Contents

Note 1

Propositional logic

Welcome! In these notes we want to show you various aspects of mathematics. You have all had mathematics before, but now you started at DTU. Therefore, we will make sure that the mathematics you already know will become sharper tools in your mind and of course teach you a lot of new mathematics as well. A very important reason for this is that you all will need mathematics in one way or another later in your studies. However, another reason is that mathematics acts as a universal language in the natural sciences and that mathematics will enable you to interact with other engineers and scientists. We also hope that we will convince you that mathematics is beautiful. So let us begin!

1.1 Prologue: A logic problem on labels and jars

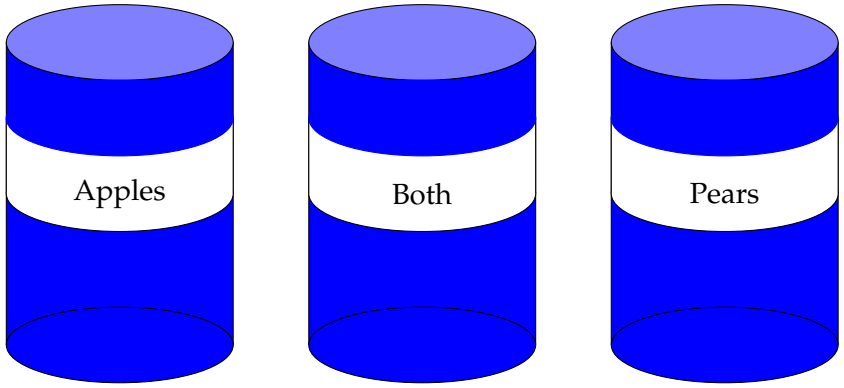
Mathematics is all about solving problems involving objects like sets, functions, numbers, derivatives, integrals, and so on. The goal of this chapter is to train and enhance your problem solving skills in general, by explaining you some tools from mathematical logic. To identify and motivate these tools, we consider as an example the following problem:

Example 1.1

Problem: Given are three jars. You cannot see what is inside the bottles, but they are labelled with “Apples”, “Both” and “Pears”. The label “Both” simply means that the jar contains both apples and pears. However, the problem is that someone switched the labels in such a way that no label is on the right jar anymore. In other words: We know that for any jar, it holds

that its label is "Apples" or "Both" or "Pears". Also we know that currently all labels are wrong, which implies that the left jar has true label "Both" or "Pears", the middle jar has true label "Apples" or "Pears", while the right jar has true label "Apples" or "Both".

To figure out where the labels really should be placed, you can draw fruit from each jar. How many times would you need to draw from the jars in order to figure out where the labels were originally?



We will solve this puzzle later, but feel free to think about it already now!

1.2 Getting started with propositional logic

Now the point of the puzzle with the jars and labels is not that it is ground breaking mathematics, but that thinking about it identifies several key ingredients that are more generally useful when thinking about a mathematical problem. One uses words like "and", "or", "not", "if ... then" when attacking problems of this sort. Let us therefore introduce some notation from what is known as *propositional logic*. First of all, it is practical to formulate short statements that can be either true or false. An example of this is: the label of jar number one is "Apples". We will call such a statements a logical *proposition*. Here are three more examples of such propositions: $x = 10$, $1 < y$, $a \neq p$. We will typically use variables like P, Q and so on, to denote such propositions. Saying that a proposition can be true of false, is more formally stated as: P can take the value T (T for True), or the value F (F for false). It is also common to use the number 1 instead of T and 0 instead of F, but we will stick to T and F.

Sometimes a proposition can be broken into smaller, simpler ones. For example, the proposition '$x = 10$ and $1 < y$', consists of the two simpler propositions '$x = 10$', '$1 < y$' combined with the word 'and'. In propositional logic, one writes $(x = 10) \wedge (1 < y)$.

To be very precise on what $\wedge$ means, let us describe exactly when an expression of the form $P \wedge Q$ is true. We will do this in the following definition

|||| Definition 1.2

Let P and Q be two propositions. Then $P \wedge Q$, pronounced as P and Q, is true precisely if P is true and Q is true. In table form:

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

The table in this definition is called a *truth table* for the logical proposition $P \wedge Q$. Since $P \wedge Q$ contains two variables P and Q and both can take the value T and F independently, the truth table of $P \wedge Q$ should handle four cases, one in each row. Each row describes what happens if P and Q take specific values. Other logical propositions can also have a truth table. Here is one more example:

|||| Example 1.3

Let P, Q, R be three logical propositions. Now consider the logical proposition $P \wedge (Q \wedge R)$. We have put parentheses around $Q \wedge R$ to clarify that we consider P combined with $Q \wedge R$ using $\wedge$. The logical proposition $(P \wedge Q) \wedge R$ may look similar, but is strictly speaking not the same as $P \wedge (Q \wedge R)$!

To determine when $P \wedge (Q \wedge R)$ is true and when it is false, we use Definition 1.2 and compute its truth table. Since we have three variables now, the truth table will contain eight rows: one row for each possible value taken by P, Q, and R. Therefore the table starts like this:

P	Q	R
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

Since $P \wedge (Q \wedge R)$ consists of P and $Q \wedge R$, it is convenient to first add a column concerning $Q \wedge R$. Using Definition 1.2, we then obtain:

P	Q	R	$Q \wedge R$
T	T	T	T
T	T	F	F
T	F	T	F
T	F	F	F
F	T	T	T
F	T	F	F
F	F	T	F
F	F	F	F

Indeed, even though in Definition 1.2 the logical propositions were called P and Q, we can also apply it for the logical proposition Q and R. Next we add a column for $P \wedge (Q \wedge R)$. Suppose for example that P, Q, R take the values F, T, T. In that case, we see from the column that we have just computed, that $Q \wedge R$ takes the value T. But then applying Definition 1.2 for the logical proposition P and $Q \wedge R$, we see that $P \wedge (Q \wedge R)$ takes the value F. Continuing like this, we can compute the final column for $P \wedge (Q \wedge R)$ and complete the truth table:

P	Q	R	$Q \wedge R$	$P \wedge (Q \wedge R)$
T	T	T	T	T
T	T	F	F	F
T	F	T	F	F
T	F	F	F	F
F	T	T	T	F
F	T	F	F	F
F	F	T	F	F
F	F	F	F	F

We can think of $\wedge$ as a logical operator: given two logical propositions P and Q, no matter how complicated P and Q already are, it produces a new propositions $P \wedge Q$. In this light $\wedge$ is sometimes called the *conjunction* and $P \wedge Q$ called the conjunction of P and Q. Let us now introduce more logical operators. In Example 1.1, we knew that all labels were wrong initially. Hence, the first jar on the left does not have label "Apples". This means that it has label "Both" or "Pears". This is formalized in the next definition:

Definition 1.4

Let P and Q be two propositions. Then $P \vee Q$, pronounced as P or Q, is defined by the following truth table:

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

The operator $\vee$ is called *disjunction* and $P \vee Q$ the disjunction of P and Q. A further logical operator is the negation of a logical proposition. We have already used this as well in Example 1.1. There we said that the labels were wrong. In particular we know that the true label of the middle jar was not "Both". Also a proposition like $x \neq 0$ is simply the negation of the proposition $x = 0$. We now formally define the negation operator.

Definition 1.5

Let P be a proposition. Then $\neg P$, pronounced as not P, is defined by the following truth table:

P	$\neg P$
T	F
F	T

As operator, $\neg$ is called the *negation*, and $\neg P$ is therefore also called the negation of P. We now already have enough ingredients to create various logical propositions. Let us consider an example.

|||| Example 1.6

Consider the logical proposition $P \vee (Q \wedge \neg P)$. We determine its truth table. Having only two variables P and Q, it is enough to consider four rows. Further $P \vee (Q \wedge \neg P)$ contains the simpler proposition $Q \wedge \neg P$, which in turn contains the proposition $\neg P$. Therefore, when computing the truth table of $P \vee (Q \wedge \neg P)$, it makes sense to add a column for $\neg P$ and one for $Q \wedge \neg P$. Then the result is the following:

P	Q	$\neg P$	$Q \wedge \neg P$	$P \vee (Q \wedge \neg P)$
T	T	F	F	T
T	F	F	F	T
F	T	T	T	T
F	F	T	F	F

Note that the truth table of $P \vee Q$ from Definition 1.4 is identical to that of $P \vee (Q \wedge \neg P)$. To be more precise, if we take the three columns of the truth table we just computed corresponding to P, Q and $P \vee (Q \wedge \neg P)$, then we get precisely the same table as the truth tabel from Definition 1.4. Apparently, two different looking logical propositions, can have the same truth tables.

1.3 Logical consequence and equivalence

The logical operators we introduced so far, $\neg$, $\wedge$, and $\vee$ allow us to write down many statements in a precise way. However, we have not discussed logical reasoning yet. We would like to be able to say something like, if P is true, then we may conclude that Q also is true. For example, if $x > 0$, then also $x > -1$. To formalize this, we use the logical symbol $\Rightarrow$, called an *implication*, and write $P \Rightarrow Q$.

We define it by giving its truth table.

|||| Definition 1.7

The logical proposition $P \Rightarrow Q$ is defined by the following truth table:

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

.

In language one often pronounces $P \Rightarrow Q$ as "P implies Q" or "if P then Q". It is sometimes convenient to write the logical proposition $P \Rightarrow Q$ as $Q \Leftarrow P$.

There are two special types of logical propositions that are simply denoted by **T** and **F**. The logical proposition **T** simply stands for a statement that is always true, like for example the statement $5 = 5$. Such a logical proposition is called a *tautology*. By contrast, the logical proposition **F**, stands for a statement that is always false, like for example $5 \neq 5$. This is called a *contradiction*. Going back to implications, saying that $P \Rightarrow Q$ is always true, really means that we claim that $P \Rightarrow Q$ is a tautology. In other words, if $P \Rightarrow Q$ is a tautology, then necessarily, P is true implies that Q is true as well.

If $P \Rightarrow Q$ is a tautology, then one says that Q is a *logical consequence* of P, or alternatively that Q is implied by P. This explains why the symbol $\Rightarrow$ is called an implication.

Stronger than an implication is what is known as a *bi-implication*, denoted by $\Leftrightarrow$ and defined as:

|||| Definition 1.8

The logical proposition $P \Leftrightarrow Q$, pronounced as "P if and only if Q", is defined by the following truth table:

P	Q	$P \Leftrightarrow Q$
T	T	T
T	F	F
F	T	F
F	F	T

.

The phrase "P if and only if Q" for the logical proposition $P \Leftrightarrow Q$ can be broken up in two parts "P if Q" and "P only if Q". The first part, "P if Q" is just a way of saying that $P \Leftarrow Q$, while "P only if Q" boils down to the statement $P \Rightarrow Q$. This explains that name bi-implication for the symbol $\Leftrightarrow$: it in fact combines two implications in one symbol. We will see later in Theorem 1.15, equation (1-22) in a more formal way that a bi-implication can indeed in this way be expressed as two implications.

|||| Example 1.9

In Example 1.6 we noted that the truth tables of $P \vee Q$ is identical to that of $P \vee (Q \wedge \neg P)$. What does this mean for the truth table of the logical proposition $(P \vee Q) \Leftrightarrow (P \vee (Q \wedge \neg P))$? Using Definition 1.4 and Example 1.6, we see that the following table is correct:

P	Q	$P \vee Q$	$P \vee (Q \wedge \neg P)$	
T	T	T	T	
T	F	T	T	
F	T	T	T	
F	F	F	F	

Now let us add a column to this table for the logical proposition $(P \vee Q) \Leftrightarrow (P \vee (Q \wedge \neg P))$ and use Definition 1.8. We obtain:

P	Q	$P \vee Q$	$P \vee (Q \wedge \neg P)$	$(P \vee Q) \Leftrightarrow (P \vee (Q \wedge \neg P))$
T	T	T	T	T
T	F	T	T	T
F	T	T	T	T
F	F	F	F	T

Since the rightmost column only contains T, we can conclude that $(P \vee Q) \Leftrightarrow (P \vee (Q \wedge \neg P))$ is a tautology.

The point now is that if $R \Leftrightarrow S$ is a tautology for some, possibly complicated, logical propositions R and S, then the truth tables of R and S are the same. In other words: if R is true, then S is true as well, but also the converse holds: if S is true, then R is true as well. Therefore, if $R \Leftrightarrow S$ is a tautology, one says that the logical propositions R and S are *logically equivalent*. From Example 1.9, we can conclude that the logical propositions $P \vee Q$ and $P \vee (Q \wedge \neg P)$ are logically equivalent. The point of this example is that it shows that sometimes one can rewrite a logical statement in a simpler form. There are several convenient tautologies that can be used to rewrite logical propositions in a simpler form. We start by giving some involving conjunction, disjunction and negation.

Theorem 1.10

Let P, Q and R be logical propositions. Then all the following expressions are tautologies.

$$P \wedge P \Leftrightarrow P \tag{1-1}$$
$$P \vee P \Leftrightarrow P \tag{1-2}$$
$$P \vee Q \Leftrightarrow Q \vee P \tag{1-3}$$
$$P \wedge Q \Leftrightarrow Q \wedge P \tag{1-4}$$
$$P \vee (Q \vee R) \Leftrightarrow (P \vee Q) \vee R \tag{1-5}$$
$$P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R \tag{1-6}$$
$$P \wedge (Q \vee R) \Leftrightarrow (P \wedge Q) \vee (P \wedge R) \tag{1-7}$$
$$P \vee (Q \wedge R) \Leftrightarrow (P \vee Q) \wedge (P \vee R) \tag{1-8}$$

Proof. To prove that one of the mentioned logical proposition is a tautology, we compute a truth table for it. Doing this for all of them would fill quite a few pages, but let us consider one of them, namely (1-6). We need to show that $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ is a tautology. In Example 1.3, we already computed the truth table of $P \wedge (Q \wedge R)$, so we do not have to redo that here. What we will need to do is to compute the truth table of $(P \wedge Q) \wedge R$, in a way similar to what we did for $P \wedge (Q \wedge R)$ in Example 1.3, and then in the last step compute the truth table of $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ using Definition 1.8. The result is the following:

P	Q	R	$P \wedge (Q \wedge R)$	$P \wedge Q$	$(P \wedge Q) \wedge R$	$P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$
T	T	T	T	T	T	T
T	T	F	F	T	F	T
T	F	T	F	F	F	T
T	F	F	F	F	F	T
F	T	T	F	F	F	T
F	T	F	F	F	F	T
F	F	T	F	F	F	T
F	F	F	F	F	F	T

We see that $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ only takes the value T, no matter what values P, Q and R take. Hence we can conclude that $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ is a tautology.

All the other items in the theorem can be shown similarly, but we will not do so here.

Readers are encouraged to prove at least one other item themselves. □

Equation (1-5) really says that when taking the disjunction of three logical propositions, it does not matter how you place the parentheses. Therefore, it is common to write $P \vee Q \vee R$ and leave the parentheses out completely. Similarly Equation (1-6) says that for the conjunction of three logical propositions, you can place the parentheses as you want. Therefore, on can write $P \wedge Q \wedge R$ without any ambiguity. This situation changes if both conjunction and disjunction occur in the same expression. Then parentheses do matter. We consider an example.

|||| Example 1.11

Consider the logical propositions $(P \wedge Q) \vee R$ and $P \wedge (Q \vee R)$. We claim that these are not logically equivalent. To show this, we could compute their truth tables, but in fact to show that two logical propositions are not logically equivalent, all we need to do is to find values for P, Q and R such that $(P \wedge Q) \vee R$ and $P \wedge (Q \vee R)$ are not both true. Let us for example find out when $(P \wedge Q) \vee R$ is false. This happens precisely if $P \wedge Q$ is false and R is false. Hence $(P \wedge Q) \vee R$ is false precisely if P and Q are not both true and R is false. However, $P \wedge (Q \vee R)$ will be false whenever P is false. Hence if (P, Q, R) take the values $(\mathrm{F}, \mathrm{T}, \mathrm{T})$, then $(P \wedge Q) \vee R$ is true, but $P \wedge (Q \vee R)$ is false. This means that in the truth table of the two expressions, there is a row looking as follows:

P	Q	R	$\ldots$	$(P \wedge Q) \vee R$	$P \wedge (Q \vee R)$
$\vdots$					
F	T	T	$\ldots$	T	F
$\vdots$					

Hence the logical propositions $(P \wedge Q) \vee R$ and $P \wedge (Q \vee R)$ are not logically equivalent. Indeed, if they would be, the logical proposition $(P \wedge Q) \vee R \Leftrightarrow P \wedge (Q \vee R)$ would be a tautology and hence only take the value T, but its truth table actually contains the following row:

P	Q	R	$\ldots$	$(P \wedge Q) \vee R$	$P \wedge (Q \vee R)$	$(P \wedge Q) \vee R \Leftrightarrow P \wedge (Q \vee R)$
$\vdots$						
F	T	T	$\ldots$	T	F	F
$\vdots$						

There are a few more tautologies that are useful when dealing with logical propositions. Apart from the conjunction $\wedge$ and disjunction $\vee$, these also involve the negation $\neg$. We leave the proofs to the reader.

Theorem 1.12

Let P, Q and R be logical propositions. Then all the following expressions are tautologies.

$$P \vee \neg P \Leftrightarrow \mathbf{T} \tag{1-9}$$
$$P \wedge \neg P \Leftrightarrow \mathbf{F} \tag{1-10}$$
$$P \Leftrightarrow \neg(\neg P) \tag{1-11}$$
$$\neg(P \vee Q) \Leftrightarrow \neg P \wedge \neg Q \tag{1-12}$$
$$\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q \tag{1-13}$$
$$\neg\mathbf{T} \Leftrightarrow \mathbf{F} \tag{1-14}$$
$$\neg\mathbf{F} \Leftrightarrow \mathbf{T} \tag{1-15}$$

Identities (1-12) and (1-13) are called the *De Morgan's laws*. Finally, there are a few tautologies describing how $\wedge$ and $\vee$ interact with tautologies and contradictions. Again, we leave the proofs of these to the reader.

Theorem 1.13

Let P, Q and R be logical propositions. Then all the following expressions are tautologies.

$$P \vee \mathbf{F} \Leftrightarrow P \tag{1-16}$$
$$P \wedge \mathbf{T} \Leftrightarrow P \tag{1-17}$$
$$P \wedge \mathbf{F} \Leftrightarrow \mathbf{F} \tag{1-18}$$
$$P \vee \mathbf{T} \Leftrightarrow \mathbf{T} \tag{1-19}$$

Using the list of tautologies in Theorems 1.10, 1.12 and 1.13 one can rewrite logical proposition in a logically equivalent form. Let us consider an example.

Example 1.14

As in Examples 1.6 and 1.9, consider the logical proposition $P \vee (Q \wedge \neg P)$. We have already seen that it is logically equivalent to $P \vee Q$, but let us now show this using Theorem 1.10 and

not by computing truth tables. First of all, using (1-8), we see that

$$P \vee (Q \wedge \neg P) \Leftrightarrow (P \vee Q) \wedge (P \vee \neg P).$$

Using (1-9), we conclude that

$$P \vee (Q \wedge \neg P) \Leftrightarrow (P \vee Q) \wedge \mathbf{T},$$

which by (1-17) can be simplified to

$$P \vee (Q \wedge \neg P) \Leftrightarrow P \vee Q.$$

In other words, using Theorem 1.10, one can prove logical equivalences without having to compute truth tables. Of course when proving this theorem, one needs to compute several truth tables, but this only needs to be done once. Generally speaking in mathematics, the point of a theorem is that it contains one or several useful results with a proof. Once the proof is given, one can use the result in the theorem whenever needed without having to prove the theorem again.

The tautologies in Theorem 1.10 only involve negation, conjunction and disjunction. Here are three very useful ones that involve implication and bi-implication as well.

|||| Theorem 1.15

Let P and Q be logical propositions. Then all the following expressions are tautologies.

$$(P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q) \tag{1-20}$$
$$(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P) \tag{1-21}$$
$$(P \Leftrightarrow Q) \Leftrightarrow (P \Rightarrow Q) \wedge (Q \Rightarrow P) \tag{1-22}$$
$$P \Leftrightarrow (\neg P \Rightarrow \mathbf{F}) \tag{1-23}$$

Proof. As in Theorem 1.10, these items can be shown by computing truth tables for each of them. We will do this for the second item and leave the others to the reader:

P	Q	$P \Rightarrow Q$	$\neg P$	$\neg Q$	$\neg Q \Rightarrow \neg P$	$(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$
T	T	T	F	F	T	T
T	F	F	F	T	F	T
F	T	T	T	F	T	T
F	F	T	T	T	T	T

Since the right column only contains T, we conclude that $(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$ indeed is a tautology. □

Equation (1-20) means that in principle, an implication can be expressed using negation and disjunction. Equation (1-21) is called *contraposition*. It means that if one wants to prove that Q is a logical consequence of P, it is also fine to show that $\neg P$ is a logical consequence of $\neg Q$. Equation (1-22) says that two logical propositions are logically equivalent precisely if they are logical consequences of each other. Finally, equation 1-23 is sometimes used to prove logical statements: instead of showing that P is true, one assumes that P is false and then tries to obtain a contradiction. If one does obtain a contradiction, one can conclude that $\neg P \Rightarrow \mathbf{F}$ is true. But then by equation 1-23, P is also true. This method is called a proof by contradiction.

1.4 Use of logic in mathematics

Logic can help to solve mathematical problems and to clarify the mathematical reasoning. In this section, we give a number of examples of this.

|||| Example 1.16

Question: Determine all real numbers x such that $-x \leq 0 \leq x - 1$.

Answer: $-x \leq 0 \leq x - 1$ is really shorthand for the logical proposition

$$-x \leq 0 \quad \wedge \quad 0 \leq x - 1.$$

The first inequality is logically equivalent to the inequality $x \geq 0$, while the second one is equivalent to $x \geq 1$. Hence a real number x is a solution if and only if

$$x \geq 0 \quad \wedge \quad x \geq 1.$$

The answer is therefore all real numbers x such that $x \geq 1$.

|||| Example 1.17

Question: determine all real numbers x such that $2|x| = 2x + 1$. Here $|x|$ denotes the absolute value of x.

Answer: if $x < 0$, then $|x| = -x$, while if $x \geq 0$, then $|x| = x$. Hence it is convenient to consider the cases $x < 0$ and $x \geq 0$ separately. More formally, we have the following sequence of logically equivalent statements:

$$
\begin{array}{l}
2|x| = 2x + 1 \\
\Leftrightarrow \\
2|x| = 2x + 1 \quad \wedge \quad (x < 0 \quad \vee \quad x \geq 0) \\
\Leftrightarrow \\
(2|x| = 2x + 1 \quad \wedge \quad x < 0) \quad \vee \quad (2|x| = 2x + 1 \quad \wedge \quad x \geq 0) \\
\Leftrightarrow \\
(-2x = 2x + 1 \quad \wedge \quad x < 0) \quad \vee \quad (2x = 2x + 1 \quad \wedge \quad x \geq 0) \\
\Leftrightarrow \\
(-4x = 1 \quad \wedge \quad x < 0) \quad \vee \quad (0 = 1 \quad \wedge \quad x \geq 0) \\
\Leftrightarrow \\
(x = -1/4 \quad \wedge \quad x < 0) \quad \vee \quad (\mathbf{F} \quad \wedge \quad x \geq 0) \\
\Leftrightarrow \\
x = -1/4 \quad \vee \quad \mathbf{F} \\
\Leftrightarrow \\
x = -1/4
\end{array}
$$

Hence the only solution to the equation $2|x| = 2x + 1$ is $x = -1/4$.

|||| Example 1.18

Question: Determine all nonnegative real numbers such that $\sqrt{x} = -x$.

Observation: It is tempting to take the square on both sides, one then obtains $x = x^2$, and then to conclude that $x = 0$ and $x = 1$ are the solutions to the equation $\sqrt{x} = -x$. However, $x = 0$ is indeed a solution, but $x = 1$ is not, since $\sqrt{1} \neq -1$. What went wrong?

Answer: The reasoning actually shows that if x satisfies the equation $\sqrt{x} = -x$, then $x = x^2$, which in turn implies that $x = 0$ or $x = 1$. Hence the following statement is completely correct:

$$(\sqrt{x} = -x) \Rightarrow (x = 0 \vee x = 1).$$

In that sense, nothing went wrong and any solution to the equation $\sqrt{x} = -x$ must indeed be either $x = 0$ or $x = 1$. What may cause confusion is that this does not at all mean that

$x = 0$ and $x = 1$ both are solutions to the equation $\sqrt{x} = -x$. This would namely amount to the statement

$$(x = 0 \vee x = 1) \Rightarrow (\sqrt{x} = -x),$$

which is different from what we have shown and actually is not true. To solve the question, all we need to do it to check if the potential solutions $x = 0$ and $x = 1$ really are solutions. We then obtain that $x = 0$ is the only solution.

1.5 Epilogue: the logic problem on labels and jars

Let us return to the problem of jars and labels from the first section.

|||| Example 1.19

Let us denote by $P_1(A)$ the statement that the left jar has true label "Apples". Similarly, let us write $P_1(B)$, respectively $P_1(P)$, for the statement that the left jar has true label "Both", respectively "Pears". We then know that $P_1(B) \vee P_1(P)$ is always true, since the left jar cannot have label "Apples". Similarly for the middle jar, we can introduce $P_2(A)$, $P_2(B)$, and $P_2(P)$ for the statements that the middle jar has true label "Apples", "Both", "Pears" and conclude that $P_2(A) \vee P_2(P)$ is a tautology. Similarly for the right jar, we obtain that $P_3(A) \vee P_3(B)$ is a tautology. In conclusion,

$$(P_1(B) \vee P_1(P)) \wedge (P_2(A) \vee P_2(P)) \wedge (P_3(A) \vee P_3(B)) \tag{1-24}$$

is a tautology. Using De Morgan's laws repeatedly, we can rewrite this to the logically equivalent statement

$$\begin{array}{lllll}
(P_1(B) \wedge P_2(A) \wedge P_3(A)) & \vee & (P_1(B) \wedge P_2(A) \wedge P_3(B)) & \vee \\
(P_1(B) \wedge P_2(P) \wedge P_3(A)) & \vee & (P_1(B) \wedge P_2(P) \wedge P_3(B)) & \vee \\
(P_1(P) \wedge P_2(A) \wedge P_3(A)) & \vee & (P_1(P) \wedge P_2(A) \wedge P_3(B)) & \vee \\
(P_1(P) \wedge P_2(P) \wedge P_3(A)) & \vee & (P_1(P) \wedge P_2(P) \wedge P_3(B)). &
\end{array}$$

This statement is still a tautology, since it is logically equivalent to the tautology from Equation 1-24. Since we know that in the correct labelling each label has to be used exactly once, a statement like $P_1(B) \wedge P_2(A) \wedge P_3(A)$ where the same label occurs twice, cannot be correct, that is to say that it is a contradiction. Using that disjunction absorbs contradictions, see Equation (1-16), we therefore conclude that

$$(P_1(B) \wedge P_2(P) \wedge P_3(A)) \quad \vee \quad (P_1(P) \wedge P_2(A) \wedge P_3(B)) \tag{1-25}$$

is a tautology.

What this shows is that there are only two possible correct ways to label the jars. This is already very helpful, since we did not even draw any fruit yet! Now let us investigate what the effect of drawing from a jar is. If we draw from the left jar, we do not learn much about the label of that jar. Indeed, since the true label is "Both" or "Pears", if we draw an apple from it, we know the true label cannot be "Pears", but if we draw a pear from it, the true label could still be "Both" or "Pears". Similarly drawing from the right jar, may not determine its true label. The situation is different for the middle jar. Since the true label of the middle jar is "Apples" or "Pears", if we draw an apple from it, its true label cannot be "Pears". Apparently, it must be "Apples" in that case. Similarly, if we draw a pear from the middle jar, its true label is "Pears". We arrive at the following solution for the problem:

Solution:

Step 1: Draw from the middle jar. Since we know all labels are wrong, the middle jar, that has label "Both", contains either only apples, or only pears. If we draw an apple from the middle jar, then we can conclude the correct label should have been "Apples," while if we draw a pear from the middle jar, then we can conclude that that correct label should have been "Pears".

Step 2: We know that the logical proposition in Equation 1-25 is a tautology. This implies that if we found in Step 1 that the correct label for the middle jar is "Apples", then $P_1(P) \wedge P_2(A) \wedge P_3(B)$ is true, while if the correct label of the middle jar was identified as "Pears" in Step 1, then $P_1(B) \wedge P_2(P) \wedge P_3(A)$ is true.

Conclusion: We only need to draw once! After that we can identify all three labels correctly. Moreover, we have actually found a simple step-by-step procedure to determine the correct labelling. This is an example of what we later will call an algorithm. To make it look more like a computer algorithm, we give it as follows:

Algorithm 1 Label Identifier

```
Draw from the jar labelled "Both" and denote the result by R.
if R = apple then
    Identify the labels of the jars as "Pears","Apples","Both",
else
    Identify the labels of the jars as "Both","Pears","Apples".
```

There are many puzzles of this type. Here is another one. Feel free to try to solve it yourself before reading the solution.

|||| Example 1.20

A police officer is investigating a burglary and was able to narrow the number of suspects down to three. He is absolutely sure that one of these three committed the crime. His questioning of the three suspects yields him the following statements:

Suspect1:	"Suspect2 did it";
	"I wasn't there";
	"I am innocent"
Suspect2:	"Suspect3 is innocent";
	"everything Suspect 1 said is a lie";
	"I didn't do it"
Suspect3:	"I didn't do it";
	"Suspect1 is lying if he said that he wasn't there";
	"Suspect2 is lying if he said that everything that Suspect1 said is a lie"

Confused, the police officer goes to his boss, the police commissioner. The police commissioner says: "I know these suspects quite well and every single one of them always lies at least once in their statements." Can you help the police officer to figure out which suspect is guilty of the burglary?

Solution Let us introduce some logical proposition to analyze the situation. First of all, P_1 is the statement "Suspect1 did it" and similarly P_2 stands for "Suspect2 did it", P_3 for "Suspect3 did it". With this notation in place, we know that

$$P_1 \vee P_2 \vee P_3$$

is a tautology, since the police officer is absolutely sure that one of the three suspects committed the burglary.

Now let us analyze the statements from the suspects:

Statements from Suspect1:	
"Suspect2 did it";	this is just P_2
"I wasn't there";	we call this R_1
"I am innocent";	this amounts to $\neg P_1$

Now let us consider the insight from the police commissioner. He says that any of the three suspects has lied at least once in their statements. In particular, Suspect1 is lying, which means that $\neg P_2 \vee \neg R_1 \vee \neg(\neg P_1)$ is a tautology. Using Equation (1-11), we conclude that

$$\neg P_2 \vee \neg R_1 \vee P_1$$

is a tautology.

Statements from Suspect2:	
"Suspect3 is innocent";	this is $\neg P_3$
"everything Suspect 1 said is a lie";	this amount to $\neg P_2 \wedge \neg R_1 \wedge P_1$
"I didn't do it";	this is $\neg P_2$

Now let us again consider the insight from the police commissioner. For Suspect 2 we obtain that $P_3 \vee \neg(\neg P_2 \wedge \neg R_1 \wedge P_1) \vee P_2$ is a tautology. One can simplify this expression using Theorem 1.10. First of all, using Equation (1-13), the proposition $\neg(\neg P_2 \wedge \neg R_1 \wedge P_1)$ is logically equivalent to $\neg(\neg P_2) \vee \neg(\neg R_1) \vee \neg P_1)$, which in turn is logically equivalent to $P_2 \vee R_1 \vee \neg P_1$ using Equation (1-11). Substituting this in the original tautology, we see that $P_3 \vee (P_2 \vee R_1 \vee \neg P_1) \vee P_2$ is a tautology. Simplifying $P_2 \vee P_2$ to P_2 using Equation (1-2), we obtain that

$$P_3 \vee P_2 \vee R_1 \vee \neg P_1$$

is a tautology.

