftrightarrow (2b+a)\cos(2t) + (b-2a-1)\sin(2t) + (k-1)\,1 = 0\,. \end{aligned}$$

Since the set $(\cos(2t), \sin(2t), 1)$ is linearly independent, this equation is satisfied exactly when

$$2b + a = 0\,,\; b - 2a - 1 = 0 \text{ and } k = 1 \;\Leftrightarrow a = -\frac{2}{5}\,,\; b = \frac{1}{5} \text{ and } k = 1\,.$$

By this we have found a particular solution

$$x_0(t) = 1 - \frac{2}{5}\cos(2t) + \frac{1}{5}\sin(2t)\,,\; t \in \mathbb{R}\,.$$

Since the corresponding homogeneous differential equation

$$x'(t) + x(t) = 0, \quad t \geq 0 \tag{16-25}$$

evidently has the general solution

$$x(t) = ce^{-t}, \quad t \geq 0\,, \tag{16-26}$$

we get the general solution to the given inhomogeneous differential equation by use of the structure theorem:

$$L_{inhom} = x_0(t) + L_{hom} = \left\{ 1 - \tfrac{2}{5}\cos(2t) + \tfrac{1}{5}\sin(2t) + ce^{-t}\,,\; t \in \mathbb{R} \;\middle|\; c \in \mathbb{R} \right\}.$$

As demonstrated in the three previous examples it makes sense to use the guess method in the inhomogeneous cases, when you already know a particular solution or easily can find one. It only requires that you can find an antiderivative $P(t)$ for the coefficient function $p(t)$.

Otherwise if you do not have an immediate particular solution, you must use the general solution formula (see below) instead. Here you get rid of the guesswork, but you must find two antiderivatives, one is $P(t)$ as above, while the other often is somewhat more difficult (if not impossible) to find, since you must integrate a product of functions. In the following section we establish the general solution formula and discuss the said problems.

16.5 The General Solution Formula

Now we consider the general first order linear differential equation in the standard form

$$x'(t) + p(t)x(t) = q(t), \quad t \in I, \tag{16-27}$$

We can determine the general solution using the following general formula.

|||| Theorem 16.16 The General Solution Formula

Let $p(t)$ and $q(t)$ be continuous functions on an open real interval I, and let $P(t)$ be an arbitrary antiderivative to $p(t)$. The differential equation

$$x'(t) + p(t)x(t) = q(t), \quad t \in I \tag{16-28}$$

then has the general solution

$$x(t) = \mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t)dt + c\mathrm{e}^{-P(t)}, \quad t \in I \tag{16-29}$$

where c is an arbitrary real number.

|||| Proof

The second term in the solution formula (16-29) we identify as L_{hom}. If we can show that the first term is a particular solution to the differential equation, then it follows from the structure theorem that the solution formula is the general solution to the differential equation.

First we must of course ask ourselves whether the indefinite integral that is part of the solution formula even exists. It does! See a detailed reasoning for this in the proof of the existence and uniqueness Theorem 16.24. That the first term

$$x_0(t) = \mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t) dt$$

is a particular solution we show by testing. We insert the term in left-hand side of the differential equation and see that the result is equal to the right-hand side.

$$\begin{aligned} x_0'(t) + p(t)x_0(t) &= \left(\mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t)\,\mathrm{d}t\right)' + p(t)\,\mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t)\,\mathrm{d}t \\ &= -p(t)\mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t)\,\mathrm{d}t + \mathrm{e}^{-P(t)}\mathrm{e}^{P(t)} q(t) + p(t)\mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t) dt \\ &= q(t)\,. \end{aligned}$$

By this the proof is completed.

■

|||| Remark 16.17

Using Remark 16.10, it is straightforward to show that Theorem 16.16 is also valid if $p(t)$ and $q(t)$ are continuous *complex*-valued functions, with the modification that the arbitrary constant c is a complex constant.

If one inserts $q(t) = 0$ in the general solution formula (16-29), the first term disappears, and what is left is the second term that is the formula (16-12) for homogeneous equation. Therefore the formula (16-29) is a "‘general formula"’ that covers both the homogeneous and the inhomogeneous case.

|||| Exercise 16.18

The solution formula (16-29) includes the indefinite integral $\int \mathrm{e}^{P(t)} q(t) dt$, that represents an arbitrary antiderivative of $e^{P(t)} q(t)$. Explain why it does not matter to the solution set which antiderivative you choose to use, when you apply the formula.

Now we give a few examples using the general solution formula. Since it contains an indefinite integral of a product of functions you will often need *integration by parts*, which the second example demonstrates.

|||| Example 16.19 Solution Using the General Formula

Given the differential equation

$$x'(t)+\frac{2}{t}x(t)=8t-\frac{10}{t},\quad t>0\,. \tag{16-30}$$

With the symbols in the general solution formula we have $p(t)=\frac{2}{t}$ and $q(t)=8t-\frac{10}{t}$. An antiderivative for $p(t)$ is given by:

$$P(t)=2\ln t\,. \tag{16-31}$$

We then have

$$\mathrm{e}^{-P(t)}=\mathrm{e}^{-2\ln t}=\mathrm{e}^{\ln(t^{-2})}=t^{-2}=\frac{1}{t^2}\,. \tag{16-32}$$

From this it follows that $\mathrm{e}^{P(t)}=t^2$. Now we use the general solution formula:

$$\begin{aligned}x(t)&=\mathrm{e}^{-P(t)}\int \mathrm{e}^{P(t)}q(t)\,\mathrm{d}t+c\mathrm{e}^{-P(t)}\\&=\frac{1}{t^2}\int t^2\left(8t-\frac{10}{t}\right)\,\mathrm{d}t+c\frac{1}{t^2}\\&=\frac{1}{t^2}\int(8t^3-10t)\,\mathrm{d}t+c\frac{1}{t^2}\\&=\frac{1}{t^2}\left(2t^4-5t^2+c\right)\\x(t)&=2t^2-5+\frac{c}{t^2}\,,\quad t>0.\end{aligned} \tag{16-33}$$

The general solution consists of these functions where c is an arbitrary real number. In short:

$$L_{inhom}=\left\{\,x(t)=2t^2-5+\frac{c}{t^2}\,,\ t>0\;\middle|\;c\in\mathbb{R}\,\right\}. \tag{16-34}$$

|||| Example 16.20 Solution Using the General Formula

We will solve the differential equation

$$x'(t)-\frac{1}{t}x(t)=t^2\sin(2t),\quad t>0\,. \tag{16-35}$$

With the symbols in the general solution formula we have $p(t)=-\frac{1}{t}$ and $q(t)=t^2\sin(2t)$. An antiderivative for $p(t)$ is given by:

$$P(t)=-\ln t\,. \tag{16-36}$$

We then have

$$\mathrm{e}^{-P(t)} = \mathrm{e}^{\ln t} = t \;\text{ and }\; \mathrm{e}^{P(t)} = \mathrm{e}^{-\ln t} = (\mathrm{e}^{\ln t})^{-1} = \frac{1}{t}\,. \tag{16-37}$$

Now we use the general solution formula::

$$\begin{aligned} x(t) &= \mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t)\,\mathrm{d}t + c\mathrm{e}^{-P(t)} \\ &= t \int \frac{1}{t} t^2 \sin(2t)\,\mathrm{d}t + ct \\ &= t \int t \sin(2t)\,\mathrm{d}t + ct\,. \end{aligned}$$

Now we perform an intermediate computation where we use integration by parts to find the antiderivative.

$$\begin{aligned} \int t\sin(2t) dt &= -\frac{1}{2} t\cos(2t) - \int -\frac{1}{2}\cos(2t)\,\mathrm{d}t \\ &= -\frac{1}{2} t\cos(2t) + \frac{1}{2}\int \cos(2t)\,\mathrm{d}t \\ &= -\frac{1}{2} t\cos(2t) + \frac{1}{4}\sin(2t)\,. \end{aligned}$$

And return to the computation

$$\begin{aligned} x(t) &= t \int t\sin(2t)\,\mathrm{d}t + ct \\ &= t\left(-\frac{1}{2} t\cos(2t) + \frac{1}{4}\sin(2t)\right) + ct \\ x(t) &= -\frac{1}{2} t^2\cos(2t) + \frac{1}{4} t\sin(2t) + ct \quad t > 0\,. \end{aligned}$$

The general solution consists of these functions where c is an arbitrary real number. In short:

$$L_{inhom} = \left\{\, x(t) = -\tfrac{1}{2} t^2\cos(2t) + \tfrac{1}{4} t\sin(2t) + ct\,,\; t > 0 \;\middle|\; c \in \mathbb{R} \,\right\}. \tag{16-38}$$

Until now we have considered the general solution to the differential equation. Often one is interested in a particular solution that for a given value of t assumes a desired functional value, a so-called initial value problem. We treat this in the next section.

16.6 Initial Value Problems

We consider a first order linear differential equation in its standard form

$$x'(t) + p(t)x(t) = q(t), \quad t \in I\,. \tag{16-39}$$

If we need a solution to the equation that for a given value of t assumes a desired functional value, the following questions arise: 1) Is there even a solution that satisfies the desired properties and 2) If yes, how many solutions are there? Before we answer these question generally, we consider a couple of examples.

|||| Example 16.21 An Initial Value Problem

In the Example 16.13 we found the general solution to the differential equation

$$x'(t) + tx(t) = t, \quad t \in \mathbb{R}. \tag{16-40}$$

viz.

$$x(t) = 1 + c\mathrm{e}^{-\frac{1}{2}t^2}, \; t \in \mathbb{R}$$

where c is an arbitrary real number.

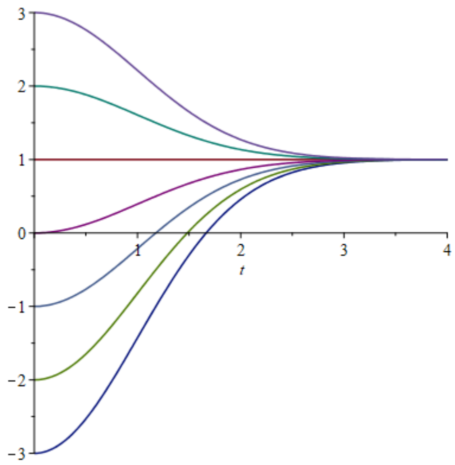
Now we will find the solution $x_0(t)$ that satisfies the initial value condition $x_0(0) = 3$. This is done by insertion of the initial value in the general solution, whereby we determine c:

$$x_0(0) = 1 + c\mathrm{e}^{-\frac{1}{2}\cdot 0^2} = 1 + c = 3 \Leftrightarrow \; c = 2. \tag{16-41}$$

Therefore the conditioned solution function to the differential equation is given by

$$x_0(t) = 1 + 2\mathrm{e}^{-\frac{1}{2}t^2}, \; t \in \mathbb{R}. \tag{16-42}$$

The figure below shows the graphs for the seven solutions that correspond to initial value conditions $x_0(0) = b$ where $b \in \{-3, -2, -1, 0, 1, 2, 3\}$. The solution we just found is the uppermost. The others are found in a similar way.



|||| Example 16.22 An Initial Value Problem

In Example 16.20 we found the general solution to the differential equation

$$x'(t) + \frac{2}{t}x(t) = 8t - \frac{10}{t}, \quad t > 0, \tag{16-43}$$

viz.

$$x(t) = 2t^2 - 5 + \frac{c}{t^2}, \quad t > 0$$

where c is an arbitrary real number.

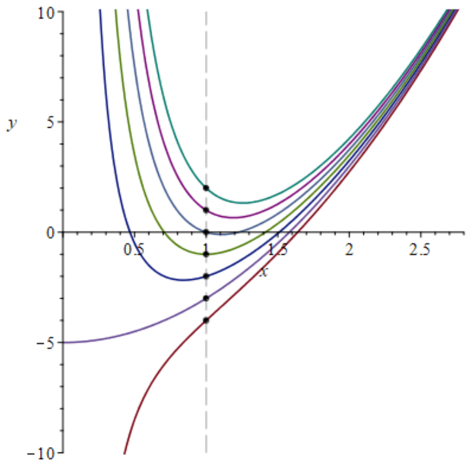
Now we will find the particular solution $x_0(t)$ that satisfies the initial value condition $x_0(1) = 2$. It is done by insertion of initial value in the general solution, whereby we determine c:

$$x_0(1) = 2 \cdot 1^2 - 5 + \frac{c}{1^2} 2 - 5 + c = 2 \;\Leftrightarrow\; c = 5\,. \tag{16-44}$$

Therefore the conditioned solution function to the differential equation is given by

$$x_0(t) = 2t^2 - 5 + \frac{5}{t^2}\,, \quad t > 0\,. \tag{16-45}$$

The figure below shows the graphs for the seven solutions that correspond to initial value conditions $x_0(0) = b$ where $b \in \{-4, -3, -2, -1, 0, 1, 2\}$. The solution we just found is the uppermost. The others are found in a similar way.



|||| Example 16.23 The Stationary Response

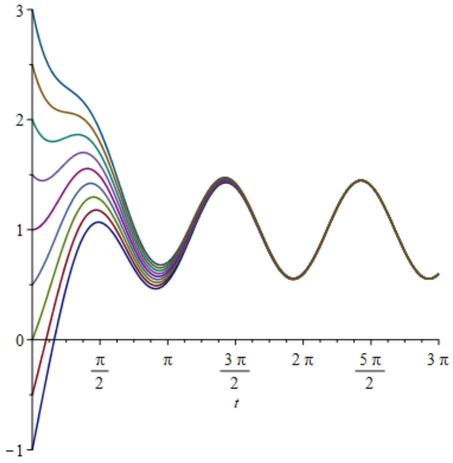
In Example 16.15 we found the general solution to the differential equation

$$x'(t) + x(t) = 1 + \sin(2t), \quad t \geq 0 \tag{16-46}$$

viz.

$$x(t) = 1 - \frac{2}{5}\cos(2t) + \frac{1}{5}\sin(2t) + ce^{-t}, \quad t \geq 0\,. \tag{16-47}$$

Here we show a series of solutions with the initial values from -1 to 3 for $t = 0$:



The figure indicates that all solutions approach a periodic oscillation when $t \to \infty$. That this is the case is seen from the general solution of the differential equation where the fourth term ce^{-t} regardless of the choice for c is negligible due to the negative exponent. The first three terms constitute the *the stationary response*.

