Acknowledgements

When beginning to write these lecture notes, I got a lot of encouragement and positive reactions from many colleagues at DTU Compute. I would like to thank them for their support and positive energy. As always when writing teaching material, mistakes have a tendence to slip in unintentionally. Fortunately, many people were kind enough to point some of them out. As such, I would like to thank the following people: Ulrik Engelund Pedersen, Steeven Hegelund Spangsdorf, Jakob Lemvig, David Brander, Michael Lund, Kristian Ulldal Kristiansen, Jens Gravesen, Steen Toft and Kim Larsen. If I forgot to mention someone here, sorry about that! Just let me know and I will add your name in a new version. Also a special thanks to Mikael Hjermitslev Hoffmann, who generously volunteered to help making figures for Chapters 6 to 12.

Last but not least, a big shout out to all the students who gave me feedback and constructive comments on older versions of this teaching material. Thank you very much for that. For new students: just keep the feedback and comments coming!

Kgs. Lyngby, July 2024 Peter Beelen

Preface

Welcome! In this book 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. You will become familiar with basic concepts from logic, complex numbers, linear algebra and systems of differential equations.

A very important reason for learning all this is that you all will need mathematics in one way or another later in your studies. Another important 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. Apart from all this, we also hope that you will find that mathematics is beautiful!

Contents

List of Figures

Chapter 1

Propositional logic

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 1.1.1

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, that thinking about it identifies several key ingredients that are useful in general, when thinking about a mathematical problem. One uses words like "and", "or", "not", "if ... then" when attacking problems of this sort. Let

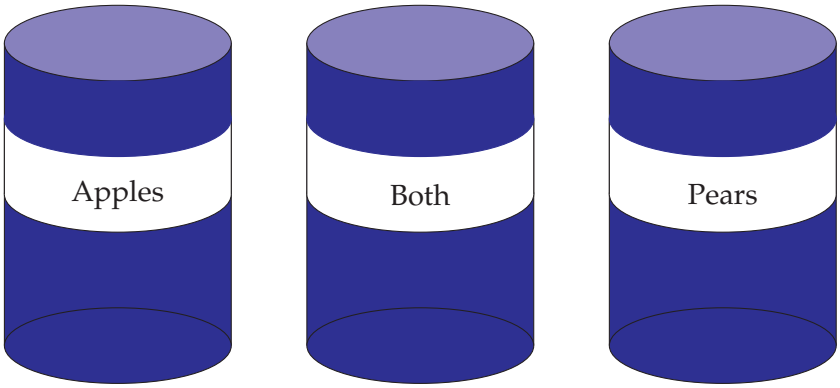


Figure 1.1: A problem with labels and jars

us therefore introduce some notation from what is known as *propositional logic*. First of all, it deals with combinations of short statements that can be either true or false. An example of such a short statement is: the label of jar number one is "Apples". We will call such a statement 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 logical proposition P 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 in this text we will stick to T and F.

Sometimes a proposition can be broken into smaller, simpler ones. For example, the proposition

$$x = 10 \quad \text{and} \quad 1 < y,$$

consists of the two simpler propositions $x = 10$, $1 < y$ combined with the word 'and'. In propositional logic, one writes

$$x = 10 \quad \wedge \quad 1 < y.$$

To improve readability, one can place paranteses around parts of the expression and for example write:

$$(x = 10) \wedge (1 < y).$$

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 1.2.1

Let P and Q be two logical 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
F	T	F
T	F	F
F	F	F

The table in this definition is called a *truth table* for the logical proposition $P \wedge Q$. Let us explain in more detail how such a truth table works. The two variables P and Q can both be true or false independently of each other. In other words: P and Q can take the value T and F independently. Therefore there are in total four cases to consider:

1) P and Q both take the value T,
2) P takes the value F and Q takes the value T,
3) P takes the value T and Q takes the value F,
4) P and Q both take the value F.

For each of these four possibilities the truth table of $P \wedge Q$ specifies which value $P \wedge Q$ takes. For example if P takes the value T and Q takes the value F, we read off from the third row of the truth table that $P \wedge Q$ takes the value F This is why the truth table of $P \wedge Q$ has four rows. Each row specifies which value $P \wedge Q$ takes if P and Q take specific values.

More complicated logical propositions also have a truth table. Here is one example:

Example 1.2.1

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 1.2.1 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
F	T	T
T	F	T
F	F	T
T	T	F
F	T	F
T	F	F
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 $Q \wedge R$. To fill out the values $Q \wedge R$ takes in each of the eight rows, we use Definition 1.2.1. Indeed, even though in Definition 1.2.1 the logical propositions were called P and Q, we can also apply it for the logical proposition Q and R. For example, in the first two rows, both Q and R take the value T, which according to Definition 1.2.1 means that then also $Q \wedge R$ takes the value T. In the third and fourth row Q takes the value F and R the value T. Hence in these rows, $Q \wedge R$ takes the value F. Continuing like this, we then obtain:

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

Next we add a column for $P \wedge (Q \wedge R)$ and determine the truth values it takes for each of the eight rows. Suppose for example that P, Q, R take the values F, T, T. This correspond to the values specified in the second row of the truth table. 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 1.2.1 for the logical propositions 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
F	T	T	T	F
T	F	T	F	F
F	F	T	F	F
T	T	F	F	F
F	T	F	F	F
T	F	F	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 logical proposition $P \wedge Q$. 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 1.1.1, 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 1.2.2

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
F	T	T
T	F	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 1.1.1. 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 1.2.3

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 1.2.2

Consider the logical proposition $P \vee (Q \wedge \neg P)$. We determine its truth table. Having only two variables P and Q, this truth table will contain four rows. Further $P \vee (Q \wedge \neg P)$ contains the simpler logical proposition $Q \wedge \neg P$, which in turn contains the logical 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 $Q \wedge \neg P$ and in this way gradually work our way towards computing the truth values of the whole logical proposition $P \vee (Q \wedge \neg P)$. 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
F	T	T	T	T
T	F	F	F	T
F	F	T	F	F

Let us compare the truth table we just computed and the truth table of $P \vee Q$ from Definition 1.2.2. The comparison shows that for given truth values of P and Q, the truth values of $P \vee (Q \wedge \neg P)$ and $P \wedge Q$ are always the same! In other words: 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 1.2.2. 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 a variety of logical propositions in a precise way. However, the whole point of logic is to make arguments and reasoning more precise. 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 1.3.1

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

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

In common 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 for certain logical propositions P and Q, really means that we claim that $P \Rightarrow Q$ is a tautology. If $P \Rightarrow Q$ is a tautology, then the truth table of the implication from Definition 1.3.1, shows that 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. Let us consider an example of a logical consquence.

Example 1.3.1

Let P and Q be logical propositions. Then we claim that the logical proposition $P \vee Q$ is a logical consequence of P. To show this, we need to verify that the logical proposition $P \Rightarrow (P \vee Q)$ is always true. In other words, we need to show that $P \Rightarrow (P \vee Q)$ is a tautology. In order to compute the truth table of $P \Rightarrow (P \vee Q)$, we proceed in the usual way. First we write down all combinations of truth values of P and Q:

P	Q
T	T
F	T
T	F
F	F

Next, we add a column for $P \vee Q$ for convenience, since it occurs in the more complicated proposition $P \Rightarrow (P \vee Q)$ we are looking at. Using Definition 1.2.2, we then find the following:

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

Now we add a column for $P \Rightarrow (P \vee Q)$ and use Definition 1.3.1 to compute its truth values from the ones of P and $P \vee Q$. The result is the truth table we wanted to compute:

P	Q	$P \vee Q$	$P \Rightarrow (P \vee Q)$
T	T	T	T
F	T	T	T
T	F	T	T
F	F	F	T

Since the column below the expression $P \Rightarrow (P \vee Q)$ only contains T's, we can indeed conclude that $P \Rightarrow (P \vee Q)$ is a tautology. In particular, we can now be sure that P implies $P \vee Q$. Another way of saying this is that $P \vee Q$ is a logical consequence of P.

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

Definition 1.3.2

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
F	T	F
T	F	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 1.3.4, Equation (1.22) in a more formal way that a bi-implication can indeed in this way be expressed as two implications.

Example 1.3.2

In Example 1.2.2 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 1.2.2 and Example 1.2.2, we see that the following table is correct:

P	Q	$P \vee Q$	$P \vee (Q \wedge \neg P)$
T	T	T	T
F	T	T	T
T	F	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 1.3.2. 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
F	T	T	T	T
T	F	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 1.3.2, 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 1.3.1

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

$$P \wedge P \Leftrightarrow P \tag{1.1}$$
$$P \vee P \Leftrightarrow P \tag{1.2}$$
$$P \wedge Q \Leftrightarrow Q \wedge P \tag{1.3}$$
$$P \vee Q \Leftrightarrow Q \vee P \tag{1.4}$$
$$P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R \tag{1.5}$$
$$P \vee (Q \vee R) \Leftrightarrow (P \vee Q) \vee R \tag{1.6}$$
$$P \wedge (Q \vee R) \Leftrightarrow (P \wedge Q) \vee (P \wedge R) \tag{1.7}$$
$$P \vee (Q \wedge R) \Leftrightarrow (P \vee Q) \wedge (P \vee R) \tag{1.8}$$

Proof. To prove that one of the mentioned logical propositions 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 Equation (1.5). We need to show that $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ is a tautology. In Example 1.2.1, 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 1.2.1, and then in the last step compute the truth table of $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ using Definition 1.3.2. 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
F	T	T	F	F	F	T
T	F	T	F	F	F	T
F	F	T	F	F	F	T
T	T	F	F	T	F	T
F	T	F	F	F	F	T
T	F	F	F	F	F	T
F	F	F	F	F	F	T

We see that the logical proposition $P \wedge (Q \wedge R) \Leftrightarrow (P \wedge Q) \wedge R$ only takes the truth 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. □

In words, Equation (1.6) states 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 (1.5) 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 1.3.3

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	...	$(P \wedge Q) \vee R$	$P \wedge (Q \vee R)$
⋮					
F	T	T	...	T	F
⋮					

This is in fact enough to conclude that 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 based on the previous, we see that its truth table actually contains the following row:

P	Q	R	...	$(P \wedge Q) \vee R$	$P \wedge (Q \vee R)$	$(P \wedge Q) \vee R \Leftrightarrow P \wedge (Q \vee R)$
⋮						
F	T	T	...	T	F	F
⋮						

This shows that $(P \wedge Q) \vee R \Leftrightarrow P \wedge (Q \vee R)$ is not a tautology and therefore that the logical propositions $(P \wedge Q) \vee R$ and $P \wedge (Q \vee R)$ indeed are not logically equivalent.

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 1.3.2

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 1.3.3

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 1.3.1, 1.3.2 and 1.3.3 one can rewrite logical proposition in a logically equivalent form. Let us consider an example.

Example 1.3.4

As in Examples 1.2.2 and 1.3.2, 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 1.3.1 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 1.3.1, 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 1.3.1 only involve negation, conjunction and disjunction. Here are three very useful ones that involve implication and bi-implication as well.

Theorem 1.3.4

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 1.3.1, 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
F	T	T	T	F	T	T
T	F	F	F	T	F	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 $\neg Q$. Let us consider a small example of contraposition.

Example 1.3.5

Consider the statement that for any real numbers x and y it holds that

$$(x \cdot y = 0) \Rightarrow ((x = 0) \vee (y = 0)).$$

This is a true statement, but in this example we do not want to prove it, but simply to figure out what the contraposition of this statement is.

First of all, the given statement is a logical proposition of the form $P \Rightarrow Q$, where P is the equation $x \cdot y = 0$ and Q the proposition $(x = 0) \vee (y = 0)$. In words, the implication $P \Rightarrow Q$ can be phrased as: if for some real numbers x and y, the equation $x \cdot y = 0$ holds, then $x = 0$ or $y = 0$.

What is the contraposition of this? According to Equation (1.21) it is $\neg Q \Rightarrow \neg P$. When used directly, we therefore find that the contraposition we are looking for, is

$$\neg((x = 0) \vee (y = 0)) \Rightarrow \neg(x \cdot y = 0).$$

However, we can simplify this a bit. First of all, one can rewrite $\neg(x \cdot y = 0)$ as $x \cdot y \neq 0$. Moreover, using Equation (1.12), one of the DeMorgan laws, we can rewrite $\neg((x = 0) \vee (y = 0))$ as $\neg(x = 0) \wedge \neg(y = 0)$, which in turn can be written as $(x \neq 0) \wedge (y \neq 0)$. Therefore the contraposition of

$$(x \cdot y = 0) \Rightarrow (x = 0) \vee (y = 0)$$

can be given as

$$((x \neq 0) \wedge (y \neq 0)) \Rightarrow (x \cdot y \neq 0).$$

More in words: the contraposition of the statement "if $x \cdot y = 0$, then $x = 0$ or $y = 0$" simply is "if $x \neq 0$ and $y \neq 0$, then $x \cdot y \neq 0$".