The statements of Suspect3 are a bit involved, so before putting them in a table, let us consider the last two statements. The second statement of Suspect3 is that "Suspect1 is lying if he said that he wasn't there". In other words: "Suspect1 wasn't there" $\Rightarrow$ "Suspect1 is lying". However, the police commissioner already told us that the statement "Suspect1 is lying" always is true. This means that the implication, "Suspect1 wasn't there" $\Rightarrow$ "Suspect1 is lying", is a tautology. Similarly, the third statement from Suspect3, "Suspect2 is lying if he said that everything that Suspect1 said is a lie", is a tautology. Hence the second and third statements from Suspect3 do not give us any information that we did not already know.

Statements from Suspect3:	
"I didn't do it";	this is $\neg P_3$
"Suspect1 is lying if he said that he wasn't there";	this is a tautology **T**
"Suspect2 is lying if he said that everything that Suspect1 said is a lie;	this is a tautology **T**

Now let us for the third time consider the insight from the police commissioner. Suspect3 lied and hence $P_3 \vee \neg\mathbf{T} \vee \neg\mathbf{T}$ is a tautology. Since $\neg\mathbf{T}$ is logically equivalent to **F** by Equation 1-14, Equation (1-16) implies that P_3 is a tautology.

Collecting everything together, we have obtained the following tautologies: $P_1 \vee P_2 \vee P_3$, $\neg P_2 \vee \neg R_1 \vee P_1$, $P_3 \vee P_2 \vee R_1 \vee \neg P_1$, P_3. The fact that P_3 is a tautology immediately implies that the only possibility is that Suspect3 has committed the burglary and that as a consequence Suspect1 and Suspect2 are innocent. However, we still need to check that in this case all the other tautologies are indeed true. If not, this would mean that no solution exists and that the police officer or the police commissioner is wrong. First of all, if P_3 takes the value T, then

$P_1 \vee P_2 \vee P_3$ and $P_3 \vee P_2 \vee R_1 \vee \neg P_1$ will be tautologies, since T $\vee$ S is logically equivalent to T for any logical proposition S. This leaves $\neg P_2 \vee \neg R_1 \vee P_1$. Since Suspect2 is innocent, P_2 takes the value F and as a consequence, $\neg P_2$ takes the value T. Hence indeed $\neg P_2 \vee \neg R_1 \vee P_1$ is a tautology. This means that there is nothing contradictory. The police should arrest Suspect3!

Note 2

Sets and functions

2.1 Sets

The notion of a *set* is very fundamental in mathematics and therefore we will discuss some terminology and notation concerning sets in this section.

Basically, a set A is a way to "bundle" elements together in one object. If we for example want to write down a set consisting of the numbers 0 and 1, we simply write $\{0,1\}$. This would be an example of a set with two elements. Elements do not have to be numbers, but could in principle be anything. Repetition of elements does not make a set larger in the sense that if an element occurs twice or more times in a set, all its duplicates can be removed. For example, one has $\{0,0,1\} = \{0,1\}$ and $\{1,1,1,1\} = \{1\}$. Also the order in which the elements of a set are written down is not important. Hence for example $\{0,1\} = \{1,0\}$.

Some sets of numbers are used so often, that there is a standard notation for them:

$\mathbb{N} = \{1,2,\dots\}$	the set of *natural numbers*,
$\mathbb{Z} = \{\dots,-2,-1,0,1,2,\dots\}$	the set of *integers*,
	and
$\mathbb{R}$	the set of all *real numbers*.

Saying that a is an element of A is expressed as: $a \in A$. Some authors prefer to write the set first and then the element, writing $A \ni a$ instead of $a \in A$. If an element a is not in the set A, one can use the negation from propositional logic and write $\neg(a \in A)$. It is also common though to write $a \notin A$ for the statement that a is not an element of the

set A. If two elements are in the same set, say $a_1 \in A$ and $a_2 \in A$, it is common to write $a_1, a_2 \in A$.

|||| Example 2.1

We have $1 \in \mathbb{N}$ and $-1 \in \mathbb{Z}$, while $-1 \notin \mathbb{N}$. Further $\pi \in \mathbb{R}$, but $\pi \notin \mathbb{Z}$, since $\pi \approx 3.1415$ is not an integer.

A set is determined by its elements, meaning that two sets A and B are equal, $A = B$, if and only if they contain the same elements. In other words $A = B$ if and only if for all elements a, it holds that $a \in A \quad \Leftrightarrow \quad a \in B$. If A and B are sets, then B is called a *subset* of A, if any element of B is also an element of A. A common notation for this is $B \subseteq A$. In other words, the statement $B \subseteq A$ is by definition true if and only if the statement $a \in B \Rightarrow a \in A$ is true for all elements a. In particular $A \subseteq A$, since for all a the implication $a \in A \Rightarrow a \in A$ is true. Instead of writing $B \subseteq A$, one may also write $A \supseteq B$.

The *empty set* is the set not containing any elements at all. It is commonly denoted by $\emptyset$, inspired by the letter Ø from the Danish and Norwegian alphabet. Some authors use $\{\}$ for the empty set, but we will always use the notation $\emptyset$ for it. The empty set $\emptyset$ is a subset of any other set A.

If one wants to stress that a set B is a subset of A, but not equal to all of A, one writes $B \subsetneq A$ or alternatively $A \supsetneq B$. Finally, if you want to express in a formula that B is not a subset of A, it is possible to use the logical negation symbol $\neg$ and write that $\neg(B \subseteq A)$, but it is more customary to write $B \nsubseteq A$ or alternatively $A \nsupseteq B$.

|||| Example 2.2

Since every natural number is an integer, we have $\mathbb{N} \subseteq \mathbb{Z}$. Every integer $n \in \mathbb{Z}$ is also a real number. Therefore $\mathbb{Z} \subseteq \mathbb{R}$. In fact, we even have $\mathbb{N} \subsetneq \mathbb{Z}$ and $\mathbb{Z} \subsetneq \mathbb{R}$. Indeed to show $\mathbb{N} \subsetneq \mathbb{Z}$, we just have to check that $\mathbb{N} \subseteq \mathbb{Z}$ (which we already observed) and that $\mathbb{N} \neq \mathbb{Z}$. However, since $-1 \in \mathbb{Z}$, but $-1 \notin \mathbb{N}$, we can indeed conclude that $\mathbb{N} \neq \mathbb{Z}$. Similarly $\mathbb{Z} \subsetneq \mathbb{R}$, since $\pi \in \mathbb{R}$ and $\pi \notin \mathbb{Z}$.

A common way to construct subsets of a set A is by selecting elements from it for which some logical expression is true. For the sake of notation, let us denote this logical expression by $P(a)$. Then $\{a \in A \mid P(a)\}$ denotes the subset of A consisting of precisely those elements $a \in A$ for which the logical expression $P(a)$ is true.

|||| Example 2.3

Let $\mathbb{Z}$ as before be the set of integers. Then $\{a \in \mathbb{Z} \mid a \geq 1\}$ is just the set $\{1, 2, 3, 4, \dots\}$ and $\{a \in \mathbb{Z} \mid a \leq 3\} = \{\dots, -1, 0, 1, 2, 3\}$. Also $\{a \in \mathbb{Z} \mid 1 \leq a \leq 3\} = \{1, 2, 3\}$.

|||| Example 2.4

Apart from the standard notations $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ that we already introduced, a further example is the set $\mathbb{Q}$: the set of all *rational numbers*, that is to say, the set of fractions of integers. More precisely we have

$$\mathbb{Q} = \left\{ \frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0 \right\}.$$

This simply means that an element of $\mathbb{Q}$ is of the form a/b, where both a and b are integers, where b is not zero. Note that fractions like $1/2$ and $2/4$ are the same, since $2/4$ can be simplified to $1/2$ by dividing both numerator and denominator by 2. More generally, two fractions a/b and c/d are the same if and only if $ad = bc$.

Since any integer $n \in \mathbb{Z}$ can be written as $n/1$, we see that $\mathbb{Z} \subseteq \mathbb{Q}$. In fact, since $1/2 \in \mathbb{Q}$ and $1/2 \notin \mathbb{Z}$, we have $\mathbb{Z} \subsetneq \mathbb{Q}$. Further, any fraction of integers is a real number, so that $\mathbb{Q} \subseteq \mathbb{R}$. It turns out that $\mathbb{Q} \subsetneq \mathbb{R}$. A way to see this is to find a real number that cannot be written as a fraction of integers. One example of such a real number is $\sqrt{2}$, but we will not show here why $\sqrt{2} \notin \mathbb{Q}$.

Given two real numbers a and b such that $a < b$, one can define several standard subsets of $\mathbb{R}$ called *intervals*. These are:

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\},$$

$$[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\},$$

$$]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$$

and

$$]a, b[= \{x \in \mathbb{R} \mid a < x < b\}.$$

Intervals of the form $[a, b]$ are called *closed* , while intervals of the form $]a, b[$ are called *open*.

It is also customary to define

$$\mathbb{R}_{\geq a} = \{x \in \mathbb{R} \mid x \geq a\},$$

$$\mathbb{R}_{>a} = \{x \in \mathbb{R} \mid x > a\},$$
$$\mathbb{R}_{\leq a} = \{x \in \mathbb{R} \mid x \leq a\}$$

and

$$\mathbb{R}_{<a} = \{x \in \mathbb{R} \mid x < a\}.$$

Example 2.5

The interval $]0,1]$ consists of all real numbers x satisfying $0 < x \leq 1$. This interval is not closed and not open either. The set $\mathbb{R}_{\geq 0}$ is the set of all nonnegative real numbers, while $\mathbb{R}_{>0}$ is the set of all positive real numbers. The notation $\mathbb{R}_+$ is also often used to denote the set of all positive real numbers.

It is intuitive that two sets are equal if and only if they are subsets of each other. Let us be more precise as to why this is true and state this as a lemma.

Lemma 2.6

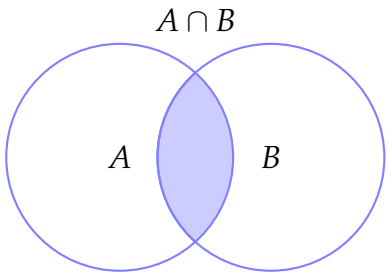
Let A and B be two sets. Then $A = B$ if and only if $A \subseteq B$ and $A \supseteq B$.

Proof. The statement $A = B$ for two sets A and B, is logically equivalent to the statement $a \in A \Leftrightarrow a \in B$ for all a. Using Equation (1-22), we can split the bi-implication up in two implications. Then we obtain the logically equivalent statement $(a \in A \Rightarrow a \in B) \quad \wedge \quad (a \in A \Leftarrow a \in B)$ for all a. But this is equivalent to saying that $A \subseteq B \quad \wedge \quad A \supseteq B$. $\square$

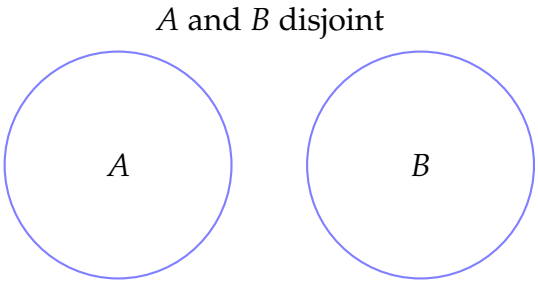
Instead of $\subseteq$ and $\supseteq$, some authors prefer the symbols $\subset$ and $\supset$. However, yet other authors, use the symbols $\subset$ and $\supset$ in the meaning of $\subsetneq$ and $\supsetneq$, inspired by the use of $<$ and $>$ in the setting of strict inequalities. To avoid confusion, we will not use the symbols $\subset$ or $\supset$.

There are several basic definitions and operations involving sets that we will use later on. We illustrate them in Example 2.7. Fist of all, if A and B are two sets, then we define the *intersection* of A and B, denoted by $A \cap B$, to be the set consisting of all elements that are both in A and in B. In other words:

$$A \cap B = \{a \mid a \in A \wedge a \in B\}. \tag{2-1}$$

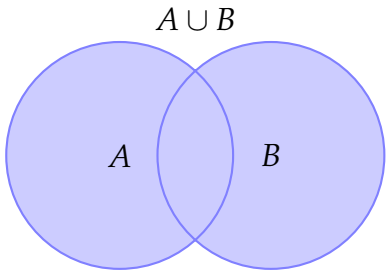


Two sets A and B are called *disjoint*, if $A \cap B = \emptyset$.

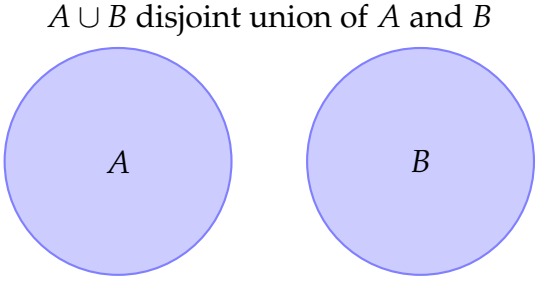


The *union* of A and B is defined as:

$$A \cup B = \{a \mid a \in A \vee a \in B\}. \tag{2-2}$$

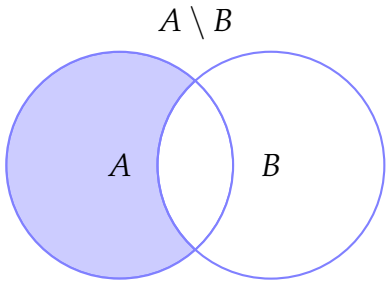


The union $A \cup B$ is called a *disjoint union* of A and B if $A \cap B = \emptyset$.



The *set difference* of A and B, often pronounced as A minus B, is defined to be:

$$A \setminus B = \{a \mid a \in A \wedge a \notin B\}.$$



Finally, the *Cartesian product* of A and B is the set:

$$A \times B = \{(a, b) \mid a \in A \wedge b \in B\}.$$

In other words, the Cartesian product of two sets A and B, is simply the set of all pairs (a, b), whose first coordinate is from A and whose second coordinate is from B. The Cartesian product of a set A with itself is sometimes denote as A^2. In other words: $A^2 = A \times A$.

Later on we will mainly use the Cartesian product of two sets, but it is not hard to define the Cartesian product of more than two sets. One simply uses more coordinates, one for each set in the Cartesian product. For example $A \times B \times C = \{(a, b, c) \mid a \in A, b \in B, \text{ and } c \in C\}$. More generally, if n is a positive integer and $A_1, \ldots, A_n$ are sets, then

$$A_1 \times \cdots \times A_n = \{(a_1, \ldots, a_n) \mid a_1 \in A_1, \ldots, a_n \in A_n\}.$$

If all sets are equal, say $A_1 = A, \ldots, A_n = A$, then one often writes A^n for their Cartesian product. In other words

$$A^n = \{(a_1, \ldots, a_n) \mid a_1 \in A, \ldots, a_n \in A\}. \tag{2-3}$$

Let us illustrate the introduced concepts for sets in an example.

|||| Example 2.7

Let 1, 2, 3, and 4 be the first four positive integers. Then:

1. $\{1,2\} \subseteq \{1,2,3\}$ and in fact $\{1,2\} \subsetneq \{1,2,3\}$,
2. $\{1,2\} \supseteq \{2\}$ and in fact $\{1,2\} \supseteq \{2\}$,
3. $\{1,4\} \not\subseteq \{1,2,3\}$,
4. $\{1,2,3\} \cap \{2,3,4\} = \{2,3\}$,
5. $\{1,2\}$ and $\{3\}$ are disjoint sets,
6. $\{1,2,3\} \cup \{2,3,4\} = \{1,2,3,4\}$,
7. $\{1,2,3,4\}$ is the disjoint union of $\{1,2\}$ and $\{3,4\}$,
8. $\{1,2,3\} \setminus \{2,3,4\} = \{1\}$,
9. $\{2,3,4\} \setminus \{1,2,3,4\} = \emptyset$,
10. $\{1,2\} \times \{3,4\} = \{(1,3),(1,4),(2,3),(2,4)\}$,
11. $\{1,2\}^2 = \{(1,1),(1,2),(2,1),(2,2)\}$.

In Equations (2-1) and (2-2), the logical opeators $\wedge$ and $\vee$ came in very handy. In Theorem 1.10 we have seen various properties of these two logical operators. These can now be used to show similar properties of intersections and unions of sets:

|||| Theorem 2.8

Let A, B and C be sets. Then

$$A \cap A = A \tag{2-4}$$
$$A \cup A = A \tag{2-5}$$
$$A \cup B = B \cup A \tag{2-6}$$
$$A \cap B = B \cap A \tag{2-7}$$
$$A \cup (B \cup C) = (A \cup B) \cup C \tag{2-8}$$
$$A \cap (B \cap C) = (A \cap B) \cap C \tag{2-9}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{2-10}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{2-11}$$

Proof. Let us prove the last item, that is to say Equation (2-11). Proving the remaining items is left to the reader. According to Equation (2-1), we have

$$B \cap C = \{a \mid a \in B \wedge a \in C\}.$$

On the other hand, applying Equation (2-2) to the sets A and $B \cap C$, we see that

$$A \cup (B \cap C) = \{a \mid a \in A \vee a \in B \cap C\}.$$

Combining these two equations and using Equation 1-8, we then obtain the following:

$$\begin{aligned} A \cup (B \cap C) &= \{a \mid a \in A \vee (a \in B \wedge a \in C)\} \\ &= \{a \mid (a \in A \vee a \in B) \wedge (a \in A \vee a \in C)\} \\ &= \{a \mid (a \in A \cup B) \wedge (a \in A \cup C)\} \\ &= (A \cup B) \cap (A \cup C). \end{aligned}$$

□

Theorem 2.8 shows that propositional logic can be used to rewrite intersections and unions of sets. We give one example involving the difference of some sets. Here Theorems 1.12 and 1.13 will come in handy.

|||| Example 2.9

Let A, B and C be three sets. In this example we show that $A \cap (B \setminus C) = (A \cap B) \setminus (A \cap C)$. First of all, we have

$$\begin{aligned} A \cap (B \setminus C) &= \{a \mid a \in A \wedge a \in B \setminus C\} \\ &= \{a \mid a \in A \wedge (a \in B \wedge \neg(a \in C))\}. \end{aligned}$$

On the other hand

$$\begin{aligned} (A \cap B) \setminus (A \cap C) &= \{a \mid a \in A \cap B \wedge \neg(a \in A \cap C)\} \\ &= \{a \mid (a \in A \wedge a \in B) \wedge \neg(a \in A \wedge a \in C)\} \\ &= \{a \mid (a \in A \wedge a \in B) \wedge (\neg(a \in A) \vee \neg(a \in C))\} \\ &= \{a \mid (a \in A \wedge a \in B) \wedge \neg(a \in A) \quad \vee \quad (a \in A \wedge a \in B) \wedge \neg(a \in C)\} \\ &= \{a \mid \mathbf{F} \quad \vee \quad (a \in A \wedge a \in B) \wedge \neg(a \in C)\} \\ &= \{a \mid (a \in A \wedge a \in B) \wedge \neg(a \in C)\} \\ &= \{a \mid a \in A \wedge (a \in B \wedge \neg(a \in C))\}. \end{aligned}$$

We can conclude that indeed $A \cap (B \setminus C) = (A \cap B) \setminus (A \cap C)$.

2.2 Functions

A very important concept in mathematics is a *function*. For two given sets A and B, a function f from A to B assigns to any $a \in A$ an element $b \in B$. Instead of the phrase "assigns to a an element b" one usually just says that "f maps a to b". For this reason a function is sometimes also called a *map*. Instead of saying that "f maps a to b" one can also say that "f evaluated in a is equal to b".

The set A is called the *domain* of the function, while the set B is called the *co-domain*. There is a compact notation to capture all this information, namely $f : A \to B$. The value of a function f in a specific element a will be denoted by $f(a)$. In words, $f(a)$ is often called the image of a under f or sometimes also the evaluation of f in a. Instead of saying that f maps the value a in A to $f(a)$, one can also briefly write $a \mapsto f(a)$. All the notation so far for a function f can compactly be given as follows:

$$\begin{array}{rcl} f : A & \to & B \\ a & \mapsto & f(a) \end{array}$$

For example, the function sending a real number to its square can be given as:

$$\begin{array}{rcl} f : \mathbb{R} & \to & \mathbb{R} \\ x & \mapsto & x^2 \end{array}$$

A function like the previous is often also given as $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = x^2$. What is also done quite often is to simply say that the function is defined as $f(x) = x^2$. In such cases, it is left to the reader to figure out what the domain and the co-domain of the function is. Whenever possible, we will clearly indicate the domain and co-domain of functions. If the domain and the co-domain are chosen to be the same set A, one can define the *identity function* id_A on A. This is the function $\mathrm{id}_A : A \to A$ such that $a \mapsto a$.

The *image* of a function $f : A \to B$ is an important notion, which is defined as the set $\{f(a) \mid a \in A\}$. The image of a function $f : A \to B$ is a subset of its co-domain B, but we will see in Example 2.10 that image and co-domain do not have to be equal. Common notations for the image of a function $f : A \to B$ are $f(A)$ or $\mathrm{image}(f)$. Let us consider some examples:

|||| Example 2.10

Let us again consider the function

$$\begin{array}{rccc} f: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & x^2 \end{array}$$

This function has domain $\mathbb{R}$ and co-domain $\mathbb{R}$. We claim that $f(\mathbb{R}) = \{r \in \mathbb{R} \mid r \geq 0\}$. In other words, we claim that $f(\mathbb{R}) = \mathbb{R}_{\geq 0}$. Using Lemma 2.6, it is enough to show that $f(\mathbb{R}) \subseteq \mathbb{R}_{\geq 0}$ and $\mathbb{R}_{\geq 0} \subseteq f(\mathbb{R})$.

First of all, note that $f(\mathbb{R}) \subseteq \mathbb{R}_{\geq 0}$, since the square of a real number cannot be negative. Conversely, if $r \in \mathbb{R}_{\geq 0}$, then $\sqrt{r}$ is defined and $r = (\sqrt{r})^2 = f(\sqrt{r})$. This shows that any non-negative real number r is in the image of f. In other words, we have shown that $\mathbb{R}_{\geq 0} \subseteq f(\mathbb{R})$. Using Lemma 2.6, we may indeed conclude that $f(\mathbb{R}) = \mathbb{R}_{\geq 0}$.

This example shows that the image of a function does not have to be equal to its co-domain.

When considering the squaring function as we just did, we could of course right from the start have defined it as $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$, with $x \mapsto x^2$. Here the only difference is that we changed the co-domain from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$. For this modified function the image is the same as the co-domain, so why do we make such a distinction between the image and the co-domain of a function in the general theory? One reason is that it is convenient not to have to keep track of the image of a function all the time. If we know a function maps real numbers to real numbers, we simply can set the co-domain equal to $\mathbb{R}$ without worrying further. For complicated functions, it may be even be very difficult to compute its image.

Two functions $f : A \to B$ and $g : C \to D$ are equal precisely if they have the same domain, the same co-domain, and they assign the same values to each of the elements of their domain A. In formulas:

$$f = g \iff A = C \quad \wedge \quad B = D \quad \wedge \quad f(a) = g(a) \text{ for all } a \in A.$$

|||| Example 2.11

Consider the functions

$$\begin{array}{rccc} f: & \{0,1\} & \to & \{0,1\} \\ & a & \mapsto & a \end{array}$$

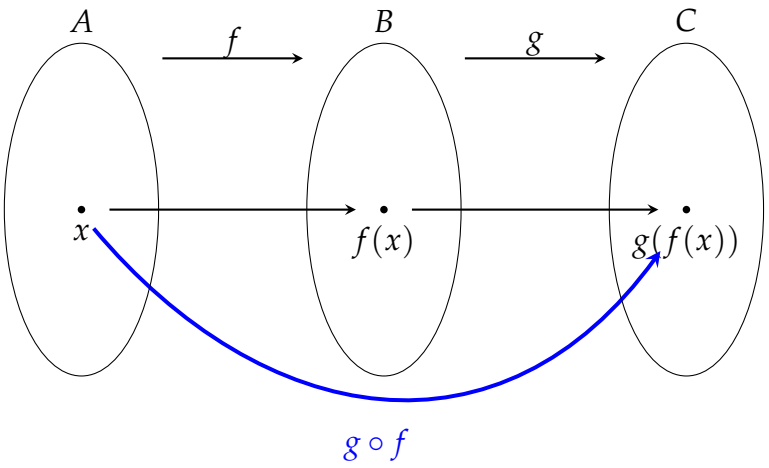


Figure 2.1: composition of the functions $f : A \to B$ and $g : B \to C$

and

$$\begin{array}{rcl} g : \{0,1\} & \to & \{0,1\} \\ a & \mapsto & a^2 \end{array}$$

The functions f and g have the same domain and co-domain. Moreover, $f(0) = 0$, $f(1) = 1$, while $g(0) = 0^2 = 0$ and $g(1) = 1^2 = 1$. Hence $f = g$.

This example shows that two functions may be the same even if they are described using different formulas.

If two functions $f : A \to B$ and $g : B \to C$ are given, it makes sense to consider the function

$$\begin{array}{rcl} h : A & \to & C \\ a & \mapsto & g(f(a)) \end{array}$$

The reason that in this definition the co-domain of the function f needs to be the same as the domain of the function g, is to guarantee that $g(f(a))$ is always defined: for any $a \in A$, we know that $f(a) \in B$, so that it indeed makes sense to use the elements $f(a)$ as input for the function g, since the domain of g is assumed to be B.

The function $h : A \to C$ obtained in this way is usually denoted by $g \circ f$ (pronounce: g **after** f) and is called the composition of g and f. Hence we have $(g \circ f)(a) = g(f(a))$.

|||| Example 2.12

Let us denote by $\mathbb{R}_{>0}$ the set of all positive real numbers. Suppose that $f : \mathbb{R} \to \mathbb{R}_{>0}$ is defined by $f(x) = x^2 + 1$ and $g : \mathbb{R}_{>0} \to \mathbb{R}$ is defined by $g(x) = \log_{10}(x)$, where $\log_{10}$ denotes the logarithm with base 10. Then $g \circ f : \mathbb{R} \to \mathbb{R}$ is the function sending $x \in \mathbb{R}$ to $\log_{10}(x^2 + 1)$. In other words:

$$\begin{array}{rccl} g \circ f : \mathbb{R} & \to & \mathbb{R} \\ x & \mapsto & \log_{10}(x^2 + 1) \end{array}$$

For example $(g \circ f)(3) = \log_{10}(3^2 + 1) = \log_{10}(10) = 1$.

|||| Lemma 2.13

Let A, B, C, and D be sets and suppose that we are given functions $h : A \to B$, $g : B \to C$, and $f : C \to D$. Then we have $(f \circ g) \circ h = f \circ (g \circ h)$.

Proof. First of all note that both $(f \circ g) \circ h$ and $f \circ (g \circ h)$ are functions from A to D, so they have the same domain and codomain. To prove the lemma it is therefore enough to show that for all $a \in A$, we have $((f \circ g) \circ h)(a) = (f \circ (g \circ h))(a)$. By definition of the composition $\circ$, we have

$$(f \circ (g \circ h))(a) = f((g \circ h)(a)) = f(g(h(a))),$$

while

$$((f \circ g) \circ h)(a) = (f \circ g)(h(a)) = f(g(h(a))).$$

We conclude that for any $a \in A$ it holds that $(f \circ (g \circ h))(a) = ((f \circ g) \circ h)(a)$, which is what we needed to show. □

The result from this lemma is usually stated as: composition of functions is an *associative* operation. Because of Lemma 2.13, it is common to simplify formulas involving composition of several functions, by leaving out the parentheses. For example, one simply writes $f \circ g \circ h$, when taking the composite of three functions.

Given a function $f : A \to B$, we say that the function f is *injective*, precisely if any two distinct elements from A are mapped to distinct elements of B. Writing this in terms of

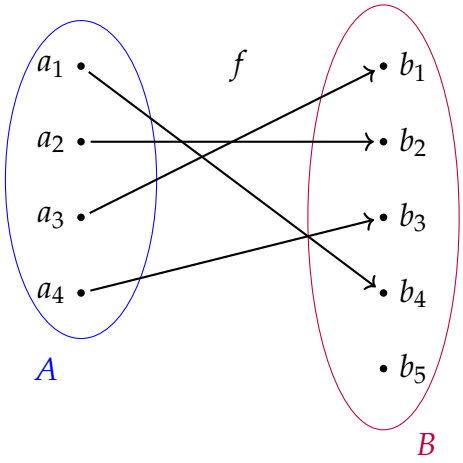


Figure 2.2: injective function $f : A \to B$

logical expressions, this means that:

$f : A \to B$ is injective if and only if for all $a_1, a_2 \in A$, $(a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2))$.

Using (1-21), it is logically equivalent to write:

$f : A \to B$ is injective if and only if for all $a_1, a_2 \in A$, $(f(a_1) = f(a_2) \Rightarrow a_1 = a_2)$.

This reformulation can be convenient in practice.

A function $f : A \to B$ is called *surjective* precisely if any element from B is in the image of f, that is:

$f : A \to B$ is surjective if and only if for all $b \in B$, there exists an $a \in A$ such that $b = f(a)$.

Using as before the notation $f(A)$ for the image of f, this can compactly be restated as: a function $f : A \to B$ is called surjective precisely if $f(A) = B$.

|||| Example 2.14

An example of a function that is injective, but not surjective, is $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}$ given by $f(x) = 1/x$. This function is not surjective, since its image actually is $\mathbb{R}\backslash\{0\}$, while its co-domain is $\mathbb{R}$. It is injective, since if $f(a) = f(b)$, that is if $1/a = 1/b$, then $a = b$.

An example of a function that is surjective, but not injective is $g : \mathbb{R} \to [-1, 1]$ given by $g(x) = \sin(x)$. This function is not injective, since for example 0 and π are both mapped to 0 by the sine function.

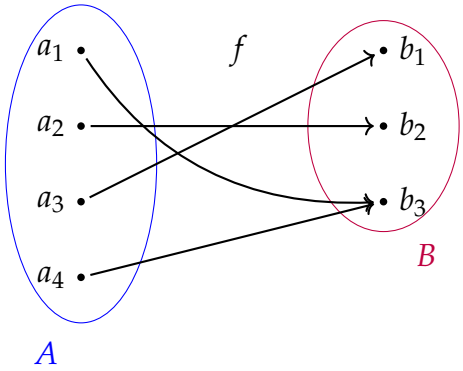


Figure 2.3: surjective function $f : A \to B$

A function $f : A \to B$ is called *bijective* if it is both injective and surjective. A bijective function is also called a *bijection*. Combining the definitions of injective and surjective, we see that function $f : A \to B$ is bijective precisely if for each $b \in B$ there exists a unique $a \in A$ such that $f(a) = b$. In the next section, we will see several examples of functions, but let us give an example here as well.

|||| Example 2.15

Consider the function $h : \{0,1,2\} \to \{3,4,5\}$ given by $h(x) = 5 - x$. Note that $h(0) = 5$, $h(1) = 4$ and $h(2) = 3$. Hence for any $b \in \{3,4,5\}$, there exists a unique $a \in \{0,1,2\}$ such that $h(a) = b$. We can conclude that h is a bijective function.

There is a very practical connection between bijective functions and inverse functions. Let us for completeness first define what the inverse of a function is.

|||| Definition 2.16

Let $f : A \to B$ be a function. A function $g : B \to A$ is called the *inverse function* of f if $f \circ g = \mathrm{id}_B$ (the identity function on B) and $g \circ f = \mathrm{id}_A$ (the identity function on A). The inverse of f will be denoted by f^{-1}.

Now we show that a function has an inverse precisely if it is a bijective function.