In the three preceding examples we did not have any difficulties in finding a solution to the differential equation that satisfied a given initial condition. In fact we saw that, for each of the initial value conditions considered, exactly one solution that satisfied the condition exists. That this applies in general we show in the following theorem.

Theorem 16.24 Existence and Uniqueness of Solutions

Given the differential equation

$$x'(t) + p(t)x(t) = q(t), \quad t \in I \tag{16-48}$$

where I is an open interval and $p(t)$ and $q(t)$ are continuous functions on I.

Then: for every number pair (t_0, b) exactly one (particular) solution $x_0(t)$ to the differential equation exists that satisfies the *inital value condition*

$$x_0(t_0) = b\,. \tag{16-49}$$

Proof

From Theorem 16.16 we know that the set of solutions to the differential equation (16-48) is given by

$$x(t) = \mathrm{e}^{-P(t)} \int \mathrm{e}^{P(t)} q(t) dt + c\mathrm{e}^{-P(t)} \tag{16-50}$$

where c is an arbitrary real number.

Let us first investigate the indefinite integral that is included in the formula. Does it exist? This is equivalent to asking: does an antiderivative for the function under the integration sign exist? We must start with $p(t)$. Since it is continuous, it has an antiderivative which we call $P(t)$. Being an antiderivative, $P(t)$ is differentiable and thus continuous. Since the exponential function is also continuous the composite function $\mathrm{e}^{P(t)}$ is continuous. Finally since $q(t)$ is continuous, the product $\mathrm{e}^{P(t)}q(t)$ is continuous.

By this we have shown that the function under the integration sign is continuous. Therefore it has an antiderivative, in fact infinitely many antiderivatives that only differ from each other by constants. We choose an arbitrary antiderivative and call it $F(t)$. Now we can reformulate the solution formula as

$$x(t) = \mathrm{e}^{-P(t)} F(t) + c\mathrm{e}^{P(t)} \tag{16-51}$$

where c is an arbitrary real number. Then we insert the initial value condition:

$$x(t_0) = \mathrm{e}^{-P(t_0)} F(t_0) + c\mathrm{e}^{-P(t_0)} = b \;\Leftrightarrow\; c = F(t_0) + b\mathrm{e}^{-P(t_0)}$$

where we first multiplied by $\mathrm{e}^{P(t_0)}$ on both sides of the equality sign and then isolated c. Thus in the general solution set exactly one solution exists that satisfies the initial value condition, viz. the one that emerge when we in (16-51) insert the found value of c.

By this the proof is completed.

■

Exercise 16.25

Again let us consider the linear map $f : C^1(I) \rightarrow C^0(I)$ that represents the left-hand side of a first order linear differential equation:

$$f(x(t)) = x'(t) + p(t)x(t) \tag{16-52}$$

We know that $\ker(f)$ is one dimensional and has the basis vector $\mathrm{e}^{-P(t)}$. But what is the image space (the range) for f?

We end this section by an example that shows how it is possible to “go backwards” from a given general solution to the differential equation it solves.

Example 16.26 From Solution to the Differential Equation

The general solution to a first order inhomogeneous differential equation is given by

$$L_{inhom} = \left\{ \, x(t) = t\mathrm{e}^{-5t} + ct \, , \;\; t > 0 \;\middle|\; c \in \mathbb{R} \, \right\} . \tag{16-53}$$

Determine the corresponding differential equation that has the form

$$x'(t) + p(t)x(t) = q(t) \, . \tag{16-54}$$

(That is, determine $p(t)$ and $q(t)$).

First we consider the corresponding homogeneous differential equation. With the structure theorem in mind we immediately see that

$$L_{hom} = \left\{ \, x(t) = ct \, , \;\; t > 0 \;\middle|\; c \in \mathbb{R} \, \right\}$$

By insertion of $x(t) = ct$ in the homogeneous equation $x'(t) + p(t)x(t) = 0$ we get

$$c + p(t)ct = 0\,, \tag{16-55}$$

and since this equation must hold for all c

$$p(t) = -\frac{1}{t}\,. \tag{16-56}$$

Since we now know $p(t)$, it only remains to determine the right-hand side $q(t)$. We find this by insertion of the particular solution $x(t) = t\mathrm{e}^{-5t}$ into the left-hand side of the equation.

$$\mathrm{e}^{-5t} - 5t\mathrm{e}^{-5t} - \frac{1}{t}\cdot t\mathrm{e}^{-5t} = -5t\mathrm{e}^{-5t} = q(t)\,. \tag{16-57}$$

Now since both $p(t)$ and $q(t)$ are determined, the whole differential equation is determined as:

$$x'(t) - \frac{1}{t}x(t) = -5t\mathrm{e}^{-5t}\,,\ t > 0\,. \tag{16-58}$$

16.7 Finite Dimensional Domain

In some cases we know in advance what type of solutions to the differential equation are of interest. Therefore one can choose to restrict the domain $C^1(\mathbb{R})$. We end this eNote with an example where the domain is a finite dimensional subset of $C^1(\mathbb{R})$ which leads to the introduction of matrix methods.

|||| Example 16.27 Solution by Matrix Compution

Consider the differential equation

$$x'(t) + (1-2t)x(t) = 7t - 4t^3\,. \tag{16-59}$$

In this example we are only interested in solutions that belong to the polynomial space $P_2(\mathbb{R})$, i.e. the subset of $C^1(\mathbb{R})$ that has the monomial base $(1, t, t^2)$.

To find the range $f\,(P_2(\mathbb{R}))$ of the linear map f that represents the left-hand side of the differential equation, we first determine the images of the basis vectors:

$$f(1) = 1 - 2t\,,\ f(t) = 1 + t - 2t^2 \ \text{ and }\ f(t^2) = 2t + t^2 - 2t^3\,.$$

Since $P_3(\mathbb{R})$ has the monomial base $(1, t, t^2, t^3)$, and the found images lie in their span, we see that the range $f\,(P_2(\mathbb{R}))$ is a subspace of $P_3(\mathbb{R})$.

We want to solve the equation

$$f(x(t)) = 7t - 4t^3,$$

which can be expressed in matrix form as

$$\mathbf{F}\,\mathbf{x} = \mathbf{b},$$

where $\mathbf{F}$ is the mapping matrix for f with respect to the monomial bases in $P_2(\mathbb{R})$ and $P_3(\mathbb{R})$, $\mathbf{x}$ is the coordinate matrix for the unknown polynomial with respect to the monomial basis in $P_2(\mathbb{R})$, and $\mathbf{b}$ is the coordinate matrix for the right-hand side of the differential equation with respect to the monomial basis in $P_3(\mathbb{R})$.

Thus, when restricted to $P_2(\mathbb{R})$, the differential equation becomes an inhomogeneous system of linear equations. The first three columns of the augmented matrix $\mathbf{T}$ of the system are given by $\mathbf{F}$, while the fourth column is $\mathbf{b}$:

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ -2 & 1 & 2 & 7 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & -4 \end{bmatrix} \rightarrow \operatorname{rref}(\mathbf{T}) = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since the rank of $\mathbf{T}$ is seen to be 3, the differential equation has only one solution. Since the fourth column in $\operatorname{rref}(\mathbf{T})$ states the coordinate vector of the solution with respect to the monomial basis in $P_2(\mathbb{R})$, the solution can immediately be stated as:

$$x_0(t) = -1 + t + 2t^2\,.$$

|||| Exercise 16.28

1. Solve the differential equation in Example 16.27 by the guess method or the general solution formula.
2. How does the general solution differ from the one found in the example?

|||| Exercise 16.29

Replace the right-hand side in the differential equation in Example 16.27 by the function $q(t) = 1$.

1. Show, using matrix computation, that the differential equation does not have a solution in the subspace $P_2(\mathbb{R})$ given in the example.
2. Using Maple (or other software), find the solution $x_0(t)$ to the differential equation that satisfies the initial value condition $x_0(t) = 0$ and draw its graph.

eNote 17

Systems of Linear First-Order Differential Equations

This eNote describes systems of linear first-order differential equations with constant coefficients and shows how these can be solved. The eNote is based on eNote 16, which describes linear differential equations in general. Thus it is a good idea to read that eNote first. Moreover eigenvalues and eigenvectors are used in the solution procedure, see eNotes 13 and 14. (Updated: 9.11.21 by David Brander).

Here we consider coupled homogeneous linear first-order differential equations with constant coefficients (see Explanation 17.1). Such a collection of ***coupled differential equations*** is called a ***system of differential equations***. A system of n first-order differential equations with constant coefficients looks like this:

$$\begin{array}{ccccccccc} x_1'(t) & = & a_{11}x_1(t) & + & a_{12}x_2(t) & + & \dots & + & a_{1n}x_n(t) \\ x_2'(t) & = & a_{21}x_1(t) & + & a_{22}x_2(t) & + & \dots & + & a_{2n}x_n(t) \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ x_n'(t) & = & a_{n1}x_1(t) & + & a_{n2}x_2(t) & + & \dots & + & a_{nn}x_n(t) \end{array} \tag{17-1}$$

On the left hand side of the system the derivatives of the n unknown functions $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ are written. Every right hand side is a linear combination of the n unknown functions. The coefficients (the a's) are real constants. In matrix form the system can be written like this:

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_n'(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \tag{17-2}$$

Even more compactly it can be written like this

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t) \tag{17-3}$$

A is called the ***system matrix***. It is now the aim to solve such a system of differential equations, that is, we wish to determine $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))$.

|||| Explanation 17.1 What Is a System of Differential Equations?

Systems of differential equations are collections of differential equations. The reason we do not consider the differential equations individually, is that they cannot be solved independently, because the unknown functions are present in more equations, that is, the equations are *coupled*. A single differential equation from a system can e.g. look like this:

$$x_1'(t) = 4x_1(t) - x_2(t) \tag{17-4}$$

It is not possible to determine neither $x_1(t)$ nor $x_2(t)$, since there are two unknown functions, but only one differential equation.

In order to be able to find the full solution to such an equation one should have as many equations as one has unknown equations (with corresponding derivatives). Thus the second equation in the system might be:

$$x_2'(t) = -6x_1(t) + 2x_2(t) \tag{17-5}$$

We now have as many equations (two), as we have unknown functions (two), and it is now possible to determine both $x_1(t)$ and $x_2(t)$.

For greater clarity we write the system of differential equations in matrix form. The system above looks like this:

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \begin{bmatrix} 4 & -1 \\ -6 & 2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{x}'(t) = \begin{bmatrix} 4 & -1 \\ -6 & 2 \end{bmatrix} \mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) \tag{17-6}$$

Disregarding that are operating with vectors and matrices the system of equations looks like something we have seen before: $x'(t) = A \cdot x(t)$, something we were able to solve in high school. The solution to this differential equation is trivial: $x(t) = ce^{At}$, where c is an arbitrary constant. Below we find that the solution to the corresponding system of differential equations is similar in structure to $x(t) = ce^{At}$.

We now solve the system of differential equations in the following Theorem 17.2. The theorem contains requirements that are not always satisfied. The special cases where the theorem is not valid are investigated later. The proof uses a well-known method, the so-called ***diagonalization method***.

Theorem 17.2

Let $\mathbf{A} \in^{n\times n}$. A system of linear differential equations consisting of n equations with a total of n unknown functions is given by

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in \mathbb{R}\,. \tag{17-7}$$

If $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ corresponding to (not necessarily different) eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the general solution of the system is determined by

$$\mathbf{x}(t) = c_1 \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + c_2 \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 + \ldots + c_n \mathrm{e}^{\lambda_n t}\mathbf{v}_n, \quad t \in \mathbb{R}\,, \tag{17-8}$$

where $c_1, c_2, \ldots, c_n$ are arbitrary complex constants.

Note that it is not always possible to find n linearly independent eigenvectors. Therefore Theorem 17.2 cannot always be applied to the solution of systems of first-order differential equations.

In the theorem we use the general complex solution for the system of differential equations. Therefore the general real solution can be found as the real subset of the complex solution.

Proof

We guess that a solution to the system of differential equations $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$ is a vector $\mathbf{v}$ multiplied by $\mathrm{e}^{\lambda t}$, λ being a constant, such that $\mathbf{x}(t) = \mathrm{e}^{\lambda t}\mathbf{v}$. We then have the derivative

$$\mathbf{x}'(t) = \lambda \mathrm{e}^{\lambda t}\mathbf{v}\,. \tag{17-9}$$

If this expression for $\mathbf{x}'(t)$ is substituted into (17-7) together with the expression for $\mathbf{x}(t)$ we get:

$$\lambda \mathrm{e}^{\lambda t}\mathbf{v} = \mathbf{A}\mathrm{e}^{\lambda t}\mathbf{v} \Leftrightarrow \mathbf{A}\mathbf{v} - \lambda\mathbf{v} = 0 \Leftrightarrow (\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = 0 \tag{17-10}$$

$\mathrm{e}^{\lambda t}$ is non-zero for every $t \in \mathbb{R}$, and can thus be eliminated. The resulting equation is an eigenvalue problem. λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{v}$ is the corresponding eigenvector. They can both be determined. We have now succeeded in finding that $\mathrm{e}^{\lambda t}\mathbf{v}$ is one solution to the system of differential equations, when λ is an eigenvalue and $\mathbf{v}$ the corresponding eigenvector of $\mathbf{A}$.

