Introduction to the Theory of Computing I.

Lecture notes by Dávid Tóth based on the notes of Dávid Szeszlér

2020

Contents

Introduction

These notes are based on the lecture notes [10] written by Dávid Szeszlér in Hungarian and cover the material of the course Introduction to the Theory of Computing I. given in every fall semester at the Faculty of Electrical Engineering and Informatics of Budapest University of Technology and Economics. The text follows closely the structure of the Hungarian version, many parts of it are just translations of the original.

The material is divided into two chapters, the first one covers the basics of number theory and also some applications. In the second one we discuss the basics of linear algebra. We will only see a special case of a general theory, though this is without doubt the most important special case. Not only that it provides a very useful tool in almost every branch of mathematics, but it has a fundamental role in many parts of computer science.

I would like to thank Rita Csákány for reading these notes and making comments.

1 Number Theory

Number theory is one of the oldest branches of mathematics. It investigates the properties of the integers, many basic notions of it were defined and named by the ancient Greeks. It provided many of the most famous problems of mathematics, some of them turned out to be very challenging and deep. After hundreds and thousands of years there are still several unsolved questions among them.

Despite all this there was no general interest towards number theory outside mathematics until the first important application appeared in 1977, when Ronald Rivest, Adi Shamir and Len Adleman discovered the so called RSA algorithm (what was named after the initials of its creators). It is used to encrypt and decrypt messages with the help of public keys, i.e. keys that can be given to anyone without endangering the privacy. The connection with cryptography made this branch very important in computer science, especially in the age of the internet. In this chapter we discuss the basics of number theory and describe some of its applications, including the RSA algorithm.

1.1 Basic Notions and the Fundamental Theorem of Arithmetic

In this section we discuss the basic notions of number theory. Most of the definitions and theorems should be familiar to anyone from high school, but here we also give the exact proofs of the claims. Unless it is told otherwise, every variable denotes an integer in this chapter.

Definition 1.1.1. If $a, b \in \mathbb{Z}$ are integers, then we say that a is a *divisor* of b (or a *divides* b, b is a *multiple* of a) if there is an integer $c \in \mathbb{Z}$ such that $b = ac$. This is denoted by $a \mid b$. If a does not divide b, then we write $a \nmid b$. The number a is a *proper divisor* of b if $a \mid b$ and $1 < |a| < |b|$ hold.

Note that other authors may not exclude the number 1 from the set of proper divisors. One checks easily that $13 \mid 91$, $-7 \mid 63$, $2 \mid 0$ and $-8 \nmid -36$ hold. At first sight it is maybe surprising that $0 \mid 0$ holds too since $0 = 0 \cdot c$ for every $c \in \mathbb{Z}$. But this does not mean that the operation "dividing by zero" is defined. The divisors of 10 are $\pm 1, \pm 2, \pm 5$ and ± 10 while the proper divisors of 10 are ± 2 and ± 5.

Definition 1.1.2. The integer $p \in \mathbb{Z}$ is called *prime* if $|p| > 1$ and p does not have a proper divisor. In other words: $p = ab$ holds if and only if $a = \pm 1$ or $b = \pm 1$. If $|p| > 1$ and p is not prime, then it is called a *composite number*. The numbers 0 and ± 1 are neither prime nor composite.

Examples of prime numbers are 3, 103 and -7. The negative primes are just the opposites of the positive primes.

Remark. Many authors call the above defined numbers *irreducibles* and define the notion of prime numbers by the property that if $p \mid ab$ holds for a product, then $p \mid a$ or $p \mid b$ must also hold. Since these two definitions give the same notion for integers, we do not follow this practice. The reason why others do it is that number theory can be worked out in "larger domains" and in general the two notions may differ. We will see such examples later but aside from these we restrict ourselves to the set of integers and recommend the book [6] to the interested reader.

The following theorem has a crucial role in number theory (which is reflected in its name) and also shows the importance of primes:

Theorem 1.1.1 (Fundamental Theorem of Arithmetic)**.** *Every integer different from* 0 *and* ± 1 *can be represented as a product of primes. This representation is unique up to the order and the sign of the factors.*

For example two different representations of the number 100 are $2{\cdot}2{\cdot}5{\cdot}5$ and $(-5){\cdot}2{\cdot}(-2){\cdot}5$, which shows that uniqueness cannot be achieved in the theorem without disregarding the order and the sign of the prime factors. We can also see why it is useful to exclude the numbers ± 1 from the set of primes. Otherwise the representation would not be unique since we could write $4 = 2 \cdot 2 = 1 \cdot 2 \cdot 2$. On the other hand, the numbers 0 and ± 1 can not be written as product of primes, they must be excluded in the theorem. Note that prime numbers can be considered as products that have only one factor and then the statement of the theorem remains true for them too.

Proof of existence of the factorization in Theorem 1.1.1. We give a simple process which provides the factorization for any $n \in \mathbb{Z}$ with $|n| > 1$. We will store a factorization all along, initially this will be the number n itself (a product with one factor). Once we have a product $n = a_1 a_2 \dots a_k$ where all the a_i's are prime numbers we stop. If at least one of the factors, say a_i is composite, then it has a proper divisor. That is, we can choose some $b, c \in \mathbb{Z}$ with $|b|, |c| > 1$ such that $a_i = bc$. We replace the factor a_i with bc in the product and proceed. In every step we increase the number of factors by 1 and the absolute value of every factor is at least 2. Hence after at most $\log_2 |n|$ steps our procedure ends and gives the required factorization. □

Before we complete the proof of the fundamental theorem, we make some remarks and show some (counter)examples. First note that the (at this point still unproved) uniqueness part is the "powerful" part of the fundamental theorem. Namely, it assures that the obtained factorization gives the arithmetic structure of the numbers and this way it makes possible to calculate all of their divisors, for example.

Although the fundamental theorem may seem evident, it is not too hard to give such "domains" where it does not hold. For instance, let us forget about the odd numbers for a moment. The set of even numbers is similar to the integers. By this we mean that the sum, difference and product of two even numbers is also even. Moreover, the notion of divisibility can be defined the same way as before. But here we do not have a unique factorization: for example $36 = 2 \cdot 18 = 6 \cdot 6$ and none of these representations can be split up further. The reader may notice that our definition for the prime numbers is not applicable here, because the number 1 is not an element of our set (i.e. it is not even). However, it is not hard to modify the definition so that it yields the right notion.

A more sophisticated example is the set of complex numbers of the form $a+ib\sqrt{5}$, where a and b are integers and i is the imaginary unit, i.e. $i^2 = -1$. Again, this is closed under addition and multiplication, but also contains the number 1. It is true that $9 = 3^2 = (2-i\sqrt{5})(2+i\sqrt{5})$ but these factors do not have "proper divisors". Of course we should clarify what a proper divisor means here, but we do not go into the details, we refer to the book [6] instead.

As a final remark, we mention that though these domains may seem artificial for the first sight, still examples similar to the last one occur naturally in number theory. For example, they play a major role in problems like Fermat's Last Theorem which was formulated in 1637 and was proved by Andrew Wiles in 1994. The theorem states that for any exponent $n \in \mathbb{N}$

greater than 2 the equation $x^n + y^n = z^n$ does not have an integer solution. Many special cases and similar problems can be treated relatively easily, but they are beyond the scope of these notes.

Proof of uniqueness of the factorization in Theorem 1.1.1. It is clearly enough to show that every positive integer greater than 1 can be written uniquely (up to order) as a product of primes. So assume that $n \in \mathbb{N}$, $n > 1$. We prove by induction. The assertion is true for every prime, in particular for $n = 2$, so assume that $n > 2$ is composite and the assertion is true for every $1 < n' < n$. If $n = p_1 \dots p_r = q_1 \dots q_s$ such that the p_i's and q_j's are primes, then $r, s \geq 2$ (since n is not a prime). If $p_i = q_j$ holds for some i and j, then dividing n by this prime we get two non-empty products giving a smaller number n'. By induction the remaining primes on the two sides of the equality differ by order only, hence the same holds for the original products.

It remains to handle the case when $p_i \neq q_j$ for every i and j. After a possible relabeling we may assume that $p_1 \leq p_i$ and $p_1 \leq q_j$ hold for every i and j. Let us define then $n' = (q_1 - p_1)q_2 \dots q_s$. We have assumed $q_1 \geq p_1$ and $q_1 \neq p_1$, hence $n > n' > q_1 - p_1 \geq 1$ follows (since $s \geq 2$). We now show n' has a factorization which contains p_1 and another one without p_1. This contradicts our hypothesis and this contradiction shows that this case is impossible and the theorem is proved. If $q_1 - p_1 = 1$, then we can simply omit this factor from the product to obtain an appropriate representation of n'. Otherwise $q_1 - p_1$ can be written uniquely (up to order) as a product of primes by induction. Replacing this factor by this product in the definition of n' above we get a factorization of n'. Since $p_1 \nmid q_1$ (because q_1 is prime) we also have that $p_1 \nmid q_1 - p_1$. So p_1 does not occur among the primes in the factorization of $q_1 - p_1$. Recall that $p_1 \neq q_j$ is also true, hence we get a factorization without the prime p_1.

Finally,

$$\begin{aligned} n' &= (q_1 - p_1)q_2 \dots q_s = q_1 q_2 \dots q_s - p_1 q_2 \dots q_s \\ &= p_1 p_2 \dots p_r - p_1 q_2 \dots q_s = p_1(p_2 \dots p_r - q_2 \dots q_s). \end{aligned}$$

Replacing $p_2 \dots p_r - q_2 \dots q_s$ by an optional prime factorization of it or simply omit this factor in the case when it equals 1 we get a prime factorization of n' including p_1. This is a contradiction, and the proof of the theorem is now complete. □

The fundamental theorem was proved for the set of integers, but then it follows also for the natural numbers: every positive integer greater than 1 has a prime factorization which is unique up to order. This makes it possible to define the *canonical representation* of the positive integers. We obtain this by collecting the identical primes in the factorization into powers and by ordering the powers by the magnitude of the bases. That is, we get the form $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$, where $p_1 < p_2 < \dots < p_k$ are primes and $\alpha_1, \dots, \alpha_k$ are positive integers. Observe that this canonical representation is unique, though many times we only require that the prime bases in this representation are pairwise different (and not necessarily ordered by magnitude). Hopefully this causes no confusion in the future. As an example, the canonical representation of the number 600 is $2^3 \cdot 3^1 \cdot 5^2$ (of course we often omit the exponent 1).

Many times it is useful to allow the exponent zero in the representation. For example it makes possible to use the same primes in the representations of two different numbers, as in the following

Proposition 1.1.2. *Let us assume that $p_1, \dots, p_k$ are pairwise different positive primes and $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$, where $\alpha_1, \dots, \alpha_k$ are non-negative integers. Then the positive integer m divides n if and only if $m = p_1^{\beta_1} \dots p_k^{\beta_k}$, where $0 \leq \beta_1 \leq \alpha_1, \dots, 0 \leq \beta_k \leq \alpha_k$ are integers.*

Proof. If m is of the form given in the proposition, then $n = ml$, where $l = p_1^{\alpha_1 - \beta_1} \dots p_k^{\alpha_k - \beta_k}$, hence $m \mid n$.

Now assume that $m \mid n$ and that the canonical representation of m is $q_1^{\gamma_1} \dots q_s^{\gamma_s}$. Then $n = ml$ for some $l \in \mathbb{Z}$. We can get a factorization of n by multiplying the factorization of m and l. But then by the uniqueness part of the fundamental theorem every q_i must coincide with some p_j. This means that m can be written as $p_1^{\beta_1} \dots p_k^{\beta_k}$ where some of the exponents may be 0. Assume that an exponent, say β_i is strictly bigger than α_i, then

$$p_i^{\beta_i - \alpha_i} \mid \frac{n}{p^{\alpha_i}} = p_1^{\alpha_1} \dots p_{i-1}^{\alpha_{i-1}} p_{i+1}^{\alpha_{i+1}} \dots p_k^{\alpha_k},$$

where $\beta_i - \alpha_i \geq 1$. The same way as before we get that p_i must coincide with some p_j, $j \neq i$. But this is impossible, since the primes $p_1, \dots, p_k$ are pairwise distinct. □

This last result makes it possible to give a formula for the number of divisors. For a positive integer n the number of its divisors is denoted by $d(n)$ (note that other notations like $\nu(n)$, $\tau(n)$ and $\sigma_0(n)$ are also common).

Corollary 1.1.3. *If $n > 1$ is an integer and its canonical representation is $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$, then*

$$d(n) = (\alpha_1 + 1) \dots (\alpha_k + 1).$$

Proof. The product given in the statement is the number of products of the form $p_1^{\beta_1} \dots p_k^{\beta_k}$, where $0 \leq \beta_1 \leq \alpha_1, \dots, 0 \leq \beta_k \leq \alpha_k$. By the previous proposition these products give all the divisors of n, and by the uniqueness of the prime factorization they give every divisor only once. □

Proposition 1.1.2 also helps us to determine the greatest common divisor and the least common multiple of two numbers. Although these notions are basically defined by their names, we give the formal definitions:

Definition 1.1.3. If $n, m \in \mathbb{Z}$ are integers and at least one of them is non-zero, then their *greatest common divisor* (often abbreviated by gcd) is the largest positive integer which divides both n and m. The greatest common divisor of n and m is denoted by (n, m) or $\gcd(n, m)$. The integers n and m are called *co-prime* if $(n, m) = 1$ holds.

Definition 1.1.4. If $n, m \in \mathbb{Z} \setminus \{0\}$ are non-zero integers, then their *least common multiple* (abbreviated by lcm) is the smallest positive number that is divisible by both n and m. The least common multiple of n and m is denoted by $[n, m]$ or $\operatorname{lcm}(n, m)$.

Note that if n is and integer, then the divisors and multiples of n and $-n$ are the same, hence we have $(n, m) = (|n|, |m|)$ and $[n, m] = [|n|, |m|]$. Also, for any positive integer n we have $(n, 0) = n$. Hence for the rest of this section we restrict ourselves to the case when n and m are positive integers.

Now we are going to use the prime factorization of the numbers to compute their greatest common divisor and least common multiple (we will address the effectiveness of this method later).

Proposition 1.1.4. *If $p_1, \dots, p_k$ are pairwise different positive primes, $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$ and $m = p_1^{\beta_1} \dots p_k^{\beta_k}$, where $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k$ are non-negative integers, then*

$$(n, m) = p_1^{\min\{\alpha_1, \beta_1\}} \dots p_k^{\min\{\alpha_k, \beta_k\}},$$
$$[n, m] = p_1^{\max\{\alpha_1, \beta_1\}} \dots p_k^{\max\{\alpha_k, \beta_k\}}.$$

Before the proof we show an example. If $n = 600$ and $m = 84$, then their canonical representations are $600 = 2^3 \cdot 3^1 \cdot 5^2$ and $84 = 2^2 \cdot 3^1 \cdot 7^1$. Observe that there are different primes in these factorizations, hence to apply the previous proposition we have to write them differently, using all the primes that occur in the two products. That is, $600 = 2^3 \cdot 3^1 \cdot 5^2 \cdot 7^0$ and $84 = 2^2 \cdot 3^1 \cdot 5^0 \cdot 7^1$. Of course, unlike in the case of the *canonical* representation here it is necessary to allow the exponent zero. Now the formulae above are applicable: $(600, 84) = 2^2 \cdot 3^1 \cdot 5^0 \cdot 7^0 = 12$ and $[600, 84] = 2^3 \cdot 3^1 \cdot 5^2 \cdot 7^1 = 4200$.

Proof of Proposition 1.1.4. By Proposition 1.1.2, for any positive integer d the properties $d \mid n$ and $d \mid m$ hold simultaneously if and only if $d = p_1^{\gamma_1} \dots p_k^{\gamma_k}$, where $0 \le \gamma_i \le \alpha_i$ and $0 \le \gamma_i \le \beta_i$, i.e. $0 \le \gamma_i \le \min\{\alpha_i, \beta_i\}$ for every i. This holds also for (n, m), and since (n, m) is the greatest among the positive divisors, we must have equality in the previous inequalities, otherwise we could get a greater divisor by increasing an exponent. The proof of the other claim is similar and left to the reader. □

Note that this proof gives more. Namely, the greatest common divisor of two numbers has the following special property:

Corollary 1.1.5. *Let $n, m \in \mathbb{N}^+$ be positive integers. Then the common divisors of n and m are the divisors of their greatest common divisor, i.e. $d \mid n$ and $d \mid m$ holds simultaneously if and only if $d \mid (n, m)$.*

Proof. The greatest common divisor of n and m divides both numbers, i.e. $n = (n, m) \cdot c_1$ and $m = c_2 \cdot (n, m)$ for some c_1, c_2 integers. If $d \mid (n, m)$, then $(n, m) = de$, so $n = d(ec_1)$ and $m = d(ec_2)$, that is, both $d \mid n$ and $d \mid m$ hold.

On the other hand, if both $d \mid n$ and $d \mid m$ hold, then the formula for (n, m) in the previous statement and the first sentence of the previous proof together with Proposition 1.1.2 give that $d \mid (n, m)$. □

Exercise 1.1.1. Assume that $n, m \in \mathbb{N}^+$ are positive integers and let $\langle n, m \rangle$ denote the least positive integer for which both $n \mid m \cdot \langle n, m \rangle$ and $m \mid n \cdot \langle n, m \rangle$ hold. Give a formula for $\langle n, m \rangle$ that is similar to the ones in Proposition 1.1.4.

We close this section by a basic theorem about the number of primes:

Theorem 1.1.6. *The number of primes is infinite.*

Proof. It is enough to prove that there are infinitely many positive primes. So in the proof every prime is assumed to be positive.

Assume on the contrary that the number of primes is finite, say k. Let $p_1, \dots, p_k$ be the list of all primes. Then $N = p_1 \dots p_k + 1$ is bigger than 1, hence it has a prime factorization. Since N is not divisible by any of the primes $p_1, \dots, p_k$, every prime in the factorization of N must be different from them, and this is a contradiction. □

1.2 Congruences

The set of integers is closed under addition, subtraction and multiplication, but this is not the case with the fourth basic operation. The result of a division is not always an integer (and we cannot divide by 0 at all). What we can do is *division with remainders*. Namely, for every $a, b \in \mathbb{Z}$, $b \neq 0$ there exist integers q, r such that $a = qb + r$ where $0 \leq r \leq |b| - 1$. This is obvious since if we regard the integers below a (and also a itself), then we can find one within the distance $|b| - 1$ which is divisible by b. Since among $|b|$ consecutive numbers there is *exactly* one which is divisible by b, we get that the number r (and then also q) is determined uniquely. The number r is called the *remainder* (and q is the *quotient*). For example, if we divide -30 by 9, then the remainder is 6 (since $-30 = (-4) \cdot 9 + 6$). This makes it possible to define the *congruence relation*:

Definition 1.2.1. Let $a, b, m \in \mathbb{Z}$ be integers and $m \neq 0$. We say that a and b are *congruent* (or a is congruent to b) modulo m if they give the same remainder when we divide them by m. This is denoted by $a \equiv b \pmod{m}$ or $a \equiv b\ (m)$. The number m is called the *modulus* of the congruence.

For example, $17 \equiv 52 \pmod{7}$ (because both of them gives the remainder 3) and $33 \equiv -30 \pmod{9}$ (here the remainder is 6). The notation of the congruence resembles the notation of equality, and this is not a coincidence. It expresses that we consider a and b the same when we count with the remainders. The following equivalent definition of the congruence is often useful:

Proposition 1.2.1. *If $a, b, m \in \mathbb{Z}$, $m \neq 0$, then $a \equiv b \pmod{m}$ if and only if $m \mid a - b$.*

Proof. Let us denote the remainder of a modulo m by r_a. Similarly, let r_b be the remainder of b. Then $a = q_a m + r_a$ and $b = q_b m + r_b$ for some q_a, q_b integers. If $r_a = r_b$, then $m \mid a - b = (q_a - q_b)m$. On the other hand, if $r_a \neq r_b$, then $a - b = (q_a - q_b)m + r_a - r_b$, where $0 \neq |r_a - r_b| < m$, and hence $m \nmid a - b$ (because the distance between two multiples of m is at least m). □

The following proposition shows why using the congruence relation makes the computations often easier:

Proposition 1.2.2. *Assume that $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$ hold for some integers $a, b, c, d, m \in \mathbb{Z}$, $m \neq 0$ and let $k \in \mathbb{Z}$ be an arbitrary integer. Then the following hold:*

(i) $a + c \equiv b + d \pmod{m}$,

(ii) $a - c \equiv b - d \pmod{m}$,

(iii) $ac \equiv bd \pmod{m}$,

(iv) $a^k \equiv b^k \pmod{m}$.

Proof. By the previous proposition our assumption is equivalent to the conditions $m \mid a - b$ and $m \mid c - d$. From these we get that $m \mid (a - b) + (c - d) = (a + c) - (b + d)$, which means that $a + c \equiv b + d \pmod{m}$ (again, by the previous proposition). Similarly, we have $m \mid (a - b) - (c - d) = (a - c) - (b - d)$, hence $a - c \equiv b - d \pmod{m}$ hold. This proves (i) and (ii).

To show (iii) we note that if $m \mid a-b$, then $m \mid c(a-b) = ac-bc$ follows. The same way, we get from $m \mid c-d$ that $m \mid b(c-d) = bc-bd$. But the sum of numbers divisible by m is again divisible by m, hence we have $m \mid ac-bc+bc-bd = ac-bd$, and this is equivalent to $ac \equiv bd \pmod{m}$.

Finally, (iv) follows from (iii): if we set $c = a$ and $d = b$, then (iii) gives $a^2 \equiv b^2 \pmod{m}$. Now we apply (iii) to the latter congruence and to $a \equiv b \pmod{m}$, and this way we obtain $a^3 \equiv b^3 \pmod{m}$. Continuing this way we get $a^k \equiv b^k \pmod{m}$ after $k-1$ steps. □

We often use the previous statements in the special case when $c = d$. As obviously $c \equiv c \pmod{m}$, we get that if $a \equiv b \pmod{m}$, then also $a \pm c \equiv b \pm c \pmod{m}$ and $ac \equiv bc \pmod{m}$. But the analogous claim does not hold for the division. Of course to be able to divide a congruence by a number c we must have integers on both sides which are divisible by c. But one has to be careful even in that case, for example $40 \equiv 64 \pmod{12}$, but dividing by 8 we get $5 \equiv 8 \pmod{12}$, which is false. The right form of the division rule is the following:

Theorem 1.2.3. *Let $a, b, c, m \in \mathbb{Z}$ be integers, $m \neq 0$ and $d = (c, m)$ (the greatest common divisor of m and c). Then $ac \equiv bc \pmod{m}$ if and only if $a \equiv b \pmod{\frac{m}{d}}$.*

Proof. If $c' = \frac{c}{d}$ and $m' = \frac{m}{d}$, then c' and m' are integers since d is a common divisor of c and m. Moreover $(c', m') = 1$, otherwise the number $d \cdot (c', m')$ would be a common divisor of m and c which is bigger than d, and this contradicts the definition of the greatest common divisor.

Now $ac \equiv bc \pmod{m}$ if and only if $m \mid ac-bc = c(a-b)$ by Proposition 1.2.1. That is, we have $c(a-b) = mk$ for some integer k. Dividing both sides by d we get the equivalent equation $c'(a-b) = m'k$. If $m' \nmid a-b$, then at least one prime divisor of m' must divide c' by the fundamental theorem, but since m' and c' are co-prime (which means that their greatest common divisor is 1), this is impossible. It follows that $m' \mid a-b$, i.e. $a \equiv b \pmod{m'}$.

On the other hand, if $a \equiv b \pmod{m'}$, then $m' \mid a-b$ and hence $m' \mid c'(a-b)$. This means that $c'(a-b) = m'k$ for some integer k, and we have already seen that this is equivalent to $ac \equiv bc \pmod{m}$. □

Corollary 1.2.4. *Assume that $a, b, c, m \in \mathbb{Z}$, $m \neq 0$ and $(m, c) = 1$ (that is, c and m are co-prime). Then $ac \equiv bc \pmod{m}$ if and only if $a \equiv b \pmod{m}$.*

Exercise 1.2.1. What is the remainder when we divide

a) 100^{100} by 11; b) 654^{321} by 655; c) 111^{41} by 35?

Solution. We use the properties of the congruence relation that are given in Proposition 1.2.2.

a) Since $11 \mid 99$ we have $100 \equiv 1 \pmod{11}$. Raising both sides to the power 100 and using property (iv) we get that $100^{100} \equiv 1^{100} = 1 \pmod{11}$ (and hence the remainder is 1).

b) Observe that $654 \equiv -1 \pmod{655}$, hence $654^{321} \equiv (-1)^{321} = -1 \pmod{655}$ by property (iv). The remainder of 654^{321} is then 654.

c) First note, that $111 \equiv 6 \pmod{35}$, so $111^{41} \equiv 6^{41} \pmod{35}$. At this point the result is not clear, but notice that $6^2 \equiv 1 \pmod{35}$. From this we obtain that $6^{40} = (6^2)^{20} \equiv 1^{20} = 1 \pmod{35}$, and then $6^{41} = 6^{40} \cdot 6 \equiv 1 \cdot 6 \pmod{35}$, i.e. the remainder is 6. □

1.3 The Euler-Fermat Theorem

The aim of this section to show that for appropriate values of a, m and k the congruence $a^k \equiv 1 \pmod{m}$ holds. We make use of this later in the RSA algorithm. One must be careful though, since if $(a, m) = d > 1$, then of course $d \nmid a^k - 1$ for any integer $k > 0$ (because $d \mid a^k$). On the other hand, in the case when a and m are co-prime we can find an appropriate k which depends only on m and not on a. To be able to formulate the precise statement we will need a tool which we introduce below.

1.3.1 Euler's Phi Function

Two numbers that are congruent to each other behave similarly from many points of view. The following statement says that even their greatest common divisor with m agrees:

Proposition 1.3.1. *Assume that $a, b, m \in \mathbb{Z}$ and $m \neq 0$. If $a \equiv b \pmod{m}$ holds, then $(a, m) = (b, m)$.*

Proof. Assume that $a \equiv b \pmod{m}$, i.e. $m \mid a - b$. This means that $b = a + km$ for some $k \in \mathbb{Z}$. If $d = (a, m)$, then since $d \mid a$ and $d \mid km$, we get that $d \mid a + km = b$. In other words, d is a common divisor of b and m. It follows that $d = (a, m) \leq (b, m)$, because the latter number is the greatest among the positive common divisors. Since the role of a and b is symmetric, we have $(b, m) \leq (a, m)$ as well, and the claim follows. □

Corollary 1.3.2. *If $a \equiv b \pmod{m}$, then $(a, m) = 1$ if and only if $(b, m) = 1$.*

Definition 1.3.1. If $n \geq 1$, then we denote by $\varphi(n)$ the number of those integers in the interval $[1, n]$ which are co-prime to n, that is,

$$\varphi(n) = \left|\{k \in \mathbb{N} : 1 \leq k \leq n,\ (k, n) = 1\}\right|.$$

The function φ is called *Euler's phi function*.

The congruence relation modulo n divides the set of integers into disjoint classes, these are called *residue classes* modulo n. Two integers belong to the same class if and only if they are congruent. The system of residue classes modulo n is complete in the sense that every integer belongs to a class. Since every class contains exactly one element in the interval $[1, n]$, we get by the previous Corollary that $\varphi(n)$ is the number of the residue classes modulo n which contain numbers that are co-prime to n.

We determine the value of $\varphi(10)$. Among the numbers $1, 2, \ldots, 10$ the even numbers and the multiples of 5 have a common divisor with 10 greater than 1, but the remaining numbers are co-prime to 10. These are 1, 3, 7 and 9, hence $\varphi(10) = 4$. If $n = p$ is prime, then all the numbers $1, \ldots, p-1$ are co-prime to p, so $\varphi(p) = p - 1$. It is also easy to determine the value of φ for prime powers:

Lemma 1.3.3. *If p is a prime and $\alpha \geq 1$ is a positive integer, then $\varphi(p^\alpha) = p^\alpha - p^{\alpha-1}$.*

Proof. The numbers among $1, \ldots, p^\alpha$ that are co-prime to p^α are the ones which are not divisible by p. So we exclude the numbers kp, where k is a positive integer and $kp \leq p^\alpha$, i.e. $k \leq p^{\alpha-1}$. This proves the claim. □

The computation of φ based on the definition becomes tiresome for a general composite number. However, we can use the following lemma and the canonical form of the number to give a formula for $\varphi(n)$.

Lemma 1.3.4. *If a and b are co-prime positive integers, then $\varphi(ab) = \varphi(a)\varphi(b)$.*

Remark. A function defined on the set of positive integers is called *multiplicative* if it has the property described in the lemma. As co-prime numbers have no common primes in their canonical representations, it follows easily from Corollary 1.1.3 that the function $d(n)$ defined in the first section is multiplicative. To learn more about multiplicative arithmetic functions see e.g. [5].

We postpone the proof of the lemma and first apply it to give a formula for $\varphi(n)$:

Theorem 1.3.5. *If $n > 1$ is a positive integer with canonical representation $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$, then*

$$\varphi(n) = (p_1^{\alpha_1} - p_1^{\alpha_1 - 1}) \dots (p_k^{\alpha_k} - p_k^{\alpha_k - 1}).$$

Proof. First note that form the previous lemma if follows by induction that if $a_1, \dots, a_k$ are pairwise co-prime numbers (i.e. $(a_i, a_j) = 1$ for every $1 \leq i, j \leq k$, $i \neq j$), then $\varphi(a_1 \dots a_k) = \varphi(a_1) \dots \varphi(a_k)$. Indeed, the lemma gives this for $k = 2$. Assume that $k > 2$ and the statement is true for $k - 1$. If $a_1, \dots, a_k$ are pairwise co-prime numbers, then $(a_1 \dots a_{k-1}, a_k) = 1$, because if a prime divides both a_k and the product, then this prime occurs in the canonical representation of some a_i where $1 \leq i \leq k - 1$. But this is impossible since $(a_i, a_k) = 1$ holds. Then $\varphi(a_1 \dots a_{k-1}a_k) = \varphi(a_1 \dots a_{k-1})\varphi(a_k)$ by the previous lemma, and using the assumption for $k - 1$ numbers we get the claim.

Now we apply this for the numbers $a_i = p_i^{\alpha_i}$ which are pairwise co-prime, and hence $\varphi(n) = \varphi(p_1^{\alpha_1}) \dots \varphi(p_k^{\alpha_k})$ holds. Finally, applying Lemma 1.3.3 we get the statement of the theorem. □

Proof of Lemma 1.3.4. First note that for the positive integers $x, a, b \in \mathbb{N}^+$ $(x, ab) = 1$ holds if and only if both $(x, a) = 1$ and $(x, b) = 1$ hold. Indeed, we get the prime factorization of ab by multiplying the factorizations of a and b, so if a prime divides x and ab, then it divides a or b. That is, if $(x, ab) > 1$, then $(x, a) > 1$ or $(x, b) > 1$ must hold. On the other hand, if x and a or x and b have a common prime divisor, then it divides ab as well. It follows that $\varphi(ab)$ is the number of those integers between 1 and ab that are co-prime to both a and b.

We write the numbers $1, 2, \dots, ab$ in a table so that the intersection of the ith row and jth column contains the number $m_{ij} = (i-1)b + j$, where $1 \leq i \leq a$, $1 \leq j \leq b$. In this table we will search for numbers that are co-prime to both a and b. The following table shows the case $a = 3$ and $b = 8$.

$$\begin{bmatrix} \boxed{1} & 2 & 3 & 4 & \boxed{5} & 6 & \boxed{7} & 8 \\ 9 & 10 & \boxed{11} & 12 & \boxed{13} & 14 & 15 & 16 \\ \boxed{17} & 18 & \boxed{19} & 20 & 21 & 22 & \boxed{23} & 24 \end{bmatrix}$$

First note that $m_{i,j} = (i-1)b + j \equiv j \pmod{b}$ for every i, j, hence by Proposition 1.3.1 we have that $(m_{ij}, b) = (j, b)$. In particular, $(m_{ij}, b) = 1$ holds if and only if $(j, b) = 1$. This means that the numbers in the table that are co-prime to b are those which lie in the jth column for some j co-prime to b. This narrows down the scope of our search to $\varphi(b)$ columns.

Now we are going to count the numbers in the jth column that are co-prime to a. In fact, we show that any two different numbers in the jth column are not congruent to each other modulo a, and since there are a rows in our table, it follows that the numbers in he jth column form a complete residue system modulo a and hence there are $\varphi(a)$ numbers among them that are co-prime to a. Putting this and the result of the previous paragraph together we get the claim.

So assume that $m_{ij} = (i-1)b+j \equiv m_{kj} = (k-1)b+j \pmod{a}$ for some $1 \leq i, k \leq a$. By property (ii) in Proposition 1.2.2 we can subtract j from both sides, and since $(a, b) = 1$, it follows from Corollary 1.2.4 that we can divide the congruence by b. We get that $i-1 \equiv k-1 \pmod{a}$, that is, $i \equiv k \pmod{a}$. But since $1 \leq i, k \leq a$, we have $i = k$, i.e. $m_{ij} = m_{kj}$. This completes the proof of the lemma. □

1.3.2 Residue Systems

Every residue class modulo m can be represented by any one of its members. That is, any member of a class identifies it. We often represent a class by the smallest non-negative integer which belongs to the class, i.e. every class modulo m can be represented by an integer $0 \leq c \leq m-1$. Moreover, these non-negative integers represent every class exactly once. The subsets of integers with this property are called *complete residue systems*. We introduce another important residue system. Recall that if a residue class modulo m contains a number which is co-prime to m, then every member of that class has the same property by Corollary 1.3.2. We have seen that the number of these classes is $\varphi(m)$. If each of these classes is represented exactly once, then we call the system *reduced*.

Definition 1.3.2. The system $R = \{c_1, \ldots, c_k\}$ of integers is called a *reduced residue system* modulo m if the following hold:

(i) $(c_i, m) = 1$ hold for every $1 \leq i \leq k$,

(ii) $c_i \not\equiv c_j \pmod{m}$ for any $1 \leq i, j \leq k$, $i \neq j$,

(iii) $k = \varphi(m)$.

The systems $\{1, 3, 7, 9\}$, $\{21, 43, 67, 89\}$ and $\{1, -1, 3, -3\}$ are reduced modulo 10.

Proposition 1.3.6. *Assume that $R = \{c_1, \ldots, c_k\}$ is a reduced residue system modulo m and $a \in \mathbb{Z}$ is an arbitrary integer with $(a, m) = 1$. Then $R' = \{ac_1, \ldots, ac_k\}$ is also a reduced residue system modulo m.*

Proof. We are going to show that the properties (i), (ii) and (iii) in the previous definition hold for R'. To see (i) we set $d_i = (ac_i, m)$. If $p \mid d_i$ is a prime, then it occurs in the prime factorization of both m and ac_i. As we get the prime factorization of ac_i by multiplying the factorization of a and c_i, p must divide at least one of them (and also m). But this contradicts the assumption $(a, m) = (c_i, m) = 1$, and it follows that $(ac_i, m) = 1$.

Assume now that $ac_i \equiv ac_j \pmod{m}$ for some $1 \leq i, j \leq k$. Then by Corollary 1.2.4 this is equivalent to $c_i \equiv c_j \pmod{m}$, because $(a, m) = 1$ holds. Since R is a reduced residue system, this can hold if and only if $i = j$, so property (ii) is proved.

Finally, the number of the elements of the systems R' and R is the same, hence (iii) follows for R'. □

1.3.3 The Euler-Fermat Theorem

After this preparation we are in the position to state and prove the so called Euler-Fermat theorem:

Theorem 1.3.7 (Euler-Fermat theorem)**.** *If $a, m \in \mathbb{Z}$ are integers, $m \neq 0$ and $(a, m) = 1$, then $a^{\varphi(m)} \equiv 1 \pmod{m}$ holds, where φ is Euler's phi function.*

Proof. Let $R = \{c_1, \dots, c_k\}$ be an arbitrary reduced residue system modulo m. Since $(a, m) = 1$, we have by Proposition 1.3.6 that $R' = \{ac_1, \dots, ac_k\}$ is also a reduced residue system modulo m. For every remainder $0 \le r \le m-1$ with $(r, m) = 1$ there is exactly one number in both R and R' which is congruent to r. Hence we can pair the numbers in R and R' so that the pairs are congruent to each other. Then by property (iii) in Proposition 1.2.2 we can multiply the numbers in R and R' and this way we still get numbers that are congruent to each other:

$$c_1 \dots c_k \equiv (ac_1) \dots (ac_k) = a^{\varphi(m)} c_1 \dots c_k \qquad (\text{mod } m),$$

where we used that $k = \varphi(m)$. Since $(c_i, m) = 1$, it follows from Corollary 1.2.4 that we can divide the previous congruence by c_i for every $1 \le i \le k$. After doing this for every i we get the statement of the theorem. □

Corollary 1.3.8 (Fermat's little theorem)**.** *If p is a positive prime and $a \in \mathbb{Z}$ is an arbitrary integer, then $a^p \equiv a \pmod{p}$.*

Proof. If $p \mid a$, then $p \mid a^p$ also holds, hence $a^p \equiv 0 \equiv a \pmod{p}$. If $p \nmid a$, then $(a, p) = 1$, because p is a prime. Then by the previous theorem we have $a^{\varphi(p)} = a^{p-1} \equiv 1 \pmod{p}$. Multiplying both sides by a we get the statement. □

Exercise 1.3.1. What is the remainder when we divide a) 11^{111} by 63 b) $51^{41^{32}}$ by 140?

Solution. a) Since $(11, 63) = 1$, we can apply the Euler-Fermat theorem, which gives that $11^{\varphi(63)} = 11^{36} \equiv 1 \pmod{63}$ (as $\varphi(63) = (7^1 - 7^0)(3^2 - 3) = 6 \cdot 6 = 36$). Now we apply property (iv) of Proposition 1.2.2 for $k = 3$. That is, we raise both sides to the 3rd power to get that $(11^{36})^3 = 11^{108} \equiv 1^3 = 1 \pmod{36}$. That is, $11^{111} = 11^{108} \cdot 11^3 \equiv 1 \cdot 11^3 \pmod{63}$, so it remains to determine the remainder of 11^3. As $11^2 = 121 \equiv -5 \pmod{63}$, we obtain that $11^3 = 11^2 \cdot 11 \equiv (-5) \cdot 11 = -55 \equiv 8 \pmod{63}$, and hence the remainder is 8.

b) We will apply the Euler-Fermat theorem for the numbers $a = 51$ and $m = 140$. This can be done since $51 = 3 \cdot 17$ and $140 = 2^2 \cdot 5 \cdot 7$, and hence $(51, 140) = 1$. We also have that $\varphi(140) = (2^2 - 2)(5 - 1)(7 - 1) = 2 \cdot 4 \cdot 6 = 48$, so $51^{48} \equiv 1 \pmod{140}$ holds by the Euler-Fermat theorem. Maybe it is not clear at first sight how this can be used in this situation. But as before, we have $51^{48k} \equiv 1^k = 1 \pmod{140}$ for every $k \ge 1$. Although the exponent is not of the form $48k$ we still can divide it by 48 with a remainder. That is, we are looking for the smallest non-negative integer r such that $41^{32} \equiv r \pmod{48}$. Luckily, $(41, 48) = 1$ holds, hence we can apply the Euler-Fermat theorem again. As $\varphi(48) = (2^4 - 2^3)(3 - 1) = 16$, we have $41^{16} \equiv 1 \pmod{48}$ and hence $(41^{16})^2 = 41^{32} \equiv 1 \pmod{48}$. This can be written as $41^{32} = 48k + 1$ for some integer k, and then $51^{41^{32}} = 51^{48k+1} = 51^{48k} \cdot 51 \equiv 51 \pmod{140}$, i.e. the remainder is 51. □

1.4 Linear Congruences

In this section we address the following question: if $a, b, m \in \mathbb{Z}$, $m \neq 0$ are given, then what are the numbers for which the congruence $ax \equiv b \pmod{m}$ holds? This problem is called a *linear congruence*, because we have information about the first power of the unknown number x.

First we note, that if a linear congruence has a solution x_0, then $ax_0 \equiv ax_1 \pmod{m}$ holds for every x_1 which is congruent to x_0 modulo m. In other words, if x_0 is a solution, then every number in its residue class modulo m is also a solution. Hence the set of the solutions is a union of residue classes, and we will give the solutions by giving only one representative

from each class which contains solutions, that is, we will write $x \equiv x_0 \pmod{m}$ (and give this way the whole class of x_0).

For example, let us examine the congruence $3x \equiv 2 \pmod{5}$. Multiplying by 2 we get $6x \equiv 4 \pmod{5}$. But $6x \equiv x \pmod{5}$, hence the only option for the solution is the class $x \equiv 4 \pmod{5}$. This is indeed a solution since $3 \cdot 4 \equiv 12 \equiv 2 \pmod{5}$.

Let us try to solve the congruence $10x \equiv 5 \pmod{30}$. If we look at this congruence, we may observe that a number of the form $10x$ has a zero in the end when we write it in the decimal system. On the other hand, if a number gives the remainder 5 when we divide it by 30, then it must end with the digit 5. This means that this congruence has no solutions.

In this section we determine the conditions that are sufficient and necessary for a linear congruence or a system of linear congruences to have a solution. We will also determine the number of the solutions. We give a method in the next section, which determines the solutions "efficiently". The word "efficiently" will also get a more or less precise meaning in the next section.

1.4.1 Existence of solutions

In the last example above we did not have a solution for a linear congruence, and the true reason for this is that the modulus and the coefficient of x had a common divisor which did not divide the right hand side. We formalize this in the following

Theorem 1.4.1. *The linear congruence $ax \equiv b \pmod{m}$ is solvable if and only if $(a, m) \mid b$. If this condition holds, then (a, m) is the number of the different residue classes which contain all the solutions.*

We usually say briefly that the number of solutions modulo m is (a, m).

Proof. First we show that if the congruence is solvable, then $d := (a, m) \mid b$. Let x_0 be a solution of the congruence. Then $m \mid ax_0 - b$ holds, and as $d \mid m$, we have that $d \mid ax_0 - b$. But $d \mid a \mid ax_0$ holds as well, hence $d \mid ax_0 - (ax_0 - b) = b$ follows.

Next we show that if $(a, m) = 1$, then the congruence is solvable. We set $x_0 = a^{\varphi(m)-1}b$, then by the Euler-Fermat theorem we get that $ax_0 = a^{\varphi(m)}b \equiv b \pmod{m}$, i.e. x_0 is indeed a solution.

Now assume that $d = (a, m) \mid b$ and set $a' = a/d$, $b' = b/d$ and $m' = m/d$. Then a', b' and m' are integers, and $(a', m') = 1$ (otherwise $(a', m') \cdot d$ would be a common divisor of a and m which is greater than d). By Theorem 1.2.3 the congruence $ax \equiv b \pmod{m}$ is equivalent to $a'x \equiv b' \pmod{m'}$, and by the previous paragraph this latter congruence has a solution, and hence so does the original congruence.

Now we turn to the number of solutions. Assume that x_1 is an arbitrary solution of the congruence. Now x_2 is another one if and only if $ax_1 \equiv b \equiv ax_2 \pmod{m}$. By Theorem 1.2.3 this is equivalent to $x_1 \equiv x_2 \pmod{m'}$. So every solution is of the form $x_1 + km'$ for some $k \in \mathbb{Z}$, and any of these numbers is a solution. Now $x_1 + k_1m' \equiv x_1 + k_2m' \pmod{m}$ holds if and only if $k_1 \equiv k_2 \pmod{m/m'}$, and as $m/m' = d$, this means that the solutions of the original congruence come from d distinct residue classes modulo m. □

Note that the last paragraph of the proof gives the set of all solutions once we have found one single solution. Namely, if x_1 is a solution, then $x_1 + km'$ $(k = 0, 1, \ldots, (a, m) - 1)$ are the representatives of all distinct residue classes modulo m which contain the solutions, each of them is represented only once.

One may observe that the second and third paragraph of the proof also gives a method to determine a first solution, however this is not useful in practice, because it is often hopeless to make the calculations fast. But the first part of this method is important from the practical point of view. Namely, given a congruence $ax \equiv b \pmod{m}$ with $d = (a, m) \mid b$, we only have to solve the equivalent congruence $a'x \equiv b' \pmod{m'}$, where $a' = a/d$, $b' = b/d$, $m' = m/d$ and $(a', m') = 1$. The solution of this congruence will be a solution of the original one as well.

Exercise 1.4.1. Solve the following congruences:

$$a)\ 68x \equiv 12 \pmod{98}, \qquad b)\ 59x \equiv 4 \pmod{222}.$$

Solution. a) Both sides of the congruence are divisible by 4, and $(4, 98) = 2$, so this congruence is equivalent to

$$17x \equiv 3 \pmod{49}$$

by Theorem 1.2.3. That is, we divided both sides by 4, but we had to divide the modulus by the greatest common divisor of 4 and 98 as well. Now we multiply both sides by 3 to obtain

$$51x \equiv 9 \pmod{49}.$$

Observe that $51 \equiv 2 \pmod{49}$ and hence $51x \equiv 2x \pmod{49}$ holds. Also, $9 \equiv 58 \pmod{49}$, so from the previous congruence we infer

$$2x \equiv 58 \pmod{49},$$

and dividing both sides by 2 we have

$$x \equiv 29 \pmod{49}.$$

There are two residue classes modulo 98 which contain numbers that are congruent to 29 modulo 49, namely the class of 29 and the class of $29 + 49 = 78$. One checks easily that these numbers satisfy the the original congruence (and then so does every number in their classes). So the solutions are $x \equiv 29$ and $x \equiv 78 \pmod{98}$.