This last statement is a true statement, since it is logically equivalent to the true statement that we started with in this example.

Equation (1.22) states that two logical propositions are logically equivalent precisely if they are logical consequences of each other. Quite often it is easier to show that $P \Rightarrow Q$ and $Q \Rightarrow P$ are true separately, then to show directly that $P \Leftrightarrow Q$ is true. Also 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.

In later chapters, we will regularly use Equations (1.21), (1.22) and (1.23), when investigating various mathematical statements. In the next section, we will also show uses of logic in mathematics.

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 1.4.1

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 1.4.2

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:

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

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

Example 1.4.3

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 1.5.1

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 true statement. Similarly for the right jar, we obtain that $P_3(A) \vee P_3(B)$ is a true statement. 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 always true. Using Equation (1.7) 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 valid, since it is logically equivalent to the statement 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 necessarily always true.

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 always true. 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 one calls 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 1.5.2

A police officer is investigating a burglary and was able to narrow the number of suspects down to three. The officer is absolutely sure that one of these three committed the crime and that the perpetrator worked alone. When questioning the three suspects, the following statements are made by the suspects:

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 the 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 true, 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: 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 true statement. Using Equation (1.11), we conclude that

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

is a true statement as well.

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 true statement. One can simplify this expression using Theorem 1.3.1. 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 statement, we see that $P_3 \vee (P_2 \vee R_1 \vee \neg P_1) \vee P_2$ is a true statement. 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 true statement as well.

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 true statement. Similarly, the third statement from Suspect3, "Suspect2 is lying if he said that everything that Suspect1 said is a lie", is true. 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";	
"Suspect2 is lying if he said that everything that Suspect1 said is a lie;	

Now let us for the third time consider the insight from the police commissioner. First of all, the given insight implies that the second are third statements from Suspect3 are true, since we know that Suspect1 and Suspect2 are lying. Since by the same insight, Suspect3 lied, we conclude that $\neg P_3$ must be a lie. In other words, P_3 must be true.

Collecting everything together, we have determined that the following are all true: $P_1 \vee P_2 \vee P_3$, $\neg P_2 \vee \neg R_1 \vee P_1$, $P_3 \vee P_2 \vee R_1 \vee \neg P_1$, P_3. The fact that P_3 is true, 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 statement we obtained 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 true by the definition of the disjunction. 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 true statement. This means that there is nothing contradictory. The police should arrest Suspect3!

Chapter 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 2.1.1

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 $B \not\subseteq A$ or alternatively $A \not\supseteq B$.

Example 2.1.2

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 2.1.3

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 2.1.4

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 2.1.5

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 2.1.1

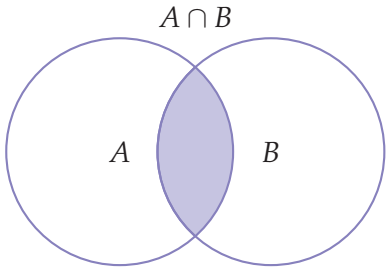
Let A and B be two sets. Then $A = B$ if and only if $A \subseteq B$ and $A \supseteq B$.

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 $A \subseteq B \quad \wedge \quad A \supseteq B$. □

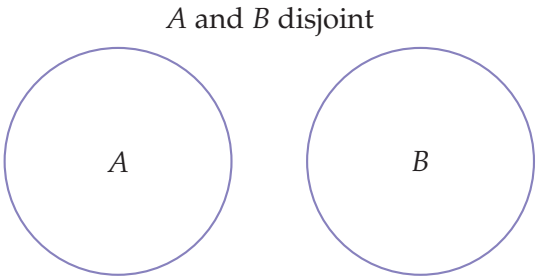
Instead of $\subseteq$ and $\supseteq$, some authors prefer the symbols $\subset$ and $\supset$. However, yet other authors, use the symbols $\subset$ and $\supset$ in the meaning of $\subsetneq$ and $\supsetneq$, inspired by the use of $<$ and $>$ in the setting of strict inequalities. To avoid confusion, 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 2.1.6. 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:

$$A \cap B = \{a \mid a \in A \wedge a \in B\}. \tag{2.1}$$

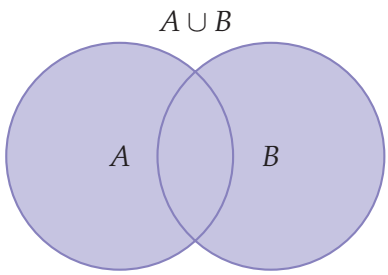


Two sets A and B are called *disjoint*, if $A \cap B = \emptyset$.



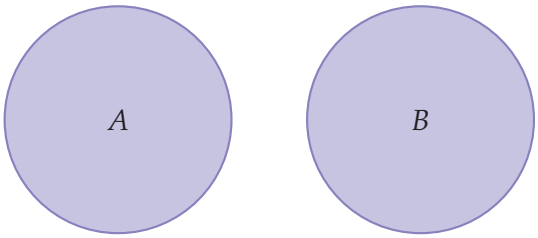
The *union* of A and B is defined as:

$$A \cup B = \{a \mid a \in A \vee a \in B\}. \tag{2.2}$$



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

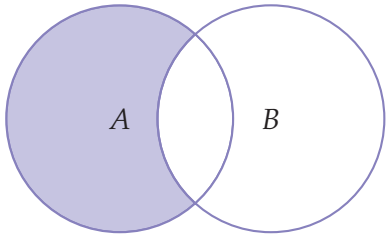
$A \cup B$ disjoint union of A and B



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

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

$A \setminus B$



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 2.1.6

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

(a) $\{1,2\} \subseteq \{1,2,3\}$ and in fact $\{1,2\} \subsetneq \{1,2,3\}$,

(b) $\{1,2\} \supseteq \{2\}$ and in fact $\{1,2\} \supsetneq \{2\}$,

(c) $\{1,4\} \not\subseteq \{1,2,3\}$,

(d) $\{1,2,3\} \cap \{2,3,4\} = \{2,3\}$,

(e) $\{1,2\}$ and $\{3\}$ are disjoint sets,

(f) $\{1,2,3\} \cup \{2,3,4\} = \{1,2,3,4\}$,

(g) $\{1,2,3,4\}$ is the disjoint union of $\{1,2\}$ and $\{3,4\}$,

(h) $\{1,2,3\} \setminus \{2,3,4\} = \{1\}$,

(i) $\{2,3,4\} \setminus \{1,2,3,4\} = \emptyset$,

(j) $\{1,2\} \times \{3,4\} = \{(1,3),(1,4),(2,3),(2,4)\}$,

(k) $\{1,2\}^2 = \{(1,1),(1,2),(2,1),(2,2)\}$.

In Equations (2.1) and (2.2), the logical opeators $\wedge$ and $\vee$ came in very handy. In Theorem 1.3.1 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 2.1.2

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:

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

□

Theorem 2.1.2 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 1.3.2 and 1.3.3 will come in handy.

Example 2.1.7

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

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

On the other hand

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

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 2.2.1 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 2.2.1

Let us again consider the function

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

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 2.1.1, 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 2.1.1, 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 2.2.2

Consider the functions

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

and

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

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

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

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 2.2.3

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:

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

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

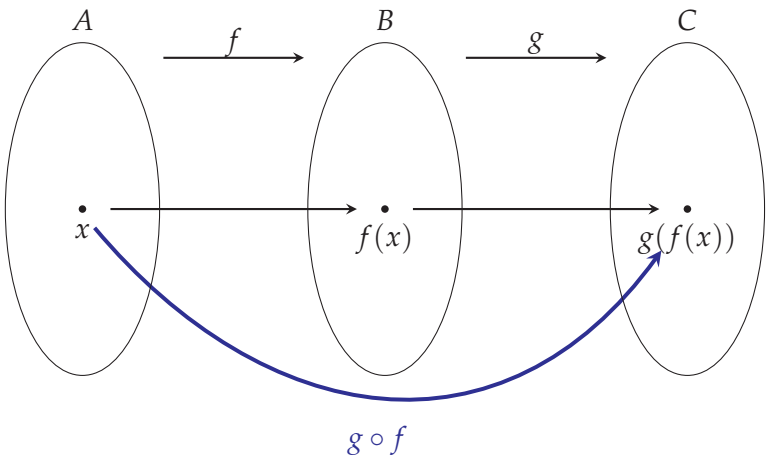


Figure 2.1: Composition of the functions $f : A \to B$ and $g : B \to C$

Lemma 2.2.1

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 2.2.1, 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 logical expressions, this means that:

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

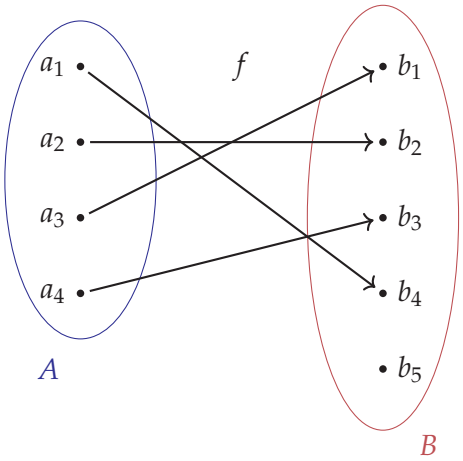


Figure 2.2: Injective function $f : A \to B$

Using (1.21), it is logically equivalent to write:

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

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:

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

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 2.2.4

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.

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.

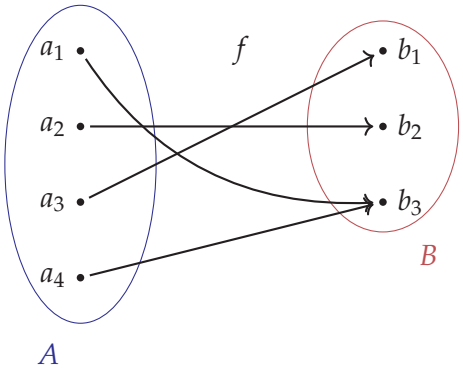


Figure 2.3: Surjective function $f : A \to B$

Example 2.2.5

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 2.2.1

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 2.2.2

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 2.2.1 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 2.2.6

Let us again consider the function $h : \{0,1,2\} \to \{3,4,5\}$ given by $h(x) = 5 - x$ from Example 2.2.5. We have seen that the function h is bijective. Hence by Lemma 2.2.2, 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.

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 2.2.7

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 1.4.2, 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: $x \in \mathbb{R}$
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 2.3.1

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 2.3.1 can be summarized as: a strictly monotone function is injective.

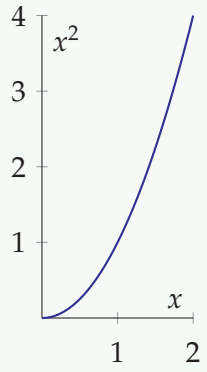
Example 2.3.1

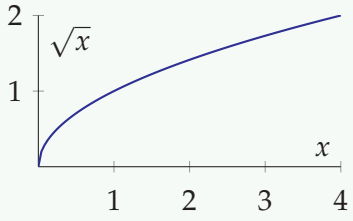
Consider the function $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = x^2$. We have already seen in Example 2.2.1 that the image of this function equals $\mathbb{R}_{\geq 0}$. In other words, $f(\mathbb{R}) = \mathbb{R}_{\geq 0}$. The function f is therefore not surjective. In fact, it is not injective 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 2.3.1 injective. We already have seen that h is surjective, so we may conclude that it is bijective. By Lemma 2.2.2, 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.