In order to find the general solution we use the so-called ***diagonalization method***:

We suppose that $\mathbf{A} = \mathbf{A}_{n\times n}$ has n linearly independent (real or complex) eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. We now introduce the invertible matrix $\mathbf{V}$, that contains all the eigenvectors:

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \tag{17-11}$$

Furthermore we introduce the function $\mathbf{y}$ with $\mathbf{y}(t) = (y_1(t), y_2(t), \ldots, y_n(t))$ such that

$$\mathbf{x}(t) = \mathbf{V}\mathbf{y}(t) \tag{17-12}$$

We then get $\mathbf{x}'(t) = \mathbf{V}\mathbf{y}'(t)$. If these expressions for $\mathbf{x}(t)$ og $\mathbf{x}'(t)$ are substituted into Equation (17-7) we get

$$\mathbf{V}\mathbf{y}'(t) = \mathbf{A}\mathbf{V}\mathbf{y}(t) \Leftrightarrow \mathbf{y}'(t) = \mathbf{V}^{-1}\mathbf{A}\mathbf{V}\mathbf{y}(t) = \mathbf{\Lambda}\mathbf{y}(t), \tag{17-13}$$

where $\mathbf{\Lambda} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_n)$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$.

We now get the equation $\mathbf{y}'(t) = \mathbf{\Lambda}\mathbf{y}(t)$, which can be written in the following way:

$$\begin{aligned} y_1'(t) &= \lambda_1 y_1(t) \\ y_2'(t) &= \lambda_2 y_2(t) \\ &\vdots \\ y_n'(t) &= \lambda_n y_n(t) \end{aligned} \tag{17-14}$$

since $\mathbf{\Lambda}$ only has non-zero elements in the diagonal. In this system the single equations are uncoupled: each of the equations only contains one function and its derivative. Therefore we can solve them independently and the general solution for every equation is $y(t) = c\mathrm{e}^{\lambda t}$ for all $c \in \mathbb{C}$. In total this yields the general solution consisting of the functions below for all $c_1, c_2, \ldots, c_n \in \mathbb{C}$:

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} = \begin{bmatrix} c_1\mathrm{e}^{\lambda_1 t} \\ c_2\mathrm{e}^{\lambda_2 t} \\ \vdots \\ c_n\mathrm{e}^{\lambda_n t} \end{bmatrix} \tag{17-15}$$

Since we now have the solution $\mathbf{y}(t)$ we can also find the solution $\mathbf{x}(t) = \mathbf{V}\mathbf{y}(t)$:

$$\begin{aligned} \mathbf{x}(t) &= \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} c_1\mathrm{e}^{\lambda_1 t} \\ c_2\mathrm{e}^{\lambda_2 t} \\ \vdots \\ c_n\mathrm{e}^{\lambda_n t} \end{bmatrix} \\ &= c_1\mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + c_2\mathrm{e}^{\lambda_2 t}\mathbf{v}_2 + \ldots + c_n\mathrm{e}^{\lambda_n t}\mathbf{v}_n \,. \end{aligned} \tag{17-16}$$

Now we have found the general complex solution to the system of equations in Equation (17-7) consisting of the functions $\mathbf{x}(t)$ for all $c_1, c_2, \ldots, c_n \in \mathbb{C}$.

■

|||| Example 17.3

Given the system of differential equations

$$\begin{aligned} x_1'(t) &= \ \ x_1(t) + 2x_2(t) \\ x_2'(t) &= 3x_1(t) \end{aligned} \tag{17-17}$$

Which in matrix form is

$$\mathbf{x}'(t) = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix} \mathbf{x}(t) = \mathbf{A}\mathbf{x}(t). \tag{17-18}$$

It can be shown that $\mathbf{A}$ has the eigenvalues $\lambda_1 = 3$ and $\lambda_2 = -2$ with the eigenvectors $\mathbf{v}_1 = (1, 1)$ and $\mathbf{v}_2 = (2, -3)$ (try for yourself!). Therefore the general real solution to the system of differential equations is given by the functions below for all $c_1, c_2 \in \mathbb{R}$:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_1 \mathrm{e}^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \mathrm{e}^{-2t} \begin{bmatrix} 2 \\ -3 \end{bmatrix}, \quad t \in \mathbb{R} \tag{17-19}$$

The solution is found using Theorem 17.2. Another way of writing the solution is to separate the system of equations so that

$$\begin{aligned} x_1(t) &= c_1 \mathrm{e}^{3t} + 2c_2 \mathrm{e}^{-2t} \\ x_2(t) &= c_1 \mathrm{e}^{3t} - 3c_2 \mathrm{e}^{-2t} \end{aligned} \tag{17-20}$$

constitutes the general solution, where $t \in \mathbb{R}$, for all $c_1, c_2 \in \mathbb{R}$.

17.1 Two Coupled Differential Equations

Given a linear homogeneous first order system of differential equations with constant coefficients with n equations and n unknown functions

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t) . \quad t \in \mathbb{R} \tag{17-21}$$

If the system matrix $\mathbf{A}$ has n linearly independent eigenvectors, the real solution can be found using Theorem 17.2. If the eigenvalues are real then the real solution can be written directly following formula (17-8) in the theorem, where the n corresponding linearly independent eigenvectors are real and the arbitrary constants are stated as being real. If the system matrix has eigenvalues that are not real then the real solution can be found by extracting the real subset of the complex solution. Also in this case the solution can be written as a linear combination of n linearly independent real solutions to the system of differential equations.

We are left with the special case in which the system matrix does not have n linearly independent eigenvectors. Also in this case the real solution will be a linear combination of n linearly independent real solutions to the system of differential equations. Here the method of diagonalization obviously cannot be used and one has to resort to other methods.

In this section we show the three cases above for systems consisting of $n = 2$ coupled differential equations with 2 unknown functions.

|||| Method 17.4

The general real solution to the system of differential equations

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in \mathbb{R}, \tag{17-22}$$

consisting of $n = 2$ equations with 2 unknown functions can be written as

$$\mathbf{x}(t) = c_1\mathbf{u}_1(t) + c_2\mathbf{u}_2(t)\,, \quad t \in \mathbb{R}, \tag{17-23}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are real linearly independent particular solutions and $c_1, c_2 \in \mathbb{R}$.

First determine the eigenvalues of $\mathbf{A}$. For the roots of the characteristic polynomial $\mathbf{A}$ there are three possibilities:

- **Two real single roots.** In this case both of the eigenvalues λ_1 and λ_2 have the algebraic multiplicity 1 and geometric multiplicity 1 and we can put

$$\mathbf{u}_1(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 \quad \text{and} \quad \mathbf{u}_2(t) = \mathrm{e}^{\lambda_2 t}\mathbf{v}_2, \tag{17-24}$$

  where $\mathbf{v}_1$ and $\mathbf{v}_2$ are proper eigenvectors of λ_1 and λ_2, respectively.

- **Two complex roots.** The two eigenvalues λ and $\bar{\lambda}$ are then conjugate complex numbers. We then determine $\mathbf{u}_1$ and $\mathbf{u}_2$ using Method 17.5.

- **One double root.** Here the eigenvalue λ has the algebraic multiplicity 2. If the geometric multiplicity of λ is 1, $\mathbf{u}_1$ and $\mathbf{u}_2$ are determined using method 17.7.

In the first case in Method 17.4 with two different real eigenvalues, Theorem 17.2 can be used directly with the arbitrary constants chosen as real, see Example 17.3.

Now follows the method that covers the case with two complex eigenvalues.

|||| Method 17.5

Two linearly independent real solutions to the system of equations

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in \mathbb{R}, \tag{17-25}$$

where $\mathbf{A}$ has the complex pair of eigenvalues $\lambda = \alpha + \beta i$ and $\bar{\lambda} = \alpha - \beta i$ with corresponding eigenvectors $\mathbf{v}$ and $\bar{\mathbf{v}}$, are

$$\begin{aligned} \mathbf{u}_1(t) &= \mathrm{Re}\left(\mathrm{e}^{\lambda t}\mathbf{v}\right) = \mathrm{e}^{\alpha t}\left(\cos(\beta t)\mathrm{Re}(\mathbf{v}) - \sin(\beta t)\mathrm{Im}(\mathbf{v})\right) \\ \mathbf{u}_2(t) &= \mathrm{Im}\left(\mathrm{e}^{\lambda t}\mathbf{v}\right) = \mathrm{e}^{\alpha t}\left(\sin(\beta t)\mathrm{Re}(\mathbf{v}) + \cos(\beta t)\mathrm{Im}(\mathbf{v})\right) \end{aligned} \tag{17-26}$$

|||| Example 17.6

Given the system of differential equations

$$\mathbf{x}'(t) = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) \tag{17-27}$$

We wish to determine the general real solution.

The eigenvalues are determined as $\lambda = 1 + i$ and $\bar{\lambda} = 1 - i$, respectively, with the corresponding eigenvectors $\mathbf{v} = (-i, 1)$ and $\bar{\mathbf{v}} = (i, 1)$, respectively. We see that there are two complex eigenvalues and their corresponding complex eigenvectors. With $\lambda = 1 + i$ we get

$$\mathbf{v} = \begin{bmatrix} -i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + i\begin{bmatrix} -1 \\ 0 \end{bmatrix} = \mathrm{Re}(\mathbf{v}) + i\mathrm{Im}(\mathbf{v}) \tag{17-28}$$

If we use Method 17.5 we then get the two solutions:

$$\mathbf{u}_1(t) = \mathrm{e}^t\left(\cos(t)\begin{bmatrix} 0 \\ 1 \end{bmatrix} - \sin(t)\begin{bmatrix} -1 \\ 0 \end{bmatrix}\right) = \mathrm{e}^t\begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix} \tag{17-29}$$

$$\mathbf{u}_2(t) = \mathrm{e}^t\left(\sin(t)\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \cos(t)\begin{bmatrix} -1 \\ 0 \end{bmatrix}\right) = \mathrm{e}^t\begin{bmatrix} -\cos(t) \\ \sin(t) \end{bmatrix} \tag{17-30}$$

The general real solution to the system of differential equations (17-27) is then given by the following functions for all $c_1, c_2 \in \mathbb{R}$:

$$\mathbf{x}(t) = c_1\mathbf{u}_1(t) + c_2\mathbf{u}_2(t) = \mathrm{e}^t\left(c_1\begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix} + c_2\begin{bmatrix} -\cos(t) \\ \sin(t) \end{bmatrix}\right), \quad t \in \mathbb{R} \tag{17-31}$$

found using Method 17.4.

Finally we describe the method that can be used if the system matrix has the eigenvalue λ with $\mathrm{am}(\lambda) = 2$ and $\mathrm{gm}(\lambda) = 1$, that is when diagonalization is not possible.

|||| Method 17.7

If the system matrix $\mathbf{A}$ to the system of differential equations

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in \mathbb{R}, \tag{17-32}$$

has one eigenvalue λ with algebraic multiplicity 2, but the corresponding eigenvector space only has geometric multiplicity 1, there are two linearly independent solutions to the system of differential equations of the form:

$$\begin{aligned} \mathbf{u}_1(t) &= \mathrm{e}^{\lambda t}\mathbf{v} \\ \mathbf{u}_2(t) &= t\mathrm{e}^{\lambda t}\mathbf{v} + \mathrm{e}^{\lambda t}\mathbf{b}, \end{aligned} \tag{17-33}$$

where $\mathbf{v}$ is the eigenvector corresponding to λ and $\mathbf{b}$ is a solution to the following linear system:

$$(\mathbf{A} - \lambda\mathbf{E})\mathbf{b} = \mathbf{v} \tag{17-34}$$

|||| Proof

It is evident that one solution to the system of differential equations is $\mathbf{u}_1(t) = \mathrm{e}^{\lambda t}\mathbf{v}$. The difficulty is to find another solution.