Lemma 2.17

Suppose that A and B are sets and let $f : A \to B$ be a function. Then f is bijective if and only if f has an inverse function.

Proof. Suppose that $f : A \to B$ is a bijection. As we have seen, a function $f : A \to B$ is bijective precisely if for any $b \in B$ there exists a unique $a \in A$ such that $f(a) = b$. The uniqueness of a implies that we can define a function $g : B \to A$ as $b \mapsto a$. We will show that g is the inverse function of f. Indeed if $b = f(a)$, we have

$$(f \circ g)(b) = f(g(b)) = f(a) = b \text{ and } (g \circ f)(a) = g(f(a)) = g(b) = a.$$

But this shows that $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$, which by Definition 2.16 means that $g = f^{-1}$.

Conversely, if f has an inverse function, then the equation $f(a) = b$ implies that $f^{-1}(f(a)) = f^{-1}(b)$. Since $a = (f^{-1} \circ f)(a) = f^{-1}(f(a))$, we see that $a = f^{-1}(b)$. Hence for any $b \in B$, there exists a unique element $a \in A$ such that $f(a) = b$ (namely $a = f^{-1}(b)$). This shows that f is bijective. □

Example 2.18

Let us again consider the function $h : \{0,1,2\} \to \{3,4,5\}$ given by $h(x) = 5 - x$ from Example 2.15. We have seen that the function h is bijective. Hence by Lemma 2.17, it has an inverse $h^{-1} : \{3,4,5\} \to \{0,1,2\}$. Recall that $h(0) = 5$, $h(1) = 4$ and $h(2) = 3$. The inverse of h simply sends the images back to the original values: $h^{-1}(5) = 0$, $h^{-1}(4) = 1$, and $h^{-1}(3) = 2$.

Note that actually the previous calculations show that $h^{-1}(x) = 5 - x$ for all $x \in \{3,4,5\}$. Hence $h^{-1} : \{3,4,5\} \to \{0,1,2\}$ is given by $h^{-1}(x) = 5 - x$. A small warning: the inverse of a function does not have to look similar to the function itself. Later we will see examples of inverse functions where this indeed is not the case.

2.2.1 Computational aspects of functions

The way we have looked at a function $f : A \to B$, we completely ignored more practical aspects like: given some $a \in A$, how do you actually compute $f(b)$? For the general

mathematical theory of functions, this is not an issue and the "inner workings" of the function f are then treated as a black box. However, for applications of the theory, it can be very important to know how to compute function values.

Fortunately, many useful functions can be computed using an *algorithm*. We will not go into the precise details on how to define what an algorithm really is, but take an intuitive view. Basically, an algorithm is a set of instructions that one could easily transform into a computer program if one would want to. These simple instructions involve "simple" operations like multiplication and addition. Moreover, intermediate results can be stored in memory and used later on in the algorithm if needed. More philosophically, an algorithm for a function f opens the black box and shows its "inner workings". Let us consider the example of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$. A first attempt to describe an algorithm that given x, computes $f(x)$ could be:

Step 1. Compute $x \cdot x$ and remember the outcome of this computation.

Step 2. Take the outcome of Step 1 and multiply it by x.

Step 3. Return the value from Step 2.

A bit more formally, we can rewrite this as:

Step 0. Denote by x the given input.

Step 1. Compute $x \cdot x$ and store the outcome under the name y.

Step 2. Compute $x \cdot y$ and store the outcome under the name z.

Step 3. Return z.

To make the description look even more like a computer algorithm, we will write it in what is known as *pseudo-code*. The main difference with the previous description is that a phrase like "Compute $x \cdot x$ and store the outcome under the name y" is compactly written as "$y \leftarrow x \cdot x$". The algorithmic pseudo-code description of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ then becomes:

Algorithm 2 for $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$

Input: $x \in \mathbb{R}$

1: $y \leftarrow x \cdot x$
2: $z \leftarrow x \cdot y$
3: **return** z

Let us consider another example:

Example 2.19

Let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be defined by $x \mapsto |x|$. Here $|x|$ denotes the absolute value of x. Just as we observed in Example 1.17, we have that if $x < 0$, then $|x| = -x$, while if $x \geq 0$, then $|x| = x$. For this reason the absolute value is often defined in the following way:

$$|x| = \begin{cases} -x & \text{if } x < 0, \\ x & \text{otherwise.} \end{cases}$$

When defining a function by cases like this, it is important to check: 1) that all elements of the domain of the function appear in one of the cases and 2) that an element of the domain of the function appears in no more than one of the cases. Here the domain of the function is $\mathbb{R}$. First of all $\mathbb{R}$ is the union of $\mathbb{R}_{<0}$ and $\mathbb{R}_{\geq 0}$, so 1) is satisfied. Moreover, $\mathbb{R}_{<0}$ and $\mathbb{R}_{\geq 0}$ are disjoint sets, so that 2) is satisfied. In other words: 1) and 2) are satisfied, because the domain of the function, $\mathbb{R}$, is the disjoint union of $\mathbb{R}_{<0}$ and $\mathbb{R}_{\geq 0}$. The given description of the absolute value function can easily be reformulated as an algorithm in pseudo-code:

Algorithm 3 to compute $|x|$ for $x \in \mathbb{R}$

```
    Input: x ∈ ℝ
1:  if x < 0 then
2:      return −x
3:  else
4:      return x
```

2.3 Examples of functions

To exemplify the theory of functions as developed above, let us now consider some elementary functions $f : A \to B$, where A and B are subsets of $\mathbb{R}$. To help us to show injectivity of such functions, we use the following lemma:

Lemma 2.20

Let $f : A \to B$ be a function and assume that A and B are subsets of $\mathbb{R}$. Suppose that either

$$\text{for all } a_1, a_2 \in A \text{ it holds that: } a_1 < a_2 \Rightarrow f(a_1) < f(a_2) \tag{2-12}$$

or

$$\text{for all } a_1, a_2 \in A \text{ it holds that: } a_1 < a_2 \Rightarrow f(a_1) > f(a_2). \tag{2-13}$$

Then f is an injective function.

Proof. Assume that the function f satisfies Equation (2-12). Let a_1 and a_2 be distinct elements of A. Since $a_1 \neq a_2$, we know that either $a_1 < a_2$ or $a_2 < a_1$. If $a_1 < a_2$, Equation (2-12) implies that $f(a_1) < f(a_2)$. If $a_2 < a_1$, Equation (2-12) implies $f(a_2) < f(a_1)$. In either case, we may conclude that $f(a_1) \neq f(a_2)$. Hence f is injective. If the function f satisfies Equation (2-13), a similar reasoning shows that f is injective as well. □

A function f satisfying Equation (2-12) or Equation (2-13) is called *strictly monotone*. More precisely, a function f satisfying Equation (2-12) is called *strictly increasing*, while if a function f satisfies Equation (2-13), it is called *strictly decreasing*. Hence Lemma 2.20 can be summarized as: a strictly monotone function is injective.

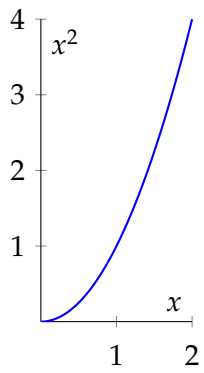
Example 2.21

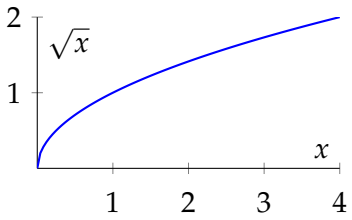
Consider the function $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = x^2$. We have already seen in Example 2.10 that the image of this function equals $\mathbb{R}_{\geq 0}$. In other words, $f(\mathbb{R}) = \mathbb{R}_{\geq 0}$. The function f is therefore not surjective. In fact, it is not injective either, since for example $f(-1) = 1$ and $f(1) = 1$.

Since the function f is not bijective, it does not have an inverse. Nonetheless, we can modify the domain and the co-domain of f so that the resulting function is bijective. First of all, we can create a function $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by $g(x) = x^2$. The difference between the functions f and g is subtle: only their co-domains are different. Therefore, even though for any real number x, it is true that $f(x) = g(x)$, we still consider the functions f and g to be two different functions. The reason for introducing the function g is that g is surjective, since $g(\mathbb{R}) = \mathbb{R}_{\geq 0}$ and $\mathbb{R}_{\geq 0}$ is the co-domain of g. However, g still does not have an inverse, since g is not injective. Indeed, the reason is the same as why f was not injective. We still have for example that $g(1) = 1$ and $g(-1) = 1$. What we do next is to introduce yet another function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by $h(x) = x^2$. The function h has the same co-domain as the function

g, but note that the domain of the function h is a subset of that of g. Indeed, the domain of h is $\mathbb{R}_{\geq 0}$, which is a strict subset of $\mathbb{R}$, the domain of g. Now one can show that the function h is strictly monotone and therefore by Lemma 2.20 injective. We already have seen that h is surjective, so we may conclude that it is bijective. By Lemma 2.17, the function h therefore has an inverse. Since for any $x \in \mathbb{R}_{\geq 0}$, it holds that $\sqrt{x^2} = x$ and $(\sqrt{x})^2 = x$, we see that the inverse of h is the function $h^{-1} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by $h^{-1}(x) = \sqrt{x}$.

To illustrate the situation, we have plotted (parts of) the graphs of the functions h and its inverse h^{-1}. Note that the graph of h^{-1} is the mirror image of the graph of h in the line $y = x$. From the graph of h we can also see the it is a strictly increasing function.

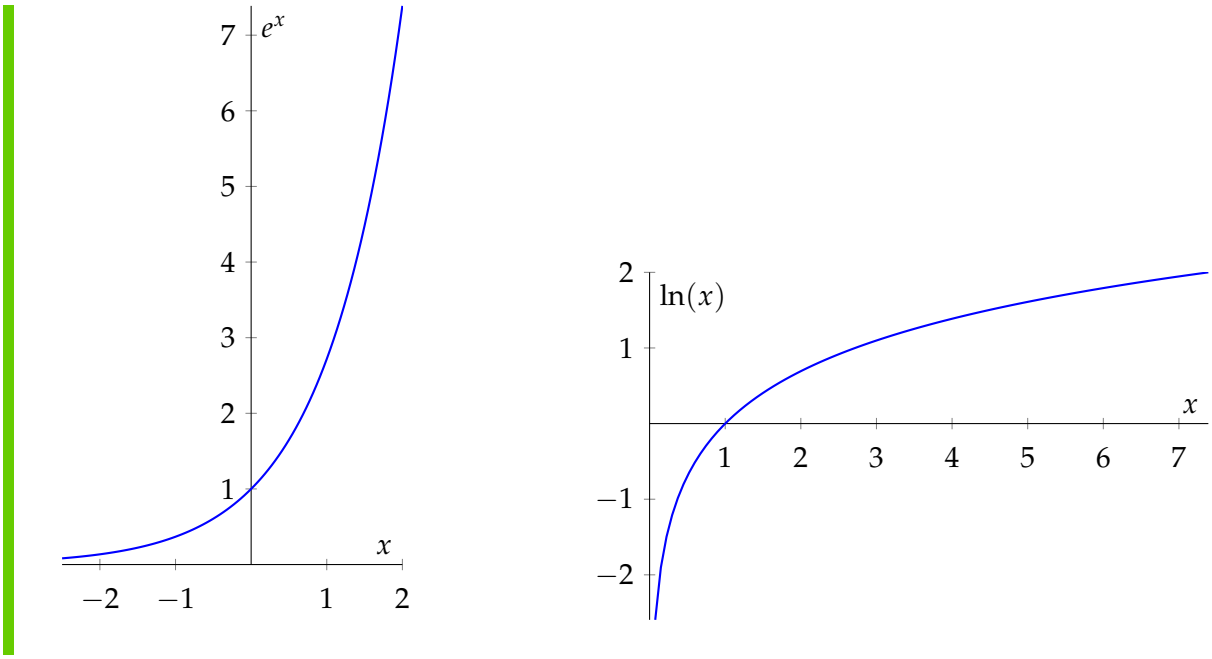




Example 2.22

Let e denote the base of the natural logarithm. The constant e is sometimes called Euler's number and is approximately equal to 2.71828. The exponential function $\exp : \mathbb{R} \to \mathbb{R}_{>0}$ is defined by $x \mapsto e^x$. It is a strictly increasing function and therefore injective. Further, the image of the exponential function is $\mathbb{R}_{>0}$, which implies that it is surjective. Combining this we see that exp is a bijective function. Its inverse is commonly denoted by $\ln : \mathbb{R}_{>0} \to \mathbb{R}$. In particular, we have $\ln(e^x) = x$ for all $x \in \mathbb{R}$ and $e^{\ln(x)} = x$ for all $x \in \mathbb{R}_{>0}$.

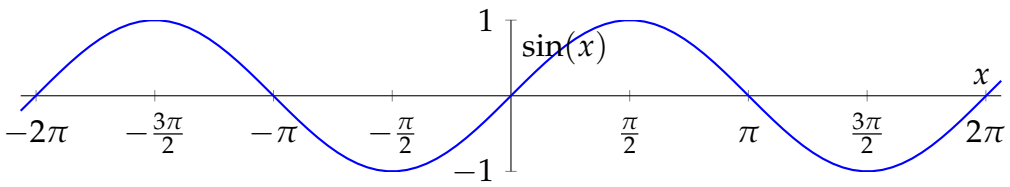
We plot the graphs of the functions exp and ln to illustrate the situation.



2.3.1 The trigonometric functions sin, cos and tan.

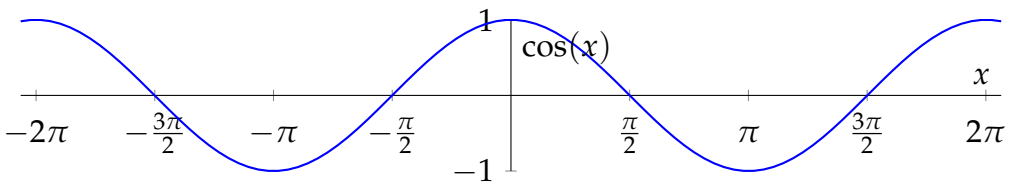
The *trigonometric functions* sine, cosine and tangent are extremely useful examples of functions and will appear again in various contexts later on. Therefore we briefly revisit them in this subsection.

First of all, the sine function is usually denoted by sin, but let us in light of our definition of functions specify which domain and co-domain it has. First of all, we define the sine function $\sin : \mathbb{R} \to [-1,1]$ to be the function such that $x \mapsto \sin(x)$. The image of sin is $[-1,1]$, meaning that sin is a surjective function. It is not an injective function, since distinct real numbers can have the same value under the sine function. For example, one has $\sin(0) = \sin(\pi) = 0$. The graph of the sine function is as follows:

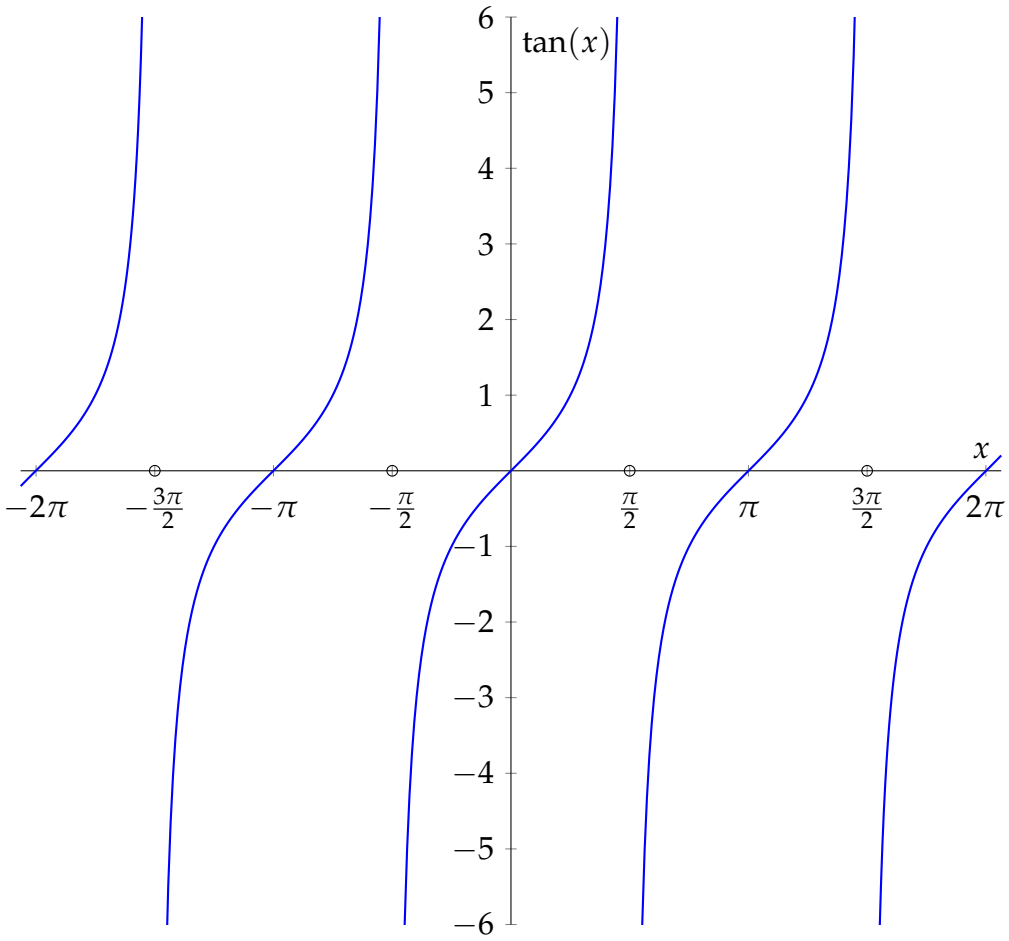


Similarly, we define $\cos : \mathbb{R} \to [-1,1]$. Again, the co-domain is chosen to be the closed

interval $[-1,1]$, which means that the function cos will be surjective. It is not injective though, since for example $\cos(-\pi/2) = \cos(\pi/2) = 0$. The graph of the cosine function is:



A third commonly used trigonometric function is the tangent function. Loosely speaking, we have $\tan(x) = \sin(x)/\cos(x)$, but this formula only makes sense for $x \in \mathbb{R}$ such that $\cos(x) \neq 0$. Therefore, we can define $\tan : \{x \in \mathbb{R} \mid \cos(x) \neq 0\} \to \mathbb{R}$, where $\tan(x) = \sin(x)/\cos(x)$. Since $\{x \in \mathbb{R} \mid \cos(x) \neq 0\} = \mathbb{R} \setminus \{x \in \mathbb{R} \mid \cos(x) = 0\}$ and $\{x \in \mathbb{R} \mid \cos(x) = 0\} = \{\dots, -3\pi/2, -\pi/2, \pi/2, 3\pi/2, \dots\}$, we can also say that the domain of the tangent function is the set $\mathbb{R} \setminus \{\dots, -3\pi/2, -\pi/2, \pi/2, 3\pi/2, \dots\}$. The graph of the tangent function is as follows:

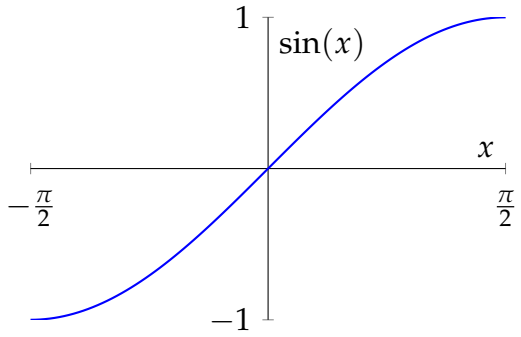


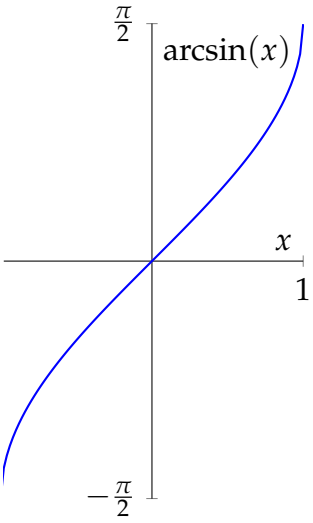
The small circles on the x-axis indicate the values of x for which the tangent function is not defined. The tangent function is surjective, since its image is $\mathbb{R}$. Just as the sine and cosine functions, it is not injective. We have for example $\tan(0) = 0$, but also $\tan(\pi) = 0$.

2.3.2 The inverse trigonometric functions

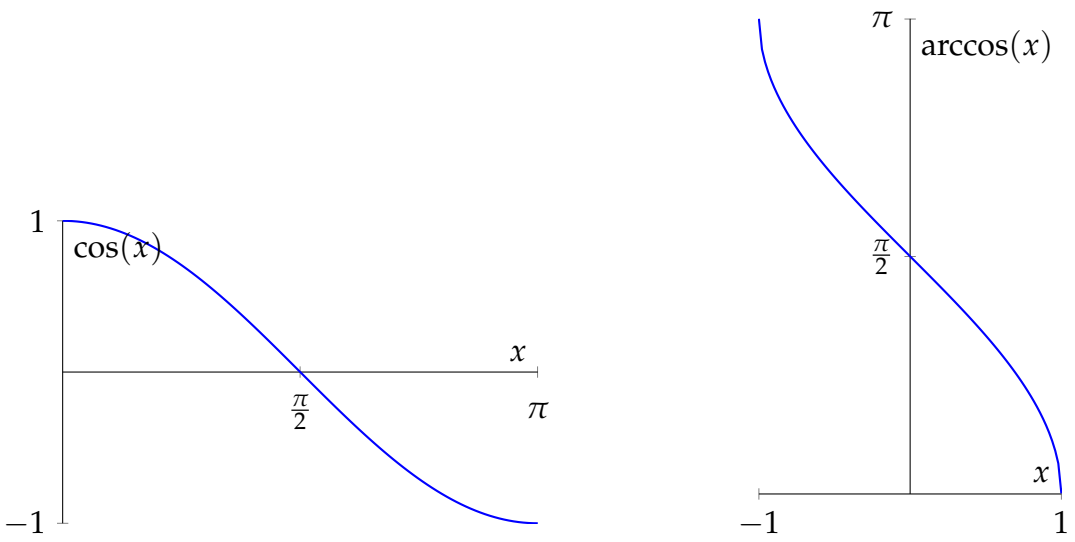
Since none of the trigonometric functions sin, cos and tan discussed in the previous subsection are bijections, we cannot find inverses for these functions. However, just as in Example 2.21, we can modify the domain of these functions and obtain functions that do have an inverse. These inverses are known as the *inverse trigonometric functions* (sometimes also as the arcus functions). In this subsection, we give the details of how these are defined.

First of all, if the domain of the function $\sin : \mathbb{R} \to [-1,1]$ is restricted to the closed interval $[-\pi/2, \pi/2]$, one obtains a function $f : [-\pi/2, \pi/2] \to [-1,1]$ defined by $f(x) = \sin(x)$. The function f is a bijective function, since the graph of the sine function is strictly increasing on the interval $[-\pi/2, \pi/2]$ with values from -1 to 1. The inverse of this function is called the *arcsine* and usually in mathematical formulas denoted by arcsin. Hence $\arcsin : [-1,1] \to [-\pi/2, \pi/2]$ is the inverse of the sine function whose domain has been restricted to $[-\pi/2, \pi/2]$. The graphs of these two functions look as follows:

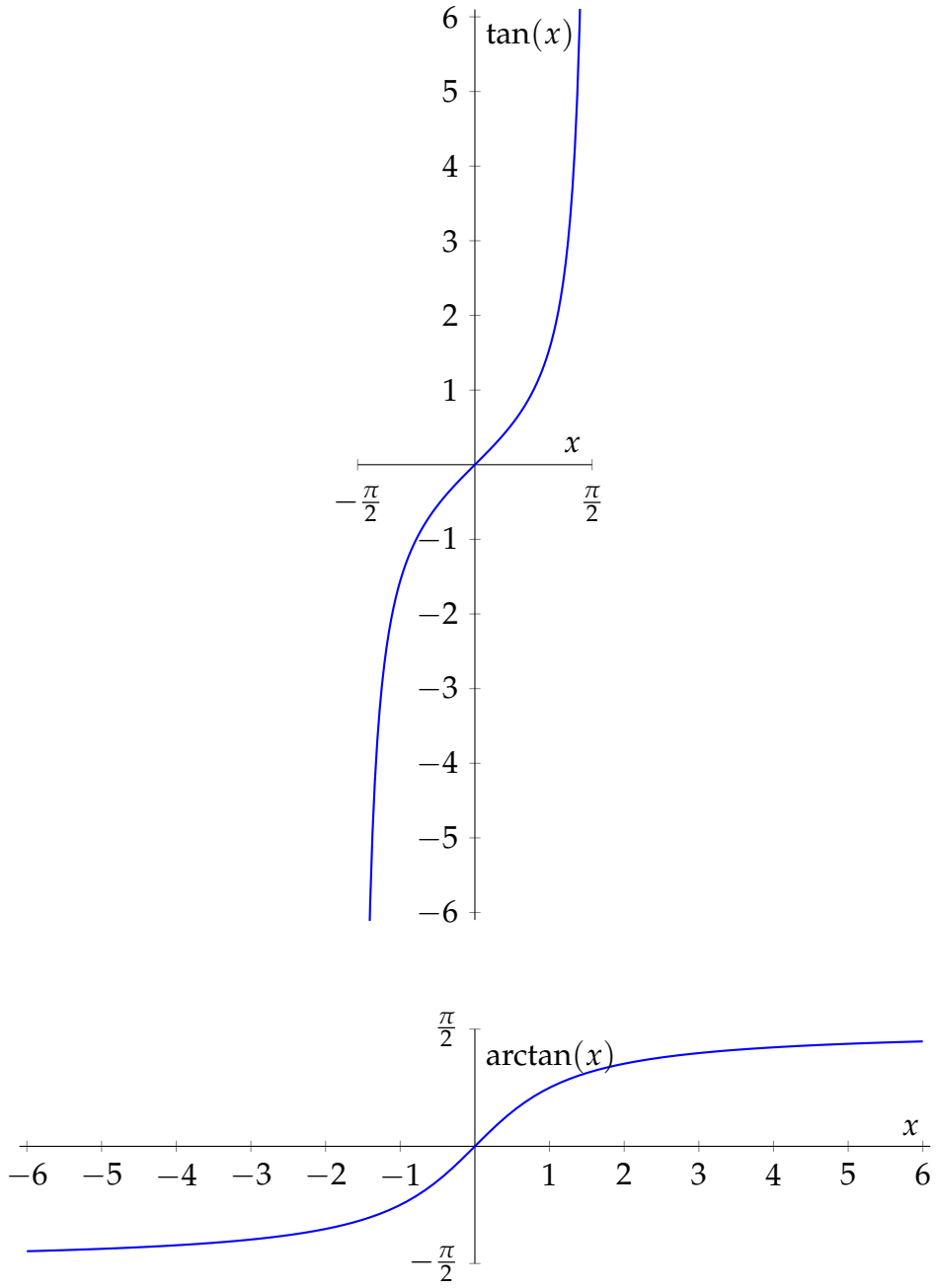




In a very similar way, we can define the *arccosine* function. First we restrict the domain of the usual cosine function to the closed interval $[0, \pi]$. The resulting function $g : [0, \pi] \to [-1,1]$, where $g(x) = \cos(x)$, is strictly decreasing as well as surjective and thus bijective. The inverse of g is the arccosine function. It is usually denoted by arccos. Hence $\arccos : [-1,1] \to [0, \pi]$ is the inverse of the cosine function when its domain is restricted to $[0, \pi]$. We illustrate the situation by showing the graphs of these two functions:



Finally, we discuss the tangent function. In this case, we simply consider the function $h :]-\pi/2, \pi/2[\to \mathbb{R}$, where $h(x) = \tan(x)$. In other words, the function h is simply the tangent function with its domain restricted to the open interval $]-\pi/2, \pi/2[$. The function h is a strictly increasing function with image $\mathbb{R}$, which implies that h is a bijection. The inverse of h is called the *arctangent* function, commonly denoted in formulas as arctan. More precisely, $\arctan : \mathbb{R} \to]-\pi/2, \pi/2[$ is the inverse of the tangent function with domain restricted to $]-\pi/2, \pi/2[$. As before, we illustrate the situation by showing the graphs of these functions:



|||| Example 2.23

Let us determine some values of the inverse trigonometric functions. Since $\sin(0) = 0$, we have $\arcsin(0) = 0$. However, even though $\sin(\pi) = 0$, we do not have $\arcsin(0) = \pi$. Indeed a function cannot take two distinct values for the same input! The issue is that arcsin is the inverse of the sine function with domain restricted to $[-\pi/2, \pi/2]$. Therefore $\sin(x) = y$ only

implies $\arcsin(y) = x$ as long as $x \in [-\pi/2, \pi/2]$. For example, since $\sin(\pi/4) = \sqrt{2}/2$, we have $\arcsin(\sqrt{2}/2) = \pi/4$.

For the arccos, we have a similar phenomenon. One has $\cos(-\pi/4) = \sqrt{2}/2$, but this does not imply $\arccos(\sqrt{2}/2) = -\pi/4$. This time the issue is that the domain of the cosine function was restricted to $[0, \pi]$, when defining the arccos function. On the interval $[0, \pi]$ the cosine does take the value $\sqrt{2}/2$, namely for $x = \pi/4$. Therefore $\arccos(\sqrt{2}/2) = \pi/4$.

As a final example, we have $\cos(\pi/3) = 1/2$ and $\sin(\pi/3) = \sqrt{3}/2$. Therefore $\tan(\pi/3) = \sin(\pi/3)/\cos(\pi/3) = \sqrt{3}$. The arctan function is the inverse of the tangent function with its domain restricted to $]-\pi/2, \pi/2[$. Since $\pi/3 \in]-\pi/2, \pi/2[$, we may therefore conclude that $\arctan(\sqrt{3}) = \pi/3$.

Note 3

Complex numbers

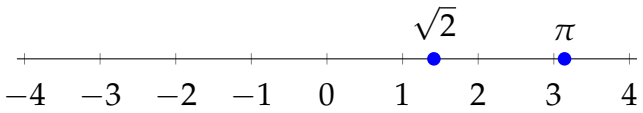
3.1 Introduction to the complex numbers

In this chapter we will introduce the set of *complex numbers*, commonly denoted by $\mathbb{C}$. These complex numbers turn out to be extremely useful and no modern scientist or engineer can do without them anymore. Let us first take a short look at some other sets of numbers in mathematics. The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ have, as their name already suggests, a very natural interpretation. They come up when one wants to count things. The integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ came around when differences of natural numbers were needed. We have also seen the set of rational numbers $\mathbb{Q}$ in Example 2.4, which consists of fractions of integers.

One may think that the set of rational numbers $\mathbb{Q}$ contains all numbers one would ever need, but this is not the case. For example, it turns out that the equation $z^2 = 2$ does not have a solution in $\mathbb{Q}$. Instead of saying that such an equation simply does not have any solutions, mathematicians extended the set of rational numbers $\mathbb{Q}$ to the set of real numbers $\mathbb{R}$. Within $\mathbb{R}$, the equation $z^2 = 2$ has two solutions, namely $\sqrt{2}$ and $-\sqrt{2}$. The set $\mathbb{R}$ is very large and contains many interesting numbers, such as e, the base of the natural logarithm, and π. Often, the real numbers $\mathbb{R}$ are represented as a straight line, which we will call the *real line*. Every point on the real line corresponds to a real number (see Figure 3.1).