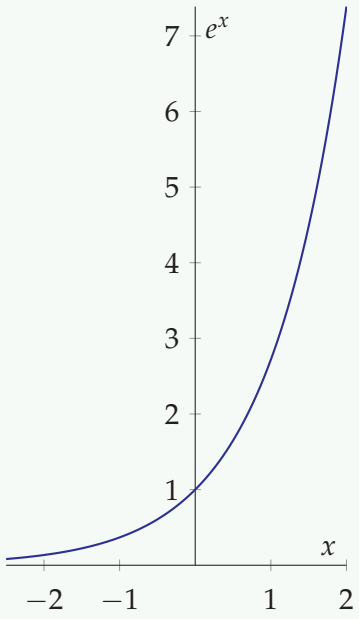


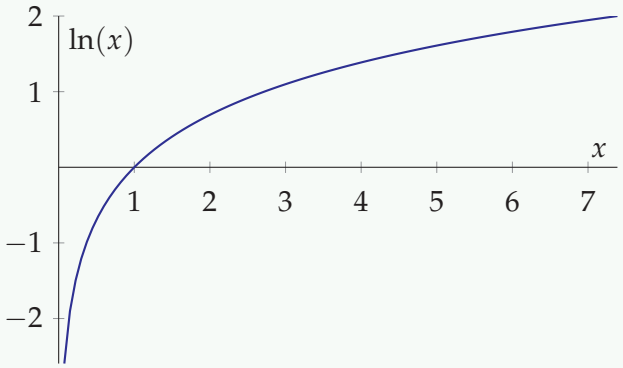


Example 2.3.2

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.

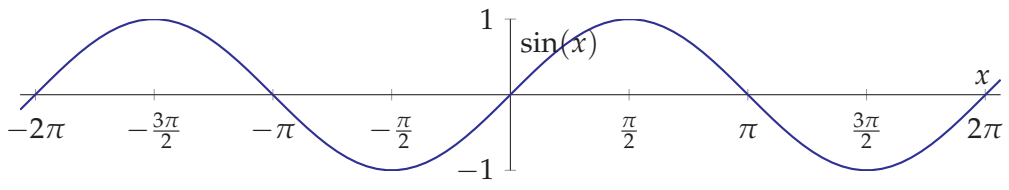




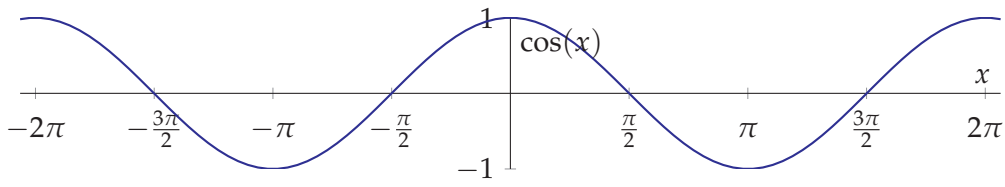
The trigonometric functions sin, cos and tan.

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.

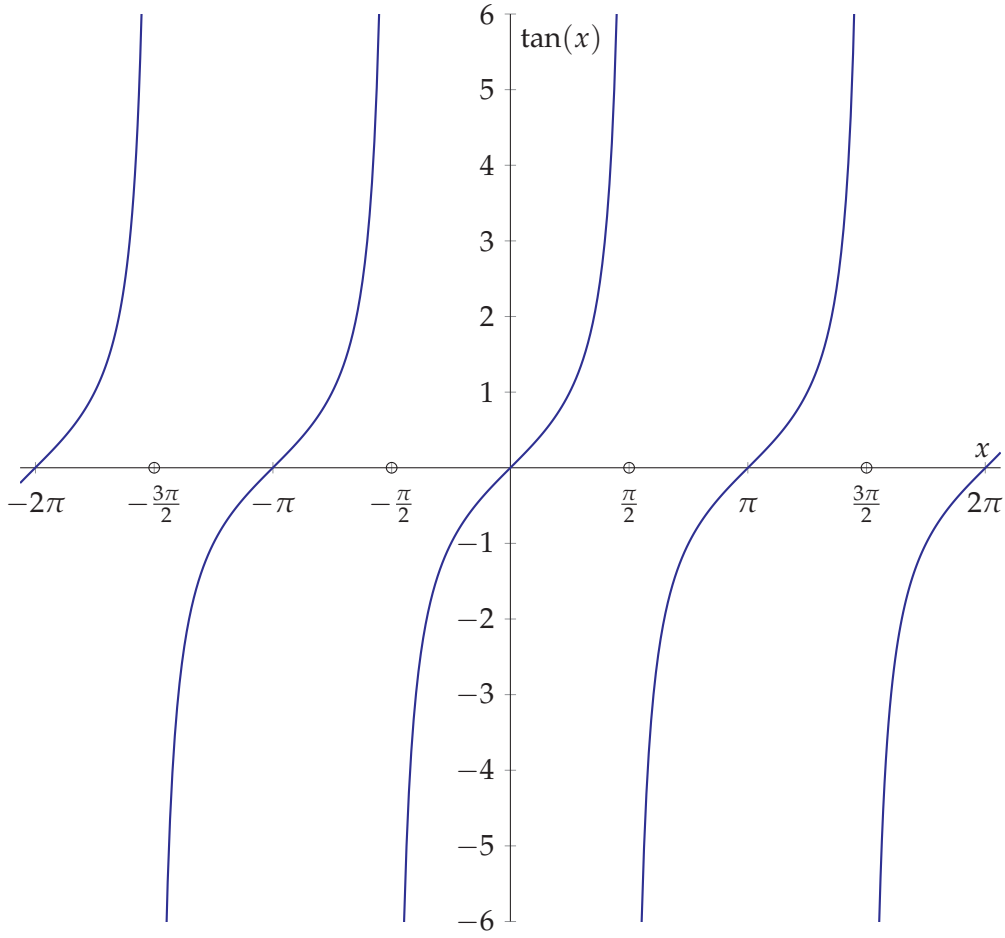
First of all, the sine function is usually denoted by sin, but let us in light of our definition of functions specify which domain and co-domain it has. First of all, we define the sine function $\sin : \mathbb{R} \to [-1,1]$ to be the function such that $x \mapsto \sin(x)$. The image of sin is $[-1,1]$, 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 $\{x \in \mathbb{R} \mid \cos(x) = 0\} = \{\ldots, -3\pi/2, -\pi/2, \pi/2, 3\pi/2, \ldots\}$, we can also say that the domain of the tangent function is the set $\mathbb{R} \setminus \{\ldots, -3\pi/2, -\pi/2, \pi/2, 3\pi/2, \ldots\}$. The graph of the tangent function is as follows:



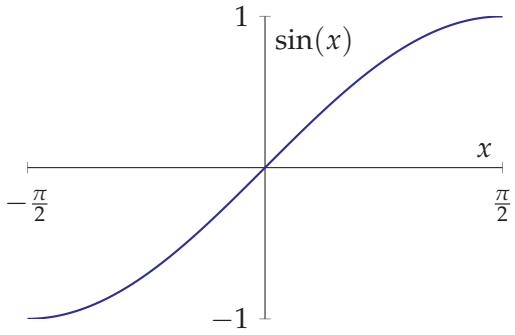
The small circles on the x-axis indicate the values of x for which the tangent function is not defined. The tangent function is surjective, since its image is $\mathbb{R}$. Just as the sine and cosine

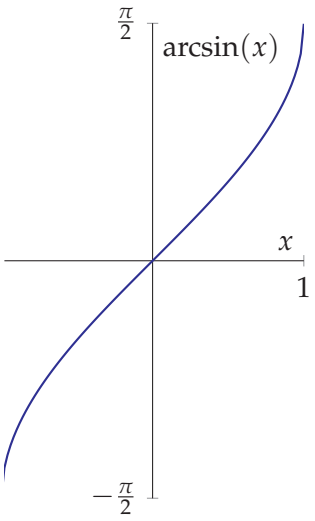
functions, it is not injective. We have for example $\tan(0) = 0$, but also $\tan(\pi) = 0$.

The inverse trigonometric functions

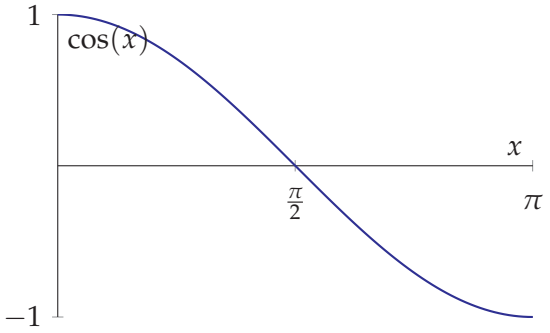
Since none of the trigonometric functions sin, cos and tan discussed in the previous subsection are bijections, we cannot find inverses for these functions. However, just as in Example 2.3.1, we can modify the domain of these functions and obtain functions that do have an inverse. These inverses are known as the *inverse trigonometric functions* (sometimes also as the arcus functions). In this subsection, we give the details of how these are defined.

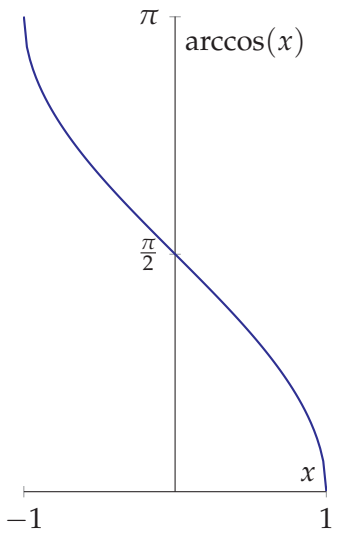
First of all, if the domain of the function $\sin : \mathbb{R} \to [-1, 1]$ is restricted to the closed interval $[-\pi/2, \pi/2]$, one obtains a function $f : [-\pi/2, \pi/2] \to [-1, 1]$ defined by $f(x) = \sin(x)$. The function f is a bijective function, since the graph of the sine function is strictly 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:



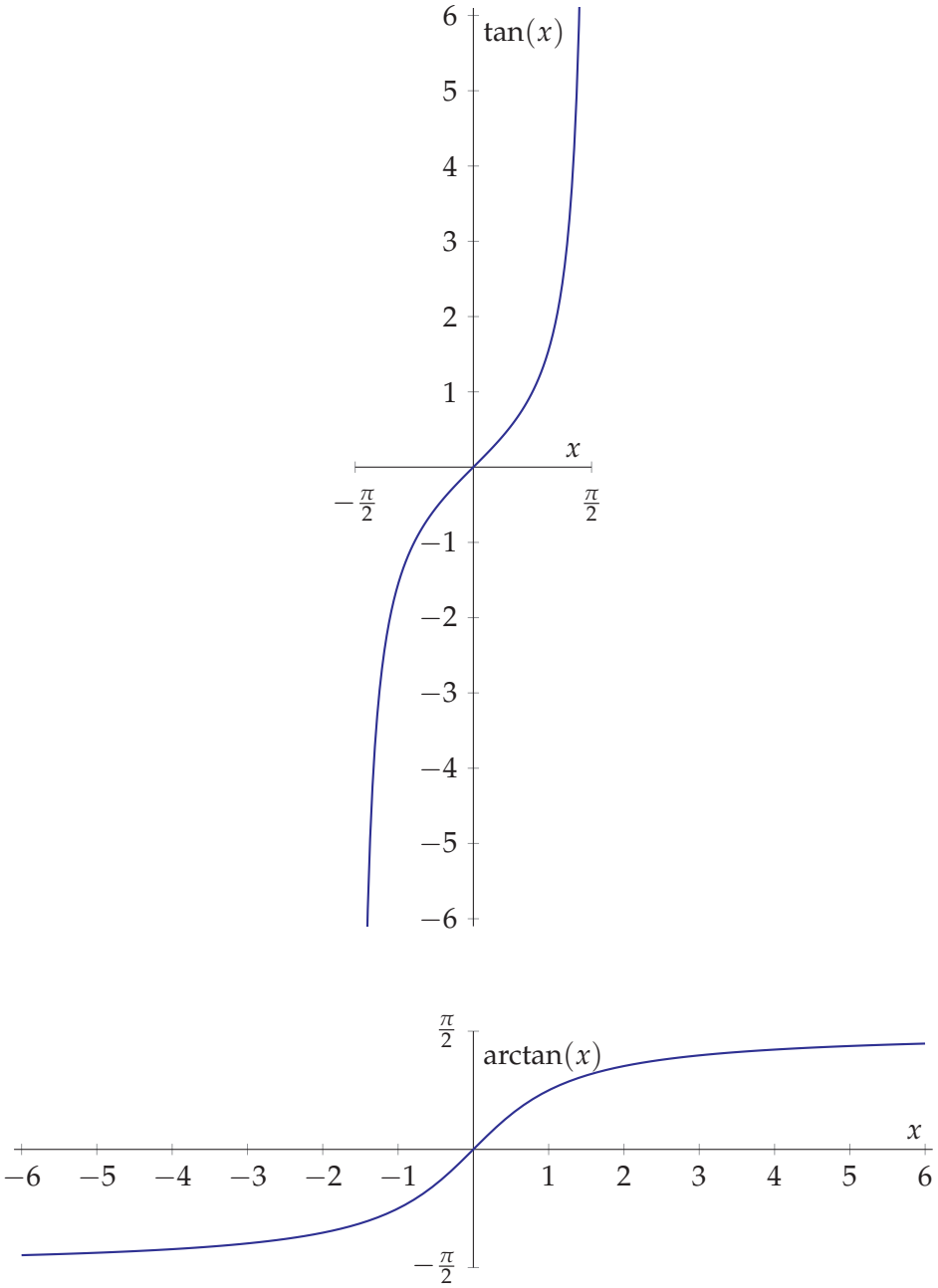


In a very similar way, we can define the *arccosine* function. First we restrict the domain of the usual cosine function to the closed interval $[0, \pi]$. The resulting function $g : [0, \pi] \to [-1, 1]$, where $g(x) = \cos(x)$, is strictly decreasing as well as surjective and thus bijective. The inverse 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 2.3.3

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 $\arctan(\sqrt{3}) = \pi/3$.