One may observe that all steps that we made gave an equivalent form of the former congruence (and not just a consequence of the former ones). We emphasized this at the first step, but then we multiplied and divided by a number which was co-prime to the modulus, so the result was equivalent to the former congruence. Hence it is fact superfluous to check our solutions, all of them must satisfy the original congruence. Also note that Theorem 1.4.1 gives us the number of solutions modulo 98 at the beginning, there are $(98, 68) = 2$ of them. We could also refer to this, and then if we get only two possibilities for the solutions, then both of them must be correct.

b) First we multiply the congruence by 4 to get

$$236x \equiv 16 \pmod{222},$$

and since $236 \equiv 14 \pmod{222}$, we can write this as

$$14x \equiv 16 \pmod{222}.$$

Dividing by 2 (and using Theorem 1.2.3) we get that

$$7x \equiv 8 \pmod{111}.$$

Now we multiply this last congruence by 16:

$$112x \equiv 128 \qquad (\text{mod } 111),$$

and since $112 \equiv 1$ and $128 \equiv 17 \pmod{11}$, we conclude

$$x \equiv 17 \qquad (\text{mod } 111).$$

We get two classes modulo 222, one of them is represented by 17 while the other one by 128. However, a computation shows that $59 \cdot 128 \equiv 4 \pmod{222}$ holds but $59 \cdot 17 \equiv 115 \pmod{222}$. How is this possible? Did we make a mistake? We can find the answer at the first step. It was right in the sense that $236x \equiv 16 \pmod{236}$ follows from the original congruence but it is *not equivalent* to it. But this latter congruence is equivalent to $59x \equiv 4 \pmod{111}$ by Theorem 1.2.3, and the set of the solutions of this latter one is larger (because here $59x - 4$ must be divisible only by 111 and not by 222). Also, Theorem 1.4.1 tells us that the number of solutions modulo 222 is $(59, 222) = 1$, so if we somehow obtain more possibilities, then only one of them can solve the original congruence. Note that this phenomenon occurs every time when we make a non-equivalent transformation at some of the steps. □

1.4.2 Simultaneous Congruences

In many applications of number theory we are faced with problems where many congruences must hold simultaneously. In the remaining part of the section we handle this problem. We start by solving two congruences at the same time.

Theorem 1.4.2. *The system of congruences $x \equiv a_1 \pmod{m_1}$ and $x \equiv a_2 \pmod{m_2}$ is solvable if and only if $(m_1, m_2) \mid a_1 - a_2$. If this condition holds, then solutions form a single residue class modulo $[m_1, m_2]$ (where $[m_1, m_2]$ is the least common multiple of m_1 and m_2).*

Proof. The system of congruences is solvable if and only if there is an x of the form $m_2y + a_2$ such that $m_2y + a_2 \equiv a_1 \pmod{m_1}$. This is equivalent to the solvability of the congruence $m_2y \equiv a_1 - a_2 \pmod{m_1}$. By Theorem 1.4.1 this is solvable if and only if $(m_1, m_2) \mid a_1 - a_2$.

Now assume that this latter condition holds, then the congruence $m_2y \equiv a_1 - a_2 \pmod{m_1}$ has (m_1, m_2) different solutions modulo m_1. If y_0 is a solution, then the other solutions modulo m_1 are $y_0 + km_1/d$, where $d = (m_1, m_2)$ and $0 \leq k \leq m_1 - 1$. This means that the solutions form exactly 1 residue class modulo m_1/d, so they are of the form $y_0 + km_1/d$, where $k \in \mathbb{Z}$. Then the solutions of the original system are of the form $m_2(y_0 + km_1/d) + a_2 = m_2y_0 + km_1m_2/d + a_2$, that is, they form a residue class modulo $m_1m_2/d = [m_1, m_2]$. This last equality is an easy consequence of Proposition 1.1.4. □

Corollary 1.4.3 (Chinese remainder theorem)**.** *Assume that $m_1, \ldots, m_k$ are pairwise coprime integers, then the system of congruences $x \equiv a_1 \pmod{m_1}$, $\ldots$, $x \equiv a_k \pmod{m_k}$ is solvable, and the solutions form a single residue class modulo $m_1 \ldots m_k$.*

Proof. We prove the statement by induction. For $k = 2$ this is a special case of the previous theorem (because $(m_1, m_2) = 1$). Assume that $k > 2$ and the statement is true for $k-1$. Then the system that consists of the first $k - 1$ congruences is equivalent to a single congruence $x \equiv a_0 \pmod{m_1 \ldots m_{k-1}}$. Together with $x \equiv a_k \pmod{m_k}$ this forms a system which is solvable by the previous theorem, and there is exactly 1 solution modulo $m_1 \ldots m_k$. Here we used that $(m_1 \ldots m_{k-1}, m_k) = 1$, this follows the same way like the analogous claim in the proof of Theorem 1.3.5. □

Exercise 1.4.2. Solve the following system of congruences:

$$x \equiv 11 \pmod{42} \qquad \text{and} \qquad x \equiv 10 \pmod{199}.$$

Solution. Since $(42, 199) = 1$, we get from the previous theorem that there is one single solution modulo $42 \cdot 199 = 8358$. By the first congruence we can write $x = 42y + 11$ for some integer y. Substituting this in the second congruence we get $42y + 11 \equiv 10 \pmod{199}$, that is, $42y \equiv -1 \equiv 198 \pmod{199}$. We can divide by 6 because $(6, 199) = 1$. We obtain $7y \equiv 33 \equiv 630 \pmod{199}$. Finally, dividing this by 7 we get $y \equiv 90 \pmod{199}$. Since we made the transformations of the congruences in every step so that the latter congruence was equivalent to the former one, we get that y must be of the form $199z + 90$. Then $x = 42y + 11 = 42(199z + 90) + 11 = 8358z + 3791$, i.e. $x \equiv 3791 \pmod{8358}$ is the only solution modulo 8358. □

1.5 Number-theoretic Algorithms

1.5.1 Efficiency of Algorithms

At the design of an algorithm one of the first questions which has to be dealt with is the expected running time of an implementation. This question is not always easy to answer, different running times are acceptable for different tasks. Sometimes every millisecond matters while in other cases the program can run for days. Of course the running time always depends on the hardware, but what is more important that in general a program runs longer for a bigger input. Here we regard the running time as a function of the size of the input.

As a first example we examine the following task which we call *prime factorization*: the input is an integer N and we are looking for its prime factorization. There is a simple method which gives the result: starting from 2 we try to divide N by every integer, and if we find a divisor p, then we continue the procedure for the number N/p (and it is enough to start searching from the number p). Note that every divisor that we find this way will be a prime number. When N is composite, then $N = ab$ for some $1 < a \leq b$, and hence $a^2 \leq ab = N$. This means that N has a divisor which is at most $\sqrt{N}$, so if we do not find a divisor until $\sqrt{N}$, then N is prime. This procedure clearly gives the expected result, it is easy to perform it for small numbers even without a calculator, but computers can determine the prime factorization this way for numbers with 10-20 digits. This may look satisfactory for the first sight, but in practice we often work with much larger numbers. For example if N has 81 digits in its decimal representation, then $N \geq 10^{80}$, so $\sqrt{N} \geq 10^{40}$. This means that if N is a prime, then our program makes at least 10^{40} divisions before it terminates. The fastest supercomputer today makes less than 10^{18} elementary floating point operations in one second, which means than it would take more time for that computer to run this algorithm than the age of the universe.

Of course this does not mean that it is impossible to give an algorithm for this task which has an acceptable running time - but unfortunately no one was able to find one yet. The situation changes a lot when we only want to decide if our number is prime. That is, the output here is "prime" or "composite", and we may have no information about the divisors in the latter case. We will learn about algorithms which solve this problem for numbers with several hundred digits in a reasonable time.

Now we try to describe what an "efficient" algorithm is. There is a definition which is more or less satisfactory both for theory and applications (leaving many questions unanswered though): we consider an algorithm efficient if it has *polynomial running time*.

Definition 1.5.1. For an algorithm **A** the size of its input is the number of bits that are used to store the input. The algorithm **A** is said to be *of polynomial (running) time* (or shortly: *polynomial*) if its running time (i.e. the number of steps of **A**) can be bounded from above by a polynomial of the size of its input, that is, if there exist a positive real number $c \in \mathbb{R}^+$ and a positive integer $k \in \mathbb{N}^+$ such that for every input of size $n \geq 1$ the algorithm **A** terminates after at most cn^k steps.

One may observe that the definition above is not precise from a mathematical point of view. First of all, it is not clear what we mean by a step of an algorithm (even the notion of algorithm is undefined). Also, the memory of a computer is not a mathematical object, so the size of an input is not accurately defined. For now, we work with this somewhat intuitive definition and leave the precise work for a later course. We will give an algorithm by a pseudocode or by a C programming code. We will also assume that executing a line of our code means a series of bit operations made by the processor of the computer and the number of these operations is called the number of steps then.

Let us return to our prime factorization algorithm. What can we say about its running time? Of course there are cases when the algorithm finds the prime divisors fast, for example when N is a power of 2. But a polynomial algorithm must run in polynomial time for every input. The size of the input is the number of digits of N written in the numeral system of base 2. This is exactly $n = \lfloor \log_2 N \rfloor + 1$, and hence $2^{n-1} \leq N$ holds. If N is a prime number, then our algorithm makes $\lfloor \sqrt{N} \rfloor$ divisions. Now

$$\lfloor \sqrt{N} \rfloor \geq \lfloor (\sqrt{2})^{n-1} \rfloor \geq (\sqrt{2})^{n-1} - 1 \geq 0.7 \cdot 1.4^n$$

if n is big enough (here we used that $\sqrt{2} > 1.4$ and $(\sqrt{2})^{-1} > 0.7$). That is, the number of steps can be bounded from below by an exponential function of the input size when N is a prime. Since there are infinitely many primes by Theorem 1.1.6, there are arbitrary large N's for which this bound holds. As an exponential function grows faster than any polynomial function, this algorithm cannot be polynomial.

This method will be applied many times when we show that an algorithm is not polynomial. Namely, in many cases one can give a lower bound for the number of steps in terms of the input size which grows faster than any polynomial. In these notes we will always use exponential lower bounds for this purpose, but of course in general there are cases when other type of functions are needed.

In this chapter the input of an algorithm is always a set of integers so the size of the input is the sum of the number of digits of these numbers (represented in the binary system). As we have already seen, for a single number N this is $\lfloor \log_2 N \rfloor + 1$. But since $\log_2 N = \log_2 10 \cdot \log_{10} N$, the notion of polynomial algorithm does not change if we regard the size of the input as the number of digits in the decimal representation. Moreover, this holds for a numeral system of any base, though we usually work with the binary or the decimal system. In short: an algorithm is polynomial in terms of the number of decimal digits if and only if it is polynomial in the sense of Definition 1.5.1.

As a final remark of this introductory section we mention that although from a theoretical point of view an algorithm with input size n and running time cn^k is polynomial and hence said to be "effective" for any $c \in \mathbb{R}^+$ and $k \in \mathbb{N}^+$, in practice the exponent is required to be small (e.g 1 or 2), otherwise the algorithm becomes too slow for the applications even for relatively small inputs.

Exercise 1.5.1. The following pseudocode gives an algorithm which computes the least common multiple of the numbers $a, b > 0$. Decide if it is polynomial or not.

```
1    x ← a
2    while b ∤ x do
3        x ← x + a
4    end while
5    print "The least common multiple=", x
```

1.5.2 Basic Arithmetic Operations

For those who have some experience with programming languages it may seem evident that the basic arithmetic operations are built into the languages or even into the hardwares. Still we elaborate on this topic, since there are algorithms behind these built-in operations as well, and what is more important, the number-theoretic algorithms of this chapter often have inputs with thousands of digits, and usually there are no built-in functions that treat such large numbers - we may have to write them.

By basic arithmetic operations we mean the addition, subtraction and multiplication of two numbers, and the division of them with a remainder. Fortunately we know effective algorithms for these tasks, namely the ones that we learned in elementary school, when we performed these operations by hand.

First we take a closer look at the addition. Assume that the number of digits of a and b are k and l, respectively where $k \geq l$ (we can assume this, because in the other case we can interchange a and b), so the size of the input is $k + l$. For simplicity we may write b as a k-digit number (writing $k - l$ zeros at the beginning of the number). Then we can carry out the addition in one loop. We go along the digits of the numbers from right to left and do the same operations in the body of the loop, namely we add the actual digits and the remainder carried over from the previous run of the body (this remainder is set to be zero at the beginning), fill the actual digit in the result and overwrite the remainder. This sum can be stored in a table (as the summands and the remainder are bounded), so the body of the loop only makes at most c bitwise operations for some constant c. We repeat the loop k times, so the procedure stops after at most $ck \leq cn$ steps. This means that the algorithm is polynomial, moreover, it is very efficient even among the polynomial algorithms, because the exponent of the input size in the bound is 1. The algorithms with this property are said to be of *linear running time*.

The subtraction can be implemented similarly and we also get a linear running time. After the previous example it is not hard to see that the multiplication can be accomplished via k multiplications of a number by a one-digit number and $k - 1$ additions. This yields at most $c(k + l)^2 = cn^2$ steps, so the multiplication is also polynomial (though this algorithm is not linear but quadratic). Note that there are faster algorithms for this task. These are more complicated and in practice one saves time only for large inputs, but in many applications, especially in cryptography they are useful. Historically the first one of these was Karatsuba's algorithm which terminates after at most $cn^{\log_2 3} \approx cn^{1.58}$ steps. There are even (asymptotically) faster methods, see e.g. the Toom-Cook algorithm or the Schönhage-Strassen algorithm. We just mention that the division can also be done in at most cn^2 steps by the usual algorithm that we use when calculating by hand. We leave the details to the reader.

1.5.3 Modular Exponentiation

It is easy to see that one cannot raise to powers in polynomial running time. Indeed, if the exponent is (not fixed, but) also part of the input, then even the number of the digits of the result is bigger than any polynomial of the input size, and hence the result cannot be written down in polynomial time. For example, if $a = 2$ and b has k decimal digits, then the number of digits of $a^b = 2^b$ is greater than

$$\log_{10} 2^b = b \cdot \log_{10} 2 \geq 10^{k-1} \cdot \log_{10} 2 > 0.03 \cdot 10^k.$$

That is, the number of digits of 2^b is bounded from below by an exponential function of the input size.

Although we cannot determine the power itself, for the RSA algorithm it is essential to calculate the remainder of powers modulo m. This task is called *modular exponentiation*. Here inputs are positive integers $a, b, m \in \mathbb{N}^+$, and the output is the remainder of a^b modulo m. Then at least the output size cannot be an exponential function of the input size, since the output is smaller than m. As we cannot calculate a^b and then divide it by m, we could try to calculate first the remainder of a and then multiply it by a and calculate the remainder of a^2, and continuing this way, we could obtain the remainder of the numbers $a^3, a^4, \ldots, a^b$. This requires the calculation of b remainders, which means that the number of steps is still exponential.

Instead of this we apply *repeated squaring*, that is, we only determine the remainders of the powers a^{2^k} for some k's, and then multiply some of them. Before the precise description we show an example. We calculate the remainder of 13^{53} modulo 97. In the following calculation we always square the congruence in the previous row:

$$\begin{aligned}
13^1 &\equiv 13 && \pmod{97},\\
13^2 &\equiv 169 \equiv 72 && \pmod{97},\\
13^4 = (13^2)^2 &\equiv 72^2 = 5184 \equiv 43 && \pmod{97},\\
13^8 = (13^4)^2 &\equiv 43^2 = 1849 \equiv 6 && \pmod{97},\\
13^{16} = (13^8)^2 &\equiv 6^2 = 36 && \pmod{97},\\
13^{32} = (13^{16})^2 &\equiv 36^2 = 1296 \equiv 35 && \pmod{97}.
\end{aligned} \tag{1}$$

It is not necessary to continue, since 13^{64} is already greater than 13^{53}. Now we use the binary representation of the exponent $53 = 110101_2 = 32 + 16 + 4 + 1$, and we write the original power as a product: $13^{53} = 13^1 \cdot 13^4 \cdot 13^{16} \cdot 13^{32}$. Multiplying the remainders of these factors we get the remainder of the original power:

$$\begin{aligned}
13^5 = 13^1 \cdot 13^4 &\equiv 13 \cdot 43 = 559 \equiv 74 && \pmod{97},\\
13^{21} = 13^5 \cdot 13^{16} &\equiv 74 \cdot 36 = 2664 \equiv 45 && \pmod{97},\\
13^{53} = 13^{21} \cdot 13^{32} &\equiv 45 \cdot 35 = 1575 \equiv 23 && \pmod{97}.
\end{aligned}$$

Note that we could do the last 3 steps in parallel with the determination of the remainders of the powers in (1), and then it is not necessary to store these remainders.

Now we describe the algorithm for the general positive integers a, b, m. As a first step we may calculate the remainder of a modulo m, so it is enough to give the algorithm for integers $0 < a < m$, $0 < b$ (if the remainder is 0, then so is the remainder of the power). We store the remainder of a product modulo m, which is set to be 1 initially (the value of the

empty product). We calculate the remainders of a^{2^k} for the exponents $k = 0, 1, \ldots, \lfloor \log_2 b \rfloor$. In parallel we multiply the stored product by the actual remainder modulo m if b has a digit 1 in the $(k+1)$th place (from the right) in its binary representation. We do not assume that the binary representation of b is given, it will be determined along the way. We use that the first digit of the binary form of a number b is its remainder modulo 2 and the other digits form the binary representation of $\lfloor \frac{b}{2} \rfloor$. So the algorithm is the following:

MODULAR EXPONENTIATION

Input: the positive integers a, b and m with $0 < a < m$, $0 < b$

```
1     c ← 1
2     while true do
3         if b is odd, then
4             c ← c · a mod m
5         b ← ⌊b/2⌋
6         if b = 0, then
7             print "a^b mod m=",c; stop
8         a ← a^2 mod m
9     end while
```

Now we show that this algorithm gives us the right result. Let a_0 and b_0 denote the numbers a and b, respectively, and we also set $c_0 = 1$. For a positive integer $k > 0$ let a_k, b_k and c_k be the value of the variables a, b and c, respectively after the kth run of the body of the loop. The remainder of a^b modulo m will be denoted by r. We are going to show by induction that $a_k^{b_k} c_k \equiv r \pmod{m}$ holds for every $k \in \mathbb{N}$. This is obviously true for $k = 0$. Now assume that $k > 0$ and that the congruence holds for $k - 1$. If b_{k-1} is even, then

$$r \equiv a_{k-1}^{b_{k-1}} c_{k-1} = (a_{k-1}^2)^{\frac{b_{k-1}}{2}} c_{k-1} \equiv a_k^{b_k} c_k \qquad (\text{mod } m)$$

as $b_k = b_{k-1}/2$ and $c_k = c_{k-1}$ in this case, and $a_k \equiv a_{k-1}^2 \pmod{m}$ holds independently from the parity of b_{k-1}. On the other hand, if b_{k-1} is odd, then $b_k = (b_{k-1}-1)/2$ and $c_k = a_{k-1}c_{k-1}$, hence

$$r \equiv a_{k-1}^{b_{k-1}} c_{k-1} = (a_{k-1}^2)^{\frac{b_{k-1}-1}{2}} a_{k-1} c_{k-1} \equiv a_k^{b_k} c_k \qquad (\text{mod } m).$$

The algorithm stops in the kth loop when $b_k = 0$, then $b_{k-1} = 1$, and the output is c_k. But we have just proved that $r \equiv a_{k-1}^{b_{k-1}} c_{k-1} \pmod{m}$ holds, moreover, $a_{k-1}^{b_{k-1}} c_{k-1} = a_{k-1} c_{k-1} = c_k$, so we are done.

Finally, we prove that the algorithm is polynomial. Let j, k and l denote the number of digits of a, b and m, respectively. Then the size of the input is $n = j + k + l$. In the body of the loop we do at most 2 multiplications with inputs less than m, 2 divisions with remainder with inputs less than m^2, and we calculate the half of a number which is at most b. So if we use the algorithms of the previous section (in the following we will always do so), then it follows that in the body of the loop we make at most $c_1(l^2 + k)$ steps for some constant c_1, and it runs $\lfloor \log_2 b \rfloor + 1$ times. This latter number is the number of the digits in the binary representation of b, hence it is at most $c_2 k$ for some constant c_2. Then the number of steps if at most $c_1 c_2 (l^2 + k)k \leq c_1 c_2 n^3$. Hence this is indeed a polynomial algorithm, but it is important to note that is is still too slow for the applications in cryptography. There are faster variants of this method, but we do not give any details here.

1.5.4 The Calculation of the Greatest Common Divisor

Proposition 1.1.4 gives us a formula for the greatest common divisor using the canonical representation of the numbers. As we have seen earlier, is hard to determine the prime factorization in general, so this formula is not applicable in practice. Luckily, there is a much more effective method for this task: the so-called *Euclidean algorithm*. It is contained in the book "Elements" which was written by the ancient Greek mathematician Euclid ca. 300 BC.

In this task the input consists of the numbers a and m, and we can assume that $0 < a < m$ holds (the cases when $a = 0$ or $a = m$ can be handled easily). To determine (a, m) we are going to make repeated divisions with remainders: first we divide m by a, then in the next step we divide a by the remainder taken from the first step, and in the ith step we divide the remainder from the $(i-2)$th step by the remainder from the $(i-1)$th step. We stop when we get 0 as a remainder, and the output will be the last non-zero remainder (or a if we stop in the first step). First we show an example: we calculate the greatest common divisor of 567 and 1238.

$$\begin{aligned}
&(1) && 1238 = 2 \cdot 567 + 104,\\
&(2) && 567 = 5 \cdot 104 + 47,\\
&(3) && 104 = 2 \cdot 47 + 10,\\
&(4) && 47 = 4 \cdot 10 + 7,\\
&(5) && 10 = 1 \cdot 7 + 3,\\
&(6) && 7 = 2 \cdot 3 + 1,\\
&(7) && 3 = 3 \cdot 1 + 0.
\end{aligned}$$

The result is the last non-zero remainder, that is $(567, 1238) = 1$. Now we write the previous steps with a general m and a (assuming that $a \nmid m$):

$$\begin{aligned}
&(1) && m = q_1 a + r_1 && (0 < r_1 < a),\\
&(2) && a = q_2 r_1 + r_2 && (0 < r_2 < r_1),\\
&(3) && r_1 = q_3 r_2 + r_3 && (0 < r_3 < r_2),\\
&(4) && r_2 = q_4 r_3 + r_4 && (0 < r_4 < r_3),\\
&\vdots && \vdots && \vdots\\
&(k) && r_{k-2} = q_k r_{k-1} + r_k && (0 < r_k < r_{k-1}),\\
&(k+1) && r_{k-1} = q_{k+1} r_k + 0, &&
\end{aligned}$$

Here the output is r_k (i.e. the last non-zero remainder).

Proposition 1.5.1. *The output of the previous algorithm is* (a, m).

Proof. The statement holds obviously when $a \mid m$, so we assume that this is not the case. By the first step we have that $m \equiv r_1 \pmod{a}$, and hence $(m, a) = (r_1, a)$ by Proposition 1.3.1. The second step gives similarly that $(a, r_1) = (r_2, r_1)$. Continuing this way we get

$$(m, a) = (a, r_1) = (r_1, r_2) = \cdots = (r_{k-1}, r_k) = (r_k, 0) = r_k,$$

and this proves the claim. □

For the computation of r_i we only have to store r_{i-1} and r_{i-1}, which makes the implementation simpler. Here is a pseudocode for the algorithm:

EUCLIDEAN ALGORITHM

Input: the positive integers a and m with $0 < a < m$

```
1    while true do
2        r ← m mod a
3        if r = 0, then
4            print "(a, m)=",a; stop
5        m ← a; a ← r
6    end while
```

In every step the remainder is always less than the number we divide by. That is, we have $a > r_1 > r_2 > \ldots$, hence the algorithm terminates after at most a loops. However, this does not show that the number of steps is polynomial, since a can be exponential in terms of the input size (note that this is not always the case, if a is small and m is big enough, then the last statement does not hold, but the point is that there is such a case when the size of a is comparable to the size of the input). But in fact the sequence of the remainders decreases faster, namely the following statement hods:

Proposition 1.5.2. *The Euclidean algorithm stops after a most* $2\lceil \log_2 a \rceil$ *loops.*

Proof. For making the notation simpler, we set $r_{-1} = m$ and $r_0 = a$. Then in every loop we make a division with a remainder: $r_{i-2} = t_i r_{i-1} + r_i$, where $r_{i-2} > r_{i-1} > r_i$. Since $r_{i-2} > r_{i-1}$ holds, we must have $t_i \geq 1$ (because t_i is a non-negative integer). It follows that $r_{i-2} \geq r_{i-1} + r_i > 2r_i$. If the algorithm does not stop after the $2k$th step, then we have

$$a = r_0 > 2r_2 > 4r_4 > \cdots > 2^k r_{2k} \geq 2^k \cdot 1,$$

i.e. $k < \log_2 a$. In other words, it is impossible that we do not stop after $2\lceil \log_2 a \rceil$ steps. □

After this preparation we are in the position to show that the Euclidean algorithm is polynomial. If m has k digits and a has l digits, then the size of the input is $n = k + l$. Since $a < m$ and hence every $r_i < m$, we perform every division on numbers with at most k digits, so we make at most $c_1 k^2$ steps in the body of the loop. By the previous proposition we run the loop at most $2\lceil \log_2 a \rceil \leq c_2 k$ times, so the total number of steps is at most $ck^3 \leq cn^3$.

As a final remark we note that the least common multiple can also be determined in polynomial time using the formula $(a, m) \cdot [a, m] = am$.

1.5.5 Solution of Linear Congruences

By Theorem 1.4.1 we know that the linear congruence $ax \equiv b \pmod{m}$ is solvable if and only if $d = (a, m) \mid b$, and that the number of solutions modulo m is d. Hence we can use the Euclidean algorithm to decide if a congruence is solvable. We also determined all the solutions using an arbitrary one, so it remains to find the first solution. By Theorem 1.2.3 we only have to solve the equivalent linear congruence $a'x \equiv b' \pmod{m'}$, where $a' = a/d$, $b' = b/d$ and $m' = m/d$, and here a' and m' are co-prime. So in this section we are going to assume that $(a, m) = 1$.

It turns out that a modification of the Euclidean algorithm can be used for solving a linear congruence. We illustrate this on an example, we are going to solve the congruence $567x \equiv 123 \pmod{1238}$. First we write down the congruence $1238x \equiv 0 \pmod{1238}$ which is obviously true for all integers and will be denoted by (A). After that we write down the original congruence $567x \equiv 123 \pmod{1238}$, this will be denoted by (I) (expressing that this is the input). Now we repeat the steps of the Euclidean algorithm, we divide 1238 by 567 with a remainder, and subtract from (A) the congruence (I) multiplied by the quotient. This way we get a linear congruence, where the coefficient of x is the remainder. We calculate the smallest remainder modulo 1238 on the right hand side to keep the numbers bounded during the process. Continuing this way, after the 6th step the coefficient of x will be the greatest common divisor, namely 1.

$$
\begin{array}{rrll}
(A) & 1238x \equiv 0 & \pmod{1238} & \\
(I) & 567x \equiv 123 & \pmod{1238} & \\
(A) - 2 \cdot (I) : (1) & 104x \equiv -246 \equiv 992 & \pmod{1238} & [1238 = 2 \cdot 567 + 104] \\
(I) - 5 \cdot (1) : (2) & 47x \equiv -4837 \equiv 115 & \pmod{1238} & [567 = 5 \cdot 104 + 47] \\
(1) - 2 \cdot (2) : (3) & 10x \equiv 762 & \pmod{1238} & [104 = 2 \cdot 47 + 10] \\
(2) - 4 \cdot (3) : (4) & 7x \equiv -2933 \equiv 781 & \pmod{1238} & [47 = 4 \cdot 10 + 7] \\
(3) - 1 \cdot (4) : (5) & 3x \equiv -19 \equiv 1219 & \pmod{1238} & [10 = 1 \cdot 7 + 3] \\
(4) - 2 \cdot (5) : (6) & x \equiv -1657 \equiv 819 & \pmod{1238} & [7 = 2 \cdot 3 + 1] \\
 & & & [3 = 3 \cdot 1 + 0]
\end{array}
$$

Now we repeat this for the general congruence $ax \equiv b \pmod{m}$, where $(a, m) = 1$, and add the congruence $mx \equiv 0 \pmod{m}$ in the beginning (which is true for every $m \neq 0$). It is clear, that in this case the last non-zero remainder in the Euclidean algorithm is $r_k = 1$ for some k. Observe, that because of this we get a congruence of the form $x \equiv c_k \pmod{m}$ in the kth step, which gives us the solution of the original linear congruence.

$$
\begin{array}{rrll}
(A) & mx \equiv 0 & \pmod{m} & \\
(I) & ax \equiv b & \pmod{m} & \\
(A) - q_1 \cdot (I) : (1) & r_1 x \equiv -q_1 b \equiv c_1 & \pmod{m} & [m = q_1 a + r_1] \\
(I) - q_2 \cdot (1) : (2) & r_2 x \equiv b - q_2 c_1 \equiv c_2 & \pmod{m} & [a = q_2 r_1 + r_2] \\
(1) - q_3 \cdot (2) : (3) & r_3 x \equiv c_1 - q_3 c_2 \equiv c_3 & \pmod{m} & [r_1 = q_3 r_2 + r_3] \\
(2) - q_4 \cdot (3) : (4) & r_4 x \equiv c_2 - q_4 c_3 \equiv c_4 & \pmod{m} & [r_2 = q_4 r_3 + r_4] \\
\vdots & \vdots & & \vdots \\
(k) & x \equiv c_{k-2} - q_k c_{k-1} \equiv c_k & \pmod{m} & [r_{k-2} = q_k r_{k-1} + r_k = q_k r_{k-1} + 1] \\
 & & & [r_{k-1} = q_{k+1} r_k + 0]
\end{array}
$$

In every step we get a congruence which follows from the previous ones and hence from $ax \equiv b \pmod{m}$. Thus, if x_0 is a solution of this congruence, then $x_0 \equiv c_k \pmod{m}$ must hold, i.e. we get the modulo m unique solution. Note though that the congruences that we obtain during the process are not necessarily equivalent to the original one. For example the congruence $104x \equiv 992 \pmod{1238}$ in the example above has 2 different solutions modulo

1238 (while the original congruence has only 1). Also, in the ith step we compute the least positive remainder modulo m (this is denoted by c_i) keeping the occurring numbers bounded by m.

By Proposition 1.5.2 we get to the last congruence after at most $2\lceil \log_2 a \rceil$ divisions. Note that after determining the remainders of a and b modulo m we can reach $0 < a < m$ and $0 < b < m$, and then every further number that occurs during the process is bounded by cm^2 for some constant c. Using all this it is not hard to see that the algorithm is polynomial, the details are left to the reader.

We give this algorithm also by a pseudocode:

EUCLIDEAN ALGORITHM FOR THE SOLUTION OF A CONGRUENCE

Task: solution of the congruence $ax \equiv b \pmod{m}$ with $(a, m) = 1$

Input: the positive integers a, b and m with $0 < a, b < m$

```
1    M ← m; c ← 0; d ← b
2    while true do
3        q ← ⌊m/a⌋; r ← m mod a
4        if r = 0, then
5            print "The solution is x ≡", d (mod M); stop
6        t ← c − qd mod M
7        m ← a; a ← r; c ← d; d ← t
8    end while
```

The variables m and a store the previous remainders (and initially m and a, respectively), while c and d store the right hand sides of the previous two congruences (which are 0 and b at the beginning, respectively). We also store the value of m in the variable M, because we need it in every loop (in line 5 and 6). Finally, we make the code readable by storing the quotient in the variable q since we need it in line 6 (though this is not necessary).

1.5.6 Primality tests

We have mentioned in Section 1.5.1 that even though we do not know an efficient algorithm which determines the prime factorization of a number, we can still decide if a number is prime or not. It is maybe surprising that this creates a situation which makes the application of some cryptographic techniques possible, as we will see in the next section.

The Fermat test

One of the simplest primality tests is based on the Euler-Fermat theorem (Theorem 1.3.7): if m is prime and $1 \leq a \leq m-1$ is an integer, then $\varphi(m) = m - 1$ and $a^{m-1} \equiv 1 \pmod{m}$. This means that if we are able to find an a such that $a^{m-1} \not\equiv 1 \pmod{m}$, then m is not prime. The so-called *Fermat test* works the following way: it generates numbers between 1 and $m-1$ randomly and computes the remainder of a^{m-1} modulo m. If this remainder is not 1, then either $(a, m) > 1$ holds, or $(a, m) = 1$ but $\varphi(m) \neq m - 1$. No matter which case applies, m cannot be prime. Note that we can calculate (a, m) fast, so if we are lucky enough to have the former case, even a divisor of m can be determined.

Of course it can happen that we pick an a such that $a^{m-1} \equiv 1 \pmod{m}$ even for a composite modulus (and then $(a, m) = 1$). If m is composite, then such an a is called a

Fermat liar. On the other hand, if $(a, m) = 1$ and $a^{m-1} \not\equiv 1 \pmod{m}$, then a is called a *Fermat witness* for the compositeness of m. So if a is a liar, then we may repeat the test several times and hope for finding a witness. It is still not obvious that we find a witness with high probability, but the following theorem assures this if there exists at least one witness:

Theorem 1.5.3. *If $m \in \mathbb{N}^+$ is composite and it has a Fermat witness (i.e. a number a between 1 and m which is co-prime to m and for which $a^{m-1} \not\equiv 1 \pmod{m}$ holds), then at least half of the numbers co-prime to m between 1 and m are witnesses.*

Proof. Let a be a witness of m and assume that $c_1, \ldots, c_k$ are all the liars of m between 1 and m (that is, $(c_i, m) = 1$ and $c_i^{m-1} \equiv 1 \pmod{m}$ for every i). Let a_i be the least positive number for which $a_i \equiv ac_i \pmod{m}$ holds. We show that $a_1, \ldots, a_k$ are pairwise different witnesses of m, and hence the number of witnesses between 1 and m is at least the number of liars in this interval. Since all the c_i's are co-primes to m, the statement follows.

First observe that since $(a, m) = 1$ and $(c_i, m) = 1$ for every $1 \leq i \leq k$, it follows from the fundamental theorem of arithmetic that $(ac_i, m) = 1$. Then $ac_i \equiv a_i \pmod{m}$ and Proposition 1.3.1 implies that $(a_i, m) = (ac_i, m) = 1$, that is, every a_i is co-prime to m. Moreover, if we raise the congruence $a_i \equiv ac_i \pmod{m}$ to the $(m-1)$th power, then we get

$$a_i^{m-1} \equiv (ac_i)^{m-1} = a^{m-1}c_i^{m-1} \equiv a^{m-1} \not\equiv 1 \qquad (\text{mod } m),$$

since c_i is a liar and a is a witness. That is, we have proved that a_i is a witness for every $1 \leq i \leq k$.

It is left to show that the numbers $a_1, \ldots, a_k$ are pairwise different. So assume that $a_i = a_j$ for some $1 \leq i, j \leq k$, and then $ac_i \equiv ac_j \pmod{m}$. Dividing both sides by a we get that $c_i \equiv c_j \pmod{m}$, where the modulus does not change because $(a, m) = 1$. Since $1 \leq c_i, c_j \leq m$ holds, we must have $c_i = c_j$. But the c_i's are pairwise different, so $i = j$ follows. □

Assume that we give the output "m is prime" if after 100 tests we do not find a witness. If m is composite and it has a witness, then we go wrong with probability at most 2^{-100}. Although this number is positive, it is so small, that it is negligible in practice. But there is a bigger problem: there are numbers which do not have any witnesses.

Definition 1.5.2. The positive integer $m \in \mathbb{N}^+$ is called a *Carmichael number* if it is composite and for every integer $a \in \mathbb{N}^+$ with $1 \leq a \leq m$ and $(a, m) = 1$ the congruence $a^{m-1} \equiv 1 \pmod{m}$ holds.

If we run the test for a Carmichael number 100 times, then we get that m is prime with very high probability. We get the output "m is composite" only if we pick a proper divisor of m at least once out of 100 tries, but this is very unlikely. And even though the Carmichael numbers are rare (the smallest one is 561, the next one is 1105, and there are only 43 below one million), unfortunately there are infinitely many of them (see [2]).

There are modifications of this method that solve this problem, among them the most popular is the so-called Miller-Rabin test (see below). We also note that there exists a primality test with polynomial running time which does not use randomness (it is *deterministic*, i.e. it always gives the right result). This was shown in [1] by Agrawal, Kayal and Saxena in 2002. However, their method is too slow for applications and hence it is not used in practice.

The Miller-Rabin test

The Miller-Rabin test is similar to the Fermat test in structure, only a few modifications are needed. The criterion $a^{m-1} \equiv 1 \pmod{m}$ will be substituted by a stricter one. We will use the following simple observation:

Proposition 1.5.4. *Assume that m is prime, then $x^2 \equiv 1$ (mod m) holds if and only if $x \equiv \pm 1 \pmod{m}$.*

Proof. If $x \equiv \pm 1 \pmod{m}$, then squaring this congruence we have $x^2 \equiv 1 \pmod{m}$, and this is true for every m.

Now assume that m is prime and $x^2 \equiv 1 \pmod{m}$ holds, i.e. $m \mid x^2 - 1 = (x-1)(x+1)$. Then by the fundamental theorem of arithmetic we must have $m \mid x-1$ or $m \mid x+1$, which is equivalent to $x \equiv \pm 1 \pmod{m}$. □

In the following we may assume that $m > 2$ is odd (and then $m-1$ is even), otherwise m is composite. In the test we choose an integer a in the interval $[1, m]$ which is co-prime to m, and check if $a^{m-1} \equiv 1 \pmod{m}$ holds. If not, then m cannot be prime. But unlike in the Fermat test, now we do not say that a is a liar right away in the other case. Instead, we check if the congruence $a^{\frac{m-1}{2}} \equiv \pm 1 \pmod{m}$ holds. If this is not true, then by the previous proposition we get that m is composite. Now if $\frac{m-1}{2}$ is odd or if $a^{\frac{m-1}{2}} \equiv -1 \pmod{m}$, then we choose another a and start the test from the beginning. However, if $\frac{m-1}{2}$ is even and $a^{\frac{m-1}{2}} \equiv 1 \pmod{m}$, then the previous proposition gives that $a^{\frac{m-1}{4}} \equiv \pm 1 \pmod{m}$ must also hold. If not, then we get that m is composite. Otherwise we say that a is a liar if the exponent $\frac{m-1}{4}$ is odd or the remainder is -1. From here we continue the same way with the exponents $\frac{m-1}{8}, \frac{m-1}{16}, \ldots$ until we get an odd exponent or a remainder different from 1. If this remainder is also different from -1, then m is composite, otherwise we choose another a.

An integer a co-prime to m is called a *Miller-Rabin witness* if choosing a in the test above we conclude that m is composite. Observe that if a is a Fermat witness, then of course it is automatically a Miller-Rabin witness, since the first step of the test is the same. Then it follows immediately from Theorem 1.5.3 that if m is composite and it is not a Carmichael number, then at least the half of the numbers co-prime to m in the interval $[1, m]$ are Miller-Rabin witnesses. The advantage of this method is that there are no Carmichael number-type exceptions here, moreover, we can be sure that there are even more witnesses than in the case of the Fermat test. Namely, the following is true:

Theorem 1.5.5. *If $m > 4$ is an integer, then at least three-quarters of the integers in the interval $[1, m-1]$ are either Miller-Rabin witnesses or not co-prime to m.*

The proof of this theorem can be found for example in [9]. We also note that it is conjectured that the least witness is relatively small. More precisely, if the so-called Extended Riemann Hypothesis is true, then we can find an integer $1 \leq a \leq 2(\ln m)^2$ so that either $(a, m) > 1$ holds or a is a Miller-Rabin witness of m. If this was true, then it would also give a deterministic polynomial algorithm for this task, because it would be enough to run the test for the least $2(\ln m)^2$ positive integers. For the details see [3].

Generation of primes

Finally, we say a few words about the generation of prime numbers. This is important because big primes play a crucial role in public key cryptography, as we will see in the

next section. A simple method for this task is that we generate numbers randomly and use a primality test to check if they are primes. It can be shown that there are enough primes among the integers so that this algorithm finds a prime number within a reasonable time. However, this requires advanced techniques, but some basic theorems are proved about the number of primes for example in [5]. This book is recommended for the interested reader because it uses only limited tools and contains the "elementary" proof of the following theorem (elementary means that it does not use the theory of analytic functions, but the argument is involved nonetheless). For a positive number x let $\pi(x)$ denote the number of positive primes that are at most x. For example $\pi(5) = 3$, $\pi(10) = 4$.

Theorem 1.5.6 (Prime number theorem)**.**

$$\lim_{x\to\infty} \frac{\pi(x)}{\frac{x}{\ln x}} = 1.$$

Roughly speaking: $\pi(n) \approx \frac{n}{\ln n}$, i.e. among the positive integers below n every $(\ln n)$th number is prime. This statement is very far from being precise, but we do not give further details here.

1.5.7 Public Key Cryptography

The method that is described in this section is based on the following: if p and q are big primes (e.g. with 300 digits) and their product $N = pq$ is public (but p and q are not), then no one is able to calculate the factors p and q within a reasonable time.

One of the main tasks of cryptography is to give methods that assure secure communication. Security means basically the following: a message sent between some participants must be readable for them but should be left hidden for anyone else who is able to read (some part of) the information that goes through the channel which connects the participants. In other words, the sender has to encrypt the message so that only the receiver can decrypt it. Encrypting and decrypting a message means nothing else than the application of two functions that are inverses to each other: the message x is encrypted with the function E giving the data $E(x)$ which is decrypted by the inverse function D of E, that is, $D(E(x)) = x$. Our goal is to find the appropriate functions E and D.

In a version of this method the functions E and D are kept secret, only the participants know them. A drawback of this that in this case they have to share these functions with each other, and many times this is inconvenient if not impossible. Public key cryptography solves this problem the following way: the function E is made public while the function D is kept secret. If for example A wants to send a message to B, then A can encrypt it with the encryption function E_B that is made public by B, while B decrypts it with the function D_B (which is known only for B). On the other hand, if B wants to answer this message, then the function E_A is used (the one that is made public by A), and A decrypts the answer with the functions D_A (that is kept secret by A).

How is it possible that the function E is known, but one cannot determine its inverse $D = E^{-1}$? The situation is similar to the following example: given a text in German and an English-German dictionary. Theoretically it is possible to translate the text with the help of this dictionary, but it is quite tedious work the find all the German words, simply because it is ordered alphabetically by the letters of the English words. Returning to the functions: the domain of E (and D) will be a set of integers $\{0, 1, \ldots, N-1\}$ for some N. The number N can be chosen so big (e.g. bigger than 10^{600}), that it takes billions of years

even for supercomputers to go through its elements and calculate all the function values. So the formula for E can be made public as long as the values of E and D can be calculated easily while it is practically impossible to determine a formula for the inverse function D from the public formula of E.

It is not obvious to give such functions that work. We describe an example below: the RSA algorithm, which was invented by Rivest, Shamir and Adleman. For this we need the following

Proposition 1.5.7. *If p and q are distinct positive primes and $N = pq$ then for every $x \in \mathbb{Z}$ integer and $k \in \mathbb{N}^+$ positive integer we have $x^{k\varphi(N)+1} \equiv x \pmod{N}$.*

Proof. This follows easily from the Euler-Fermat theorem (Theorem 1.3.7) in the case when x is co-prime to N, since then we have $x^{\varphi(N)} \equiv 1 \pmod{N}$, so raising this congruence to the kth power and multiplying by x we get the statement.

If $(x, N) \neq 1$, then $p \mid x$ or $q \mid x$. If both divisions hold, then $N \mid x$ and hence $x^{k\varphi(N)+1} \equiv 0 \equiv x \pmod{N}$. So assume for example that $p \nmid x$ and $q \mid x$ (in the other case the proof is practically the same). Then $(p, x) = 1$, since p is a prime, and by the Euler-Fermat theorem we get that $x^{\varphi(p)} = x^{p-1} \equiv 1 \pmod{p}$. We raise this congruence to the $k(q-1)$th power and multiply both sides by x. Using that $\varphi(N) = (p-1)(q-1)$ we get that $x^{k\varphi(N)+1} = x^{k(p-1)(q-1)+1} \equiv x \pmod{p}$. But since $q \mid x$, this congruence holds modulo q as well. Finally, as $p \mid x^{k\varphi(N)+1} - x$ and $q \mid x^{k\varphi(N)+1} - x$ hold, we obtain that $pq = N \mid x^{k\varphi(N)+1} - x$ because p and q are distinct primes. □

As the first step of the RSA algorithm we choose two primes p and q with (for example) 300 digits. We set $N = pq$ and choose also a c with $(c, \varphi(N)) = 1$. Then the encryption function will be the following: $E(x) = x^c \bmod N$. The values of E can easily be calculated with repeated squaring. Moreover, it turns out that the inverse D of E is of the same form: $D(y) = y^d \bmod N$ for some integer d. As D is the inverse of E, we must have

$$x = D(E(x)) \equiv E(x)^d \equiv x^{cd} \qquad (\text{mod } N)$$

for every $0 \leq x \leq N-1$. By the previous proposition it is enough to find a d for which $cd = k\varphi(N) + 1$ for some positive integer k. In other words, we have to solve the congruence $cx \equiv 1 \pmod{\varphi(N)}$. This congruence is solvable since $(c, \varphi(N)) = 1$ by the choice of c. The solution can be calculated efficiently with the methods described in the previous sections.