Again for some time it was thought that the set of real numbers $\mathbb{R}$ would contain all numbers one would ever want to use. But what about an equation like $z^2 = -1$? It is clear that within the set of real numbers, this equation does not have any solutions. We are again in the same situation as before with the equation $z^2 = 2$ before the real

Figure 3.1: The real line.



numbers were introduced. We simply try to find a set of numbers even larger than $\mathbb{R}$ that does contain a solution to the equation $z^2 = -1$. It would be natural to denote a solution to $z^2 = -1$ by $\sqrt{-1}$, but it is more common to write i instead. Hence we want that $i^2 = -1$. Now we simply define the complex numbers as follows.

Definition 3.1

The set $\mathbb{C}$ of complex numbers is defined as:

$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}.$$

The complex number i satisfies the rule

$$i^2 = -1.$$

The expression $a + bi$ should simply be thought of as a polynomial in the variable i. Hence it holds for example that $a + bi = a + ib$. Also, it makes no difference to write $a + b \cdot i$ instead of $a + bi$. Hence we have for all $a, b \in \mathbb{R}$:

$$a + bi = a + b \cdot i = a + i \cdot b = a + ib.$$

Finally, just like for polynomials, $a + bi$ denotes exactly the same complex number as $bi + a$.

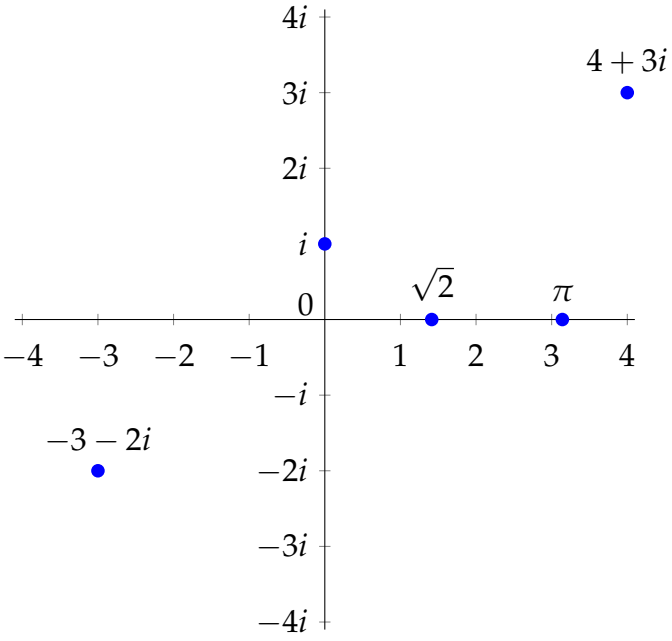
For any $a, b, c, d \in \mathbb{R}$, the two complex numbers $a + bi$ and $c + di$ are the same if and only if $a = c$ and $b = d$. If $a = 0$ it is customary to simplify $0 + bi$ to bi. In other words $0 + bi = bi$. Similarly, if $b = 0$, one typically writes a instead of $a + 0i$. Finally, if $b = 1$, the 1 in front of the i is often omitted. For example, $5 + 1i = 5 + i$. Using all the above, one has for example $i = 1i = 0 + 1i = 0 + 1 \cdot i$. The set of complex numbers $\mathbb{C}$ contains the set of real numbers $\mathbb{R}$, because for $a \in \mathbb{R}$, we have $a = a + 0i$. In other words: $\mathbb{R} \subseteq \mathbb{C}$. In fact $\mathbb{R} \subsetneq \mathbb{C}$, since $i \in \mathbb{C}$, while $i \notin \mathbb{R}$.

The complex numbers can represented graphically, but now as a plane called the *complex plane*. A complex number $a + bi$ is represented as the point (a, b) in that plane. This means that the number i has coordinates $(0, 1)$ and therefore will lie on the second axis. The number i and some other complex numbers have been drawn in the complex plane in Figure 3.2.

The axes in the complex plane have a special name. The horizontal axis is called the *real axis*, because all real numbers lie on it. Indeed, a number on the real axis in the complex plane will be of the form $a + 0i$ for some $a \in \mathbb{R}$.

The vertical axis is called the *imaginary axis*. In fact, the symbol i is an abbreviation of the word imaginary. The numbers that lie on the vertical axis are called *purely imaginary numbers*. The expressions "complex numbers" and "imaginary numbers" are historical and show that at some point in time scientists struggled to understand these numbers. Nowadays, the complex numbers are completely standard.

Figure 3.2: The complex plane.



The coordinates for a complex number $z \in \mathbb{C}$ in the complex plane have a special name. The first coordinate is called the *real part* of z (denoted by $\mathrm{Re}(z)$), while the second coordinate of z is called the *imaginary part* (denoted by $\mathrm{Im}(z)$). If one knows $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$, one can compute the number z, because it holds that

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

If a complex number z is written in the form $\mathrm{Re}(z) + \mathrm{Im}(z)i$, then one says that the number z is written in *rectangular form*. For a given complex number z, the pair $(\mathrm{Re}(z), \mathrm{Im}(z))$ is called the *rectangular coordinates* of z.

|||| Example 3.2

Compute the rectangular coordinates of the following complex numbers:

1. $2 + 3i$
2. $\sqrt{2}$
3. i

Answer:

1. The number $2 + 3i$ is in rectangular form. Therefore, we can read off the real and imaginary part directly. We have $\mathrm{Re}(2 + 3i) = 2$ and $\mathrm{Im}(2 + 3i) = 3$. Hence the rectangular coordinates of the complex number $2 + 3i$ are $(2, 3)$.
2. The number $\sqrt{2}$ is a real number, but we can also view it as a complex number, since $\sqrt{2} = \sqrt{2} + 0i$. From this we see that $\mathrm{Re}(\sqrt{2}) = \sqrt{2}$ and $\mathrm{Im}(\sqrt{2}) = 0$. All real numbers have in fact imaginary part equal to 0. The rectangular coordinates of $\sqrt{2}$ are $(\sqrt{2}, 0)$.
3. The number i is a purely imaginary number and one could also write $i = 0 + 1 \cdot i$. Therefore we have $\mathrm{Re}(i) = 0$ and $\mathrm{Im}(i) = 1$. All purely imaginary numbers have real part 0. The rectangular coordinates of i are $(0, 1)$.

3.2 Arithmetic with complex numbers

Now that we have introduced the complex numbers, we can start to investigate how much structure they have. We are used to being able to add two numbers, subtract them, multiply them and divide them. It is not clear at this point if this can be done with complex numbers, but we will see that this is possible.

We start by defining an addition and a subtraction.

Definition 3.3

Let $a, b, c, d \in \mathbb{R}$ and let $a + bi$ and $c + di$ be two complex numbers in $\mathbb{C}$ written in rectangular form. Then we define:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$(a + bi) - (c + di) = (a - c) + (b - d)i.$$

The addition or subtraction of two complex numbers is very similar to the addition or subtraction of two polynomials of degree one (polynomials will be defined more precisely in Definition 4.1). One simply collect the terms not involving i and the terms involving i. One can therefore remember the addition by for example adding the following intermediate steps:

$$\begin{aligned}(a + bi) + (c + di) &= a + bi + c + di \\ &= a + c + bi + di \\ &= (a + c) + (b + d)i\end{aligned}$$

The subtraction can be explained similarly. Graphically, the addition of complex numbers is like the addition of two vectors in the plane, see Figure 3.3. Note that $(a + bi) + (c + di) = (c + di) + (a + bi)$. Hence, when adding several complex numbers, the order in which one adds these numbers does not matter.

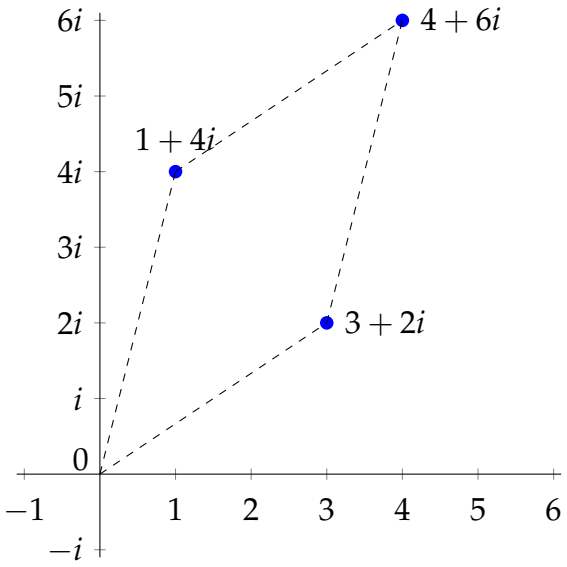
Example 3.4

Simplify the following expressions and write the outcome in rectangular form.

1. $(3 + 2i) + (1 + 4i)$
2. $(3 + 2i) - (1 + 4i)$
3. $(5 - 7i) - i$
4. $(5 - 7i) - (-10 + i)$

Answer:

Figure 3.3: Addition of complex numbers. Here it is shown graphically that $(3+2i)+(1+4i)=4+6i$.



1. $(3+2i)+(1+4i)=(3+1)+(2+4)i=4+6i$
2. $(3+2i)-(1+4i)=(3-1)+(2-4)i=2-2i$
3. $(5-7i)-i=5+(-7-1)i=5-8i$
4. $(5-7i)-(-10+i)=(5-(-10))+(-7-1)i=15-8i$

Now that we have the addition and subtraction of complex numbers in place, let us take a look at their multiplication. Suppose for example that we would want to multiply the complex numbers $a+bi$ and $c+di$, where as usual $a,b,c,d\in\mathbb{R}$. First of all, let us see what happens if we simply multiply these expressions viewed as polynomials in the variable i:

$$(a+bi)\cdot(c+di)=a\cdot(c+di)+bi\cdot(c+di)=a\cdot c+a\cdot d\,i+b\cdot c\,i+b\cdot d\,i^2.$$

Till now, the only thing we have done is to simplify the product to get rid of the parentheses. But now we should remember that the whole point of introducing i was that it is a solution to the equation $z^2=-1$. Hence $i^2=-1$. If we use this, we get

$$(a+bi)\cdot(c+di)=a\cdot c+a\cdot di+b\cdot ci+b\cdot d\cdot(-1)=(a\cdot d-b\cdot c)+(a\cdot d+b\cdot c)i.$$

We arrived again at a complex number! All we needed to use were the usual rules of computation (when we got rid of the parentheses) and the formula $i^2=-1$. Let us

therefore take the formula we just found and put it as the formal definition of multiplication of complex numbers.

Definition 3.5

Let $a, b, c, d \in \mathbb{R}$ and let $a + bi$ and $c + di$ be two complex numbers in $\mathbb{C}$ given in rectangular form. We define:

$$(a + bi) \cdot (c + di) = (a \cdot c - b \cdot d) + (b \cdot c + a \cdot d)i.$$

There is no need to memorize the above definition. To calculate a product of two complex numbers in rectangular form, all one needs to do is to remember how we obtained it: we simplified the product by multiplying out all terms and then used that $i^2 = -1$. Note that $(a + bi) \cdot (c + di) = (c + di) \cdot (a + bi)$, so the order of the complex numbers does not matter in a multiplication. One says that multiplication of complex numbers is *commutative*. We will see in Section 3.3 that the multiplication of two complex numbers also can be described geometrically.

Example 3.6

Simplify the following expression and write the result in rectangular form.

1. $(1 + 2i) \cdot (3 + 4i)$
2. $(4 + i) \cdot (4 - i)$

Answer:

1.

$$\begin{aligned}(1 + 2i)(3 + 4i) &= 1 \cdot 3 + 1 \cdot 4i + 2i \cdot 3 + 2i \cdot 4i \\ &= 3 + 4i + 6i + 8i^2 \\ &= 3 + 10i - 8 \\ &= -5 + 10i.\end{aligned}$$

2.

$$\begin{aligned}(4+i)\cdot(4-i) &= 4\cdot 4+4\cdot(-i)+i\cdot 4-i^2\\ &= 16-4i+4i-(-1)\\ &= 17+0i\\ &= 17.\end{aligned}$$

In this case the outcome is actually a real number.

Part two of this example shows that the product of two nonreal numbers can be a real number. This example is actually a special case of the following lemma:

Lemma 3.7

Let $a, b \in \mathbb{R}$ and $z = a + bi$ a complex number in rectangular form. Then

$$(a+bi)\cdot(a-bi) = a^2+b^2.$$

Proof. We have

$$\begin{aligned}(a+bi)\cdot(a-bi) &= a\cdot a+a\cdot(-bi)+(bi)\cdot a-b\cdot bi^2\\ &= a^2-abi+abi-b^2\cdot(-1)\\ &= a^2+b^2.\end{aligned}$$

□

Motivated by this lemma, we introduce the following:

Definition 3.8

Let $z \in \mathbb{C}$ be a complex number. Suppose that $z = a + bi$ in rectangular form. Then we define the complex conjugate of z as $\overline{z} = a - bi$. The function from $\mathbb{C}$ to $\mathbb{C}$ defined by $z \mapsto \overline{z}$ is called the *complex conjugation* function.

Note that directly from this definition, we see that $\mathrm{Re}(\overline{z}) = \mathrm{Re}(z)$ and $\mathrm{Im}(\overline{z}) = -\mathrm{Im}(z)$. Hence,

$$\overline{z} = \mathrm{Re}(z) - \mathrm{Im}(z)i.$$

Therefore Lemma 3.7 implies that

$$z \cdot \overline{z} = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2. \tag{3-1}$$

Note that this equation implies that for any $z \in \mathbb{C}$, the product $z \cdot \overline{z}$ is a real number.

Complex conjugation turns out to be useful for defining division of complex numbers. We would like to be able to divide any complex number by any nonzero complex number. Note that we already are able to divide a complex number $a + bi \in \mathbb{C}$ by a nonzero real number $c \in \mathbb{R}$ by defining:

$$\frac{a+bi}{c} = \frac{a}{c} + \frac{b}{c}\,i \quad a, b \in \mathbb{R} \text{ and } c \in \mathbb{R} \setminus \{0\}.$$

The trick to divide any complex number $z_1 = a + bi$ by any nonzero complex number $z_2 = c + di$ is to observe the following:

$$\frac{z_1}{z_2} = \frac{a+bi}{c+di} = \frac{a+bi}{c+di} \cdot \frac{c-di}{c-di} = \frac{(a+bi)\cdot(c-di)}{c^2+d^2}. \tag{3-2}$$

The numerator of the righthand side in this equation is just a product of two complex numbers, which we know how to handle already. The denominator is a nonzero real number, namely $c^2 + d^2$, and we also already know how to divide a complex number by a real number. Let us make sure that the denominator $c^2 + d^2$ indeed is nonzero real number. First of all, it is a real number, since c and d are real numbers. Second of all, since the square of a real number cannot be a negative, we see that $c^2 \geq 0$, $d^2 \geq 0$. The only way $c^2 + d^2 = 0$ can hold is therefore if both $c^2 = 0$ and $d^2 = 0$. But then $c = 0$ and $d = 0$, implying that $c + di = 0$, contrary to our assumption that we were attempting to divide by a nonzero complex number.

Looking back at the way we defined division by a complex number, we see that the main ingredient was that if $z_1 \in \mathbb{C}$ and $z_2 \in \mathbb{C}\setminus\{0\}$, then the main idea for computing z_1/z_2 was to multiply both numerator and denominator with the complex conjugate of z_2, since then the denominator becomes $z_2 \cdot \overline{z_2}$, which is a real number. Equation (3-2) allows us therefore to divide by nonzero complex numbers. A special case of Equation (3-2) is the following:

$$\frac{1}{c+di} = \frac{1}{c+di} \cdot \frac{c-di}{c-di} = \frac{c-di}{c^2+d^2} = \frac{c}{c^2+d^2} - \frac{d}{c^2+d^2}\,i. \tag{3-3}$$

Now, let us consider some examples:

|||| Example 3.9

Simplify the following expressions and write the result in rectangular form.

1. $1/(1+i)$
2. $\dfrac{1+2i}{3+4i}$

Answer:

1. Note that $1/(1+i)$ is just a different way to write $\frac{1}{1+i}$. Hence we obtain using Equation (3-2), or alternatively Equation (3-3):

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

2. Using Equation (3-2), we find

$$\begin{aligned}\frac{1+2i}{3+4i} &= \frac{(1+2i)(3-4i)}{(3+4i)(3-4i)} = \frac{3-4i+6i-8i^2}{3^2+4^2} \\ &= \frac{3+2i+8}{9+16} = \frac{11+2i}{25} = \frac{11}{25} + \frac{2}{25}i.\end{aligned}$$

Let us collect various properties of multiplication and addition together in one theorem. We will not prove the theorem, though several of the statements have actually already been shown in the previous.

Theorem 3.10

Let $\mathbb{C}$ be the set of complex numbers and let $z_1, z_2, z_3 \in \mathbb{C}$ be chosen arbitrarily. Then the following properties are satisfied:

1. Addition and multiplication are *associative*: $z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3$, and $z_1 \cdot (z_2 \cdot z_3) = (z_1 \cdot z_2) \cdot z_3$.
2. Addition and multiplication are *commutative*: $z_1 + z_2 = z_2 + z_1$, and $z_1 \cdot z_2 = z_2 \cdot z_1$.
3. *Distributivity* of multiplication over addition holds: $z_1 \cdot (z_2 + z_3) = z_1 \cdot z_2 + z_1 \cdot z_3$.

Further one has for complex numbers, similarly as for the real numbers, the following properties:

Theorem 3.11

1. Addition and multiplication have a neutral element: the elements 0 and 1 in $\mathbb{C}$ satisfy $z + 0 = z$ and $z \cdot 1 = z$ for all $z \in \mathbb{C}$.
2. Additive inverses exist: for every $z \in \mathbb{C}$, there exists an element in $\mathbb{C}$, denoted $-z$, called the additive inverse of z, such that $z + (-z) = 0$.
3. Multiplicative inverses exist: for every $z \in \mathbb{C} \setminus \{0\}$, there exists an element in $\mathbb{C}$, denoted by z^{-1} or $1/z$, called the multiplicative inverse of z, such that $z \cdot z^{-1} = 1$.

Note that point two and three of Theorem 3.11 guarantee the existence of additive and multiplicative inverses. It does not state how to compute these inverses though. However, we have already seen how to compute these. To illustrate the computational method algorithmically, let us write down exactly how to compute $-z$ and $1/z$ in pseudo-code in the following example:

|||| Example 3.12

A possible algorithm that finds $-z$ for a given complex number z can be described as follows: first write z in rectangular form, which essentially means that it finds $a, b \in \mathbb{R}$ such that $z = a + bi$. Then $-z = -a - bi$. In pseudo-code:

Algorithm 4 for computing the "additive inverse of $z \in \mathbb{C}$".

Input: $z \in \mathbb{C}$
1: $a \leftarrow \mathrm{Re}(z)$
2: $b \leftarrow \mathrm{Im}(z)$
3: **return** $-a - bi$

To find $1/z$, we use Equation (3-3). Note that $1/z$ does not exist if $z = 0$. Therefore the algorithm first checks if $z = 0$.

Algorithm 5 for computing the "multiplicative inverse of $z \in \mathbb{C}$".

Input: $z \in \mathbb{C}$
1: **if** $z = 0$ **then**
2: **return** "0 has no multiplicative inverse!"
3: **else**
4: $c \leftarrow \mathrm{Re}(z)$,
5: $d \leftarrow \mathrm{Im}(z)$,
6: $N \leftarrow c^2 + d^2$,
7: **return** $\frac{c}{N} - \frac{d}{N}i$.

3.3 Modulus and argument

We have seen in Section 3.1 that a complex number z can be uniquely determined by its real part $\mathrm{Re}(z)$ and its imaginary part $\mathrm{Im}(z)$, since for any $z \in \mathbb{C}$ it holds that $z = \mathrm{Re}(z) + \mathrm{Im}(z)i$. We called the pair $(\mathrm{Re}(z), \mathrm{Im}(z))$ the rectangular coordinates of z. In this section we will introduce another way to describe a complex number. Given a complex number z, we can draw a triangle in the complex plane with vertices in the complex numbers 0, $\mathrm{Re}(z)$ and z (see Figure 3.4). The distance from z to 0 is called the *modulus* or *absolute value* of z and is denoted by $|z|$. The angle from the positive part of the real axis to the vector from 0 to z is called the *argument* of z and is denoted by $\arg(z)$.

We will always give the argument (and indeed any angle) in radians. Since the angle 2π denotes a full turn, one can always add an integer multiple of 2π to an angle. For example the angle $-\pi/4$ can also be given as $7\pi/4$, since $-\pi/4 + 2\pi = 7\pi/4$. For this reason one says that the argument of a complex number is determined only up to a multiple of 2π. A formula like "$\arg(z) = 5\pi/4$" should therefore be read as: "$5\pi/4$ is an argument of z". It is always possible to find an argument of a complex number z in the interval $]-\pi, \pi]$. This value is sometimes called the *principal value* of the argument and denoted by $\mathrm{Arg}(z)$.

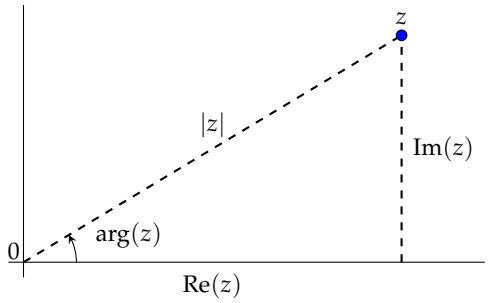


Figure 3.4: Modulus and argument of a complex number z.

From Figure 3.4 we can deduce that

$$\mathrm{Re}(z) = |z|\cos(\arg(z)) \quad \text{and} \quad \mathrm{Im}(z) = |z|\sin(\arg(z)). \tag{3-4}$$

Therefore, given $|z|$ and $\arg(z)$, we can compute z's rectangular coordinates. This implies that the pair $(|z|, \arg(z))$ completely determines the complex number z, since

$$z = |z|\,(\cos(\arg(z)) + \sin(\arg(z))i\,). \tag{3-5}$$

The pair $(|z|, \mathrm{Arg}(z))$ is called the *polar coordinates* of a complex number $z \in \mathbb{C}$. If a complex number z is written in the form $z = r\,(\cos(\alpha) + i\sin(\alpha)\,)$, with r a positive real number, it holds that $|z| = r$ and $\arg(z) = \alpha$. Moreover, if $\alpha \in]-\pi, \pi]$, then $\mathrm{Arg}(z) = \alpha$. Again from Figure 3.4 we can deduce that

$$|z| = \sqrt{\mathrm{Re}(z)^2 + \mathrm{Im}(z)^2} \quad \text{and} \quad \tan(\arg(z)) = \mathrm{Im}(z)/\mathrm{Re}(z), \text{ if } \mathrm{Re}(z) \neq 0. \tag{3-6}$$

This equation is the key to compute the polar coordinates of a number from its rectangular coordinates. More precisely, using the inverse tangent function arctan discussed in Subsection 2.3.2, we have the following:

Theorem 3.13

If a complex number z different from zero has polar coordinates (r, α), then

$$\mathrm{Re}(z) = r\cos(\alpha) \quad \text{and} \quad \mathrm{Im}(z) = r\sin(\alpha).$$

Conversely, if a complex number z different from zero has rectangular coordinates (a, b), then:

$$|z| = \sqrt{a^2 + b^2} \quad \text{and} \quad \mathrm{Arg}(z) = \begin{cases} \arctan(b/a) & \text{if } a > 0, \\ \pi/2 & \text{if } a = 0 \text{ and } b > 0, \\ \arctan(b/a) + \pi & \text{if } a < 0 \text{ and } b \geq 0, \\ -\pi/2 & \text{if } a = 0 \text{ and } b < 0, \\ \arctan(b/a) - \pi & \text{if } a < 0 \text{ and } b < 0. \end{cases}$$

Proof. Given the polar coordinates of z, we can use Equation (3-4) to compute its rectangular coordinates. Conversely, given the rectangular coordinates (a, b) of z, we get from Equation (3-6) that $|z| = \sqrt{a^2 + b^2}$. If $a = 0$, the number z lies on the imaginary axis. In this case we have that $\mathrm{Arg}(z) = \pi/2$ if $b > 0$ and $\mathrm{Arg}(z) = -\pi/2$ if $b < 0$. If $a \neq 0$, it holds according to Equation (3-6) that $\tan(\mathrm{Arg}(z)) = b/a$. Therefore it then holds that $\mathrm{Arg}(z) = \arctan(b/a) + n\pi$ for some integer $n \in \mathbb{Z}$. If z lies in the first or fourth quadrant, then $\mathrm{Arg}(z)$ lies in the interval $]-\pi/2, \pi/2[$. In this case we therefore get that $\mathrm{Arg}(z) = \arctan(b/a)$. If z lies in the second quadrant, its argument lies in the interval $]\pi/2, \pi]$. Therefore we then find that $\mathrm{Arg}(z) = \arctan(b/a) + \pi$. Similarly, if z lies in the third quadrant, we find that $\mathrm{Arg}(z) = \arctan(b/a) - \pi$. □

The modulus can be seen as a function $f : \mathbb{C} \to \mathbb{R}$, where $f(z) = |z|$. It plays a similar role for the complex numbers as the absolute value function from Example 2.19. In fact, if $z = a + 0i$ is a real number, it holds that $|z| = \sqrt{a^2 + 0^2}$ if we apply the modulus function. However, $\sqrt{a^2} = |a|$, where now $|a|$ denotes the absolute value of a real number. Hence the modulus, when applied to a real number a, gives exactly the same output as the absolute value applied to a. This explains why it makes sense to use exactly the notation $|a|$ both for the usual absolute value of a real number and for the modulus of a complex number. Indeed, $|z|$ is in fact often also called the absolute value of a complex number. Finally, observe that $|z|^2 = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2 = z \cdot \bar{z}$, the final equality following from equation (3-1).

The formula for the argument of a complex number $a + bi$ depends on in which quadrant of the complex plane the number lies (see Figure 3.5).

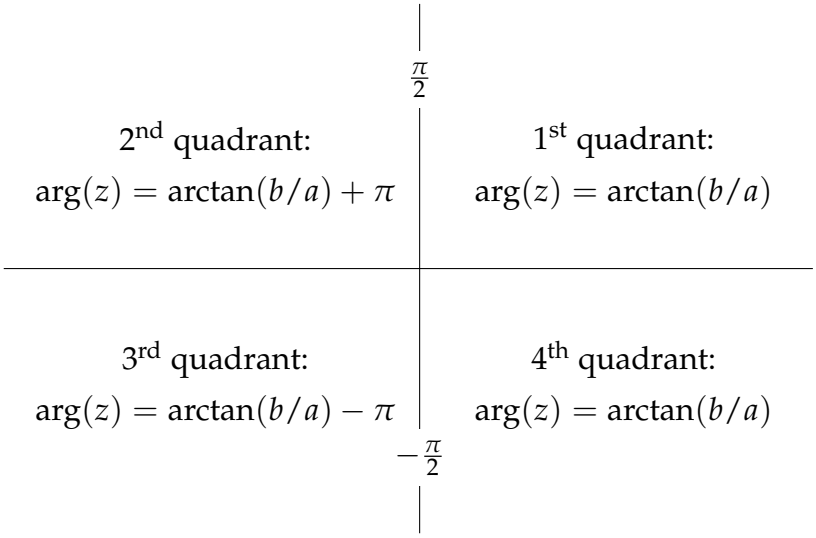


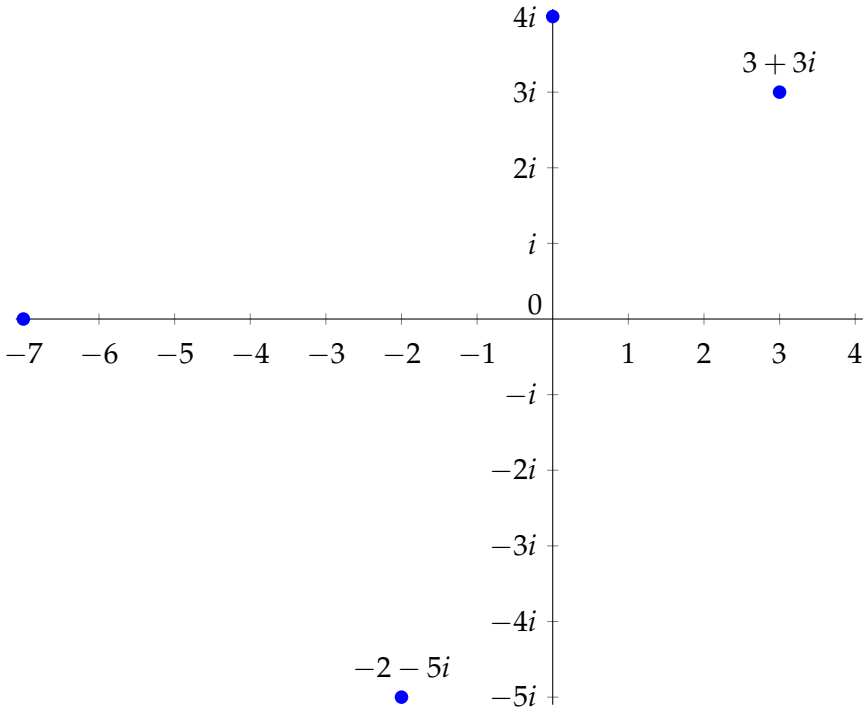
Figure 3.5: Formulas for the argument of $z = a + bi$.

|||| Example 3.14

Compute the polar coordinates of the following complex numbers:

1. $4i$
2. -7
3. $3 + 3i$
4. $-2 - 5i$

Answer: We can find the modulus and argument using Theorem 3.13. Figure 3.5 is useful when computing the argument. Therefore, we first plot the four given complex numbers in the complex plane.



1. $|4i| = |0 + 4i| = \sqrt{0^2 + 4^2} = 4$ and $\text{Arg}(4i) = \pi/2$. Therefore the polar coordinates of $4i$ are $(4, \pi/2)$.

2. $|-7| = \sqrt{(-7)^2 + 0^2} = 7$ and $\text{Arg}(-7) = \arctan(0/(-7)) + \pi = \pi$. Therefore the polar coordinates of -7 are $(7, \pi)$.

3. $|3 + 3i| = \sqrt{3^2 + 3^2} = 3\sqrt{2}$ and $\text{Arg}(3 + 3i) = \arctan(3/3) = \pi/4$. Therefore the polar coordinates of $3 + 3i$ are $(3\sqrt{2}, \pi/4)$.

4. $|-2 - 5i| = \sqrt{(-2)^2 + (-5)^2} = \sqrt{29}$

   and

   $\text{Arg}(-2 - 5i) = \arctan((-5)/(-2)) - \pi = \arctan(5/2) - \pi$. Therefore the polar coordinates of $-2 - 5i$ are $(\sqrt{29}, \arctan(5/2) - \pi)$.

|||| Example 3.15

The following polar coordinates are given. Compute the corresponding complex numbers and write those numbers in rectangular form.

1. $(2, \pi/3)$

2. $(10, \pi)$
3. $(4, -\pi/4)$
4. $(2\sqrt{3}, -2\pi/3)$
5. $(3, 2)$

Answer: We use Equation (3-5) to compute the complex numbers z corresponding to the given polar coordinates. Afterwards we express these complex numbers in rectangular form.

1. $z = 2 \cdot (\cos(\pi/3) + \sin(\pi/3)i) = 2 \cdot (1/2 + \sqrt{3}/2i) = 1 + \sqrt{3}i$.
2. $z = 10 \cdot (\cos(\pi) + \sin(\pi)i) = -10 + 0i = -10$.
3. $z = 4 \cdot (\cos(-\pi/4) + \sin(-\pi/4)i) = 4 \cdot (\sqrt{2}/2 - \sqrt{2}/2i) = 2\sqrt{2} - 2\sqrt{2}i$.
4. $z = 2\sqrt{3} \cdot (\cos(-2\pi/3) + \sin(-2\pi/3)i) = 2\sqrt{3} \cdot (-1/2 - \sqrt{3}/2i) = -\sqrt{3} - 3i$.
5. $z = 3 \cdot (\cos(2) + \sin(2)i) = 3\cos(2) + 3\sin(2)i$.