Chapter 3

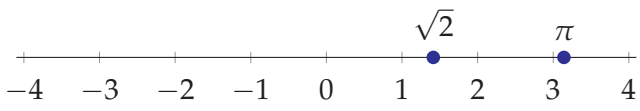
Complex numbers

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 2.1.4, 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).

Figure 3.1: The real line.



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 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 3.1.1

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 $i \notin \mathbb{R}$.

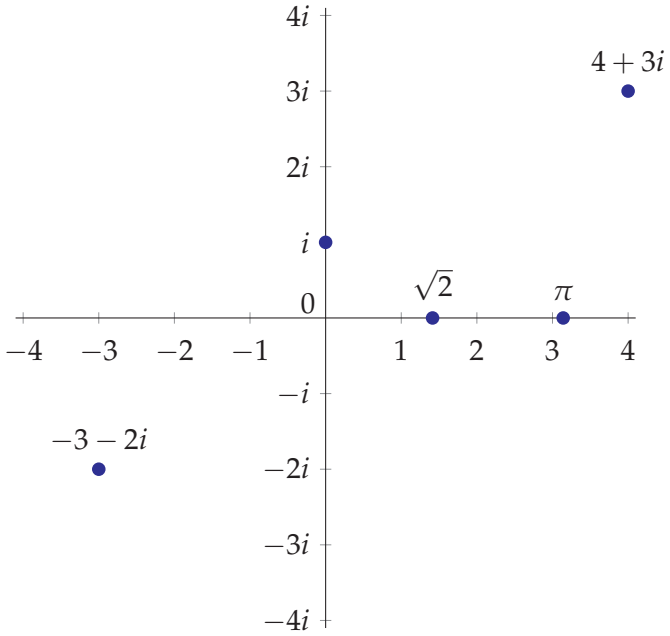
The complex numbers can represented graphically, but now as a plane called the *complex plane*. A complex number $a + bi$ is represented as the point (a, b) in that plane. This means that the number i has coordinates $(0, 1)$ and therefore will lie on the second axis. The number i and some other complex 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 3.1.1

Compute the rectangular coordinates of the following complex numbers:

(a) $2 + 3i$

(b) $\sqrt{2}$

(c) i

Answer:

(a) 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 $(2,3)$.

(b) 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 $(\sqrt{2},0)$.

(c) 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 3.2.1

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

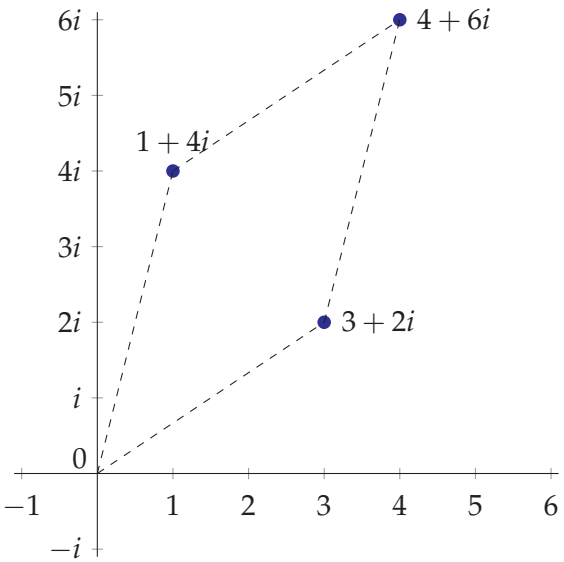
$$(a+bi)-(c+di) = (a-c)+(b-d)i.$$

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 4.1.1). 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.

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



Example 3.2.1

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

(a) $(3+2i)+(1+4i)$

(b) $(3+2i)-(1+4i)$

(c) $(5-7i)-i$

(d) $(5-7i)-(-10+i)$

Answer:

(a) $(3+2i)+(1+4i)=(3+1)+(2+4)i=4+6i$

(b) $(3+2i)-(1+4i)=(3-1)+(2-4)i=2-2i$

(c) $(5-7i)-i=5+(-7-1)i=5-8i$

(d) $(5-7i)-(-10+i)=(5-(-10))+(-7-1)i=15-8i$

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

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

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 3.2.2

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 3.2.2

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

(a) $(1+2i)\cdot(3+4i)$

(b) $(4+i)\cdot(4-i)$

Answer:

(a)

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

(b)

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

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 3.2.1

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 3.2.3

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 3.2.1 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 $z_2 \in \mathbb{C}\backslash\{0\}$, 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 3.2.3

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

(a) $1/(1+i)$

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

Answer:

(a) 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):

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

(b) 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 3.2.2

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:

(i) Addition and multiplication are *associative*: $z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3$, and $z_1 \cdot (z_2 \cdot z_3) = (z_1 \cdot z_2) \cdot z_3$.

(ii) Addition and multiplication are *commutative*: $z_1 + z_2 = z_2 + z_1$, and $z_1 \cdot z_2 = z_2 \cdot z_1$.

(iii) *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 3.2.3

(i) 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 $z \in \mathbb{C}$.

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

(iii) 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 3.2.3 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 3.2.4

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 $\mathrm{Arg}(z)$.

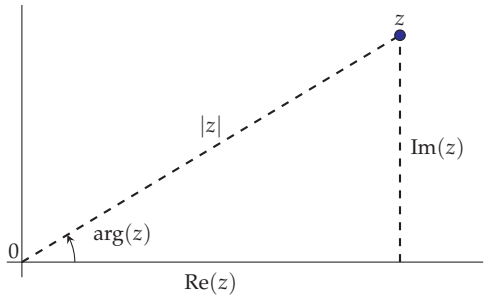


Figure 3.4: Modulus and argument of a complex number z.

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, we have the following:

Theorem 3.3.1

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 2.2.7. 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).

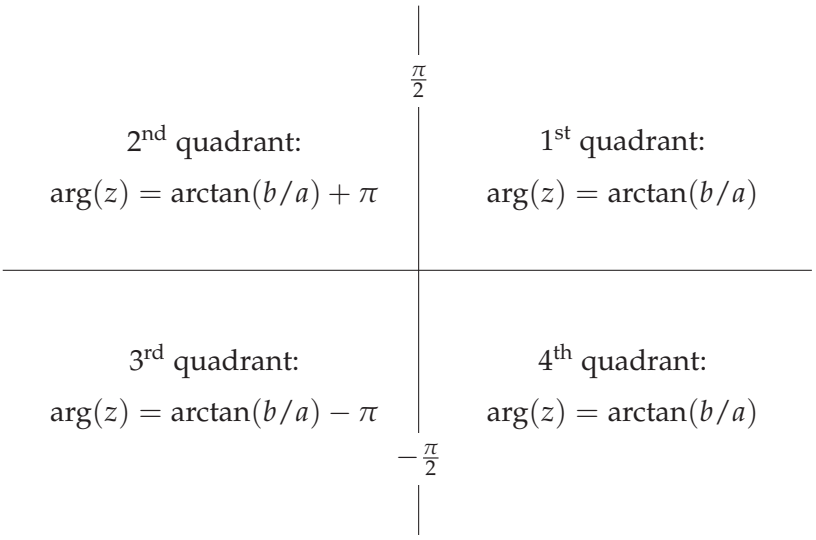


Figure 3.5: Formulas for the argument of $z = a + bi$.

Example 3.3.1

Compute the polar coordinates of the following complex numbers:

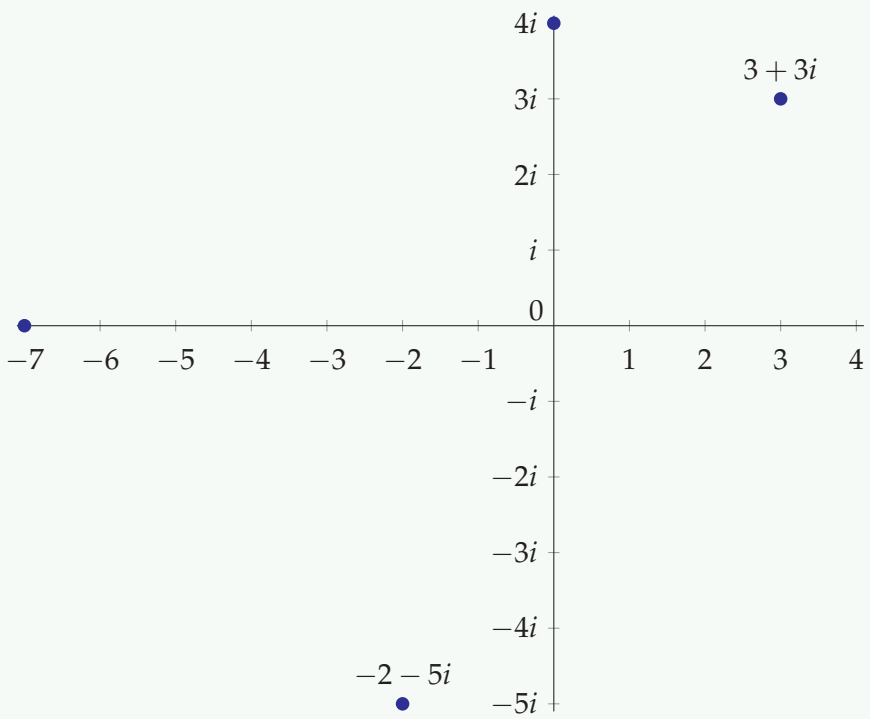
(a) $4i$

(b) -7

(c) $3 + 3i$

(d) $-2 - 5i$

Answer: We can find the modulus and argument using Theorem 3.3.1. Figure 3.5 is useful when computing the argument. Therefore, we first plot the four given complex numbers in the complex plane.



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

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

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

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

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 3.3.2

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

(a) $(2, \pi/3)$

(b) $(10, \pi)$

(c) $(4, -\pi/4)$

(d) $(2\sqrt{3}, -2\pi/3)$

(e) $(3, 2)$

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.

(a) $z = 2 \cdot (\cos(\pi/3) + \sin(\pi/3)i) = 2 \cdot (1/2 + \sqrt{3}/2i) = 1 + \sqrt{3}i$.

(b) $z = 10 \cdot (\cos(\pi) + \sin(\pi)i) = -10 + 0i = -10$.

(c) $z = 4 \cdot (\cos(-\pi/4) + \sin(-\pi/4)i) = 4 \cdot (\sqrt{2}/2 - \sqrt{2}/2i) = 2\sqrt{2} - 2\sqrt{2}i$.

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

(e) $z = 3 \cdot (\cos(2) + \sin(2)i) = 3\cos(2) + 3\sin(2)i$.

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 3.4.1

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 3.4.1

Write the following expressions in rectangular form:

(a) e^2

(b) e^{1+i}

(c) $e^{\pi i}$

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

(e) $e^{2\pi i}$

Answer: We use Definition 3.4.1 and simplify till we find the desired rectangular form.

(a) Since e^2 is a real number, it is already in rectangular form. If we use Definition 3.4.1 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.

(b) $e^{1+i} = e^1 \cdot (\cos(1) + \sin(1)i) = e\cos(1) + e\sin(1)i$.

(c) $e^{\pi i} = e^{0+\pi i} = e^0 \cdot (\cos(\pi) + \sin(\pi)i) = 1 \cdot (-1 + 0i) = -1$.

(d) $e^{\ln(2)+i\pi/4} = e^{\ln(2)} \cdot (\cos(\pi/4) + \sin(\pi/4)i) = 2(\sqrt{2}/2 + \sqrt{2}/2i) = \sqrt{2} + \sqrt{2}i$.