Observe that we need the value of $\varphi(N) = (p-1)(q-1)$ to determine the value of d, and for this we need the prime factorization of N. Hence if we keep p, q, d and $\varphi(N)$ secret, then we can make c and N public, and one cannot determine the function D, at least by the method described above.

It could happen that someone determines the function D in some other form (based on c and N only), or that someone finds an efficient algorithm for the factorization of integers. However, this seems very unlikely, and the experience of decades shows that this method is safe. But many problems must be handled by the implementation of the algorithm. For example, the system can be attacked by analyzing the time of the encryption and decryption or by measuring the energy consumption of the computer. Also, the number N and c must be chosen carefully. Just to mention one difficulty: there are such numbers that can be factorized easier than a general N. But we do not address these questions, these topics are beyond the scope of these notes.

2 Linear Algebra

Linear algebra is undoubtedly one of the most important branches of mathematics. It is hard to give an exhausting list of its applications inside and outside mathematics. It provides a basic tool also in computer science, and from the countless problems whose solutions require the usage of this theory we address the solution of systems of equations in this chapter to give an introduction to this topic. We also get a glimpse of the connection of linear algebra with geometry through linear transformations. Finally, we are going to introduce the notions of eigenvectors and eigenvalues. All these play a crucial role for example in image processing and visualization, just to mention two evident examples.

We emphasize that in these notes we introduce only one basic example of a more general algebraic structure, although evidently the most important one. But while for many applications this suffices, there are plenty of them which require deeper understanding of the theory. There are countless books and notes in this topic, and many of them build upon some knowledge in abstract algebra. Instead of that we give here a more elementary introduction and recommend some other books for the interested reader. It is important to note that an abstract notion can be understood much easier via examples which makes our approach advantageous. For further reading we recommend the book [7] which still concentrates on important special cases of the general theory. For an introduction to abstract algebra we recommend for example the book [4].

2.1 Analytic Geometry in the Space

Analytic geometry in the plane can be familiar from high school. It uses algebraic tools to handle geometric objects: points, lines and different curves. This introductory section extends this theory to the space, where we use triples instead of pairs to describe the points. We restrict ourselves to the description of lines and planes, one can read more in [7] about methods which allow us to handle other surfaces.

2.1.1 The Coordinate System

On the plane one fixes two orthogonal lines, a positive direction and a unit on each of the lines to obtain a unique representation of every point. One can extend this method to the space where we fix three pairwise orthogonal lines - the axes x, y and z - which intersect in one point and determine the point of origin O this way. We also fix a point different from O on each axis and these points determine three segments whose other endpoint is the origin and also three directions from the origin towards the selected points. For simplicity we may choose a unit segment so that length and distance can be measured in the space, and in this section we assume that each of the three segments above has length 1. In other words, we fix three unit vectors on the axes which together with O form a *coordinate system*. Note that we can still choose their directions on the lines. Then every point P of the space determines uniquely a (maybe degenerate) rectangular cuboid whose edges are parallel to the three unit vectors and the section OP is its diagonal (by a degenerate cuboid we mean that its vertices are co-planar). As the *directed* units are fixed on each of the three axes, we can measure the *signed* length of the edges of the cuboid and obtain the *coordinate triple* (x_0, y_0, z_0) for the point P (so that x_0, y_0 and z_0 are the signed length of the edges parallel to the lines x, y and z, respectively). Note that this is a one-to-one correspondence between the points of the space and the ordered triples of real numbers.

The coordinate system can be oriented in two different ways, right or left. It is said to be *right-oriented* if once the right thumb points along the z axis in the positive direction, then the right index finger points along the x axis and the middle finger points along the y axis, both of them in the positive direction. Otherwise the coordinate system is called *left-oriented*. Note that we usually use right-oriented systems.

We used the word "vector" when we fixed the units on the axes. Now we give its precise meaning: similarly as on the plane, by a *space vector* we mean a directed line segment in the space so that any two segments with the same length and direction are considered to be the same vector. If the initial point and the endpoint of the segment coincide, then we get the *zero vector* whose direction is not determined (but its length is 0). If a coordinate system is fixed, then any vector can be given by its coordinates, that is, by the coordinates of the point which is the endpoint of the representative whose initial point is the origin. Such a representative is called a *position vector*. Vectors are usually denoted by underlined lower case letters or by the triples of their coordinates, e.g. $\underline{v} = (7, 2, 3)$ denotes a space vector. We also use another notation: if a vector (or more precisely a representative of it) points from A to B where A and B are some points of the space, then this vector can be denoted by $\overrightarrow{AB}$.

We have seen that once a coordinate system is chosen, the set of space vectors can be identified with the ordered triples of real numbers. By ordered we mean that the order of numbers is fixed (but they are not necessarily ordered by magnitude). The set of triples is denoted by $\mathbb{R}^3$. Here we mention that in analytic geometry we usually write the triples in a row (at least we do so in this section), while in later sections the elements of $\mathbb{R}^3$ will be written in a column. Hopefully this will not be confusing in the following.

Vector operations

We can define the addition, subtraction and scalar multiplication of vectors like one does in the case of plane vectors. If $\underline{u}$ and $\underline{v}$ are space vectors, then $\underline{u}+\underline{v}$ is defined the following way: we first take an arbitrary representative of $\underline{u}$ and then the representative of $\underline{v}$ whose initial point and the endpoint of the representative of $\underline{u}$ agree (we also say somewhat inaccurately that we translate the vector $\underline{v}$ to the endpoint of $\underline{u}$). Then the sum is determined by the representative which points from the initial point of (the representative of) $\underline{u}$ to the endpoint of (the representative of) $\underline{v}$.

Also, if $\lambda \in \mathbb{R}$ is a non-zero real number and $\underline{v}$ is a non-zero space vector, then we define $\lambda\underline{v}$ the following way: we multiply the length of $\underline{v}$ by $|\lambda|$ and the direction of the product will be the same if $\lambda > 0$ and the opposite if $\lambda < 0$. If $\lambda = 0$ or $\underline{v} = \underline{0}$ is the zero vector, then the result is the zero-vector. In this situation λ is called a *scalar* and this operation is called *scalar multiplication*. Finally, the difference of $\underline{u}$ and $\underline{v}$ is defined by $\underline{u} - \underline{v} := \underline{u} + (-1) \cdot \underline{v}$.

Like in the case of the plane, the basic properties of these operations remain true. The proofs of the following claims are basically the same as the ones of the analogous statements for plane vectors.

Theorem 2.1.1. *If $\underline{u}$, $\underline{v}$ and $\underline{w}$ are space vectors, then*

(i) $(\underline{u} + \underline{v}) + \underline{w} = \underline{u} + (\underline{v} + \underline{w})$ *(the addition is associative),*

(ii) $\underline{u} + \underline{v} = \underline{v} + \underline{u}$ *(the addition is commutative),*

(iii) $\underline{u} + \underline{0} = \underline{u}$,

(iv) there is an additive inverse for any vector, namely $\underline{u} + (-1) \cdot \underline{u} = \underline{0}$ *holds.*

Moreover, if $\lambda, \mu \in \mathbb{R}$, *then*

(v) $\lambda(\underline{u} + \underline{v}) = \lambda\underline{u} + \lambda\underline{v}$,

(vi) $(\lambda + \mu)\underline{u} = \lambda\underline{u} + \mu\underline{u}$,

(vii) $\lambda(\mu\underline{u}) = (\lambda\mu)\underline{u}$,

(viii) $1\underline{u} = \underline{u}$.

If we fix a coordinate system, then we can easily connect these operations with the coordinates of the vectors. As before, we omit the proof of the following theorem and refer to the case of the plane instead.

Theorem 2.1.2. *If* $\underline{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ *and* $\underline{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ *are space vectors and* $\lambda \in \mathbb{R}$ *is a scalar, then*

(i) $\underline{u} + \underline{v} = (u_1 + v_1, u_2 + v_2, u_3 + v_3)$,

(ii) $\underline{u} - \underline{v} = (u_1 - v_1, u_2 - v_2, u_3 - v_3)$,

(iii) $\lambda\underline{u} = (\lambda u_1, \lambda u_2, \lambda u_3)$.

For the next definition we need to introduce a notation: the length of the vector $\underline{u}$ is denoted by $|\underline{u}|$. If $\underline{u}$ and $\underline{v}$ are non-zero vectors, then their *scalar product* is defined by $\underline{u} \cdot \underline{v} = |\underline{u}| \cdot |\underline{v}| \cdot \cos\varphi$, where φ is the angle of the vectors (i.e. the angle of the lines determined by some representatives of the vectors). If any one of the vectors are zero, then the scalar product is defined to be zero.

The scalar product can be used to decide if two non-zero vectors are orthogonal. Namely, if $\underline{u}$ and $\underline{v}$ are non-zero, then $\underline{u} \cdot \underline{v} = |\underline{u}| \cdot |\underline{v}| \cdot \cos\varphi = 0$ holds if and only if $\cos\varphi = 0$, i.e. $\varphi = 90°$. What makes this latter observation useful is that the scalar product can be expressed easily with the help of the coordinates:

Theorem 2.1.3. *If* $\underline{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ *and* $\underline{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ *are space vectors, then* $\underline{u} \cdot \underline{v} = u_1v_1 + u_2v_2 + u_3v_3$.

Again, the proof of this theorem is basically the same as the analogous one about plane vectors.

2.1.2 Equations of a Line

A line on the plane is determined by one of its points and a vector which is parallel or orthogonal (perpendicular) to it. In the space the situation is different: if we fix a point of a line and a vector which is orthogonal to it, then this does not determine the line uniquely. Indeed, if we rotate the line around the axis which is parallel to the vector and contains the point, then we obtain other lines with the same property, and the union of these lines is a plane which is orthogonal to the vector.

But the other option still works, so assume that l is a line of the space, then we fix a point $P_0(x_0, y_0, z_0) \in l$ and a non-zero vector $\underline{v} = (a, b, c)$ which is parallel to l. We are going to construct all the points of the line. Assume that $P(x, y, z)$ is an arbitrary point of the space, then it lies on l if and only if the vector $\overrightarrow{P_0P}$ is parallel to v or it is the zero vector. To avoid many cases we say that the zero vector is parallel to every vector. If $\underline{p}_0$ is the vector that

points from the origin to P_0, and similarly, $\underline{p}$ is the vector that points from the origin to P, then $\overrightarrow{P_0P} = \underline{p} - \underline{p}_0$. Now two vectors are parallel to each other if and only if they have the same or the opposite direction, or if one of them is zero. We obtain that $P \in l$ holds if and only if $\underline{p} - \underline{p}_0 = \lambda \underline{v}$ for some $\lambda \in \mathbb{R}$, or equivalently, $\underline{p} = \underline{p}_0 + \lambda \underline{v}$. By Theorem 2.1.2 we have that $\underline{p}_0 + \lambda \underline{v} = (x_0 + \lambda a, y_0 + \lambda b, z_0 + \lambda c)$, hence our condition is equivalent to

$$\begin{aligned} x &= x_0 + \lambda a, \\ y &= y_0 + \lambda b, \\ z &= z_0 + \lambda c, \end{aligned} \tag{2}$$

where $\lambda \in \mathbb{R}$ is an appropriate real number. These equations are called the *parametric equations* of the line l. When the parameter λ runs through the set of real numbers, the triples (x, y, z) run through the points of l.

While these equations give exactly the points of l, they are inconvenient when we want to decide if a given point $P(x, y, z)$ is on l, because we first have to compute the parameters for which the first, second and third equation of (2) are true. If the same λ suits for all of them, that means that the point P is on l. However, this argument gives another description, which is often better for our goals. We summarize this in the following

Theorem 2.1.4. *Assume that l is a line in the space parallel to the vector $\underline{v} = (a, b, c)$, and the point $P_0(x_0, y_0, z_0)$ lies on l. Then an arbitrary point $P(x, y, z)$ lies on l if and only if*

a) $a \neq 0$, $b \neq 0$ and $c \neq 0$, and

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c}, \qquad \textit{or}$$

b) $a \neq 0$, $b \neq 0$ and $c = 0$, and

$$\frac{x - x_0}{a} = \frac{y - y_0}{b}, \ \textit{and } z = z_0,$$

or an analogous condition holds if exactly one of a, b and c is 0, or

c) $a = b = 0$ and $c \neq 0$, and

$$x = x_0, \ y = y_0,$$

or an analogous condition holds if exactly one of a, b and c is non-zero.

Proof. We already have that $P \in l$ holds if and only if the equations in (2) hold simultaneously for some $\lambda \in \mathbb{R}$. If $a \neq 0$, $b \neq 0$ and $c \neq 0$, then we can express λ from the equations and we get that this system of equations is solvable if and only if the condition in a) holds.

If $a \neq 0$ and $b \neq 0$ but $c = 0$, then the third equation in (2) gives $z = z_0$ while expressing λ from the first two equations we get the same value if and only if the system is solvable. This gives the condition in b). The cases where exactly one of a, b and c is zero can be handled the same way.

Finally, if $a = b = 0$ and $c \neq 0$, then the first first two equations of (2) give $x = x_0$ and $y = y_0$, while the third equation holds automatically for $\lambda = \frac{z - z_0}{c}$ and hence it does not give further restrictions, so we get the condition in c). The cases where exactly one of a, b and c is non-zero can be handled the same way. □

2.1.3 Equation of a Plane

We have already seen that a point and a vector determine the plane that contains the point and orthogonal to the vector. Now we describe this plane, i.e. we give a condition which can be used to decide if a point is contained in the plane. Given a plane S, a vector $\underline{n} \neq 0$ is called a *normal vector* of S if it is orthogonal to it. Note that a normal vector of a plane is not unique, every non-zero scalar multiple of it is also a normal vector of the same plane, moreover, we obtain all normal vectors of the plane this way.

Theorem 2.1.5. *Let S be a plane which contains the point $P_0(x_0, y_0, z_0)$ and assume that $\underline{n} = (a, b, c)$, $\underline{n} \neq \underline{0}$ is a normal vector of S. Then an arbitrary point $P(x, y, z)$ lies on S if and only if $ax + by + cz = ax_0 + by_0 + cz_0$ holds.*

Proof. Let $\underline{p} = (x, y, z)$ and $\underline{p}_0 = (x_0, y_0, z_0)$ be the vectors that point from the origin to P and P_0, respectively. Now $P \in S$ is and only if $\overrightarrow{P_0P} = \underline{p} - \underline{p}_0 = (x - x_0, y - y_0, z - z_0)$ is parallel to S and hence orthogonal to $\underline{n}$. That is, we have the equivalent condition

$$0 = (\underline{p} - \underline{p}_0) \cdot \underline{n} = a(x - x_0) + b(y - y_0) + c(z - z_0)$$

by Theorem 2.1.3. Reordering this equation we get the statement of the theorem. □

It follows also that every equation of the form $ax + by + cz = d$ determines a plane, where $a, b, c, d \in \mathbb{R}$ are real numbers and at least one of a, b and c is non-zero. Indeed, assume for example that $a \neq 0$, the other cases are similar. Now the plane which contains the point $(d/a, 0, 0)$ and orthogonal to the vector (a, b, c) is given by the equation above.

2.2 The Space $\mathbb{R}^n$

In this section we generalize the notion of the plane and the space. While it is hard two visualize more than three pairwise orthogonal lines, the identification of the points with the set of coordinate tuples provides an appropriate starting point for this work.

2.2.1 The Notion of $\mathbb{R}^n$

In the case of the plane and the space we used pairs and triples of real numbers to describe the points. In the following step we forget about the geometric background (at least for a while) and proceed in the following way: we are going to work with n-tuples for a general positive integer n. The points represented vectors before, and accordingly we call the n-tuples vectors and use the analogues of the formulae in Theorem 2.1.2 to *define* operations on them:

Definition 2.2.1. Let $n \geq 1$ be a positive integer, then the set of n-tuples (i.e. sequences of length n) of real numbers is denoted by $\mathbb{R}^n$. We write the elements of $\mathbb{R}^n$ in columns, and define the addition operation and the scalar multiplication for a scalar $\lambda \in \mathbb{R}$ by the following formulae:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}, \qquad \text{and} \qquad \lambda \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

The elements of $\mathbb{R}^n$ are called *vectors*, and they are often denoted by underlined lower case letters. The numbers that form the n-tuples are called the *coordinates* of the vectors.

From now on by a vector we do not mean a directed segment, they are simply columns of numbers. The terms "plane vector" and "space vector" are used to refer to the geometric objects.

As before, we define the difference of the vectors $\underline{u}$ and $\underline{v}$ by $\underline{u} - \underline{v} := \underline{u} + (-1)\underline{v}$, which means coordinate-wise difference. The vector with only zero coordinates are called the *zero vector* and it is often denoted by $\underline{0}$. The statement of the following theorem is basically the same as the statement of Theorem 2.1.1:

Theorem 2.2.1. *If $\underline{u}, \underline{v}, \underline{w} \in \mathbb{R}^n$ and $\lambda, \mu \in \mathbb{R}$, then*

(i) $(\underline{u} + \underline{v}) + \underline{w} = \underline{u} + (\underline{v} + \underline{w})$ *(the addition is associative),*

(ii) $\underline{u} + \underline{v} = \underline{v} + \underline{u}$ *(the addition is commutative),*

(iii) $\underline{u} + \underline{0} = \underline{u}$,

(iv) there is an additive inverse for any vector, namely $\underline{u} + (-1) \cdot \underline{u} = \underline{0}$ *holds.*

(v) $\lambda(\underline{u} + \underline{v}) = \lambda\underline{u} + \lambda\underline{v}$,

(vi) $(\lambda + \mu)\underline{u} = \lambda\underline{u} + \mu\underline{u}$,

(vii) $\lambda(\mu\underline{u}) = (\lambda\mu)\underline{u}$,

(viii) $1\underline{u} = \underline{u}$.

Proof. All these properties follow easily from the properties of the addition and multiplication of real numbers. We prove (v) as an example and leave the proof of the other statements to the reader. So assume that $\lambda \in \mathbb{R}$, $\underline{u}, \underline{v} \in \mathbb{R}^n$,

$$\underline{u} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{and} \qquad \underline{v} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Then

$$\lambda(\underline{u} + \underline{v}) = \lambda \cdot \left[\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \right] = \lambda \cdot \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

$$= \begin{pmatrix} \lambda(x_1 + y_1) \\ \lambda(x_2 + y_2) \\ \vdots \\ \lambda(x_n + y_n) \end{pmatrix} = \begin{pmatrix} \lambda x_1 + \lambda y_1 \\ \lambda x_2 + \lambda y_2 \\ \vdots \\ \lambda x_n + \lambda y_n \end{pmatrix}$$

$$= \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix} + \begin{pmatrix} \lambda y_1 \\ \lambda y_2 \\ \vdots \\ \lambda y_n \end{pmatrix} = \lambda \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \lambda \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \lambda\underline{u} + \lambda\underline{v}.$$

□

Remark. Observe that the previous theorem and Theorem 2.1.1 contain the same statement about completely different objects. That is, here we talk about column vectors while in Section 2.1.1 we worked with directed line segments. But these theorems enlighten that these object are very similar from the algebraic point of view. We say that they form a *vector space* over $\mathbb{R}$. This means simply that they satisfy the statement of these theorems. We just want to point out that although we will work with $\mathbb{R}^n$ in this section, we have already seen another example of a vector space, and the truth is that this kind of structures occur many times in many different situations, in fact we will see some other examples in this chapter. We also mention that in Section 2.1.1 we identified the set of space vectors with $\mathbb{R}^3$ after we fixed a coordinate system. This example foreshadows the important role of $\mathbb{R}^n$, but we still cannot say that the set of space vectors and $\mathbb{R}^3$ are basically the same, because we have to choose a coordinate system for the identification. In other words, the coordinates of a space vector look different for different choices of the coordinate system. Speaking loosely, we can reach that the space vectors look like $\mathbb{R}^3$ but they do not look like $\mathbb{R}^3$ naturally. It is probably very hard to understand this concept at first sight, but in fact it is not necessary, since we do not lean on the the notion of vector space later, we concentrate only on some special cases (like $\mathbb{R}^n$) instead and recommend the book [7] for the interested reader.

2.2.2 Subspaces of $\mathbb{R}^n$

In geometry it is clear that a plane contains infinitely many copies of a line and the space contains infinitely many copies of a plane. They are in some sense "smaller", but we can still restrict the vector operations to these subsets. In the following we are going to study the subsets of $\mathbb{R}^n$ which have this property.

Definition 2.2.2. Assume $\emptyset \neq V \subseteq \mathbb{R}^n$ is a non-empty subset if $\mathbb{R}^n$. We say that V is a *subspace* of $\mathbb{R}^n$ if the following hold:

(i) if $\underline{u}, \underline{v} \in V$, then $\underline{u} + \underline{v} \in V$,

(ii) if $\underline{u} \in V$ and $\lambda \in \mathbb{R}$, then $\lambda \underline{u} \in V$.

If V is a subspace of $\mathbb{R}^n$, then this is denoted by $V \leq \mathbb{R}^n$.

In other words, V is a subspace of $\mathbb{R}^n$ if it is non-empty and closed under addition and scalar multiplication. The subsets $V = \mathbb{R}^n$ and $V = \{\underline{0}\}$ satisfy these conditions and hence they are subspaces. They are called the *trivial subspaces* of $\mathbb{R}^n$. We also get that $\underline{0} \in V$ for every subspace V. Indeed, if $\underline{u} \in V$ is an arbitrary vector (note that there is a $\underline{u}$ in V since V is non-empty), then by property (ii) we have $0\underline{u} = \underline{0} \in V$.

Exercise 2.2.1. Decide if the the following sets of $\mathbb{R}^2$ are subspaces or not:

a) $V_1 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 : x \geq 0, y \geq 0 \right\}$,

b) $V_2 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 : x = y \right\}$,

c) $V_3 = \left\{ \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \in \mathbb{R}^4 : x + y + z + w = 0 \right\}$.

Solution. a) If the coordinates of the vector $\underline{u} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$ are positive, then $\underline{u} \in V_1$, but $-x < 0$ and $-y < 0$, so $(-1)\underline{u} \notin V_1$. Hence V_1 is not closed under scalar multiplication, so it is not a subspace of $\mathbb{R}^2$. Note that V_1 is still closed under addition, since the sum of non-negative numbers is non-negative.

b) If $\underline{u}, \underline{v} \in V_2$, then $\underline{u} = \begin{pmatrix} x \\ x \end{pmatrix}$ and $\underline{v} = \begin{pmatrix} y \\ y \end{pmatrix}$ for some $x, y \in \mathbb{R}$, and hence their sum is $\begin{pmatrix} x+y \\ x+y \end{pmatrix}$, which is also in V_2. Also, for any $\lambda \in \mathbb{R}$ the product $\lambda\underline{u} = \begin{pmatrix} \lambda x \\ \lambda x \end{pmatrix}$ is in V_2, hence V_2 is a subspace of $\mathbb{R}^2$.

c) If $\underline{u}, \underline{v} \in V_3$, where

$$\underline{u} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ w_1 \end{pmatrix} \qquad \text{and} \qquad \underline{v} = \begin{pmatrix} x_2 \\ y_2 \\ z_2 \\ w_2 \end{pmatrix},$$

then $w_i = -x_i - y_i - z_i$ for $i = 1, 2$. Hence

$$\underline{u} + \underline{v} = \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \\ z_1 + z_2 \\ w_1 + w_2 \end{pmatrix} \in V_3,$$

since $w_1 + w_2 = -(x_1 + x_2) - (y_1 + y_2) - (z_1 + z_2)$. Similarly, if $\lambda \in \mathbb{R}$, then $\lambda\underline{u} \in V_3$, because $\lambda w_1 = -\lambda x_1 - \lambda y_1 - \lambda z_1$. Thus, V_3 is a subspace of $\mathbb{R}^4$. □

Exercise 2.2.2. Show that the lines in $\mathbb{R}^2$ that contain the origin are subspaces of $\mathbb{R}^2$. Show that the lines and planes in $\mathbb{R}^3$ that contain the origin are subspaces of $\mathbb{R}^3$.

It will turn out later that these are the only subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$ beside the trivial ones.

2.2.3 Generated Subspace

It is a well-known fact that if two plane vectors are not parallel, then all plane vectors can be expressed from them with the vector operations. The analogue of this fact holds also in the space:

Proposition 2.2.2. *If $\underline{a}, \underline{b} \in \mathbb{R}^3$ are space vectors that are not parallel to each other and lie on the plane S which contains the origin, then every vector $\underline{v} \in \mathbb{R}^3$ that lies on S can be expressed in the form $\alpha\underline{a} + \beta\underline{b}$.*

If $\underline{a}, \underline{b}, \underline{c} \in \mathbb{R}^3$ are space vectors such that they do not lie on a plane that contains the origin, then every vector $\underline{v} \in \mathbb{R}^3$ can be expressed in the form $\underline{v} = \alpha\underline{a} + \beta\underline{b} + \gamma\underline{c}$.

Proof. Assume first that $S \leq \mathbb{R}^3$ is a plane that contains the origin (note that S is a subspace by Exercise 2.2.2), and $\underline{a}, \underline{b} \in S$ are vectors in S that are not parallel to each other (and hence both of them are non-zero). If $\underline{v} = \overrightarrow{OP} \in S$, where O is the origin, then let e be the line which goes through O and parallel to the vector $\underline{a}$, and let f be the line which goes through P and parallel to $\underline{b}$. Since the lines e and f lie on the same plane, they intersect each other in a point Q. Then $\underline{v} = \overrightarrow{OQ} + \overrightarrow{QP}$, and since $\overrightarrow{OQ}$ and $\overrightarrow{QP}$ are parallel to $\underline{a}$ and $\underline{b}$, respectively,

we have that $\overrightarrow{OQ} = \alpha\underline{a}$ and $\overrightarrow{QP} = \beta\underline{b}$ for some $\alpha, \beta \in \mathbb{R}$, hence the first claim of the theorem follows.

For the second part let $\underline{a}, \underline{b}, \underline{c} \in \mathbb{R}^3$ be vectors that do not lie on a plane. Then none of them is zero, and the origin together with the endpoints of any two of them determines a plane. So if $\underline{v} \in \mathbb{R}^3$ is an arbitrary vector, then let S be the plane going through the origin which is spanned by a and b. The line which goes through P and is parallel to $\underline{c}$ intersects S in the point R (because it is not parallel to S). Now $\underline{v} = \overrightarrow{OR} + \overrightarrow{RP}$, where $\overrightarrow{RP} = \gamma\underline{c}$ for some $\gamma \in \mathbb{R}$ (since it is parallel to $\underline{c}$) and $\overrightarrow{OR} = \alpha\underline{a} + \beta\underline{b}$ for some $\alpha, \beta \in \mathbb{R}$ by the first paragraph. □

The following definition generalizes the expression $\alpha\underline{a}+\beta\underline{b}+\gamma\underline{c}$ that occurs in the previous theorem:

Definition 2.2.3. If $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k \in \mathbb{R}^n$ are vectors and $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ are scalars, then the *linear combination* of the vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k$ with the scalars $\lambda_1, \lambda_2, \ldots, \lambda_k$ is the vector $\lambda_1\underline{v}_1 + \lambda_2\underline{v}_2 + \cdots + \lambda_k\underline{v}_k$.

Note that in the definition above the number of the vectors and scalars can be 1, that is, for a vector $\underline{v}$ and a scalar λ the vector $\lambda\underline{v}$ is a linear combination of $\underline{v}$. Moreover, we also define the linear combination of an empty set of vectors to be the zero vector $\underline{0}$.

Now the statement of Proposition 2.2.2 can be rephrased the following way: if three vectors in $\mathbb{R}^3$ do not lie on a plane, then every vector in $\mathbb{R}^3$ can be written as a linear combination of those vectors.

Theorem 2.2.3. *Let $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k \in \mathbb{R}^n$ be arbitrary vectors for some $k \in \mathbb{N}$. If $W \subset \mathbb{R}^n$ is the set of vectors that can be expressed as a linear combination of the vectors $\underline{v}_1, \ldots, \underline{v}_k$, then W is a subspace in $\mathbb{R}^n$.*

Proof. If $k = 0$, then the only vector which is a linear combination of the empty set of vectors is $\underline{0}$, hence $W = \{\underline{0}\}$, which is indeed a subspace. So assume that $k \geq 1$. We have to show that $W \neq \emptyset$ and that it is closed under addition and scalar multiplication. First note that taking (for example) the linear combination $0\underline{v}_1 + 0\underline{v}_2 + \cdots + 0\underline{v}_k = \underline{0}$ we have that $\underline{0} \in W$ and hence $W \neq \emptyset$. Now assume that $\underline{w}_1, \underline{w}_2 \in W$, then by the definition of W we have that

$$\underline{w}_1 = \alpha_1\underline{v}_1 + \cdots + \alpha_k\underline{v}_k \qquad \text{and} \qquad \underline{w}_2 = \beta_1\underline{v}_1 + \cdots + \beta_k\underline{v}_k$$

for some $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k \in \mathbb{R}$ scalars. Now using the properties of the vector operations in Theorem 2.2.1 we get that

$$\underline{w}_1 + \underline{w}_2 = (\alpha_1 + \beta_1)\underline{v}_1 + \cdots + (\alpha_k + \beta_k)\underline{v}_k \in W,$$

hence it is a linear combination of the vectors $\underline{v}_1, \ldots, \underline{v}_k$. Similarly, for every $\lambda \in \mathbb{R}$ we have

$$\lambda\underline{w}_1 = (\lambda\alpha_1)\underline{v}_1 + \cdots + (\lambda\alpha_k)\underline{v}_k,$$

which is again a linear combination of the vectors $\underline{v}_1, \ldots, \underline{v}_k$ and hence it is in W. This completes the proof of the theorem. □

Definition 2.2.4. If $\underline{v}_1, \ldots, \underline{v}_k \in \mathbb{R}^n$ are arbitrary vectors, then the subspace W that consists of the linear combinations of these vectors are called the *span* of $\underline{v}_1, \ldots, \underline{v}_k$ and it is denoted by $W = \operatorname{span}\{\underline{v}_1, \ldots, \underline{v}_k\}$. We also say that W is *spanned* or *generated* by the vectors $\underline{v}_1, \ldots, \underline{v}_k$, and we call the vectors $\underline{v}_1, \ldots, \underline{v}_k$ a *generating system* of W.

We can rephrase the statement of Proposition 2.2.2 again: if three vectors of $\mathbb{R}^3$ do not lie on a plane, then they span the whole space $\mathbb{R}^3$, or $\mathbb{R}^3$ is generated by them. Note that the vectors $\underline{v}_1, \dots, \underline{v}_k$ are also in the space spanned by them, since for every $1 \le i \le k$ we have

$$\underline{v}_i = 0\underline{v}_1 + \dots + 0\underline{v}_{i-1} + 1\underline{v}_i + 0\underline{v}_{i+1} + \dots + \underline{v}_k.$$

Remark. The notation $\operatorname{span}\{\underline{v}_1, \dots, \underline{v}_k\}$ expresses that the subspace is spanned by the elements of the set $S = \{\underline{v}_1, \dots, \underline{v}_k\}$. For an arbitrary set $S \subset \mathbb{R}^n$ one can define the subspace $\operatorname{span} S$ to be the set of all linear combinations of finitely many vectors from S. One can show similarly as in the proof of the previous theorem that $\operatorname{span} S$ is indeed a subspace for every (not necessarily finite) subset S of $\mathbb{R}^n$. Also, it is easy to see that this new definition gives the same subspace for a finite set $S = \{\underline{v}_1, \dots, \underline{v}_k\}$ as Definition 2.2.4, since every linear combination of some elements of S can be completed to a linear combination of all of its elements by adding the missing vectors multiplied by 0. We also note that $\operatorname{span} \emptyset = \{\underline{0}\}$.

Exercise 2.2.3. Describe the subspace $\operatorname{span}\{\underline{u}, \underline{v}\} \le \mathbb{R}^3$, where

$$\text{a) } \underline{u} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix}, \ \underline{v} = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}, \qquad \text{b) } \underline{u} = \begin{pmatrix} 1 \\ 6 \\ 1 \end{pmatrix}, \ \underline{v} = \begin{pmatrix} 3 \\ 4 \\ -1 \end{pmatrix}.$$

Solution. a) A linear combination of $\underline{u}$ and $\underline{v}$ with the scalars $\alpha, \beta \in \mathbb{R}$ is

$$\alpha \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ 5\alpha - 2\beta \end{pmatrix}.$$

If a vector $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ can be expressed in this form, then we have to choose the scalars $\alpha = x$ and $\beta = y$, and then $5x - 2y = z$ must hold. Also, if this relation holds between the coordinates, then the vector can be expressed as a linear combination of $\underline{u}$ and $\underline{v}$ choosing $\alpha = x$ and $\beta = y$. Reordering this equation we get that the subspace spanned by $\underline{u}$ and $\underline{v}$ is nothing else than the plane $5x - 2y - z = 0$.

b) This problem can be handled like part a), but now we show another method. Since we are in $\mathbb{R}^3$, 3-dimensional geometry can be applied (and in this solution we write the vectors in a row again). Since $\underline{u}$ and $\underline{v}$ are not parallel, they span a plane S by Proposition 2.2.2. We are going to determine a normal vector $\underline{n} = (a, b, c)$ of S using the scalar product. A normal vector is orthogonal to every vector on S, in particular to $\underline{u}$ and $\underline{v}$. We have seen in the previous section that this is equivalent to $\underline{n} \cdot \underline{u} = 0$ and $\underline{n} \cdot \underline{v} = 0$. By Theorem 2.1.3 this gives that

$$\begin{aligned} a + 6b + c &= 0, \\ 3a + 4b - c &= 0. \end{aligned}$$

If we express c from the second equation and substitute in the first one, then we obtain $4a + 10b = 0$. Then $a = 5$ and $b = -2$ is a solution of this, and the vector $(5, -2, 7)$ satisfies both equations, hence it is a normal vector of S. Using that $\underline{0} \in S$ we get by Theorem 2.1.5 that the equation of the plane is $5x - 2y + 7z = 0$. □

2.2.4 Linear Independence

It can happen that among the vectors that span a subspace W there are "superfluous" elements, which means that some of the vectors may be omitted while the spanned subspace remains the same. Assume for example that $\underline{a}, \underline{b} \in \mathbb{R}^n$ and $\underline{c}$ is the linear combination of $\underline{a}$ and $\underline{b}$, i.e. $\underline{c} = \lambda\underline{a} + \mu\underline{b}$ for some $\lambda, \mu \in \mathbb{R}$. Then every vector in span $\{\underline{a}, \underline{b}, \underline{c}\}$ is of the form

$$\alpha\underline{a} + \beta\underline{b} + \gamma\underline{c} = \alpha\underline{a} + \beta\underline{b} + \gamma(\lambda\underline{a} + \mu\underline{b}) = (\alpha + \gamma\lambda)\underline{a} + (\beta + \gamma\mu)\underline{b},$$

which is a linear combination of $\underline{a}$ and $\underline{b}$. It follows that span $\{\underline{a}, \underline{b}, \underline{c}\} \subset$ span $\{\underline{a}, \underline{b}\}$. On the other hand, every linear combination of $\underline{a}$ and $\underline{b}$ can be written as a linear combination of $\underline{a}$, $\underline{b}$ and $\underline{c}$, since $\alpha\underline{a} + \beta\underline{b} = \alpha\underline{a} + \beta\underline{b} + 0\underline{c}$. Then span $\{\underline{a}, \underline{b}\} \subset$ span $\{\underline{a}, \underline{b}, \underline{c}\}$, and hence

$$\text{span}\,\{\underline{a}, \underline{b}\} = \text{span}\,\{\underline{a}, \underline{b}, \underline{c}\}.$$

If there is no superfluous vector in he above sense, then we say that the vectors are independent:

Definition 2.2.5. The collection of vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_k \in \mathbb{R}^n$ is called *linearly independent*, if no one of them can be written as a linear combination of the others. If there is a vector among them, which is a linear combination of the others, then the collection of vectors $\underline{v}_1, \ldots, \underline{v}_k$ is said to be *linearly dependent*.

Note that the empty set is defined to be linearly independent. If $k = 1$, then the definition above gives that a single vector is linearly independent if and only if it is not the linear combination of the empty set, i.e. if and only if it is non-zero. For $k = 2$ we get that two vectors are linearly dependent if and only if one of them is the scalar multiple of the other. Note that if the zero vector is among the vectors, then they are linearly dependent, since multiplying any other vector by 0 we get the zero vector as a linear combination of the others.

If $\underline{a}, \underline{b}, \underline{c} \in \mathbb{R}^3$ are space vectors such that they do not lie on a plane that contains the origin, then none of them can be written as a linear combination of the other two. Indeed, any two of them generates a plane which goes through the origin by Proposition 2.2.2, and the third one cannot lie on it. For the same reason we get that if $\underline{a}$, $\underline{b}$ and $\underline{c}$ lie on a plane which contains the origin, then they are linearly dependent.

It is important to note that the linear independence or dependence is the property of the whole collection and not of the single vectors. We use the word "collection" instead of "set" in the definition to handle the situation when a vector appears more then once among $\underline{v}_1, \ldots, \underline{v}_k$, because a vector can be an element of a set only once. Note that in the above case this collection will be dependent automatically, because if $\underline{v}_i = \underline{v}_j$ for some $i \neq j$, then both of them are linear combinations of the other vectors. For example, to express $\underline{v}_i$ we choose the scalar 1 as the coefficient of $\underline{v}_j$ and multiply the other vectors by 0. On the other hand, if the vectors are pairwise distinct (e.g. when they are linearly independent), then they form a set, so we may say that a set of vectors is independent or dependent. Also, we may omit the word collection or set, and simply say (somewhat loosely) that the vectors are independent or dependent. It follows immediately from the definition that if a set of vectors is independent, then any subset of them is also independent.

The following theorem gives an equivalent condition to the linear independence. It is particularly useful when one wants to decide if a collection of vectors is independent. Note that many authors use it to *define* linear independence.

Theorem 2.2.4. *The collection of the vectors $\underline{v}_1, \dots, \underline{v}_k \in \mathbb{R}^n$ is linearly independent if and only if the equation $\lambda_1\underline{v}_1 + \cdots + \lambda_k\underline{v}_k = \underline{0}$ holds only for the scalars $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$.*

Proof. Assume first that $\lambda_1\underline{v}_1 + \cdots + \lambda_k\underline{v}_k = \underline{0}$ holds only in the case $\lambda_1 = \cdots = \lambda_k = 0$. If one of the vectors, say $\underline{v}_i$ can be expressed as the linear combination of the others, then

$$\underline{v}_i = \alpha_1\underline{v}_1 + \cdots + \alpha_{i-1}\underline{v}_{i-1} + \alpha_{i+1}\underline{v}_{i+1} + \cdots + \alpha_k\underline{v}_k$$

for some $\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_k \in \mathbb{R}$, and by reordering the equation we get that

$$\alpha_1\underline{v}_1 + \cdots + \alpha_{i-1}\underline{v}_{i-1} - 1\underline{v}_i + \alpha_{i+1}\underline{v}_{i+1} + \cdots + \alpha_k\underline{v}_k = \underline{0},$$

which contradicts our assumption (since the coefficient of $\underline{v}_i$ is nonzero), so the collection of the vectors must be independent.

Now assume that the collection of the vectors $\underline{v}_1, \dots, \underline{v}_k$ is linearly independent. Now if $\lambda_1\underline{v}_1 + \cdots + \lambda_k\underline{v}_k = \underline{0}$ holds for some $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ such that not all of them are zero, then we choose the index i such that $\lambda_i \neq 0$. But then

$$\underline{v}_i = -\frac{\lambda_1}{\lambda_i}\underline{v}_1 - \cdots - \frac{\lambda_{i-1}}{\lambda_i}\underline{v}_{i-1} - \frac{\lambda_{i+1}}{\lambda_i}\underline{v}_{i+1} - \cdots - \frac{\lambda_k}{\lambda_i}\underline{v}_k$$

contradicting the linear independence (since $\underline{v}_i$ can be expressed as a linear combination of the other vectors). This means that $\lambda_1\underline{v}_1 + \cdots + \lambda_k\underline{v}_k = \underline{0}$ can hold only if all of the coefficients are 0. □

The linear combination $0\underline{v}_1 + 0\underline{v}_2 + \cdots + 0\underline{v}_k = \underline{0}$ is called the *trivial linear combination* of the vectors $\underline{v}_1, \dots, \underline{v}_k$. The previous statement gives that a collection of vectors is independent if and only if there is no linear combination of the vectors which gives the zero vector other than the trivial one.

Exercise 2.2.4. Decide if the following sets of vectors are linearly independent or not.

a) $\begin{pmatrix} 1 \\ 2 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 4 \\ 0 \\ 1 \end{pmatrix}$, b) $\begin{pmatrix} 1 \\ 2 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \\ 1 \end{pmatrix}$.

Solution. a) Let us denote the vectors in order by $\underline{a}$, $\underline{b}$, $\underline{c}$ and $\underline{d}$. We are going to use the previous theorem, that is, we have to decide if the equation $\alpha\underline{a} + \beta\underline{b} + \gamma\underline{c} + \delta\underline{d} = \underline{0}$ has a non-trivial solution (i.e. different from $\alpha = \beta = \gamma = \delta = 0$). Substituting the vectors and using the definitions of the vector operations we get that

$$\alpha\underline{a} + \beta\underline{b} + \gamma\underline{c} + \delta\underline{d} = \begin{pmatrix} \alpha + \beta \\ 2\alpha + 2\beta + 4\delta \\ 2\alpha + 2\beta + 3\gamma \\ 5\beta + \gamma + \delta \end{pmatrix}.$$

This linear combination gives the zero vector if and only if the following equations hold:

$$\begin{aligned} \alpha + \beta &= 0, \\ 2\alpha + 2\beta + 4\delta &= 0, \\ 2\alpha + 2\beta + 3\gamma &= 0, \\ 5\beta + \gamma + \delta &= 0. \end{aligned}$$

If we multiply the first equation by 2 and subtract it from the the second and third equation, the we obtain $4\delta = 0$ and $3\gamma = 0$, which means that δ and γ must be 0. Substituting this in the fourth equation we have $5\beta = 0$, i.e. $\beta = 0$, and then $\alpha = 0$ follows from the first equation. It follows that the vectors are linearly independent.

b) We can start the same way as in part a) and infer the system of equations

$$\begin{aligned} \alpha + \beta + 2\delta &= 0, \\ 2\alpha + 2\beta + 4\delta &= 0, \\ 2\alpha + 2\beta + 3\gamma &= 0, \\ 5\beta + \gamma + \delta &= 0. \end{aligned}$$

Comparing this to the system in part a) one can notice that only the first equation changed. Also, in this case we get the second equation if we multiply the first one by 2. Hence we get an equivalent system if we omit (for example) the first one. Now we can take the difference of the first two equations in the new system to obtain $3\gamma = 4\delta$, i.e. $\delta = \frac{3}{4}\gamma$. Substituting this in the last equation we get $\beta = -\frac{7}{20}\gamma$, and then $\alpha = -\beta - \frac{3}{2}\gamma = (\frac{7}{20} - \frac{3}{2})\gamma = -\frac{23}{20}\gamma$. We have no other information about the variables. Indeed, if we express every other variable in terms of γ and substitute them in the equations, then the coefficient of γ becomes 0. This means simply that the value of γ can be chosen freely, and then the other values are determined. That is, we do have a non-trivial solution of the system (e.g. $\alpha = -23$, $\beta = -7$, $\gamma = 20$, $\delta = 15$), and hence the vectors are dependent. □

2.2.5 The I-G Inequality

In this section we prove a result which will have a crucial role in the following. We have already seen that 3 vectors in $\mathbb{R}^3$ can form a generating system of $\mathbb{R}^3$ (see Proposition 2.2.2) but 2 vectors cannot. Also, 3 vectors of $\mathbb{R}^3$ can be independent if they do not lie on a plane containing the origin, and it is easy to see that 4 space vectors cannot be independent. That is, a set of independent vectors contains at most as many elements as a generating system. This latter statement holds in general (not only in $\mathbb{R}^3$):

Theorem 2.2.5 (I-G inequality)**.** *Let $V \leq \mathbb{R}^n$ be a subspace. If $\underline{f}_1, \ldots, \underline{f}_k \in V$ is a set of independent vectors and $\underline{g}_1, \ldots, \underline{g}_m \in V$ is a generating system in V, then $k \leq m$.*

For the proof we will use the following lemmas:

Lemma 2.2.6. *Assume that $\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_k, \underline{f}_{k+1} \in \mathbb{R}^n$ such that the collection $\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_k$ is linearly independent, while the collection $\underline{f}_1, \underline{f}_2, \ldots, \underline{f}_k, \underline{f}_{k+1}$ is linearly dependent. Then $\underline{f}_{k+1} \in \operatorname{span}\{\underline{f}_1, \ldots, \underline{f}_k\}$ (i.e. $\underline{f}_{k+1}$ can be expressed as the linear combination of the other vectors).*

Proof. As the vectors $\underline{f}_1, \ldots, \underline{f}_k, \underline{f}_{k+1}$ are linearly dependent, by Theorem 2.2.4 we have scalars $\lambda_1, \ldots, \lambda_k, \lambda_{k+1} \in \mathbb{R}$ such that at least one of them is non-zero and

$$\lambda_1 \underline{f}_1 + \cdots + \lambda_k \underline{f}_k + \lambda_{k+1} \underline{f}_{k+1} = \underline{0}.$$

Here $\lambda_{k+1} \neq 0$ must hold, otherwise $\underline{0}$ would be a non-trivial linear combination of the vectors $\underline{f}_1, \ldots, \underline{f}_k$ contradicting the linear independence of this them. Reordering this equation we obtain

$$\underline{f}_{k+1} = -\frac{\lambda_1}{\lambda_{k+1}} \underline{f}_1 - \cdots - \frac{\lambda_k}{\lambda_{k+1}} \underline{f}_k,$$

and the statement is proved. □

Lemma 2.2.7 (The exchange lemma)**.** *Assume that $V \leq \mathbb{R}^n$ is a subspace. If the collection $\underline{f}_1, \ldots, \underline{f}_k \in V$ is linearly independent and $\underline{g}_1, \ldots, \underline{g}_m$ is a generating system of V, then for every $1 \leq i \leq k$ we can find a $1 \leq j \leq m$ such that the vectors $\underline{f}_1, \ldots, \underline{f}_{i-1}, \underline{g}_j, \underline{f}_{i+1}, \ldots, \underline{f}_k$ are linearly independent.*

Proof. After a possible renumbering we may assume $i = k$. Let us replace $\underline{f}_k$ with $\underline{g}_j$ for some $1 \leq j \leq m$. If the collection $\underline{f}_1, \ldots, \underline{f}_{k-1}, \underline{g}_j$ is independent, then we are done. On the other hand, if it is linearly dependent, then we get by the previous lemma that $\underline{g}_j \in \operatorname{span}\{\underline{f}_1, \ldots, \underline{f}_{k-1}\}$ (since the vectors $\underline{f}_1, \ldots, \underline{f}_{k-1}$ are independent because they form a subset of a set of independent vectors).