We guess at a solution in the form

$$\mathbf{u}_2(t) = t\mathrm{e}^{\lambda t}\mathbf{v} + \mathrm{e}^{\lambda t}\mathbf{b} = \mathrm{e}^{\lambda t}(t\mathbf{v} + \mathbf{b}), \tag{17-35}$$

where $\mathbf{v}$ is an eigenvector corresponding to λ. We then have

$$\mathbf{u}_2{}'(t) = (\mathrm{e}^{\lambda t} + \lambda t\mathrm{e}^{\lambda t})\mathbf{v} + \lambda\mathrm{e}^{\lambda t}\mathbf{b} = \mathrm{e}^{\lambda t}((1+\lambda t)\mathbf{v} + \lambda\mathbf{b}) \tag{17-36}$$

We check whether $\mathbf{u}_2(t)$ is a solution by substitution into $\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t)$:

$$\begin{aligned} \mathbf{u}_2{}'(t) &= \mathbf{A}\mathbf{u}_2(t) \Leftrightarrow \\ (1+\lambda t)\mathbf{v} + \lambda\mathbf{b} &= \mathbf{A}(t\mathbf{v} + \mathbf{b}) \Leftrightarrow \\ t(\lambda\mathbf{v} - \mathbf{A}\mathbf{v}) + (\mathbf{v} + \lambda\mathbf{b} - \mathbf{A}\mathbf{b}) &= \mathbf{0} \Leftrightarrow \\ \lambda\mathbf{v} - \mathbf{A}\mathbf{v} = \mathbf{0} \wedge \mathbf{v} + \lambda\mathbf{b} - \mathbf{A}\mathbf{b} &= \mathbf{0} \end{aligned} \tag{17-37}$$

The first equation can easily be transformed into $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, which is seen to be true, since $\mathbf{v}$ is an eigenvector corresponding to λ. The other equation is transformed into:

$$\begin{aligned} \mathbf{v} + \lambda\mathbf{b} - \mathbf{A}\mathbf{b} = \mathbf{0} &\Leftrightarrow \\ \mathbf{A}\mathbf{b} - \lambda\mathbf{b} = \mathbf{v} &\Leftrightarrow \\ (\mathbf{A} - \lambda\mathbf{E})\mathbf{b} = \mathbf{v} & \end{aligned} \tag{17-38}$$

If $\mathbf{b}$ satisfies the given system of equations, $\mathbf{u}_2(t)$ will also be a solution to the system of differential equations. We now have found two solutions and we have to find out whether these are linearly independent. This is done by a normal linearity criterion: If the equation $k_1\mathbf{u}_1 + k_2\mathbf{u}_2 = \mathbf{0}$ only has the solution $k_1 = k_2 = 0$ then $\mathbf{u}_1$ and $\mathbf{u}_2$ are linearly independent.

$$\begin{aligned} k_1\mathbf{u}_1 + k_2\mathbf{u}_2 = \mathbf{0} &\Rightarrow \\ k_1 e^{\lambda t}\mathbf{v} + k_2(te^{\lambda t}\mathbf{v} + \mathbf{b}e^{\lambda t}) = \mathbf{0} &\Leftrightarrow \\ t(k_2\mathbf{v}) + (k_1\mathbf{v} + k_2\mathbf{b}) = \mathbf{0} &\Leftrightarrow \\ k_2\mathbf{v} = \mathbf{0} \wedge k_1\mathbf{v} + k_2\mathbf{b} = \mathbf{0} & \end{aligned} \tag{17-39}$$

Since $\mathbf{v}$ is an eigenvector, it is not the zero-vector, and hence $k_2 = 0$ according to the first equation. Thus the other equation is reduced to $k_1\mathbf{v} = \mathbf{0}$, and with the same argument we get $k_1 = 0$. Therefore the two solutions are linearly independent, and thus the method has been proved.

■

|||| Example 17.8

Given the system of differential equations

$$\mathbf{x}'(t) = \begin{bmatrix} 16 & -1 \\ 4 & 12 \end{bmatrix}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t). \tag{17-40}$$

The eigenvalues for $\mathbf{A}$ are determined:

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{E}) = \begin{vmatrix} 16-\lambda & -1 \\ 4 & 12-\lambda \end{vmatrix} &= (16-\lambda)(12-\lambda) + 4 \\ &= \lambda^2 - 28\lambda + 196 = (\lambda - 14)^2 = 0 \end{aligned} \tag{17-41}$$

There is only one eigenvalue, viz. $\lambda = 14$, even though it is a 2×2-system. The eigenvectors are determined:

$$\mathbf{A} - 14\mathbf{E} = \begin{bmatrix} 16-14 & -1 \\ 4 & 12-14 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \to \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 0 \end{bmatrix} \tag{17-42}$$

We then obtain the eigenvector $(\frac{1}{2}, 1)$ or $\mathbf{v} = (1, 2)$. We can then conclude that the eigenvalue λ has the algebraic multiplicity 2, but that the corresponding eigenvector space has the geometric multiplicity 1. In order to determine two independent solutions to the system of differential equations we can use Method 17.7.

First we solve the following system of equations:

$$(\mathbf{A}-\lambda\mathbf{E})\mathbf{b}=\mathbf{v}\Rightarrow\begin{bmatrix}2 & -1\\ 4 & -2\end{bmatrix}\mathbf{b}=\begin{bmatrix}1\\ 2\end{bmatrix} \tag{17-43}$$

$$\begin{bmatrix}2 & -1 & 1\\ 4 & -2 & 2\end{bmatrix}\rightarrow\begin{bmatrix}1 & -\frac{1}{2} & \frac{1}{2}\\ 0 & 0 & 0\end{bmatrix} \tag{17-44}$$

This yields $\mathbf{b}=(1,1)$, if the free parameter is put at1. The two solutions then are

$$\begin{aligned}\mathbf{u}_1(t)&=\mathrm{e}^{14t}\begin{bmatrix}1\\ 2\end{bmatrix}\\ \mathbf{u}_2(t)&=t\mathrm{e}^{14t}\begin{bmatrix}1\\ 2\end{bmatrix}+\mathrm{e}^{14t}\begin{bmatrix}1\\ 1\end{bmatrix}.\end{aligned} \tag{17-45}$$

By use of Method 17.4 the general solution can be determined to the following functions for all $c_1, c_2 \in \mathbb{R}$:

$$\mathbf{x}(t)=c_1\mathbf{u}_1(t)+c_2\mathbf{u}_2(t)=c_1\mathrm{e}^{14t}\begin{bmatrix}1\\ 2\end{bmatrix}+c_2\mathrm{e}^{14t}\left(t\begin{bmatrix}1\\ 2\end{bmatrix}+\begin{bmatrix}1\\ 1\end{bmatrix}\right). \tag{17-46}$$

17.2 n-Dimensional Solution Space

In the preceding section we have considered coupled systems consisting of two linear equations with two unknown functions. The solution space is two-dimensional, since it can be written as a linear combination of two linearly independent solutions. This can be generalized to arbitrary systems with $n \geq 2$ coupled linear differential equations with n unknown functions: The solution is a linear combination of exactly n linearly independent solutions. This is formulated in a general form in the following theorem.

|||| Theorem 17.9

Given the linear homogeneous first order system of differential equations with constant real coefficients

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in \mathbb{R}, \tag{17-47}$$

consisting of n equations and with n unknown functions. The general real solution to the system is n-dimensional and can be written as

$$\mathbf{x}(t) = c_1\mathbf{u}_1(t) + c_2\mathbf{u}_2(t) + \cdots + c_n\mathbf{u}_n(t), \tag{17-48}$$

where $\mathbf{u}_1(t), \mathbf{u}_2(t), \ldots, \mathbf{u}_n(t)$ are linearly independent real solutions to the system of differential equations and $c_1, c_2, \ldots, c_n \in \mathbb{R}$.

Below is an example with a coupled system of three differential equations that exemplifies Theorem 17.9.

|||| Example 17.10 Advanced

Given the system of differential equations

$$\mathbf{x}'(t) = \begin{bmatrix} -9 & 10 & 0 \\ -3 & 1 & 5 \\ 1 & -4 & 6 \end{bmatrix} \mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) \tag{17-49}$$

We wish to determine the general real solution to the system of differential equations. Eigenvalues and eigenvectors can be determined and are as follows:

$$\begin{aligned} \lambda_1 = -4 \quad &: \quad \mathbf{v}_1 = (10, 5, 1) \\ \lambda_2 = 1 \quad &: \quad \mathbf{v}_2 = (5, 5, 3) \end{aligned}$$

Moreover λ_2 has the algebraic multiplicity 2, but the corresponding eigenvector space has the geometric multiplicity 1. Because $n = 3$ we need 3 linearly independent solutions to construct the general solution, as seen in 17.9. The eigenvalues are considered separately:

1) The first eigenvalue, $\lambda_1 = -4$, has both geometric and algebraic multiplicity equal to 1. This yields exactly one solution

$$\mathbf{u}_1(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 = \mathrm{e}^{-4t}\begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix} \tag{17-50}$$

2) The other eigenvalue, $\lambda_2 = 1$, has algebraic multiplicity 2, but geometric multiplicity 1.

Therefore we can use method 17.7 in order to find two solutions. First **b** is determined:

$$(\mathbf{A}-\lambda_2\mathbf{E})\mathbf{b}=\mathbf{v}_2 \Rightarrow \begin{bmatrix} -10 & 10 & 0 \\ -3 & 0 & 5 \\ 1 & -4 & 5 \end{bmatrix}\mathbf{b}=\begin{bmatrix} 5 \\ 5 \\ 3 \end{bmatrix} \tag{17-51}$$

A particular solution to this system of equations is $\mathbf{b}=(0,\frac{1}{2},1)$. With this knowledge we have two additional linearly independent solutions to the system of differential equations:

$$\begin{aligned} \mathbf{u}_2(t) &= \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 = \mathrm{e}^{t}\begin{bmatrix} 5 \\ 5 \\ 3 \end{bmatrix} \\ \mathbf{u}_3(t) &= t\mathrm{e}^{\lambda_2 t}\mathbf{v}_2 + \mathrm{e}^{\lambda_2 t}\mathbf{b} = t\mathrm{e}^{t}\begin{bmatrix} 5 \\ 5 \\ 3 \end{bmatrix} + \mathrm{e}^{t}\begin{bmatrix} 0 \\ \frac{1}{2} \\ 1 \end{bmatrix} \end{aligned} \tag{17-52}$$

We leave it to the reader to show that all three solutions are linearly independent.

According to Method 17.9 the general real solution consists of the following linear combination for all $c_1, c_2, c_3 \in \mathbb{R}$:

$$\mathbf{x}(t) = c_1\mathbf{u}_1(t) + c_2\mathbf{u}_2(t) + c_3\mathbf{u}_3(t) \tag{17-53}$$

Thus this yields

$$\mathbf{x}(t) = c_1\mathrm{e}^{-4t}\begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix} + c_2\mathrm{e}^{t}\begin{bmatrix} 5 \\ 5 \\ 3 \end{bmatrix} + c_3\left(t\mathrm{e}^{t}\begin{bmatrix} 5 \\ 5 \\ 3 \end{bmatrix} + \mathrm{e}^{t}\begin{bmatrix} 0 \\ \frac{1}{2} \\ 1 \end{bmatrix}\right) \tag{17-54}$$

where $t \in \mathbb{R}$ and all $c_1, c_2, c_3 \in \mathbb{R}$.

17.3 Existence and Uniqueness of Solutions

According to the Structural Theorem 17.9 the general solution to a system of differential equations with n equations contains n arbitrary constants. If we have n ***initial conditions***, then the constants can be determined, and we then get a unique solution. This is formulated in the following ***existence and uniqueness theorem***.

|||| Theorem 17.11

A first order system of differential equations consisting of n equations in n unknown functions with constant coefficients is given by

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t), \quad t \in I\,. \tag{17-55}$$

For every $t_0 \in I$ and every number set $\mathbf{y}_0 = (y_1, y_2, \ldots, y_n)$ exactly one solution exists $\mathbf{x}(t) = (x_1(t), x_2(t) \ldots, x_n(t)\,)$ satisfying the initial conditions

$$\mathbf{x}(t_0) = \mathbf{y}_0\,, \tag{17-56}$$

that is

$$x_1(t_0) = y_1,\ x_2(t_0) = y_2,\ \ldots,\ x_n(t_0) = y_n. \tag{17-57}$$

|||| Example 17.12

In Example 17.3 we found the general solution to the system of differential equations

$$\mathbf{x}'(t) = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}\mathbf{x}(t), \quad t \in \mathbb{R}, \tag{17-58}$$

viz.

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = c_1 e^{3t}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{-2t}\begin{bmatrix} 2 \\ -3 \end{bmatrix}, \quad t \in \mathbb{R} \tag{17-59}$$

Now we wish to determine the unique solution $\mathbf{x}(t) = (x_1(t), x_2(t))$ that satisfies the initial condition $\mathbf{x}(0) = (x_1(0), x_2(0)) = (6, 6)$. This yields the system of equations

$$\begin{bmatrix} 6 \\ 6 \end{bmatrix} = c_1 e^{0}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{0}\begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & -3 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \tag{17-60}$$

By ordinary Gauss-Jordan elimination we get

$$\begin{bmatrix} 1 & 2 & 6 \\ 1 & -3 & 6 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & 6 \\ 0 & -5 & 0 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 6 \\ 0 & 1 & 0 \end{bmatrix} \tag{17-61}$$

Thus we obtain the solution $(c_1, c_2) = (6, 0)$, and the unique conditional solution is therefore

$$\mathbf{x}(t) = 6e^{3t}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad t \in \mathbb{R}\,, \tag{17-62}$$

which is equivalent to

$$x_1(t) = 6e^{3t} \quad x_2(t) = 6e^{3t}\,. \tag{17-63}$$

In this particular case the two functions are identical.

17.4 Transformation of Linear *n*'th Order Homogeneous Differential Equations to a First Order System of Differential Equations

With a bit of ingenuity it is possible to transform a homogeneous nth order differential equation with constant coefficients to a system of differential equations that can be solved using the methods in this eNote.

|||| **Method 17.13**

An nth order linear differential equation

$$x^{(n)}(t) + a_{n-1}x^{(n-1)}(t) + a_{n-2}x^{(n-2)}(t) + \cdots + a_1 x'(t) + a_0 x(t) = 0 \qquad (17\text{-}64)$$

for $t \in \mathbb{R}$, can be transformed into a first order system of differential equations and the system will look like this:

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_{n-1}'(t) \\ x_n'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} \qquad (17\text{-}65)$$

and $x_1(t) = x(t)$.

The proof of this rewriting is simple but gives a good understanding of the transformation.

|||| **Proof**

Given an nth order differential equation as in Equation (17-64). We introduce n functions in this way:

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= x_1'(t) &&= x'(t) \\ x_3(t) &= x_2'(t) &&= x''(t) \\ &\vdots && \vdots \\ x_{n-1}(t) &= x_{n-2}'(t) &&= x^{(n-2)}(t) \\ x_n(t) &= x_{n-1}'(t) &&= x^{(n-1)}(t) \end{aligned} \qquad (17\text{-}66)$$

These new expressions are substituted into the differential equation (17-64):

$$x_n'(t) + a_{n-1}x_n(t) + a_{n-2}x_{n-1}(t) + \ldots + a_1x_2(t) + a_0x_1(t) = 0 \tag{17-67}$$

Now this equation can together with equations (17-66) be written in matrix form.