(e) $e^{2\pi i} = \cos(2\pi) + \sin(2\pi)i = 1 + 0i = 1$. Note that also $e^0 = 1$. This shows that the complex exponential function is not injective.

Directly from Definition 3.4.1, 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 3.4.1

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 3.4.2

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

$$\begin{aligned} e^{z_1} \cdot e^{z_2} &= e^{a_1} \cdot (\cos(b_1) + \sin(b_1)i) \cdot e^{a_2} \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1} \cdot e^{a_2} \cdot (\cos(b_1) + \sin(b_1)i) \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1+a_2} \cdot (\cos(b_1) + \sin(b_1)i) \cdot (\cos(b_2) + \sin(b_2)i) \\ &= e^{a_1+a_2} \cdot (\cos(b_1 + b_2) + \sin(b_1 + b_2)i) \text{ (using Lemma 3.4.1)} \\ &= e^{a_1+a_2+(b_1+b_2)i} = e^{z_1+z_2}. \end{aligned}$$

□

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 3.4.1. 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 3.5.1

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.

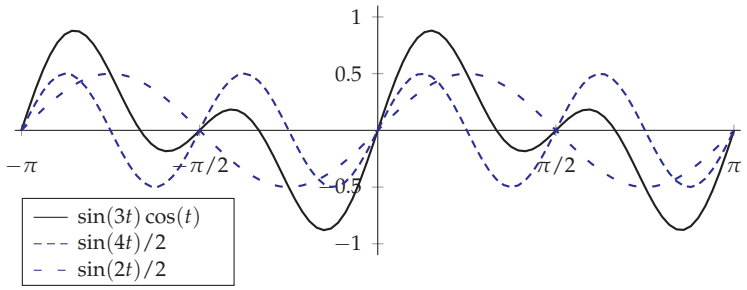


Figure 3.6: Illustration of the identity $\sin(3t)\cos(t) = \frac{\sin(4t)}{2}+\frac{\sin(2t)}{2}$.

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

Theorem 3.5.1

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 3.5.2

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 3.5.3

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 3.4.1, 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 3.6.1

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.

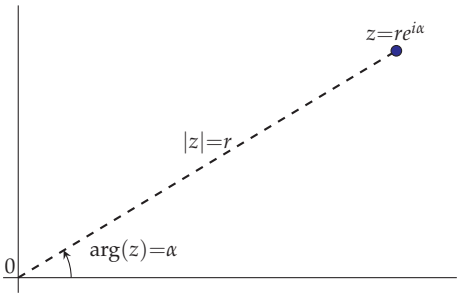


Figure 3.7: Polar form of a complex number z.

Example 3.6.1

Write the following complex numbers in polar form:

(a) $-1+i$

(b) $2+5i$

(c) e^{7+3i}

(d) $e^{7+3i}/(-1+i)$

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.

(a) $|-1+i| = \sqrt{1+1} = \sqrt{2}$ and $\arg(-1+i) = \arctan(1/-1) + \pi = 3\pi/4$. In polar form the number is therefore given by $\sqrt{2}e^{i3\pi/4}$.

(b) $|2+5i| = \sqrt{4+25} = \sqrt{29}$ and $\arg(2+5i) = \arctan(5/2)$. We therefore find that $2+5i$ has the following polar form: $\sqrt{29}e^{i \arctan(5/2)}$.

(c) $e^{7+3i} = e^7 e^{3i}$. 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.

(d) 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 3.4.1, 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 3.6.1

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 3.4.2 to prove the following theorem.

Theorem 3.6.2

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 3.4.2 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 3.6.2 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).

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 3.6.2

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

(a) $(1+i)^{13}$.

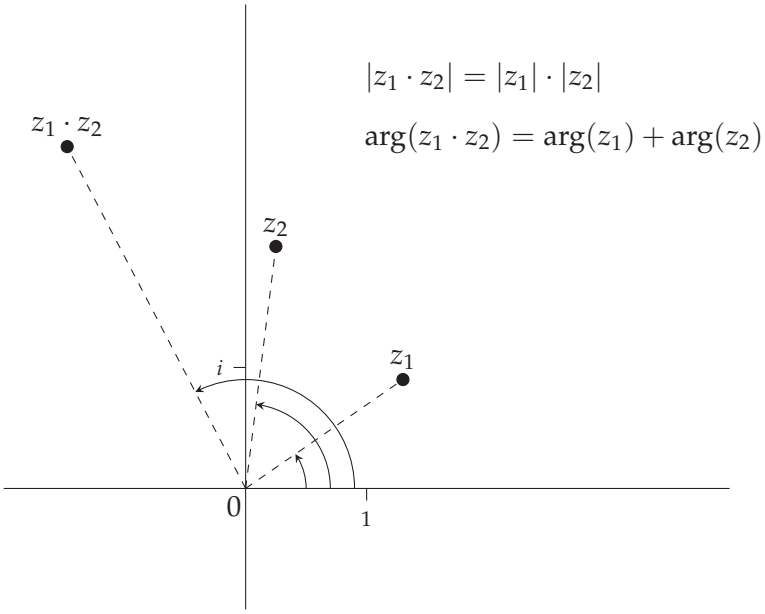


Figure 3.8: Graphic illustration of Theorem 3.6.2.

(b) $(-1-\sqrt{3}i)^{15}$.

Answer:

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

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

(b) First we calculate modulus and argument $-1-\sqrt{3}i$. According to Theorem 3.3.1 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

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

Chapter 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 4.1.1

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 4.1.1

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.

(a) $1 + Z^2$

(b) $Z^{-1} + 1 + Z^3$

(c) i

(d) $\sin(Z) + Z^{12}$

(e) $1 + 2Z + 5Z^{10} + 0Z^{11}$

(f) $1 + Z + Z^{2.5}$

(g) $(1 + Z)^2$

Answer:

(a) $1 + Z^2$ 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 4.1.1, 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.

(b) $Z^{-1} + 1 + Z^3$ 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.

(c) The complex number i can be interpreted as a polynomial in $\mathbb{C}[Z]$. One chooses $n = 0$ and $a_0 = i$ in Definition 4.1.1. The polynomial i has therefore degree 0 and leading coefficient i.

(d) $\sin(Z) + Z^{12}$ is not a polynomial because of the term $\sin(Z)$.

(e) $1 + 2Z + 5Z^{10} + 0Z^{11}$ is a polynomial in $\mathbb{C}[Z]$. The term $0Z^{11}$ 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.

(f) $1 + Z + Z^{2.5}$ is not a polynomial, because of the term $Z^{2.5}$. The exponents of Z must be natural numbers.

(g) $(2+Z)^2$ is a polynomial in $\mathbb{C}[Z]$, though it is not written in the form as in Definition 4.1.1. 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 4.1.2

Write the following polynomials in the form as in Definition 4.1.1.

(a) $(Z+5) \cdot (Z+6)$.

(b) $(3Z+2) \cdot (3Z-2)$.

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

Answer:

(a) $(Z+5) \cdot (Z+6) = Z \cdot (Z+6) + 5 \cdot (Z+6) = Z^2+6Z+5Z+30 = Z^2+11Z+30$.

(b) $(3Z+2) \cdot (3Z-2) = (3z)^2 - 6Z + 6Z - 2^2 = 9Z^2 - 4$.

(c) 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 4.1.2

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 of degree two with real coefficients

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 4.2.1

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 4.4.1, 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 4.2.1

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

(i) two real roots $z=\dfrac{-b\pm\sqrt{D}}{2a}$ if $D>0$,

(ii) one real root $z=\dfrac{-b}{2a}$ if $D=0$,

(iii) two non-real roots $z=\dfrac{-b\pm i\sqrt{|D|}}{2a}$ if $D<0$.

The description of the roots in Theorem 4.2.1 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 $\deg(p(Z))=2$
1: $a\leftarrow$ coefficient of Z^2 in $p(Z)$
2: $b\leftarrow$ coefficient of Z^1 in $p(Z)$
3: $c\leftarrow$ coefficient of Z^0 in $p(Z)$
4: $D\leftarrow b^2-4ac$
5: **if** $D\geq 0$ **then**
6: **return** $\dfrac{-b+\sqrt{D}}{2a}$ and $\dfrac{-b-\sqrt{D}}{2a}$
7: **else**
8: **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.

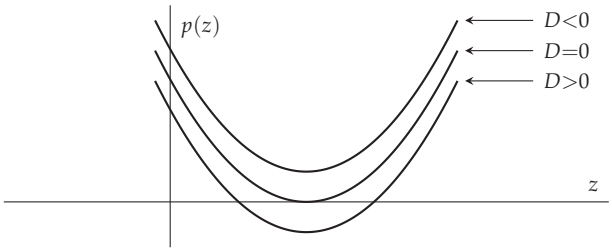


Figure 4.1: A degree two polynomial $p(Z) \in \mathbb{R}[Z]$ has two real roots if $D > 0$, a double root if $D = 0$, and two complex, two non-real roots if $D < 0$.

Example 4.2.1

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 4.2.1 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 4.2.1 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 3.2.3, will play an important role. Complex conjugation has several nice properties. We list some of these in the following lemma.

Lemma 4.3.1

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

(i) $\overline{\overline{z}} = z$,

(ii) $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$,

(iii) $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$,

(iv) $\overline{1/z} = 1/\overline{z}$ provided $z \neq 0$,

(v) $\overline{z^n} = (\overline{z})^n$, 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 4.3.1

Express the following complex numbers on rectangular form.

(a) $\overline{-3 + 6i}$

(b) $\overline{\pi}$

(c) $\overline{-97i}$

Answer:

(a) From the definition of the complex conjugate we find $\overline{-3+6i} = -3-6i$.

(b) $\overline{\pi} = \overline{\pi + 0i} = \pi - 0i = \pi$. This illustrates the more general fact that $\overline{z} = z$, if z is a real number.

(c) $\overline{-97i} = -(-97i) = 97i$. 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 4.3.2

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

(i) $\overline{e^z} = e^{\overline{z}}$,

(ii) $\overline{e^{i\alpha}} = e^{-i\alpha}$,

(iii) $\overline{z} = |z|e^{-i\arg(z)}$.

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

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

□

Example 4.3.2

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 $\overline{z} = |z|e^{-i\arg(z)}$.

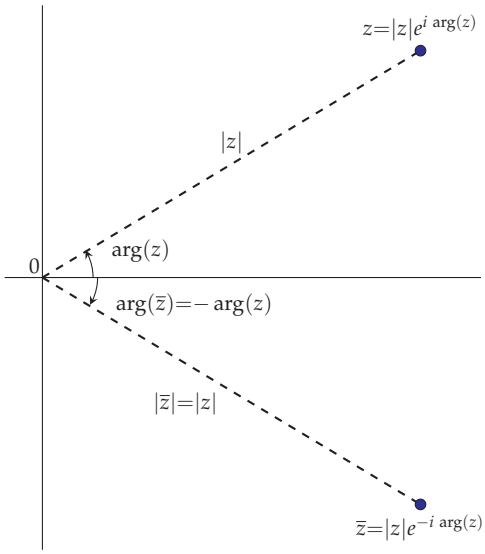


Figure 4.2: Polar form of a complex number z and its complex conjugate $\overline{z}$.

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

Lemma 4.3.3

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 4.3.1, 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 4.3.3 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 4.2.1 that one of its roots is $1 + 2i$. Lemma 4.3.3 implies that the complex number $1 - 2i$ then is a root of $2Z^2 - 4Z + 10$ as well. We have seen in Example 4.2.1 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 4.4.1

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 3.6.1, 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 4.4.1

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 4.4.1, 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 4.4.1 all solutions to $z^4 = -8 + i8\sqrt{3}$ are given by:

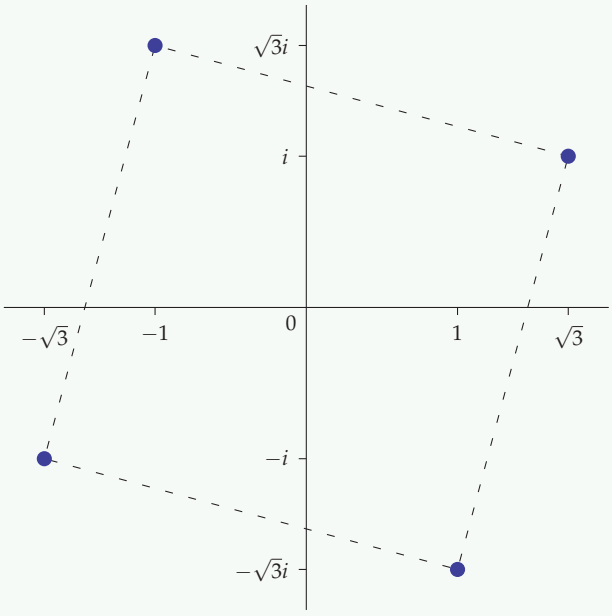
$$z = \sqrt[4]{16}e^{i\left(\frac{2\pi}{3\cdot 4} + p\frac{2\pi}{4}\right)}, \text{where } p \text{ can be chosen freely from the set } \{0, 1, 2, 3\}, \text{ so}$$

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

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 4.4.1, 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.