3.4 The complex exponential function

We have seen that many computations one can do with real numbers, like addition, subtraction, multiplication and division, also can be done with complex numbers. We will see in this section that also the exponential function $\exp : \mathbb{R} \to \mathbb{R}_{>0}$, where $\exp(t) = e^t$ can be defined for complex numbers as well. The resulting function is called the *complex exponential function*.

|||| Definition 3.16

Let $z \in \mathbb{C}$ be a complex number whose rectangular form is given by $z = a + bi$ for certain $a, b \in \mathbb{R}$. Then we define

$$e^z = e^a \cdot (\cos(b) + \sin(b)i).$$

The complex exponential function is usually again denoted by exp. This time the domain of the function is $\mathbb{C}$ though. More precisely, the complex exponential function

is the function $\exp : \mathbb{C} \to \mathbb{C}$. Note that if z is a real number, say $z = a + 0i$, then $e^z = e^a \cdot (\cos(0) + \sin(0)i) = e^a$. So the complex exponential function, when evaluated in a real number, gives exactly the same as the usual exponential function would have given. This is the reason why it makes sense to denote both the exponential function $\exp : \mathbb{R} \to \mathbb{R}_{>0}$ and the complex exponential function $\exp : \mathbb{C} \to \mathbb{C}$ with the same symbol exp.

|||| Example 3.17

Write the following expressions in rectangular form:

1. e^2
2. e^{1+i}
3. $e^{\pi i}$
4. $e^{\ln(2)+i\pi/4}$ (whenever we write ln, we mean the logarithm with base e)
5. $e^{2\pi i}$

Answer: We use Definition 3.16 and simplify till we find the desired rectangular form.

1. Since e^2 is a real number, it is already in rectangular form. If we use Definition 3.16 anyway, we find $e^2 = e^{2+0i} = e^2 \cdot (\cos(0) + \sin(0)i) = e^2 \cdot (1 + 0i) = e^2$, which again shows that e^2 already was in rectangular form. It is also fine to write $e^2 = e^2 + 0i$ and then to return $e^2 + 0i$ as answer.
2. $e^{1+i} = e^1 \cdot (\cos(1) + \sin(1)i) = e\cos(1) + e\sin(1)i$.
3. $e^{\pi i} = e^{0+\pi i} = e^0 \cdot (\cos(\pi) + \sin(\pi)i) = 1 \cdot (-1 + 0i) = -1$.
4. $e^{\ln(2)+i\pi/4} = e^{\ln(2)} \cdot (\cos(\pi/4) + \sin(\pi/4)i) = 2(\sqrt{2}/2 + \sqrt{2}/2i) = \sqrt{2} + \sqrt{2}i$.
5. $e^{2\pi i} = \cos(2\pi) + \sin(2\pi)i = 1 + 0i = 1$. Note that also $e^0 = 1$. This shows that the complex exponential function is not injective.

Directly from Definition 3.16, we see that for any $z \in \mathbb{C}$:

$$\mathrm{Re}(e^z) = e^{\mathrm{Re}(z)}\cos(\mathrm{Im}(z)) \quad \text{and} \quad \mathrm{Im}(e^z) = e^{\mathrm{Re}(z)}\sin(\mathrm{Im}(z)).$$

The complex exponential function has many properties in common with the usual real exponential function. To show those, we will use the following lemma.

Lemma 3.18

Let $\alpha_1, \alpha_2 \in \mathbb{R}$. We have

$$(\cos(\alpha_1) + \sin(\alpha_1)i) \cdot (\cos(\alpha_2) + \sin(\alpha_2)i) = \cos(\alpha_1 + \alpha_2) + \sin(\alpha_1 + \alpha_2)i.$$

Proof. By multiplying out the parentheses, we can compute the real and imaginary part of the product $(\cos(\alpha_1) + \sin(\alpha_1)i) \cdot (\cos(\alpha_2) + \sin(\alpha_2)i)$. It turns out that the real part is given by $\cos(\alpha_1)\cos(\alpha_2) - \sin(\alpha_1)\sin(\alpha_2)$ and the imaginary part by $\cos(\alpha_1)\sin(\alpha_2) + \sin(\alpha_1)\cos(\alpha_2)$. Using the additions formulas for the cosine and sine functions the lemma follows. □

Theorem 3.19

Let z, z_1 and z_2 be complex numbers and n an integer. Then it holds that

$$e^z \neq 0$$

$$1/e^z = e^{-z}$$

$$e^{z_1}e^{z_2} = e^{z_1+z_2}$$

$$e^{z_1}/e^{z_2} = e^{z_1-z_2}$$

$$(e^z)^n = e^{nz}$$

Proof. We will show the third item: $e^{z_1}e^{z_2} = e^{z_1+z_2}$. First we write z_1 and z_2 in rectangu-

lar form: $z_1 = a_1 + b_1 i$ and $z_2 = a_2 + b_2 i$. Then we find that

$$\begin{aligned} e^{z_1} \cdot e^{z_2} &= e^{a_1} \cdot (\cos(b_1) + \sin(b_1)i) \cdot e^{a_2} \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1} \cdot e^{a_2} \cdot (\cos(b_1) + \sin(b_1)i) \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1+a_2} \cdot (\cos(b_1) + \sin(b_1)i) \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1+a_2} \cdot (\cos(b_1 + b_2) + \sin(b_1 + b_2)i) \text{ (using Lemma 3.18)} \\ &= e^{a_1+a_2+(b_1+b_2)i} = e^{z_1+z_2}. \end{aligned}$$

□

3.5 Euler's formula

The complex exponential function gives a connection between trigonometry and complex numbers. We will explore this connection in this section.

Let t be a real number. The formula

$$e^{it} = \cos(t) + i\sin(t) \tag{3-7}$$

is known as *Euler's formula* and is a consequence of Definition 3.16. It implies that

$$e^{-it} = \cos(-t) + i\sin(-t) = \cos(t) - i\sin(t). \tag{3-8}$$

Equations (3-7) and (3-8) can be seen as equations in the unknowns $\cos(t)$ and $\sin(t)$. Solving for $\cos(t)$ and $\sin(t)$ gives:

$$\cos(t) = \frac{e^{it} + e^{-it}}{2} \text{ and } \sin(t) = \frac{e^{it} - e^{-it}}{2i}. \tag{3-9}$$

Equation (3-9) can be used to rewrite products of cos- and sin-functions to a sum of cos- and sin-functions (that is to say, as a sum of purely harmonic functions). This kind of computations are standard in frequency analysis, where one tries to write arbitrary functions as a sum of purely harmonic functions. It can also be useful to compute integrals of trigonometric expressions as we can see in the following example.

|||| Example 3.20

Compute $\int \sin(3t)\cos(t)dt$.

Answer: First we use Euler's formulas to rewrite the expression $\sin(3t)\cos(t)$:

$$\sin(3t)\cos(t) = \frac{e^{i3t}-e^{-i3t}}{2i}\cdot\frac{e^{it}+e^{-it}}{2} = \frac{(e^{i3t}-e^{-i3t})(e^{it}+e^{-it})}{4i}$$

$$= \frac{e^{i4t}+e^{i2t}-e^{-i2t}-e^{-i4t}}{4i} = \frac{1}{2}\left(\frac{e^{i4t}-e^{-i4t}}{2i}+\frac{e^{i2t}-e^{-i2t}}{2i}\right)$$

$$= \frac{\sin(4t)}{2}+\frac{\sin(2t)}{2}.$$

Now we get

$$\int \sin(3t)\cos(t)dt = \int \frac{\sin(4t)}{2}+\frac{\sin(2t)}{2}dt = -\frac{\cos(4t)}{8}-\frac{\cos(2t)}{4}+c, \quad c\in\mathbb{R}.$$

In Figure 3.6 the identity $\sin(3t)\cos(t) = \frac{\sin(4t)}{2}+\frac{\sin(2t)}{2}$ from the previous example is illustrated.

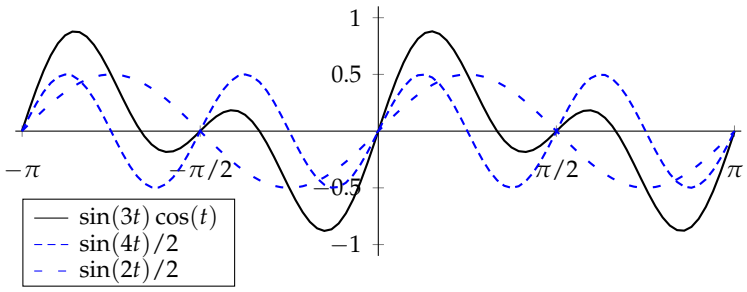


Figure 3.6: It holds that $\sin(3t)\cos(t) = \frac{\sin(4t)}{2}+\frac{\sin(2t)}{2}$.

Another application of Euler's formula is given in the following theorem.

Theorem 3.21

Let $n \in \mathbb{N}$ be a natural number. Then the following formulas hold:

$$\cos(nt) = \mathrm{Re}((\cos(t) + \sin(t)i)^n)$$

and

$$\sin(nt) = \mathrm{Im}((\cos(t) + \sin(t)i)^n)$$

Proof. The key is the following equation:

$$\cos(nt) + \sin(nt)i = e^{int} = (e^{it})^n = (\cos(t) + \sin(t)i)^n.$$

The theorem follows by taking real and imaginary parts on both side of this equality. □

The expressions in this theorem are known as *DeMoivre's formula*. Let us consider some examples.

Example 3.22

Express $\cos(2t)$ and $\sin(2t)$ in $\cos(t)$ and $\sin(t)$.

Answer: According to DeMoivre's formula for $n = 2$, we have $\cos(2t) = \mathrm{Re}((\cos(t) + \sin(t)i)^2)$ and $\sin(2t) = \mathrm{Im}((\cos(t) + \sin(t)i)^2)$. Since

$$\begin{aligned}(\cos(t) + \sin(t)i)^2 &= \cos^2(t) + 2\cos(t)\sin(t)i + \sin^2(t)i^2 \\ &= \cos^2(t) + 2\cos(t)\sin(t)i - \sin^2(t) \\ &= \cos^2(t) - \sin^2(t) + 2\cos(t)\sin(t)i,\end{aligned}$$

we find that

$$\cos(2t) = \cos(t)^2 - \sin(t)^2$$

and

$$\sin(2t) = 2\cos(t)\sin(t).$$

|||| Example 3.23

Express $\cos(3t)$ and $\sin(3t)$ in $\cos(t)$ and $\sin(t)$.

Answer: According to DeMoivre's formula for $n = 3$, we have $\cos(3t) = \mathrm{Re}((\cos(t) + i\sin(t))^3)$ and $\sin(3t) = \mathrm{Im}((\cos(t) + i\sin(t))^3)$. After some computations we find that $(\cos(t) + i\sin(t))^3 = (\cos(t)^3 - 3\cos(t)\sin(t)^2) + i(3\cos(t)^2\sin(t) - \sin(t)^3)$. Apparently the following holds:

$$\cos(3t) = \cos(t)^3 - 3\cos(t)\sin(t)^2$$

and

$$\sin(3t) = 3\cos(t)^2\sin(t) - \sin(t)^3.$$

3.6 The polar form of a complex number

Let r be a positive, real number and α a real number. Then from Definition 3.16, we see that $r \cdot e^{\alpha i} = r \cdot (\cos(\alpha) + \sin(\alpha)i)$. As we have seen in and after Equation (3-5), the number $r \cdot e^{\alpha i}$ then has modulus r and an argument equal to α (see Figure 3.7). Also we can rewrite Equation (3-5) as $z = |z|e^{i\arg(z)}$. This way to write a complex number has a special name:

|||| Definition 3.24

Let $z \in \mathbb{C} \setminus \{0\}$ be a non-zero complex number. Then the righthand side of the equation

$$z = |z| \cdot e^{i\arg(z)}$$

is called the *polar form* of z.

If $z \neq 0$, we can from the polar coordinates (r, α) of z directly write z in polar form, namely $z = re^{i\alpha}$. Conversely, given an expression of the form $z = re^{i\alpha}$, with $r > 0$ a positive real number and $\alpha \in]-\pi, \pi]$ a real number, we can read off that the polar coordinates of z are given by (r, α). See Figure 3.7 for an illustration.

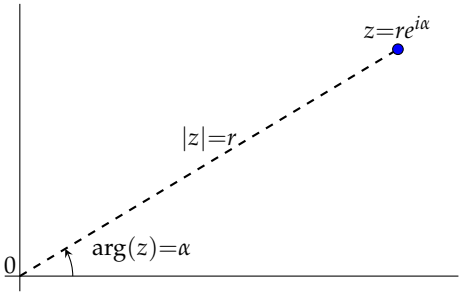


Figure 3.7: Polar form of a complex number z.

|||| Example 3.25

Write the following complex numbers in polar form:

1. $-1+i$
2. $2+5i$
3. e^{7+3i}
4. $e^{7+3i}/(-1+i)$

Answer: In principle, one can for each of the given numbers calculate its modulus and its argument. Once these have been calculated, one can write the number in polar form.

1. $|-1+i| = \sqrt{1+1} = \sqrt{2}$ and $\arg(-1+i) = \arctan(1/-1)+\pi = 3\pi/4$. In polar form the number is therefore given by $\sqrt{2}e^{i3\pi/4}$.
2. $|2+5i| = \sqrt{4+25} = \sqrt{29}$ and $\arg(2+5i) = \arctan(5/2)$. We therefore find that $2+5i$ has the following polar form: $\sqrt{29}e^{i\arctan(5/2)}$.
3. $e^{7+3i} = e^7 e^{3i}$. The righthand side of this equation is already the polar form of the number, since it is of the form $re^{i\alpha}$ (with $r > 0$ and $\alpha \in \mathbb{R}$). We can read off that the modulus of the number e^{7+3i} equals e^7, while its argument equals 3.
4. We have seen in the first part of this example that $-1+i = \sqrt{2}e^{i3\pi/4}$. Then we get that:
$$\frac{e^{7+3i}}{-1+i} = \frac{e^7 e^{3i}}{\sqrt{2}e^{i3\pi/4}} = \frac{e^7}{\sqrt{2}}\frac{e^{3i}}{e^{i3\pi/4}} = \frac{e^7}{\sqrt{2}}e^{(3-3\pi/4)i}.$$
The last expression is the desired polar form. We can read off that the number $e^{7+3i}/(-1+i)$ has modulus $e^7/\sqrt{2}$ and argument $3-3\pi/4$.

In the previous example, we saw that the modulus of the number e^{7+3i} equalled e^7, while its argument was given by 3. In general it holds that

$$|e^z| = e^{\mathrm{Re}(z)} \quad \text{and} \quad \arg(e^z) = \mathrm{Im}(z). \tag{3-10}$$

In the last item of Example 3.17, we have seen that the complex exponential function is not injective, since the equation $e^z = 1$ has several solutions, for example 0 and $2\pi i$. Using what we have learned so far, let us investigate more generally how to solve this type of equation:

|||| Lemma 3.26

Let $w \in \mathbb{C}$ be a complex number. If $w = 0$, then the equation $e^z = w$ has no solutions. If $w \neq 0$, then the solutions to equation $e^z = w$ are precisely those $z \in \mathbb{C}$ of the form $z = \ln(|w|) + \arg(w)i$, where $\arg(w)$ can be any argument of w.

Proof. Equation (3-10) implies that $|e^z|$ cannot be zero, since $e^{\mathrm{Re}(z)} > 0$ for all $z \in \mathbb{C}$. Hence the equation $e^z = 0$ has no solutions. Now assume that $w \neq 0$. If $e^z = w$, then Equation (3-10) implies that $|w| = |e^z| = e^{\mathrm{Re}(z)}$ and therefore that $\mathrm{Re}(z) = \ln(|w|)$. Similarly, using the second part of Equation (3-10), $e^z = w$ implies that $\arg(w) = \mathrm{Im}(z)$. Note though that there are infinitely many possible values for $\arg(w)$, since we can always modify it by adding an integer multiple of 2π to it. So far, we have showed that if $w \neq 0$, then any solution of the equation $e^z = w$ has to be of the form $z = \ln(|w|) + \arg(w)i$. Conversely, given any z satisfying $z = \ln(|w|) + \arg(w)i$, where $\arg(w)$ is any argument of w, then $e^z = e^{\ln(|w|)+\arg(w)i} = e^{\ln(|w|)} \cdot e^{i\arg(w)} = |w| \cdot e^{i\arg(w)} = w$, where the last equality follows since $|w|e^{i\arg(w)}$ is simply the polar form of w. □

A direct consequence of this lemma is that the image of the complex exponential function $\exp : \mathbb{C} \to \mathbb{C}$ with $z \mapsto e^z$, satisfies $\exp(\mathbb{C}) = \mathbb{C} \setminus \{0\}$. Indeed, the equation $e^z = 0$ has no solutions, implying that 0 is not in the image, while for any nonzero complex number w, the lemma explains how to find complex numbers z that are mapped to w by the complex exponential function.

We can now revisit polar coordinates and use the properties of the complex exponential function as given in Theorem 3.19 to prove the following theorem.

Theorem 3.27

Let $z_1, z_2 \in \mathbb{C} \setminus \{0\}$ be two complex numbers both different from zero. Then the following holds:

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|$$

and

$$\arg(z_1 \cdot z_2) = \arg(z_1) + \arg(z_2).$$

We also have

$$|z_1/z_2| = |z_1|/|z_2|$$

and

$$\arg(z_1/z_2) = \arg(z_1) - \arg(z_2).$$

Finally, let $n \in \mathbb{Z}$ be an integer and $z \in \mathbb{C} \setminus \{0\}$ a non-zero complex number. Then

$$|z^n| = |z|^n$$

and

$$\arg(z^n) = n \arg(z).$$

Proof. We only show the first two parts of the theorem. Let us write $r_1 = |z_1|$, $r_2 = |z_2|$, $\alpha_1 = \arg(z_1)$ and $\alpha_2 = \arg(z_2)$. According to Equation (3-5) we have

$$\begin{aligned} z_1 \cdot z_2 &= r_1 \cdot e^{\alpha_1 i} \cdot r_2 \cdot e^{\alpha_2 i} \\ &= r_1 \cdot r_2 \cdot e^{\alpha_1 i} \cdot e^{\alpha_2 i} \\ &= r_1 \cdot r_2 \cdot e^{\alpha_1 i + \alpha_2 i} \\ &= r_1 \cdot r_2 \cdot e^{(\alpha_1 + \alpha_2) i} \end{aligned}$$

We used the third item of Theorem 3.19 in the third equality. We can now conclude that

$$|z_1 \cdot z_2| = r_1 \cdot r_2 = |z_1| \cdot |z_2| \quad \text{and} \quad \arg(z_1 \cdot z_2) = \alpha_1 + \alpha_2 = \arg(z_1) + \arg(z_2).$$

□

Theorem 3.27 gives a geometric way to describe the multiplication of two complex numbers: the length of a product is the product of the lengths and the argument of a product is the sum of the arguments (see Figure 3.8).

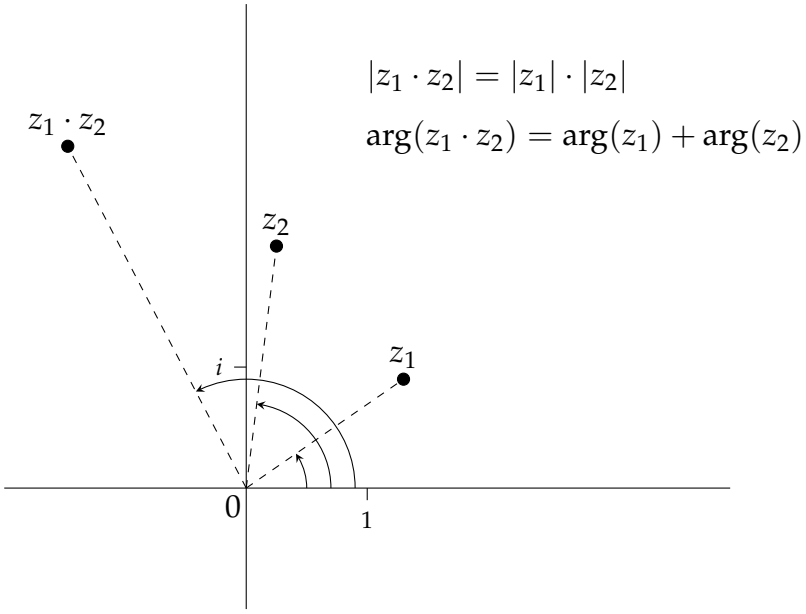


Figure 3.8: Graphic illustration of Theorem 3.27.

The polar form of a complex number can be very useful for the computation of an integer power of a complex number. Let us look at an example.

|||| Example 3.28

Write the following complex numbers in rectangular form. Hint: use polar forms.

1. $(1+i)^{13}$.
2. $(-1-\sqrt{3}i)^{15}$.

Answer:

1. The number $1+i$ has argument $\pi/4$ and modulus $\sqrt{2}$. Hence $1+i = \sqrt{2}\cdot e^{i\pi/4}$. Hence

$$\begin{aligned}(1+i)^{13} &= \left(\sqrt{2}\cdot e^{i\pi/4}\right)^{13}\\ &= \sqrt{2}^{13}\cdot e^{i13\pi/4}\\ &= \sqrt{2}^{13}\cdot(\cos(13\pi/4)+\sin(13\pi/4)i)\\ &= \sqrt{2}^{13}\cdot(\cos(-3\pi/4)+\sin(-3\pi/4)i)\\ &= 64\sqrt{2}\cdot(\cos(-3\pi/4)+\sin(-3\pi/4)i)\\ &= 64\sqrt{2}\cdot\left(-\sqrt{2}/2-i\sqrt{2}/2\right)\\ &= -64-64i.\end{aligned}$$

2. First we calculate modulus and argument $-1-\sqrt{3}i$. According to Theorem 3.13 it holds that

$$\arg(-1-\sqrt{3}i) = \arctan((-\sqrt{3})/(-1)) - \pi = -2\pi/3$$

and

$$|-1-\sqrt{3}i| = \sqrt{(-1)^2+(-\sqrt{3})^2} = 2.$$

Hence $-1-\sqrt{3}i = 2\cdot e^{-i2\pi/3}$. Therefore

$$\begin{aligned}(-1-\sqrt{3}i)^{15} &= \left(2\cdot e^{-i2\pi/3}\right)^{15}\\ &= 2^{15}\cdot e^{-i30\pi/3}\\ &= 2^{15}\cdot(\cos(-30\pi/3)+\sin(-30\pi/3)i)\\ &= 2^{15}\cdot(\cos(-10\pi)+\sin(-10\pi)i)\\ &= 2^{15}\cdot(\cos(0)+\sin(0)i)\\ &= 2^{15}\cdot(1+0i)\\ &= 2^{15}.\end{aligned}$$

Note 4

Polynomials

4.1 Definition of polynomials

In this chapter we will investigate a certain type of expressions called polynomials. Polynomials will come up again later, when we discuss differential equations, examples of vector spaces, and eigenvalues of a matrix, but that is for later. For now, we start by defining what a polynomial is.

Definition 4.1

A *polynomial* $p(Z)$ in a variable Z is an expression of the form:

$$p(Z) = a_0 Z^0 + a_1 Z^1 + a_2 Z^2 + \cdots + a_n Z^n, \text{ with } n \in \mathbb{Z}_{\geq 0} \text{ a non-negative integer.}$$

Here the symbols $a_0, a_1, a_2, \ldots, a_n \in \mathbb{C}$ denote complex numbers, which are called the *coefficients* of $p(Z)$. The expressions $a_0 Z^0, a_1 Z^1, \ldots, a_n Z^n$ are called the *terms* of the polynomial $p(Z)$. The largest i for which $a_i \neq 0$ is called the *degree* of $p(Z)$ and is denoted by $\deg(p(Z))$. The corresponding coefficient is called the *leading coefficient*. Finally, the set of all polynomials in Z with complex coefficients is denoted by $\mathbb{C}[Z]$.

It is common not to write Z^0 and to write Z instead of Z^1. Then a polynomial is simply written as $p(Z) = a_0 + a_1 Z + a_2 Z^2 + \cdots + a_n Z^n$. A polynomial of degree zero can then just be interpreted as a nonzero constant a_0, while a polynomial of degree one has the form $a_0 + a_1 Z$. The polynomial all of whose coefficients are zero is called the *zero*

polynomial and denoted by 0. It is customary to define the degree of the zero polynomial to be $-\infty$, minus infinity.

By definition, the coefficients completely determine a polynomial. In other words: two polynomials $p_1(Z) = a_0 + a_1 Z \cdots + a_n Z^n$ of degree n and $p_2(Z) = b_0 + b_1 Z \cdots + b_m Z^m$ of degree m are equal if and only if $n = m$ and $a_i = b_i$ for all i. The order of the terms is not important. For example, the polynomials $Z^2 + 2Z + 3$, $Z^2 + 3 + 2Z$ and $3 + 2Z + Z^2$ are all the same. The notation $\mathbb{C}[Z]$ for the set of all polynomials with coefficients in $\mathbb{C}$ is standard, but the symbol used to indicate the variable, in our case Z, varies from book to book. We have chosen Z, since we have been using z for complex numbers. Other sets of polynomials can be obtained by replacing $\mathbb{C}$ by something else. For example, we will frequently use $\mathbb{R}[Z]$, which denotes the set of all polynomials with coefficients in $\mathbb{R}$. Note that $\mathbb{R}[Z] \subseteq \mathbb{C}[Z]$, since $\mathbb{R} \subseteq \mathbb{C}$.

|||| Example 4.2

Indicate which of the following expressions is an element of $\mathbb{C}[Z]$. If the expression is a polynomial, give its degree and leading coefficient.

1. $1 + Z^2$
2. $Z^{-1} + 1 + Z^3$
3. i
4. $\sin(Z) + Z^{12}$
5. $1 + 2Z + 5Z^{10} + 0Z^{11}$
6. $1 + Z + Z^{2.5}$
7. $(1 + Z)^2$

Answer:

1. $1 + Z^2$ is a polynomial in Z. If we want to write it in the form $a_0 + a_1 Z + a_2 Z^2 + \cdots + a_n Z^n$ as in Definition 4.1, we can write it as $1 + 0Z + 1Z^2$. Hence $n = 2$, $a_0 = a_2 = 1$ and $a_1 = 0$. Because $a_2 \neq 0$, the polynomial is of degree 2, while its leading coefficient is a_2, which is equal to 1.
2. $Z^{-1} + 1 + Z^3$ is not a polynomial in Z because of the term Z^{-1}. The exponents of Z of the terms in a polynomial may not be negative.

3. The complex number i can be interpreted as a polynomial in $\mathbb{C}[Z]$. One chooses $n = 0$ and $a_0 = i$ in Definition 4.1. The polynomial i has therefore degree 0 and leading coefficient i.
4. $\sin(Z) + Z^{12}$ is not a polynomial because of the term $\sin(Z)$.
5. $1 + 2Z + 5Z^{10} + 0Z^{11}$ is a polynomial in $\mathbb{C}[Z]$. The term of degree eleven can be left out though, since the coefficient of Z^{11} is 0. The highest power of Z with a coefficient different from zero is therefore 10. This means that $\deg(1 + 2Z + 5Z^{10} + 0Z^{11}) = 10$, while its leading coefficient is 5.
6. $1 + Z + Z^{2.5}$ is not a polynomial, because of the term $Z^{2.5}$. The exponents of Z must be natural numbers.
7. $(2 + Z)^2$ is a polynomial in $\mathbb{C}[Z]$, though it is not written in the form as in Definition 4.1. However, it can be rewritten in this form, since $(2 + Z)^2 = 4 + 4Z + Z^2 = 4 + 4Z + 1Z^2$. We have that $\deg((2 + Z)^2) = 2$. The leading coefficient of $(1 + Z)^2$ is 1.

Given a polynomial $p(Z) \in \mathbb{C}[Z]$, one can evaluate the polynomial in any complex number $z \in \mathbb{C}$. More precisely, if $p(Z) = a_0 + a_1 Z + \cdots + a_n Z^n \in \mathbb{C}[Z]$ and $z \in \mathbb{C}$, then we can define $p(z) = a_0 + a_1 \cdot z + \cdots + a_n \cdot z^n \in \mathbb{C}$. In this way, any polynomial $p(Z) \in \mathbb{C}[Z]$ gives rise to a function $p : \mathbb{C} \to \mathbb{C}$, defined by $z \mapsto p(z)$. A function $f : \mathbb{C} \to \mathbb{C}$ is called a *polynomial function*, if there exists a polynomial $p(Z) \in \mathbb{C}[Z]$ such that for all $z \in \mathbb{C}$ it holds that $f(z) = p(z)$. Similarly, a function $f : \mathbb{R} \to \mathbb{R}$ is called a polynomial function, if there exists a polynomial $p(Z) \in \mathbb{R}[Z]$ such that for all $x \in \mathbb{R}$ it holds that $f(x) = p(x)$.

Two polynomials $p_1(Z) = a_0 + a_1 Z \cdots + a_n Z^n$ and $p_2(Z) = b_0 + b_1 Z \cdots + b_m Z^m$ can be multiplied by adding all the terms $a_i b_j Z^{i+j}$, where $0 \leq i \leq n$ and $0 \leq j \leq m$. This simply means that in order to compute $p_1(Z) \cdot p_2(Z)$, one simply multiplies each term in $p_1(Z)$ with each term in $p_2(Z)$ and then adds up the resulting terms. Let us look at some examples.

|||| Example 4.3

Write the following polynomials in the form as in Definition 4.1.

1. $(Z + 5) \cdot (Z + 6)$.
2. $(3Z + 2) \cdot (3Z - 2)$.
3. $(Z - 1) \cdot (Z^2 + Z + 1)$.

Answer:

1. $(Z+5)\cdot(Z+6) = Z\cdot(Z+6)+5\cdot(Z+6) = Z^2+6Z+5Z+30 = Z^2+11Z+30$.
2. $(3Z+2)\cdot(3Z-2) = (3z)^2-6Z+6Z-2^2 = 9Z^2-4$.
3. In this example, the only difference from the previous two is that there will be more terms when multiplying, but otherwise there is no difference:
$$\begin{aligned}(Z-1)\cdot(Z^2+Z+1) &= Z\cdot(Z^2+Z+1)-(Z^2+Z+1)\\ &= Z^3+Z^2+Z-Z^2-Z-1\\ &= Z^3-1.\end{aligned}$$

Note that if a polynomial is a product of two other polynomials, say $p(Z) = p_1(Z)\cdot p_2(Z)$, then $\deg p(Z) = \deg p_1(Z) + \deg p_2(Z)$. In other words:

$$p(Z) = p_1(Z)\cdot p_2(Z) \quad \Rightarrow \quad \deg p(Z) = \deg p_1(Z) + \deg p_2(Z). \tag{4-1}$$

If $p(Z) \in \mathbb{C}[Z]$ is a polynomial, then the equation $p(z) = 0$ is called a *polynomial equation*. Solutions to a polynomial equation have a special name:

|||| Definition 4.4

Let $p(Z) \in \mathbb{C}[Z]$ be a polynomial. A complex number $\lambda \in \mathbb{C}$ is called a *root* of $p(Z)$ precisely if $p(\lambda) = 0$.

Note that by definition, a complex number is a root of a polynomial $p(Z)$ if and only if it is a solution to the polynomial equation $p(z) = 0$.