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_{n-1}'(t) \\ x_n'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} \tag{17-68}$$

The method is thus proved.

■

|||| Example 17.14

Given a linear differential equation of third order with constant coefficients:

$$x'''(t) - 4x''(t) - 7x'(t) + 10x(t) = 0, \quad t \in \mathbb{R}. \tag{17-69}$$

We wish to determine the general solution. Therefore the following functions are introduced

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= x_1'(t) = x'(t) \\ x_3(t) &= x_2'(t) = x''(t) \end{aligned} \tag{17-70}$$

In this way we can rewrite the differential equation as

$$x_3'(t) - 4x_3(t) - 7x_2(t) + 10x_1(t) = 0 \tag{17-71}$$

And we can then gather the last three equations in a system of equations.

$$\begin{aligned} x_1'(t) &= x_2(t) \\ x_2'(t) &= x_3(t) \\ x_3'(t) &= -10x_1(t) + 7x_2(t) + 4x_3(t) \end{aligned} \tag{17-72}$$

This is written in matrix form in this way:

$$\mathbf{x}'(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -10 & 7 & 4 \end{bmatrix} \mathbf{x}(t) \tag{17-73}$$

The eigenvalues are determined to be $\lambda_1 = -2$, $\lambda_2 = 1$ and $\lambda_3 = 5$. The general solution to the system of differential equations according to Theorem 17.2 is given by the following functions for all the arbitrary constants $c_1, c_2, c_3 \in \mathbb{R}$:

$$\mathbf{x}(t) = c_1 \mathrm{e}^{-2t}\mathbf{v}_1 + c_2\mathrm{e}^{t}\mathbf{v}_2 + c_3\mathrm{e}^{5t}\mathbf{v}_3, \quad t \in \mathbb{R}, \tag{17-74}$$

where $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are the respective eigenvectors.

But we need only the solution of $x_1(t) = x(t)$, and we isolate this from the general solution to the system. Furthermore we introduce three new arbitrary constants $k_1, k_2, k_3 \in \mathbb{R}$, that are equal to the product of the c's and the first coordinates of the eigenvectors. The result is

$$x(t) = x_1(t) = k_1\mathrm{e}^{-2t} + k_2\mathrm{e}^{t} + k_3\mathrm{e}^{5t}, \quad t \in \mathbb{R} \tag{17-75}$$

This constitutes the general solution to the differential equation (17-69). If the first coordinate in $\mathbf{v}_1$ is 0, we put $k_1 = 0$; otherwise k_1 can be an arbitrary real number. Similarly for k_2 and k_3.

eNote 18

Linear Second-Order Differential Equations with Constant Coefficients

Following eNotes 16 and 17 about differential equations, we now present this eNote about second-order differential equations. Parts of the proofs closely follow the preceding notes and a knowledge of these notes is therefore a prerequisite. In addition, complex numbers are used.

Updated: 15.11.21 David Brander

Linear second-order differential equations with constant coefficients look like this:

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, q : I \to \mathbb{R} \tag{18-1}$$

$a_0, a_1 \in \mathbb{R}$ are constant coefficients of $x(t)$ and $x'(t)$, respectively. The right hand side $q(t)$ is a continuous real function, with the domain being an interval I (which could be all of $\mathbb{R}$). The equation is called homogeneous if $q(t) = 0$ for all $t \in I$ and otherwise inhomogeneous.

The left hand side is linear in x, i.e., the map $f : C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R})$ given by

$$f\big(\,x(t)\,\big) = x''(t) + a_1 x'(t) + a_0 x(t) \tag{18-2}$$

satisfies the linearity requirements L_1 and L_2 . The method used in this eNote for solving the inhomogeneous equation exploits this linearity.

|||| Method 18.1 Solutions and their structure

1. The general solution L_{hom} for a homogeneous linear second-order differential equation
$$x''(t) + a_1 x'(t) + a_0 x(t) = 0, \quad t \in I \tag{18-3}$$
where $a_0, a_1 \in \mathbb{R}$, can be determined using Theorem 18.2.

2. The general solution set L_{inhom} for an inhomogeneous linear second-order differential equation
$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I\,, q : I \to \mathbb{R}, \tag{18-4}$$
where $a_0, a_1 \in \mathbb{R}$, can, using Theorem 12.14, be split into two:

   (a) First the general solution L_{hom} to the *corresponding homogeneous equation* is determined. This is produced by setting $q(t) = 0$ in (18-4).

   (b) Then a particular solution $x_0(t)$ to (18-4) is determined e.g. by guessing. Concerning this see section 18.2.

   The general solution then has the following structure
$$L_{inhom} = x_0(t) + L_{hom}\,. \tag{18-5}$$

18.1 The Homogeneous Equation

We now consider the linear homogeneous second-order differential equation
$$x''(t) + a_1 x'(t) + a_0 x(t) = 0, \quad t \in \mathbb{R}\,, \tag{18-6}$$
where a_0 and a_1 are real constants. We wish to determine the general solution. This can be accomplished using exact formulas that depend on the appearance of the equation.

|||| Theorem 18.2 Solution to the Homogeneous Equation

The homogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = 0, \quad t \in \mathbb{R}, \tag{18-7}$$

has the so-called ***characteristic equation***

$$\lambda^2 + a_1 \lambda + a_0 = 0. \tag{18-8}$$

The type of roots to this equation determines how the general solution L_{hom} to the homogeneous differential equation will appear.

- **Two different real roots** λ_1 and λ_2 yield the solution

$$x(t) = c_1 \mathrm{e}^{\lambda_1 t} + c_2 \mathrm{e}^{\lambda_2 t}, \quad t \in \mathbb{R}. \tag{18-9}$$

- **Two complex roots** $\lambda = \alpha \pm \beta i$, with $\mathrm{Im}(\lambda) = \pm\beta \neq 0$, yield the real solution

$$x(t) = c_1 \mathrm{e}^{\alpha t} \cos(\beta t) + c_2 \mathrm{e}^{\alpha t} \sin(\beta t), \quad t \in \mathbb{R}. \tag{18-10}$$

- **The double root** λ yields the solution

$$x(t) = c_1 \mathrm{e}^{\lambda t} + c_2 t \mathrm{e}^{\lambda t}, \quad t \in \mathbb{R}. \tag{18-11}$$

In all three cases the respective functions for all $c_1, c_2 \in \mathbb{R}$ constitute the general solution L_{hom}.

In Section 17.4 you find the theory for rewriting this type of differential equation as a system of first-order differential equations.This method works here. The system will then look like this:

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{18-12}$$

where $x_1(t) = x(t)$ and $x_2(t) = x_1'(t) = x'(t)$. The problem can now be solved using the theory and methods outlined in that section.

|||| Proof

The homogeneous second-order linear differential equation (18-7) is rewritten as a system of first-order differential equations:

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{18-13}$$

where $x_1(t) = x(t)$ is the wanted solution that constitutes the general solution. The proof begins with the theorems and methods in Section 17.1. For the proof we need the eigenvalues of the system matrix $\mathbf{A}$:

$$\det(\mathbf{A} - \lambda \mathbf{E}) = \begin{vmatrix} -\lambda & 1 \\ -a_0 & -a_1 - \lambda \end{vmatrix} = \lambda^2 + a_1\lambda + a_0 = 0, \tag{18-14}$$

which is also the characteristic equation for the differential equation. The type of roots of this equation determines the solution $x(t) = x_1(t)$, which gives the following three parts of the proof:

First part
The characteristic equation has two different real roots: λ_1 and λ_2. By using Method 17.4 we obtain two linearly independent solutions $\mathbf{u}_1(t) = \mathbf{v}_1 \mathrm{e}^{\lambda_1 t}$ and $\mathbf{u}_2(t) = \mathbf{v}_2 \mathrm{e}^{\lambda_2 t}$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors corresponding to the two eigenvalues , respectively. The general solution is then spanned by:

$$\mathbf{x}(t) = k_1\mathbf{u}_1(t) + k_2\mathbf{u}_2(t) = k_1 \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + k_2 \mathrm{e}^{\lambda_2 t}\mathbf{v}_2, \tag{18-15}$$

for all $k_1, k_2 \in \mathbb{R}$. The first coordinate $x_1(t) = x(t)$ is the solution wanted:

$$x_1(t) = x(t) = c_1 \mathrm{e}^{\lambda_1 t} + c_2 \mathrm{e}^{\lambda_2 t}, \tag{18-16}$$

which for all the arbitrary constants $c_1, c_2 \in \mathbb{R}$ constitutes the general solution. c_1 and c_2 are two new arbitrary constants and they are the products of the k-constants and the first coordinates of the eigenvectors: $c_1 = k_1 v_{1_1}$ and $c_2 = k_2 v_{2_1}$.

Second part
The characteristic equation has the complex pair of roots $\lambda = \alpha + \beta i$ and $\bar{\lambda} = \alpha - \beta i$. It is possible to find the general solution using Method 17.5.

$$\begin{aligned} \mathbf{x}(t) &= k_1\mathbf{u}_1(t) + k_2\mathbf{u}_2(t) \\ &= k_1 \mathrm{e}^{\alpha t}\left(\cos(\beta t)\mathrm{Re}(\mathbf{v}) - \sin(\beta t)\mathrm{Im}(\mathbf{v})\right) + k_2 \mathrm{e}^{\alpha t}\left(\sin(\beta t)\mathrm{Re}(\mathbf{v}) + \cos(\beta t)\mathrm{Im}(\mathbf{v})\right) \\ &= \mathrm{e}^{\alpha t}\cos(\beta t)\cdot\left(k_1\mathrm{Re}(\mathbf{v}) + k_2\mathrm{Im}(\mathbf{v})\right) + \mathrm{e}^{\alpha t}\sin(\beta t)\cdot\left(-k_1\mathrm{Im}(\mathbf{v}) + k_2\mathrm{Re}(\mathbf{v})\right). \end{aligned} \tag{18-17}$$

$\mathbf{v}$ is an eigenvector corresponding to λ and k_1 and k_2 are arbitrary constants. The first coordinate $x_1(t) = x(t)$ is the wanted solution, and is according to the above given by

$$x_1(t) = x(t) = c_1 \mathrm{e}^{\alpha t}\cos(\beta t) + c_2 \mathrm{e}^{\alpha t}\sin(\beta t). \tag{18-18}$$

For all $c_1, c_2 \in \mathbb{R}$, $x(t)$ constitutes the general solution. c_1 and c_2 are two new arbitrary constants given by $c_1 = k_1\mathrm{Re}(v_1) + k_2\mathrm{Im}(v_1)$ and $c_2 = -k_1\mathrm{Im}(v_1) + k_2\mathrm{Re}(v_1)$. v_1 is the first coordinate of $\mathbf{v}$.

Third part
The characteristic equation has the double root λ. Because of the appearance of the system matrix (the matrix is equivalent to an upper triangular matrix) it is possible to see that the geometric multiplicity of the corresponding eigenvector space is 1, and it is then possible to use Method 17.7 to find the general solution.

$$\mathbf{x}(t) = k_1\mathbf{u}_1(t) + k_2\mathbf{u}_2(t) = k_1\mathrm{e}^{\lambda t}\mathbf{v} + k_2(t\mathrm{e}^{\lambda t}\mathbf{v} + \mathrm{e}^{\lambda t}\mathbf{b}) = \mathrm{e}^{\lambda t}(k_1\mathbf{v} + k_2\mathbf{b}) + k_2t\mathrm{e}^{\lambda t}\mathbf{v}, \qquad (18\text{-}19)$$

where $\mathbf{v}$ is an eigenvector corresponding to λ, $\mathbf{b}$ is the solution to the system of equations $(\mathbf{A} - \lambda\mathbf{E})\mathbf{b} = \mathbf{v}$, and k_1, k_2 are two arbitrary constants. Taking the first coordinate we get

$$x(t) = c_1\mathrm{e}^{\lambda t} + c_2t\mathrm{e}^{\lambda t}, \qquad (18\text{-}20)$$

which for all $c_1, c_2 \in \mathbb{R}$ constitutes the general solution. c_1d and c_2 are two new arbitrary constants, given by $c_1 = k_1v_1 + k_2b_1$ and $c_2 = k_2v_1$, in which v_1 is the first coordinate in $\mathbf{v}$, as b_1 is the first coordinate in $\mathbf{b}$.

All the three different cases of roots of the characteristic equation have now been treated thus proving the theorem.

Notice that it is also possible to arrive at the characteristic equation by guessing a solution to the differential equation of the form $x(t) = \mathrm{e}^{\lambda t}$. One then gets:

$$x''(t) + a_1x'(t) + a_0x(t) = 0 \quad \Rightarrow \quad \lambda^2\mathrm{e}^{\lambda t} + a_1\lambda\mathrm{e}^{\lambda t} + a_0\mathrm{e}^{\lambda t} = 0 \qquad (18\text{-}21)$$

Dividing this equation by $\mathrm{e}^{\lambda t}$, which is non-zero for all values of t, yields the characteristic equation.

■

|||| Example 18.3 Solution to the Homogeneous Equation

Given the homogeneous differential equation

$$x''(t) + x'(t) - 20x(t) = 0, \quad t \in \mathbb{R}, \qquad (18\text{-}22)$$

which has the characteristic equation

$$\lambda^2 + \lambda - 20 = 0. \qquad (18\text{-}23)$$

We wish to determine the general solution L_{hom} to this homogeneous equation.