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 4.4.1 that the equation $s^2 = b^2 - 4ac$ 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 4.4.2

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 4.4.2

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 4.4.1, 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 4.4.2, 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 4.5.1

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 4.5.1

(a) Decide if the polynomial $Z+3$ is a factor of the polynomial $2Z^2+3Z-9$.

(b) Decide if the polynomial $Z+4$ is a factor of the polynomial $3Z^3+2Z+1$.

(c) Decide if the polynomial $2Z^2+Z+3$ is a factor of the polynomial $6Z^4+3Z^3+19Z^2+5Z+15$.

Answer:

(a) 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 $q(Z) = b_1Z+b_0$ for certain numbers $b_1, b_0 \in \mathbb{C}$. We first try to find b_1. Without simplifying the product $(Z+3)\cdot(b_1Z+b_0)$ 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 $(Z+3)\cdot(b_1Z+b_0) = b_1Z^2 +$ terms of degree less than 2. On the other hand we want that $(Z+3)\cdot(b_1Z+b_0) = 2Z^2+3Z-9$. 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 4.5.1 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 $2Z^2+3Z-9$:

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

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 the expression $2Z^2+3Z-9-2Z\cdot(Z+3)$ before our simplicifations. 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:

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

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 $2Z^2+3Z-9=(Z+3)\cdot(2Z-3)$.

(b) 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:

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

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

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

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

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

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 $3\cdot(-4)^3+2\cdot(-4)+1 = -199$.

(c) We state the schematic procedure only this time:

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

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:

(i) $p(Z) = d(Z)q(Z)+r(Z)$.

(ii) $r(Z) = 0 \quad \vee \quad deg(r(z)) < \deg(d(z))$.

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:

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

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:

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

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:

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

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:

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

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 4.6.1 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 4.4.2. 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 4.6.2
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 4.6.1

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

Example 4.6.1

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:

(a) 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 4.5.1 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.

(b) We have $p_2(-3) = 1$. Therefore -3 is not a root of $p_2(Z)$.

(c) 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.

(d) 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.

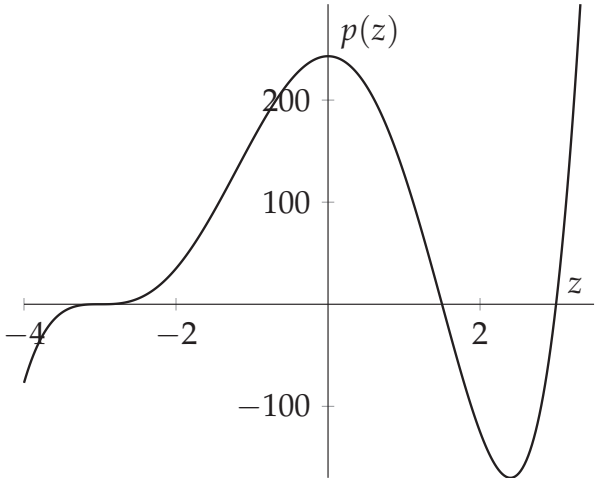


Figure 4.3: The graph of the polynomial function $p : \mathbb{R} \to \mathbb{R}$, where $p(z) = 2z^5 + 9z^4 - 18z^3 - 108z^2 + 243$.

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 4.6.1) says that each polynomial of degree at least 1 has a root. This has the following consequence:

Theorem 4.6.3

Let $p(Z) = a_nZ^n + a_{n-1}Z^{n-1} + \cdots + a_1Z + a_0$ 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 4.6.2, 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 4.6.2, 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 4.6.2

As an example we take the polynomial $p_4(Z) = 2Z^5 + 9Z^4 - 18Z^3 - 108Z^2 + 243$ from Example 4.6.1. We wish to write this polynomial as in Theorem 4.6.3. 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 4.6.3 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 4.6.3 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 4.6.3 has the following consequence

Corollary 4.6.4

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 4.6.3 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 4.3.3 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 4.6.3 into a factorization of $p(Z)$ in $\mathbb{R}[Z]$ in first and second degree factors with real coefficients. □

Example 4.6.3

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

(a) $p_1(Z) = Z^3 - Z^2 + Z - 1$

(b) $p_2(Z) = Z^4 + 4$

Answer:

(a) 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:

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

(b) 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 4.6.4 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

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

Chapter 5

Recursion and induction

5.1 Examples of 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 5.1.1

The *factorial* function $\text{fac} : \mathbb{N} \to \mathbb{N}$ 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 $\text{fac}(n)$. We have for example $\text{fac}(1) = 1$, $\text{fac}(2) = 1 \cdot 2 = 2$, $\text{fac}(3) = 1 \cdot 2 \cdot 3 = 6$, $\text{fac}(4) = 1 \cdot 2 \cdot 3 \cdot 4 = 24$, etcetera. Now note that, if we want to compute the next value, $\text{fac}(5)$, we can use that we already know what $\text{fac}(4)$ is. Indeed,

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

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

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

Input: $n \in \mathbb{Z}_{\geq 1}$.
1: **if** $n = 1$ **then**
2: **return** 1
3: **else**
4: **return** $\text{fac}(n-1) \cdot n$.

This algorithm simply uses itself to compute $\text{fac}(n)$. More precisely, if $n = 1$, it directly

returns 1 as the value for $\text{fac}(1)$, as prescribed in line 2 of the algorithm. Graphically, we can illustrate this as follows:

$\text{fac}(1)$

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

return 1

If $n = 2$, the algorithm will go to line 4 and attempt to return $\text{fac}(2-1) \cdot 2$. However, this requires that first the value of $\text{fac}(1)$ 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 $\text{fac}(1) = 1$, it can revisit line 4 and compute that $\text{fac}(2) = \text{fac}(1) \cdot 2 = 1 \cdot 2 = 2$. Hence the algorithm returns 2. Graphically, the situation is:

$\text{fac}(2)$

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

$\text{fac}(1)$

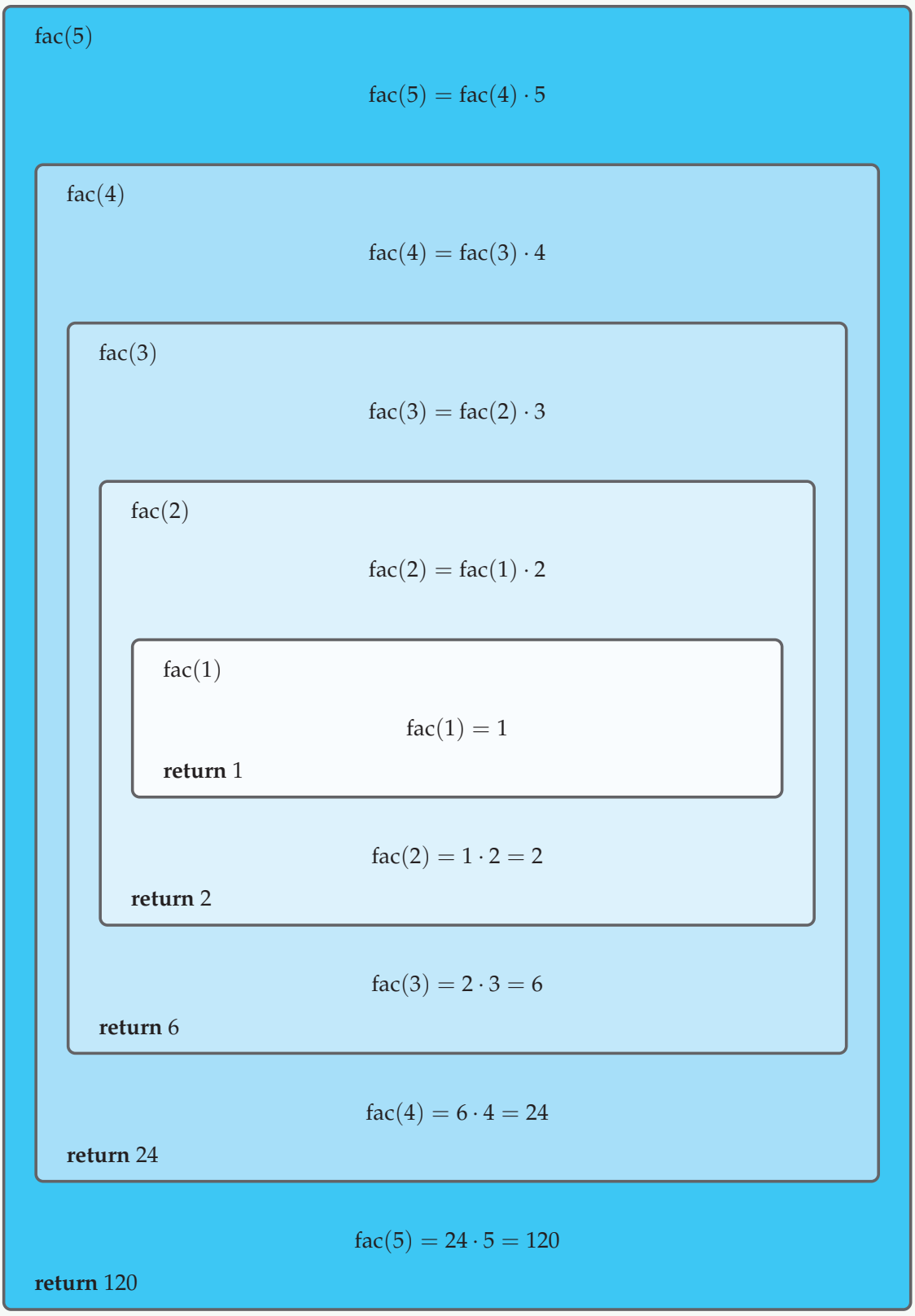
$$\text{fac}(1) = 1$$

return 1

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

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 $n = 5$:


fac(5)
$\mathrm{fac}(5) = \mathrm{fac}(4) \cdot 5$
fac(4)
$\mathrm{fac}(4) = \mathrm{fac}(3) \cdot 4$
fac(3)
$\mathrm{fac}(3) = \mathrm{fac}(2) \cdot 3$
fac(2)
$\mathrm{fac}(2) = \mathrm{fac}(1) \cdot 2$
fac(1)
$\mathrm{fac}(1) = 1$
return 1
$\mathrm{fac}(2) = 1 \cdot 2 = 2$
return 2
$\mathrm{fac}(3) = 2 \cdot 3 = 6$
return 6
$\mathrm{fac}(4) = 6 \cdot 4 = 24$
return 24
$\mathrm{fac}(5) = 24 \cdot 5 = 120$
return 120

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:

$$\mathrm{fac}(n) = \begin{cases} 1 & \text{if } n = 1, \\ \mathrm{fac}(n-1) \cdot n & \text{if } n \geq 2. \end{cases} \tag{5.1}$$

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 5.4.2, 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 5.1.2

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 sequence 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.

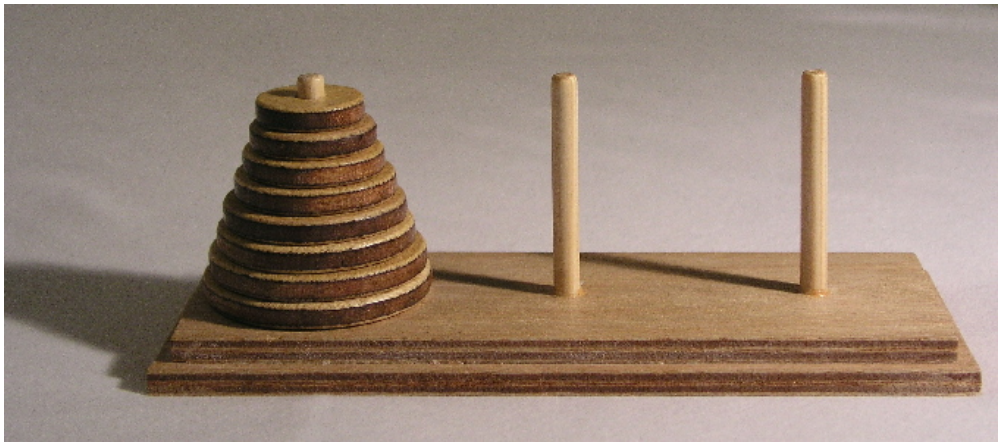
5.2 The towers of Hanoi

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. The picture in this figure was taken from wikipedia; see https://commons.wikimedia.org/wiki/File:Tower_of_Hanoi.jpeg for more details.