4.2 Polynomials in $\mathbb{R}[Z]$ of degree two

To see why complex numbers were introduced in the first place, we will explain in this section how to find the roots of a polynomial $p(Z) \in \mathbb{R}[Z]$ of degree two. Note that we are assuming that $p(Z) \in \mathbb{R}[Z]$ so that the polynomial $p(Z)$ has real coefficients. Such a polynomial $p(Z)$ can therefore be written in the form

$$p(Z) = aZ^2 + bZ + c,$$

where $a, b, c \in \mathbb{R}$ and $a \neq 0$. To find its roots, we need to solve the polynomial equation $az^2 + bz + c = 0$. Now the following holds:

$$\begin{aligned} az^2 + bz + c = 0 &\Leftrightarrow 4a^2z^2 + 4abz + 4ac = 0 \\ &\Leftrightarrow (2az)^2 + 2(2az)b + b^2 = b^2 - 4ac \\ &\Leftrightarrow (2az + b)^2 = b^2 - 4ac. \end{aligned} \tag{4-2}$$

The expression $b^2 - 4ac$ is called the discriminant of the polynomial $aZ^2 + bZ + c$. We will denote it by D. From Equation (4-2) it follows that in order to compute the roots of the polynomial $aZ^2 + bZ + c$, we need to take the square root of its discriminant D. If $D \geq 0$, one can use the usual square root, but now we will define the square root of any real number:

|||| Definition 4.5

Let D be a real number. Then we define

$$\sqrt{D} = \begin{cases} \sqrt{D} & \text{if } D \geq 0, \\ i\sqrt{|D|} & \text{if } D < 0. \end{cases}$$

If $D \geq 0$, then $\sqrt{D}$ is exactly what we are used to and it holds that $\sqrt{D}^2 = D$. If $D < 0$, it holds that $\sqrt{D}^2 = (i\sqrt{|D|})^2 = i^2\sqrt{|D|}^2 = (-1)|D| = D$. Therefore, for all real numbers D it holds that $\sqrt{D}^2 = D$. This is exactly the property that we would like the square root symbol to have. Moreover, all solutions to the equation $z^2 = D$ can now be given: they are $z = \sqrt{D}$ and $z = -\sqrt{D}$. Later, in Theorem 4.13, we will even be able to describe all the solutions to equations of the form $z^n = w$ for any $n \in \mathbb{N}$ and $w \in \mathbb{C}$. We now return to the computation of the roots of the polynomial $p(z) = az^2 + bz + c$. Using the extended square root and Equation (4-2) we find that

$$\begin{aligned} az^2 + bz + c = 0 &\Leftrightarrow (2az + b)^2 = b^2 - 4ac \\ &\Leftrightarrow (2az + b) = \sqrt{b^2 - 4ac} \quad \vee \quad (2az + b) = -\sqrt{b^2 - 4ac} \\ &\Leftrightarrow z = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \vee \quad z = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \end{aligned} \tag{4-3}$$

We get the usual formula to solve an equation of degree two, but the square root of the discriminant is now also defined if the discriminant is negative. In fact we now have shown the following theorem.

Theorem 4.6

The polynomial $p(Z) = aZ^2 + bZ + c \in \mathbb{R}[Z]$ with $a \neq 0$, has precisely the following roots in $\mathbb{C}$:

$$\frac{-b+\sqrt{D}}{2a} \text{ and } \frac{-b-\sqrt{D}}{2a}, \text{ where } D = b^2 - 4ac.$$

To be more precise, the polynomial has

1. two real roots $z = \dfrac{-b \pm \sqrt{D}}{2a}$ if $D > 0$,
2. one real root $z = \dfrac{-b}{2a}$ if $D = 0$,
3. two non-real roots $z = \dfrac{-b \pm i\sqrt{|D|}}{2a}$ if $D < 0$.

The description of the roots in Theorem 4.6 is very algorithmic in nature. In fact, let us write some pseudo-code for an algorithm:

Algorithm 6 for computing the roots of $p(Z) \in \mathbb{R}[Z]$ of degree two.

Input: $p(Z) \in \mathbb{R}[Z]$, with $\deg(p(Z)) = 2$
$a \leftarrow$ coefficient of Z^2 in $p(Z)$
$b \leftarrow$ coefficient of Z^1 in $p(Z)$
$c \leftarrow$ coefficient of Z^0 in $p(Z)$
$D \leftarrow b^2 - 4ac$
if $D \geq 0$ **then**
    return $\dfrac{-b+\sqrt{D}}{2a}$ and $\dfrac{-b-\sqrt{D}}{2a}$
else
    return $\dfrac{-b+i\sqrt{|D|}}{2a}$ and $\dfrac{-b-i\sqrt{|D|}}{2a}$

In Figure 4.1, we have drawn the graphs of some second degree polynomials. Real roots of a second degree polynomial correspond to intersection points of the x-axis and its graph. If there are no intersection points, the polynomial does not have real roots, but complex roots. If $D = b^2 - 4ac = 0$, the polynomial equation $az^2 + bz + c = 0$ has one solution and we say in this case that the polynomial has a *double root*, or a root of multiplicity two. If $D \neq 0$, one says that the roots have multiplicity one. We see that any polynomial of degree two has two roots if the roots are counted with their multiplicities.

We will return to roots and multiplicities in more detail in Section 4.6. If we consider the graph of a polynomial function $f : \mathbb{R} \to \mathbb{R}$ coming from a degree two polynomial in $\mathbb{R}[Z]$, then this graph intersects the horizontal axis twice if $D > 0$, once if $D = 0$ and not at all if $D < 0$. See Figure 4.1 for an illustration.

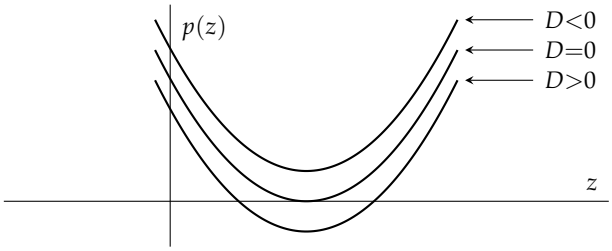


Figure 4.1: A degree two polynomial $p(Z) \in \mathbb{R}[Z]$ has two real roots if $D > 0$, a double root if $D = 0$, and two complex, two non-real roots if $D < 0$.

|||| Example 4.7

Compute all complex roots of the polynomial $2Z^2 - 4Z + 10 = 0$.

Answer: The discriminant of the polynomial $2Z^2 - 4Z + 10$ equals

$$D = (-4)^2 - 4 \cdot 2 \cdot 10 = -64.$$

According to Definition 4.5 we then find that

$$\sqrt{D} = \sqrt{-64} = i\sqrt{64} = 8i.$$

Therefore the polynomial equation $2z^2 - 4z + 10 = 0$ has two non-real roots, namely

$$z = \frac{-(-4) + 8i}{2 \cdot 2} = 1 + 2i \quad \vee \quad z = \frac{-(-4) - 8i}{2 \cdot 2} = 1 - 2i.$$

Although Theorem 4.6 guarantees that $1 + 2i$ and $1 - 2i$ are the roots of the polynomial $2Z^2 - 4Z + 10$, let us check that $1 + 2i$ is a root by hand:

$$\begin{aligned} 2 \cdot (1 + 2i)^2 - 4 \cdot (1 + 2i) + 10 &= 2 \cdot (1^2 + 4i + (2i)^2) - 4 \cdot (1 + 2i) + 10 \\ &= 2 \cdot (1 - 4 + 4i) - 4 \cdot (1 + 2i) + 10 \\ &= 2 \cdot (-3 + 4i) - 4 \cdot (1 + 2i) + 10 \\ &= (-6 + 8i) - (4 + 8i) + 10 \\ &= 0. \end{aligned}$$

Hence indeed, just as the theory predicts, $1 + 2i$ is a root of $2Z^2 - 4Z + 10$.

4.3 Polynomials with real coefficients

In the previous section, we studied degree two polynomials with real coefficients. Many of the polynomials we will encounter later on will have real coefficients. In this section we will therefore collect some facts about such polynomials. Complex conjugation as introduced in Definition 3.8, will play an important role. Complex conjugation has several nice properties. We list some of these in the following lemma.

|||| Lemma 4.8

Let $z, z_1, z_2 \in \mathbb{C}$ be complex numbers. Then it holds that

1. $\overline{\overline{z}} = z$,
2. $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$,
3. $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$,
4. $\overline{1/z} = 1/\overline{z}$ provided $z \neq 0$,
5. $\overline{z^n} = (\overline{z})^n$, where $n \in \mathbb{Z}$.

Proof. We will prove the second and third item of the lemma. Proving the remaining items is left to the reader. For a sum of two complex numbers $z_1 = a + bi$ and $z_2 = c + di$ on rectangular form it holds that

$$\overline{z_1 + z_2} = \overline{(a + c) + (b + d)i} = (a + c) - (b + d)i = (a - bi) + (c - di) = \overline{z_1} + \overline{z_2}.$$

For a product of two complex numbers $z_1 = a + bi$ and $z_2 = c + di$ on rectangular form we have $z_1 \cdot z_2 = (ac - bd) + (ad + bc)i$. Therefore

$$\overline{z_1 \cdot z_2} = (ac - bd) - (ad + bc)i.$$

On the other hand,

$$\begin{aligned} \overline{z_1} \cdot \overline{z_2} &= (a - bi) \cdot (c - di) \\ &= ac - adi - bci + (-b) \cdot (-d)i^2 \\ &= ac - (ad + bc)i + bd \cdot (-1) \\ &= ac - bd - (ad + bc)i. \end{aligned}$$

This shows that $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$. □

|||| Example 4.9

Express the following complex numbers on rectangular form.

1. $\overline{-3+6i}$
2. $\overline{\pi}$
3. $\overline{-97i}$

Answer:

1. From the definition of the complex conjugate we find $\overline{-3+6i} = -3-6i$.
2. $\overline{\pi} = \overline{\pi + 0i} = \pi - 0i = \pi$. This illustrates the more general fact that $\overline{z} = z$, if z is a real number.
3. $\overline{-97i} = -(-97i) = 97i$. It turns out that more generally $\overline{z} = -z$ for all purely imaginary numbers.

Complex conjugation also interacts well with the complex exponential function.

|||| Lemma 4.10

Let $z \in \mathbb{C}$ be a complex number and $\alpha \in \mathbb{R}$ a real number. It holds that

1. $\overline{e^z} = e^{\overline{z}}$,
2. $\overline{e^{i\alpha}} = e^{-i\alpha}$,
3. $\overline{z} = |z|e^{-i\arg(z)}$.

Proof. We prove the first two parts of the lemma. The third part of the lemma is illustrated in Figure 4.2. Suppose that $z = a + bi$ is the rectangular form of z. From the

definition of the complex exponential function we find that

$$\overline{e^z} = \overline{e^a \cos(b) + e^a \sin(b)i} = e^a \cos(b) - e^a \sin(b)i$$

$$= e^a \cos(-b) + e^a \sin(-b)i = e^{a-bi} = e^{\overline{z}}.$$

If $z = i\alpha$ (with $\alpha \in \mathbb{R}$) we get the special case

$$\overline{e^{i\alpha}} = e^{\overline{i\alpha}} = e^{-i\alpha}.$$

□

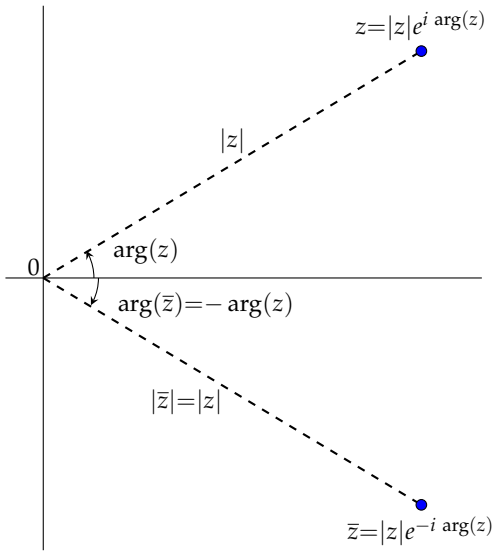


Figure 4.2: Polar form of a complex number z and its complex conjugate $\overline{z}$.

|||| Example 4.11

Write the complex number $\overline{5e^{i\pi/3}}$ in polar form.

Answer:

$\overline{5e^{i\pi/3}} = \overline{5}\,\overline{e^{i\pi/3}} = 5e^{-i\pi/3}$. This illustrates the third part of the previous lemma, which says that $\overline{z} = |z|e^{-i\arg(z)}$.

Now let us return to our discussion of polynomials with real coefficients. The reason we have introduced complex conjugation is the following property:

|||| Lemma 4.12

Let $p(Z) \in \mathbb{R}[Z]$ be a polynomial with real coefficients and let $\lambda \in \mathbb{C}$ be a root of $p(z)$. Then the complex number $\overline{\lambda} \in \mathbb{C}$ is also a root of $p(Z)$.

Proof. Let us write $p(Z) = a_n Z^n + \cdots + a_1 Z + a_0$. Since $p(Z)$ has real coefficients, it holds that $a_n, \ldots, a_0 \in \mathbb{R}$. It is given that $\lambda \in \mathbb{C}$ is a root of $p(Z)$ and therefore it holds that

$$0 = a_n \lambda^n + \cdots + a_1 \lambda + a_0.$$

We will now show that $\overline{\lambda}$ is a root of $p(Z)$ as well, by taking the complex conjugate in this equation. We find that

$$0 = \overline{a_n \lambda^n + \cdots + a_1 \lambda + a_0}.$$

Using this and the properties given in Lemma 4.8, we get:

$$\begin{aligned}
0 &= \overline{a_n \lambda^n + a_{n-1}\lambda^{n-1} \cdots + a_1 \lambda + a_0} \\
&= \overline{a_n \lambda^n} + \overline{a_{n-1}\lambda^{n-1}} + \cdots + \overline{a_1 \lambda} + \overline{a_0} \\
&= \overline{a_n}\,\overline{\lambda^n} + \overline{a_{n-1}}\,\overline{\lambda^{n-1}} + \cdots + \overline{a_1}\,\overline{\lambda} + \overline{a_0} \\
&= \overline{a_n}(\overline{\lambda})^n + \overline{a_{n-1}}(\overline{\lambda})^{n-1} + \cdots + \overline{a_1}\,\overline{\lambda} + \overline{a_0} \\
&= a_n(\overline{\lambda})^n + a_{n-1}(\overline{\lambda})^{n-1} + \cdots + a_1\,\overline{\lambda} + a_0 \\
&= p(\overline{\lambda})
\end{aligned}$$

In the fifth equality we have used that the coefficients of the polynomial $p(Z)$ are real numbers, so that $\overline{a_j} = a_j$ for all j between 0 and n. We have now shown that $p(\overline{\lambda}) = 0$ and hence can conclude that $\overline{\lambda}$ is a root of the polynomial $p(Z)$ as well. □

Lemma 4.12 has the following consequence: non-real roots of a polynomial with real coefficients come in pairs. Take for example the polynomial $2Z^2 - 4Z + 10$. We have seen in Example 4.7 that one of its roots is $1 + 2i$. Lemma 4.12 implies that the complex number $1 - 2i$ then is a root of $2Z^2 - 4Z + 10$ as well. We have seen in Example 4.7 that this indeed is the case.

4.4 Binomials

In this section we look at polynomials of the form $Z^n - w$ for some natural number $n \in \mathbb{N}$ and a complex number $w \in \mathbb{C}$ different from 0. The number n is the degree of the polynomial $Z^n - w$. Because a polynomial of the form $Z^n - w$ only has two terms, namely Z^n and $-w$, it is often called a *binomial*. The corresponding equation $z^n = w$ is called a *binomial equation*. We will give an exact expression for all roots of a binomial $Z^n - w \in \mathbb{C}[Z]$. This means that we have to compute all $z \in \mathbb{C}$ satisfying the equation $z^n = w$. It turns out that the polar form of the complex number w is of great help.

|||| Theorem 4.13

Let $w \in \mathbb{C}\backslash\{0\}$. The equation $z^n = w$ has exactly n different solutions, namely:

$$z = \sqrt[n]{|w|}e^{i(\frac{\arg(w)}{n} + p\frac{2\pi}{n})},\ p \in \{0, \ldots, n-1\}.$$

Here $\sqrt[n]{|w|}$ denotes the unique positive real number satisfying $\left(\sqrt[n]{|w|}\right)^n = |w|$.

Proof. The main idea of this proof is to try to find all solutions z to the equation $z^n = w$ in polar form. Therefore we write $z = |z|e^{iu}$ and we will try to determine the possible values of $|z|$ and u such that $z^n = |w|e^{i\alpha}$. In the first place we have $z^n = (|z|e^{iu})^n = |z|^n e^{inu}$ and this expression should be equal to $|w|e^{i\alpha}$. This holds if and only if $|w| = |z|^n$ and $e^{inu} = e^{i\alpha}$, or in other words, if and only if $|w| = |z|^n$ and $e^{i(nu-\alpha)} = 1$. The equation $|w| = |z|^n$ has exactly one solution for $|z| \in \mathbb{R}_{>0}$, namely $|z| = \sqrt[n]{|w|}$, while according to Lemma 3.26, the equation $e^{i(nu-\alpha)} = 1$ is satisfied if and only if $nu - \alpha = \arg(1)$. The possible arguments of 1 are precisely the integral multiples of 2π, that is to say, $\arg(1) = p2\pi$ for some integer $p \in \mathbb{Z}$.

All solutions to $z^n = w$ are therefore of the form $z = \sqrt[n]{|w|}e^{i(\frac{\alpha}{n} + p\frac{2\pi}{n})}$, where $p \in \mathbb{Z}$. In principle, we find a solution for any choice of $p \in \mathbb{Z}$, but when p runs through the set $\{0, \ldots, n-1\}$ we already get all different possibilities for z. □

When drawn in the complex plane, the solutions to the equation $z^n = w$ form the vertices of a regular n-gon with center in 0. Let us illustrate this in an example.

|||| Example 4.14

In this example we will find all roots of the polynomial $Z^4 + 8 - i8\sqrt{3}$ and write them in rectangular form.

Answer: We can use Theorem 4.13, with $n = 4$ and $w = -(8 - i8\sqrt{3})$. First, we need to write the complex number $-(8 - i8\sqrt{3} = -8 + i8\sqrt{3})$ in polar form. We have

$$|-8 + i8\sqrt{3}| = \sqrt{(-8)^2 + (8\sqrt{3})^2} = 16$$

and

$$\arg(-8 + i8\sqrt{3}) = \arctan(8\sqrt{3}/(-8)) + \pi = 2\pi/3.$$

Therefore we find that $-8 + i8\sqrt{3} = 16e^{i2\pi/3}$, which is the desired polar form. According to Theorem 4.13 all solutions to $z^4 = -8 + i8\sqrt{3}$ are given by:

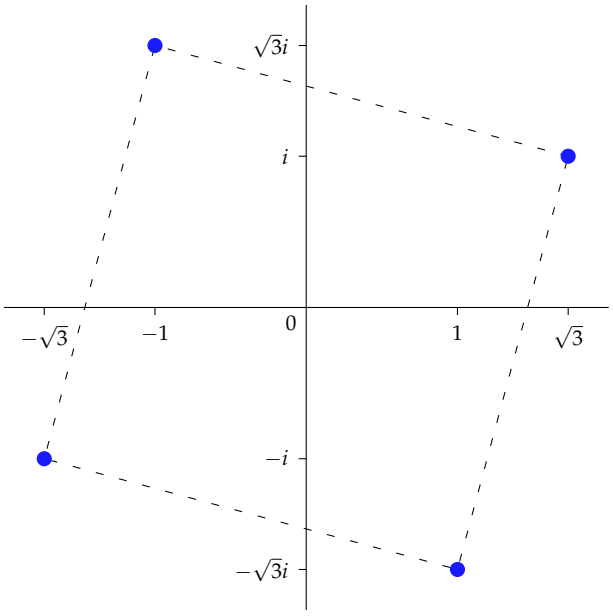
$$z = \sqrt[4]{16}e^{i\left(\frac{2\pi}{3\cdot 4} + p\frac{2\pi}{4}\right)}, \text{where } p \text{ can be chosen freely from the set } \{0,1,2,3\}, \text{ so}$$

$$z = 2e^{i\frac{\pi}{6}} \quad \vee \quad z = 2e^{i\frac{2\pi}{3}} \quad \vee \quad z = 2e^{i\frac{7\pi}{6}} \quad \vee \quad z = 2e^{i\frac{5\pi}{3}}.$$

Now we still need to write these roots in rectangular form. Using the formula $e^{it} = \cos(t) + i\sin(t)$ we get:

$$z = \sqrt{3} + i \quad \vee \quad z = -1 + i\sqrt{3} \quad \vee \quad z = -\sqrt{3} - i \quad \vee \quad z = 1 - i\sqrt{3}.$$

As remarked after Theorem 4.13, these solutions form the vertices of a regular 4-gon (that is to say, a square) with center in zero. This is indeed the case as shown in the following figure.



4.4.1 Polynomials in $\mathbb{C}[Z]$ of degree two

In Section 4.2, we have seen how to find the roots of a degree two polynomials in $\mathbb{R}[Z]$. Now that we know how to find the roots of binomial polynomials, we can find the roots of a degree two polynomials in $\mathbb{C}[Z]$ without much additional effort. The main observation is that for any polynomial $aZ^2 + bZ + c \in \mathbb{C}[Z]$ such that $a \neq 0$, Equation (4-3) is still valid. Hence $az^2 + bz + c = 0 \Leftrightarrow (2az + b)^2 = b^2 - 4ac$. We know from Theorem 4.13 that the equation $t^2 = b^2 - 4ac$ has exactly two solutions, say s and $se^{i\pi} = -s$. Then $az^2 + bz + c = 0 \Leftrightarrow 2az + b = s \vee 2az + b = -s$. Solving for z, we then obtain the following result:

|||| Theorem 4.15

Let $p(Z) = aZ^2 + bZ + c \in \mathbb{C}[Z]$ be a polynomial of degree two. Further, let $s \in \mathbb{C}$ be a solution to the binomial equation $s^2 = b^2 - 4ac$. Then $p(Z)$ has precisely the following roots:

$$\frac{-b+s}{2a} \text{ and } \frac{-b-s}{2a}.$$

|||| Example 4.16

As an example, let us find the roots of the polynomial $Z^2 + 2Z + 1 - i$.

Answer: The discriminant of the polynomial $Z^2 + 2Z + 1 - i$ is equal to $2^2 - 4 \cdot 1 \cdot (1 - i) = 4i$. Therefore, we first need to solve the binomial equation $s^2 = 4i$. We have $|4i| = 4$ and $\mathrm{Arg}(4i) = \pi/2$. Using Theorem 4.13, we see that the equation $s^2 = 4i$ has solutions

$$2 \cdot e^{\pi/4\, i} = 2 \cdot (\cos(\pi/4) + i \sin(\pi/4)) = 2 \cdot (\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} i) = \sqrt{2} + \sqrt{2}\, i$$

and

$$2 \cdot e^{(\pi/4+\pi)\, i} = 2 \cdot (\cos(5\pi/4) + i \sin(5\pi/4)) = 2 \cdot (-\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2} i) = -\sqrt{2} - \sqrt{2}\, i.$$

Hence using Theorem 4.15, we obtain that the roots of the polynomial $Z^2 + Z + 1 - i$ are given by

$$\frac{-2 + \sqrt{2} + i\sqrt{2}}{2} = -1 + \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} i \quad \text{and} \quad \frac{-2 - \sqrt{2} - i\sqrt{2}}{2} = -1 - \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2} i.$$

4.5 The division algorithm

In the previous section, we have seen how to find the roots of some specific polynomials. To study the behaviour of roots for more general polynomials, we begin with the following observation:

|||| Lemma 4.17

Let $p(Z) \in \mathbb{C}[Z]$ be a polynomial and suppose that $p(Z) = p_1(Z) \cdot p_2(Z)$ for certain polynomials $p_1(Z), p_2(Z) \in \mathbb{C}[Z]$. Further, let $\lambda \in \mathbb{C}$. Then λ is a root of $p(z)$ if and only if λ is a root of $p_1(z)$ or of $p_2(z)$.

Before proving this lemma, let us relate the statement of the lemma to propositional logic from Note 1 to clarify what really is stated. A statement like

$$\text{“}\lambda \text{ is a root of } p(z) \text{ if and only if } \lambda \text{ is a root of } p_1(z) \text{ or of } p_2(z)\text{”}$$

in a mathematical text, is just a way to express a statement from propositional logic into more common language. Reformulating everything in propositional logic, we simply get the statement

$$\lambda \text{ is a root of } p(z) \quad \Leftrightarrow \quad \lambda \text{ is a root of } p_1(z) \ \vee \ \lambda \text{ is a root of } p_2(z).$$

We can even go further and remove all words:

$$p(\lambda) = 0 \quad \Leftrightarrow \quad p_1(\lambda) = 0 \ \vee \ p_2(\lambda) = 0.$$

It is a good habit to make sure that you understand what a mathematical statement, when formulated in common language, really means. Here it is for example perfectly possible that λ is a root of both $p_1(Z)$ and $p_2(Z)$, even though in language “or” often is used in the meaning of “either one or the other, but not both”. In mathematical texts, “or” typically has the same meaning as “$\vee$”. With this in mind, let us continue to the proof of the lemma:

Proof. The number λ is a root of $p(Z)$ if and only if $p(\lambda) = 0$. Since $p(Z) = p_1(Z)p_2(Z)$ this is equivalent to saying that $p_1(\lambda)p_2(\lambda) = 0$ and therefore with the statement that $p_1(\lambda) = 0 \vee p_2(\lambda) = 0$. This statement is logically equivalent to saying that λ is a root of $p_1(z)$ or of $p_2(z)$. □

If one wants to find all roots of a polynomial, the above lemma suggests that it is always a good idea to try to write the polynomial as a product of polynomials of lower degree. If $p(Z) = p_1(Z) \cdot p_2(Z)$ as in the previous lemma, one says that $p_1(Z)$ and $p_2(Z)$ are *factors* of the polynomial $p(Z)$. It is therefore useful to have an algorithm that allows one to decide whether or not a given polynomial $p_1(Z) \in \mathbb{C}[Z]$ is a factor of a given second polynomial $p(Z) \in \mathbb{C}[Z]$. Equation (4-1) is already of some help, since it implies that $p(Z) = p_1(Z) \cdot p_2(Z)$ can only be true if $\deg p(Z) = \deg p_1(Z) + \deg p_2(Z)$. In particular, $p_1(Z)$ cannot be a factor of $p(Z)$ if $\deg p_1(Z) > \deg p(Z)$. However, this still leaves the case $\deg p_1(Z) \leq \deg p(Z)$ open. Before giving the algorithm that solves the problem completely, let us first consider a few examples.

|||| Example 4.18

1. Decide if the polynomial $Z + 3$ is a factor of the polynomial $2Z^2 + 3Z - 9$.
2. Decide if the polynomial $Z + 4$ is a factor of the polynomial $3Z^3 + 2Z + 1$.
3. Decide if the polynomial $2Z^2 + Z + 3$ is a factor of the polynomial $6Z^4 + 3Z^3 + 19Z^2 + 5Z + 15$.

Answer:

1. We will try to find a polynomial $q(Z) \in \mathbb{C}[Z]$ such that $(Z + 3) \cdot q(Z) = 2Z^2 + 3Z - 9$. If $q(Z)$ exists, it should have degree 1 using Equation (4-1). Hence if $q(Z)$ exists, it should be of the form $q(Z) = b_1 Z + b_0$ for certain numbers $b_1, b_0 \in \mathbb{C}$. We first try to find b_1. Without simplifying the product $(Z + 3) \cdot (b_1 Z + b_0)$ we can already see that the highest power of Z in the product is 2 and that the coefficient of Z^2 in the product is b_1. This means that $(Z + 3) \cdot (b_1 Z + b_0) = b_1 Z^2 +$ terms of degree less than 2. On the other hand we want that $(Z + 3) \cdot (b_1 Z + b_0) = 2Z^2 + 3Z - 9$. We see that b_1 has to be 2. Now that we know that $b_1 = 2$, we will determine b_0. On the one hand we want that $(Z + 3) \cdot (2Z + b_0) = 2Z^2 + 3Z - 9$, but on the other hand we can write $(Z + 3) \cdot (2Z + b_0) = (Z + 3) \cdot 2Z + (Z + 3) \cdot b_0$. Therefore, we can conclude that

$$(Z + 3) \cdot b_0 = 2Z^2 + 3Z - 9 - (Z + 3) \cdot 2Z = -3Z - 9. \tag{4-4}$$

   The important observation here is that previously we have chosen b_1 in such a way that the Z^2 term in Equation (4-4) is gone. By looking at the coefficients of Z, we conclude that $b_0 = -3$. We have shown the implication $(Z + 3) \cdot q(Z) = 2Z^2 + 3Z - 9 \Rightarrow q(Z) = 2Z - 3$. A direct check verifies that indeed $2Z^2 + 3Z - 9 = (Z + 3) \cdot (2Z - 3)$. We can conclude that indeed $Z + 3$ is a factor of $2Z^2 + 3Z - 9$. Since -3 is the root of $Z + 3$,

Lemma 4.17 then implies that -3 is also a root of the polynomial $2Z^2+3Z-9$. Indeed, we have $2\cdot(-3)^2+3\cdot(-3)-9=0$.

There is a more convenient way to write down the calculations we just carried out. The first step was to calculate b_1 and to subtract $b_1\cdot(Z+3)$ from $2Z^2+3Z-9$:

$$\begin{array}{l|l|l} Z+3 & 2Z^2+3Z-9 & 2Z \\ \hline & \underline{2Z^2+6Z} & \\ & -3Z-9 & \end{array}$$

The first line contains the polynomials we start with $Z+3$ and $2Z^2+3Z-9$ as well as all terms of $q(Z)$ we have calculated in the first step. The second line consists of the multiple of $Z+3$ which we subtracted from $2Z^2+3Z-9$ in Equation (4-4). The third line gives, after some simplifications, the expression $2Z^2+3Z-9-2Z\cdot(Z+3)$. We also got this in the righthand side of Equation (4-4). The next step was to determine the b_0. We again get that $b_0=-3$ and update the above scheme as follows:

$$\begin{array}{l|l|l} Z+3 & 2Z^2+3Z-9 & 2Z-3 \\ \hline & \underline{2Z^2+6Z} & \\ & -3Z-9 & \\ & \underline{-3Z-9} & \\ & 0 & \end{array}$$

This just means that $2Z^2+3Z-9-(Z+3)\cdot(2Z-3)=0$. This zero on the righthand side comes from the last line in the above scheme. The conclusion is therefore that $Z+3$ is a factor of the polynomial $2Z^2+3Z-9$. More than that we can even write the factorization down, since we showed that $2Z^2+3Z-9=(Z+3)\cdot(2Z-3)$.

2. This time, let us investigate if the polynomial $Z+4$ is a factor of the polynomial $3Z^3+2Z+1$. We try to find a polynomial $q(Z)$ such that $(Z+4)\cdot q(Z)=3Z^3+2Z+1$. We see that $q(Z)$ should have degree 2, that is to say $q(Z)=b_2Z^2+b_1Z+b_0$, and we want to determine its three coefficients. By looking at the highest power of Z we see that $b_2=3$. This time we directly use the schematic procedure we described in the first part of this example. First we get:

$$\begin{array}{l|l|l} Z+4 & 3Z^3\qquad +2Z+1 & 3Z^2 \\ \hline & \underline{3Z^3+12Z^2} & \\ & -12Z^2+2Z+1 & \end{array}$$

Now we can see that the coefficient of Z in $q(Z)$ should be -12 and we find:

$$\begin{array}{l|l|l} Z+4 & 3Z^3\qquad +\ 2Z+1 & 3Z^2-12Z \\ \hline & \underline{3Z^3+12Z^2} & \\ & -12Z^2+\ 2Z+1 & \\ & \underline{-12Z^2-48Z} & \\ & 50Z+1 & \end{array}$$

We can now read of that the constant term b_0 of $q(Z)$ should be 50 and we get:

$$\begin{array}{l|l|l} Z+4 & 3Z^3 \quad + \; 2Z + \; 1 & 3Z^2-12Z+50 \\ \hline & \underline{3Z^3+12Z^2} & \\ & -12Z^2+\; 2Z+\; 1 & \\ & \underline{-12Z^2-48Z} & \\ & \qquad 50Z+\; 1 & \\ & \qquad \underline{50Z+200} & \\ & \qquad\quad -199 & \end{array}$$

This time we do not get a zero in the last line. What the above scheme actually shows is that $3Z^3+2Z+1-(Z+4)\cdot(3Z^2-12Z+50) = -199$. This means that $Z+4$ cannot be a factor of $3Z^3+2Z+1$, since then $Z+4$ would also be a factor of $3Z^3+2Z+1-(Z+4)\cdot(3Z^2-12Z+50) = -199$. This would be impossible, since $\deg(Z+4) = 1 > 0 = \deg(-199)$. Note that -4 is not a root of the polynomial $3Z^3+2Z+1$, since $3\cdot(-4)^3+2\cdot(-4)+1 = -199$.