The characteristic equation has the roots $\lambda_1 = -5$ and $\lambda = 4$, since $-5 \cdot 4 = -20$ and $-(-5 + 4) = 1$ are the coefficients of the characteristic equation. Therefore the general solution to the homogeneous equation is

$$L_{hom} = \left\{ c_1 e^{-5t} + c_2 e^{4t}, t \in \mathbb{R} \;\middle|\; c_1, c_2 \in \mathbb{R} \right\}, \tag{18-24}$$

that has been found using 18.2.

|||| Example 18.4 Solution to the Homogeneous Equation

A homogeneous second-order differential equation with constant coefficients is given by:

$$x''(t) - 8x'(t) + 16x(t) = 0, \quad t \in \mathbb{R}. \tag{18-25}$$

We wish to determine L_{hom}, the general solution to the homogeneous equation. The characteristic equation is

$$\lambda^2 - 8\lambda + 16 = 0 \Leftrightarrow (\lambda - 4)^2 = 0 \tag{18-26}$$

Thus we have the double root $\lambda = 4$, and the general solutions set is composed of the following function for all $c_1, c_2 \in \mathbb{R}$:

$$x(t) = c_1 e^{4t} + c_2 t e^{4t}, \quad t \in \mathbb{R}. \tag{18-27}$$

The result is determined using Theorem 18.2.

As can be seen from the two preceding examples it is relatively simple to determine the solution to the homogeneous equation. In addition it is possible to determine the differential equation from the solution, that is "go backwards". This is illustrated in the following example.

|||| Example 18.5 From Solution to Equation

The solution to a differential equation is known:

$$x(t) = c_1 e^{2t} \cos(7t) + c_2 e^{2t} \sin(7t), \quad t \in \mathbb{R}, \tag{18-28}$$

which with the arbitrary constants c_1, c_2 constitute the general solution.

Since the solution only includes terms with arbitrary constants, the equation must be homogeneous. Furthermore it is seen that the solution structure is similar to the solution structure in equation (18-10) in Theorem 18.2. This means that the characteristic equation of the second-order differential equation has two complex roots: $\lambda = 2 \pm 7i$. The characteristic equation given these roots reads:

$$\begin{gathered}(\lambda - 2 + 7i)(\lambda - 2 - 7i) = (\lambda - 2)^2 - (7i)^2 = \\ \lambda^2 - 4\lambda + 4 + 49 = \lambda^2 - 4\lambda + 53 = 0\end{gathered} \tag{18-29}$$

Directly from coefficients of the characteristic equation we can write the differential equation as:

$$x''(t) - 4x'(t) + 53x(t) = 0, \quad t \in \mathbb{R}. \tag{18-30}$$

This can also be seen from Theorem 18.2.

18.2 The Inhomogeneous Equation

In this section we wish to determine a particular solution $x_0(t)$ to the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, q : I \to \mathbb{R}. \tag{18-31}$$

We wish to find a particular solution, because it is part of the general solution L_{inhom} together with the general solution L_{hom} to the corresponding homogeneous equation cf. Method 18.1.

In this eNote we do not use a specific solution formula. Instead we use different methods depending on the form of $q(t)$. In general one might guess that a particular solution $x_0(t)$ has a form that somewhat resembles $q(t)$, as will appear from the following methods. Notice that these methods cover some frequently occurring forms of $q(t)$, but certainly not all.

Furthermore the concept of *superposition* will be treated. Superposition is a basic quality of linear equations and linear differential equations. The point is to split the equation into more equations in which the left hand sides stay the same while the sum of the right hand sides is equal to the right hand side of the original equation. If the original equation has the right hand side $q(t) = \sin(2t) + 2t^2$, it may be a good idea to split the equation into two, where the right hand sides become $q_1(t) = \sin(2t)$ and $q_2(t) = 2t^2$ respectively. It is easier to determine particular solutions to the two equations. A particular solution to the original equation will then be the sum of the two particular solutions.

Finally we will introduce *the complex guess method*. The complex guess method can be used if the right hand side $q(t)$ of the equation is the real part of a simple complex expression, e.g. $q(t) = \mathrm{e}^t \sin(3t)$ that is the real part of $-i\mathrm{e}^{(1+3i)t}$. Solving an equation with a simple right hand side is easier, and therefore the corresponding complex equation is solved instead. The solutions to the real equation and to the corresponding complex equation are closely related.

18.2.1 General Solution Methods

|||| Method 18.6 Polynomial

Given the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, \tag{18-32}$$

where q is an n-th degree polynomial. If $a_0 \neq 0$ a polynomial of degree n that is a particular solution to the equation exists. In general a polynomial of degree $n+2$ at the most, that is a particular solution to the equation, exists. A particular solutions of the form mentioned is found by insertion of polynomials of a suitable degree with unknown coefficients in the left-hand side of the equation and tune this to the right-hand side q, cf the identity theorem for polynomials, eNote 2, Theorem 2.15.

|||| Example 18.7 Polynomial

Given the inhomogeneous second-order differential equation with constant coefficients

$$x''(t) - 3x'(t) + x(t) = 2t^2 - 16t + 25, \quad t \in \mathbb{R}. \tag{18-33}$$

We wish to determine a particular solution $x_0(t)$ to the inhomogeneous equation. Since the right hand side is a second degree polynomial we insert an unknown polynomial of second degree in the left-hand side of the equation and equate this with the right-hand side:

$$x_0(t) = b_2 t^2 + b_1 t + b_0, \quad t \in \mathbb{R}. \tag{18-34}$$

The coefficients are determined by substituting the expression into the differential equation together with $x_0'(t) = 2b_2 t + b_1$ og $x_0''(t) = 2b_2$.

$$\begin{aligned} 2b_2 - 3(2b_2 t + b_1) + b_2 t^2 + b_1 t + b_0 &= 2t^2 - 16t + 25 \Leftrightarrow \\ (b_2 - 2)t^2 + (-6b_2 + b_1 + 16)t + (2b_2 - 3b_1 + b_0 - 25) &= 0 \Leftrightarrow \\ b_2 - 2 = 0 \text{ og } -6b_2 + b_1 + 16 = 0 \text{ og } 2b_2 - 3b_1 + b_0 - 25 &= 0 \end{aligned} \tag{18-35}$$

From the first equation it is evident that $b_2 = 2$, and by substituting this in the second equation we get $b_1 = -4$. Finally the last equation yields $b_0 = 9$. Therefore a particular solution to Equation (18-33) is given by

$$x_0(t) = 2t^2 - 4t + 9, \quad t \in \mathbb{R}. \tag{18-36}$$

|||| Exercise 18.8 Polynomium

Given the following differential equation where the right-hand side is a first degree polynomial:

$$x''(t) = t + 1\,, \quad t \in \mathbb{R}. \tag{18-37}$$

Show that you have to go to the third degree in order to find a polynomial that is a particular solution to the equation.

|||| Method 18.9 Trigonometric

A particular solution $x_0(t)$ to the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, \tag{18-38}$$

where $q(t) = a\cos(\omega t) + b\sin(\omega t)$, is of the same form:

$$x_0(t) = A\sin(\omega t) + B\cos(\omega t), \quad t \in I, \tag{18-39}$$

where A and B are determined by substitution of the expression for $x_0(t)$ as a solution into the inhomogeneous equation.

It is also possible to determine a particular solution to a differential equation like the one in Method 18.9 using *the complex guess method*, cf. e.g. section 18.2.3.

Example 18.10 Trigonometric

Given the differential equation

$$x''(t) + x'(t) - x(t) = -20\sin(3t) + 6\cos(3t), \quad t \in \mathbb{R}. \tag{18-40}$$

We wish to determine a particular solution $x_0(t)$. By the use of Method 18.9 a particular solution is

$$x_0(t) = A\sin(\omega t) + B\cos(\omega t) = A\sin(3t) + B\cos(3t). \tag{18-41}$$

In addition we have

$$\begin{aligned} x_0'(t) &= 3A\cos(3t) - 3B\sin(3t) \\ x_0''(t) &= -9A\sin(3t) - 9B\cos(3t) \end{aligned} \tag{18-42}$$

This is substituted into the equation

$$\begin{aligned} (-9A\sin(3t) - 9B\cos(3t)) + (3A\cos(3t) - 3B\sin(3t)) - (A\sin(3t) + B\cos(3t)) \\ = -20\sin(3t) + 6\cos(3t) \Leftrightarrow \\ (-9A - 3B - A + 20)\sin(3t) + (-9B + 3A - B - 6)\cos(3t) = 0 \Leftrightarrow \\ -9A - 3B - A + 20 = 0 \ \text{ og } \ -9B + 3A - B - 6 = 0 \end{aligned} \tag{18-43}$$

This is two equations in two unknowns. Substituting $A = -\frac{3}{10}B + 2$ from the first equation in the second yields

$$-9B + 3\left(-\frac{3}{10}B + 2\right) - B - 6 = 0 \Leftrightarrow -10B - \frac{9}{10}B = 0 \Leftrightarrow B = 0 \tag{18-44}$$

From this we get that $A = 2$, and a particular solution to the differential equation is then

$$x_0(t) = 2\sin(3t), \quad t \in \mathbb{R}. \tag{18-45}$$

Note that the number $\omega = 3$ is the same in the arguments of both cosine and sine in Example 18.10, and this is the only case that Method 18.9 facilitates. If two different numbers are present Method 18.9 does not apply, e.g. $q(t) = 3\sin(t) + \cos(10t)$. But either *superposition* or *the complex guess method* can be applied, and they will be described in section 18.2.2 and section 18.2.3, respectively.

|||| Method 18.11 Exponential Function

A particular solution $x_0(t)$ to the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, \tag{18-46}$$

where $q(t) = \beta e^{\alpha t}$ og $\alpha, \beta \in \mathbb{R}$, is also an exponential function:

$$x_0(t) = \gamma e^{\alpha t}, \quad t \in I, \tag{18-47}$$

where γ is determined by substituting the expression for $x_0(t)$ as a solution into the inhomogeneous equation. We emphasize that α must not be a root of the characteristic equation for the differential equation.

As commented by the end of Method 18.11 the exponent α must not be a root of the characteristic equation. If this is the case the guess will be a solution to the corresponding homogeneous equation c.f. Theorem 18.2. This is a "problem" for all orders of differential equations.

|||| Example 18.12 Exponential Function

Given the differential equation

$$x''(t) + 11x'(t) + 5x(t) = -20e^{-t}, \quad t \in \mathbb{R}. \tag{18-48}$$

We wish to determine a particular solution $x_0(t)$. According to Method 18.11 a particular solution is given by $x_0(t) = \gamma e^{\alpha t} = \gamma e^{-t}$. We do not yet know whether $\alpha = -1$ is a root in the characteristic equation, but if it is possible to find γ, it is not a root. We have $x_0'(t) = -\gamma e^{-t}$ and $x_0''(t) = \gamma e^{-t}$, and this is substituted into the differential equation:

$$\gamma e^{-t} + 11(-\gamma e^{-t}) + 5\gamma e^{-t} = -20e^{-t} \Leftrightarrow -5\gamma = -20 \Leftrightarrow \gamma = 4 \tag{18-49}$$

Thus we have succeeded in finding γ, and therefore we have a particular solution to the differential equation:

$$x_0(t) = 4e^{-t}, \quad t \in \mathbb{R}. \tag{18-50}$$

|||| Method 18.13 Exponential Function Belonging to L_{hom}

A particular solution $x_0(t)$ to the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, \tag{18-51}$$

where $q(t) = \beta e^{\lambda t}$, $\beta \in \mathbb{R}$ and λ is a root in the characteristic equation of the differential equation, has the following form:

$$x_0(t) = \gamma t e^{\lambda t}, \quad t \in I, \tag{18-52}$$

where γ is determined by substitution of the expression for $x_0(t)$ as a solution into the inhomogeneous equation.

|||| Example 18.14 Exponential Function Belonging to L_{hom}

Given the differential equation

$$x''(t) - 7x'(t) + 10x(t) = -3e^{2t}, \quad t \in \mathbb{R}. \tag{18-53}$$

We wish to determine a particular solution. First we try to use Method 18.11, and guess a solution of the form $x_0(t) = \gamma e^{\alpha t} = \gamma e^{2t}$. One then has $x_0'(t) = 2\gamma e^{2t}$ and $x_0''(t) = 4\gamma e^{2t}$, which by substitution into the equation gives

$$4\gamma e^{2t} - 7 \cdot 2\gamma e^{2t} + 10\gamma e^{2t} = -3e^{2t} \Leftrightarrow 0 = -3 \tag{18-54}$$

It is seen that γ does not appear in the last equation, and that the equation otherwise is false. Therefore $\alpha = \lambda$ must be a root in the characteristic equation. The characteristic equation looks like this:

$$\lambda^2 - 7\lambda + 10 = 0 \tag{18-55}$$

This second degree equation has the roots 2 and 5, since $2 \cdot 5 = 10$ and $-(2+5) = -7$. It is true that $\alpha = 2$ is a root.

Consequently we use Method 18.13, and we guess a solution of the form $x_0(t) = \gamma t e^{\lambda t} = \gamma t e^{2t}$. We then have

$$\begin{aligned} x_0'(t) &= \gamma e^{2t} + 2\gamma t e^{2t} \\ x_0''(t) &= 2\gamma e^{2t} + 2\gamma e^{2t} + 4\gamma t e^{2t} = 4\gamma e^{2t} + 4\gamma t e^{2t} \end{aligned} \tag{18-56}$$

This is substituted into the equation in order to determine γ.

$$\begin{aligned} 4\gamma e^{2t} + 4\gamma t e^{2t} - 7(\gamma e^{2t} + 2\gamma t e^{2t}) + 10\gamma t e^{2t} = -3e^{2t} \;\Leftrightarrow \\ (4\gamma - 14\gamma + 10\gamma)t + (4\gamma - 7\gamma + 3) = 0 \;\Leftrightarrow \\ \gamma = 1 \end{aligned} \tag{18-57}$$

We have now succeeded in finding γ, and therefore a particular solution to the equation is

$$x_0(t) = t e^{2t}, \quad t \in \mathbb{R}. \tag{18-58}$$

18.2.2 Superposition

Within all types of linear equations the concept of ***superposition*** exists. We present the concept here for second-order linear differential equations with constant coefficients. Superposition is here used in order to determine a particular solution to the inhomogeneous equation, when the right hand side ($q(t)$) is a combination (addition) of more types of functions, e.g. a sine function added to a polynomial.

|||| Theorem 18.15 Superposition

Let $q_1, q_2, \ldots, q_n$ be continuous functions on an interval I. If $x_{0_i}(t)$ is a particular solution to the inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q_i(t) \tag{18-59}$$

for every $i = 1, \ldots, n$, then

$$x_0(t) = x_{0_1}(t) + x_{0_2}(t) + \ldots + x_{0_n}(t) \tag{18-60}$$

is a particular solution to

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t) = q_1(t) + q_2(t) + \ldots + q_n(t), \tag{18-61}$$

|||| Proof

Superposition is a consequence of the differential equation being linear. We will here give a general proof for all types of linear differential equations.