Assume that we get a dependent collection for every j in the previous paragraph. This means that $\underline{g}_1, \ldots, \underline{g}_m \in \operatorname{span}\{\underline{f}_1, \ldots, \underline{f}_{k-1}\}$, and hence every linear combination of the $\underline{g}_j$'s is in this span, since it is closed under addition and scalar multiplication. But every element of V is a linear combination of the $\underline{g}_j$'s, because they span V. As $\underline{f}_k \in V$, we obtain that $\underline{f}_k \in \operatorname{span}\{\underline{f}_1, \ldots, \underline{f}_{k-1}\}$, and this is impossible, because the vectors $\underline{f}_1, \ldots, \underline{f}_k$ are independent. This contradiction completes the proof of the lemma. □

Proof of Theorem 2.2.5. We apply the previous lemma first to $\underline{f}_1$ and get the set $\underline{g}_j, \underline{f}_2, \ldots, \underline{f}_k$ of independent vectors for some $1 \leq j \leq m$. In the next step we apply the exchange lemma for this set and the generating system $\underline{g}_1, \ldots, \underline{g}_m$, and replace $\underline{f}_2$ with some $\underline{g}_j$ still obtaining an independent set $\underline{g}_j, \underline{g}_l, \underline{f}_3, \ldots, \underline{f}_k$ for some $1 \leq l \leq m$. Continuing this way we can replace all the $\underline{f}_i$'s such that the result is and independent collection of k vectors consisting of some of the $\underline{g}_j$'s. Moreover, in this collection the vectors are different, because they are independent. Since the cardinality of the set $\{\underline{g}_1, \ldots, \underline{g}_m\}$ is m, we must have $k \leq m$. □

2.2.6 Basis and Dimension

The triples of vectors in $\mathbb{R}^3$ that do not lie on a plane containing the origin have a special property: they are independent and also span the whole space $\mathbb{R}^3$. The sets of vectors with this property have an important role.

Definition 2.2.6. Assume that $V \leq \mathbb{R}^n$ is a subspace. The set of the vectors $\underline{b}_1, \ldots, \underline{b}_k \in V$ is called a *basis* of V if it is linearly independent and spans V.

Theorem 2.2.8. *Assume that $V \leq \mathbb{R}^n$ is a subspace. If $\underline{b}_1, \ldots, \underline{b}_k$ and $\underline{c}_1, \ldots, \underline{c}_m$ are bases in V, then $k = m$.*

Proof. We apply the I-G inequality (Theorem 2.2.5) for the independent set $\underline{b}_1, \ldots, \underline{b}_k$ and the generating system $\underline{c}_1, \ldots, \underline{c}_m$ in V and obtain that $k \leq m$. Changing the roles of the two bases and applying the I-G inequality again we get $m \leq k$ and the assertion follows. □

Definition 2.2.7. Let $\underline{b}_1, \ldots, \underline{b}_k$ be a basis in the subspace $V \leq \mathbb{R}^n$. Then the number k is called the *dimension* of V. The dimension of the subspace V is denoted by $\dim V$.

Theorem 2.2.8 assures that the previous definition is correct, since if there is a finite basis in a subspace, then the number of the vectors in it is uniquely determined. But at this point we do not know if there is always a basis in a subspace. Luckily this is the case, i.e. every subspace of $\mathbb{R}^n$ has a dimension (which is finite), but for the proof we need some preparation.

The standard basis

Notation. In the following we use the notation $\underline{e}_i \in \mathbb{R}^n$ for the vector whose coordinates are 0 except for the ith one which is 1.

Proposition 2.2.9. *The set of the vectors $\underline{e}_1, \dots, \underline{e}_n$ is a basis of $\mathbb{R}^n$.*

Proof. First we write down the linear combination of the vectors $\underline{e}_1, \dots, \underline{e}_n$ with the scalars $\lambda_1, \dots, \lambda_n$:

$$\lambda_1 \underline{e}_1 + \dots + \lambda_n \underline{e}_n = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \dots + \lambda_n \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

It follows immediately that if this linear combination gives the zero vector, then every λ_i must be zero, hence the $\underline{e}_i$'s are independent. Also, if $\underline{v} \in \mathbb{R}^n$, then we can choose λ_i to be the ith coordinate of $\underline{v}$, and this way the linear combination above gives the vector $\underline{v}$, i.e. the $\underline{e}_i$'s span $\mathbb{R}^n$. □

Definition 2.2.8. The basis $\underline{e}_1, \dots, \underline{e}_n \in \mathbb{R}^n$ defined above is called the *standard basis* of $\mathbb{R}^n$. It is denoted by E_n or E (if n is clear from the context).

It follows immediately that $\dim \mathbb{R}^n = n$. This is in accordance with our intuition, as one calls the space $\mathbb{R}^3$ three-dimensional. The reason for this that in $\mathbb{R}^3$ there are 3 independent directions, often represented by the directions of the axes which correspond to the vectors $\underline{e}_1$, $\underline{e}_2$ and $\underline{e}_3$ defined above. Though $\mathbb{R}^3$ is three-dimensional also in the sense of Definition 2.2.7, it is still not quite right to call it *the* three-dimensional space, since in general there are other subspaces with this property (note that despite that we are going to do so in some cases).

Exercise 2.2.5. Show that $\mathbb{R}^m$ has an n-dimensional subspace for every $0 \leq n \leq m$.

Exercise 2.2.6. Let $V \leq \mathbb{R}^4$ be the subspace of $\mathbb{R}^4$ which consists of the vectors in $\mathbb{R}^4$ for which the sum of their coordinates is zero (see part c) of Exercise 2.2.1). Give a basis in V.

Solution. If

$$\underline{b}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \qquad \underline{b}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \underline{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix},$$

then $\underline{b}_1, \underline{b}_2, \underline{b}_3 \in V$. We show that these vectors form a basis in V. Now as in the previous proof, we have that

$$\lambda_1 \underline{b}_1 + \lambda_2 \underline{b}_2 + \lambda_3 \underline{b}_3 = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ -\lambda_1 - \lambda_2 - \lambda_3 \end{pmatrix}$$

for every $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. So if this linear combination gives the zero vector, then clearly $\lambda_1 = \lambda_2 = \lambda_3 = 0$ must hold, hence the $\underline{b}_i$'s are independent.

On the other hand, if $\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \in V$, then $w = -x - y - z$, hence we get this vector as a linear combination of the $\underline{b}_i$'s by choosing the coefficients $\lambda_1 = x$, $\lambda_2 = y$, $\lambda_3 = z$. That is, the $\underline{b}_i$'s span V and hence they form a basis in V. □

Existence of a basis in subspaces

Theorem 2.2.10. *If $V \leq \mathbb{R}^n$ is a subspace, then there exists a basis of V.*

Proof. If $V = \{\underline{0}\}$, then the empty set is a basis of V, since it is independent and the zero vector is a linear combination of the empty set by definition (and hence $\dim V = 0$). Otherwise there is a non-zero vector $\underline{0} \neq \underline{v} \in V$, which constitutes a linearly independent set (with 1 element). Hence the statement follows from the next theorem. □

Theorem 2.2.11. *Assume that $V \leq \mathbb{R}^n$ is a subspace. If $\underline{f}_1, \ldots, \underline{f}_k$ is an independent set of vectors in V (where k is a non-negative integer), then it can be completed to a basis of V by adding finitely many (possibly zero) elements.*

Proof. If $W = \operatorname{span}\{\underline{f}_1, \ldots, \underline{f}_k\}$, then $W \subset V$, because V is a subspace, so every linear combination of the $\underline{f}_i$'s must be in V (note that for $k = 0$ we have $W = \{\underline{0}\}$). If $W = V$, then we are done. Otherwise there is a $\underline{v} \in V \setminus W$. Then by Lemma 2.2.6 the collection $\underline{f}_1, \ldots, \underline{f}_k, \underline{v}$ must be independent, otherwise $\underline{v}$ would be in the span of the $\underline{f}_i$'s. If this larger set already generates V, then we are done. Otherwise we continue the same way.

It remains to show that this procedure stops after finitely many steps. But this is true, since by Proposition 2.2.9 there is a generating system in $\mathbb{R}^n$ with n elements, and hence a set of independent vectors in $\mathbb{R}^n$ can contain at most n elements by the I-G inequality. □

Corollary 2.2.12. *Assume that $V \leq \mathbb{R}^n$ is a subspace with $\dim V = k$. If the vectors $\underline{f}_1, \ldots, \underline{f}_k \in V$ are linearly independent, then they constitute a basis in V.*

Proof. By the previous theorem the set of the vectors $\underline{f}_1, \ldots, \underline{f}_k$ can be completed to a basis by adding finitely many (possibly zero) elements. But since $\dim V = k$, every basis has exactly k elements, so the vectors above form a basis. □

An analogous statement holds with a generating system instead of independent vectors:

Theorem 2.2.13. *Assume that $V \leq \mathbb{R}^n$ is a subspace with $\dim V = k$. If the vectors $\underline{g}_1, \ldots, \underline{g}_k \in V$ span V, then they constitute a basis in V.*

Proof. As $\dim V = k$, there are vectors $\underline{f}_1, \ldots, \underline{f}_k \in V$ which form a basis, and hence they are linearly independent. If we repeat the proof of Theorem 2.2.5 with the $\underline{f}_i$'s as independent vectors and the $\underline{g}_j$'s as the vectors that span the subspace, we get the statement. □

Exercise 2.2.7. Let V be the subspace of those vectors in $\mathbb{R}^4$ for which the sum of their coordinates is zero (we have seen in Exercise 2.2.6 that this is indeed a subspace). Give a basis of V which contains the vector $\underline{f} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ -6 \end{pmatrix}$.

Solution. We have seen in Exercise 2.2.6 that $\dim V = 3$, so by Corollary 2.2.12 it is enough to give 3 independent vectors in V such that one of them is $\underline{f}$. We are going to add $\underline{b}_1$ and $\underline{b}_2$ from the solution of Exercise 2.2.6. Note that the two vectors $\underline{f}, \underline{b}_1$ form an independent set since they are not the scalar multiples of each other. Every linear combination of them is of the form

$$\alpha\underline{f} + \beta\underline{b}_1 = \alpha \begin{pmatrix} 1 \\ 2 \\ 3 \\ -6 \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} \alpha+\beta \\ 2\alpha \\ 3\alpha \\ -6\alpha-\beta \end{pmatrix}.$$

It is easy to see that $\underline{b}_2 \notin \operatorname{span}\{\underline{f}, \underline{b}_1\}$. Indeed, in the other case we would have $2\alpha = 1$ and $3\alpha = 0$, which is impossible. Then by Lemma 2.2.6 we get that the vectors $\underline{f}, \underline{b}_1, \underline{b}_2$ are independent, because otherwise $\underline{b}_2$ would be the linear combination of the other two vectors. As we mentioned above, it follows from this that they form a basis. □

Coordinate Vectors

If $\underline{b}_1, \ldots, \underline{b}_k$ is a basis in the subspace V, then it spans V, that is, every vector can be expressed as a linear combination of the basis vectors. What makes bases special among the generating systems of V is that this representation of the vectors is unique:

Theorem 2.2.14. *Assume that $V \leq \mathbb{R}^n$ is a subspace. Then the vectors $\underline{b}_1, \ldots, \underline{b}_k \in V$ form a basis of V if and only if every $\underline{v} \in V$ can be expressed uniquely as a linear combination of them (i.e. if $\underline{v} = \lambda_1\underline{b}_1 + \cdots + \lambda_k\underline{b}_k = \mu_1\underline{b}_1 + \cdots + \mu_k\underline{b}_k$ holds, then $\lambda_i = \mu_i$ for every $1 \leq i \leq k$).*

Proof. Assume first, that every vector in V can be written uniquely as the linear combination of the $\underline{b}_i$'s. Then of course the $\underline{b}_i$'s generate V. Moreover, since the trivial linear combination of them gives the zero vector (which is in V), no other linear combination can be the zero vector by our assumption, which means by Theorem 2.2.4 that the vectors $\underline{b}_1, \ldots, \underline{b}_k$ are independent, i.e. they form a basis in V.

Now assume that the $\underline{b}_i$'s form a basis in V. Then they span V by definition, so every $\underline{v} \in V$ is a linear combination of them. Assume that for a $\underline{v} \in V$ we have

$$\underline{v} = \lambda_1\underline{b}_1 + \cdots + \lambda_k\underline{b}_k = \mu_1\underline{b}_1 + \cdots + \mu_k\underline{b}_k,$$

then reordering this equality we get that

$$\underline{0} = (\lambda_1 - \mu_1)\underline{b}_1 + \cdots + (\lambda_k - \mu_k)\underline{b}_k.$$

But since the $\underline{b}_i$'s are linearly independent, we obtain by Theorem 2.2.4 that all of the coefficients above are zero, that is, $\lambda_i = \mu_i$ for every $1 \leq i \leq k$. □

Now we can fix a basis $B = \{\underline{b}_1, \ldots, \underline{b}_k\}$ in any subspace $V \leq \mathbb{R}^n$. We remark that this notation is somewhat misleading since it suggests that this basis is a set. Which is true of course, but here the order of the basis vectors will also be important for us (and not just the elements of the set B). From now on, once we say that we fix a basis we mean that we fix an *ordered* basis, i.e. the set B *and* the order of the vectors in B. Still we stick to this (unprecise) notation above since it is common in the literature.

Once a(n ordered) basis is fixed in a subspace V, one can represent every vector of V uniquely as a linear combination of the basis elements, and the coefficients of the basis vectors can be assigned to the vector that is represented. That is, every basis determines a "coordinate system" in $\mathbb{R}^n$. The n coefficients can be written as a column vector, i.e. as an element of $\mathbb{R}^n$:

Definition 2.2.9. Assume that $V \leq \mathbb{R}^n$ is a subspace, $\underline{v} \in V$ and $B = \{\underline{b}_1, \ldots, \underline{b}_k\}$ is a basis in V. If $\underline{v} = \lambda_1\underline{b}_1 + \cdots + \lambda_k\underline{b}_k$ (and then the coefficients $\lambda_1, \ldots, \lambda_k$ are determined uniquely), then the vector

$$[\underline{v}]_B := \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix} \in \mathbb{R}^k$$

is called the *coordinate vector* of $\underline{v}$ relative to B.

If $V = \mathbb{R}^n$ and $B = E_n$ is the standard basis, then $\underline{v} = [\underline{v}]_B$ for every $\underline{v} \in V$ (this follows easily from the proof of Proposition 2.2.9). On the other hand, for any other basis B the coordinate vector relative to B can be different from the vector as we are going to see in the next example.

Exercise 2.2.8. Let

$$B = \left\{ \underline{b}_1 = \begin{pmatrix} 1 \\ 6 \\ 1 \end{pmatrix}, \underline{b}_2 = \begin{pmatrix} 3 \\ 4 \\ -1 \end{pmatrix}, \underline{b}_3 = \begin{pmatrix} 1 \\ -8 \\ 2 \end{pmatrix} \right\} \subset \mathbb{R}^3, \qquad \underline{v} = \begin{pmatrix} 3 \\ 4 \\ 9 \end{pmatrix} \in \mathbb{R}^3.$$

Show that B is a basis in $\mathbb{R}^3$ and determine the coordinate vector $[\underline{v}]_B$.

Solution. At this point we have plenty of available tools, so we show three different solutions for the first part of the exercise (and mention a fourth way). In the first two solutions we show that B spans $\mathbb{R}^3$. Since there are 3 elements in B and $\dim \mathbb{R}^3 = 3$, we obtain by Theorem 2.2.13 that B is a basis.

One can observe that $\frac{1}{7}(2\underline{b}_2 + \underline{b}_3) = \underline{e}_1$, so $\underline{e}_1 \in \operatorname{span}\{\underline{b}_1, \underline{b}_2, \underline{b}_3\} = V$. As the spanned subspace V is closed under addition and scalar multiplication, we have that $\frac{1}{10}(\underline{b}_1 + \underline{b}_2 - 4\underline{e}_1) = \underline{e}_2 \in V$, and then $\underline{b}_1 - \underline{e}_1 - 6\underline{e}_2 = \underline{e}_3 \in V$. So every linear combination of $\underline{e}_1, \underline{e}_2, \underline{e}_3$ is in V, but this is the standard basis of $\mathbb{R}^3$, hence $\operatorname{span}\{\underline{e}_1, \underline{e}_2, \underline{e}_3\} = \mathbb{R}^3 \subset V \subset \mathbb{R}^3$, i.e. $V = \mathbb{R}^3$.

One can show this using 3-dimensional geometry. In Exercise 2.2.3 we calculated the equation of the plane that is spanned by $\underline{b}_1$ and $\underline{b}_2$. This equation is $5x - 2y + 7z = 0$, and substituting the coordinates of $\underline{b}_3$ we see that it is not on this plane, hence by Proposition 2.2.2 the vectors in B span $\mathbb{R}^3$. Note that we could also use the method that was shown in part a) of Exercise 2.2.3 to show that B spans the whole space.

Now we apply Corollary 2.2.12 to show that B is a basis. Again, because of the cardinality of B it is enough to show that the vectors $\underline{b}_1$, $\underline{b}_2$ and $\underline{b}_3$ are linearly independent. As in Exercise 2.2.4, we have to solve the following system of equations:

$$\begin{aligned} \alpha + 3\beta + \gamma &= 0, \\ 6\alpha + 4\beta - 8\gamma &= 0, \\ \alpha - \beta + 2\gamma &= 0. \end{aligned}$$

Multiplying the last equation by 4 and adding it to the second one we get that $10\alpha = 0$, i.e. $\alpha = 0$. Substituting this in the third equation we obtain $\beta = 2\gamma$, while from the first equation we infer $\gamma = -3\beta = -6\gamma$, and hence $\beta = \gamma = 0$, so the vectors in B are linearly independent by Theorem 2.2.4 and hence form a basis.

It remains to calculate the coordinate vector of $\underline{v}$ relative to B. For this we have to solve the equation $\alpha\underline{b}_1 + \beta\underline{b}_2 + \gamma\underline{b}_3 = \underline{v}$, which leads us (by equating the coordinates on the two sides) to the system

$$\begin{aligned} \alpha + 3\beta + \gamma &= 3, \\ 6\alpha + 4\beta - 8\gamma &= 4, \\ \alpha - \beta + 2\gamma &= 9. \end{aligned}$$

Again, we multiply the last equation by 4 and add it to the second one to get $10\alpha = 40$, i.e. $\alpha = 4$. Substituting this in the first and the third equation we get that

$$\begin{aligned} 3\beta + \gamma &= -1, \\ -\beta + 2\gamma &= 5. \end{aligned}$$

Now we multiply the second equation by 3 and add it to the first one to obtain $7\gamma = 14$, that is, $\gamma = 2$, and then $\beta = -1$. Hence the coordinate vector of $\underline{v}$ is

$$[\underline{v}]_B = \begin{pmatrix} 4 \\ -1 \\ 2 \end{pmatrix}.$$

□

2.3 Systems of Linear Equations

In the previous section we encountered systems of equations several times. Namely, every time we wanted to express a vector as a linear combination of some other vectors we got linear equations by equating the coordinates of the linear combination and the vector in question.

By a *linear equation* we mean an equation of the form $a_1x_2 + a_2x_2 + \cdots + a_nx_n = b$, where $x_1, x_2, \ldots, x_n$ are variables, and the coefficients $a_1, a_2, \ldots, a_n$ and the constant term b are real numbers. A *system of linear equations* consists of finitely many linear equations, where the same variables (say $x_1, \ldots, x_n$) occur in every equation. By a solution of this system we mean the real numbers $y_1, \ldots, y_n$ that satisfy all the equations at the same time when substituted in the place of the variables.

These kind of systems occur in many applications (beside the one that we have already seen), so we give a general algorithm for their solution called the *Gaussian elimination*. With the help of this algorithm we will be able to decide if a system of linear equations is solvable, and if it is, then the algorithm will give all of its solutions in a manageable way. Our method will also make it possible for us to prove general statements about systems of equations.

2.3.1 Examples of Gaussian Elimination

Before describing the general algorithm we give some examples as an introduction. Let us consider the following system of linear equations with 3 variables and 4 equations (in the left column):

$$\begin{array}{l} 2x_1 - x_2 + 6x_3 = 12 \\ 2x_1 + 2x_2 + 3x_3 = 24 \\ 6x_1 - x_2 + 17x_3 = 46 \\ 4x_1 - x_2 + 13x_3 = 32 \end{array} \quad \rightarrow \quad \begin{array}{l} x_1 - \frac{1}{2}x_2 + 3x_3 = 6 \\ 2x_1 + 2x_2 + 3x_3 = 24 \\ 6x_1 - x_2 + 17x_3 = 46 \\ 4x_1 - x_2 + 13x_3 = 32 \end{array} \quad \rightarrow \quad \begin{array}{l} x_1 - \frac{1}{2}x_2 + 3x_3 = 6 \\ 3x_2 - 3x_3 = 12 \\ 2x_2 - x_3 = 10 \\ x_2 + x_3 = 8 \end{array}$$

For the elimination of the first variable x_1 from the last three equations we first divide the first equation by 2 and hence the coefficient of x_1 becomes 1, while the other equations are left unchanged (this can be seen in the middle column above). Then we subtract a suitable multiple of the (new) first equation from the others so that x_1 does not occur in the resulting equations (see the third column).

Now we can repeat these steps for the system of the last three equations, where the number of the variables is less than in the original system (since x_1 has already been eliminated). But before this we rewrite the steps above in a form which is practical for storing this kind of data. Namely, the names of the variables and the signs of the operations are superfluous (at least for the computer). Hence we omit them and write the coefficients and the constant term on the right hand sides in a so-called *augmented coefficient matrix*. This is nothing else than a table of numbers where every row belongs to an equation and we write the coefficients of the variables in the (fixed) order of the variables in every row, while we separate the constant term in the end of every line by a vertical line. Here are the 3 matrices for the original system and for the ones after the first two steps:

$$\left(\begin{array}{rrr|r} 2 & -1 & 6 & 12 \\ 2 & 2 & 3 & 24 \\ 6 & -1 & 17 & 46 \\ 4 & -1 & 13 & 32 \end{array}\right) \sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 2 & 2 & 3 & 24 \\ 6 & -1 & 17 & 46 \\ 4 & -1 & 13 & 32 \end{array}\right) \sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 3 & -3 & 12 \\ 0 & 2 & -1 & 10 \\ 0 & 1 & 1 & 8 \end{array}\right)$$

Observe that the multiplication of an equation by a number α (i.e. the division by $1/\alpha$) corresponds to the multiplication of the elements in a row by α in the augmented coefficient matrix. Similarly, adding (or subtracting) a multiple of an equation to another corresponds to adding (or subtracting) a scalar multiple of a row to another, where we make the operations element-wise (like in the case of column vectors).

Now we continue the elimination using this notation. In the next step we divide the second row by 3 (i.e. multiply it by 1/3), then we subtract from the third and the fourth row the (new) second row multiplied by 2 and 1, respectively:

$$\sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 1 & -1 & 4 \\ 0 & 2 & -1 & 10 \\ 0 & 1 & 1 & 8 \end{array}\right) \sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 2 & 4 \end{array}\right) \sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

In the third row of the resulting matrix the first non-zero number is 1, so we only subtract 2 times this row from the last one. This way every number in the last row becomes 0, and this row corresponds to the equation $0x_1 + 0x_2 + 0x_3 = 0$. This equation holds regardless of how the values of the variables x_1, x_2 and x_3 are chosen. Hence the solutions of the system remain the same if we omit this equation:

$$\sim \left(\begin{array}{rrr|r} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 1 & 2 \end{array}\right)$$

The result above is called the *row echelon form* of the system: there is a non-zero element in every row before the vertical line, and the first one among these is 1 (in every row). Moreover if $i < j$, then the first non-zero number in the jth row is on the right of the first non-zero element in the ith row. The first non-zero element in a row on the left side of the vertical line is called the *leading coefficient* of that row. Observe that one may read the

solution of the equation from this. The last row corresponds to $0x_1 + 0x_2 + 1x_3 = 2$, i.e. $x_3 = 2$. From the second row we get $x_2 - x_3 = 4$, hence $x_2 = 4 + x_3 = 6$. We get similarly from the first row that $x_1 = 3$.

Instead of this we can continue the process, and eliminate all non-zero numbers in the matrix above the leading coefficient. That is, we add the last row to the second one and subtract 3 times the last row from the first one. Note that this does not effect the first two columns of the matrix, because the first two elements in the last row are zero. Finally, we add $\frac{1}{2}$ times the (new) second row to the first one (effecting only the second and the last column, because every other element in the second row is 0):

$$\sim \left(\begin{array}{ccc|c} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 6 \\ 0 & 0 & 1 & 2 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 6 \\ 0 & 0 & 1 & 2 \end{array}\right)$$

The result above is the so-called *reduced row echelon form*: it is a row echelon form (and hence the leading coefficient of every row is 1) and every other element in the column of the leading coefficients is zero. One can read the solution of the system from it directly: the last column contains the values of x_1, x_2 and x_3 in this order.

Let us change the right hand side of the last equation in the original system to 33. If we repeat the steps that we made above until we got the identically zero last row, here we obtain the following matrix:

$$\left(\begin{array}{ccc|c} 1 & -\frac{1}{2} & 3 & 6 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

Here the last row corresponds to the equation $0x_1 + 0x_2 + 0x_3 = 1$, which does not hold for any values of x_1, x_2, x_3. As this equation follows from the original ones, this means that the system is *inconsistent*, i.e. it has no solution. Hence our algorithm can stop whenever a row occurs where every element is zero except the last one (we may call this a *forbidden row*).

We show another example for the Gaussian elimination, where the steps that we made above cannot always be accomplished and we need a refinement of our method. Also, in this case we will have infinitely many solutions. So let us consider the system

$$\begin{aligned} x_1 + x_2 + 2x_3 + 2x_4 + x_5 &= -1, \\ 4x_1 + 4x_2 + 8x_3 + 9x_4 + x_5 &= -7, \\ 2x_1 + 5x_2 + 13x_3 + x_4 + 26x_5 &= 10, \\ x_1 + 3x_2 + 8x_3 + 2x_4 + 11x_5 &= 1, \\ 2x_1 + x_2 + x_3 + 2x_4 + 3x_5 &= 3. \end{aligned}$$

First we make the augmented coefficient matrix of the system. The boxed element of the matrix will indicate the current phase of the process. These boxed elements will be set to 1 and these will be the leading coefficients of the rows. The first coefficient of the first row is 1 originally, so we do not need to multiply this row, we only change the numbers below it to

0 by adding an appropriate multiple of the first row to the other rows:

$$\left(\begin{array}{ccccc|c} \boxed{1} & 1 & 2 & 2 & 1 & -1 \\ 4 & 4 & 8 & 9 & 1 & -7 \\ 2 & 5 & 13 & 1 & 26 & 10 \\ 1 & 3 & 8 & 2 & 11 & 1 \\ 2 & 1 & 1 & 2 & 3 & 3 \end{array}\right) \sim \left(\begin{array}{ccccc|c} \boxed{1} & 1 & 2 & 2 & 1 & -1 \\ 0 & 0 & 0 & 1 & -3 & -3 \\ 0 & 3 & 9 & -3 & 24 & 12 \\ 0 & 2 & 6 & 0 & 10 & 2 \\ 0 & -1 & -3 & -2 & 1 & 5 \end{array}\right)$$

Now we change the position of the box: we step to the next row and then to the next column. But in this example we cannot continue as before, because the boxed element is zero and hence we cannot multiply the row so that this number becomes 1. But this problem can be solved easily: we may swap the second and the third rows (equations) - and leave the position of the box unchanged:

$$\sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & \boxed{0} & 0 & 1 & -3 & -3 \\ 0 & 3 & 9 & -3 & 24 & 12 \\ 0 & 2 & 6 & 0 & 10 & 2 \\ 0 & -1 & -3 & -2 & 1 & 5 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & \boxed{3} & 9 & -3 & 24 & 12 \\ 0 & 0 & 0 & 1 & -3 & -3 \\ 0 & 2 & 6 & 0 & 10 & 2 \\ 0 & -1 & -3 & -2 & 1 & 5 \end{array}\right)$$

Now we continue as above: we divide the second row by 3 and add an appropriate multiple of the (new) second row to the others below it such that every number below the leading coefficient of the second row becomes zero:

$$\sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & \boxed{1} & 3 & -1 & 8 & 4 \\ 0 & 0 & 0 & 1 & -3 & -3 \\ 0 & 2 & 6 & 0 & 10 & 2 \\ 0 & -1 & -3 & -2 & 1 & 5 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & \boxed{1} & 3 & -1 & 8 & 4 \\ 0 & 0 & 0 & 1 & -3 & -3 \\ 0 & 0 & 0 & 2 & -6 & -6 \\ 0 & 0 & 0 & -3 & 9 & 9 \end{array}\right)$$

Next we change the position of the box as before, and we get that the boxed element is zero again. But unlike in the previous case every element below the boxed number is zero and we cannot swap the rows. Hence we cannot change the boxed element to 1 without changing the first two rows (which we will keep fixed for a while). Instead of this, we change the position of the box again: we step to the next element in the same row:

$$\sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & 1 & 3 & -1 & 8 & 4 \\ 0 & 0 & \boxed{0} & 1 & -3 & -3 \\ 0 & 0 & 0 & 2 & -6 & -6 \\ 0 & 0 & 0 & -3 & 9 & 9 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & 1 & 3 & -1 & 8 & 4 \\ 0 & 0 & 0 & \boxed{1} & -3 & -3 \\ 0 & 0 & 0 & 2 & -6 & -6 \\ 0 & 0 & 0 & -3 & 9 & 9 \end{array}\right)$$

The current leading coefficient is 1, so we do not have to multiply the third row, we simply change the numbers below the boxed element to zero. This way every element of the last two rows becomes zero, so we can omit these rows and obtain the echelon form of the matrix:

$$\sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & 1 & 3 & -1 & 8 & 4 \\ 0 & 0 & 0 & \boxed{1} & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & -1 \\ 0 & 1 & 3 & -1 & 8 & 4 \\ 0 & 0 & 0 & \boxed{1} & -3 & -3 \end{array}\right)$$

The following steps of the algorithm lead to the reduced echelon form. We are going to change every element above the leading coefficients to zero. We begin this process in the last row by adding the last row to the second one and subtracting two times the last row from the first one. Then we step back to the second row and subtract it from the first one:

$$\sim \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 0 & 7 & 5 \\ 0 & 1 & 3 & 0 & 5 & 1 \\ 0 & 0 & 0 & 1 & -3 & -3 \end{array}\right) \sim \left(\begin{array}{ccccc|c} 1 & 0 & -1 & 0 & 2 & 4 \\ 0 & 1 & 3 & 0 & 5 & 1 \\ 0 & 0 & 0 & 1 & -3 & -3 \end{array}\right)$$

We have reached the reduced echelon form and the algorithm stops. This form gives us all the solutions of the system in a manageable form. To see this we consider the equations that correspond to the rows of the last matrix:

$$\begin{aligned} x_1 \quad - \quad x_3 \quad + 2x_5 &= 4, \\ x_2 + 3x_3 \quad + 5x_5 &= 1, \\ x_4 - 3x_5 &= -3. \end{aligned}$$

Easy to see that the values of x_3 and x_5 can be chosen freely, and after that the values of the other variables can be expressed in terms of these values uniquely. Hence the solutions of the system can be written in the following form:

$$\begin{aligned} &x_3 = \alpha \in \mathbb{R},\ x_5 = \beta \in \mathbb{R}, \\ &x_1 = 4 - 2\beta + \alpha, \\ &x_2 = 1 - 5\beta - 3\alpha, \\ &x_4 = -3 + 3\beta. \end{aligned}$$

We call the variables x_3 and x_5 *free parameters* (since their values can be chosen freely). They are those variables for which there is no leading coefficient in the corresponding column of the coefficient matrix.

2.3.2 Gaussian Elimination

For a system of equations with n variables and k equations we fix the notation x_j for the variables ($1 \leq j \leq n$) and $a_{i,j}$ for the coefficient of x_j in the ith equation ($1 \leq i \leq k$, $1 \leq j \leq n$). The constant on the right hand side of the ith equation will be denoted by b_i. Also, we say that the system is of size $(k \times n)$, and its equations and augmented coefficient matrix are the following:

$$\begin{array}{rcl} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n &=& b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n &=& b_2 \\ \vdots & \vdots & \vdots \\ a_{k,1}x_1 + a_{k,2}x_2 + \cdots + a_{k,n}x_n &=& b_k \end{array} \qquad \left(\begin{array}{cccc|c} a_{1,1} & a_{1,2} & \dots & a_{1,n} & b_1 \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,n} & b_k \end{array}\right)$$

Definition 2.3.1. If a system of equations of size $(k \times n)$ is given with its augmented coefficient matrix, then we call the following operations *elementary row operations*: for every $1 \leq i, j \leq k$, $i \neq j$ and $\lambda \in \mathbb{R}$,

(i) the (element-wise) multiplication of a row by λ if $\lambda \neq 0$,

(ii) replacement of the ith row of the matrix by the sum of itself and λ times the jth row,

(iii) swapping of the ith and the jth rows,

(iv) omission of a row which contains only zero elements.

Proposition 2.3.1. *The operations given in the previous definition are equivalent transformations of the coefficient matrix, i.e. the numbers $y_1, \dots, y_n$ constitute a solution of the system given by the coefficient matrix before the operations if and only if they give a solution after the operations.*

Proof. We prove the statement only for operation (ii), the remaining part of the proof is left to the reader. Assume that $y_1, \dots, y_n$ is a solution of the system given by the matrix, then $a_{i,1}y_1 + \dots + a_{i,n} = b_i$ and $a_{j,1}y_1 + \dots + a_{j,n} = b_j$ hold. Adding λ times the second equation to the first we obtain that $(a_{i,1} + \lambda a_{j,1})y_1 + \dots (a_{i,n} + \lambda a_{j,n})y_n = b_i + \lambda b_j$. But this means that the equation belonging to the ith row of the matrix after the operation holds. As the other rows do not change, we have that $y_1, \dots, y_n$ is a solution of the new system.

On the other hand, if $y_1, \dots, y_n$ is the solution of the system that is described by the coefficient matrix after the operation, then $(a_{i,1} + \lambda a_{j,1})y_1 + \dots (a_{i,n} + \lambda a_{j,n})y_n = b_i + \lambda b_j$ and $a_{j,1}y_1 + \dots + a_{j,n} = b_j$ hold (these correspond to the ith and jth row of the new matrix, respectively). Multiplying the latter equation by λ and subtracting the result from the former one we get $a_{i,1}y_1 + \dots + a_{i,n} = b_i$, hence $y_1, \dots, y_n$ satisfy the ith equation of the original system, and since the other equations does not change, $y_1, \dots, y_n$ is a solution of the original system. □

Definition 2.3.2. If a system of equations of size $(k \times n)$ is given with its augmented coefficient matrix, then we say that it is of *row echelon form*, when the following hold:

(i) every row of the matrix contains a non-zero element before the vertical line, and the first non-zero element (the so-called *leading coefficient*) of the row is 1,

(ii) if $1 \leq i < j \leq k$, and the leading coefficient of the ith row is in the lth column, while the leading coefficient in the jth row is in the mth column, then $l < m$ (and then every element below a leading coefficient in the corresponding column is zero, moreover, every element on the left of a leading coefficient in its row and in the rows below it are also zero).

We say that the coefficient matrix is of *reduced row echelon form*, if the following holds beside (i) and (ii):

(iii) every element above a leading coefficient in the corresponding column is zero (i.e. a column of a leading coefficient contains only one non-zero element, namely the leading coefficient itself).

Here is an example of a matrix of row echelon form and another one which is of reduced row echelon form (every $*$ denotes an arbitrary real number):

$$\left(\begin{array}{cccccc|c} 1 & * & * & \dots & * & * & * \\ 0 & 0 & 1 & \dots & * & * & * \\ 0 & 0 & 0 & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & * & * \end{array}\right), \qquad \left(\begin{array}{cccccc|c} 1 & * & 0 & \dots & 0 & * & * \\ 0 & 0 & 1 & \dots & 0 & * & * \\ 0 & 0 & 0 & \dots & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & * & * \end{array}\right).$$

If a system is of reduced row echelon form, then it is easy to find all of its solutions. Indeed, if every column contains a leading coefficient (which is 1), then the unique solution is given by the values on the right of the vertical line (as in the first example of the previous section). If there are columns which do not contain a leading coefficient, then they correspond to *free parameters*, i.e. variables whose values can be chosen freely, and then the values of the other variables can be expressed in terms of the free parameters and the values on the right of the vertical line (see the last example of the previous section).

The *Gaussian elimination* works in the following way: if a system of equations is given (by an augmented coefficient matrix), then we apply elementary row operations so that either we get a *forbidden row*, that is, a row which contains only zero elements on the left side of the vertical line and a non-zero last element (and then the system is *inconsistent*, i.e. it has no solution), or a matrix of reduced row echelon form is obtained (and then we can read all of the solutions of the system). All this is ensured by Proposition 2.3.1. The process is divided into two phases. In the first one we reach a matrix of row echelon form or we get a row whose elements are all zero except for the last one. In the latter case we stop and give the output "there is no solution". Otherwise we continue with the second phase where we reach a matrix of reduced row echelon form.

Gaussian Elimination - First Phase

Assume that the size of the system is $(k \times n)$. We store the number of the row (in the variable i) and the number of the column (in the variable j) where the next leading coefficient is supposed to be. Initially we set $i = j = 1$. In the first part of this phase we run a loop, whose body is described by the following two paragraphs.

If $a_{i,j} \neq 0$, then we multiply the ith row by $1/a_{i,j}$, and for every $i < l \leq k$ we multiply the (new) ith row by $(-a_{l,j})$ and add it to the lth row (obtaining zeros below the current leading coefficient). Now if $i < k$ and $j < n$, then we increase i and j by 1 continue from the beginning of the body, otherwise we break the loop and jump to the second part of the first phase (detailed below later).

On the other hand, if $a_{i,j} = 0$ and $a_{l,j} \neq 0$ for some $i < l \leq k$, then we choose an l with this property (e.g. the least one) and swap the ith and the lth rows and continue as in the previous paragraph. If there is no such l, then we increase j by 1 when it is smaller than n and go back to the beginning of the body (i.e. to the previous paragraph). If we cannot increase j, then we decrease i by 1, break the loop and jump to the second part of the first phase.

In the second part of the first phase we do the following. If $i = k$, then this means that we reached the last line of the matrix and set the leading coefficient in it to 1. In this case the matrix is of row echelon form, so we finish the first phase. If $i < k$, then either we reached the nth column and set the leading coefficient to 1 in the ith row and nth column and eliminated the non-zero elements below it, or we had $a_{l,n} = 0$ for every $i + 1 \leq l \leq k$ anyway (but then $a_{i,n}$ is not necessarily 1, or in the (degenerate, but still possible) case when $i = -1$ it is not even defined). In all of these cases we have only zeros in the lth row on the left side of the vertical line for every $i < l \leq k$. So if $b_l \neq 0$ for some $i < l \leq k$, then there are no solutions and the algorithm stops. Otherwise we omit the lth row for every $i < l \leq k$.

We give the steps of this phase also in the form of a pseudocode:

GAUSSIAN ELIMINATION - FIRST PHASE

Input: a matrix A with k rows and $n+1$ columns (the augmented coefficient matrix of a system of linear equations with n variables and k equations)

```
1   i ← 1; j ← 1;
2   while true do
3       if a_{i,i} ≠ 0, then
4           multiply the ith row by 1/a_{i,j}
5           if i < k, then
6               for every i < l ≤ k add (−a_{l,j}) times the ith row to the lth row
7           if i = k or j = n, then
8               goto SECOND PART
9           else
10              i ← i + 1; j ← j + 1
11      else
12          if i < k and a_{l,j} ≠ 0 for some i < l ≤ k, then
13              swap the ith and the lth rows
14          else
15              if j = n, then
16                  i ← i − 1
17                  goto SECOND PART
18              else
19                  j ← j + 1
20  end while
21  SECOND PART:
22  if i < k, then
23      if b_l ≠ 0 for some i < l ≤ k, then
24          print "The system has no solution."; stop
25      else
26          omit the lth row for every i < l ≤ k
27  print "The matrix is of echelon form."; stop
```

Gaussian Elimination - Second Phase

In the second phase of the algorithm the input is an augmented coefficient matrix which is of row echelon form. Here we simply eliminate the non-zero elements above the leading coefficients. For example, if $a_{i,j} = 1$ is the leading coefficient of the ith row, then for every $1 \leq l < i$ we multiply the ith row by $a_{l,j}$ and subtract the result from the lth row (element-wise). We begin this phase with the last row and go backwards. Note that this is not necessary but this way we decrease the number of operations (because of the zeros that we produced in the previous steps of this phase). It is obvious that the resulting matrix is of reduced echelon form. We summarize this in the following theorem:

Theorem 2.3.2. *If a system of equation is given by its augmented coefficient matrix and we apply Gaussian elimination, then exactly one of the following cases holds:*

(i) We get a line at the end of the first phase whose elements are zero except for the last one. In this case the system has no solution.

(ii) We obtain a matrix of reduced row echelon form such that there is a leading coefficient in every column. Then the system has a unique solution.

(iii) We obtain a matrix of reduced row echelon form such that there are columns without a leading coefficient. Then the system has infinitely many solutions.

In the cases (ii) and (iii) the solutions can be read from the reduced echelon form as discussed above.

Corollary 2.3.3. *If a system of equations with k equations and n variables has a unique solution, then $k \geq n$.*

Proof. As the system has a solution, the Gaussian elimination produces a matrix of reduced row echelon form with say k' rows. Then $k' \leq k$ since the algorithm does not increase the number of the lines. Since the solution is unique, every column of the resulting matrix contains a leading coefficient, but the number of the leading coefficients is the same as the number of the rows, hence $k' = n$ and the claim follows. □

Finally, we turn to the running time of the Gaussian elimination. It is not hard to see that in the case of a system with k equations and n variables the algorithm makes at most ck^2n basic operations for some constant c, but the running time of these operations largely depends on how we store the numbers that are obtained during the process as results of previous operations, and also on how we actually implement these operations.

If all the inputs are rational numbers, then it would be possible to store the numerator and the denominator of them. But in this case we have to simplify the fractions after performing the operations on them, otherwise the magnitude of the numbers can get so large that the running time of the algorithm becomes exponential. This simplification can be done (for example) with the Euclidean algorithm, which still gives a polynomial running time, but unfortunately it is not fast enough for applications.

Hence in practice an approximation (typically a floating-point format) is used giving a reasonable running time. The drawback of this is that the errors of the approximations can accumulate during the process resulting in an unacceptable outcome. This can happen for example when we divide by numbers which are very close to zero. Without giving any further details we summarize this as follows: the Gaussian elimination is an efficient algorithm when implemented carefully.