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.

Figure 5.1: The tower of Hanoi with eight discs.



- 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:

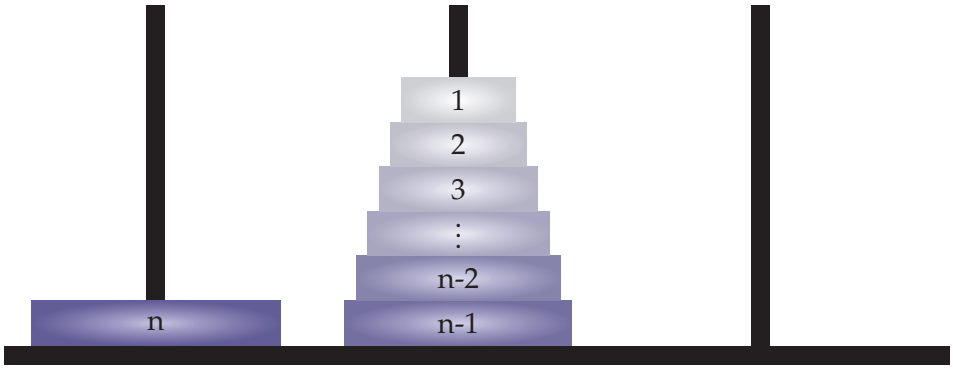
Move disc 1 from stick 1 to stick 2

Move disc 2 from stick 1 to stick 3

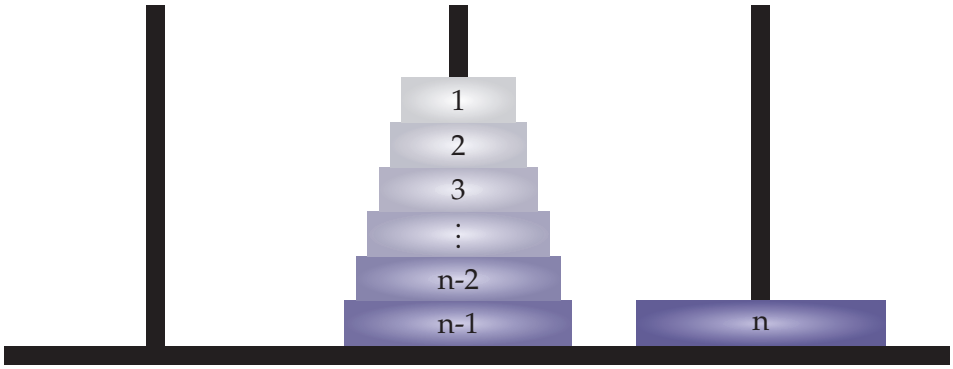
Move disc 1 from stick 2 to stick 3

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:

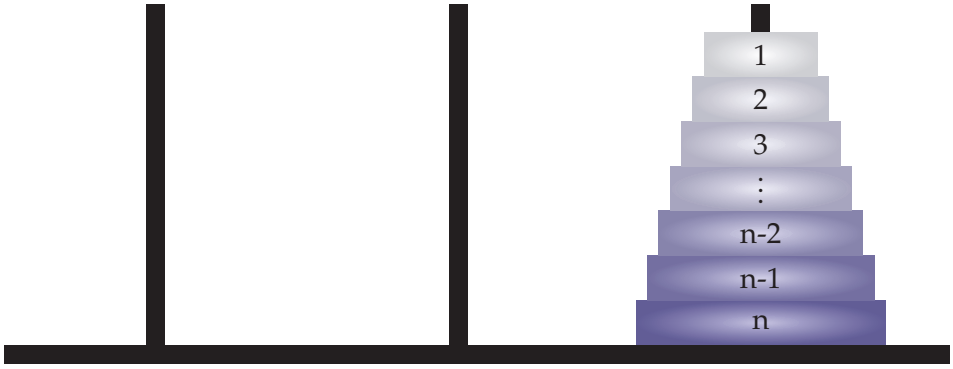
Using that we know how to move $n-1$ discs to another stick, move the stack of discs 1 to $n-1$ from stick 1 to stick 2.



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

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: $\sum_{k=1}^{n} z_k$. Using a recursive definition, we can be very precise:

$$\sum_{k=1}^{n} z_k = \begin{cases} z_1 & \text{if } n = 1, \\ \left(\sum_{k=1}^{n-1} z_k\right) + z_n & \text{if } n > 1. \end{cases} \tag{5.5}$$

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	$\sum_{k=1}^{n} z_k$
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 $\sum_{k=1}^{n} f(k)$. Consider for example the expression $\sum_{k=1}^{n} k$, that is to say, the sum of the first n natural numbers. Similarly as in Table 5.6, we obtain the following:

n	$\sum_{k=1}^{n} k$
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.

The variable k in an expression like $\sum_{k=1}^{n} z_k$ is called the summation index. There is no reason to use the variable k as such and using another variable is completely fine. In particular one has for example $\sum_{k=1}^{n} z_k = \sum_{j=1}^{n} z_j$, since both summations amount to adding up the numbers $z_1, \ldots, z_n$. Also one may index the numbers that are to be added in a different way. In particular, if we want to add up the numbers $z_2, \ldots, z_{10}$, one can simply write $\sum_{k=2}^{10} z_k$.

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 5.4.1 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 5.4.1. Such consequences are often called "corollaries" in mathematical texts and we will use the same terminology.

Corollary 5.4.2

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 5.4.1, 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 5.4.1 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 5.4.1 are satisfied. Therefore by Theorem 5.4.1, 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 5.4.1, 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 5.4.2 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 5.4.1

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 $S(n) = \sum_{k=1}^{n} k$. 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

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

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 5.4.2

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 $\text{fac} : \mathbb{N} \to \mathbb{N}$ 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 $\text{fac} : \mathbb{N} \to \mathbb{N}$, 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 $\text{fac}(n)$ after finitely many steps.

Therefore, let $P(n)$ be the statement that $\text{fac}(n)$ 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 $\text{fac}(1) = 1$. Now let $n \geq 2$ be an arbitrary natural number and assume as induction hypothesis that $\text{fac}(n-1)$ can be computed in finitely many steps using Equation (5.1). Since $n \geq 2$, Equation (5.1) implies that $\text{fac}(n) = \text{fac}(n-1)\cdot n$. Hence given $\text{fac}(n-1)$, all we need is one multiplication with n to compute $\text{fac}(n)$. Hence $\text{fac}(n)$ can be computed in finitely many steps, if $\text{fac}(n-1)$ 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 4.1.1. 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 5.5.1

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 5.4.2 to the logical statements $Q(n)$. The first requirement from Corollary 5.4.2 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 5.4.2 becomes that for all $n \in \mathbb{N}_{\geq 2}$: $\quad Q(n-1) \Rightarrow Q(n)$. 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 5.4.2 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 5.4.2. 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 5.5.1

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 5.5.1 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 $n \in \mathbb{Z}_{\geq 5}$.

Chapter 6

Systems of linear equations

6.1 Structure of systems of linear equations

When dealing with an equation in one variable, it is very common to use the variable x. In Example 1.4.2, we studied for example the equation $2|x| = 2x + 1$. Often there is not just one variable, but several. If there are two variables, one often uses x and y, if there are three x, y and z, but what to do if there are more variables, say five variables? In such cases it is common to use variables x_1, x_2, etcetera. For example, if we need five variables, we just use x_1, x_2, x_3, x_4 and x_5. We can even leave the precise number of variables unspecified and say that we have n variables for some natural number $n \in \mathbb{N}$. One says that one has an equation in the n variables $x_1, \ldots, x_n$.

A *linear equation* in the n variables $x_1, \ldots, x_n$ is an equation of the form

$$a_1 \cdot x_1 + \cdots a_n \cdot x_n = b,$$

where $a_1, \ldots, a_n, b$ are constants. These constants will typically be real or complex numbers, depending on the situation. To avoid having to specify all the time if we are working with real or complex numbers, let us introduce the following definition:

Definition 6.1.1

A set $\mathbb{F}$ is called a *field*, if there is an addition $+$ and multiplication $\cdot$ defined for all pairs of elements of $\mathbb{F}$ in such a way that the following rules are satisfied:

(i) Addition and multiplication are *associative*: $a_1 + (a_2 + a_3) = (a_1 + a_2) + a_3$, and $a_1 \cdot (a_2 \cdot a_3) = (a_1 \cdot a_2) \cdot a_3$ for all $a_1, a_2, a_3 \in \mathbb{F}$.

(ii) Addition and multiplication are *commutative*: $a_1 + a_2 = a_2 + a_1$, and $a_1 \cdot a_2 = a_2 \cdot a_1$ for all $a_1, a_2, a_3 \in \mathbb{F}$.

(iii) *Distributivity* of multiplication over addition holds: $a_1 \cdot (a_2 + a_3) = a_1 \cdot a_2 + a_1 \cdot a_3$ for all $a_1, a_2, a_3 \in \mathbb{F}$.

(iv) Addition and multiplication have a neutral element. More precisely, there exist two distinct elements in $\mathbb{F}$ usually denoted by 0 and 1 that satisfy $a + 0 = a$ and $a \cdot 1 = a$ for all $a \in \mathbb{F}$./

(v) Additive inverses exist: for every $a \in \mathbb{F}$, there exists an element in $\mathbb{F}$, denoted by $-a$ and called the additive inverse of a, such that $a + (-a) = 0$.

(vi) Multiplicative inverses exist: for every $a \in \mathbb{F} \setminus \{0\}$, there exists an element in $\mathbb{F}$, denoted by a^{-1} or $1/a$ and called the multiplicative inverse of a, such that $a \cdot a^{-1} = 1$.

Theorems 3.2.2 and 3.2.3 together simply state that the complex numbers form a field. Also the real numbers $\mathbb{R}$ with the usual addition and multiplication form a field. There are many more possible examples of fields, but whenever we use the symbol $\mathbb{F}$ or write something like "the field $\mathbb{F}$", you can just think of $\mathbb{R}$ or $\mathbb{C}$. Just to show that there exist more fields, we give two examples.

Example 6.1.1

Let $\mathbb{F} = \mathbb{Q}$ be the set of rational numbers, see Example 2.1.4. This set, equipped with the usual addition and multiplication, is a field. It is called the field of rational numbers.

Example 6.1.2

Let $\mathbb{F}_2 = \{0, 1\}$ and define addition and multiplication as follows: $0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1, 1 + 1 = 0$ and $0 \cdot 0 = 0, 0 \cdot 1 = 0, 1 \cdot 0 = 0, 1 \cdot 1 = 1$. Then with this addition and multiplication, $\mathbb{F}_2$ is a field. It is called the field of bits, the binary field, or also the finite field with two elements.

Returning to our study of linear equations, we can now give a more precise definition.

Definition 6.1.2

A linear equation over a field $\mathbb{F}$ in the n variables $x_1, \ldots, x_n$, is an equation of the form

$$a_1 \cdot x_1 + \cdots + a_n \cdot x_n = b,$$

where $a_1, \ldots, a_n, b \in \mathbb{F}$.

A solution to this linear equation is an n-tuple $(v_1, \ldots, v_n) \in \mathbb{F}^n$ such that $a_1 \cdot v_1 + \cdots + a_n \cdot v_n = b$.

We have seen the notation $\mathbb{F}^n$ in this definition before in Section 2.1, see equation (2.3). It is the Cartesian product of $\mathbb{F}$ with itself n times. More down to earth, $\mathbb{F}^n$ is simply the set of all n-tuples $(v_1, \ldots, v_n)$, where each coordinate is an element from $\mathbb{F}$. Sometimes the multiplication between the constant and variables are omitted. For example $2x_1$ has the same meaning as $2 \cdot x_1$.

There is a subtlety in Definition 6.1.2 that is easy to miss. If we say that we consider a linear equation over $\mathbb{F}$, we are only interested in solutions $(v_1, \ldots, v_n)$ that lie in $\mathbb{F}^n$. In other words, by specifying that the linear equation is over $\mathbb{F}$, we implicitly say that all the coordinates of a solution $(v_1, \ldots, v_n)$ must lie in $\mathbb{F}$. Let us consider a few examples.