The left hand side of a differential equation is called $f(x(t))$. We now posit n differential equations:

$$f(x_{0_1}(t)) = q_1(t), \qquad f(x_{0_2}(t)) = q_2(t), \qquad \ldots, \qquad f(x_{0_n}(t)) = q_n(t) \tag{18-62}$$

where $x_{0_1}, x_{0_2}, \ldots, x_{0_n}$ are particular solutions to the respective inhomogeneous differential equations. Define $x_0 = x_{0_1} + x_{0_2} + \ldots + x_{0_n}$ and substitute this into the left hand side:

$$\begin{aligned} f(x_0(t)) &= f(x_{0_1}(t) + x_{0_2}(t) + \ldots + x_{0_n}(t)) \\ &= f(x_{0_1}(t)) + f(x_{0_2}(t)) + \ldots + f(x_{0_n}(t)) \\ &= q_1(t) + q_2(t) + \ldots + q_n(t) \end{aligned} \tag{18-63}$$

On the right hand side we get the sum of the functions $q_1, q_2, \ldots, q_n$, which sum we call q. The Theorem is thus proven.

■

|||| Example 18.16 Superposition

Given the inhomogeneous differential equation

$$x''(t) - x'(t) - 3x(t) = 9\mathrm{e}^{4t} + 3t - 14, \quad t \in \mathbb{R}. \tag{18-64}$$

We wish to determine a particular solution $x_0(t)$. It is seen that the right hand side is a combination of an exponential function ($q_1(t) = 9\mathrm{e}^{4t}$) and a polynomial ($q_2(t) = 3t - 14$). Therefore we use superposition 18.15 and the equation is split into two parts.

$$x''(t) - x'(t) - 3x(t) = 9\mathrm{e}^{4t} = q_1(t) \tag{18-65}$$

$$x''(t) - x'(t) - 3x(t) = 3t - 14 = q_2(t) \tag{18-66}$$

First we treat (18-65), for which we use Method 18.11. A particular solution then has the form $x_{0_1}(t) = \gamma \mathrm{e}^{\alpha t} = \gamma \mathrm{e}^{4t}$. We have $x'_{0_1}(t) = 4\gamma \mathrm{e}^{4t}$ and $x''_{0_1}(t) = 16\gamma \mathrm{e}^{4t}$. This is inserted into the equation.

$$16\gamma \mathrm{e}^{4t} - 4\gamma \mathrm{e}^{4t} - 3\gamma \mathrm{e}^{4t} = 9\mathrm{e}^{4t} \Leftrightarrow \gamma = 1 \tag{18-67}$$

Therefore $x_{0_1}(t) = \mathrm{e}^{4t}$.

Now we treat Equation (18-66), where a particular solution is a polynomial of at the most first degree, cf. Method 18.6, thus $x_{0_2}(t) = b_1 t + b_0$. Hence $x'_{0_2}(t) = b_1$ and $x''_{0_2}(t) = 0$. This is substituted into the differential equation.

$$0 - b_1 - 3(b_1 t + b_0) = 3t - 14 \Leftrightarrow (-3b_1 - 3)t + (-b_1 - 3b_0 + 14) = 0 \tag{18-68}$$

Thus we have two equations in two unknowns, and we find that $b_1 = -1$, and therefore that $b_0 = 5$. Thus a particular solution is $x_{0_2}(t) = -t + 5$. The general solution to (18-64) is then found as the sum of the already found particular solutions to the two split equations:

$$x_0(t) = x_{0_1}(t) + x_{0_2}(t) = \mathrm{e}^{4t} - t + 5, \quad t \in \mathbb{R}. \tag{18-69}$$

18.2.3 The Complex Guess Method

The complex guess method is used when it is easy to rewrite the right hand side of the differential equation as a complex expression, such that the given real right hand side is the real part of the complex.

If e.g. the original right hand side is $2\mathrm{e}^{2t}\cos(3t)$, adding $i(-2\mathrm{e}^{2t}\sin(3t))$, we get

$$2\mathrm{e}^{2t}(\cos(3t) - i\sin(3t)) = 2\mathrm{e}^{(2-3i)t}\,. \tag{18-70}$$

Here it is evident that $\mathrm{Re}(2\mathrm{e}^{(2-3i)t}) = 2\mathrm{e}^{2t}\cos(3t)$. One then finds a complex particular solution with complex right hand side. The wanted real particular solution to the original equation is then the real part of the found complex solution.

Note that this method can be used because the equation is linear. It is exactly the linearity that secures that the real part of the complex solution found is the wanted real solution. This is shown by interpreting the left hand side of the equation as linear map $f(z(t))$ in the set of complex functions of one real variable and using the following general theorem:

|||| Theorem 18.17

Given a linear map $f : (C^\infty(\mathbb{R}),\mathbb{C}) \to (C^\infty(\mathbb{R}),\mathbb{C})$ and the equation

$$f(z(t))) = s(t)\,. \tag{18-71}$$

If we state $z(t)$ and $s(t)$ in rectangular form as $z(t) = x(t) + i\cdot y(t)$ and $s(t) = q(t) + i\cdot r(t)$, then (18-71) is true and if and only if

$$f(x(t)) = q(t) \quad \text{and} \quad f(y(t)) = r(t)\,. \tag{18-72}$$

|||| Proof

Given the function $z(t)$ and letting the linear map f and the functions $z(t)$ and $s(t)$ be given as in Theorem 18.17. As a consequence of the qualities of a linear map, cf. Definition **??**, the following applies:

$$\begin{aligned} f(z(t)) = s(t) &\Leftrightarrow \\ f(x(t) + i \cdot y(t)) = q(t) + i \cdot r(t) &\Leftrightarrow \\ f(x(t)) + i \cdot f(y(t)) = q(t) + i \cdot r(t) &\Leftrightarrow \\ f(x(t)) = q(t) \text{ and } f(y(t)) = r(t)\,. \end{aligned} \tag{18-73}$$

Thus the theorem is proven.

■

|||| Method 18.18 The Complex Guess Method

A particular solution $x_0(t)$ to the real inhomogeneous differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in \mathbb{R}, \tag{18-74}$$

where a_0 og a_1 are real coefficients and

$$q(t) = \mathrm{Re}\Big((a+bi)\mathrm{e}^{(\alpha+\omega i)t}\Big) = a\mathrm{e}^{\alpha t}\cos(\omega t) - b\mathrm{e}^{\alpha t}\sin(\omega t), \tag{18-75}$$

is initially determined by the corresponding complex particular solution to the following complex equation

$$z''(t) + a_1 z'(t) + a_0 z(t) = (a+bi)\mathrm{e}^{(\alpha+\omega i)t}, \quad t \in \mathbb{R}, \tag{18-76}$$

The complex particular solution has the form $z_0(t) = (c+di)\mathrm{e}^{(\alpha+\omega i)t}$, where c and d are determined by substitution of $z_0(t)$ into Equation (18-76).

Then a particular solution to equation (18-74) is given by

$$x_0(t) = \mathrm{Re}(z_0(t))\,. \tag{18-77}$$

A decisive reason for using the complex guess method is that it is so easy to determine the derivative of the exponential function, even when it is complex.

|||| Example 18.19 The Complex Guess Method

Given a second-order inhomogeneous differential equation:

$$x''(t) - 2x'(t) - 2x(t) = 19\mathrm{e}^{4t}\cos(t) - 35\mathrm{e}^{4t}\sin(t), \quad t \in \mathbb{R}. \tag{18-78}$$

We wish to determine a particular solution. It is evident that we can use *the complex guess method* in Method 18.18. Initially the following is true for the right hand side:

$$q(t) = 19\mathrm{e}^{4t}\cos(t) - 35\mathrm{e}^{4t}\sin(t) = \mathrm{Re}\Big((19+35i)\mathrm{e}^{(4+i)t}\Big)\,. \tag{18-79}$$

We shall now instead of the original problem find a complex particular solution to the differential equation

$$z''(t) - 2z'(t) - 2z(t) = (19+35i)\mathrm{e}^{(4+i)t}, \quad t \in \mathbb{R}. \tag{18-80}$$

by guessing that $z_0(t) = (c+di)\mathrm{e}^{(4+i)t}$ is a solution. We also have

$$\begin{aligned} z_0'(t) &= (c+di)(4+i)\mathrm{e}^{(4+i)t} = (4c-d+(c+4d)i)\,\mathrm{e}^{(4+i)t} \quad \text{and} \\ z_0''(t) &= (4c-d+(c+4d)i)(4+i)\mathrm{e}^{(4+i)t} = (15c-8d+(8c+15d)i)\mathrm{e}^{(4+i)t} \end{aligned} \tag{18-81}$$

These expressions are substituted into the complex equation in order to determine c and d:

$$\begin{aligned} (15c-8d+(8c+15d)i)\mathrm{e}^{(4+i)t} - 2(4c-d+(c+4d)i)\mathrm{e}^{(4+i)t} - 2(c+di)\mathrm{e}^{(4+i)t} \\ = (19+35i)\mathrm{e}^{(4+i)t} \Leftrightarrow \\ 15c-8d+(8c+15d)i - 2(4c-d+(c+4d)i) - 2(c+di) = 19+35i \Leftrightarrow \\ 5c-6d+(6c+5d)i = 19+35i \Leftrightarrow \\ 5c-6d = 19 \;\text{ og }\; 6c+5d = 35 \end{aligned} \tag{18-82}$$

These are two equations in two unknowns. The augmented matrix of the system of equations is written:

$$\left[\begin{array}{cc|c} 5 & -6 & 19 \\ 6 & 5 & 35 \end{array}\right] \to \left[\begin{array}{cc|c} 1 & -\frac{6}{5} & \frac{19}{5} \\ 0 & \frac{61}{5} & \frac{61}{5} \end{array}\right] \to \left[\begin{array}{cc|c} 1 & 0 & 5 \\ 0 & 1 & 1 \end{array}\right]. \tag{18-83}$$

thus we have that $c=5$ and $d=1$, which yields $z_0(t) = (5+i)\mathrm{e}^{(4+i)t}$. Therefore a particular solution to the equation (18-78) is

$$x_0(t) = \mathrm{Re}(z_0(t)) = \mathrm{Re}\Big((5+i)\mathrm{e}^{(4+i)t}\Big) = 5\mathrm{e}^{4t}\cos(t) - \mathrm{e}^{4t}\sin(t), \quad t \in \mathbb{R}. \tag{18-84}$$

18.3 Existence and Uniqueness

Here we formulate a theorem about ***existence and uniqueness*** for differential equations of the second order with constant coefficients. We need two *initial value conditions*: The value of the function and its first derivative at the chosen initial point.

|||| Theorem 18.20 Existence and Uniqueness

For every 3-tuple (t_0, x_0, v_0) (*double initial value condition*), there exists exactly one solution $x(t)$ to the differential equation

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t), \quad t \in I, q : I \to \mathbb{R}, \tag{18-85}$$

such that

$$x(t_0) = x_0 \quad \text{and} \quad x'(t_0) = v_0, \tag{18-86}$$

where $t_0 \in I$, $x_0 \in \mathbb{R}$ and $v_0 \in \mathbb{R}$.

|||| Example 18.21 Exsistence and Uniqueness

Given the differential equation

$$x''(t) - 5x'(t) - 36x(t) = 0, \quad t \in \mathbb{R}. \tag{18-87}$$

It is seen that the equation is homogeneous. It has the characteristic equation

$$\lambda^2 - 5\lambda - 36 = 0. \tag{18-88}$$

We wish to determine a function $x(t)$ that is a solution to the differential equation and has the initial value condition $(t_0, x_0, v_0) = (0, 5, 6)$. The characteristic equation has the roots $\lambda_1 = -4$ and $\lambda_2 = 9$, since $-4 \cdot 9 = -36$ and $-(9 + (-4)) = 5$ are the coefficients of the equation. Therefore the general solution for the homogeneous equation (using Theorem 18.2) is spanned by the following functions for all $c_1, c_2 \in \mathbb{R}$:

$$x(t) = c_1 e^{-4t} + c_2 e^{9t}, \quad t \in \mathbb{R}. \tag{18-89}$$

One then has

$$x'(t) = -4c_1 e^{-4t} + 9c_2 e^{9t} \tag{18-90}$$

if the initial value condition ($x(0) = 5$ and $x'(0) = 6$) is substituted into the two equations one can solve for (c_1, c_2).

$$\begin{aligned} 5 &= c_1 + c_2 \\ 6 &= -4c_1 + 9c_2 \end{aligned} \tag{18-91}$$

since $\mathrm{e}^0 = 1$. If $c_2 = 5 - c_1$ is substituted into the second equation one gets

$$6 = -4c_1 + 9(5 - c_1) = -13c_1 + 45 \Leftrightarrow c_1 = \frac{6-45}{-13} = 3 \tag{18-92}$$

Therefore $c_2 = 5 - 3 = 2$ and the conditional solution is

$$x(t) = 3\mathrm{e}^{-4t} + 2\mathrm{e}^{9t}, \quad t \in \mathbb{R} \tag{18-93}$$

Note that one can determine a unique and conditional solution to a homogeneous differential equation, as in this case. The right hand side needs not be different from zero. The general solution for the equation is $L_{inhom} = L_{hom}$, since $x_0(t) = 0$.