2.4 The Determinant

We have seen in Corollary 2.3.3 that if a system of equations has a unique solution, then the number of the equations is at least as large as the number of the variables. An important special case is when these two numbers are the same. One may expect that in this case the other direction of the statement of Corollary 2.3.3 holds, i.e. a system with n variables and n equations has a unique solution. However, this is not the case. But we will see later that if the solution is not unique, then this is due to some kind of "coincidence". The tool that is used to describe this phenomenon precisely is the determinant, which will be defined and investigated below.

As an introduction we describe the case of the system with 2 equations and 2 variables:

$$a_{1,1}x_1 + a_{1,2}x_2 = b_1,$$
$$a_{2,1}x_1 + a_{2,2}x_2 = b_2.$$

It is easy to see that if $a_{1,1}a_{2,2} - a_{1,2}a_{2,1} \neq 0$, then

$$x_1 = \frac{a_{2,2}b_1 - a_{1,2}b_2}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}, \qquad x_2 = \frac{a_{1,1}b_2 - a_{2,1}b_1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}$$

is the unique solution of the system. It is also not hard to show that if $a_{1,1}a_{2,2} - a_{1,2}a_{2,1} = 0$, then the system has infinitely many or no solutions depending on the numbers b_1 and b_2. We will handle this problem for a general n, and our results will contain this special case as well. But we encourage the reader to try to prove these statements above.

2.4.1 The Inversion Number of Permutations

By a *permutation* we mean a sequence of length n (for some positive integer n) which contains each of the numbers $1, 2, \ldots, n$ exactly once. In other words, it is a way of arranging the numbers $1, 2 \ldots, n$. The permutations will be denoted by Greek letters. For a permutation π we denote the number in the ith place by π_i. For example, for $n = 8$ a permutation can be given by $\pi = (5, 3, 1, 8, 4, 2, 6, 7)$, and here $\pi_1 = 5, \pi_2 = 3, \ldots, \pi_8 = 7$.

Remark. To be precise, a permutation is defined as a function from the set $\{1, 2, \ldots, n\}$ onto itself, i.e. it is a one-to-one map (or *bijection*), and for a permutation π the number π_i is nothing else than the function value $\pi(i)$. This approach is very useful from many points of view, since the composition of functions provides an operation on the set of permutations, which - for example - makes it possible to introduce the notions below in a natural way. We choose another way instead which is maybe less expressive, but it is shorter and does not lead us to sidetracks.

Definition 2.4.1. Assume that $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ is a permutation. If for some $i < j$ we have $\pi_i > \pi_j$, then the pair of elements (π_i, π_j) is called an *inversion* of π. The *inversion number* of a permutation is the number of inversions (π_i, π_j) of π, and it is denoted by $I(\pi)$.

In the example above the inversions are the pairs $(5, 3)$, $(5, 1)$, $(5, 4)$, $(5, 2)$, $(3, 1)$, $(3, 2)$, $(8, 4)$, $(8, 2)$, $(8, 6)$, $(8, 7)$ and $(4, 2)$, hence $I(\pi) = 11$.

Proposition 2.4.1. *Assume that $\pi = (\pi_1, \pi_2, \ldots, \pi_i, \ldots, \pi_j, \ldots, \pi_n)$ is a permutation, and let us interchange the numbers π_i and π_j. Then this way we obtain another permutation $\pi' = (\pi_1, \pi_2, \ldots, \pi_j, \ldots, \pi_i, \ldots, \pi_n)$, and the parity of $I(\pi)$ and $I(\pi')$ is different (i.e. one of them is even while the other one is odd).*

Proof. We first prove the proposition in the case when we interchange elements in consecutive places, i.e. when $j = i + 1$. Then the relation of π_i and π_j changes: if (π_i, π_j) *is* an inversion of π, then $\pi_i > \pi_j$, and hence $(\pi_j, \pi_i) = (\pi'_i, \pi'_j)$ is *not* an inversion of π'. Similarly, if (π_i, π_j) is not an inversion of π, then $\pi_i < \pi_j$, and then $(\pi_j, \pi_i) = (\pi'_i, \pi'_j)$ is an inversion of π'. For any indices $k \neq l$ different form i and j we have $\pi_k = \pi'_k$ and $\pi_l = \pi'_l$, so (π_k, π_l) is an inversion of π if and only if (π'_k, π'_l) is an inversion of π'. If $k < i$, then $k < j$ holds as well and (π_k, π_i) is an inversion of π if and only if $(\pi'_k, \pi'_j) = (\pi_k, \pi_i)$ is an inversion of π'. The remaining pairs can be handled similarly and we get that the exchange of two neighboring elements increases or decreases the the inversion number by 1, that is, $I(\pi') = I(\pi) \pm 1$, and then the parity of $I(\pi)$ and $I(\pi')$ are different.

Now we turn to the general case. We accomplish the replacement of the elements π_i and π_j by a series of exchanges of neighboring elements. As $i < j$, we first interchange π_j with

π_{j-1}, then with π_{j-2}, and we continue until the elements π_i and π_j become neighbors, i.e. until, say after t exchanges π_j gets in the $(i+1)$th position. Starting with the permutation $\pi = \pi^{(0)}$ we obtain this way a sequence of permutations $\pi^{(1)}, \pi^{(2)}, \dots, \pi^{(t)}$. Then we swap the elements π_i and π_j and get the permutation $\pi^{(t+1)}$, and finally we move the element π_i in the original position of π_j via t exchanges of consecutive elements resulting in the permutations $\pi^{(t+2)}, \dots, \pi^{(2t+1)} = \pi'$. By the first paragraph the parity of $I(\pi^{(k)})$ and $I(\pi^{(k+1)})$ is different for every $0 \le k \le 2t$, and we change this parity $2t+1$ times, hence the parity of $I(\pi)$ and $I(\pi')$ must be different. □

2.4.2 The Definition of the Determinant

Let us consider the following problem: we would like to place 8 rooks on a chessboard in a non-attacking arrangement (i.e. in a configuration such that none of them can capture any other rook in one step). This means that any row or any column can contain at most 1 rook, and since the number of rows (and columns) on the board is 8, we get that every row (and column) must contain exactly 1 rook.

The problem can be solved in the following way: we place the rooks on the board in 8 steps. In the ith step we place one rook in the ith row, and hence every row will contain exactly one of them. Once we place a rook on the board, we exclude exactly one column from the possible choices in the next steps. That is, in the first step we can choose a square in the first row freely. But in the second step the column of the first rook is excluded, so there are 7 possibilities left. Similarly, in the third step we can choose from 6 squares for the third rook, etc. We obtain that the number different non-attacking rook arrangements is $8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 8! = 40320$.

Observe that we can encode the rook arrangements above with a permutation. Namely, we can assign to the rook in the ith row the number of its column and obtain a sequence of 8 numbers this way. Since there is exactly one rook in every column, every number between 1 and 8 occurs exactly once, that is, we get a permutation. This method works in the opposite direction as well, i.e. every permutation $\pi = (\pi_1, \dots, \pi_8)$ gives a rook arrangement: in the ith step we place a rook in the ith row and the π_ith column. Since every number between 1 and 8 occurs exactly once in the sequence $\pi_1, \dots, \pi_8$, we get a non-attacking rook arrangement indeed. These maps between the arrangements and the permutations are inverses to each other, that is, if we encode an arrangement with a permutation and then obtain an arrangement from this permutation like we described above, then we get back the original arrangement.

Note that the number 8 has no significance here. Assume that a table of real numbers is given with n rows and n columns. Such a table is called a *matrix* of size $n \times n$ and the numbers in the table are called the *entries* of the matrix. The entry of a matrix A in the ith row and jth column is usually denoted by $a_{i,j}$.

We can choose n entries in the matrix so that any two of them are in different rows and different columns. In other words, every row and every column contains exactly one entry from the chosen ones. Such a set of n entries will be called a rook arrangement. As before, we can encode the rook arrangements with a permutation of n numbers.

Definition 2.4.2. Let us choose n entries of a matrix A of size $n \times n$ so that every row and every column of A contains exactly one of them. Then the set of these entries is called a *rook arrangement* of A. We say that a permutation of the numbers $1, 2, \dots, n$ *corresponds to a rook arrangement* if in the arrangement the π_ith entry is chosen from the ith row.

It follows in the same way as above that the number of the rook arrangements for a matrix of size $n \times n$ is $n!$.

Definition 2.4.3. Assume that A is a matrix of size $n \times n$. For every rook arrangement of A we multiply the entries of it, and we multiply the product by $(-1)^{I(\pi)}$, where π is the permutation that corresponds to the arrangement (i.e. we keep the sign of the product if the permutation π has even inversion number, and change the sign otherwise). The sum of the $n!$ products that are obtained this way is called the *determinant* of A and it is denoted by $\det A$ or $|A|$. It can be expressed by the following formula:

$$\det A = \sum_{\pi} (-1)^{I(\pi)} a_{1,\pi_1} a_{2,\pi_2} \dots a_{n,\pi_n}, \tag{3}$$

where we sum over all permutations π of the set $\{1, 2, \dots, n\}$.

It is an important part of the definition that the matrix A is a *square matrix*, i.e. the number of its rows is the same as the number of its columns. Of course we can talk about a matrix of size $k \times n$ with k rows and n columns, but its determinant is defined only if $k = n$.

In general it is rather tiresome to calculate the determinant of a matrix for even a relatively small n. For example, if $n = 5$, then we have $5! = 120$ rook arrangements, so we have to multiply 5 numbers and calculate the inversion number of a permutation 120 times. Also, the value $n!$ grows so fast (faster than an exponential function) that a computer cannot accomplish the calculation in a reasonable time for bigger n's. Later we give a polynomial algorithm for this task.

On the other hand, for $n = 2$ and $n = 3$ it is not hard to memorize the formula that the definition gives. The case $n = 2$ is so simple and important that we give the details of its calculation. If

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}, \tag{4}$$

then there are only two rook arrangements and we get two products: $a_{1,1}a_{2,2}$ and $a_{1,2}a_{2,1}$. The first one belongs to the permutation $(1, 2)$ whose inversion number is 0, and the second one belongs to $(2, 1)$ with inversion number 1. Hence

$$\det A = \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}. \tag{5}$$

That is, the determinant of a 2×2 matrix is the difference of the products of the entries in the *diagonals* of the matrix. The product of the entries in the so-called *main diagonal* (the diagonal that consists of entries with the same row and column index) has the positive sign, while we take the product of the entries in the *antidiagonal* with a negative sign. It is important that this formula holds *only* for a 2×2 matrix.

Exercise 2.4.1. Determine the formula for the determinant in the case of a 3×3 matrix.

As final remark of this section we mention that once we give a matrix with its entries (like on the right hand side of (4)), then in the notation of its determinant we often omit the parentheses (like in (5)).

2.4.3 The Basic Properties of the Determinant

Although the definition of the determinant cannot be used in general for its calculation, there are special cases when the formula in (3) simplifies a lot. These special determinants turn out to be useful in general since (as we will see later) an arbitrary determinant can be transformed so that its calculation becomes easy.

The matrix A of size $n \times n$ is called *upper triangular* if every entry of it below the main diagonal is 0. That is, for every $1 \leq i, j \leq n$, $i > j$ we have $a_{i,j} = 0$. Similarly, A is called a *lower triangular matrix* if every entry above its main diagonal is 0, i.e. for every $1 \leq i, j \leq n$, $i < j$ we have $a_{i,j} = 0$.

Upper triangular matrix:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & \dots & a_{1,n} \\ 0 & a_{2,2} & a_{2,3} & a_{2,4} & \dots & a_{2,n} \\ 0 & 0 & a_{3,3} & a_{3,4} & \dots & a_{3,n} \\ 0 & 0 & 0 & a_{4,4} & \dots & a_{4,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & a_{n,n} \end{pmatrix}$$

Lower triangular matrix:

$$\begin{pmatrix} a_{1,1} & 0 & 0 & 0 & \dots & 0 \\ a_{2,1} & a_{2,2} & 0 & 0 & \dots & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} & 0 & \dots & 0 \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & a_{n,4} & \dots & a_{n,n} \end{pmatrix}$$

Theorem 2.4.2. *Assume that A is a matrix of size $n \times n$.*

(i) If A has a row or column whose entries are all 0, then $\det A = 0$.

(ii) If A is a upper triangular or a lower triangular matrix, then its determinant is the product of the entries in its main diagonal, i.e. $\det A = a_{1,1}a_{2,2}\dots a_{n,n}$.

Proof. Part (i) follows immediately from the definition of the determinant. Indeed, assume that the ith row of A contains only 0 entries. Since every term in the sum (3) is the product of some entries of the matrix and exactly one of them is from the ith row, we get that every product is 0 and hence so is the determinant. The analogous claim for a column instead of a row follows the same way (replacing the word "row" by "column" in the argument above).

For the proof of (ii) we first assume that A is an upper triangular matrix. If a rook arrangement contains a 0 entry, then the term in (3) belonging to it becomes zero. Hence we are going to identify those arrangements which does not (necessarily) contain 0 entries. From the first column we can only choose the first entry $a_{1,1}$, since all the other entries are zero. From the second column we cannot chose the first entry since it is excluded by our choice in the first column. The remaining entries are zero except for $a_{2,2}$ in the main diagonal, so we have to choose this entry. Similarly, from the third column we cannot choose the first two entries and below the third one every entry is zero, hence we choose $a_{3,3}$. Continuing this way we get that the only rook arrangement which does not (necessarily) contain a 0 entry is the one which consists of the entries in the main diagonal, and this belongs to the permutation $(1, 2, \dots, n)$ whose inversion number is 0. Thus, in (3) there is only one term left, namely $a_{1,1}a_{2,2}\dots a_{n,n}$ with a positive sign. The analogous statement for a lower triangular matrix can be proved similarly, but also follows from the next theorem. □

Definition 2.4.4. Let A be a matrix of size $k \times n$, then the *transpose* of A is a matrix of size $n \times k$ denoted by A^T whose entry in the jth row and the ith column is the same as the entry of A in the ith row and the jth column for every $1 \leq i \leq k$ and $1 \leq j \leq n$. That is, if the entry of A in the ith row and jth column is $a_{i,j}$ and the entry of A^T in the jth row and ith column is $b_{j,i}$, then $a_{i,j} = b_{j,i}$.

One may visualize this in the following way: we get the transpose of a matrix when we reflect its entries to the main diagonal. This operation swaps the rows and the columns of the matrix, i.e. the rows of a matrix are the same as the columns of its transpose (similarly, the columns of a matrix are the same as the rows of its transpose). Note that $(A^T)^T = A$ holds. An example is the following:

$$A = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 \\ 12 & 13 & 14 & 15 & 16 \end{pmatrix}, \quad A^T = \begin{pmatrix} 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \\ 6 & 11 & 16 \end{pmatrix}.$$

If A is a lower triangular matrix, then A^T is an upper triangular matrix, so the statement of (ii) in the previous theorem for a lower triangular matrix follows from the case of the upper triangular matrix and the following:

Theorem 2.4.3. *If A is a matrix of size $n \times n$, then* $\det A^T = \det A$.

Proof. Let us denote the entry of $B = A^T$ in the ith row and jth column by $b_{i,j}$ (and the entry of A in the same position by $a_{i,j}$). We are going to show that in the formula (3) for A and B we get the same products with the same signs and hence the statement follows.

Let $\pi = (\pi_1, \dots, \pi_n)$ be an arbitrary permutation. The corresponding term in the formula (3) for A is

$$(-1)^{I(\pi)} a_{1,\pi_1} a_{2,\pi_2} \dots a_{n,\pi_n} = (-1)^{I(\pi)} b_{\pi_1,1} b_{\pi_2,2} \dots b_{\pi_n,n}$$

by the definition of the matrix B. Now every number between 1 and n occurs exactly once in the sequence $\pi_1, \dots, \pi_n$, so if π' is the permutation for which $\pi'_{\pi_i} = i$, then (since the multiplication of the real numbers is commutative) the term above can be written as

$$(-1)^{I(\pi)} b_{\pi_1,\pi'_{\pi_1}} b_{\pi_2,\pi'_{\pi_2}} \dots b_{\pi_n,\pi'_{\pi_n}} = (-1)^{I(\pi)} b_{1,\pi'_1} b_{2,\pi'_2} \dots b_{n,\pi'_n}.$$

It is obvious that if the permutations π and ϱ are different, then so are the permutations π' and ϱ'. Indeed, there is an i for which $\pi_i \neq \varrho_i$, and hence $\pi'_{\pi_i} = i = \varrho'_{\varrho_i} \neq \varrho'_{\pi_i}$, since ϱ' is a permutation and takes different values for different indices. So the map $\pi \mapsto \pi'$ is one-to-one on the set of permutations (because its domain and its image has the same (finite) cardinality). So it is enough to show that $I(\pi) = I(\pi')$ for every permutation π, because then

$$(-1)^{I(\pi)} a_{1,\pi_1} a_{2,\pi_2} \dots a_{n,\pi_n} = (-1)^{I(\pi')} b_{1,\pi'_1} b_{2,\pi'_2} \dots b_{n,\pi'_n},$$

and this way we get disjoint pairs of equal terms in the formula (3) for A and B, and then $\det A = \det B$ must hold.

So assume that for some $1 \leq k < l \leq n$ we have $\pi'_k = i$ and $\pi'_l = j$. Then by definition $\pi_i = k$ and $\pi_j = l$. Now the pair (π'_k, π'_l) is an inversion of π' if and only if $i > j$. But this latter inequality means exactly that the pair $(\pi_j, \pi_i) = (l, k)$ is an inversion of π. However, if (π'_k, π'_l) is not an inversion, then $i < j$ and $(\pi_i, \pi_j) = (k, l)$ is not an inversion either. Also, for different inversions of π' we get different inversions of π, so we have a one-to-one map between the inversions of π' and π, i.e. $I(\pi) = I(\pi')$, and the proof is complete. □

Remark. The permutation π' defined in the previous proof is called the *inverse permutation* of π. The reason for this is the following: if we regard the permutations as functions from the set $\{1, 2, \dots, n\}$ onto itself, then π' is the inverse function of π, that is, $\pi'(\pi(i)) = i = \pi(\pi'(i))$

holds for every integer $1 \leq i \leq n$. The first equality is the direct consequence of the definition of π'. But also the second equation follows from the definition, since if $\pi'(i) = k$, then $i = \pi(k)$ must hold, because π' is a bijection (a one-to-one function), and hence cannot take the same values on different places, and then $\pi(\pi'(i)) = \pi(k) = i$.

To calculate a determinant of an arbitrary matrix we will transform it so that we get (for example) an upper triangular matrix and then the calculation becomes easy by Theorem 2.4.2. The following theorem describes the steps of this transformation:

Theorem 2.4.4. *Assume that A is a matrix of size $n \times n$, $\lambda \in \mathbb{R}$ is a scalar and $1 \leq i, j \leq n$, $i \neq j$ are integers.*

(i) If we multiply a row or a column of A by λ element-wise, then for the resulting matrix A' we have $\det A' = \lambda \cdot \det A$.

(ii) If we interchange two rows or columns of A, then for the resulting matrix A' we have $\det A' = (-1) \cdot \det A$.

(iii) If we replace the ith row by the (element-wise) sum of itself and λ times the jth row, then the determinant of the resulting matrix A' is the same as the determinant of A, i.e. $\det A' = \det A$. Similarly, if we replace the ith column by the (element-wise) sum of itself and λ times the jth column obtaining the matrix A', then $\det A' = \det A$.

Proof. By the previous theorem it is enough to prove the statements for row operations, because an operation on the columns means an operation on the rows of the transpose of the matrix. In other words, assume that the statements are true for row operations and we obtain the matrix A' by a column operation while we get the matrix A'' by the corresponding row operation. Then if for example this operation on the columns is of type (iii), then

$$\det A' = \det(A')^T = \det(A^T)'' = \det A^T = \det A.$$

Here the first and the last equality follows from the previous theorem. The second equality means that if we make the operation on the columns and then we reflect the matrix to the main diagonal, then this means the same as a reflection and then the corresponding operation on the rows of the transpose. Finally, the third equation follows from our assumption (that row operations of type (iii) do not change the determinant). The claim follows similarly for the other types of column operations.

For the proof of (i) let us assume (for example) that we obtain A' by multiplying the ith row of A by λ. By definition, the determinant of A' is obtained by the formula

$$\begin{aligned}\det A' &= \sum_{\pi} (-1)^{I(\pi)} a_{1,\pi_1} \dots a_{i-1,\pi_{i-1}} (\lambda a_{i,\pi_i}) a_{i+1,\pi_{i+1}} \dots a_{n,\pi_n} \\ &= \lambda \cdot \sum_{\pi} (-1)^{I(\pi)} a_{1,\pi_1} \dots a_{n,\pi_n} = \lambda \cdot \det A.\end{aligned}$$

For the proof of (ii) let us assume that we obtain the matrix A' by swapping the ith and jth row of A, where $1 \leq i < j \leq n$. We are going to pair the terms in the formula (3) for A and A'. To a term belonging to the permutation $\pi = (\pi_1, \dots, \pi_i, \dots, \pi_j, \dots, \pi_n)$ we assign the term with the corresponding permutation $\pi' = (\pi_1, \dots, \pi_j, \dots, \pi_i, \dots, \pi_n)$. We are going to show that these terms differ only in sign, and since in both sums every term has

exactly one pair (i.e. we have a one-to-one map between the terms of the sums), we get that $\det A' = -\det A$.

Let us fix the permutation π, then the corresponding term in the sum (3) is

$$(-1)^{I(\pi)} a_{1,\pi_1} \dots a_{i,\pi_i} \dots a_{j,\pi_j} \dots a_{n,\pi_n}.$$

Let π' be the assigned permutation. Then the parity of $I(\pi)$ and $I(\pi)$ are different by Proposition 2.4.1, and hence $(-1)^{I(\pi')} = -(-1)^{I(\pi)}$. Also, if $a_{k,l}$ and $a'_{k,l}$ are the elements of A and A' in their kth row and lth column, respectively, then $a_{k,\pi_k} = a'_{k,\pi'_k}$ for every $k \neq i, j$, because the kth row of A and A' are the same and $\pi_k = \pi'_k$ holds as well. Also, we have $a'_{i,\pi'_i} = a_{j,\pi_j}$ and $a'_{j,\pi'_j} = a_{i,\pi_i}$ by the definition of A' and π'. Hence the term that corresponds to π' in the formula (3) for $\det A'$ is

$$(-1)^{I(\pi')} a'_{1,\pi'_1} \dots a'_{i,\pi'_i} \dots a'_{j,\pi'_j} \dots a'_{n,\pi'_n} = -(-1)^{I(\pi)} a_{1,\pi_1} \dots a_{j,\pi_j} \dots a_{i,\pi_i} \dots a_{n,\pi_n},$$

and this is exactly that we wanted to show.

Before the proof of (iii) we are going to show the following statement (which is useful also in other situations):

Lemma 2.4.5. *Assume that the $n \times n$ matrices X, X' and X'' have the same entries outside the ith row, while the ith row of X is the element-wise sum of the ith row of X' and X'':*

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x'_{i,1} + x''_{i,1} & x'_{i,2} + x''_{i,2} & x'_{i,3} + x''_{i,3} & \cdots & x'_{i,n} + x''_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n} \end{pmatrix},$$

$$X' = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x'_{i,1} & x'_{i,2} & x'_{i,3} & \cdots & x'_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n} \end{pmatrix}, \quad X'' = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x''_{i,1} & x_{i,2} & x''_{i,3} & \cdots & x''_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n} \end{pmatrix}.$$

Then $\det X = \det X' + \det X''$. The analogous claim holds with columns instead of rows.

Proof. By the previous theorem it is enough to prove the statement for rows. For any permutation π the corresponding term in the definition of $\det X$ is

$$\begin{aligned} (-1)^{I(\pi)} x_{1,\pi_1} \dots (x'_{i,\pi_i} + x''_{i,\pi_i}) \dots x_{n,\pi_n} = \\ = (-1)^{I(\pi)} x_{1,\pi_1} \dots x'_{i,\pi_i} \dots x_{n,\pi_n} + (-1)^{I(\pi)} x_{1,\pi_1} \dots x''_{i,\pi_i} \dots x_{n,\pi_n}, \end{aligned}$$

because it is the sum of the corresponding terms in $\det X'$ and $\det X''$. This holds for every term and hence the claim follows. □

Turning to the proof of (iii) we will apply the previous lemma for the matrices A', A and Y, where A' is the resulting matrix after the operation, A is the original matrix and Y is the matrix that is obtained by replacing the ith row of A by λ times the jth row of A. Then the lemma gives that $\det A' = \det A + \det Y$, so it remains to show that $\det Y = 0$.

First note that $\det Y = \lambda \cdot \det Y'$ by (i), where Y' is the matrix obtained from A by replacing its ith row by its jh row. Now we can apply (ii) for Y': if we swap its ith and jth row, then we get the same matrix, on the other hand, the sign of the determinant changes by (ii), i.e. $\det Y' = (-1) \cdot \det Y'$, hence $\det Y' = 0$ and therefore $\det Y = \lambda \cdot \det Y' = 0$ must hold and the claim follows. □

2.4.4 The Calculation of the Determinant

In this chapter we first show an example for the application of Theorem 2.4.4 and after that we give a general algorithm for the calculation of the determinant. The method will be familiar since these transformation rules are similar to the ones that we applied in the Gaussian elimination. There are differences though, first of all, while in the case of a system of equations the analogous steps do not change the set of solutions, here the determinant may change, so we have to keep track of these changes. On the other hand, we have more options here, since while it makes no sense to operate on the set of columns in an augmented coefficient matrix, by the calculation of the determinant this is allowed (and often can be very useful).

Let us calculate the following determinant:

$$\begin{vmatrix} 3 & 12 & -3 & -6 \\ 2 & 8 & 3 & -9 \\ 1 & 5 & -1 & 0 \\ -1 & -3 & 3 & 5 \end{vmatrix}.$$

We are going to use the row operations listed in Theorem 2.4.4 to transform this matrix to an upper triangular matrix. First we apply operation (i) on the first row. If we multiply it by $1/3$, then the determinant is also multiplied by $1/3$, or equivalently:

$$\begin{vmatrix} 3 & 12 & -3 & -6 \\ 2 & 8 & 3 & -9 \\ 1 & 5 & -1 & 0 \\ -1 & -3 & 3 & 5 \end{vmatrix} = 3 \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 2 & 8 & 3 & -9 \\ 1 & 5 & -1 & 0 \\ -1 & -3 & 3 & 5 \end{vmatrix}.$$

The determinant on the right hand side is $1/3$ times the determinant on the left hand side, and after rearranging this we get the equality above. This way we reach that the first non-zero number in the first row is 1, and we add an appropriate multiple of the first row to the other rows so that the numbers below this entry become 0. Namely, we add (-2), (-1) and 1 times the first row to the second one, third one and fourth one, respectively. This way we reach a matrix where the first column looks like in an upper triangular matrix. Also, these steps do not change the value of the determinant, so the last product above is

$$= 3 \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 0 & 0 & 5 & -5 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 2 & 3 \end{vmatrix} = (-3) \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 5 & -5 \\ 0 & 1 & 2 & 3 \end{vmatrix}.$$

After that we cannot continue with the second row as above, since there is a 0 in the second column and we cannot change it to 1 by a multiplication. So we swap the second and the third rows changing the sign of the determinant. This way we already get that the entry in

the second row and second column is 1, so we simply subtract this row from the last one and obtain the last product is

$$= (-3) \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 5 & -5 \\ 0 & 0 & 2 & 1 \end{vmatrix} = (-3) \cdot 5 \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 2 & 1 \end{vmatrix},$$

where we applied an operation of type (i) again (from Theorem 2.4.4) for the third row and for $\lambda = 1/5$. Finally, we subtract 2 times the third row from the last one. This does not change the value of the determinant, and the result is an upper triangular matrix, hence Theorem 2.4.2 is applicable. So the product above becomes

$$(-3) \cdot 5 \cdot \begin{vmatrix} 1 & 4 & -1 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 3 \end{vmatrix} = (-3) \cdot 5 \cdot (1 \cdot 1 \cdot 1 \cdot 3) = -45.$$

That is, the value of the determinant is -45.

The computation above followed the steps of an efficient algorithm which gives the value of the determinant in general. It can be regarded as a variant of the first phase of the Gaussian elimination, since we make similar row operations on the matrix to obtain an upper triangular matrix. Namely, in the ith loop we try to change the entry in the ith row and ith column (i.e. the ith number in the main diagonal) to 1 and then use this to eliminate all non-zero elements in its column below it.

There are important differences though. We have already seen that the changes of the determinant should be recorded during the process. Moreover, if in the ith row we get zero in the main diagonal and all the other entries in its column below it are also zero, then the algorithm can stop, and the determinant is zero. Indeed, we cannot swap the ith row and some other row below it to change this entry to a non-zero number. But if we continue the process with the next element of the main diagonal to obtain an upper triangular form, then the first i columns do not change in the remaining steps. This means that at the end of the process the zero entry in the ith row and ith column makes the product of the numbers in the main diagonal - that is, the determinant - zero.

After this explanation we give the steps of the algorithm by a pseudocode:

CALCULATION OF A DETERMINANT

Input: a matrix A of size $n \times n$

```
1     i ← 1; D ← 1;
2     while true do
3         if a_{i,i} ≠ 0, then
4             D ← a_{i,i} D
5             multiply the ith row by 1/a_{i,i}
6             if i < n, then
7                 for every i < j ≤ n add (−a_{j,i}) times the ith row to the jth row
8                 i ← i + 1
9             else
10                print "det A =",D; stop
11        else
```

```
12              if i < n and a_{j,i} ≠ 0 for some i < j ≤ n, then
13                  swap the ith and jth rows
14                  D ← (−1) · D
15              else
16                  print "det A = 0"; stop
17      end while
```

Exercise 2.4.2. Let A be the matrix of size $n \times n$ for which the entries of the main diagonal are all p for some $p \in \mathbb{R}$ and all its other entries are 1. Calculate the determinant of A.

2.4.5 Systems of Linear Equations and the Determinant

As we have promised in the introductory paragraphs of the section, we are going to handle now the case of a system of linear equations with n equations and n variables. The variant of the Gaussian elimination detailed above establishes the connection between these systems and the determinant which can be used to decide if the system has a unique solution:

Theorem 2.4.6. *Let $(A|\underline{b})$ be the augmented coefficient matrix of a system of linear equations with n equations and n variables. That is, A is the coefficient matrix of size $n \times n$, while $\underline{b} \in \mathbb{R}^n$ is the vector whose coordinates are constants on the right hand sides of the equations. Then the system has a unique solution if and only if* $\det A \neq 0$.

Proof. Let us run first phase of the Gaussian elimination to this system as it is described in Section 2.3.2. To be precise, we run only the first part of the first phase, until line 21 in the code on page 57. Though the steps of the algorithm change the determinant of the coefficient matrix A, but the determinant of the resulting matrix is non-zero if and only if it was non-zero originally, i.e. if and only if $\det A \neq 0$ (this follows from Theorem 2.4.4).

At line 21 of the algorithm we have 3 possibilities. If the system has no solutions, then at that point we have a forbidden row in the matrix. On the other hand, if there is infinitely many solutions, then by Theorem 2.3.2 there is a column without a leading coefficient, so there must be a row which is identically zero (otherwise the number of the leading coefficients would be the number of the rows which is the same as the number of the columns). In both cases the transformed coefficient matrix (without the constant vector on the right) has a row with only zero entries. Then its determinant is 0 by part (i) of Theorem 2.4.2 and then $\det A = 0$ holds as well.

On the other hand, if the solution of the system is unique, then there is a leading coefficient in every column and hence in every row. This means that we get an upper triangular matrix whose entries in the main diagonal are all 1, and then the determinant of the transformed coefficient matrix is non-zero, therefore $\det A \neq 0$ follows. □

2.4.6 The Expansion Theorem for Determinants

In this section we prove a theorem called the Laplace expansion which reduces the calculation of a determinant of size $n \times n$ to the calculation of n determinants of size $(n-1) \times (n-1)$. The verb "reduce" may be a little bit misleading, since it refers to the reduction of the size, not to the reduction of the number of operations that are needed for the computation. Though in a few cases it can be used to simplify the actual calculation, its main importance is revealed mostly in theoretical arguments. Before giving the formula we need some preparations:

Definition 2.4.5. Let A be a matrix of size $n \times n$. With the entry $a_{i,j}$ of A in its ith row and jth column we associate the sign $(-1)^{i+j}$ and the *sub-matrix* $M_{i,j}$ of size $(n-1)\times(n-1)$ which is obtained from A by deleting its ith row and its jth column. Then the *cofactor associated with* $a_{i,j}$ is the product $C_{i,j} = (-1)^{i+j} \det M_{i,j}$. The determinant of a sub-matrix $M_{i,j}$ is called a *(first) minor* of A.

Theorem 2.4.7 (Laplace expansion)**.** *If A is a matrix of size $n \times n$ and we add the entries of a fixed row or column multiplied by the associated cofactors, then the result is the determinant of the matrix. That is,*

$$\begin{aligned}\det A &= \sum_{k=1}^{n} a_{i,k}C_{i,k} = \sum_{k=1}^{n}(-1)^{i+k}a_{i,k}\det M_{i,k} \\ &= \sum_{k=1}^{n} a_{k,j}C_{k,j} = \sum_{k=1}^{n}(-1)^{k+j}a_{k,j}\det M_{k,j}\end{aligned}$$

for every $1 \le i, j \le n$.

Proof. Assume first that the statement of the theorem holds for rows. Let $b_{j,i}$ denote the entry of A^T in its jth row and in its ith column. If $N_{j,i}$ is the associated sub-matrix for $b_{j,i}$ and $M_{i,j}$ is the associated sub-matrix for $a_{i,j}$, then $N_{j,i} = M_{i,j}^T$ holds. So for every $1 \le j \le n$ we have

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{k+j}a_{k,j}\det M_{k,j} &= \sum_{k=1}^{n}(-1)^{k+j}a_{k,j}\det M_{k,j}^T \\ &= \sum_{k=1}^{n}(-1)^{j+k}b_{j,k}\det N_{j,k} \\ &= \det A^T = \det A\end{aligned}$$

by Theorem 2.4.3. Hence it is enough to prove the theorem for rows.

We prove the theorem in three steps. In the first step we show that if the first $n-1$ entries of the last row of A are zero, then the statement holds for the last row, that is,

$$\det A = \sum_{k=1}^{n} a_{n,k}C_{n,k} = \sum_{k=1}^{n}(-1)^{n+k}a_{n,k}\det M_{n,k} = (-1)^{n+n}a_{n,n}\det M_{n,n} = a_{n,n}\det M_{n,n}.$$

For this, observe that a term belonging to a permutation π in the formula (3) is non-zero only if $\pi_n = n$, so it is enough to sum over these permutations:

$$\det A = \sum_{\pi,\,\pi_n=n}(-1)^{I(\pi)}a_{1,\pi_1}\dots a_{n-1,\pi_{n-1}}a_{n,n} = a_{n,n}\sum_{\pi,\,\pi_n=n}(-1)^{I(\pi)}a_{1,\pi_1}\dots a_{n-1,\pi_{n-1}}.$$

Hence it remains to show that this last sum is $\det M_{n,n}$. The entries $a_{1,\pi_1},\dots,a_{n-1,\pi_{n-1}}$ form a rook arrangement of the minor $M_{n,n}$, because $\pi_i \ne n$ for $i = 1,\dots,n-1$. For the same reason, $\pi' = (\pi_1,\dots,\pi_{n-1})$ is a permutation of the numbers $1,\dots,n-1$. Also, for every permutation π' of the numbers $1,\dots,n-1$ we can associate a permutation of the numbers $1,\dots,n$ for which $\pi_n = n$, namely $\pi = (\pi'_1,\dots,\pi'_{n-1},n)$. These maps constitute a one-to-one correspondence between the permutations of $1,\dots,n$ with $\pi_n = n$ and the permutations of $1,\dots,n-1$, which means that we get every rook arrangement of $M_{n,n}$ in

the form $a_{1,\pi_1}, \ldots, a_{n-1,\pi_{n-1}}$. Also, if $\pi_n = n$, then π_n does not occur in any inversion of π, so $I(\pi) = I(\pi')$. It follows that

$$a_{n,n} \sum_{\pi,\, \pi_n = n} (-1)^{I(\pi)} a_{1,\pi_1} \ldots a_{n-1,\pi_{n-1}} = a_{n,n} \sum_{\pi'} (-1)^{I(\pi')} a_{1,\pi'_1} \ldots a_{n-1,\pi'_{n-1}}$$

where in the last sum we sum over the permutations of $1, \ldots, n-1$. But the last sum is $\det M_{n,n}$ by definition and the proof of the first step is complete.

In the second step we prove that if in the ith row of A there is at most 1 non-zero entry, then the statement holds for the ith row. Assume for example that we have $a_{i,k} = 0$ for every $1 \leq k \leq n$, $k \neq j$, where $1 \leq j \leq n$. We have to show that

$$\det A = \sum_{k=1}^{n} (-1)^{i+k} a_{i,k} \det M_{i,k} = (-1)^{i+j} a_{i,j} \det M_{i,j}.$$

Now we transform the matrix A in the following way: we swap the ith and row and the next one, then we swap the new $(i+1)$th row (which is the ith row of A) and the next one, and we continue this way until the ith row of A becomes the last row of the transformed matrix A'. By part (ii) of Theorem 2.4.4 we have $\det A' = (-1)^{n-i} \det A$, since we have made $n - i$ swaps. Similarly, we swap the jth column of A' and the next one, then $(j+1)$th column of the resulting matrix and the following one, and we continue this until the jth column of A' becomes the last one of the transformed matrix A''. As before, we have

$$\det A'' = (-1)^{n-j} \det A' = (-1)^{2n-i-j} \det A = (-1)^{i+j} \det A,$$

i.e. $\det A = (-1)^{i+j} \det A''$. But the first $n-1$ entries of the last row of A'' are zero, so by the first part of the proof we can expand it along its last row:

$$\det A'' = \sum_{k=1}^{n} (-1)^{n+k} a''_{n,k} \det M''_{n,k} = a''_{n,n} \det M''_{n,n},$$

where $a''_{i,j}$ is the entry of A'' in the ith row and the jth column, and $M''_{i,j}$ is the assigned minor. But because of the way that A'' was obtained from A we have $a''_{n,n} = a_{i,j}$ and $M''_{n,n} = M_{i,j}$, so the statement follows for the ith row of A.

Finally, we prove the statement for a general matrix $A \in \mathbb{R}^{n \times n}$. We fix an index $1 \leq i \leq n$ and apply Lemma 2.4.5 for the ith row $(n-1)$ times to obtain

$$\det A = \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & a_{i,3} & \ldots & a_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \ldots & a_{n,n} \end{vmatrix}$$

$$= \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \ldots & a_{n,n} \end{vmatrix} + \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{i,2} & a_{i,3} & \ldots & a_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \ldots & a_{n,n} \end{vmatrix} = \ldots$$

$$= \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{vmatrix} + \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{i,2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{vmatrix} +$$

$$+ \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{i,3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{vmatrix} + \dots + \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{i,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{vmatrix}.$$

By the second part of the proof the statement of the theorem holds for the latter determinants, and the statement follows. □

There is an easy consequence of this theorem which will be useful for us later. Assume that a matrix A is given. Let us fix two different indices $1 \leq i \neq j \leq n$ and construct the matrix A' from A so that we replace its jth row by its ith row. As $i \neq j$, the ith and the jth rows of the matrix A' are identical. If we subtract one of them from the other one, then the determinant of the matrix does not change, but the result contains an identically zero row, so its determinant is zero by part (i) of Theorem 2.4.2 and hence $\det A' = 0$ follows.

Let us expand the matrix A' along the jth row. Then by the previous theorem we have

$$0 = \det A' = \sum_{k=0}^{n} (-1)^{j+k} a'_{j,k} \det M'_{j,k} = \sum_{k=0}^{n} (-1)^{j+k} a_{i,k} \det M_{j,k} = \sum_{k=1}^{n} a_{i,k} C_{j,k},$$

since the ith row of A and the jth row of A' are the same, and the matrices A and A' agree outside their jth row, but the minors $M_{j,k}$ and $M'_{j,k}$ do not contain elements from the jth row of the corresponding matrix, so they are the same. A similar argument (or the application of Theorem 2.4.3) gives the analogous result for columns instead of rows. We summarize this in the following

Corollary 2.4.8. *Assume that $A \in \mathbb{R}^{n \times n}$. Then for indices $1 \leq i \neq j \leq n$ we have*

$$\begin{aligned} 0 &= \sum_{k=1}^{n} a_{i,k} C_{j,k} = \sum_{k=1}^{n} (-1)^{j+k} a_{i,k} \det M_{j,k} \\ &= \sum_{k=1}^{n} a_{k,i} C_{k,j} = \sum_{k=1}^{n} (-1)^{k+j} a_{k,i} \det M_{k,j}. \end{aligned}$$

2.4.7 Three-dimensional Analytic Geometry and the Determinant

In this section we introduce the notion of the *cross product* of vectors in $\mathbb{R}^3$, and shortly address the connection of the determinant with the volume of parallelepipeds.

Remark. The three-dimensional space is easily visualized, and a big part of analytic geometry can be generalized to any dimension. For example the scalar product and also the volume have a natural generalization, and the connection between the determinant and the volume detailed below has an analogue in higher dimensions. Although here we restrict ourselves

to the three-dimensional space, we note that this connection makes it possible to *define* the determinant as the volume of a (higher dimensional) parallelepiped. A useful byproduct of this is that this method provides a natural and easy way for the proof of Theorem 2.5.5 below. On the other hand, the dimension is important in the case of the cross product. Without giving any details we just mention here that an analogue of the three-dimensional cross product can be defined in the 7-dimensional space, but not for any other dimensions (see [8]).

Definition 2.4.6. If $\underline{u}, \underline{v} \in \mathbb{R}^3$ are space vectors, then their *cross product* is denoted by $\underline{u} \times \underline{v}$, and it is the space vector defined uniquely by the following properties when both $\underline{u}$ and $\underline{v}$ are non-zero and not parallel to each other:

1. the length of $\underline{u} \times \underline{v}$ is $|\underline{u} \times \underline{v}| = |\underline{u}|\,|\underline{v}| \sin\varphi$, where $|\underline{u}|$ and $|\underline{v}|$ are the length of the vectors $\underline{u}$ and $\underline{v}$ and φ is the angle between them,
2. $\underline{u} \times \underline{v}$ is orthogonal (perpendicular) to $\underline{u}$ and $\underline{v}$,
3. the system of the vectors $\underline{u}$, $\underline{v}$ and $\underline{u} \times \underline{v}$ is right-oriented (see page 33).

If $\underline{u} = \underline{0}$ or $\underline{v} = \underline{0}$ or they are parallel, then $\underline{u} \times \underline{v} = \underline{0}$ by definition.

An immediate consequence of the definition that this operation on the space vectors is not commutative. Indeed, because of property 3. we have $\underline{v} \times \underline{u} = -\underline{u} \times \underline{v}$. The cross product is useful for example when we need a vector which is orthogonal to two other vectors that are not parallel. Luckily it is easy to determine its coordinates from the coordinates of the factors:

Theorem 2.4.9. *If $\underline{u} = (u_1, u_2, u_3)$ and $\underline{v} = (v_1, v_2, v_3)$ are space vectors, then*

$$\underline{u} \times \underline{v} = \left(\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix}, -\begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \right).$$

We omit the proof of this theorem. Note that the cross product can be expressed in the following form:

$$\begin{vmatrix} \underline{i} & \underline{j} & \underline{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},$$

where $\underline{i}$, $\underline{j}$ and $\underline{k}$ are the unit vectors of the standard basis belonging to the axes x, y and z, respectively. Of course this is just a formal notation, since we cannot write vectors in a determinant. But Theorem 2.4.7 hepls us to read this correctly: if we formally expand this "determinant" along the first row, then we get

$$\underline{u} \times \underline{v} = \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} \cdot \underline{i} - \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} \cdot \underline{j} + \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \cdot \underline{k},$$

which is the statement of the previous theorem.

Finally we mention another important theorem in geometry without a proof. Three position vectors $\underline{u}$, $\underline{v}$ and $\underline{w}$ determine (span) a parallelepiped in the space whose vertices are of the form $\alpha\underline{u} + \beta\underline{v} + \gamma\underline{w}$, where α, β and γ take the values 0 or 1. Then the signed volume of this parallelepiped can be expressed by the coordinates of the vectors (signed volume means that it takes positive sign if the system $\underline{u}$, $\underline{v}$ and $\underline{w}$ is right-oriented, and takes negative sign otherwise).

Theorem 2.4.10. *If $\underline{u} = (u_1, u_2, u_3)$, $\underline{v} = (v_1, v_2, v_3)$ and $\underline{w} = (w_1, w_2, w_3)$ are space vectors, then the signed volume of the parallelepiped spanned by them is*

$$\begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

2.5 Matrices

We have already worked with matrices in the previous sections, they are simply defined as tables of numbers. Now we are going to take a closer look at them. They turn out to be useful for many purposes. In this section we will see how they connect the systems of linear equations and the space $\mathbb{R}^n$, while later we will use them to represent the elements of a very important family of functions that are called linear maps.