3. We state the schematic procedure only this time:

$$\begin{array}{l|l|l} 2Z^2+Z+3 & 6Z^4+3Z^3+19Z^2+5Z+15 & 3Z^2+5 \\ \hline & \underline{6Z^4+3Z^3+\; 9Z^2} & \\ & \qquad\qquad 10Z^2+5Z+15 & \\ & \qquad\qquad \underline{10Z^2+5Z+15} & \\ & \qquad\qquad\qquad 0 & \end{array}$$

The conclusion is that $6Z^4+3Z^3+19Z^2+5Z+15-(2Z^2+Z+3)\cdot(3Z^2+5) = 0$ and therefore that $6Z^4+3Z^3+19Z^2+5Z+15 = (2Z^2+Z+3)\cdot(3Z^2+5)$. Hence $2Z^2+Z+3$ is a factor of the polynomial $6Z^4+3Z^3+19Z^2+5Z+15$.

The algorithm described in the above examples is called *polynomial division* or the *division algorithm* or sometimes also *long division*. Let us describe it in full generality.

Given as input are two polynomials $p(Z), d(Z) \in \mathbb{C}[Z]$, where $d(Z)$ is not the zero polynomial. What we want, is to compute two polynomials $q(Z)$ and $r(Z)$ in $\mathbb{C}[Z]$ such that:

1. $p(Z) = d(Z)q(Z) + r(Z)$.
2. $r(Z) = 0 \quad \vee \quad deg(r(z)) < \deg(d(z))$.

The produced polynomial $q(Z)$ is called the *quotient* of $p(Z)$ modulo $d(Z)$, while the

polynomial $r(Z)$ is called the *remainder* of $p(Z)$ modulo $d(Z)$. The polynomial $d(Z)$ is a factor of $p(Z)$ if and only if this remainder is the zero polynomial. Hence the division algorithm can also be used to determine if any given polynomial divides $p(Z)$.

To find the quotient and remainder, we start the following schematic procedure:

$$\begin{array}{l|ll} d(Z) & p(Z) & 0 \end{array}$$

If we are lucky, we have $\deg p(Z) < \deg d(Z)$. In this case, we can already stop the division algorithm and return the values $q(Z) = 0$ and $r(Z) = p(Z)$. Otherwise, we would start the long division and find a simple multiple of $d(Z)$ that has the same degree and leading coefficient as $p(Z)$. Now let us denote the degree of $d(Z)$ by m, the leading coefficient of $d(Z)$ by d_m, and the leading coefficient of $p(Z)$ by b. Then the polynomial $bd_m^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ has exactly the same degree and leading coefficient as $p(Z)$. Hence we update the schematic procedure as follows:

$$\begin{array}{l|ll} d(Z) & p(Z) & bd_m{}^{-1}Z^{\deg p(Z)-m} \\ & bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z) & \\ \hline & p(Z) - bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z) & \end{array}$$

Note that the degree of the polynomial $p(Z) - bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ is strictly less than $\deg p(Z)$, since the leading coefficients of $p(Z)$ and $bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ are the same and therefore cancel each other when the difference of the two polynomials is taken. If it so happens that the degree of the resulting polynomial $p(Z) - bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ is strictly less than that of $d(Z)$, we are done and can return as answer the polynomials $p(Z) - bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ for $r(Z)$ and $bd_m{}^{-1}Z^{\deg p(Z)-m} \cdot d(Z)$ for $q(Z)$, otherwise we continue to the next line.

Now suppose that we have carried out the procedure a couple of times and have arrived at the following:

$$\begin{array}{l|ll} d(Z) & p(Z) & q^*(Z) \\ & \ddots & \\ \hline & r^*(Z) & \end{array}$$

If $\deg r^*(Z) < \deg d(Z)$, then we are already done and can return $q^*(Z)$ and $r^*(Z)$ as the quotient and remainder we are looking for. Otherwise, we perform one more step in the long division and find a simple multiple of $d(Z)$ that has the same degree and leading coefficient as $r^*(Z)$. Very similarly as in the first step of the long division, now denoting by b the leading coefficient of $r^*(Z)$, we find that the polynomial $bd_m^{-1}Z^{\deg r^*(Z)-m} \cdot d(Z)$ has exactly the same degree and leading coefficient as $r^*(Z)$. Hence we update the

schematic procedure as follows:

$$\begin{array}{l|ll|l} d(Z) & p(Z) & & q^*(Z) + bd_m{}^{-1}Z^{\deg r^*(Z)-m} \\ & \ddots & & \\ \hline & r^*(Z) & & \\ & bd_m{}^{-1}Z^{\deg r^*(Z)-m} \cdot d(Z) & & \\ \hline & r^*(Z) - bd_m{}^{-1}Z^{\deg r^*(Z)-m} \cdot d(Z) & & \end{array}$$

Since at each step of the iteration, the degree of the polynomial at the bottom of the scheme decreases, we will after finitely many steps arrive at the situation:

$$\begin{array}{l|ll|l} d(Z) & p(Z) & & q(Z) \\ & \cdots & & \\ \hline & \ddots & & \\ & r(Z) & & \end{array}$$

Here $r(Z)$ is either the zero polynomial or $\deg r(Z) < \deg d(Z)$. The quotient and remainder are then the polynomials $q(Z)$ and $r(Z)$ found in the scheme. Let us for good measure also formulate this algorithm in pseudo-code. To indicate that the algorithm should keep running as long as $\deg r^*(Z) \geq \deg d(Z)$, we use what is known as a while loop in the pseudo-code.

Algorithm 7 for performing long division in $\mathbb{C}[Z]$

Input: $p(Z) \in \mathbb{C}[Z]$, $d(Z) \in \mathbb{C}[Z] \setminus \{0\}$.

$m \leftarrow \deg d(Z)$
$d_m \leftarrow$ leading coefficient of $d(Z)$
$q^*(Z) \leftarrow 0$ and $r^*(Z) \leftarrow p(Z)$
while $\deg r^*(Z) \geq m$ **do**
    $b \leftarrow$ leading coefficient of $r^*(Z)$
    $q^*(Z) \leftarrow q^*(Z) + bd_m^{-1}Z^{\deg r(Z)-m}$
    $r^*(Z) \leftarrow r^*(Z) - bd_m^{-1}Z^{\deg r^*(Z)-m} \cdot d(Z)$
return $q^*(Z), r^*(Z)$

4.6 Roots, multiplicities and factorizations

A surprising and beautiful theorem is that any polynomial $p(Z) \in \mathbb{C}[Z]$ of degree at least 1 has a root in $\mathbb{C}$. This result is often called the *fundamental theorem of algebra*. For future reference, let us state the theorem.

|||| Theorem 4.19 Fundamental theorem of algebra

Let $p(Z) \in \mathbb{C}[Z]$ be a polynomial of degree at least one. Then $p(Z)$ has a root $\lambda \in \mathbb{C}$.

We will not prove this theorem, since the proof is quite involved. We have seen that the theorem is true for degree two polynomials in Theorem 4.15. Note that not every polynomial needs to have a real root. For example, the polynomial $Z^2 + 1$ does not have a real root, but has a pair of (non-real) complex roots, namely i and $-i$.

Given a polynomial, it can be difficult or downright impossible to find a useful exact expression for its roots, but often a numerical approximation of the roots is sufficient. One can make a precise statement on the number of roots a polynomial can have though. We will see that if a polynomial has degree n, then it has n roots if we count the roots in a particular way. Now that we have the division algorithm as a tool, we start our investigation of roots of a polynomial.

|||| Lemma 4.20

Let $p(Z) \in \mathbb{C}[Z]$ be a polynomial of degree $n \geq 1$ and let $\lambda \in \mathbb{C}$ be a complex number. The number λ is a root of $p(Z)$ if and only if $Z - \lambda$ is a factor of $p(Z)$.

Proof. If $Z - \lambda$ is a factor of $p(Z)$, then there exists a polynomial $q(Z) \in \mathbb{C}[Z]$ such that $p(Z) = (Z - \lambda) \cdot q(Z)$. Therefore it then holds that $p(\lambda) = 0 \cdot q(\lambda) = 0$. This shows that λ is a root of $p(Z)$ if $Z - \lambda$ is a factor of $p(Z)$

Now suppose that λ is a root of $p(Z)$. Using the division algorithm we can find polynomials $q(Z)$ and $r(Z)$ such that

$$p(Z) = (Z - \lambda) \cdot q(z) + r(Z), \tag{4-5}$$

where $r(Z)$ is the zero polynomial, or $\deg(r(Z)) < \deg(Z - \lambda) = 1$. Since $r(Z) = 0$ or $\deg(r(Z)) < 1$, we see that $r(Z)$ actually is a constant $r \in \mathbb{C}$. By setting $Z = \lambda$ in Equation (4-5), we get that $p(\lambda) = r + 0 = r$. Therefore we actually have shown that $p(Z) = (Z - \lambda) \cdot q(Z) + p(\lambda)$. If λ is a root of $p(Z)$ (that is to say $p(\lambda) = 0$), we therefore get that $Z - \lambda$ is a factor of $p(Z)$. □

Using this lemma we can define the multiplicity of a root.

Definition 4.21

Let λ be a root of a polynomial $p(Z)$. The multiplicity of the root is defined to be the largest natural number $m \in \mathbb{N}$ such that $(Z-\lambda)^m$ is a factor of $p(Z)$. One says that λ is a root of $p(Z)$ of *multiplicity* m.

Note that Lemma 4.20 implies that any root of a polynomial has multiplicity at least 1. A root of multiplicity two is sometimes called a double root.

Example 4.22

Decide if -3 is a root of the following polynomials. If yes, determine its multiplicity.

- $p_1(Z) = 2Z^2 + 3Z - 9$.
- $p_2(Z) = Z^2 + 3Z + 1$.
- $p_3(Z) = Z^3 + 3Z^2 - 9Z - 27$.
- $p_4(Z) = (2Z^2 + 3Z - 9) \cdot (Z^3 + 3Z^2 - 9Z - 27) = 2Z^5 + 9Z^4 - 18Z^3 - 108Z^2 + 243$.

Answer:

1. We have $p_1(-3) = 18 - 9 - 9 = 0$. Therefore is -3 a root of the polynomial $2Z^2 + 3Z - 9$. We have seen in Example 4.18 that $2Z^2 + 3Z - 9 = (Z+3) \cdot (2Z-3)$. This means that the multiplicity of the root -3 equals 1. We can also see that the factor $2Z - 3$ gives rise to another root of $p_1(Z)$, namely the root $3/2$. This root also has multiplicity 1.
2. We have $p_2(-3) = 1$. Therefore -3 is not a root of $p_2(Z)$.
3. This time we have $p_3(-3) = 0$, so -3 is a root of $p_3(Z)$. Using the division algorithm, we find:

$$\begin{array}{r|l|l} Z+3 & Z^3+3Z^2-9Z-27 & Z^2-9 \\ \cline{2-2} & \underline{Z^3+3Z^2} & \\ & \quad -9Z-27 & \\ & \quad \underline{-9Z-27} & \\ & \qquad\quad 0 & \end{array}$$

Therefore it holds that $Z^3 + 3Z^2 - 9Z - 27 = (Z+3) \cdot (Z^2 - 9)$. The number -3 is also a root of the polynomial $Z^2 - 9$, so the multiplicity of the root -3 is at least 2. Actually, it holds that $Z^2 - 9 = (Z+3) \cdot (Z-3)$, so $Z^3 + 3Z^2 - 9Z - 27 = (Z+3) \cdot (Z^2-9) = (Z+3)^2 \cdot (Z-3)$. This means that the root -3 of $p_3(Z)$ has multiplicity 2. We also showed that 3 is a root of $p_3(Z)$ and that this root has multiplicity 1.

4. We have $p_4(Z) = p_1(Z)p_3(Z)$. From the first and the third part of this example, we get that $p_4(Z) = (Z+3)^3 \cdot (2Z-3) \cdot (Z-3)$. This means that the root -3 has multiplicity 3. We also see that the numbers $3/2$ and 3 are roots of $p_4(Z)$, both with multiplicity 1. The graph of real polynomial function that $p_4(Z)$ gives rise to, is given in Figure 4.3.

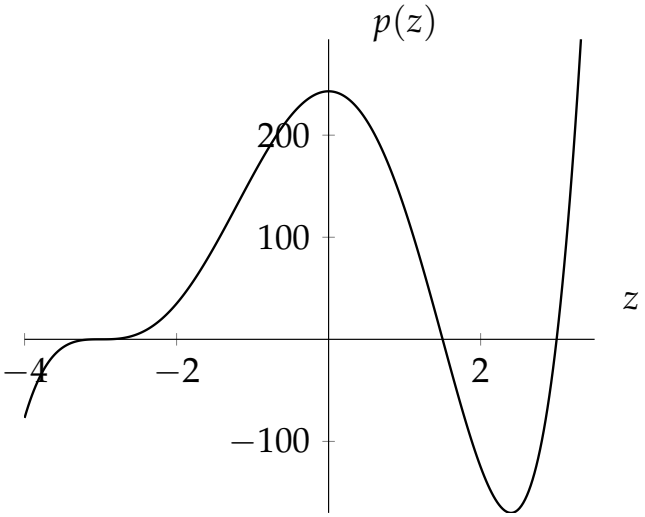


Figure 4.3: The graph of the polynomial function $p : \mathbb{R} \to \mathbb{R}$, where $p(z) = 2z^5 + 9z^4 - 18z^3 - 108z^2 + 243$.

The above example illustrates that there is a one to one correspondence between factors of degree one of a polynomial and the roots of a polynomial. The fundamental theorem of algebra (Theorem 4.19) says that each polynomial of degree at least 1 has a root. This has the following consequence:

|||| Theorem 4.23

Let $p(Z) = a_n Z^n + a_{n-1} Z^{n-1} + \cdots + a_1 Z + a_0$ be a polynomial of degree $n > 0$. Then there exist $\lambda_1, ..., \lambda_n \in \mathbb{C}$ such that

$$p(Z) = a_n \cdot (Z - \lambda_1) \cdots (Z - \lambda_n).$$

Proof. According to the fundamental theorem of algebra there exists a root $\lambda_1 \in \mathbb{C}$ of the polynomial $p(Z)$. Using Lemma 4.20, we can write $p(Z) = (Z - \lambda_1)q_1(Z)$ for a certain

polynomial $q_1(Z)$. Note that $\deg(q_1(Z)) = \deg(p(Z)) - 1$. If $q_1(Z)$ is a constant, we are done. Otherwise, we can apply the fundamental theory of algebra to the polynomial $q_1(Z)$ and find a root $\lambda_2 \in \mathbb{C}$ of $q_1(Z)$. Again using Lemma 4.20, we can write $q_1(Z) = (Z - \lambda_2) \cdot q_2(Z)$. This implies that $p(Z) = (Z - \lambda_1) \cdot (Z - \lambda_2) \cdot q_2(Z)$. Continuing in this way, we can write $p(Z)$ as a product of polynomials of degree one of the form $Z - \lambda$ times a constant c. Since the leading coefficient of $p(Z)$ is a_n, this constant c is equal to a_n. □

|||| Example 4.24

As an example we take the polynomial $p_4(Z) = 2Z^5 + 9Z^4 - 18Z^3 - 108Z^2 + 243$ from Example 4.22. We wish to write this polynomial as in Theorem 4.23. We have already seen that $p_4(Z) = (Z+3)^3 \cdot (2Z-3) \cdot (Z-3)$. By pulling out the 2 from the factor $2Z - 3$ we get:

$$p_4(Z) = 2 \cdot (Z+3)^3 \cdot (Z - 3/2) \cdot (Z-3) = 2 \cdot (Z+3) \cdot (Z+3) \cdot (Z+3) \cdot (Z-3/2) \cdot (Z-3).$$

In the notation of Theorem 4.23 we find that $\lambda_1 = -3$, $\lambda_2 = -3$, $\lambda_3 = -3$, $\lambda_4 = 3/2$, and $\lambda_5 = 3$. This illustrates once more that the multiplicities of the roots -3, $3/2$, and 3 are 3, 1, and 1. Note that the sum of all multiplicities is equal to 5, which is the degree of $p_4(Z)$.

In fact it always holds that the sum of all multiplicities of the roots of a polynomial is equal to its degree. In words one can therefore reformulate Theorem 4.23 as follows: a polynomial of degree $n \geq 1$ has exactly n roots, if the roots are counted with their multiplicities. For polynomials in $\mathbb{R}[Z]$, Theorem 4.23 has the following consequence

|||| Corollary 4.25

Any polynomial $p(Z) \in \mathbb{R}[Z]$ of degree at least one, can be written as the product of degree one and degree two polynomials from $\mathbb{R}[Z]$.

Proof. According to Theorem 4.23 any nonzero polynomial $p(Z)$ can be written as the product of the leading coefficient of $p(Z)$ and degree one factors of the form $Z - \lambda$. The $\lambda \in \mathbb{C}$ is a root of the polynomial $p(Z)$. Applying this to a polynomial $p(Z)$ with real coefficients, we see that the leading term is a real number as well, but the roots λ do not have to be real numbers. However, any real root λ gives rise to a factor of degree one with real coefficients, namely $Z - \lambda$.

Now let $\lambda \in \mathbb{C} \setminus \mathbb{R}$ be a root of $p(Z)$. Let us write $\lambda = a + bi$ in rectangular form. Since $\lambda \notin \mathbb{R}$, we know that $b \neq 0$. Lemma 4.12 implies that then the number $\overline{\lambda} = a - bi$ is also a root of $p(Z)$. Moreover, $\lambda \neq \overline{\lambda}$, since $b \neq 0$. Hence $Z - \lambda$ and $Z - \overline{\lambda}$ are two distinct factors of $p(Z)$ if we would work in $\mathbb{C}[Z]$. Now the idea is to multiply the factors $Z - \lambda$ and $Z - \overline{\lambda}$ together, since it turns out that $(Z - \lambda) \cdot (Z - \overline{\lambda})$ has real coefficients. Indeed, we have

$$\begin{aligned}(Z - \lambda) \cdot (Z - \overline{\lambda}) &= Z^2 - (\lambda + \overline{\lambda})Z + \lambda\overline{\lambda} \\ &= Z^2 - (a + bi + a - bi)Z + (a + bi) \cdot (a - bi) \\ &= Z^2 - 2aZ + (a^2 + b^2),\end{aligned}$$

which indeed is a polynomial of degree two in $\mathbb{R}[Z]$ since its coefficients are real numbers. In this way we can transform the factorization of $p(Z)$ in $\mathbb{C}[Z]$ from Theorem 4.23 into a factorization of $p(Z)$ in $\mathbb{R}[Z]$ in first and second degree factors with real coefficients. □

|||| Example 4.26

Write the following polynomials as a product of degree one and degree two polynomials with real coefficients.

1. $p_1(Z) = Z^3 - Z^2 + Z - 1$
2. $p_2(Z) = Z^4 + 4$

Answer:

1. The number 1 is a root of $p_1(Z)$, since $p(1) = 0$. Using the division algorithm, one can show that $p_1(Z) = (Z - 1) \cdot (Z^2 + 1)$. The polynomial $Z^2 + 1$ does not have any real root and therefore cannot be factorized further over the real numbers (over the complex numbers one could: $Z^2 + 1 = (Z + i) \cdot (Z - i)$). The desired factorization is therefore:
$$Z^3 - Z^2 + Z - 1 = (Z - 1) \cdot (Z^2 + 1).$$
2. Using the theory of Section 4.4 we can find all roots of the polynomial $Z^4 + 4$. In this way one can find the roots $1 + i, 1 - i, -1 + i$ and $-1 - i$. Therefore we have that
$$Z^4 + 4 = (Z - (1 + i)) \cdot (Z - (1 - i)) \cdot (Z - (-1 + i)) \cdot (Z - (-1 - i)).$$
As in the proof of Corollary 4.25 we can multiply pairs of complex conjugated factors together to get rid of the complex coefficients. Then we find that
$$(Z - (1 + i)) \cdot (Z - (1 - i)) = Z^2 - 2Z + 2$$

and

$$(Z-(-1+i))\cdot(Z-(-1-i)) = Z^2+2Z+2.$$

The desired factorization of Z^4+4 is therefore

$$Z^4+4 = (Z^2-2Z+2)\cdot(Z^2+2Z+2).$$

Note 5

Recursion and induction

5.1 Examples of a recursively defined functions

In this section, we introduce the concept of a recursively defined function. The concept of a *recursion* in this context is simply to define a function or an expression using that function or expression itself for other input values. Let us start with an example:

Example 5.1

The *factorial* function $\texttt{fac} : \mathbb{N} \to \mathbb{N}$ is defined by $n \mapsto 1 \cdot 2 \cdots n$. Hence n is mapped to the product of the first n positive integers. It is also very common to write $n!$ instead of $\texttt{fac}(n)$. We have for example $\texttt{fac}(1) = 1$, $\texttt{fac}(2) = 1 \cdot 2 = 2$, $\texttt{fac}(3) = 1 \cdot 2 \cdot 3 = 6$, $\texttt{fac}(4) = 1 \cdot 2 \cdot 3 \cdot 4 = 24$, etcetera. Now note that, if we want to compute the next value, $\texttt{fac}(5)$, we can use that we already know what $\texttt{fac}(4)$ is. Indeed,

$$\texttt{fac}(5) = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = (1 \cdot 2 \cdot 3 \cdot 4) \cdot 5 = \texttt{fac}(4) \cdot 5 = 24 \cdot 5 = 120.$$

In general, if for some $n > 1$, we already have computed $\texttt{fac}(n-1)$, we can compute the value of $\texttt{fac}(n)$ using that $\texttt{fac}(n) = \texttt{fac}(n-1) \cdot n$. This leads to the following algorithmic description of the factorial function:

Algorithm 8 $\texttt{fac}(n)$

Input: $n \in \mathbb{Z}_{\geq 1}$.

1: **if** $n = 1$ **then**
2: **return** 1
3: **else**
4: **return** $\texttt{fac}(n-1) \cdot n$.

This algorithm simply uses itself to compute $\texttt{fac}(n)$. More precisely, if $n = 1$, it directly returns 1 as the value for $\texttt{fac}(1)$, as prescribed in line 2 of the algorithm. Graphically, we can illustrate this as follows:

$\texttt{fac}(1)$

$$\texttt{fac}(1) = 1$$

$\texttt{return}\ 1$

If $n = 2$, the algorithm will go to line 4 and attempt to return $\texttt{fac}(2-1) \cdot 2$. However, this requires that first the value of $\texttt{fac}(1)$ is computed. Hence the algorithm will then start over, but now for the value 1. We have already seen that the algorithm returns 1 in that case. Now that the algorithm has arrived at the conclusion that $\texttt{fac}(1) = 1$, it can revisit line 4 and compute that $\texttt{fac}(2) = \texttt{fac}(1) \cdot 2 = 1 \cdot 2 = 2$. Hence the algorithm returns 2. Graphically, the situation is:

$\texttt{fac}(2)$

$$\texttt{fac}(2) = \texttt{fac}(1) \cdot 2$$

$\texttt{fac}(1)$

$$\texttt{fac}(1) = 1$$

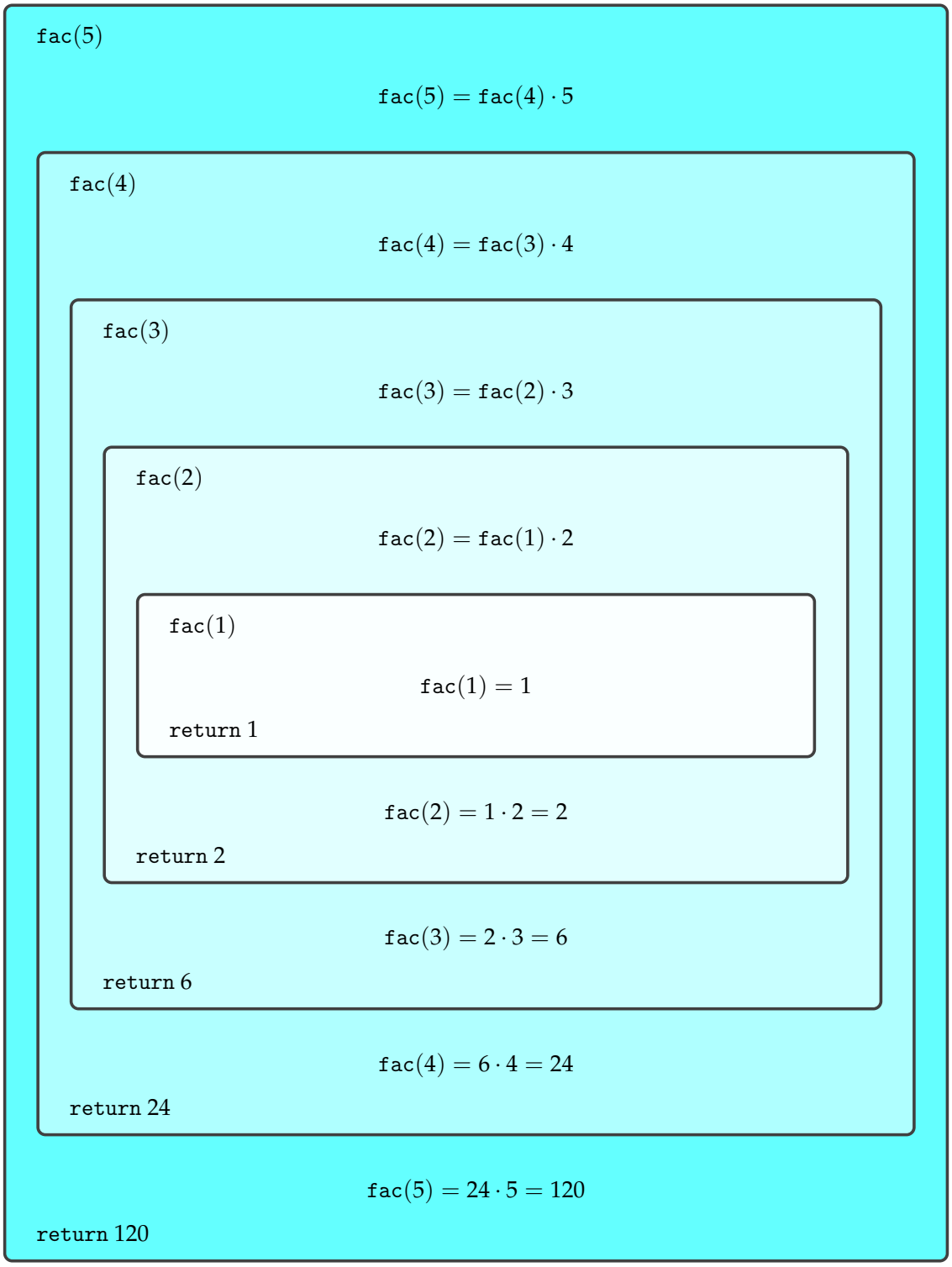
$\texttt{return}\ 1$

$$\texttt{fac}(2) = 1 \cdot 2 = 2$$

$\texttt{return}\ 2$

For larger values of n more "boxes inside other boxes" will appear, since the algorithm will need to use itself more often to compute its output for the smaller input values $n-1, n-$

$2, \ldots, 1$ before it can return its final output. For $n = 5$, the following graphical representation indicates what happens when this algorithm gets input value $n = 5$:



Since the algorithm uses itself while running (in algorithmic terms one often says that the algorithm *calls* itself), it is called a *recursive* algorithm. A recursive algorithm is simply an algorithm that might call itself for other input values in order to compute its final output value. Also in mathematics, recursions occur. In the context of this example, we have actually give a recursive definition of the factorial function:

$$\texttt{fac}(n) = \begin{cases} 1 & \text{if } n = 1, \\ \texttt{fac}(n-1) \cdot n & \text{if } n \geq 2. \end{cases} \tag{5-1}$$

What this example illustrates is the principle of a recursive definition: to define the values a function takes using that same function itself. Note by the way that it is also very common to define $0! = 1$, but that is another matter. Here is another example of a recursively defined function: Let $z \in \mathbb{C}$ be a complex number and define $f : \mathbb{N} \to \mathbb{C}$ recursively as:

$$f(n) = \begin{cases} z & \text{if } n = 1, \\ f(n-1) \cdot z & \text{if } n \geq 2. \end{cases} \tag{5-2}$$

Then $f(1) = z$, since this corresponds to the case $n = 1$ in the recursive definition. Further $f(2) = f(1) \cdot z$, since this is what the recursive definition gives for $n = 2$. Using that we already computed that $f(1) = z$, we may conclude that $f(2) = f(1) \cdot z = z \cdot z$. Finally using that $z \cdot z = z^2$, we see that $f(2) = f(1) \cdot z = z \cdot z = z^2$. Similarly, $f(3) = z^3$. Therefore it is perfectly reasonable to *define* the expression z^n for any natural number n recursively as $f(n)$. In previous chapters, we have used n-th powers of complex numbers several times. Now we have a more formal definition for it. In this light, it is also common to define $z^0 = 1$ and $z^{-n} = 1/z^n$ for any natural number n. This means that we now have defined very precisely what z^n means for any integer $n \in \mathbb{Z}$.

When attempting to define a function recursively, one should make sure afterwards that such a recursive description actually defines the function for all values from its domain. For the functions defined in equations (5-1) and (5-2) you can find a justification in Example 5.6, but feel free to skip that example on a first reading. For now, let us just show an example of a recursive description that does not work out. Let $g : \mathbb{N} \to \mathbb{C}$ be a function and suppose that

$$g(n) = \begin{cases} 1 & \text{if } n = 1, \\ g(n+1) & \text{if } n \geq 2. \end{cases}$$

By definition we see that $g(1) = 1$, but we do not have enough information to determine what $g(2)$ is. If we apply the recursive definition, we would just obtain that $g(2) =$

$g(3)$. Then attempting to compute $g(3)$, the recursion only yields that $g(3) = g(4)$. Continuing like this, we obtain that $g(2) = g(3) = g(4) = g(5) = \cdots$, but we never find out what $g(2)$ actually is.

As a final example of a recursive definition, we consider the famous Fibonacci numbers.

|||| Example 5.2

Let us now consider a recursive definition that looks slightly different. We are going to define recursively a function $F : \mathbb{N} \to \mathbb{N}$ whose values $F(1), F(2), F(3), F(4), \ldots$ are called the *Fibonacci numbers*:

$$F(n) = \begin{cases} 1 & \text{if } n = 1, \\ 1 & \text{if } n = 2, \\ F(n-1) + F(n-2) & \text{if } n \geq 3. \end{cases} \tag{5-3}$$

Let us see how this definition works in practice by computing the first Fibonacci numbers. First of all $F(1) = 1$, since if $n = 1$, the first line of equation (5-3) applies. If $n = 2$, the second line of equation (5-3) applies, so that $F(2) = 1$. For $n = 3$, the third line of equation (5-3) applies and we find that $F(3) = F(2) + F(1) = 1 + 1 = 2$. Similarly for $n = 4$, we find that $F(4) = F(3) + F(2) = 2 + 1 = 3$, using that we already have computed that $F(3) = 2$ before.