Below is an example going through the whole solution procedure for an inhomogeneous equation with a double initial value condition. After that an example is presented where the purpose is to find the differential equation given the general solution. It is analogous to example 18.5, but now we have a right hand side different from zero.

|||| Example 18.22 Accumulated Example

Given the differential equation

$$x''(t) + 6x'(t) + 5x(t) = 20t^2 + 48t + 13, \quad t \in \mathbb{R}. \tag{18-94}$$

We determine the general solution L_{inhom}. Then the conditional solution $x(t)$ that satisfies the initial value condition $(t_0, x_0, v_0) = (0, 5, -8)$, will be determined.

First we solve the corresponding homogeneous equation, and the characteristic equation looks like this:

$$\lambda^2 + 6\lambda + 5 = 0 \tag{18-95}$$

This has the roots $\lambda_1 = -5$ and $\lambda_2 = -1$, since $(\lambda + 5)(\lambda + 1) = \lambda^2 + 6\lambda + 5$. Because these roots are real and different, cf. Theorem 18.2, the general homogeneous solution set is given by

$$L_{hom} = \left\{ c_1\mathrm{e}^{-5t} + c_2\mathrm{e}^{-t} \;, t \in \mathbb{R} \;\middle|\; c_1, c_2 \in \mathbb{R} \right\}. \tag{18-96}$$

Now we determine a particular solution to the inhomogeneous equation. Since the right hand side is a second degree polynomial we guess that $x_0(t) = b_2t^2 + b_1t + b_0$, using Method 18.6. We then have that $x_0'(t) = 2b_2t + b_1$ and $x_0''(t) = 2b_2$. This is substituted into the differential equation.

$$\begin{gathered} 2b_2 + 6(2b_2t + b_1) + 5(b_2t^2 + b_1t + b_0) = 20t^2 + 48t + 13 \Leftrightarrow \\ (5b_2 - 20)t^2 + (12b_2 + 5b_1 - 48)t + (2b_2 + 6b_1 + 5b_0 - 13) = 0 \Leftrightarrow \\ 5b_2 - 20 = 0 \;\text{ og }\; 12b_2 + 5b_1 - 48 = 0 \;\text{ og }\; 2b_2 + 6b_1 + 5b_0 - 13 = 0. \end{gathered} \tag{18-97}$$

The first equation easily yields $b_2 = 4$. If this is substituted into the second equation we get $b_1 = 0$. Finally in the third equation we get $b_0 = 1$. A particular solution to the inhomogeneous equation is therefore

$$x_0(t) = 4t^2 + 1, \quad t \in \mathbb{R}. \tag{18-98}$$

Following the structural theorem, e.g. Method 18.1, the general solution to the inhomogeneous equation is given by

$$L_{inhom} = x_0(t) + L_{hom} = \left\{ 4t^2 + 1 + c_1 \mathrm{e}^{-5t} + c_2 \mathrm{e}^{-t}, t \in \mathbb{R} \;\middle|\; c_1, c_2 \in \mathbb{R} \right\} \tag{18-99}$$

We now determine the solution that satisfies the given initial value conditions. An arbitrary solution has the form

$$x(t) = 4t^2 + 1 + c_1 \mathrm{e}^{-5t} + c_2 \mathrm{e}^{-t}, \quad t \in \mathbb{R}. \tag{18-100}$$

We now determine the derivative

$$x'(t) = 8t - 5c_1 \mathrm{e}^{-5t} - c_2 \mathrm{e}^{-t}, \quad t \in \mathbb{R}. \tag{18-101}$$

If $x(0) = 5$ and $x'(0) = -8$ are substituted we get two equations

$$\begin{aligned} 5 &= \quad c_1 + c_2 + 1 \\ -8 &= -5c_1 - c_2 \end{aligned} \tag{18-102}$$

Substituting $c_1 = 4 - c_2$ from the first equation into the second we get

$$-8 = -5(4 - c_2) - c_2 \Leftrightarrow -8 + 20 = 4c_2 \Leftrightarrow c_2 = 3. \tag{18-103}$$

This yields $c_1 = 1$ and therefore the conditional solution is

$$x(t) = \mathrm{e}^{-5t} + 3\mathrm{e}^{-t} + 4t^2 + 1, \quad t \in \mathbb{R}. \tag{18-104}$$

|||| Example 18.23 From the Solution to the Equation

Given the general solution to a linear second-order differential equation with constant coefficients:

$$L_{inhom} = \left\{ c_1 \mathrm{e}^{-2t} + c_2 \mathrm{e}^{2t} - \tfrac{1}{2}\sin(2t), t \in \mathbb{R} \;\middle|\; c_1, c_2 \in \mathbb{R} \right\} \tag{18-105}$$

It is now the aim to find the differential equation, which in general looks like this:

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t) \tag{18-106}$$

Thus we have to determine a_1, a_0 and $q(t)$.

First we split the solution into a particular solution and the general homogeneous solution set:

$$x_0(t) = -\tfrac{1}{2}\sin(2t),\ t \in \mathbb{R} \quad \text{and} \quad L_{hom} = \left\{ c_1 \mathrm{e}^{-2t} + c_2 \mathrm{e}^{2t}\ t \in \mathbb{R} \,\middle|\, c_1, c_2 \in \mathbb{R} \right\} \tag{18-107}$$

Now we consider the general homogeneous solution. The looks of this complies with the first case of in Theorem 18.2. The characteristic equation has two real roots and they are $\lambda_1 = -2$ and $\lambda_2 = 2$. Therefore the characteristic equation is

$$(\lambda + 2)(\lambda - 2) = \lambda^2 - 4 = 0 \tag{18-108}$$

This determines the coefficients on the left hand side of the differential equation: $a_1 = 0$ and $a_0 = -4$. The differential equation so far looks like this:

$$x''(t) - 4x(t) = q(t), \quad t \in \mathbb{R}. \tag{18-109}$$

Since $x_0(t)$ is a particular solution to the inhomogeneous equation the right hand side $q(t)$ can be determined by substituting $x_0(t)$. We have that $x_0''(t) = 2\sin(2t)$.

$$\begin{aligned} x_0''(t) - 4x_0(t) &= q(t) \Leftrightarrow \\ 2\sin(2t) - 4(-\tfrac{1}{2}\sin(2t)) &= q(t) \Leftrightarrow \\ 4\sin(2t) &= q(t) \end{aligned} \tag{18-110}$$

Now all unknowns in the differential equation are determined:

$$x''(t) - 4x(t) = 4\sin(2t), \quad t \in \mathbb{R}. \tag{18-111}$$

In these eNotes we do not consider systems of second-order homogeneous linear differential equations with constant coefficients. We should mention, however, that with the presented theory and a bit of cleverness we can solve such problems. If we have a system of second-order homogeneous differential equations then we can consider each equation individually. By use of Section 17.4 such an equation be rewritten as two equations of first order. If this is done with all the equations in the system, we end up with double the number of equations, but those now of first-order equations. We can solve this new system with the theory presented in eNote 16. Systems of second-order homogeneous linear differential equations are seen in many places in mechanical physics, chemistry, electro-magnetism etc.

18.4 Summary

In this note linear second-order differential equations with constant coefficients are written as:

$$x''(t) + a_1 x'(t) + a_0 x(t) = q(t) \tag{18-112}$$

- This equation is solved by first determining the general solution to the corresponding homogeneous equation and then adding this to a particular solution to the inhomogeneous equation, see Method 18.1.
- The general solution to the corresponding homogeneous differential equation is determined by finding the roots of the *characteristic equation*:

  $$\lambda^2 + a_1\lambda + a_0 = 0. \tag{18-113}$$

  There are in principle three cases, see Theorem 18.2.
- A particular solution is determined by "guessing" a solution that has the same appearance as the right hand side $q(t)$. If e.g. $q(t)$ is a polynomial then $x_0(t)$ is also a polynomial of at the most same degree. In the note many examples are given, see Section 18.2.
- In particular we have *the complex guess method* for the determination of the particular solution $x_0(t)$. The complex guess method can be used when the right hand side has this appearance:

  $$q(t) = \mathrm{Re}\Big((a+bi)\mathrm{e}^{(\alpha+\omega i)t}\Big) = a\mathrm{e}^{\alpha t}\cos(\omega t) - b\mathrm{e}^{\alpha t}\sin(\omega t). \tag{18-114}$$

  The solution is then determined by rewriting the differential equation in the corresponding exponential form, see Method 18.18.
- Furthermore *superposition* is introduced. Superposition is a general principle that applies to all types of linear equations. The idea is that two particular solutions can be added. When they are substituted into the differential equation they will not influence each other, and hence the right hand side can also be split into two terms, each corresponding to one of the two solutions. This can be used to determine a particular solution, when the right hand side is the sum of e.g. a sine function and a polynomial. See e.g. Example 18.16.
- Furthermore an *existence and uniqueness theorem* is formulated, see Theorem 18.20. According to this theorem a unique conditional solution that must satisfy two particular *initial value conditions* to a second-order differential equation can be determined.

Index

$\mathbf{A}^{-1}$, inverse matrix, 174
E, identity matrix, 173
$\det(\mathbf{A})$, 184
0-row, 145

diagonal of matrix, 138
Gauss-Jordan elimination , 133
row operations , 133

$\mathbf{A}^{\top}$, 169
adjoint matrix, 197
algebraic multiplicity, 346
angle between two vectors **a** and **b**, 368
approximating polynomial, 102
area funktions, 91
arrow, 204

basis, 220, 261
basis vector, 220
bi-diagonal-matrix, 190

change of basis matrix, 276
characteristic equation, 338, 341, 446
characteristic matrix, 201
characteristic polynomial, 201
codomain, 284
column vector, 126
concept of continuity, 76
coordinate matrix, 237, 270
Coordinate vector, 224
coordinate vector, 221
coupled differential equations, 427
curly bracket, 74

derivative, 78
determinant, 184
diagonal in a matrix, 131, 148
diagonal matrix, 172, 190
diagonalization method, 429, 430
diagonalization of symmetric matrices, 380
diagonalized, 355
direction vector, 209
direction vectors, 219
domain, 71, 284

eigenspace, 337
eigenvalue, 321, 325
eigenvector, 320, 325
eigenvector space, 337
element in a matrix, 131
epsilon function, 75
epsilon-functions, 71
Euclidian vector space $(\mathbb{R}^n, \cdot)$, 364
expansion along the r-th row, 187

free parametre, 152

Gauss-Jordan elimination, 142
general solution, 128, 130
geometric multiplicity, 346
geometric vector, 204
global maximum point, 113
global maximum value, 113
global minimum point, 113
global minimum value, 113

hat vector, 243

hyperbolic cosine, 87
hyperbolic sine, 87

identity matrix, 173
identity theorem for polynomials, 265
image, 292
inconsistent equation, 128
indefinite, 397
infinite-dimensional vector spaces, 282
initial conditions, 439
intersection, 131, 153
inverse function, 84
inverse matrix, 174
invertible matrix, 174

kernel, 292

leading 1, 138
length of a vector **a**, 365
linear, 289
linear combination, 156, 213, 255
linear dependence, 215
linear equation, 127
linear independence, 215
linearly dependent, 257
linearly independent, 257
local maximum point, 116
local maximum value, 116
local minimum point, 116
local minimum value, 116

main diagonal, 172
mapping matrix, 300
matrix, 131, 159
matrix product, 166
mean value theorem, 101
mirroring the graph, 85
monomial basis, 265, 273

negative definite, 397
negative orthogonal matrix, 372
negative semi-definite, 397
norm of a vector **a**, 365

ordered set, 220
ordinary basis, 221, 223
orthogonal complement, 377
orthogonal matrix, 371
orthogonal projection, 241
orthogonal vectors, 368
orthonormal basis for $(\mathbb{R}^n, \cdot)$, 370

parallel displacement, 204
parametic representation, 209
parametric representation, 218
particular solution, 130, 157
perpendicular, 368
pivot, 138
position vector, 205
positive definite, 397
positive orthogonal matrix, 372
positive semi-definite, 397
powers of a square matrix, 180
proper local maximum value, 116
proper local minimum value, 116
proper vector, 365

quadratic form, 392

range, 71, 284, 292
rank, 145
reduced row echelon form, 138, 139
regular tetrahedron, 247
remainder function, 99
rotation matrix, 389
row operations, 133
row vector, 126
rule of conversion for equations, 129

second-order differential equations
    existence and uniqueness, 461
singular matrix, 174
smooth function, 97
solution set, 128, 130
span, 225, 256
special orthorgonal, 372

spectral theorem for symmetric matrices, 380
squaree matrices, 184
stability, 251
standard basis, 221, 223, 264, 265
standard coordinate system in 3-space, 223
standard coordinate sytem, 221
standard parameter form, 129
stationary point, 113
subspace, 277
superposition, 456
symmetric matrix, 172, 368
system matrix, 428
system of differential equations, 427
system of differential equations
    existence and uniqueness, 439
system of linear equations, 130

the characteristic matrix, 338
the characteristic polynomial, 338
the complex guess method, 458
the inner product, 364
the scalar product, 364
the tangent to the graph, 80
transpose matrix, 169
trivial equation, 128, 144

unit vector, 205

vector, 251
vector space, 251
vector space over the complex numbers, 251
vector space over the real numbersl, 251
vector space with inner product, 364

zero vector, 205
zero-extension of a function, 74