2.5.1 Matrix Operations

In the following we will define and investigate the basic operations of matrices. Two of them can be defined like in the case of column vectors:

Definition 2.5.1. For some integers $k, n \geq 1$ a *matrix* of size $k \times n$ is a table which contains k rows and n columns and whose entries are real numbers. The set of the matrices of size $k \times n$ is denoted by $\mathbb{R}^{k\times n}$, while for a matrix $A \in \mathbb{R}^{k\times n}$ we denote the entry in its ith row and jth column by $a_{i,j}$ (and similarly for the matrices $B, C, \ldots$ this entry is $b_{i,j}, c_{i,j}$, etc.). If $A, B \in \mathbb{R}^{k\times n}$, then their sum $A + B \in \mathbb{R}^{k\times n}$ is defined by the following formula:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,n} \end{pmatrix} + \begin{pmatrix} b_{1,1} & b_{1,2} & \ldots & b_{1,n} \\ b_{2,1} & b_{2,2} & \ldots & b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k,1} & b_{k,2} & \ldots & b_{k,n} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} & \ldots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} & \ldots & a_{2,n} + b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} + b_{k,1} & a_{k,2} + b_{k,2} & \ldots & a_{k,n} + b_{k,n} \end{pmatrix}.$$

If moreover $\lambda \in \mathbb{R}$ is a *scalar*, then the matrix $\lambda A \in \mathbb{R}^{k\times n}$ is called a *scalar multiple* of A and it is defined by

$$\lambda \cdot \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,n} \end{pmatrix} = \begin{pmatrix} \lambda a_{1,1} & \lambda a_{1,2} & \ldots & \lambda a_{1,n} \\ \lambda a_{2,1} & \lambda a_{2,2} & \ldots & \lambda a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{k,1} & \lambda a_{k,2} & \ldots & \lambda a_{k,n} \end{pmatrix}.$$

It is an important part of the definition that the sum of two matrices A and B is defined only if A and B have the same number of rows and the same number of columns (and then $A+B$ is the element-wise sum of them). As in the case of the vectors we define the subtraction by $A - B := A + (-1) \cdot B$.

Notice that the set $\mathbb{R}^{k\times n}$ of matrices and the set $\mathbb{R}^{k\cdot n}$ of vectors together with addition operation and the scalar multiplication are basically the same, they differ only in notation. In fact a row vector can be regarded as a matrix of size $1 \times n$, while a column vector can be regarded as a matrix of size $k \times 1$. It is then not surprising that the analogue of Theorem 2.2.1 is true for the matrices as well:

Theorem 2.5.1. *If $A, B, C \in \mathbb{R}^{k\times n}$ are arbitrary matrices and $\lambda, \mu \in \mathbb{R}$ are scalars, then*

(i) $(A + B) + C = A + (B + C)$ *(the addition of matrices is associative),*

(ii) $A + B = B + A$ *(the addition of matrices is commutative),*

(iii) $A + 0 = A$, *where* 0 *denotes the* zero matrix *(whose entries are all zero),*

(iv) there is an additive inverse for any matrix, namely $A + (-1) \cdot A = 0$ *holds, where* 0 *is the zero matrix again.*

(v) $\lambda(A + B) = \lambda A + \lambda B$,

(vi) $(\lambda + \mu)A = \lambda A + \mu A$,

(vii) $\lambda(\mu A) = (\lambda\mu)A$,

(viii) $1 \cdot A = A$.

This theorem follows easily from the definitions above and from the properties of the operations on real numbers, hence its proof is left to the reader. What makes a difference between the vectors and matrices is that another operation is defined for matrices, namely the multiplication:

Definition 2.5.2. If $A \in \mathbb{R}^{k\times n}$ and $B \in \mathbb{R}^{n\times m}$, then their *product* $C = AB$ is a matrix of size $k \times m$ whose entries are given by

$$c_{i,j} = a_{i,1}b_{1,j} + a_{i,2}b_{2,j} + \cdots + a_{i,n}b_{n,j} \tag{6}$$

for every $1 \le i \le k$ and $1 \le j \le m$.

The product of two matrices is defined only if the number of the columns in the first matrix is the same as the number of the rows in the second one, and the number of the rows of the resulting matrix is the same as for the first matrix, while the result has as many columns as the second matrix. The entry of the result in the ith row and the jth column is a "scalar product" type sum, namely we take the ith row of the first matrix and the jth column of the second one, and then multiply the first entry in the row and the first entry in the column, then we multiply the second entries, and continue this until the last entries. The sum of these products will be the entry of the resulting matrix. Note that if there is only 3 entries, then this is exactly the scalar product of two space vectors, so (6) generalizes this operation. Thus, we will call the expression on the right hand side of (6) the *scalar product* of the ith row of A and the jth column of B.

The following figure helps to visualize how the product of matrices is defined:

$$\begin{pmatrix} b_{1,1} & \dots & b_{1,j} & \dots & b_{1,m} \\ b_{2,1} & \dots & b_{2,j} & \dots & b_{2,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{n,1} & \dots & b_{n,2} & \dots & b_{n,m} \end{pmatrix} = B$$

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \dots & a_{i,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,n} \end{pmatrix} \begin{pmatrix} & \uparrow & \\ & & \\ \leftarrow & c_{i,j} & \\ & & \end{pmatrix} = C = AB.$$

Recall the definition of the *transpose* of a matrix (see Definition 2.4.4). In the following exercise we show some examples for the operations above:

Exercise 2.5.1. Let A and B be the following matrices:

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

Decide if the following operations are defined and if they are, then calculate the result:

a) $3A + 9B$, b) AB, c) BA, d) $BA - 2A$ e) $A^T B^T$.

Solution. a) As $4A \in \mathbb{R}^{2\times 3}$ and $9B \in \mathbb{R}^{2\times 2}$, they are not of the same size and hence their sum is not defined.

b) As A has 3 columns and B has only 2 rows, the operation AB is not defined.

c) The number of the columns of B is the same as the number of rows of A, so the product BA is defined and the result is in $\mathbb{R}^{2\times 3}$:

$$\begin{pmatrix} 2 & -1 & -5 \\ 1 & 4 & -3 \end{pmatrix} = A$$

$$B = \begin{pmatrix} 5 & -4 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} c_{1,1} & c_{1,2} & c_{1,3} \\ c_{2,1} & c_{2,2} & c_{2,3} \end{pmatrix} = C = BA,$$

where

$$\begin{aligned} c_{1,1} &= 5 \cdot 2 + (-4) \cdot 1 = 6, \\ c_{1,2} &= 5 \cdot (-1) + (-4) \cdot 4 = -21, \\ c_{1,3} &= 5 \cdot (-5) + (-4) \cdot (-3) = -13, \\ c_{2,1} &= (-2) \cdot 2 + 3 \cdot 1 = -1, \\ c_{2,2} &= (-2) \cdot (-1) + 3 \cdot 4 = 14, \\ c_{2,3} &= (-2) \cdot (-5) + 3 \cdot (-3) = 1. \end{aligned}$$

Hence the result is

$$BA = \begin{pmatrix} 6 & -21 & -13 \\ -1 & 14 & 1 \end{pmatrix}.$$

d) As BA and $2A$ are both in $\mathbb{R}^{2\times 3}$, we can subtract the second one from the first one:

$$BA - 2A = \begin{pmatrix} 6 & -21 & -13 \\ -1 & 14 & 1 \end{pmatrix} - \begin{pmatrix} 4 & -2 & -10 \\ 2 & 8 & -6 \end{pmatrix}$$

e) As $A^T \in \mathbb{R}^{3\times 2}$ and $B^T \in \mathbb{R}^{2\times 2}$, the product $A^T B^T$ is defined:

$$\begin{pmatrix} 5 & -2 \\ -4 & 3 \end{pmatrix} = B^T$$

$$A^T = \begin{pmatrix} 2 & 1 \\ -1 & 4 \\ -5 & -3 \end{pmatrix} \begin{pmatrix} d_{1,1} & d_{1,2} \\ d_{2,1} & d_{2,2} \\ d_{3,1} & d_{3,2} \end{pmatrix} = D = A^T B^T,$$

Observe that here we have to make the same calculations as in the case of BA, only the order is different. For example $d_{1,1} = 2 \cdot 5 + 1 \cdot (-4) = 6 = c_{1,1}$, $d_{1,2} = 2 \cdot (-2) + 1 \cdot 3 = -1 = c_{2,1}$. In general, the calculation of $d_{i,j}$ and $c_{j,i}$ requires the same operations, so the result will be the transpose of $C = AB$, that is

$$D = A^T B^T = \begin{pmatrix} 6 & -1 \\ -21 & 14 \\ -13 & 1 \end{pmatrix}.$$

□

The connection between the results of part c) and part e) follows from the general statement below:

Theorem 2.5.2. *Let A and B be matrices. Then the operation AB is defined if and only the operation $B^T A^T$ is defined, and in this case $(AB)^T = B^T A^T$.*

Proof. Let A be a matrix of size $k \times n$ (and hence equivalently $A^T \in \mathbb{R}^{n\times k}$), then the product AB is defined if and only if B is of size $n \times m$. But this is equivalent to $B^T \in \mathbb{R}^{m\times n}$ which is equivalent to the existence of the product $B^T A^T$. Hence the first part of the statement is proved.

We set $X = AB$ and $Y = B^T A^T$. Then $x_{i,j}$ is the scalar product of the ith row of A and the jth column of B. As the entries of the jth row of B^T are the same as the entries of the jth column of B, and the same hold for the ith column of A^T and the ith row of A, we get that we make the same computations when calculating the element $y_{j,i}$, hence $x_{i,j} = y_{j,i}$ for every $1 \le i \le k$ and $1 \le j \le m$, i.e. $X^T = Y$. □

Now we turn to the main properties of the matrix multiplication. We have already seen that it cannot be commutative, since it can happen that AB is defined but BA is not (see part b) of the previous exercise). But $AB = BA$ does not hold in general even in the case when both sides are defined. For example, when

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

then AB is the zero matrix but BA is not. But the other properties that are usual for the multiplication of real numbers hold for the matrix multiplication as well:

Theorem 2.5.3. *Assume that A, B and C are matrices and $\lambda \in \mathbb{R}$ is a scalar. Then for each of the following equations its left hand side is defined if and only if its right hand side is defined, and in that case the equations hold:*

(i) $(\lambda A)B = \lambda(AB) = A(\lambda B)$,

(ii) $A(B+C) = AB + AC$ *and* $(B+C)A = BA + CA$ *(distributive law for the matrix operations),*

(iii) $(AB)C = A(BC)$ *(the matrix multiplication is associative).*

Proof. We begin with the proof of (i), and we only prove the first equality, the other one follows similarly. Now $A \in \mathbb{R}^{k\times n}$ if and only if $\lambda A \in \mathbb{R}^{k\times n}$, so the existence of the result on both sides is equivalent to $B \in \mathbb{R}^{n\times m}$. So assume that B is a matrix of size $n \times m$, $X = AB$, $Y = \lambda(AB)$ and $Z = (\lambda A)B$. By the definition of the matrix multiplication we have $x_{i,j} = a_{i,1}b_{1,j} + \cdots + a_{i,n}b_{n,j}$ for every $1 \le i \le k$ and $1 \le j \le m$, and hence $y_{i,j} = \lambda(a_{i,1}b_{1,j} + \cdots + a_{i,n}b_{n,j})$. We also have $z_{i,j} = (\lambda a_{i,1})b_{1,j} + \cdots + (\lambda a_{i,n})b_{n,j}$ from the definition, and then the basic properties of the operations on the real numbers give that $y_{i,j} = z_{i,j}$.

Now we turn to the proof of (ii). Again, we show only the first equality, the proof of the second one is similar. If $A \in \mathbb{R}^{k\times n}$, then both sides are defined if and only if $B, C \in \mathbb{R}^{n\times m}$. So assume that B and C are of size $n \times m$, and set $X = AB$, $Y = AC$ and $Z = A(B + C)$. By definition, we have:

$$\begin{aligned} x_{i,j} &= a_{i,1}b_{1,j} + \cdots + a_{i,n}b_{n,j}, \\ y_{i,j} &= a_{i,1}c_{1,j} + \cdots + a_{i,n}c_{n,j}, \\ z_{i,j} &= a_{i,1}(b_{1,j} + c_{1,j}) + \cdots + a_{i,n}(b_{n,j} + c_{n,j}) \end{aligned}$$

for every $1 \le i \le k$ and $1 \le j \le m$. As $z_{i,j} = x_{i,j} + y_{i,j}$ for every i, j, we get the statement.

Finally, we turn to the proof of (iii). Let $A \in \mathbb{R}^{k\times n}$, then the left hand side is defined if and only if $B \in \mathbb{R}^{n\times m}$ and (as $AB \in \mathbb{R}^{k\times m}$, we must have) $C \in \mathbb{R}^{m\times t}$. But this is also equivalent to existence of the product $BC \in \mathbb{R}^{n\times t}$ and the product $A(BC) \in \mathbb{R}^{k\times t}$. So assume that B is of size $n \times m$ while C is of size $m \times t$, and set $X = AB$ and $Y = XC = (AB)C$. Then

$$x_{i,j} = a_{i,1}b_{1,j} + \cdots + a_{i,n}b_{n,j}$$

for every $1 \le i \le k$ and $1 \le j \le m$, so

$$\begin{aligned} y_{i,j} &= x_{i,1}c_{1,j} + \cdots + x_{i,m}c_{m,j}, \\ &= (a_{i,1}b_{1,1} + \cdots + a_{i,n}b_{n,1})c_{1,j} + (a_{i,1}b_{1,2} + \cdots + a_{i,n}b_{n,2})c_{2,j} \\ &\qquad + \cdots + (a_{i,1}b_{1,m} + \cdots + a_{i,n}b_{n,m})c_{m,j}. \end{aligned}$$

Applying the distributive law for the reals we get that $y_{i,j}$ is the sum of all the products of the form $a_{i,r}b_{r,s}c_{s,j}$ where $1 \le r \le n$ and $1 \le s \le m$. A similar computation shows that the corresponding entry of $A(BC)$ is the same sum. We omit the details of this latter computation. □

The multiplication of the real numbers has another important property. The number 1 has a special role, since $1 \cdot a = a \cdot 1 = a$ holds for every number $a \in \mathbb{R}$. There is an analogue of this property for matrices as well.

Definition 2.5.3. The matrix of size $n \times n$ whose entries in its main diagonal are 1 and all its other entries are 0 is called the *identity matrix* in $\mathbb{R}^{n\times n}$. It is denoted by I_n or (if the size of the matrix is clear from the context, then) simply by I.

Proposition 2.5.4. *If $A \in \mathbb{R}^{k\times n}$, then $I_k A = AI_n = A$.*

Proof. We prove the statement for AI_n, the proof of the another part is similar. So if $C = AI_n$, then by definition we have

$$c_{i,j} = a_{i,1} \cdot 0 + \cdots + a_{i,j-1} \cdot 0 + a_{i,j} \cdot 1 + a_{i,j+1} \cdot 0 + \cdots + a_{i,n} \cdot 0 = a_{i,j},$$

hence the statement follows. □

Finally, the following theorem connects the product of the matrices and the determinant:

Theorem 2.5.5. *If $A, B \in \mathbb{R}^{n\times n}$, then* $\det AB = \det A \cdot \det B$.

For the proof we will use an analogue of Theorem 2.4.4:

Lemma 2.5.6. *Assume that $A, B \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{R}$ is a scalar and $1 \le i, j \le n$, $i \ne j$ are integers.*

(i) If we multiply a row of A or a column of B by λ element-wise, then for the resulting matrices A' and B' we have

$$\det A'B = \det AB' = \lambda \cdot \det AB$$

and

$$\det A' \cdot \det B = \det A \cdot \det B' = \lambda \cdot \det A \cdot \det B.$$

(ii) If we interchange two rows of A or two columns of B, then for the resulting matrices A' and B' we have

$$\det A'B = \det AB' = (-1) \cdot \det AB$$

and

$$\det A' \cdot \det B = \det A \cdot \det B' = (-1) \cdot \det A \cdot \det B.$$

(iii) If we replace the ith row of A (or the ith column of B) by the (element-wise) sum of itself and λ times the jth row of A (λ times the jth column of B), then for the resulting matrices A' and B' we have

$$\det A'B = \det AB' = \det AB$$

and

$$\det A' \cdot \det B = \det A \cdot \det B' = \det A \cdot \det B.$$

Proof. First note that the second statements in (i), (ii) and (iii) are immediate consequences of Theorem 2.4.4, hence it remains to show the first statements. For the proof of (i) observe that if we multiply the ith row of A by λ getting the matrix A', then by the definition of the matrix multiplication we have that $A'B$ is obtained from AB by multiplying the ith row of the latter product by λ. Hence $\det A'B = \lambda \cdot \det AB$ follows from part (i) of Theorem 2.4.4. Similarly, if we multiply the ith column of B by λ producing the matrix B', then the product AB' is obtained from AB by multiplying its ith column by λ, and $\det AB' = \lambda \cdot \det AB$ follows as above. The proofs of the other statements are similar and left to the reader. □

Proof of Theorem 2.5.5 As in the first phase of the Gaussian elimination one can perform the steps described in part (i), (ii) and (iii) of the lemma above on the matrix A to obtain an upper triangular matrix A'. By the lemma we have that $\det A'B = c \cdot \det AB$ and $\det A' \cdot \det B = c \cdot \det A \cdot \det B$ for the same non-zero real number c.

By performing analogous steps on the columns of B one can obtain an upper triangular matrix B' so that $\det A'B' = c' \cdot \det A'B$ and $\det A' \cdot \det B' = c' \cdot \det A' \cdot \det B$ hold for the same non-zero constant c'. This can be done in the following way: we start in the last row of B and swap a non-zero entry in the last place if necessary. Then we eliminate all the other non-zero entries on the left of the last entry in the row. After this we continue in the previous row where all the entries on the left of the main diagonal will be eliminated. If the entry in the main diagonal is 0, then first we interchange this with a non-zero entry on the left of the diagonal. If every entry of a row on the left of the main diagonal is also zero, then of course we can simply continue with the previous row.

Hence

$$\det AB = cc' \cdot \det A'B', \qquad \det A \cdot \det B = cc' \cdot \det A' \cdot \det B',$$

where A' and B' are upper triangular. As c and c' above are non-zero, it remains to show the statement for upper triangular matrices. But for an upper triangular matrix its determinant is the product of the entries in its main diagonal by Theorem 2.4.2. So if the entries of the main diagonal of A' and B' are $a_{1,1}, \ldots, a_{n,n}$ and $b_{1,1} \ldots, b_{n,n}$, respectively, then

$$\det A' \cdot \det B' = a_{1,1} \ldots a_{n,n} b_{1,1} \ldots b_{n,n}.$$

On the other hand, the product $A'B'$ is also an upper triangular matrix, and the entries in its main diagonal are $a_{1,1}b_{1,1}, \ldots, a_{n,n}b_{n,n}$ by the definition of the matrix multiplication, so the product on the left hand side of the last equation is in fact $\det A'B'$. □

2.5.2 Matrix Multiplication and Systems of Linear Equations

Assume that the matrix A and the vector $\underline{b}$ are the following:

$$A = \begin{pmatrix} 2 & -1 & 6 \\ 2 & 2 & 3 \\ 6 & -1 & 17 \\ 4 & -1 & 13 \end{pmatrix}, \qquad \underline{b} = \begin{pmatrix} 12 \\ 24 \\ 46 \\ 32 \end{pmatrix}.$$

Consider the following problem: we are looking for those vectors $\underline{x}$ for which the equality $A\underline{x} = \underline{b}$ holds. That is, we would like to solve this matrix equation. First of all, as A has 3 columns, the numbers of the rows of $\underline{x}$ must be 3 as well. Also, the result is of size 4×1, so the number of the columns of $\underline{x}$ is necessarily 1, hence the equation $A\underline{x} = \underline{b}$ can hold only if $\underline{x} \in \mathbb{R}^{3\times 1}$, i.e. if $\underline{x}$ is a 3-dimensional column vector. Let us write

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

then by the definition of matrix multiplication we have

$$A\underline{x} = \begin{pmatrix} 2x_1 - x_2 + 6x_3 \\ 2x_1 + 2x_2 + 3x_3 \\ 6x_1 - x_2 + 17x_3 \\ 4x_1 - x_2 + 13x_3 \end{pmatrix}.$$

This vector equals to $\underline{b}$, then this is equivalent to the following system (by equating the coordinates of $\underline{b}$ and the product above):

$$\begin{aligned} 2x_1 - \ x_2 + \ 6x_3 &= 12 \\ 2x_1 + 2x_2 + \ 3x_3 &= 24 \\ 6x_1 - \ x_2 + 17x_3 &= 46 \\ 4x_1 - \ x_2 + 13x_3 &= 32 \end{aligned}$$

We have already solved this system in Section 2.3.1 as the first example for the Gaussian elimination, its unique solution is

$$\underline{x} = \begin{pmatrix} 3 \\ 6 \\ 2 \end{pmatrix}.$$

We see that the matrix equation $A\underline{x} = \underline{b}$ above is equivalent to a system of linear equations. These systems occurred also when we examined the spanned subspaces of vectors in $\mathbb{R}^n$. The following theorem describes these connections precisely:

Theorem 2.5.7. *Assume that $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n, \underline{b} \in \mathbb{R}^k$ are vectors and let A be the matrix whose ith column is the vector $\underline{a}_i$ for every $1 \leq i \leq n$ (and hence $A \in \mathbb{R}^{k \times n}$). Then the following are equivalent:*

(i) the matrix equation $A\underline{x} = \underline{b}$ has a solution,

(ii) the system of linear equations given by the augmented coefficient matrix $(A|\underline{b})$ is solvable,

(iii) $\underline{b} \in \operatorname{span}\{\underline{a}_1, \ldots, \underline{a}_n\}$.

Proof. The vector $\underline{b}$ is in the span of the vectors $\underline{a}_1, \ldots, \underline{a}_n$ if and only if $\lambda_1\underline{a}_1 + \cdots + \lambda_n\underline{a}_n = \underline{b}$ holds for some real numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. If the ith coordinate of the vector $\underline{a}_j$ is denoted by $a_{i,j}$, then the ith coordinate of a linear combination $\lambda_1\underline{a}_1 + \ldots, \lambda_n\underline{a}_n$ is $a_{i,1}\lambda_1 + \cdots + a_{i,n}\lambda_n$. Since $\lambda_1\underline{a}_1 + \cdots + \lambda_n\underline{a}_n = \underline{b}$ holds if and only if the coordinates on the right and left hand side are the same, respectively, it is then equivalent to $a_{i,1}\lambda_1 + \cdots + a_{i,n}\lambda_n = b_i$ for every $1 \leq i \leq k$, where b_i is the ith coordinate of $\underline{b}$. But this means exactly that the system of linear equations given by the matrix $(A|\underline{b})$ is solvable, so (ii) and (iii) equivalent to each other.

Now we turn to the equivalence of (i) and (ii). Observe that if $A\underline{x} = \underline{b}$ is solvable, then (as the product on the left hand side exists) $\underline{x} \in \mathbb{R}^{n \times 1}$ must hold, since the number of its rows must be the number of the columns of A while the number of its columns must be the number of the columns of $\underline{b}$. If we denote the jth coordinate of $\underline{x}$ by x_j for every $1 \leq j \leq n$, then the ith coordinate of the product $A\underline{x}$ is $a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n$ by definition. Hence $A\underline{x} = \underline{b}$ is solvable if and only if $\underline{x} \in \mathbb{R}^{n \times 1}$ and $a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n = b_i$ holds for every $1 \leq i \leq k$, that is, if and only if the system given by $(A|\underline{b})$ is solvable. □

Observe that the proof above gives more. The solvability of the equation $A\underline{x} = \underline{b}$ and the system $(A|\underline{b})$ is not just equivalent, but the solutions are basically the same. This means that if $x_1, \ldots, x_n$ is a solution of the system if and only if the vector $\underline{x} = (x_1, \ldots, x_n)^T$ is a solution of $A\underline{x} = \underline{b}$. Also, this holds if and only if the vector $\underline{b}$ can be expressed as the linear combination of the $\underline{a}_i$'s with the scalars $x_1, \ldots, x_n$.

Accordingly, we may use the notation $A\underline{x} = \underline{b}$ for the system given by $(A|\underline{b})$. We also note that the equivalence of (i) and (iii) together with the remark above can be expressed

in the following way: the vector $\underline{x}$ is the solution of the equation $A\underline{x} = \underline{b}$ if and only if $\underline{b}$ can be expressed as the linear combination of the columns of A with the coordinates of $\underline{x}$ as coefficients. This and Theorem 2.2.4 together give the following:

Corollary 2.5.8. *Assume that $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n \in \mathbb{R}^k$ are vectors let A be the matrix whose ith column is the vector $\underline{a}_i$ for every $1 \leq i \leq n$ (and hence $A \in \mathbb{R}^{k\times n}$). Then the following are equivalent:*

(i) the system of linear equations $A\underline{x} = \underline{0}$ has the unique solution $\underline{x} = \underline{0}$,

(ii) the vectors $\underline{a}_1, \ldots, \underline{a}_n$ are linearly independent.

This leads us to an important property of square matrices:

Theorem 2.5.9. *Assume that $A \in \mathbb{R}^{n\times n}$ is a square matrix. Then the following are equivalent:*

(i) the columns of A as vectors in $\mathbb{R}^n$ are linearly independent,

(ii) the rows of A as row vectors of length n are linearly independent,

(iii) $\det A \neq 0$.

Note that we have not defined the linear combinations and linear independence of row vectors, but it can be done in an analogous way as in the case of column vectors. Also, the statement (ii) can be understood so that the transposes of the row vectors (regarded as elements of $\mathbb{R}^n$) are linearly independent.

Proof. By the previous corollary (i) holds if and only if the system $A\underline{x} = \underline{0}$ has a unique solution. By Theorem 2.4.6 this is equivalent to $\det A \neq 0$. By Theorem 2.4.3 (and by the equality $(A^T)^T = A$ which holds for every matrix) this is equivalent to $\det A^T \neq 0$. As we have seen, this holds if and only if the columns of A^T are independent. As the columns of A^T are the transposes of the rows of A, the statement follows. □

Note that for space vectors this means that they are independent if and only if the 3×3 determinant consisting of their coordinates is non-zero. But by Theorem 2.4.10 this is nothing else than the signed volume of the parallelepiped spanned by the vectors. This means that this volume is non-zero if and only if the vectors are independent, i.e. they are not co-planar - which agrees with the natural intuition about the volume.

2.5.3 The Inverse of a Matrix and its Calculation

A system of linear equations can be written in a form $A\underline{x} = \underline{b}$ by Theorem 2.5.7, where A is the coefficient matrix and $\underline{b}$ is the vector whose coordinates are the constants on the right hand sides of the equations. As formally there is a matrix multiplication on the left hand side, it seems a natural question if there is an analogue of the division for matrices, since in that case we could hope for a solution by "dividing both sides by A". It turns out that the answer for this question is (at least in partly) positive. This means that *in some cases* we have this analogue. To understand the following notion correctly we note that the division by a real number a is nothing else than the multiplication by its reciprocal $1/a = a^{-1}$. Now we introduce the corresponding notion for matrices (and use the latter notation to emphasize the similarity):

Definition 2.5.4. Assume that $A \in \mathbb{R}^{n\times n}$, then the matrix $X \in \mathbb{R}^{n\times n}$ is called the *inverse* of A if $AX = I_n = XA$ holds. In this case we use the notation $X = A^{-1}$.

It is important part of the definition that the inverse is defined only for a square matrix. Also, if it exists, then it is unique. Indeed, if $XA = I = AX$ and $YA = I = AY$ hold for the matrices X and Y, then

$$X = XI = X(AY) = (XA)Y = IY = Y$$

by Proposition 2.5.4 and by the associativity of the matrix multiplication. So the notation A^{-1} is justified by the uniqueness, and from now on we can talk about *the* inverse of a matrix - at least if it exists. It is easy to see that there are matrices whose inverse exists, for example $I^{-1} = I$ by Proposition 2.5.4. Unfortunately this is not always the case, but the next theorem gives a complete answer for this question:

Theorem 2.5.10. *The matrix $A \in \mathbb{R}^{n\times n}$ has an inverse if and only if* $\det A \neq 0$.

Proof. Assume first that A^{-1} exists. It follows easily from the definition of the determinant (or by part (ii) of Theorem 2.4.2) that $\det I_n = 1$ for every n. Then by Theorem 2.5.5 we have $1 = \det I_n = \det(AA^{-1}) = \det A \cdot \det A^{-1}$, and hence $\det A \neq 0$ must hold.

For the other direction we need the following lemma:

Lemma 2.5.11. *If $A \in \mathbb{R}^{n\times n}$ and* $\det A \neq 0$*, then there exists a unique matrix $X \in \mathbb{R}^{n\times n}$ for which $AX = I_n$ holds.*

Proof. If $AX = I_n$ holds, then of course X must be of size $n \times n$. So let $\underline{x}_1, \ldots, \underline{x}_n \in \mathbb{R}^n$ be the columns of the matrix X. Observe that the ith column of AX is $A\underline{x}_i$ by the definition of the matrix multiplication. Hence the equation $AX = I_n$ holds if and only if the equation $A\underline{x}_i = \underline{e}_i$ holds for every $1 \leq i \leq n$, where $\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n$ are the columns of I_n, i.e. the vectors of the standard basis of $\mathbb{R}^n$ (see page 46). Since $\det A \neq 0$ by our assumption, each of these equations has a unique solution by Theorem 2.4.6, and the statement follows. □

Now we return to the proof of the theorem. By the lemma there is a unique matrix X for which $AX = I$ holds. We will show that in this case $XA = I$ holds as well. By Theorem 2.5.5 we have $1 = \det I = \det(AX) = \det A \cdot \det X$, so $\det X \neq 0$ holds and hence the lemma above is applicable for X as well. Thus, there is a unique matrix Y for which $XY = I$ holds. Now Proposition 2.5.4 and the associativity of the matrix multiplication give

$$Y = IY = (AX)Y = A(XY) = AI = A,$$

hence $XA = I$ and the theorem follows. □

Now if a system of linear equations if given by the matrix equation $A\underline{x} = \underline{b}$, where $A \in \mathbb{R}^{n\times n}$ (which means that the number of equations is the same as the number of variables), and $\det A \neq 0$ also holds, then we can multiply the equation by A^{-1} from the left to obtain

$$\underline{x} = I_n\underline{x} = (A^{-1}A)\underline{x} = A^{-1}(A\underline{x}) = A^{-1}\underline{b}, \tag{7}$$

so the unique solution of the system is $A^{-1}\underline{b}$. This means that the system can be solved by a matrix multiplication if the matrix A^{-1} is known. But observe that the proof above gives a method also for the computation of the inverse (if $\det A \neq 0$). By its last paragraph it is enough to determine the matrix X for which $AX = I$ holds, it will be automatically the

inverse of A. By the proof of Lemma 2.5.11 this can be done by solving the systems given by $A\underline{x} = \underline{e}_i$ for every $1 \leq i \leq n$ using Gaussian elimination (for example).

An advantage of this method is that it is not necessary to run the Gaussian elimination n times, the systems can be solved simultaneously. Indeed, the coefficient matrices of the systems agree, so the algorithm makes the same steps in all cases, we only have to keep track of the changes of all vectors on the right hand sides of the equations. That is, we write down the matrix $(A|\,\underline{e}_1\,\underline{e}_2\,\ldots\,\underline{e}_n)$ and run the Gaussian elimination for this matrix (note that the steps are determined only by the coefficient matrix A). If the determinant is 0, then at some point we get 0 in the main diagonal so that all the entries in its column are 0 below it and we can stop (just like in line 16 of the algorithm for the calculation of the determinant, see page 67). Otherwise, after the algorithm stops, we obtain the reduced row echelon form on the left side of the vertical line, which is the matrix I_n in this case. So the result will be of the form $(I_n|\,\underline{x}_1\,\ldots\,\underline{x}_n)$ where $\underline{x}_i$ is the solution of the system $A\underline{x} = \underline{e}_i$, that is, on the right side of the line we get the inverse of A.

We demonstrate this method by an example. Let us calculate the inverse of the matrix

$$A = \begin{pmatrix} 1 & -3 & 7 \\ -1 & 3 & -6 \\ 2 & -5 & 12 \end{pmatrix}.$$

We are going to run the Gaussian elimination for the matrix $(A|I_3)$:

$$\left(\begin{array}{ccc|ccc} 1 & -3 & 7 & 1 & 0 & 0 \\ -1 & 3 & -6 & 0 & 1 & 0 \\ 2 & -5 & 12 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & -3 & 7 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & -2 & -2 & 0 & 1 \end{array}\right) \sim$$

The leading coefficient in the first row is already 1 in the beginning, so in the first step we eliminate the non-zero elements below it: we add the first row to the second one and subtract 2 times the first row from the third one. After these steps the second entry of the second row is zero, but we can swap the second and the third row to obtain a non-zero entry in the main diagonal:

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

The arising matrix is already of row echelon form, so we turn to the second phase of the elimination. We eliminate the non-zero entries above the leading coefficient in the last row by adding 2 times the last row to the second one and subtracting 7 times the last row from the first one. Finally, we add 3 times the second row to the first one:

$$\sim \left(\begin{array}{ccc|ccc} 1 & -3 & 0 & -6 & -7 & 0 \\ 0 & 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -6 & -1 & 3 \\ 0 & 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right).$$

At this point the algorithm stops and the columns on the right side of the vertical line form the inverse of A:

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

Now assume that we want to solve the system

$$\begin{aligned} x_1 - 3x_2 + 7x_3 &= p, \\ -x_1 + 3x_2 - 6x_3 &= q, \\ 2x_1 - 5x_2 + 12x_3 &= r. \end{aligned}$$

This is equivalent to the equation $A\underline{x} = \underline{b}$ where A is the matrix above and $\underline{b} = (p, q, r)^T$. As $\det A \neq 0$ and hence its inverse exists, we get by (7) that

$$\underline{x} = A^{-1}\underline{b} = \begin{pmatrix} -6p - q + 3r \\ 2q + r \\ p + q \end{pmatrix},$$

so the unique solution of the system is $x_1 = -6p - q + 3r$, $x_2 = 2q + r$ and $x_3 = p + q$.

This algorithm above works well in practice, but now we also give a formula for the inverse which is often useful in theoretical arguments. If $A \in \mathbb{R}^{n\times n}$, then let $\hat{A} \in \mathbb{R}^{n\times n}$ be the matrix whose entry $\hat{a}_{i,j}$ in the ith row and in the jth column is the cofactor $C_{j,i}$ assigned to the entry $a_{j,i}$ of A (given in Definition 2.4.5). Note that the indeces i, j are swapped in the definition of $\hat{a}_{i,j}$. Then Theorem 2.4.7 and Corollary 2.4.8 together give that

$$A\hat{A} = \begin{pmatrix} \det A & 0 & \dots & 0 \\ 0 & \det A & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \det A \end{pmatrix}, \tag{8}$$

so we have the following:

Theorem 2.5.12. *If $A \in \mathbb{R}^{n\times n}$ and $\det A \neq 0$, then*

$$A^{-1} = \frac{1}{\det A}\hat{A}. \tag{9}$$

In general it is rather tiresome to calculate $\hat{A}$ based on its definition, but for $n = 2$ it is in fact very easy, since the minors of the matrix are 1×1 determinants whose values are just their single entry. So if

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

whose determinant (given explicitly in (5)) is non-zero, then

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} \det M_{1,1} & -\det M_{2,1} \\ -\det M_{1,2} & \det M_{2,2} \end{pmatrix} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{10}$$

We can use the theorem above to give an exact formula for the unique solution of a system $A\underline{x} = \underline{b}$, where $A \in \mathbb{R}^{n\times n}$ and $\det A \neq 0$. As we have seen in (7), this is

$$A^{-1}\underline{b} = \frac{1}{\det A}\hat{A}\underline{b},$$

so the value of the variable x_i is

$$\frac{1}{\det A}\sum_{k=1}^{n} C_{k,i}b_k.$$

If B_i is the matrix which is obtained from A when replacing its ith column by $\underline{b}$, then the latter sum is just the determinant of B_i expanded along the ith column, so we have

Theorem 2.5.13 (Cramer's rule)**.** *Assume that the system of linear equations is given by* $A\underline{x} = \underline{b}$*, where* $A \in \mathbb{R}^{n\times n}$ *and* $\det A \neq 0$*. Then the unique solution of the system is*

$$x_i = \frac{\det B_i}{\det A} \qquad (1 \leq i \leq n),$$

where B_i *is the matrix obtained from* A *when replacing its* i*th column by the vector* $\underline{b}$*.*

As in the case of Theorem 2.5.12 we have to warn the reader that this formula is not practical for the calculation of the solution in general. But for $n = 2$ it is simple enough to take a closer look at it. So if we have the system

$$a_{1,1}x_1 + a_{1,2}x_2 = b_1,$$
$$a_{2,1}x_1 + a_{2,2}x_2 = b_2,$$

then

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}, \qquad B_1 = \begin{pmatrix} b_1 & a_{1,2} \\ b_2 & a_{2,2} \end{pmatrix}, \qquad B_2 = \begin{pmatrix} a_{1,1} & b_1 \\ a_{2,1} & b_2 \end{pmatrix},$$

and if $\det A = a_{1,1}a_{2,2} - a_{1,2}a_{2,1} \neq 0$, then the unique solution is given by

$$x_1 = \frac{\det B_1}{\det A} = \frac{a_{2,2}b_1 - a_{1,2}b_2}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}, \qquad x_2 = \frac{\det B_2}{\det A} = \frac{a_{1,1}b_2 - a_{2,1}b_1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}.$$

Recall that these are the same formulae that were given at the beginning of Section 2.4.

Finally, we show another quick application of Theorem 2.5.12. Note that if the entries of the matrix A are rational numbers, then so are the entries of $\hat{A}$ (because its entries are determinants with rational entries multiplied by ± 1, so every operation that we make by the calculation of $\hat{A}$ gives a rational result). Also, the determinant of A is rational, so (9) gives that A^{-1} has rational entries. In fact this follows also from the algorithm above for the calculation of the inverse. But the formula in (9) gives even more when we repeat this argument with integer entries. Namely, if the entries of A are integers, then so are the entries of $\hat{A}$, so we immediately get

Corollary 2.5.14. *Assume that* $A \in \mathbb{R}^{n\times n}$ *and the entries of* A *are integers. If* $\det A = \pm 1$*, then the entries of* A^{-1} *are also integers.*

This is a basic (and very important) fact in number theory, but we do not go into that direction.

2.5.4 The Rank of a Matrix

The columns of a matrix A of size $k \times n$ can be regarded as a system of vectors in $\mathbb{R}^k$. Also, the rows of this matrix constitute a system of row vectors of length n. At first sight one may see no connection between these two systems. But Theorem 2.5.9 tells us that in the special case when A is a square matrix, its columns are independent if and only if its rows are independent, and both of these are equivalent to $\det A \neq 0$. In this section we generalize this result for an arbitrary matrix.

Definition 2.5.5. Assume that $A \in \mathbb{R}^{k\times n}$ and $r \leq \min\{k, n\}$. An $(r \times r)$ *square sub-matrix* of A is formed by the common entries of r arbitrary columns and r arbitrary rows of A. An $r \times r$ *minor* of A (or a *minor determinant of order* r) is the determinant of an $r \times r$ square sub-matrix of A. A minor of order 0 is defined to be 1. If $k = n$ (i.e. A is a square matrix), then an $r \times r$ minor is also called an $(n - r)$*th minor*. The zeroth minor is then the determinant of A.

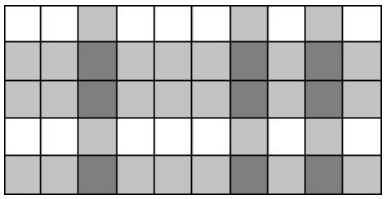

Figure 1: A sub-matrix of size 3×3 is formed by the common entries of 3 rows and 3 columns

Figure 1 illustrates the choice of a 3×3 sub-matrix M of a matrix A of size 5×10, where the 2nd, 3rd and 5th rows and the 3rd, 7th and 9th columns are chosen, so the entries of M are the following:

$$M = \begin{pmatrix} a_{2,3} & a_{2,7} & a_{2,9} \\ a_{3,3} & a_{3,7} & a_{3,9} \\ a_{5,3} & a_{5,7} & a_{5,9} \end{pmatrix}.$$

Definition 2.5.6. Assume that $A \in \mathbb{R}^{k \times n}$.

(i) The *column rank* of A is r if r columns of A can be chosen so that they are linearly independent (as vectors in $\mathbb{R}^k$), but any system of $r+1$ columns of A is linearly dependent. We use the notation $r_c(A)$ for the column rank of A.

(ii) The *row rank* of A is r if r rows of A can be chosen so that they are linearly independent (as row vectors of length n), but any system of $r+1$ rows of A is linearly dependent. We use the notation $r_r(A)$ for the row rank of A.

(iii) The *determinantal rank* of A is r if A has non-zero $r \times r$ minor, but every minor of A of order $r+1$ is zero. We use the notation $r_d(A)$ for the determinantal rank of A.

Note that the column, row and determinantal rank is determined uniquely by the previous definitions. Indeed, if there is no linearly independent system which consits of $r+1$ column vectors of A, then more than $r+1$ columns of A cannot be independent, since in that case any $r+1$ vectors of that system would be independent. So $r_c(A)$ is the maximal integer r for which there exists a linearly independent system of r column vectors of A. Similarly, $r_r(A)$ is the maximal number r for which there exists an independent system of r row vectors of A.

Turning to the determinantal rank, we show by induction that if every minor of A of order $r+1$ is zero, then so is every minor of order $r+n$ for every integer $n \geq 1$. The statement holds for $n = 1$ by assumption. It remains to show that if $n > 1$ and the statement holds for $n-1$, then it is also true for n. But this follows from the Laplace expansion, since a minor of order $r+n$, i.e. the determinant of an $(r+n) \times (r+n)$ sub-matrix M of A can be written as a sum

$$\sum_{k=1}^{n+r} (-1)^{k+1} a_k \det M_k$$

when expanded along its first row (for example). Here M_k is a $(r+n-1) \times (r+n-1)$ sub-matrix of M and hence also a sub-matrix of A, so $\det M_k$ is an $(r+n-1) \times (r+n-1)$ minor of A and hence it is zero for every k by assumption. Hence $r_d(A)$ is the maximal integer r for which A has a non-zero minor of order r.

For the zero matrix 0 we have $r_c(0) = r_r(0) = 0$, since any column or row of it is the zero vector which is itself dependent (while the empty set of vectors is independent by definition).

Also, among the minors of the zero matrix only the minor of order 0 is different from zero (by definition), so $r_d(0) = 0$ holds as well.

Now we compute these values for the matrix

$$A = \begin{pmatrix} 1 & 3 & 5 & 7 \\ 9 & 11 & 13 & 15 \\ 17 & 19 & 21 & 23 \end{pmatrix}.$$

Let us denote the ith row of A by $\underline{r}_i$ while the jth column by $\underline{c}_j$ ($1 \leq i \leq 3$ and $1 \leq j \leq 4$). Now for example $\underline{c}_1$ and $\underline{c}_2$ are independent, because they are not the scalar multiple of each other, but

$$\begin{aligned} \underline{c}_1 - 2\underline{c}_2 + \underline{c}_3 &= \underline{0}, \\ 2\underline{c}_1 - 3\underline{c}_2 + \underline{c}_4 &= \underline{0}, \\ \underline{c}_1 - 3\underline{c}_3 + 2\underline{c}_4 &= \underline{0}, \\ \underline{c}_2 - 2\underline{c}_3 + \underline{c}_4 &= \underline{0}, \end{aligned} \tag{11}$$

and hence any 3columns of A are dependent, so $r_c(A) = 2$. Similarly, $\underline{r}_1$ and $\underline{r}_2$ are independent, but $\underline{r}_1 - 2\underline{r}_2 + \underline{r}_3 = \underline{0}$ shows that the 3 rows together are dependent, and then $r_r(A) = 2$.

Finally, the first and the last rows and columns determine the minor

$$\begin{vmatrix} 1 & 7 \\ 17 & 23 \end{vmatrix} = 1 \cdot 23 - 17 \cdot 7 = -96 \neq 0,$$

but if we chose 3 columns and (all the) 3 rows, then the corresponding minor will be zero. This follows easily from the equations in (11), since they show that if an appropriate scalar multiples of the second and third chosen columns are added to the first one, then the first column will be identically zero, and then the determinant is zero as well (by part (i) of Theorem 2.4.2). This gives that $r_d(A) = 2$.

The values of the 3 different ranks of A were the same in the previous example. The following statement shows that this is not a coincidence:

Theorem 2.5.15. *For every matrix $A \in \mathbb{R}^{k \times n}$ we have $r_c(A) = r_r(A) = r_d(A)$.*

Proof. If A is the zero matrix, then all of the three types of the rank are zero. So we can assume that A is non-zero, and hence $r_c(A)$, $r_r(A)$ and $r_d(A)$ are positive integers.