When dealing with a sequences of numbers, such as the Fibonacci numbers, it is quite common to change the notation a bit: instead of writing $F(n)$, one often writes F_n. In this notation we would get $F_1 = 1$, $F_2 = 1$, $F_3 = 2$, $F_4 = 3$ and so on. It turns out that it is possible to derive a closed formula expression for the Fibonacci numbers:

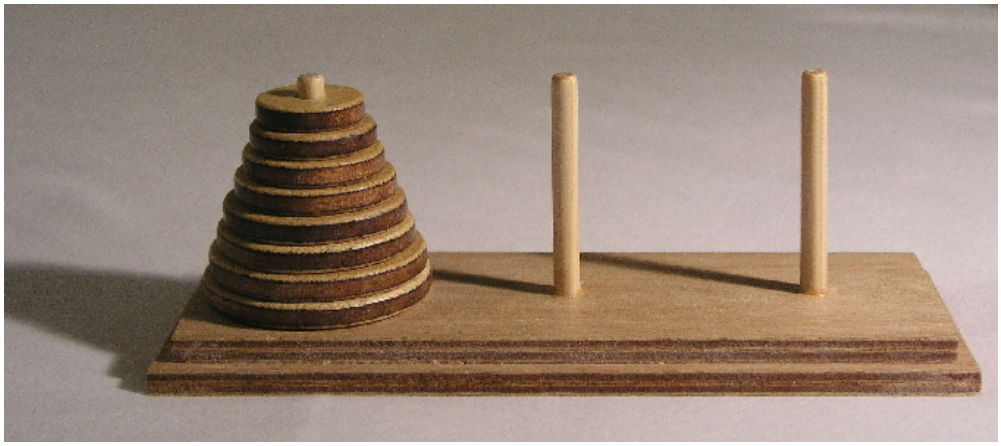
$$F_n = \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2} \right)^n . \tag{5-4}$$

We will come back to explaining how this expression comes about in a later chapter.

5.2 The towers of Hanoi

In this section, we further illustrate the usefulness of a recursive way of thinking when analyzing a puzzle called *the towers of Hanoi.* The towers of Hanoi is a puzzle on a board containing three upright sticks of equal lengths and sizes. Further there are various circular discs all of different diameter, each with a hole in the middle so they can be placed on a stick. In the starting position of the puzzle, all discs are stacked on the first stick. The disc with largest diameter is stacked first, the other discs in decreasing diameter size. The number of disks can vary. For an example with eight disks, see Figure 5.1.

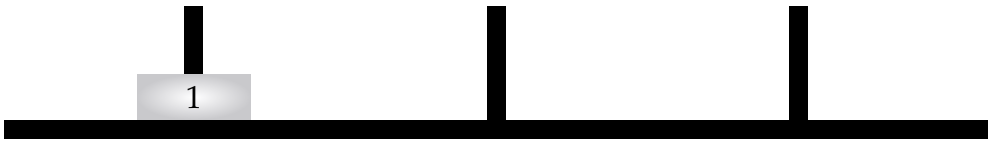
Figure 5.1: The tower of Hanoi with eight discs.



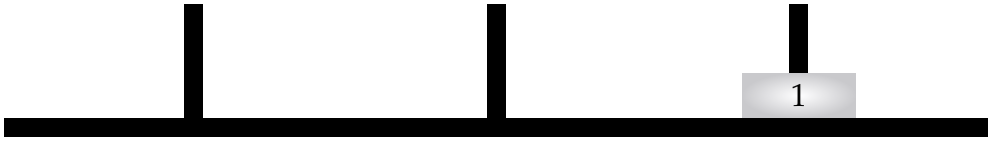
Now the goal of the puzzle is to move the stack of discs from the first to the third stick, stacked in the same way again from large to small. However, the challenge is that this has to be achieved following three rules:

- Only one disc may be moved at a time.
- Only a disc on top of a stack may be moved.
- A disc may only be placed on a larger disc.

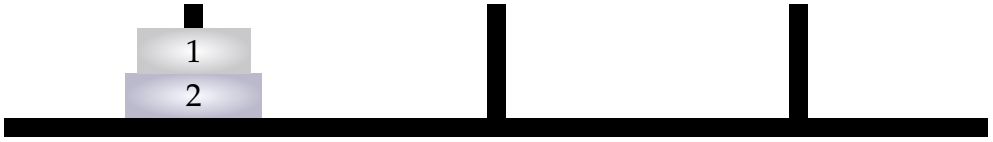
If there are only very few discs, it is not hard to solve the puzzle. If there are many discs, the game becomes more complicated and a priori it is not even clear if there always exists a solution. To get started, let us look at some examples with only a few discs. First of all, if there is only one disc, we can solve the puzzle in one move:



Move disc 1 from stick 1 to stick 3



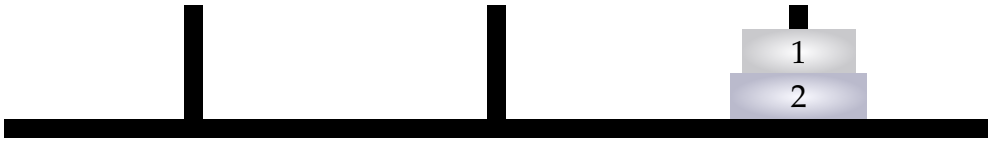
If there are two discs, the puzzle can be solved in three moves:



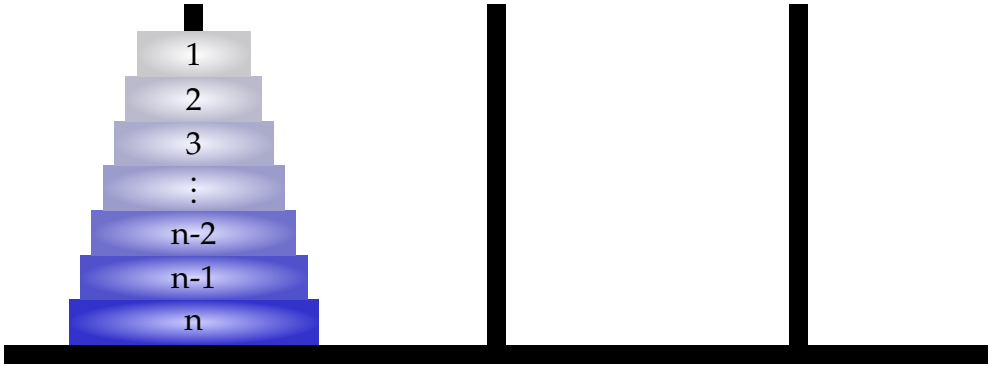
Move disc 1 from stick 1 to stick 2

Move disc 2 from stick 1 to stick 3

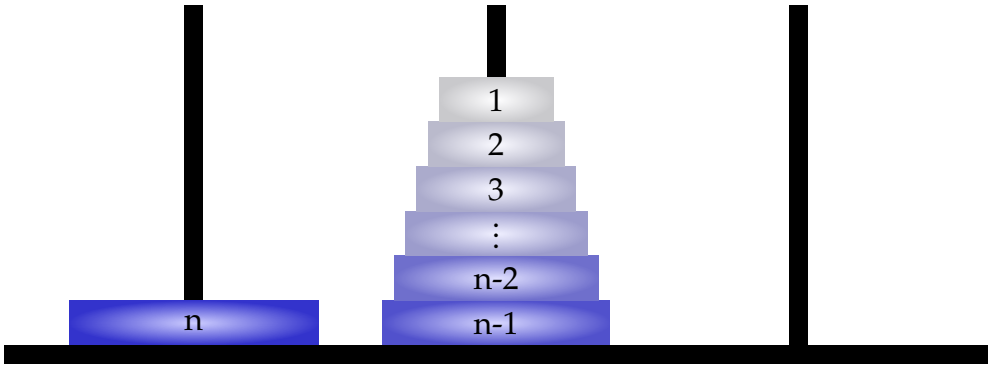
Move disc 1 from stick 2 to stick 3



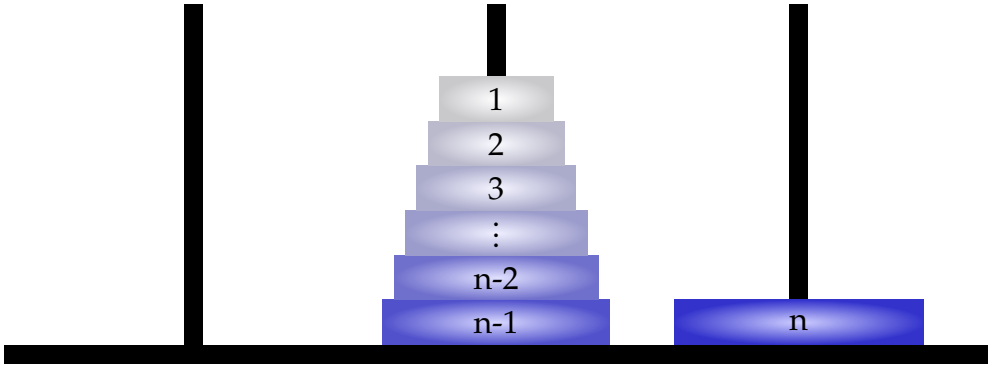
If there are three discs, it is still not so hard to solve the puzzle by some trial and error, but what if there are ten discs, or a hundred? To find a solution, let us try to think in a recursive way. We already know how to solve the puzzle for if there is only one disc (and also if there are two discs). Perhaps, just as for the factorial function, we can figure out what to do for a larger number of discs, say n discs, if we already would know what to do if there are less than n discs. Suppose therefore that $n \geq 2$ is a natural number and that we already know how to solve the puzzle if there are $n-1$ discs. This means that we know how to move a stack of $n-1$ discs from one stick to another stick. Then the following strategy works to move n discs:



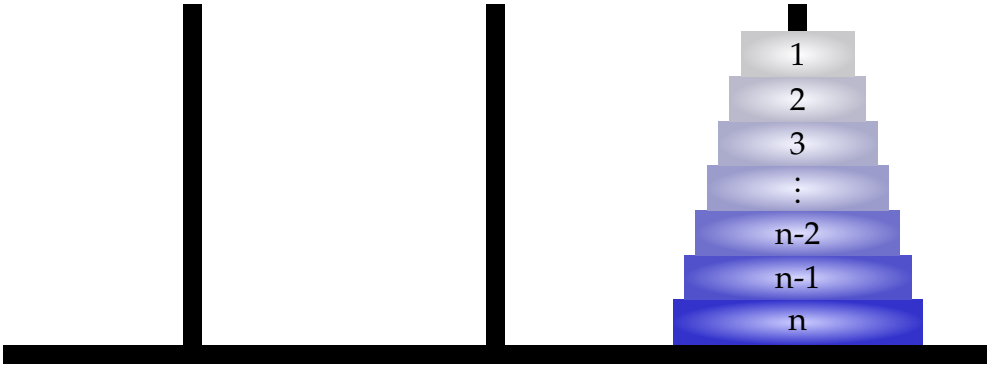
Using that we know how to move $n-1$ discs to another stick, move the stack of discs 1 to $n-1$ from stick 1 to stick 2.



Now move disc n from stick 1 to stick 3. This just takes one move.



Again using that we know how to move $n-1$ discs to another stick, move the stack of discs 1 to $n-1$ from stick 2 to stick 3.



This shows that the puzzle can be solved recursively! In particular, there is a solution for any number of discs.

5.3 The summation symbol Σ

If n is some natural number and $z_1, \ldots, z_n$ are complex numbers, then one can denote their sum by an expression like $z_1 + z_2 + \cdots + z_n$ or $z_1 + \cdots + z_n$. However, it is sometimes more convenient to have a more compact notation for this: $\sum_{i=1}^{n} z_i$. Using a recursive definition, we can be very precise:

$$\sum_{i=1}^{n} z_i = \begin{cases} z_1 & \text{if } n = 1, \\ \left(\sum_{i=1}^{n-1} z_i\right) + z_n & \text{if } n > 1. \end{cases} \tag{5-5}$$

Using this recursive definition, we obtain precisely what we wanted. One can simply use the definition and verify that indeed for small values of n one obtains:

n	$\sum_{i=1}^{n} z_i$
1	z_1
2	$z_1 + z_2$
3	$z_1 + z_2 + z_3$
4	$z_1 + z_2 + z_3 + z_4$

(5-6)

If $f : \mathbb{N} \to \mathbb{C}$ is a function, one similarly can replace the sum $f(1) + f(2) + \cdots + f(n)$ by the more compact expression $\sum_{i=1}^{n} f(i)$. Consider for example the expression $\sum_{i=1}^{n} i$, that is to say, the sum of the first n natural numbers. Similarly as in Table 5-6, we obtain

the following:

n	$\sum_{i=1}^{n} i$
1	1
2	$1+2=3$
3	$1+2+3=6$
4	$1+2+3+4=10$

(5-7)

Having this notation, will come in handy in various later chapters, but it is also heavily used in several other areas of mathematics and natural sciences.

5.4 Induction

In the previous section, we ended by solving the towers of Hanoi puzzle completely by approaching the problem in a recursive way. The number of moves our solution requires, can also be described recursively. If we denote by $T(n)$ the number of moves our strategy has for the puzzle with n discs, then we know that $T(1) = 1$ (the puzzle with only one disc can be solved in one move), but also that $T(n) = T(n-1) + 1 + T(n-1) = 2T(n-1) + 1$ for $n \geq 2$ (our strategy involved moving a stack of $n-1$ discs twice and a single move of the nth disc). In other words, we have

$$T(n) = \begin{cases} 1 & \text{if } n = 1, \\ 2 \cdot T(n-1) + 1 & \text{if } n \geq 2. \end{cases} \tag{5-8}$$

For instance $T(2) = 2 \cdot 1 + 1 = 3$, $T(3) = 2 \cdot 3 + 1 = 7$, and $T(4) = 2 \cdot 7 + 1 = 15$. It is striking that for these small values of n, the value of $T(n)$ is always one less than 2^n. Therefore one may "guess" that $T(n) = 2^n - 1$ for all natural number n. Let us test this conjecture, the word typically used instead of "guess", by computing $T(5)$. We have $T(5) = 2 \cdot T(4) + 1 = 2 \cdot 15 + 1 = 31$. This confirms our conjecture that $T(n) = 2^n - 1$ for $n = 5$. On the downside, all we know now is that the conjecture is true for all n in the set $\{1, 2, 3, 4, 5\}$. We could or course continue to verify our conjecture for more values of n by computing $T(6), T(7)$ and so on, but since there are infinitely many natural numbers, there is no way we can verify the formula $T(n) = 2^n - 1$ for *all* natural numbers n in this way. Fortunately, there is a very intuitive property of the natural numbers that can help us out and which we state without proof:

|||| Theorem 5.3 Induction principle

Let S be a subset of the natural numbers and assume that S has the following two properties:

1. $1 \in S$,
2. if $n-1 \in S$ for some arbitrary natural number $n \geq 2$, then also $n \in S$.

In this case, we have $S = \mathbb{N}$.

The statement in this theorem is often called the *induction principle* or simply *induction*. Requirement 1. ($1 \in S$) is called the *base case of the induction*, while requirement 2. (if $n-1 \in S$ for some natural number n, then also $n \in S$) is called the *induction step*. The reason that in requirement 2., the natural number n has to be at least two, is that otherwise $n-1$ might not be a natural number. Indeed, if $n = 1$, then $n - 1 = 0$, but 0 is not in $\mathbb{N}$. Requirement 2. can be reformulated in propositional logic as follows.

2. for all $n \in \mathbb{N}_{\geq 2}$: $\quad n-1 \in S \quad \Rightarrow \quad n \in S$.

Verifying requirement 2., that is to say, verifying the induction step, is typically done by showing that $n \in S$ is true if we assume that $n-1 \in S$. When verifying the induction step $n-1 \in S \quad \Rightarrow \quad n \in S$, the assumption $n-1 \in S$ is called the *induction hypothesis*. The process of verifying the two requirements is typically called a *proof by induction* or, if the role of the variable n needs to be stressed, a *proof by induction on* n.

The induction principle is the key to understanding many statement in mathematics, but is also central in computer science, since there it can be used to show correctness of various algorithms, recursive definitions and computer programs.

In mathematics, it is convenient to use a reformulation of the induction principle, avoiding having to work with a subset $S \subseteq \mathbb{N}$. The reason is that this can be avoided using a nice consequence of Theorem 5.3. Such consequences are often called "corollaries" in mathematical texts and we will use the same terminology.

|||| Corollary 5.4

For each natural number n, let $P(n)$ be a logical proposition. Suppose that the following two statements are true:

1. $P(1)$,
2. for all $n \in \mathbb{N}_{\geq 2}$: $\quad P(n-1) \Rightarrow P(n)$.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

Proof. In order to be able to use Theorem 5.3, we use a trick by defining $S = \{n \in \mathbb{N} \mid P(n)\}$. In other words, $n \in S$ by definition precisely if $P(n)$ is true. To be able to conclude that $P(n)$ is true for all natural number n, it is enough to show that $S = \mathbb{N}$. Indeed if there would exist some natural number m such that $P(m)$ is false, then by definition of S, we would have that $m \notin S$ and therefore that $S \neq \mathbb{N}$.

Now we use Theorem 5.3 to show that $S = \mathbb{N}$. The assumption that $P(1)$ is true, just means that $1 \in S$. The assumption that for all $n \in \mathbb{N}_{\geq 2}$: $\quad P(n-1) \Rightarrow P(n)$, means that whenever $n - 1 \in S$, also $n \in S$. Hence the two requirements from Theorem 5.3 are satisfied. Therefore by Theorem 5.3, we may conclude that $S = \mathbb{N}$. This, as remarked already, just means that $P(n)$ is true for all natural numbers n. □

As in Theorem 5.3, checking that $P(1)$ is valid is called the base case of the induction, while checking that for all $n \in \mathbb{N}_{\geq 2}$: $\quad P(n-1) \Rightarrow P(n)$, is called the induction step. While carrying out the induction step, the logical proposition $P(n-1)$ is called the induction hypothesis, similarly as before. Also, the statement in Corollary 5.4 as a whole is still called the induction principle. Hence to prove a claim of the form "$P(n)$ is true for all natural numbers n," we can follow the following strategy:

(i) Inform the reader that you are going to prove the claim that "$P(n)$ is true for all natural numbers n," using induction on n.

(ii) **Base case:** Check that $P(1)$ is valid.

(iii) **Induction step:** For an arbitrary natural number $n \geq 2$, assume that $P(n-1)$ is true and use this assumption (the induction hypothesis) to show that in that case also $P(n)$ is true. The challenge here is sometimes to figure out how to use the induction hypothesis $P(n-1)$ to one's advantage.

(iv) Once the previous items are finished, inform the reader that from the induction principle one can now conclude that $P(n)$ is valid for all natural numbers n.

Now, let use use this strategy to prove our conjecture that $T(n) = 2^n - 1$. In other words, let us prove the following:

Claim: Let $T : \mathbb{N} \to \mathbb{N}$ satisfy the recursion

$$T(n) = \begin{cases} 1 & \text{if } n = 1, \\ 2 \cdot T(n-1) + 1 & \text{if } n \geq 2. \end{cases}$$

Then for all $n \in \mathbb{N}$ we have $T(n) = 2^n - 1$.

Proof. Let $P(n)$ be the statement $T(n) = 2^n - 1$. We will show the claim using induction on n.

Base case: We have $T(1) = 1$. Since $2^1 - 1 = 1$, we see that $T(1) = 2^1 - 1$. Hence $P(1)$ is valid.

Induction step: Let $n \geq 2$ be an arbitrary natural number. The induction hypothesis is $P(n-1)$, which in our case just means the equation $T(n-1) = 2^{n-1} - 1$. Assuming this, we should derive that $P(n)$ is valid. In other words, assuming that $T(n-1) = 2^{n-1} - 1$, we should derive that $T(n) = 2^n - 1$. From the recursive definition of $T(n)$, using that $n \geq 2$, we know that $T(n) = 2 \cdot T(n-1) + 1$. Combining this with the induction hypothesis, we see that

$$T(n) = 2 \cdot T(n-1) + 1 = 2 \cdot (2^{n-1} - 1) + 1 = 2 \cdot 2^{n-1} - 2 \cdot 1 + 1 = 2^n - 1.$$

This is exactly what we needed to show.

Now that we have carried out the base case of the induction as well as the induction step, we can conclude from the induction principle that the statement $T(n) = 2^n - 1$ is true for all natural numbers n. □

One can actually show that the strategy we found in Section 5.2 is the best possible. In other words, any solution of the puzzle with n discs will take at least $T(n)$ moves. We see that solving a ten disc version of the towers of Hanoi, already would take $2^{10} - 1 = 1023$ moves.

The best way to get the hang of proofs by induction is to look at several examples and then to try to do an inductive proof yourself. Let us therefore look at some more examples. Here is a famous one:

|||| Example 5.5

Let us denote by $S(n)$ the sum of the first n natural numbers. Informally, one often writes $1 + 2 + \cdots + n$ for this sum, while we can also use the summation sign and write $S(n) = \sum_{i=1}^{n} i$. As we saw in Table 5-7, we have for example $S(1) = 1$, $S(2) = 1+2 = 3$, $S(3) = 1+2+3 = 6$ and $S(4) = 1+2+3+4 = 10$. The claim is that the following equality holds for all natural number n:

$$S(n) = \frac{n \cdot (n+1)}{2}.$$

Note that $S(n)$ satisfies the following recursion:

$$S(n) = \begin{cases} 1 & \text{if } n = 1, \\ S(n-1) + n & \text{if } n \geq 2. \end{cases}$$

Indeed, we have already observed that $S(1) = 1$, while if $n \geq 2$, using equation (5-5), we obtain that

$$S(n) = 1 + \cdots + n = \sum_{i=1}^{n} i = \left(\sum_{i=1}^{n-1} i \right) + n = S(n-1) + n.$$

Now let us prove the following claim.

Claim: For $n \in \mathbb{N}$, let $S(n) = 1 + \cdots + n$, the sum of the first n natural numbers. Then $S(n) = \frac{n \cdot (n+1)}{2}$.

Proof. We prove the claim using induction on n.

Base case: If $n = 1$, then $S(1) = 1$, while $\frac{1 \cdot (1+1)}{2} = 1$. Hence the formula $S(n) = \frac{n \cdot (n+1)}{2}$ is valid for $n = 1$.

Induction step: Let $n \geq 2$ be an arbitrary natural number and assume as induction hypothesis that $S(n-1) = \frac{(n-1) \cdot (n-1+1)}{2}$. We can simplify the induction hypothesis slightly and say that it holds that $S(n-1) = \frac{(n-1) \cdot n}{2}$. Assuming the induction hypothesis and using that

$S(n) = S(n-1) + n$, we may conclude that

$$\begin{aligned} S(n) &= S(n-1) + n \\ &= \frac{(n-1)\cdot n}{2} + n \\ &= \frac{(n-1)\cdot n}{2} + \frac{2\cdot n}{2} \\ &= \frac{n^2 - n}{2} + \frac{2\cdot n}{2} \\ &= \frac{n^2 - n + 2\cdot n}{2} \\ &= \frac{n^2 + n}{2} \\ &= \frac{n\cdot(n+1)}{2}. \end{aligned}$$

This is exactly what we needed to show, completing the induction step.

Using the induction principle, we may conclude that the formula $S(n) = \frac{n\cdot(n+1)}{2}$ is valid for all natural numbers n. □

|||| Example 5.6

This example is of a more theoretical nature and can be skipped on a first reading. We want to make sure that the recursive definition we gave previously of the factorial function $\texttt{fac} : \mathbb{N} \to \mathbb{N}$ in equation (5-1), actually was correct from a mathematical point of view. The issue is that we never showed that `fac` is defined by its recursive description for *any* natural number n. In other words, when writing $\texttt{fac} : \mathbb{N} \to \mathbb{N}$, we implicitly say that the domain of the function is $\mathbb{N}$, but how do we know? What we need to do is to show that for any natural number n, the recursive description in equation (5-1) will give rise to the output value $\texttt{fac}(n)$ after finitely many steps.

Therefore, let $P(n)$ be the statement that $\texttt{fac}(n)$ can be computed in finitely many steps using equation (5-1) for any natural number n. We want to show that this statement $P(n)$ is true for all natural numbers. The base of the induction is taken care of by the observation that equation (5-1) immediately implies that $\texttt{fac}(1) = 1$. Now let $n \geq 2$ be an arbitrary natural number and assume as induction hypothesis that $\texttt{fac}(n-1)$ can be computed in finitely many steps using equation (5-1). Since $n \geq 2$, equation (5-1) implies that $\texttt{fac}(n) = \texttt{fac}(n-1)\cdot n$. Hence given $\texttt{fac}(n-1)$, all we need is one multiplication with n to compute $\texttt{fac}(n)$. Hence $\texttt{fac}(n)$ can be computed in finitely many steps, if $\texttt{fac}(n-1)$ can. This completes the induction step.

More generally a function $f : \mathbb{N} \to B$, from the natural numbers $\mathbb{N}$ to a given set B, can

be defined recursively as long as $f(1)$ is specified and for any $n \geq 2$, the value $f(n)$ can be computed from $f(n-1)$. The reason is that in such cases, a very similar reasoning as the one we just carried out for the factorial function, applies. In particular, equation (5-2) defines z^n for any natural number n.

5.5 A variant of induction

Many variants of induction exist. In this section, we would like to mention one of them: induction starting with a different base case. So far, the base case of our induction proofs always was the case $n = 1$ and after that we considered larger natural numbers n. In some cases however, a logical statement also makes sense for other values of n. Consider for example the statement:

$$\text{A polynomial } p(Z) \in \mathbb{C}[Z] \text{ of degree } n \text{ has at most } n \text{ roots in } \mathbb{C}.$$

This statement also makes sense for $n = 0$. Indeed, for $n = 0$ the statement is rather easy to verify: a polynomial $p(Z)$ of degree zero, is just a nonzero constant p_0. Indeed, the constant p_0 is nonzero precisely since in general the leading terms of a degree d polynomial is nonzero by Definition 4.1. But then $p(z) = p_0 \neq 0$ for all $z \in \mathbb{C}$, implying that the polynomial has no roots.

Conversely, there are statements that only become true for large enough values of n. Consider for example, the statement:

$$\text{There exist } n \text{ points in the plane } \mathbb{R}^2 \text{ that do not lie on a line.}$$

If $n = 1$, this is wrong, since there are many lines through any given point. Also if $n = 2$, this is wrong, since given any two points, the line connecting them will contain these points. However, for $n \geq 3$, the statement is true. Indeed, if $n \geq 3$, we can for example choose three of the points as the vertices of an equilateral triangle and the remaining $n - 3$ points arbitrarily.

Because of these kind of examples, it is convenient to have a slightly more flexible variant of induction. For a given integer $a \in \mathbb{Z}$, we denote by $\mathbb{Z}_{\geq a} = \{n \in \mathbb{Z} \mid n \geq a\}$. For example $\mathbb{Z}_{\geq -1} = \{-1, 0, 1, 2, \dots\}$. With this notation in place, we can formulate the following variant of induction, called *induction with base case b*:

Theorem 5.7

Let $b \in \mathbb{Z}$ be an integer and for each integer $n \geq b$, let $P(n)$ be a logical proposition. Suppose that the following two statements are true:

1. $P(b)$,
2. for all $n \in \mathbb{Z}_{\geq b+1}$: $\quad P(n-1) \Rightarrow P(n)$.

Then $P(n)$ is true for all $n \in \mathbb{Z}_{\geq b}$.

Proof. Let us define the logical statement $Q(n)$ to be $P(n+b-1)$. Then $Q(n)$ is defined for any natural number n. Indeed if $n \geq 1$, then $n+b-1 \geq b$. Now we apply Corollary 5.4 to the logical statements $Q(n)$. The first requirement from Corollary 5.4 then is that $Q(1)$ should be valid. However, this is fine, since $Q(1) = P(b)$ and it is given that $P(b)$ is valid. The second requirement from Corollary 5.4 becomes that for all $n \in \mathbb{N}_{\geq 2}$: $Q(n-1) \Rightarrow Q(n)$. However, since $n \geq 2$, we have $n+b-1 \geq b+1$ and therefore $n+b-1 \in \mathbb{Z}_{\geq b+1}$. Since $Q(n-1) = P(n+b-2)$ and $Q(n) = P(n+b-1)$ and the implication $P(n+b-2) \Rightarrow P(n+b-1)$ is valid (we know that $n+b-1 \in \mathbb{Z}_{\geq b+1}$), we see that the implication $Q(n-1) \Rightarrow Q(n)$ is valid. Hence the second requirement for the logical statements $Q(n)$ when applying Corollary 5.4 is also met. Hence the corollary implies that $Q(n)$ is valid for all natural numbers n. Since $Q(n) = P(n+b-1)$, this means that $P(n+b-1)$ is valid for all natural number n. In particular $P(1+b-1) = P(b)$ is valid, $P(2+b-1) = P(b+1)$ is valid, and so on. This amount to the statement that $P(n)$ is valid for all integers $n \geq b$, which is what we wanted to show. □

Note that if we choose $b = 1$, we recover Corollary 5.4. The overall structure of a proof with induction with base case b is the same as for the usual induction. One still has a base case and an induction step. Let us consider an example of a proof by induction of this type.

Example 5.8

Consider the inequality $n + 10 \leq n^2 - n$. Since a polynomial of degree two like $n^2 - n$ grows faster than a degree one polynomial like $n + 10$, one should expect that if n becomes large enough this inequality is true. Now let us denote by $P(n)$ the statement that $n + 10 \leq n^2 - n$. In this case, we can define $P(n)$ for any integer n. The statement $P(4)$ for example is the inequality $4 + 10 \leq 4^2 - 4$. This is false since in fact $14 = 4 + 10 > 4^2 - 4 = 12$. On the other

hand, for $n = 5$, the statement $P(5)$ is true, since $15 = 5 + 10 \leq 5^2 - 5 = 20$. We claim that $P(n)$ is true for any $n \in \mathbb{Z}_{\geq 5}$ and give a proof by induction using Theorem 5.7 with $b = 5$:

Base case: We have already verified that $P(5)$ is valid, so the base case is done.

Induction step: Let $n \geq 6$ be an arbitrary natural number and assume as induction hypothesis that $P(n-1)$ is valid. In particular, this means that we may assume that $(n-1) + 10 \leq (n-1)^2 - (n-1)$. Using this assumption, we should deduce that $P(n)$ is valid. Let us first rewrite the induction hypothesis in a more convenient form. We have $(n-1) + 10 = n + 9$, while $(n-1)^2 - (n-1) = n^2 - 2n + 1 - n + 1 = n^2 - 3n + 2$. Hence the induction hypothesis amounts to assuming that the inequality $n + 9 \leq n^2 - 3n + 2$ is valid. But then we can deduce:

$$\begin{aligned} n + 10 &= (n + 9) + 1 \\ &\leq (n^2 - 3n + 2) + 1 \\ &= n^2 - 3n + 3 \\ &= n^2 - n - 2n + 3 \\ &\leq n^2 - n. \end{aligned}$$

The final inequality holds, since $-2n + 3 \leq 0$ for any $n \geq 6$ (in fact even for any $n \geq 2$). We conclude that if $P(n-1)$ is true, then $n + 10 \leq n^2 - n$, that is to say $P(n)$, is true as well. This is what we needed to show, thus completing the induction step.

Using induction with base case 5, we may conclude that the inequality $n + 10 \leq n^2 - n$ is valid for all $n \in \mathbb{Z}_{\geq 5}$.

Index