First we show that $r_c(A) \geq r_d(A)$. Assume that $r_d(A) = r$, it is then enough to choose r columns of A so that they are linearly independent. As $r_d(A) = r$, there is a non-zero minor of A of order r. That is, A has a sub-matrix M of size $r \times r$ so that $\det M \neq 0$. Let A_M denote the sub-matrix of A formed by the r columns of A that are chosen by the construction of the sub-matrix M. We show that the columns of A_M are independent. Assume that a linear combination of them with the coefficients $x_1, \ldots, x_r$ gives the zero vector. Then $A_M\underline{x} = \underline{0}$ for the vector $\underline{x} = (x_1, \ldots, x_r)^T$. This matrix equation encodes a system of linear equations with the coefficient matrix A_M where the constants on the right hand sides are all zero. Let us omit some equations from this system (which means the omission of some rows of A_M). Namely, we *keep* only those equations that belong to the rows that are chosen by the construction of M, i.e. we keep only these rows of A_M and hence the resulting coefficient matrix of this new system is M itself. The remaining equations still hold for $x_1, \ldots, x_n$, so

$M\underline{x} = \underline{0}$ follows. This gives that the linear combination of the columns of M with the scalars $x_1, \dots, x_r$ is zero. As $\det M \neq 0$, we have by Theorem 2.5.9 that the columns of M are independent, hence $x_1 = \cdots = x_r = 0$ must hold by Theorem 2.2.4. We conclude that a linear combination of the columns of A_M gives the zero vector if only if every coefficient is zero, i.e. they are linear independent by Theorem 2.2.4 again.

For the proof of the inequality $r_c(A) \leq r_d(A)$ we are going to use the following

Lemma 2.5.16. *Assume that the columns of a matrix $C \in \mathbb{R}^{k\times n}$ (as vectors in $\mathbb{R}^k$) are linearly independent. If $k > n$, then there is a row of C which can be omitted so that the columns of the resulting matrix $C' \in \mathbb{R}^{(k-1)\times n}$ are still linearly independent.*

Proof. Let us denote the columns of C by $\underline{c}_1, \dots, \underline{c}_n$. If $W = \operatorname{span}\{\underline{c}_1, \dots, \underline{c}_n\}$, then there is a generating system of size n in W. Now if $k > n$, then we cannot find k independent vectors in W by the I-G inequality (Theorem 2.2.5). So there is a vector among the vectors of the standard basis of $\mathbb{R}^k$ (i.e. among the columns of the identity matrix I_k) that is not in W. Assume that the vector $\underline{e}_j$ (i.e. the vector whose jth coordinate is 1 and all the others are zero) has this property ($1 \leq j \leq k$). We show that we can omit the jth row of C so that the columns of the resulting matrix C' are independent.

Assume to the contrary that the columns of C' are dependent and hence the equation $C'\underline{x} = \underline{0}$ for some $\underline{x} \neq \underline{0}$. Than $C\underline{x} \neq \underline{0}$, because the columns of C are independent. But $C\underline{x}$ is is obtained by inserting the scalar product α of the jth row of C and $\underline{x}$ in $C'\underline{x}$ after the $(j-1)$th coordinate, hence $\alpha \neq 0$ and $C\underline{x} = \alpha\underline{e}_j$, i.e. $\alpha^{-1}C\underline{x} = C(\alpha^{-1}\underline{x}) = \underline{e}_j$ contradicting $\underline{e}_j \notin W$. □

Now we turn to the proof of $r_c(A) \leq r_d(A)$. Assume that $r_c(A) = r$, and the columns $\underline{c}_1, \dots, \underline{c}_r$ of A are linearly independent. As $A \in \mathbb{R}^{k\times n}$, the vectors $\underline{c}_j$ are in $\mathbb{R}^k$, so we must have $k \geq r$ by the I-G inequality, since there is a generating system in $\mathbb{R}^k$ which consists of k vectors. Let C be the matrix whose jth column is $\underline{c}_j$. If $k > r$, then we can omit a row of C getting the matrix $C' \in \mathbb{R}^{(k-1)\times r}$ so that its columns are still independent. If $k - 1 > r$, then we can continue this process, and after $(k-r)$ steps we get a matrix $M \in \mathbb{R}^{r\times r}$ so that its columns are linearly independent. By Theorem 2.5.9 we have $\det M \neq 0$, and hence $r_d(A) \geq r$.

We have proved that $r_c(A) \leq r_d(A)$ and $r_c(A) \geq r_d(A)$, i.e. $r_c(A) = r_d(A)$ holds for any matrix A. As the rows of A are the columns of A^T, we get $r_r(A) = r_c(A^T) = r_d(A^T)$ by the previous paragraphs. Since the square sub-matrices of A^T are the transposes of the square sub-matrices of A and hence by Theorem 2.4.3 the minors of A^T are the same as the minors of A, we obtain that the maximal order of the non-zero minors (i.e. the determinantal rank) is the same for A and A^T. This means that $r_r(A) = r_d(A^T) = r_d(A)$, and the proof of the theorem is complete. □

Definition 2.5.7. If $A \in \mathbb{R}^{k\times n}$, then the common value of $r_c(A)$, $r_r(A)$ and $r_d(A)$ is called the *rank* of A. It is denoted by $\operatorname{rk}(A)$ or $\operatorname{rank}(A)$.

Theorem 2.5.17. *Assume that $A \in \mathbb{R}^{k\times n}$ and let us denote its jth column by $\underline{a}_j$ for any $1 \leq j \leq n$. Then* $\operatorname{rank}(A)$ *is the dimension of* $\operatorname{span}\{\underline{a}_1, \dots, \underline{a}_n\}$.

Proof. If $\operatorname{rank}(A) = r$, then we can choose r vectors from the column vectors of A so that they are independent, but any $r+1$ of them are dependent. After a possible renumbering we may assume that $\underline{a}_1, \dots, \underline{a}_r$ are the chosen vectors. It is enough to show that these form a basis in the subspace $W = \operatorname{span}\{\underline{a}_1, \dots, \underline{a}_n\}$.

The vectors $\underline{a}_1, \ldots, \underline{a}_r$ are independent by our choice, so it remains to show that they span W, that is, if $U = \operatorname{span}\{\underline{a}_1, \ldots, \underline{a}_r\}$, then $U = W$. It is obvious that $U \subset W$ since every linear combination of $\underline{a}_1, \ldots, \underline{a}_r$ is also a linear combination of $\underline{a}_1, \ldots, \underline{a}_n$. Hence we need to show that $W \subset U$. For every $r < i \leq n$ the system $\underline{a}_1, \ldots, \underline{a}_r, \underline{a}_i$ is dependent (by the definition of the rank r) and hence by Lemma 2.2.6 we have $\underline{a}_i \in U$. But $\underline{a}_1, \ldots, \underline{a}_r \in U$ holds as well, because they span U, so we obtain that $\underline{a}_i \in U$ for every $1 \leq i \leq n$. But U is a subspace of $\mathbb{R}^k$, so it is closed under addition and scalar multiplication, so every element of W is in U and we are done. □

Exercise 2.5.2. Assume that A and B are matrices and the product AB is defined. Show that $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$.

The Computation of the Rank

Now we give an effective algorithm for the computation of the rank. As in many cases before, a version of the Gaussian elimination is applicable here. We will apply its steps for an arbitrary matrix (instead of an augmented coefficient matrix), and the following proposition tells us that these steps does not change the rank.

Proposition 2.5.18. *The elementary row operations (see Definition 2.3.1) does not change the rank of a matrix.*

Proof. Assume that $\underline{c}_1, \ldots, \underline{c}_m$ are some of the columns of a matrix A, and they form the sub-matrix A' of A. By Corollary 2.5.8 the columns of A' are independent if and only if the system $A'\underline{x} = \underline{0}$ has the unique solution $\underline{x} = \underline{0}$. Assume that we apply an elementary row operation on A. Parallel to this, let us apply the same operation on the augmented coefficient matrix $(A'|\underline{0})$. By Proposition 2.3.1 this does not change the set of the solutions of the system $(A'|\underline{0})$, so the solution of the original system is unique if and only if the resulting system has a unique solution, or equivalently, if and only if the columns of the resulting coefficient matrix are independent (this last equivalence follows from Corollary 2.5.8 again). But the columns of the resulting coefficient matrix are the same as the columns that are obtained from $\underline{c}_1, \ldots, \underline{c}_m$ after the row operation on the matrix A (because A' is formed by the columns $\underline{c}_1, \ldots \underline{c}_m$), let us denote them by $\underline{d}_1, \ldots, \underline{d}_m$. We get that $\underline{c}_1, \ldots, \underline{c}_m$ are independent if and only if $\underline{d}_1, \ldots, \underline{d}_m$ are, so the column rank is the same before and after the application of the operation, and the statement follows. □

Proposition 2.5.19. *If a matrix is of row echelon form, then its rank is the number of its rows.*

Proof. Assume that the matrix $A \in \mathbb{R}^{k \times n}$ is of row echelon form. As every row of it contains a leading coefficient which is 1, and any two of them are in different columns, we get that the matrix has at least as many columns as rows, that is, $k \leq n$. We are going to show that $\operatorname{rank}(A) = k$. Let us examine the sub-matrix M of size $k \times k$ which is obtained by the common entries of all of the rows of A and the columns which contain a leading coefficient. Let us denote these columns by $\underline{c}_1, \ldots, \underline{c}_k$ in order. Since A is of row echelon form, the ith coordinate of $\underline{c}_i$ is 1 while its jth coordinate is zero for all $i < j \leq k$. That is, M is an upper triangular matrix and all of its entries in the main diagonal are 1. Hence $\det M = 1 \neq 0$, so $r_d(A) \geq k$. But since there is only k rows of A, we have $k \geq r_r(A) = \operatorname{rank}(A) = r_d(A) \geq k$, and then $\operatorname{rank}(A) = k$. □

It is now easy to construct an algorithm for the calculation of the rank. We have already seen in Section 2.3.2 that the row echelon form can be reached by elementary row operations. Namely, we can simply run the first phase of the Gaussian elimination for the matrix with some modifications, as in this case there are no associated equations and hence we do not have to keep track of the changes of the right hand sides, accordingly, we cannot obtain a forbidden row. However, we still can get identically zero rows which can be omitted (this is also an elementary row operation). In other words, we apply the algorithm given by the a code on page 57 without the lines $23 - 25$. The resulting matrix has the same rank as the original one by Proposition 2.5.18, and this rank is the number of the rows of the result by the last proposition above.

Note that

$$\operatorname{rank}(A) = r_r(A) = r_c(A^T) = \operatorname{rank}(A^T). \tag{12}$$

So if we apply an elementary operation on columns instead of rows, then this means the same as the application of the corresponding operation for the rows of A^T and then taking the transpose again, hence it does not change the rank of the matrix. To be precise, let $T_c(A)$ be the result of an elementary operation on the columns of a matrix A while we define $T_r(A)$ to be the result of the corresponding operation on the rows. Then

$$\operatorname{rank}(T_c(A)) = \operatorname{rank}([T_r(A^T)]^T) = \operatorname{rank}(T_r(A^T)) = \operatorname{rank}(A^T) = \operatorname{rank}(A).$$

Here the first equality means that if we apply the operation on the columns, then we get the same matrix as if we apply the corresponding row operation on the transpose and then transpose the result back. The second equality follows from our observation (12) above, while the next one is just the application of Proposition 2.5.18 for A^T, and finally we apply (12) again. Hence Proposition 2.5.18 holds for also for elementary column operations. This often makes the calculation easier in practice. We repeat the statement in

Corollary 2.5.20. *The elementary column operations does not change the rank of a matrix.*

Finally, we address the following problem: given a matrix A, we are looking for a maximal independent set of the column vectors of it. By the proof of Proposition 2.5.18 we get that a set of the columns of A is independent if and only if it is independent after the application of an elementary row operation. This means that if we choose a maximal independent set of columns from the resulting matrix after the first phase of the Gaussian elimination, then the corresponding columns of A form a maximal independent subset of the columns of A.

We show that if A is of row echelon form, then the columns that contain a leading coefficient form a maximal independent set of column vectors of A. This, together with the previous paragraph gives an algorithm for our task. First of all, if A has k rows, than $\operatorname{rank}(A) = k$ by Proposition 2.5.19. As the number of the leading coefficient is the same as the number of rows (i.e. k) and any two of them are in different columns, the number of the columns that contain a leading coefficient is k. Hence, it remains to show that they are independent, since then they are automatically maximal among the independent sets of columns. But as in the proof of Proposition 2.5.19, we get that they form an upper triangular matrix whose determinant is non-zero (it is in fact 1), and hence they are independent by Theorem 2.5.9.

Note that this method works only if we apply elementary *row* operations exclusively. The corresponding operations on the columns - although they leave the rank of the matrix unchanged - may change the independence of some systems of the columns.

2.6 Linear Maps

In many parts of mathematics and its applications we have to handle functions that assign a vector to another one. The notation $f : \mathbb{R}^n \to \mathbb{R}^k$ means that f is a function whose domain is (the whole set) $\mathbb{R}^n$ and its range is a subset of $\mathbb{R}^k$ (so maybe not the whole set). Among these, the so-called linear functions are particularly significant. They appear for example by the study of geometric transformations and also in multivariable calculus, namely, they can be used by the local approximation of more complex functions - as the tangential line (if it exists) can be used for the local approximation of a function $f : \mathbb{R} \to \mathbb{R}$. The study of linear functions is one of the most important parts of linear algebra.

2.6.1 Basic Properties and Examples

Definition 2.6.1. A function $f : \mathbb{R}^n \to \mathbb{R}^k$ is called a *linear map* if the following hold for any $\underline{x}, \underline{y} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$:

(i) $f(\underline{x} + \underline{y}) = f(\underline{x}) + f(\underline{y})$ (f is *additive*),

(ii) $f(\lambda \underline{x}) = \lambda f(\underline{x})$ (f is *homogeneous* (of degree 1)).

Note that the additions on the left and right hand sides of (i) are different. On the left, we add two vectors in $\mathbb{R}^n$ and apply the function f for the sum, while on the right we apply the addition in $\mathbb{R}^k$ for the *images* of the two vectors $\underline{x}$ and $\underline{y}$. Also, the scalar multiplication on the left hand side of (ii) is an operation in $\mathbb{R}^n$, while on the right hand side it is an operation in $\mathbb{R}^k$.

Examples.

1. The function $f_1 : \mathbb{R}^3 \to \mathbb{R}^2$ defined by

$$f_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x - y \\ x + z \end{pmatrix}$$

is linear. Indeed,

$$f_1 \left(\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} \right) = f_1 \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \\ z_1 + z_2 \end{pmatrix} = \begin{pmatrix} (x_1 + x_2) - (y_1 + y_2) \\ (x_1 + x_2) + (z_1 + z_2) \end{pmatrix}$$

$$= \begin{pmatrix} x_1 - y_1 \\ x_1 + z_1 \end{pmatrix} + \begin{pmatrix} x_2 - y_2 \\ x_2 + z_2 \end{pmatrix} = f_1 \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + f_1 \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix},$$

so f_1 is additive. Similarly, if $\lambda \in \mathbb{R}$, then

$$f_1 \left(\lambda \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right) = f_1 \begin{pmatrix} \lambda x \\ \lambda y \\ \lambda z \end{pmatrix} = \begin{pmatrix} \lambda x - \lambda y \\ \lambda x + \lambda z \end{pmatrix} = \lambda \begin{pmatrix} x - y \\ x + z \end{pmatrix} = \lambda f_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

so f_1 is homogeneous, and hence linear.

2. Let $f_2 : \mathbb{R}^2 \to \mathbb{R}^2$ be the function which assigns to every plane vector its (orthogonal) projection to the x axis. It is easy to give a formula for f_2:

$$f_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix}.$$

   The linearity of f_2 follows from this formula like in the previous case.

3. Let $f_3 : \mathbb{R}^2 \to \mathbb{R}^2$ be the rotation about the origin by the angle α, then f_3 is a linear map. Indeed, as the sum of two non-zero vectors is the vertex of the (maybe degenerate) parallelogram spanned by them which is different from the origin and the endpoints of the vectors, and the rotation takes the spanned parallelogram into the parallelogram which is spanned by the rotated vectors, we get that the image of the sum is the sum of the images (this holds obviously if at least one of the vectors is the zero vector), hence the additivity of f_3 follows. Also, the application of a dilation and a rotation after that gives the same result as the application of these transformations in reverse order, so f_3 is homogeneous and hence linear. We will give a formula for f_3 later.

4. Let $f_4 : \mathbb{R}^2 \to \mathbb{R}^2$ be the reflection in a line going through the origin. As in the case of the rotation above, it is easy to see that f_4 is linear.

5. Let $f_5 : \mathbb{R}^n \to \mathbb{R}^k$ be the identically zero map. Since

$$f_5(\underline{x}) + f_5(\underline{y}) = \underline{0} + \underline{0} = \underline{0} = f_5(\underline{x} + \underline{y}) \qquad \text{and} \qquad \lambda f_5(\underline{x}) = \lambda\underline{0} = \underline{0} = f_5(\lambda\underline{x})$$

   hold for every $\underline{x}, \underline{y} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we get that f_5 is linear.

6. Let $f_6 : \mathbb{R}^n \to \mathbb{R}^n$ be the identity map which maps every vector in $\mathbb{R}^n$ to itself. Then f_6 is obviously linear.

Here are some basic properties of a linear map:

Proposition 2.6.1. *Let $f : \mathbb{R}^n \to \mathbb{R}^k$ be a linear map. Then the following hold:*

(i) if $\underline{0}_1$ is the zero vector in $\mathbb{R}^n$ and $\underline{0}_2$ is the zero vector in $\mathbb{R}^k$, then $f(\underline{0}_1) = \underline{0}_2$,

(ii) $f(\lambda_1\underline{x}_1 + \cdots + \lambda_m\underline{x}_m) = \lambda_1 f(\underline{x}_1) + \cdots + \lambda_m f(\underline{x}_m)$ for any vectors $\underline{x}_1, \ldots, \underline{x}_m \in \mathbb{R}^n$ and for any scalars $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$.

Proof. We have $f(\underline{0}_1) = f(\underline{0}_1 + \underline{0}_1) = f(\underline{0}_1) + f(\underline{0}_1)$, because f is additive. Adding $-f(\underline{0}_1) = (-1) \cdot f(\underline{0}_1)$ to both sides we get (i). By the repeated application of the additivity and the homogeneity of f we get the second statement immediately. □

It follows from property (i) above that in the case of the rotation f_3 in the examples above it is important that we rotate about the origin. Rotation about any other point of the plane does not fix the origin and hence it is not linear. Similarly, if we reflect in a line which does not go through the origin, then this transformation is not linear.

We assign two important sets to every linear map:

Definition 2.6.2. Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map. The *kernel* of f is the set of vectors in the domain $\mathbb{R}^n$ of f that are mapped to the zero vector $\underline{0} \in \mathbb{R}^k$ of the range of f. That is,

$$\ker f = \{\underline{x} \in \mathbb{R}^n : f(\underline{x}) = \underline{0}\}.$$

The *image* of f is the set of the vectors in the range of the map f. That is,

$$\operatorname{Im} f = \{\underline{y} \in \mathbb{R}^k : \exists\, \underline{x} \in \mathbb{R}^n,\, f(\underline{x}) = \underline{y}\} = \{f(\underline{x}) : \underline{x} \in \mathbb{R}^n\}.$$

Examples. Let $f_1, \ldots, f_6$ denote the same maps as in the examples above.

1. Every vector of $\mathbb{R}^2$ is an image of f_1 since
$$\begin{pmatrix} x \\ y \end{pmatrix} = f_1 \begin{pmatrix} 0 \\ -x \\ y \end{pmatrix},$$
hence $\operatorname{Im} f_1 = \mathbb{R}^2$. The vector $(x, y, z)^T$ is in $\ker f_1$ if and only if
$$\begin{aligned} x - y &= 0 \\ x + z &= 0. \end{aligned}$$
The solutions of this system are the vectors of the form $(\alpha, \alpha, -\alpha)^T$ for some $\alpha \in \mathbb{R}$, so $\ker f_1 = \operatorname{span}\{(1, 1-1)^T\}$.

2. In the case of the projection to the x axis every value of f_2 is (obviously) on the x axis and we get every point of this axis as an image, so $\operatorname{Im} f_2$ is the x axis. The projection yields the zero vector if and only if the x coordinate of the projected vector is zero, hence $\ker f_2$ is the y axis.

3. Rotations and reflections are bijections of the plane, so $\operatorname{Im} f_3 = \operatorname{Im} f_4 = \mathbb{R}^2$. The origin is mapped to itself in both cases so $\ker f_3 = \ker f_4 = \{\underline{0}\}$.

4. In the case of f_5 every vector is mapped to $\underline{0}$, so $\operatorname{Im} f_5 = \{\underline{0}\}$ and $\ker f_5 = \mathbb{R}^n$. For the identity map f_6 we obviously have $\operatorname{Im} f_6 = \mathbb{R}^n$ and $\ker f = \{\underline{0}\}$.

The image and the kernel have special structure in the previous examples, namely, they are subspaces. As the following theorem shows, this is true in general:

Theorem 2.6.2. *If $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, then* $\ker f$ *is a subspace of* $\mathbb{R}^n$ *and* $\operatorname{Im} f$ *is a subspace of* $\mathbb{R}^k$.

Proof. The kernel of f is non-empty, since $\underline{0} \in \mathbb{R}^n$ is in it by part (i) of the previous proposition. If $\underline{x}, \underline{y} \in \ker f$, then by the additivity of f we have
$$f(\underline{x} + \underline{y}) = f(\underline{x}) + f(\underline{y}) = \underline{0} + \underline{0} = \underline{0},$$
so $\underline{x} + \underline{y} \in \ker f$. Similarly, if $\lambda \in \mathbb{R}$ then by the homogeneity of f we have
$$f(\lambda \underline{x}) = \lambda f(\underline{x}) = \lambda \underline{0} = \underline{0},$$
so $\lambda \underline{x} \in \ker f$. This means that $\ker f$ is a non-empty set of $\mathbb{R}^n$ which is closed under addition and scalar multiplication, so it is a subspace of $\mathbb{R}^n$.

Now we prove the statement for the image of f. We have $f(\underline{0}) = \underline{0} \in \operatorname{Im} f$, so it is non-empty. If $\underline{x}, \underline{y} \in \operatorname{Im} f$, then there is a $\underline{u} \in \mathbb{R}^n$ for which $f(\underline{u}) = \underline{x}$, and similarly, there is a $\underline{v} \in \mathbb{R}^n$ for which $f(\underline{v}) = \underline{y}$. Then by the additivity of f we have
$$\underline{x} + \underline{y} = f(\underline{u}) + f(\underline{v}) = f(\underline{u} + \underline{v}),$$
so $\underline{x} + \underline{y} \in \operatorname{Im} f$. Similarly, if $\lambda \in \mathbb{R}$, then by the homogeneity of f we have
$$\lambda \underline{x} = \lambda f(\underline{u}) = f(\lambda \underline{u}),$$
hence $\lambda \underline{x} \in \operatorname{Im} f$. That is, $\operatorname{Im} f$ is non-empty and closed under addition and scalar multiplication, so it is a subspace of $\mathbb{R}^k$. □

As we have seen above, a linear map is not always injective, i.e. the image of different vectors can be the same. It turns out that there is a connection between the injectivity of a linear map and its kernel:

Theorem 2.6.3. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map and $\underline{x}, \underline{y} \in \mathbb{R}^n$. Then $f(\underline{x}) = f(\underline{y})$ holds if and only if $\underline{x} - \underline{y} \in \ker f$. Consequently, f is injective if and only if $\ker f = \{\underline{0}\}$.*

Proof. If $f(\underline{x}) = f(\underline{y})$ holds, then by the linearity of f this is equivalent to

$$\underline{0} = f(\underline{x}) - f(\underline{y}) = f(\underline{x} - \underline{y}),$$

so $f(\underline{x}) = f(\underline{y})$ holds if and only if $\underline{x} - \underline{y} \in \ker f$.

If f is not injective, then there are different vectors $\underline{u}, \underline{v} \in \mathbb{R}^n$ for which $f(\underline{u}) = f(\underline{v})$, and hence $\underline{0} \neq \underline{u} - \underline{v} \in \ker f$. On the other hand, if $\ker f \neq \{\underline{0}\}$, then there is a vector $\underline{0} \neq \underline{u} \in \ker f$, and then $f(\underline{u}) = f(\underline{0}) = \underline{0}$, so f is not injective. □

2.6.2 The Dimension Theorem

We have seen that for a linear map $f : \mathbb{R}^n \to \mathbb{R}^k$ the sets $\ker f$ and $\operatorname{Im} f$ are subspaces of $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively. Note that the elements of $\ker f$ have n coordinates while the vectors in $\operatorname{Im} f$ have k coordinates, so in general they are different objects. Still, there is a relation between these subspaces:

Theorem 2.6.4 (Dimension theorem)**.** *If $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, then*

$$\dim \ker f + \dim \operatorname{Im} f = n.$$

Proof. Assume that $\dim \ker f = s$, and let $\underline{b}_1, \ldots, \underline{b}_s$ be a basis of $\ker f$ (there is a basis in $\ker f$ by Theorem 2.2.10). Since this is an independent set of vectors in $\mathbb{R}^n$, it can be completed to a basis of $\mathbb{R}^n$ by Theorem 2.2.11, so assume that $\underline{b}_1, \ldots, \underline{b}_s, \underline{b}_{s+1}, \ldots, \underline{b}_n$ is a basis of $\mathbb{R}^n$. It is enough to prove that the vectors $f(\underline{b}_{s+1}), \ldots, f(\underline{b}_n) \in \mathbb{R}^k$ form a basis of $\operatorname{Im} f$.

First we show that they span $\operatorname{Im} f$. Assume that $\underline{x} \in \operatorname{Im} f$, then there exists a vector $\underline{u} \in \mathbb{R}^n$ so that $f(\underline{u}) = \underline{x}$. The vector $\underline{u}$ can be written uniquely as a linear combination of the basis vectors $\underline{b}_1, \ldots, \underline{b}_n$, i.e. $\underline{u} = \lambda_1 \underline{b}_1 + \cdots + \lambda_n \underline{b}_n$ for some uniquely determined $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Then

$$\begin{aligned}
\underline{x} &= f(\lambda_1 \underline{b}_1 + \cdots + \lambda_n \underline{b}_n) \\
&= \lambda_1 f(\underline{b}_1) + \cdots + \lambda_s f(\underline{b}_s) + \lambda_{s+1} f(\underline{b}_{s+1}) + \cdots + \lambda_n f(\underline{b}_n) \\
&= \lambda_{s+1} f(\underline{b}_{s+1}) + \cdots + \lambda_n f(\underline{b}_n),
\end{aligned}$$

since $\underline{b}_1, \ldots, \underline{b}_s \in \ker f$ and hence $f(\underline{b}_1) = \cdots = f(\underline{b}_s) = \underline{0}$. This shows that every vector in $\operatorname{Im} f$ is a linear combination of the vectors $f(\underline{b}_{s+1}), \ldots, f(\underline{b}_n)$, that is, they span $\operatorname{Im} f$.

Now assume that

$$\lambda_{s+1} f(\underline{b}_{s+1}) + \cdots + \lambda_n f(\underline{b}_n) = \underline{0}$$

for some $\lambda_{s+1}, \ldots, \lambda_n \in \mathbb{R}$. The left hand side above is $f(\lambda_{s+1}\underline{b}_{s+1} + \cdots + \lambda_n \underline{b}_n)$ by the linearity of f, and hence $\underline{v} = \lambda_{s+1}\underline{b}_{s+1} + \cdots + \lambda_n \underline{b}_n \in \ker f$. But then $\underline{v}$ can be written uniquely as a linear combination of the basis vectors $\underline{b}_1, \ldots, \underline{b}_s$ (because they form a basis in $\ker f$):

$$\begin{aligned}
\lambda_1 \underline{b}_1 + \cdots + \lambda_s \underline{b}_s &= \underline{v} = \lambda_{s+1}\underline{b}_{s+1} + \cdots + \lambda_n \underline{b}_n, \\
\lambda_1 \underline{b}_1 + \cdots + \lambda_s \underline{b}_s - \lambda_{s+1}\underline{b}_{s+1} - \cdots - \lambda_n \underline{b}_n &= \underline{0},
\end{aligned}$$

and since $\underline{b}_1, \dots, \underline{b}_n$ is a basis in $\mathbb{R}^n$, they are independent and hence by Theorem 2.2.4 we get $\lambda_1 = \dots = \lambda_s = \lambda_{s+1} = \dots = \lambda_n = 0$. Another application of Theorem 2.2.4 gives that the vectors $f(\underline{b}_{s+1}), \dots, f(\underline{b}_n)$ are independent, therefore they form a basis of $\operatorname{Im} f$, and the proof is complete. □

Corollary 2.6.5. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map (so f maps the vectors of $\mathbb{R}^n$ into the* same *set). Then* $\operatorname{Im} f = \mathbb{R}^n$ *if and only if* $\ker f = \{\underline{0}\}$. *That is, f is injective if and only if it is surjective. Thus, if f is injective or surjective, then it is a bijection.*

Proof. Assume that $\operatorname{Im} f = \mathbb{R}^n$, then $\dim \operatorname{Im} f = n$ and hence by the previous theorem we get that $\dim \ker f = 0$, so $\ker f = \{\underline{0}\}$ must hold.

Assume now that $\ker f = \{\underline{0}\}$, then $\dim \ker f = 0$ and hence $\dim \operatorname{Im} f = n$. This means that there is a basis of size n in $\operatorname{Im} f$, and by Corollary 2.2.12 it is a basis of $\mathbb{R}^n$, thus, we have in fact $\operatorname{Im} f = \mathbb{R}^n$.

The second statement follows from the first one and from Theorem 2.6.3. □

Theorem 2.6.4 is sometimes referred to as the "dimension formula", or more often, as the "rank-nullity theorem". Accordingly, the number $\dim \operatorname{Im} f$ is called the *rank* of the linear map f and it is denoted by $\operatorname{rank}(f)$, while the number $\dim \ker f$ is called the *nullity* of f and it is denoted by $\operatorname{null}(f)$. We will see later that there is close connection between the rank of linear maps and the rank of matrices.

2.6.3 The Matrix of a Linear Map

Theorem 2.6.6. *Assume that $\underline{b}_1, \dots, \underline{b}_n \in \mathbb{R}^n$ is a* basis *in $\mathbb{R}^n$ and $\underline{c}_1, \dots, \underline{c}_n \in \mathbb{R}^k$ are* arbitrary *vectors. Then there is exactly one linear map $f : \mathbb{R}^n \to \mathbb{R}^k$ for which $f(\underline{b}_i) = \underline{c}_i$ for every $1 \le i \le n$. That is, the image of the basis elements determines a linear map uniquely.*

Proof. Assume first that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map for which $f(\underline{b}_i) = \underline{c}_i$ holds for every $1 \le i \le n$. If $\underline{x} \in \mathbb{R}^n$, then by Theorem 2.2.14 it can be written uniquely as a linear combination of the basis elements $\underline{b}_1, \dots, \underline{b}_n$, i.e.

$$\underline{x} = \lambda_1 \underline{b}_1 + \dots + \lambda_n \underline{b}_n,$$

where the scalars $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ are determined uniquely by $\underline{x}$ and the basis $B = \{\underline{b}_1, \dots, \underline{b}_n\}$. Note that the scalars λ_i are the coordinates of $\underline{x}$ relative to B. Hence by Proposition 2.6.1 we have

$$\begin{aligned} f(\underline{x}) &= f(\lambda_1 \underline{b}_1 + \dots + \lambda_n \underline{b}_n) \\ &= \lambda_1 f(\underline{b}_1) + \dots + \lambda_n f(\underline{b}_n) \\ &= \lambda_1 \underline{c}_1 + \dots + \lambda_n \underline{c}_n. \end{aligned} \tag{13}$$

This means that the values of f are determined uniquely, so there is at most one linear map $f : \mathbb{R}^n \to \mathbb{R}^k$ that satisfies $f(\underline{b}_i) = \underline{c}_i$ for every i.

We show that the map *defined* by (13) is linear and satisfies the conditions in the statement. This will complete the proof of the theorem. So for every $\underline{x} \in \mathbb{R}^n$ we define

$$f(\underline{x}) = \lambda_1 \underline{c}_1 + \dots + \lambda_n \underline{c}_n,$$

where $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ are the (uniquely determined) coordinates of $\underline{x}$ relative to B. First of all, the coordinates of a basis element $\underline{b}_i$ relative to B are zero except for $\lambda_i = 1$, hence $f(\underline{b}_i) = \underline{c}_i$ holds for every $1 \le i \le n$.

It remains to show that f is linear. Assume that $\underline{x}, \underline{y} \in \mathbb{R}^n$ and the coordinate vectors of them relative to B are $[\underline{x}]_B = (\lambda_1, \ldots, \lambda_n)^T$ and $[\underline{y}]_B = (\mu_1, \ldots, \mu_n)^T$. Then

$$\begin{aligned}\underline{x} + \underline{y} &= (\lambda_1 \underline{b}_1 + \cdots + \lambda_n \underline{b}_n) + (\mu_1 \underline{b}_1 + \cdots + \mu_n \underline{b}_n) \\ &= (\lambda_1 + \mu_1)\underline{b}_1 + \cdots + (\lambda_n + \mu_n)\underline{b}_n,\end{aligned}$$

hence we have that $[\underline{x} + \underline{y}]_B = (\lambda_1 + \mu_1, \ldots, \lambda_n + \mu_n)^T$. Therefore,

$$\begin{aligned}f(\underline{x} + \underline{y}) &= (\lambda_1 + \mu_1)\underline{c}_1 + \cdots + (\lambda_n + \mu_n)\underline{c}_k \\ &= (\lambda_1 \underline{c}_1 + \cdots + \lambda_n \underline{c}_n) + (\mu_1 \underline{c}_1 + \cdots + \mu_n \underline{c}_n) \\ &= f(\underline{x}) + f(\underline{y}),\end{aligned}$$

which means that f is additive. Similarly, if $\alpha \in \mathbb{R}$, then the coordinates of $\alpha\underline{x}$ relative to B are $\alpha\lambda_1, \ldots, \alpha\lambda_n$, hence

$$\begin{aligned}f(\alpha\underline{x}) &= \alpha\lambda_1 \underline{c}_1 + \cdots + \alpha\lambda_n \underline{c}_n \\ &= \alpha(\lambda_1 \underline{c}_1 + \cdots + \lambda_k \underline{c}_n) \\ &= \alpha f(\underline{x}),\end{aligned}$$

so f is homogeneous, and together with additivity this yields that f is linear and we are done. □

Now we are going to assign a matrix to a linear map $f : \mathbb{R}^n \to \mathbb{R}^k$ once a basis is chosen in both $\mathbb{R}^n$ and $\mathbb{R}^k$.

Definition 2.6.3. Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, $B_1 = \{\underline{v}_1, \ldots, \underline{v}_n\}$ is a basis of $\mathbb{R}^n$ and $B_2 = \{\underline{w}_1, \ldots, \underline{w}_k\}$ is a basis of $\mathbb{R}^k$. If $f(\underline{v}_i) = a_{1,i}\underline{w}_1 + \cdots + a_{k,i}\underline{w}_k$, that is, the uniquely determined coordinates of $\underline{v}_i$ relative to B_2 are $a_{1,i}, \ldots, a_{k,i}$, then the *matrix of the linear map* f with respect to the bases B_1 and B_2 is

$$[f]_{B_1,B_2} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,n} \end{pmatrix}.$$

In the special case when B_1 is the standard basis of $\mathbb{R}^n$ and B_2 is the standard basis of $\mathbb{R}^k$ we omit the indeces B_1 and B_2 in the notation and write simply $[f]$ for the matrix of f with respect to the standard bases.

The matrix above depends not just on f, but also on the chosen bases. The theorem above together with the uniqueness of the coordinates of a vector relative to a basis assures that once the bases B_1 and B_2 are fixed, then the matrix $[f]_{B_1,B_2}$ is determined uniquely by f. In other words, keeping the bases B_1, B_2 fixed, there is a one-to-one correspondence between the linear maps from $\mathbb{R}^n$ to $\mathbb{R}^k$ and the matrices in $\mathbb{R}^{k\times n}$ (but again, we get different matrices for the same linear map if we chose different bases).

This means that we can give the linear map by giving its matrix, and an advantage of this is that the matrix can be used to calculate the values of the map for an arbitrary vector:

Theorem 2.6.7. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$ is a basis of $\mathbb{R}^n$ and $B_2 = \{\underline{w}_1, \dots, \underline{w}_k\}$ is a basis of $\mathbb{R}^k$. If $\underline{x} \in \mathbb{R}^n$, then*

$$[f(\underline{x})]_{B_2} = [f]_{B_1,B_2} \cdot [\underline{x}]_{B_1}. \tag{14}$$

That is, if we multiply the matrix of f with respect to B_1 and B_2 by the coordinate vector of $\underline{x}$ relative to B_1 from the right, then we obtain the coordinate vector of the vector $f(\underline{x})$ relative to the basis B_2. In the special case when B_1 and B_2 are the standard bases in $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively, we obtain

$$f(\underline{x}) = [f] \cdot \underline{x}.$$

Proof. If

$$[\underline{x}]_{B_1} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} \qquad \text{and} \qquad [f]_{B_1,B_2} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,n} \end{pmatrix},$$

then

$$\begin{aligned} f(\underline{x}) &= f(\lambda_1 \underline{v}_1 + \dots + \lambda_n \underline{v_n}) \\ &= \lambda_1 f(\underline{v}_1) + \dots + \lambda_n f(\underline{v}_n) \\ &= \lambda_1 (a_{1,1}\underline{w}_1 + \dots + a_{k,1}\underline{w}_k) + \dots + \lambda_n (a_{1,n}\underline{w}_1 + \dots + a_{k,n}\underline{w}_k) \\ &= (a_{1,1}\lambda_1 + a_{1,2}\lambda_2 + \dots + a_{1,n}\lambda_n)\underline{w}_1 + \dots + (a_{k,1}\lambda_1 + \dots + a_{k,n}\lambda_n)\underline{w}_k, \end{aligned}$$

and this gives (14).

If B_1 is the standard basis in $\mathbb{R}^n$ and B_2 is the standard basis in $\mathbb{R}^k$, then $[\underline{x}]_{B_1} = \underline{x}$ and $[f(\underline{x})]_{B_2} = f(\underline{x})$, hence the second statement follows. □

Corollary 2.6.8. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map and $[f] \in \mathbb{R}^{k \times n}$ is its matrix with respect to the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^k$. If $\underline{c}_1, \dots, \underline{c}_n$ are the columns of $[f]$, then $\operatorname{Im} f = \operatorname{span}\{\underline{c}_1, \dots, \underline{c}_n\}$. Moreover, $\operatorname{rank}(f) = \operatorname{rank}([f])$ holds.*

Proof. Let us introduce the notation $W = \operatorname{span}\{\underline{c}_1, \dots, \underline{c}_n\}$. If $\underline{x} \in \mathbb{R}^n$, then $f(\underline{x}) = [f] \cdot \underline{x}$ by the previous theorem, and the product on the right hand side is a linear combination of the columns of $[f]$, so $\operatorname{Im} f \subset W$. On the other hand, if $\underline{y} = x_1\underline{c}_1 + \dots + x_n\underline{c}_n \in W$ is a linear combination of the columns, then $\underline{y} = [f] \cdot \underline{x} = f(\underline{x})$ for $\underline{x} = (x_1, \dots, x_n)^T$, hence $W \subset \operatorname{Im} f$ and the first statement of the theorem holds. Moreover,

$$\operatorname{rank}(f) = \dim \operatorname{Im} f = \dim W = \operatorname{rank}([f])$$

holds by Theorem 2.5.17. □

As we have promised, now we give a formula for the rotation about the origin on the plane by the angle α. More precisely, we give the matrix of this map:

Proposition 2.6.9. *Let $f_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ be the rotation about the origin by the angle α, then f_α is linear and its matrix with respect to the standard bases is*

$$[f_\alpha] = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Proof. We have already seen in Section 2.6.1 that f_α is linear. The first column of $[f_\alpha]$ is

$$f_\alpha((1,0)^T) = (\cos\alpha, \sin\alpha)^T$$

by the definition of $\cos\alpha$ and $\sin\alpha$. As $(0,1)^T$ is obtained when rotating $(1,0)$ about the origin by the angle $90°$, we get its image from $(\cos\alpha, \sin\alpha)^T$ in the same way, hence

$$f_\alpha((0,1)^T) = (-\sin\alpha, \cos\alpha)^T,$$

which is the second column of $[f_\alpha]$, so we are done. □

Consider the following problem: given a linear map f, we are looking for a basis in $\operatorname{Im} f$. We get the matrix $[f] \in \mathbb{R}^{k\times n}$ of f by writing the coordinates of $f(\underline{e}_i)$ in the ith column of a matrix $(1 \le i \le n)$, where $\underline{e}_1, \dots, \underline{e}_n \in \mathbb{R}^n$ is the standard basis. By the previous corollary we need to find a basis of the spanned subspace of the columns of $[f]$. As we have already seen in the proof of Theorem 2.5.17, this means that we have to find a maximal set of independent column vectors of $[f]$, and the details of the algorithm for this task was given at the end of Section 2.5.4.

Now we handle the same problem for the subspace $\ker f$. By Theorem 2.6.4 we have

$$\dim\ker f = n - \dim\operatorname{Im} f = n - \operatorname{rank}([f]).$$

The subspace $\ker f$ consists of those vectors $\underline{x} \in \mathbb{R}^n$ for which the equation $[f]\cdot\underline{x} = \underline{0}$ holds. Hence we need to find $n - \operatorname{rank}([f])$ independent vectors among the solutions of the equation above, which is equivalent to a system of linear equations. When we apply the Gaussian elimination for this system, we get $n - \operatorname{rank}(f)$ free parameters which can be chosen freely and after that the values of the other variables are defined uniquely. Every solution gives the coordinates of a vector $\underline{x}$ which solves the matrix equation $[f]\cdot\underline{x} = \underline{0}$. It is easy to see that if we take those vectors that come from the solutions where exactly one of the free parameters is 1 while the other free parameters are zero, then we get $n - \operatorname{rank}([f])$ independent vectors, so they form a basis in $\ker f$. Indeed, assume, that $m = n - \operatorname{rank}(f)$ and $x_{j_1}, \dots, x_{j_m}$ are the free parameters, where $1 \le j_1 < j_2 < \cdots < j_m \le k$ are the indices of them. For an $1 \le i \le m$, let $\underline{y}_i$ be the solution of $[f]\cdot\underline{x} = 0$ whose j_ith coordinate is 1 but whose j_lth coordinate is 0 for every $1 \le l \le m$, $l \ne i$. Then the matrix $Y \in \mathbb{R}^{k\times m}$ whose ith column is $\underline{y}_i$ contains the identity matrix I_m as a sub-matrix, hence $m = r_d(Y) = \operatorname{rank}(Y) = r_c(Y)$, so the columns of Y are independent.

Note that there is no significance of the special choice of the bases B_1 and B_2 here, we only made these changes for simplicity. The argument above can be told (with some appropriate minor changes) for the matrix $[f]_{B_1,B_2}$ where B_1 and B_2 are arbitrary bases of $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively. The details are left to the reader.

Example 2.6.4. Let $f : \mathbb{R}^5 \to \mathbb{R}^4$ be the linear map given by the matrix

$$[f] = A = \begin{pmatrix} 2 & 8 & 6 & 4 & 2 \\ 1 & 2 & -1 & 12 & 7 \\ -1 & -1 & 3 & -12 & 0 \\ 5 & 22 & 19 & 4 & 7 \end{pmatrix}.$$

After applying Gaussian elimination for the system given by the matrix $(A|\underline{0})$ we obtain the following reduced row echelon form:

$$\left(\begin{array}{ccccc|c} 1 & 0 & -5 & 0 & -31 & 0 \\ 0 & 1 & 2 & 0 & 7 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 \end{array}\right).$$

The columns that contain a leading coefficient are independent, and the corresponding columns of A (that is, the first, second and fifth column) give a basis of $\operatorname{Im} f$.

The free parameters are the third and the fifth variables (say x_3 and x_5). The solution that comes from $x_3 = 1$ and $x_5 = 0$ is $(5, -2, 1, 0, 0)^T$, while the solution that comes from $x_3 = 0$ and $x_5 = 1$ is $(31, -7, 0, -2, 1)^T$. These two vectors form a basis in $\ker f$.

Exercise 2.6.1. Give an alternative proof of the dimension theorem: show (without the usage of it) that the vectors in $\ker f$ that are obtained by the method above are not only independent, but in fact they span $\ker f$. Deduce the statement of the theorem from this.

2.6.4 Operations of Linear Maps

Since the addition is defined in $\mathbb{R}^k$, we can define point-wise addition for functions that map into $\mathbb{R}^k$ as in the case of real valued functions. As we also have scalar multiplication in $\mathbb{R}^k$, a scalar multiple of such a function can be defined as well.

Definition 2.6.5. If $f, g : X \to \mathbb{R}^k$ are functions from some set X into $\mathbb{R}^k$, then their *sum* is the function $f + g : X \to \mathbb{R}^k$ defined by $(f + g)(x) := f(x) + g(x)$ for every $x \in X$. Also, if $\lambda \in \mathbb{R}$, then the function $\lambda f : X \to \mathbb{R}^k$ defined by $(\lambda f)(x) = \lambda \cdot f(x)$ for every $x \in X$ is called the *scalar multiple* of f by λ.

Remark. It is not hard to see that the functions $X \to \mathbb{R}^k$ together with the addition and scalar multiplication defined above satisfy the statements of the theorems 2.1.1, 2.2.1 and 2.5.1, that is, as it was mentioned in the remark on page 38, they constitute a vector space over $\mathbb{R}$ (as the set of space vectors, the set $\mathbb{R}^n$ and the set of matrices in $\mathbb{R}^{k\times n}$ do, of course, together with the corresponding operations on them). As we are going to see in the following, there is a strong connection between these vector spaces and the space of matrices in the case when the functions are linear with the domain $\mathbb{R}^n$ for some integer n.

Theorem 2.6.10. *Assume that $f, g : \mathbb{R}^n \to \mathbb{R}^k$ are linear maps, $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$ is a basis of $\mathbb{R}^n$ and $B_2 = \{\underline{w}_1, \dots, \underline{w}_k\}$ is basis of $\mathbb{R}^k$. Then the functions $f+g$ and λf are also linear, and the matrix of $f+g$ w.r.t. (with respect to) the bases B_1 and B_2 is the sum of the matrices of f and g w.r.t. the bases B_1 and B_2. Also, the matrix of λf w.r.t. B_1 and B_2 is the matrix of f w.r.t. B_1 and B_2 multiplied by λ. That is*

$$[f + g]_{B_1,B_2} = [f]_{B_1,B_2} + [g]_{B_1,B_2} \qquad \text{and} \qquad [\lambda f]_{B_1,B_2} = \lambda [f]_{B_1,B_2}.$$

Proof. First we show that $f + g : \mathbb{R}^n \to \mathbb{R}^k$ is linear. If $\underline{x}, \underline{y} \in \mathbb{R}^n$, then

$$\begin{aligned}(f + g)(\underline{x} + \underline{y}) &= f(\underline{x} + \underline{y}) + g(\underline{x} + \underline{y}) \\ &= f(\underline{x}) + f(\underline{y}) + g(\underline{x}) + g(\underline{y}) \\ &= f(\underline{x}) + g(\underline{x}) + f(\underline{y}) + g(\underline{y}) \\ &= (f + g)(\underline{x}) + (f + g)(\underline{y}),\end{aligned}$$

hence $f + g$ is additive. Similarly, if $\mu \in \mathbb{R}$, then

$$\begin{aligned}(f + g)(\mu\underline{x}) &= f(\mu\underline{x}) + g(\mu\underline{x}) = \mu f(\underline{x}) + \mu g(\underline{x}) \\ &= \mu(f(\underline{x}) + g(\underline{x})) = \mu(f + g)(\underline{x}),\end{aligned}$$

so $f+g$ is homogeneous. The proof of the linearity of λf for a $\lambda \in \mathbb{R}$ is similar:

$$\begin{aligned}(\lambda f)(\underline{x}+\underline{y}) &= \lambda f(\underline{x}+\underline{y}) = \lambda(f(\underline{x})+f(\underline{y})) \\ &= \lambda f(\underline{x}) + \lambda f(\underline{y}) = (\lambda f)(\underline{x}) + (\lambda f)(\underline{y}),\end{aligned}$$

and hence λf is additive. Finally,

$$(\lambda f)(\mu\underline{x}) = \lambda f(\mu\underline{x}) = \lambda(\mu f(\underline{x})) = \mu(\lambda f(\underline{x})) = \mu(\lambda f)(\underline{x}),$$

i.e. λf is homogeneous.

For the second statement we simply use Theorem 2.6.7 and that $[\underline{u}+\underline{v}]_{B_2} = [\underline{u}]_{B_2} + [\underline{v}]_{B_2}$ holds for any vectors $\underline{u}, \underline{v} \in \mathbb{R}^k$ (see the proof of Theorem 2.6.6) together with the properties of the matrix operations. That is, for any $\underline{x} \in \mathbb{R}^n$ we have

$$\begin{aligned}[(f+g)(\underline{x})]_{B_2} &= [f(\underline{x})+g(\underline{x})]_{B_2} \\ &= [f(\underline{x})]_{B_2} + [g(\underline{x})]_{B_2} \\ &= [f]_{B_1,B_2}[\underline{x}]_{B_1} + [g]_{B_1,B_2}[\underline{x}]_{B_1} \\ &= ([f]_{B_1,B_2} + [g]_{B_1,B_2})[\underline{x}]_{B_1},\end{aligned}$$

so we get the coordinate vector of $(f+g)(\underline{x})$ relative to the basis B_2 if we multiply the matrix $[f]_{B_1,B_2} + [g]_{B_1,B_2}$ by the coordinate vector of $\underline{x}$ relative to the basis B_1. We apply this for the vectors $\underline{v}_j$ $(1 \le j \le n)$ that are the vectors of the basis of B_1. But as

$$\underline{v}_u = 0\underline{v}_1 + \cdots + 0\underline{v}_{i-1} + 1\underline{v}_i + 0\underline{v}_{i+1} + \cdots + 0\underline{v}_n,$$

we have that $[\underline{v}_j]_{B_1} = \underline{e}_j$, where $\underline{e}_j$ is the vector of the standard basis of $\mathbb{R}^k$ whose jth coordinate is 1 while its other coordinates are zero. Hence

$$[(f+g)(\underline{v}_j)]_{B_2} = ([f]_{B_1,B_2} + [g]_{B_1,B_2})\underline{e}_j,$$

and the product of the right hand side gives the jth column of the matrix $([f]_{B_1,B_2}+[g]_{B_1,B_2})$ by the definition of the matrix multiplication, while the left hand side is the jth column of the (uniquely determined) matrix of $f+g$ w.r.t. B_1 and B_2. Since this holds for every $1 \le j \le n$, we get $[f+g]_{B_1,B_2} = [f]_{B_1,B_2} + [g]_{B_1,B_2}$.

The proof is basically the same for λf, since for any $\underline{u} \in \mathbb{R}^k$ we have $[\lambda\underline{u}]_{B_2} = \lambda[\underline{u}]_{B_2}$, and hence for any $\underline{x} \in \mathbb{R}^n$

$$[(\lambda f)(\underline{x})]_{B_2} = [\lambda f(\underline{x})]_{B_2} = \lambda[f(\underline{x})]_{B_2} = \lambda([f]_{B_1,B_2}[\underline{x}]_{B_1}) = (\lambda[f]_{B_1,B_2})[\underline{x}]_{B_1}.$$

If we apply this for the vectors $\underline{v}_j$ for every $1 \le j \le n$, then we get that the jth column of the matrix of λf w.r.t. B_1 and B_2 is the jth column of $\lambda[f]_{B_1,B_2}$, so they are equal. □

There is one more operation which can be defined for functions in very general situations. If $f : A \to B$ is a function which maps from a set A to B, and the function $g : B \to C$ maps from B to C, then their composition $h = g \circ f$ is a function which maps from A to C and it is defined by $h(x) = (g \circ f)(x) = g(f(x))$ for every $x \in A$. Note that here the order of the functions f and g in the definition of h is important, since g is applicable only if the element in its argument is in B. In the case when f and g are linear maps then their composition $g \circ f$ is called the *product* of them (if it is defined).

Theorem 2.6.11. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ are linear maps. Then their product $g \circ f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map. Moreover, if $B_1 = \{\underline{u}_1, \dots, \underline{u}_n\}$ is a basis of $\mathbb{R}^n$, $B_2 = \{\underline{v}_1, \dots, \underline{v}_k\}$ is a basis of $\mathbb{R}^k$ and $B_3 = \{\underline{w}_1, \dots, \underline{w}_m\}$ is a basis of $\mathbb{R}^m$, then $[g \circ f]_{B_1,B_3} = [g]_{B_2,B_3}[f]_{B_1,B_2}$ holds.*

Proof. Assume that $\underline{x}, \underline{y} \in \mathbb{R}^n$, then by the additivity of f and g we have

$$\begin{aligned}(g \circ f)(\underline{x} + \underline{y}) = g(f(\underline{x} + \underline{y})) &= g(f(\underline{x}) + f(\underline{y})) \\ &= g(f(\underline{x})) + g(f(\underline{y})) = (g \circ f)(\underline{x}) + (g \circ f)(\underline{y}),\end{aligned}$$

so $(g \circ f)$ is additive. If moreover $\lambda \in \mathbb{R}$, then by the homogeneity of f and g we get

$$(g \circ f)(\lambda \underline{x}) = g(f(\lambda \underline{x})) = g(\lambda f(\underline{x})) = \lambda g(f(\underline{x})) = \lambda (g \circ f)(\underline{x}),$$

hence $(g \circ f)$ is homogeneous, thus, it is linear.

For the second statement we apply Theorem 2.6.7 twice, that is, for any $\underline{x} \in \mathbb{R}^n$ we have

$$\begin{aligned}[(g \circ f)(\underline{x})]_{B_3} &= [g(f(\underline{x}))]_{B_3} = [g]_{B_2,B_3}[f(\underline{x})]_{B_2} \\ &= [g]_{B_2,B_3}([f]_{B_1,B_2}[\underline{x}]_{B_1}) = ([g]_{B_2,B_3}[f]_{B_1,B_2})[\underline{x}]_{B_1}. \qquad (15)\end{aligned}$$

Applying this to a basis vector $\underline{u}_j \in B_1$ (as in the previous proof) we get $[(g \circ f)(\underline{u}_j)]_{B_3}$ on the left hand side, which is by definition the jth column of the matrix $[g \circ f]_{B_1,B_3}$. But as we have seen before, we have $[\underline{u}_j]_{B_1} = \underline{e}_j$, where $\underline{e}_j$ is the vector of the standard basis of $\mathbb{R}^n$ whose jth coordinate is 1 while its other coordinates are zero. Therefore, the right hand side becomes $([g]_{B_2,B_3}[f]_{B_1,B_2})\underline{e}_j$, which is the jh column of the matrix $([g]_{B_2,B_3}[f]_{B_1,B_2})$ by the definition of the matrix multiplication. As this holds for every $1 \leq j \leq n$, the statement follows. □

Note that in the previous proof we used that the matrix multiplication is associative. But observe that this is in fact unnecessary. If we do not do the last step in (15), then we simply obtain

$$[(g \circ f)(\underline{x})]_{B_3} = [g]_{B_2,B_3}([f]_{B_1,B_2}[\underline{x}]_{B_1}).$$

Applying this (without the associativity) to the basis vector $\underline{u}_j$ we get that the jth column of $[g \circ f]_{B_1,B_3}$ is $[g]_{B_2,B_3}([f]_{B_1,B_2}\underline{e}_j)$, so this jth column is the product of $[g]_{B_2,B_3}$ and the jth column of $[f]_{B_1,B_2}$. Hence the entry of $[g \circ f]_{B_1,B_3}$ in its ith row and jth column is the scalar product of the ith row of $[g]_{B_2,B_3}$ and the jth column of $[f]_{B_1,B_2}$, i.e. the matrix of $g \circ f$ is the product of the matrix of g and the matrix of f.

We needed only the *definition* of the matrix product so far. Now it is an easy exercise that the composition of functions is associative, that is, we have $h \circ (g \circ f) = (h \circ g) \circ f$ if both sides are defined. If A, B and C are matrices so that the products $A(BC)$ and $(AB)C$ are defined, then there are uniquely determined linear maps f, g and h so that $A = [h]$, $B = [g]$ and $C = [f]$. The previous theorem together with the associativity of the composition of functions gives an alternative proof of the associativity of the matrix multiplication. Although this argument was a little bit sketchy, it is not hard to work out the missing pieces. Also, this is in fact very enlightening: we now see that the matrix multiplication is associative because it realizes a composition of functions. The computations in the proof of Theorem 2.5.3 based on the definition of the matrix multiplication hardly show anything about this.

We are going to show another application of the previous theorem to trigonometric functions. In view of Proposition 2.6.9 it is probably not surprising that the application of certain geometric transformations may connect some algebraic expressions of trigonometric functions:

Corollary 2.6.12. *If $\alpha, \beta \in \mathbb{R}$, then*

(i) $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$,

(ii) $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$.

Proof. Let $f_\alpha, f_\beta : \mathbb{R}^2 \to \mathbb{R}^2$ be the rotations about the origin by the angles α and β, respectively. Then $f_\alpha \circ f_\beta$ is the rotation about the origin by the angle $\alpha+\beta$, so we denote this product by $f_{\alpha+\beta}$. These 3 maps are linear, and Proposition 2.6.9 gives their matrices $[f_\alpha]$, $[f_\beta]$ and $[f_{\alpha+\beta}]$ w.r.t. the standard bases. By the previous theorem we get that

$$[f_{\alpha+\beta}] = [f_\alpha \circ f_\beta] = [f_\alpha][f_\beta],$$

that is,

$$\begin{pmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix}$$

$$= \begin{pmatrix} \cos\alpha\cos\beta - \sin\alpha\sin\beta & -\cos\alpha\sin\beta - \sin\alpha\cos\beta \\ \sin\alpha\cos\beta + \cos\alpha\sin\beta & -\sin\alpha\sin\beta + \cos\alpha\cos\beta \end{pmatrix}.$$

Comparing the entries (for example) in the first columns of the two sides the result follows. □

If a function $f : A \to B$ is injective (one-to-one) and surjective (onto), that is, it is a bijection, then its *inverse* $f^{-1} : B \to A$ can be defined. For a $y \in B$ the value $f^{-1}(y)$ is the unique element $x \in A$ for which $f(x) = y$ holds. Then for every $x \in A$ we have $f^{-1}(f(x)) = x$ (i.e. $f^{-1} \circ f : A \to A$ is the identity map of A), and for every $y \in B$ we have $f(f^{-1}(y)) = y$ (i.e. $f \circ f^{-1} : B \to B$ is the identity map of B). On the other hand, if f is not a bijection, then its inverse does not exist.

If $f : \mathbb{R}^n \to \mathbb{R}^n$ is linear map then by Corollary 2.6.5 it is bijective if and only if it is injective, and also, this holds if and only if f is surjective. In the following we give another equivalent condition for this. Also, we show that if the inverse exists, then it is linear, and we determine its matrix.

Theorem 2.6.13. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map and B_1, B_2 are bases of $\mathbb{R}^n$. Then the inverse of f exists if and only if $\det[f]_{B_1,B_2} \neq 0$, and in this case it is linear and we have*

$$[f^{-1}]_{B_1,B_2} = [f]^{-1}_{B_2,B_1}.$$

Proof. By Corollary 2.6.5 the inverse of f exists if and only if f is injective, and by Theorem 2.6.3 this is equivalent to $\ker f = \{\underline{0}\}$. By Theorem 2.6.7 we have

$$\underline{0} = f(\underline{x}) \Longleftrightarrow \underline{0} = [\underline{0}]_{B_2} = [f(x)]_{B_2} = [f]_{B_1,B_2}[\underline{x}]_{B_1},$$

and as $[\underline{x}]_{B_1} = \underline{0} \Longleftrightarrow \underline{x} = \underline{0}$, the inverse exists if and only if the matrix equation $[f]_{B_1,B_2}\underline{y} = \underline{0}$ has the unique solution $\underline{y} = \underline{0}$. Since $[f]_{B_1,B_2} \in \mathbb{R}^{n\times n}$, this is equivalent to $\det[f]_{B_1,B_2} \neq 0$ by Theorem 2.4.6.

Now assume that f^{-1} exists and $\underline{x}, \underline{y} \in \mathbb{R}^n$. By the surjectivity of f there are vectors $\underline{u}, \underline{v} \in \mathbb{R}^n$ so that $f(\underline{u}) = \underline{x}$ and $f(\underline{v}) = \underline{y}$. By the definition of the inverse function we have $\underline{u} = f^{-1}(f(\underline{u})) = f^{-1}(\underline{x})$ and $\underline{v} = f^{-1}(f(\underline{v})) = f^{-1}(\underline{y})$, and together with the linearity of f this gives that

$$f^{-1}(\underline{x}+\underline{y}) = f^{-1}(f(\underline{u}) + f(\underline{v})) = f^{-1}(f(\underline{u}+\underline{v})) = \underline{u}+\underline{v} = f^{-1}(\underline{x}) + f^{-1}(\underline{y}),$$

so f^{-1} is additive. Moreover, if $\lambda \in \mathbb{R}$, then

$$f^{-1}(\lambda \underline{x}) = f^{-1}(\lambda f(\underline{u})) = f^{-1}(f(\lambda \underline{u})) = \lambda \underline{u} = \lambda f^{-1}(\underline{x}),$$

hence f^{-1} is homogeneous, i.e. it is linear.

It remains to determine the matrix of f^{-1} w.r.t. B_1 and B_2. Assume that $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$. If $id_{\mathbb{R}^n} : \mathbb{R}^n \to \mathbb{R}^n$ denotes the identity map, then obviously $[id_{\mathbb{R}^n}(\underline{v}_j)]_{B_1} = [\underline{v}_j]_{B_1} = \underline{e}_j$ holds for every basis vector $\underline{v}_j \in B_1$, where $\underline{e}_j$ is the jth standard basis vector in $\mathbb{R}^n$. Hence the matrix $[id_{\mathbb{R}^n}]_{B_1,B_1}$ is the identity matrix I_n, and we get by the application of Theorem 2.6.11 for f^{-1}, f, B_1, B_2 and $B_3 = B_1$ that

$$I_n = [id_{\mathbb{R}^n}]_{B_1,B_1} = [f \circ f^{-1}]_{B_1,B_1} = [f]_{B_2,B_1}[f^{-1}]_{B_1,B_2},$$

so $[f^{-1}]_{B_1,B_2}$ is the right inverse of $[f]_{B_2,B_1}$. By the last paragraph of the proof of Theorem 2.5.10 (or by a computation similar to the previous one) this is also a left inverse, so the statement follows. □

2.6.5 Change of Basis

We have defined the matrix of a linear map with respect to some bases, but we have not told anything about the significance of the bases so far. One may think that the choice of the standard bases simplifies the calculations, but this is only an illusion caused by the simplicity of the notations. Roughly speaking, choosing a basis means the choice of a point of view, and one can have different reasons for changing the perspective. It can happen that the simplicity of the formulae is important, but maybe the viewpoint is fixed for some reason and we simply want to adjust the computations to it. In this section we show how the coordinate vectors and the matrices of a linear maps w.r.t. different bases are connected to each other.

First we take a closer look at the situation when a vector $\underline{x}$ of $\mathbb{R}^n$ is given with its coordinate vector with respect to the basis $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$ and we have to change the basis, i.e. the coordinate vector with respect to another basis $B_2 = \{\underline{v}'_1, \dots, \underline{v}'_n\}$ of $\mathbb{R}^n$ is needed. In this case we can simply apply Theorem 2.6.7 for the identity map $id_{\mathbb{R}^n}$ of $\mathbb{R}^n$ and the bases B_2 and B_1. This way we obtain

$$[\underline{x}]_{B_1} = [id_{\mathbb{R}^n}]_{B_2,B_1}[\underline{x}]_{B_2}.$$

The jth column of the matrix $[id_{\mathbb{R}^n}]_{B_2,B_1}$ comes from the equation

$$\underline{v}'_j = a_{1,j}\underline{v}_1 + \dots + a_{n,j}\underline{v}_n.$$

Now let $g : \mathbb{R}^n \to \mathbb{R}^n$ be the linear map that maps the basis elements of B_1 to the elements of B_2, that is, $g(\underline{v}_j) = \underline{v}'_j$ holds for every $1 \le j \le n$. Observe that the jth column of the matrix $[id_{\mathbb{R}^n}]_{B_2,B_1}$ is by definition the same as the jth column of $[g]_{B_1,B_1}$, so these matrices are the same. At this point we introduce a notation:

Notation. If $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map and B is a basis of $\mathbb{R}^n$, then for simplicity we write $[f]_B$ instead of $[f]_{B,B}$, and we say that $[f]_B$ is the matrix of f w.r.t. the basis B.

Using this notation we have $[g]_{B_1,B_1} = [g]_{B_1}$ and hence

$$[\underline{x}]_{B_1} = [g]_{B_1}[\underline{x}]_{B_2}.$$

As all the elements of the basis B_2 are in $\operatorname{Im} g$, and so are the linear combinations of them, thus, in fact $\operatorname{Im} g = \mathbb{R}^n$ holds. This means that g is surjective and hence by Corollary 2.6.5 it is a bijection, so its inverse g^{-1} is defined and linear by Theorem 2.6.13, moreover, $[g^{-1}]_{B_1} = [g]_{B_1}^{-1}$ holds as well. Multiplying the identity above by this matrix from the left we obtain

Theorem 2.6.14. *Assume that $\underline{x} \in \mathbb{R}^n$, and $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$, $B_2 = \{\underline{v}'_1, \dots, \underline{v}'_n\}$ are bases of $\mathbb{R}^n$. If $g : \mathbb{R}^n \to \mathbb{R}^n$ is the linear map for which $g(\underline{v}_j) = \underline{v}'_j$ holds for every $1 \le j \le n$, then*

$$[\underline{x}]_{B_2} = [g]_{B_1}^{-1}[\underline{x}]_{B_1}.$$

Now we turn to the change of the bases in the case of linear maps:

Theorem 2.6.15. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, $B_1 = \{\underline{v}_1, \dots, \underline{v}_n\}$ and $C_1 = \{\underline{v}'_1, \dots, \underline{v}'_n\}$ are bases of $\mathbb{R}^n$, while $B_2 = \{\underline{w}_1, \dots, \underline{w}_k\}$ and $C_2 = \{\underline{w}'_1, \dots, \underline{w}'_k\}$ are bases of $\mathbb{R}^k$. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be the uniquely determined linear map for which $g(\underline{v}_j) = \underline{v}'_j$ holds for every $1 \le j \le n$ (that is, g maps the vectors of the basis B_1 to the elements of the basis C_1). Also, let $h : \mathbb{R}^k \to \mathbb{R}^k$ be the uniquely determined linear map for which $h(\underline{w}_i) = \underline{w}'_i$ holds for every $1 \le i \le k$ (that is, h maps the vectors of the basis B_2 to the elements of the basis C_2). Then*

$$[f]_{C_1,C_2} = [h]_{B_2}^{-1}[f]_{B_1,B_2}[g]_{B_1}.$$

This formula above simplifies a lot when f maps from $\mathbb{R}^n$ into itself and we use only one "old" and one "new" basis for the description of the map by a matrix:

Corollary 2.6.16. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map and let $B = \{\underline{v}_1, \dots, \underline{v}_n\}$ and $C = \{\underline{v}'_1, \dots, \underline{v}'_n\}$ be bases of $\mathbb{R}^n$. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be the uniquely determined linear map for which $g(\underline{v}_j) = \underline{v}'_j$ holds for every $1 \le j \le n$. Then*

$$[f]_C = [g]_B^{-1}[f]_B[g]_B.$$

Proof of Theorem 2.6.15. Assume that $f(\underline{v}'_j) = a'_{1,j}\underline{w}'_1 + \dots + a'_{k,j}\underline{w}'_k$. This means that the jth column of $[f]_{C_1,C_2}$ is the vector $(a'_{1,j}, \dots, a'_{k,j})^T$. As $g(\underline{v}_j) = \underline{v}'_j$ for every $1 \le j \le n$ and $h(\underline{w}_i) = \underline{w}'_i$ for every $1 \le i \le k$, we have

$$\begin{aligned}(f \circ g)(\underline{v}_j) = f(g(\underline{v}_j)) = f(\underline{v}'_j) &= a'_{1,j}\underline{w}'_1 + \dots + a'_{k,j}\underline{w}'_k \\ &= a'_{1,j}h(\underline{w}_1) + \dots + a'_{k,j}h(\underline{w}_k) \\ &= h(a'_{1,j}\underline{w}_1 + \dots + a'_{k,j}\underline{w}_k).\end{aligned}$$

As the elements $\underline{w}'_1, \dots, \underline{w}'_k$ of the basis C_2 of $\mathbb{R}^k$ are in $\operatorname{Im} h$, we have $\operatorname{Im} h = \mathbb{R}^k$, that is, h is surjective and hence by Corollary 2.6.5 it is a bijection, so its inverse h^{-1} is defined and linear. Thus,

$$(h^{-1} \circ f \circ g)(\underline{v}_j) = a'_{1,j}\underline{w}_1 + \dots + a'_{k,j}\underline{w}_k,$$

which means that the jth column of the matrix $[h^{-1} \circ f \circ g]_{B_1,B_2}$ is the same as the jth column of $[f]_{C_1,C_2}$. This holds for every $1 \le j \le n$, hence

$$[f]_{C_1,C_2} = [h^{-1} \circ f \circ g]_{B_1,B_2}.$$

Now we apply Theorem 2.6.11 first for h^{-1}, $f \circ g$ and the bases B_1, B_2 and $B_3 = B_2$ to obtain $[h^{-1} \circ f \circ g]_{B_1,B_2} = [h^{-1}]_{B_2}[f \circ g]_{B_1,B_2}$. Next we apply the same theorem to the maps f and g

and the (ordered) triple of bases B_1, B_1, B_2, which yields $[f \circ g]_{B_1,B_2} = [f]_{B_1,B_2}[g]_{B_1}$. Finally, the application of Theorem 2.6.13 for $[h^{-1}]_{B_2}$ gives the statement. □

By Exercise 2.5.2 we have $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$ for any matrices A, B for which the product AB is defined. Note that then

$$\operatorname{rank}(AB) = \operatorname{rank}((AB)^T) = \operatorname{rank}(B^T A^T) \leq \operatorname{rank}(B^T) = \operatorname{rank}(B)$$

follows as well. Assume that $f : \mathbb{R}^n \to \mathbb{R}^k$ is a linear map, B_1 is a basis of $\mathbb{R}^n$ and B_2 is a basis of $\mathbb{R}^k$. If $g : \mathbb{R}^n \to \mathbb{R}^n$ is the linear map which maps the elements of the standard basis of $\mathbb{R}^n$ to the elements of B_1 and $h : \mathbb{R}^k \to \mathbb{R}^k$ is the linear map that maps the elements of the standard basis of $\mathbb{R}^k$ to the elements of the basis B_2, then

$$\operatorname{rank}([f]_{B_1,B_2}) = \operatorname{rank}([h]^{-1}[f][g]) \leq \operatorname{rank}([h]^{-1}[f]) \leq \operatorname{rank}([f])$$

by Theorem 2.6.15 and by the remarks above. Similarly, as the equation $[h][f]_{B_1,B_2}[g]^{-1} = [f]$ holds as well, $\operatorname{rank}([f]) \leq \operatorname{rank}([f]_{B_1,B_2})$ also follows, hence these ranks are the same. Thus, by Corollary 2.6.8 we infer that

$$\operatorname{rank}(f) = \operatorname{rank}([f]) = \operatorname{rank}([f]_{B_1,B_2})$$

for any bases B_1 and B_2, as we have already mentioned in Section 2.6.3.

Example 2.6.6. Finally, we show an application of these tools. Assume that $f : \mathbb{R}^2 \to \mathbb{R}^2$ is the reflection in the line that goes through the origin and is parallel to the vector $\underline{u} = (1, 2)^T$. We are looking for its matrix w.r.t. the standard basis.

The calculation of the image of the standard basis elements requires some work, but we choose another way instead and use some images that can be determined easily. For example, f fixes the vector $\underline{u}$, that is, $f(\underline{u}) = \underline{u}$. Also, after the rotation $\underline{u}$ about the origin by 90° we get $\underline{v} = (-2, 1)^T$, and obviously $f(\underline{v}) = -\underline{v}$ holds. Therefore, it is easy to give the matrix of f w.r.t. the basis $B = \{\underline{u}, \underline{v}\}$. Indeed, we have

$$[f]_B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

If $g : \mathbb{R}^2 \to \mathbb{R}^2$ is the linear map that maps the standard basis to B, then its matrix w.r.t. the standard basis is

$$[g] = \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}, \qquad \text{and} \qquad [g]^{-1} = \frac{1}{5}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$$

by the formula for the inverse in (10). By Corollary 2.6.16 we get that

$$[f]_B = [g]^{-1}[f][g] \qquad \Longleftrightarrow \qquad [g][f]_B[g]^{-1} = [f],$$

so the matrix of f w.r.t. the standard basis is

$$\frac{1}{5}\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix} = \begin{pmatrix} -3/5 & 4/5 \\ 4/5 & 3/5 \end{pmatrix}.$$

2.6.6 Eigenvalues and Eigenvectors

We saw in the last example that the matrix of reflection w.r.t. a carefully chosen basis had a simple form, namely every entry of it outside the main diagonal was zero. A matrix of this form is called a *diagonal matrix*. Accordingly, it was easy to determine the image of the basis vectors. Using the notations of Example 2.6.6 we had $f(\underline{u}) = \underline{u}$ and $f(\underline{v}) = -\underline{v}$, so the image was a scalar multiple of the original vector.

In the following we are going to see when a basis can be chosen for a linear map so that the corresponding matrix of the map w.r.t. that basis is diagonal. We are going to work with linear maps that map from $\mathbb{R}^n$ into itself, so that $\operatorname{Im} f \subset \mathbb{R}^n$ holds. We call these maps *linear transformations* (note though that in many books this expression simply refers to a linear map). In the example above the vectors $\underline{u}$ and $\underline{v}$ were useful because f transformed them in a simple way. Now we introduce the notions that generalize this phenomenon:

Definition 2.6.7. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation. The real number $\lambda \in \mathbb{R}$ is called the *eigenvalue* of f if there is a non-zero vector $\underline{0} \neq \underline{v} \in \mathbb{R}^n$ so that $f(\underline{v}) = \lambda\underline{v}$. A non-zero vector $\underline{0} \neq \underline{v} \in \mathbb{R}^n$ is called an *eigenvector* of f if there is a real number $\lambda \in \mathbb{R}$ so that $f(\underline{v}) = \lambda\underline{v}$.

Note that the zero vector must be excluded from the set of eigenvectors since we have $f(\underline{0}) = \lambda\underline{0}$ for every linear transformation f and every scalar $\lambda \in \mathbb{R}$. But an eigenvalue can be zero, and observe that $\lambda = 0$ is the eigenvalue of a linear transformation if and only if $\ker f \neq \{\underline{0}\}$, that is, if and only if f is not injective, and the corresponding eigenvectors are the non-zero vectors in $\ker f$.

If $f(\underline{v}) = \lambda\underline{v}$ for a non-zero vector $\underline{v}$, then we say that the eigenvalue λ belongs to the eigenvector $\underline{v}$, and similarly, the eigenvector $\underline{v}$ belongs to the eigenvalue λ. Clearly, there is exactly one eigenvalue that belongs to an eigenvector, since if $\mu\underline{v} = f(\underline{v}) = \lambda\underline{v}$, then obviously $\lambda = \mu$ follows (because $\underline{v} \neq \underline{0}$). But the opposite is not true, in fact the following holds:

Proposition 2.6.17. *If $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation and $\lambda \in \mathbb{R}$ is an eigenvalue of f, then the eigenvectors belonging to λ together with the zero vector constitute a subspace of $\mathbb{R}^n$.*

Proof. Let V_λ be the set that consists of the zero vector and the eigenvectors belonging to λ. Then V_λ is non-empty since the zero vector is in it, so we have to show that it is closed under addition and scalar multiplication. Assume that $\underline{u}, \underline{v} \in V_\lambda$ and $\mu \in \mathbb{R}$. Note that $f(\underline{0}) = \lambda\underline{0}$ holds also for the zero vector. By the linearity of f we have

$$f(\underline{u} + \underline{v}) = f(\underline{u}) + f(\underline{v}) = \lambda\underline{u} + \lambda\underline{v} = \lambda(\underline{u} + \underline{v}),$$

so $\underline{u}+\underline{v}$ is either an eigenvector of f belonging to λ or the zero vector, so it is in V_λ. Similarly,

$$f(\mu\underline{u}) = \mu f(\underline{u}) = \mu(\lambda\underline{u}) = \lambda(\mu\underline{u}),$$

so $\mu\underline{u} \in V_\lambda$ and the statement follows. □

A reflection on the plane in a line going through the origin has the eigenvalues 1 and -1. Every non-zero vector on the line is an eigenvector of the reflection belonging to the eigenvalue 1, while every non-zero vector that is orthogonal to the line is an eigenvector belonging to the eigenvalue -1. The only eigenvalue of the identity map is 1, every non-zero

vector is an eigenvector of it. A rotation on the plane about the origin by an angle different from $k \cdot 180^\circ$ (for some integer k) does not have any eigenvalues or eigenvectors.

Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, and B is a basis of $\mathbb{R}^n$. If $\lambda \in \mathbb{R}$ is an eigenvalue of f and $\underline{v} \in \mathbb{R}^n$ is an eigenvector that belongs to λ, then by Theorem 2.6.7 we have

$$[f]_B[\underline{v}]_B = [f(\underline{v})]_B = [\lambda\underline{v}]_B = \lambda[\underline{v}]_B,$$

so if one multiplies the vector $[\underline{v}]_B$ by the matrix of f w.r.t. B, then the matrix multiplication transforms the vector $\underline{v}$ so that becomes the scalar multiple of itself.

Definition 2.6.8. Assume that $A \in \mathbb{R}^{n\times n}$. The scalar $\lambda \in \mathbb{R}$ is an *eigenvalue* of the matrix A if there exists a non-zero vector $\underline{0} \neq \underline{x} \in \mathbb{R}^n$ so that $A\underline{x} = \lambda\underline{x}$ holds. A non-zero vector $\underline{0} \neq \underline{x} \in \mathbb{R}^n$ is an *eigenvector* of the matrix A if $A\underline{x} = \lambda\underline{x}$ for some $\lambda \in \mathbb{R}$.

Hence if λ is an eigenvalue of the linear transformation f, then it is also an eigenvalue of its matrix of w.r.t. any basis of $\mathbb{R}^n$. Also, if $\underline{v}$ is an eigenvector of f and B is a basis of $\mathbb{R}^n$, then $\underline{x} = [\underline{v}]_B$ is an eigenvector of the matrix $[f]_B$.

On the other hand, if $A \in \mathbb{R}^{n\times n}$ is the matrix of a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ w.r.t. a basis $B = \{\underline{v}_1, \ldots, \underline{v}_n\}$, that is, $A = [f]_B$, and λ is an eigenvalue of the matrix A with the corresponding eigenvector $\underline{x} = (x_1, \ldots, x_n)^T$, then for the vector $\underline{v} = x_1\underline{v}_1 + \cdots + x_n\underline{v}_n$ we have $\underline{x} = [\underline{v}]_B$, moreover,

$$[f(\underline{v})]_B = [f]_B[\underline{v}]_B = A\underline{x} = \lambda\underline{x} = \lambda[\underline{v}]_B = [\lambda\underline{v}]_B.$$

Hence λ is an eigenvalue of f with the corresponding eigenvector $\underline{v}$.

This means that one can determine the eigenvalues of a linear map by calculating the eigenvalues of its matrix w.r.t. any basis. We will give a method for this and we will show how the eigenvectors of a matrix can be determined once its eigenvalues are known. The following statement shows how all this can be applied for finding a basis so that the matrix of the map w.r.t. it has a simple form:

Proposition 2.6.18. *Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation and let $B = \{\underline{v}_1, \ldots, \underline{v}_n\}$ be a basis of $\mathbb{R}^n$. Then the matrix $[f]_B$ is diagonal if and only if B consists of eigenvectors of f. In this case the entries in the diagonal of the matrix are the eigenvalues that belong to the corresponding eigenvectors.*

Proof. Observe, that by definition

$$[f]_B = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}$$

holds if and only if $f(\underline{v}_j) = \lambda_j\underline{v}_j$ for every $1 \leq j \leq n$. The statement follows from this immediately. □

Note that there are linear transformations whose matrix is not diagonal in any basis. For example, a rotation of the plane about the origin by angle different from $k \cdot 180^\circ$ has no eigenvalues, so its matrix cannot be diagonal.

Now we turn to the determination of the eigenvalues. The following theorem makes this possible, at least in principle.

Theorem 2.6.19. *The scalar $\lambda \in \mathbb{R}$ is an eigenvalue of the matrix $A \in \mathbb{R}^{n\times n}$ if and only if* $\det(A - \lambda I_n) = 0$.

Proof. The scalar $\lambda \in \mathbb{R}$ is an eigenvalue of A if and only the equation $A\underline{x} = \lambda\underline{x}$ has a non-zero solution. This equation holds if and only if

$$\underline{0} = A\underline{x} - \lambda\underline{x} = A\underline{x} - \lambda I_n\underline{x} = (A - \lambda I_n)\underline{x}.$$

By Theorem 2.4.6 the equation $(A - \lambda I_n)\underline{x} = \underline{0}$ has a non-zero solution if and only if the determinant of the coefficient matrix is zero, hence the statement follows. □

Observe that if λ is regarded as a variable, then $\det(A - \lambda I_n)$ is a polynomial in the variable λ. By the previous statement the eigenvalues of the matrix A are exactly the roots of this polynomial.

Definition 2.6.9. If $A \in \mathbb{R}^{n\times n}$ and λ is a variable, then the polynomial $\det(A - \lambda I_n)$ is called the *characteristic polynomial* of the matrix A. It is denoted by $p_A(\lambda)$.

If the entries of A are denoted by $a_{i,j}$, then its characteristic polynomial is the following determinant:

$$\begin{vmatrix} a_{1,1} - \lambda & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} - \lambda & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} - \lambda \end{vmatrix}.$$

Observe that if we calculate this by the definition, then the variable λ occurs in every rook arrangement at most n times. Moreover, there is only one rook arrangement which contains λ exactly n times, namely the one that is obtained by choosing the entries in the main diagonal. The product of these entries is

$$\prod_{i=1}^{n}(a_{i,i} - \lambda),$$

so the coefficient of λ^n is $(-1)^n$. Hence $\deg p_A(\lambda) = n$, and its leading coefficient is $(-1)^n$. By a well-known theorem of algebra a polynomial of degree n with real coefficients can have at most n roots, but in general these can be determined only by approximate methods.

It is not hard to see that the constant term of $p_A(\lambda)$ is $\det A$. Also, one can see easily that the coefficient of λ^{n-1} is $(-1)^{n-1}$ times the sum of the entries in the main diagonal. This latter sum is called the *trace* of the matrix A and it is denoted by $\operatorname{tr} A$.

Now assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, and B_1, B_2 are bases of $\mathbb{R}^n$. Then $[f]_{B_2} = C^{-1}[f]_{B_1}C$ for some matrix $C \in \mathbb{R}^{n\times n}$ by Corollary 2.6.16, and hence by Theorem 2.5.5 we have

$$\begin{aligned} \det([f]_{B_2} - \lambda I_n) &= \det(C^{-1}[f]_{B_1}C - \lambda I_n) = \det(C^{-1}[f]_{B_1}C - \lambda C^{-1}I_nC) \\ &= \det[C^{-1}([f]_{B_1} - \lambda I_n)C] = \det(C^{-1})\det([f]_{B_1} - \lambda I_n)\det(C) \\ &= \det([f]_{B_1} - \lambda I_n)\det(C^{-1})\det(C) = \det([f]_{B_1} - \lambda I_n)\det(C^{-1}C) \\ &= \det([f]_{B_1} - \lambda I_n)\det(I_n) = \det([f]_{B_1} - \lambda I_n), \end{aligned}$$

that is, the characteristic polynomial of the matrix of f w.r.t. some basis does not depend on the choice of the basis. Thus, we can define the *characteristic polynomial of the linear transformation* f as the characteristic polynomial of its matrix w.r.t. an arbitrary basis. It

will be denoted by $p_f(\lambda)$. This shows that the trace and the determinant of a matrix of f is also independent of the choice of the basis. Note that this follows also from Corollary 2.6.16 and Exercise 2.6.3.

Remark. Theorem 2.5.5 was stated only for matrices with real entries, while here we used for matrices whose entries are in fact polynomials. It is not hard to change the argument above so that it becomes precise. We remark that Theorem 2.5.5 can be proved also for matrices with polynomial entries. But actually we do not need this. If we choose a number $\lambda \in \mathbb{R}$, then $\det([f]_{B_1} - \lambda I_n)$ is the value of the characteristic polynomial when the number λ is substituted in place of the variable. We have seen that this value is the same for the characteristic polynomials of the matrices $[f]_{B_1}$ and $[f]_{B_2}$ and for any choice of $\lambda \in \mathbb{R}$, hence the polynomials themselves must be identical.

Once we know the eigenvalues of a matrix A, we can calculate the corresponding eigenvectors by solving the equation $A\underline{x} = \lambda\underline{x}$, or equivalently, the equation $(A - \lambda I_n)\underline{x} = \underline{0}$, where λ is an eigenvalue of A (recall that this has a non-zero solution since $\det(A - \lambda I_n) = 0$). This can be done for example by using Gaussian elimination. For a linear transformation f we can choose an arbitrary basis B, calculate the roots of the characteristic polynomial of its matrix w.r.t. B to obtain the eigenvalues of f, and finally we can solve the matrix equations above to get the coordinate vectors of the corresponding eigenvectors w.r.t. B.

Example 2.6.10. We demonstrate this method by an example. Assume that the linear transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$ is give by the formula

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 3y \\ 3x + 9y \end{pmatrix}.$$

Then the image of the vector $(1, 0)^T$ is $(1, 3)^T$ while $f((0, 1)^T) = (3, 9)^T$, so the matrix of f w.r.t. the standard basis is

$$[f] = \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}.$$

The eigenvalues of f are the roots of the polynomial

$$\det([f] - \lambda I_2) = \begin{vmatrix} 1 - \lambda & 3 \\ 3 & 9 - \lambda \end{vmatrix} = (1 - \lambda)(9 - \lambda) - 9 = \lambda^2 - 10\lambda,$$

i.e. $\lambda_1 = 0$ and $\lambda_2 = 10$. To determine the corresponding eigenvectors we have solve the equations $[f]\underline{x} = \underline{0}$ and $[f]\underline{x} = 10\underline{x}$. The solutions of the first equation are the vectors of the form $(-3\alpha, \alpha)^T$, while the second equation, which is equivalent to the system

$$\begin{aligned} -9x + 3y &= 0, \\ 3x - \ y &= 0, \end{aligned}$$

gives the solutions $(\beta, 3\beta)^T$ for any $\beta \in \mathbb{R}$. Hence the matrix of f w.r.t. the basis $B = \{(3, -1)^T, (1, 3)^T\}$ is

$$[f]_B = \begin{pmatrix} 0 & 0 \\ 0 & 10 \end{pmatrix}.$$

Exercise 2.6.2. If

$$p(x) = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_1 x + a_0$$

is a polynomial with real coefficients, then we can substitute a linear transformation f in place of the variable x, since the power f^k can be defined as the product (composition) of f with itself if $k > 0$ is a positive integer, while we define f^0 as the identity map. The sum and scalar multiple of linear maps were defined is Section 2.6.4, we only note that instead of the constant term a_0 we substitute $a_0 f^0 = a_0 \cdot id_{\mathbb{R}^n}$. Also, if $A \in \mathbb{R}^{n\times n}$ is a matrix, then $p(A)$ can be defined as

$$p(A) = a_m A^m + a_{m-1} A^{m-1} + \cdots + a_1 A + a_0 I_n.$$

The results of Section 2.6.4 show that $[p(f)]_B = p([f]_B)$ holds for any basis B of $\mathbb{R}^n$. Use this to prove the following: if the matrix of f is diagonal w.r.t. some basis of $\mathbb{R}^n$, then $p_f(f)$ is the identically zero map, where $p_f(\lambda)$ is the characteristic polynomial of f. Show the same for any $f : \mathbb{R}^2 \to \mathbb{R}^2$. (Note that the statement holds for any linear transformation in any dimension.)

Exercise 2.6.3. Show that $\mathrm{tr}\,(A + B) = \mathrm{tr}\,(A) + \mathrm{tr}\,(B)$, $\mathrm{tr}\,(\lambda A) = \lambda \mathrm{tr}\,(A)$ and $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$ hold for any matrices $A, B \in \mathbb{R}^{n\times n}$ and $\lambda \in \mathbb{R}$. Deduce that the trace of the matrix of a linear transformation w.r.t. some basis does not depend on the choice of the basis.

References

[1] M. Agrawal, N. Kayal, N. Saxena, *PRIMES is in P*, Ann. Math. **160** (2004), 781-793.

[2] W. R. Alford, A. Granville, C. Pomerance, *There are infinitely many Charmichael numbers*, Ann. Math. **139** (1994), 703-722.

[3] E. Bach, J. Shallit, *Algorithmic Number Theory, Volume I: Efficient Algorithms*, MIT Press (1996)

[4] P. J. Cameron, *Introduction to Algebra*, Second Edition, Oxford Univeristy Press (2008)

[5] G. H. Hardy, E. M. Wright, R. Heath-Brown, J. Silverman, A. Wiles, *An Introduction to the Theory of Numbers* (sixth edition), Oxford University Press (2008)

[6] K. Ireland, M. Rosen, *A Classical Introduction to Modern Number Theory*, Springer-Verlag (Graduate Texts in Mathematics, vol. **84**), New York (1982)

[7] R. Kaye, R. Wilson, *Linear Algebra*, Oxford University Press (1998)

[8] W. S. Massey, *Cross products of vectors in higher dimensional Euclidean spaces*, Amer. Math. Monthly **90** (1983), 697-701.

[9] M. O. Rabin, *Probabilistic algorithm for primality testing*, J. Number Theory **12** (1980), 128-138.

[10] D. Szeszlér, *Bevezetés a számításelméletbe I* (available online *here*)

Index