Example 6.1.3 (a) Find a solution to the linear equation $3x_1 + x_2 = 5$ over $\mathbb{R}$.

(b) Consider the linear equation $x_1 + x_2 = 0$ over $\mathbb{C}$. Is $(i, -i) \in \mathbb{C}^2$ a solution to this linear equation?

(c) Consider the linear equation $x_1 + x_2 = 0$ over $\mathbb{R}$. Is $(i, -i) \in \mathbb{C}^2$ a solution to this linear equation?

(d) Consider the linear equation $x_1 + x_2 = 0$ over $\mathbb{R}$. Find a solution.

Answer:

(a) There are many possible solutions, but for example $(v_1, v_2) = (0, 5)$ is a solution, since $3 \cdot 0 + 5 = 5$.

(b) Since $i + (-i) = 0$, the pair $(i, -i) \in \mathbb{C}^2$ is indeed a solution to the linear equation $x_1 + x_2 = 0$ over $\mathbb{C}$.

(c) Even though $i + (-i) = 0$, the pair $(i, -i)$ is not a solution to the linear equation $x_1 + x_2 = 0$ over $\mathbb{R}$. The reason is that the pair $(-i, i)$ is not an element of $\mathbb{R}^2$.

(d) A possible solution is $(1, -1)$. Another solution is $(0, 0)$.

Now we arrive at the main topic of this section, namely systems of linear equations. It is simply an extension of Definition 6.1.2 by not considering only one linear equation, but several linear equations over a field $\mathbb{F}$ at the same time.

Definition 6.1.3

A system of m linear equations $R_1, \ldots, R_m$ over a field $\mathbb{F}$ in the n variables $x_1, \ldots, x_n$, is a system of m equations of the form

$$\begin{cases} R_1: & a_{11} \cdot x_1 & + & \cdots & + & a_{1n} \cdot x_n & = & b_1 \\ R_2: & a_{21} \cdot x_1 & + & \cdots & + & a_{2n} \cdot x_n & = & b_2 \\ & & & \vdots & & & & \vdots \\ R_m: & a_{m1} \cdot x_1 & + & \cdots & + & a_{mn} \cdot x_n & = & b_m \end{cases}$$

where $a_{11}, \ldots, a_{mn}, b_1, \ldots, b_m \in \mathbb{F}$.

A solution to this system of linear equations is an n-tuple $(v_1, \ldots, v_n) \in \mathbb{F}^n$ such that for all j between 1 and m it holds that $a_{j1} \cdot v_1 + \cdots + a_{jn} \cdot v_n = b_j$.

Some explanation of the notation is in order. First of all, a double index was used for the constants in front of the variables. The constant a_{ij} denotes the constant occurring in equation i in front of the variable x_j. For example, if we have at least two equations and at least three

variables, then a_{23} would denote the constant in the second equation in front of the variable x_3. In case $m = 1$ in Definition 6.1.3, we just recover the case of one linear equation as described in Definition 6.1.2.

The use of the brace $\{$ in front of the equations is just to emphasize that all equations are considered simultaneously and that a solution to the system should satisfy all equations at the same time. In logical terms, we can therefore write that an n-tuple $(v_1, \ldots, v_n) \in \mathbb{F}^n$ is a solution to the system of equations as given in Definition 6.1.3 precisely if:

$$a_{11} \cdot v_1 + \cdots + a_{1n} \cdot v_n = b_1 \quad \wedge \quad \cdots \quad \wedge \quad a_{m1} \cdot v_1 + \cdots + a_{mn} \cdot v_n = b_m.$$

Using $R_1, \ldots, R_m$ as "labels" for the equations, is not necessary and often these labels are just omitted. We will usually also omit these labels, but when developing the theory on how to solve systems of linear equations, they can be quite convenient. To digest this definition, let us immediately consider some examples.

Example 6.1.4

Determine the set of solutions to the following system of two linear equations in two variables over $\mathbb{R}$:

$$\begin{cases} x_1 + 2x_2 &= 1 \\ x_2 &= 2 \end{cases}$$

This system is quite simple to solve, since the second equation already determines x_2 (namely $x_2 = 2$). Then using this in the first equation, we see that any pair (x_1, x_2) that satisfies *both* linear equations, will satisfy $x_2 = 2$ and $x_1 = 1 - 2x_2 = 1 - 2 \cdot 2 = -3$. Hence the system has only one solution, namely $(x_1, x_2) = (-3, 2)$. The set of all solutions is therefore given by $\{(-3, 2)\}$.

Example 6.1.5

Consider the following system of linear equations over $\mathbb{R}$ in the variables $x_1, \ldots, x_4$:

$$\begin{cases} 2x_1 + 5x_2 + x_4 &= 0 \\ 3x_1 - x_3 &= 6 \end{cases}$$

Let us see how this example fits with Definition 6.1.3. First of all, we have two linear equations and hence $m = 2$. Further, the only variables occurring in these three equations are x_1, x_2, x_3 and x_4. Hence we can choose $n = 4$. To determine the a_{ij} is now a matter of reading off the constants in front of the variables. However, before we do this, it is convenient to rewrite the system of equations a bit as follows:

$$\begin{cases} 2 \cdot x_1 + 5 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 = 0 \\ 3 \cdot x_1 + 0 \cdot x_2 + (-1) \cdot x_3 + 0 \cdot x_4 = 6 \end{cases}$$

We can now read off directly that $a_{11} = 2$, $a_{12} = 5$, $a_{13} = 0$, $a_{14} = 1$, $b_1 = 0$, $a_{21} = 3$, $a_{22} = 0$, $a_{23} = -1$, $a_{24} = 0$ and $b_2 = 6$. We will determine the solutions of this system of linear equations later.

A system of m linear equations over a field $\mathbb{F}$ in the n variables $x_1, \ldots, x_n$ is called *homogeneous*, if for all i between 1 and m, it holds that $b_i = 0$. Otherwise, the system is called *inhomogeneous*. The system given in Example 6.1.5 is inhomogeneous, since in that example $b_2 \neq 0$. An example of a homogeneous system of linear equations in three variables is:

$$\begin{cases} 3 \cdot x_1 + 5 \cdot x_2 + 10 \cdot x_3 = 0 \\ 5 \cdot x_1 + 2 \cdot x_2 - 2 \cdot x_3 = 0 \end{cases}$$

Note that the all-zero tuple $(0, 0, 0)$ is a possible solution to this system. More generally, one can show that a homogeneous system of linear equations in n variables has the all-zero n-tuple $(0, \ldots, 0)$ as solution. Let us end this section by giving two structure theorems concerning the solution sets of systems of linear equations. One will be for homogeneous systems, one for inhomogeneous systems.

Theorem 6.1.1

Let a homogeneous system of m linear equations $R_1, \ldots, R_m$ over a field $\mathbb{F}$ in the n variables $x_1, \ldots, x_n$ be given, say

$$\begin{cases} a_{11} \cdot x_1 + \cdots + a_{1n} \cdot x_n = 0 \\ a_{21} \cdot x_1 + \cdots + a_{2n} \cdot x_n = 0 \\ \vdots \\ a_{m1} \cdot x_1 + \cdots + a_{mn} \cdot x_n = 0 \end{cases}$$

where $a_{11}, \ldots, a_{mn} \in \mathbb{F}$. Then

(i) The all-zero tuple $(0, \ldots, 0) \in \mathbb{F}^n$ is a solution to the system.

(ii) If $(v_1, \ldots, v_n) \in \mathbb{F}^n$ is a solution and $c \in \mathbb{F}$, then $(c \cdot v_1, \ldots, c \cdot v_n)$ is also a solution.

(iii) If $(v_1, \ldots, v_n), (w_1, \ldots, w_n) \in \mathbb{F}^n$ are solutions, then $(v_1 + w_1, \ldots, v_n + w_n)$ is also a solution.

Proof. We have already remarked that the all-zero tuple is a solution to a homogeneous system. We will prove the third statement and leave proving the second statement to the reader. If $(v_1, \ldots, v_n), (w_1, \ldots, w_n) \in \mathbb{F}^n$ are solutions, then we know that for all j between 1 and m that:

$$a_{j1} \cdot v_1 + \cdots + a_{jn} \cdot v_n = 0 \text{ and } a_{j1} \cdot w_1 + \cdots + a_{jn} \cdot w_n = 0.$$

Adding these equations, we find that

$$a_{j1} \cdot v_1 + a_{j1} \cdot w_1 + \cdots + a_{jn} \cdot v_n + a_{jn} \cdot w_n = 0,$$

which can be rewritten as

$$a_{j1} \cdot (v_1 + w_1) + \cdots + a_{jn} \cdot (v_n + w_n) = 0.$$

The reader is encouraged to think about which properties of a field from Definition 6.1.1 we have used here. Since this is true for any j, we may conclude that $(v_1 + w_1, \ldots, v_n + w_n)$ is also a solution to the given homogeneous system of linear equations. □

Theorem 6.1.2

Let an inhomogeneous system of m linear equations $R_1, \ldots, R_m$ over a field $\mathbb{F}$ in the n variables $x_1, \ldots, x_n$ be given, say

$$\begin{cases} a_{11} \cdot x_1 & + & \cdots & + & a_{1n} \cdot x_n & = & b_1 \\ a_{21} \cdot x_1 & + & \cdots & + & a_{2n} \cdot x_n & = & b_2 \\ & & \vdots & & & & \vdots \\ a_{m1} \cdot x_1 & + & \cdots & + & a_{mn} \cdot x_n & = & b_m \end{cases}$$

where $a_{11}, \ldots, a_{mn}, b_1, \ldots, b_m \in \mathbb{F}$ and not all b_i are zero. If the system does have a solution, say $(v_1, \ldots, v_n) \in \mathbb{F}^n$, then any other solution is of the form $(v_1 + w_1, \ldots, v_n + w_n)$, where $(w_1, \ldots, w_n) \in \mathbb{F}^n$ is a solution to the corresponding homogeneous system:

$$\begin{cases} a_{11} \cdot x_1 & + & \cdots & + & a_{1n} \cdot x_n & = & 0 \\ a_{21} \cdot x_1 & + & \cdots & + & a_{2n} \cdot x_n & = & 0 \\ & & \vdots & & & & \vdots \\ a_{m1} \cdot x_1 & + & \cdots & + & a_{mn} \cdot x_n & = & 0 \end{cases}$$

Proof. Suppose that the system has a solution, say $(v_1, \ldots, v_n) \in \mathbb{F}^n$. Let $(v'_1, \ldots, v'_n) \in \mathbb{F}^n$ be any other solution. First of all, if we define $w_i = v'_i - v_i$, then by definition of the w_i, we obtain that $(v'_1, \ldots, v'_n) = (v_1 + w_1, \ldots, v_n + w_n)$. Hence what we need to show is that the tuple $(w_1, \ldots, w_n)$ is a solution to the homogeneous system stated in the theorem. However, we know that for all j:

$$a_{j1} \cdot v'_1 + \cdots + a_{jn} \cdot v'_n = b_j \text{ and } a_{j1} \cdot v_1 + \cdots + a_{jn} \cdot v_n = b_j.$$

Taking the difference of these two equations, we obtain:

$$a_{j1} \cdot v'_1 - a_{j1} \cdot v_1 + \cdots + a_{jn} \cdot v'_n - a_{jn} \cdot v_n = b_j - b_j,$$

which can be rewritten as

$$a_{j1} \cdot (v'_1 - v_1) + \cdots + a_{jn} \cdot (v'_n - v_n) = 0.$$

Since $w_i = v'_i - v_i$, we obtain that for all j it holds that

$$a_{j1} \cdot w_1 + \cdots + a_{jn} \cdot w_n = 0.$$

This is exactly the same as saying that $(w_1, \ldots, w_n)$ is a solution to the homogeneous system given in the theorem. □

It is not a coincidence that Theorem 6.1.2 is formulated as it is. Indeed, the theorem holds if there exists a solution to the inhomogeneous system, but there is no guarantee that such a solution actually exists. A solution to an inhomogeneous system of linear equations is

sometimes called a *particular solution*. Theorem 6.1.2 can then in words be described as stating that all solutions to an inhomogeneous system can be obtained as the sum of a given particular solution (if it exists) and the solutions to the corresponding homogeneous system.

We illustrate Theorem 6.1.2 in Figure 6.1 for a small inhomogeneous system of equations over the field $\mathbb{R}$. In this figure, the green line indicates all solutions to the system, while the blue line indicates all solutions to the corresponding homogeneous system.

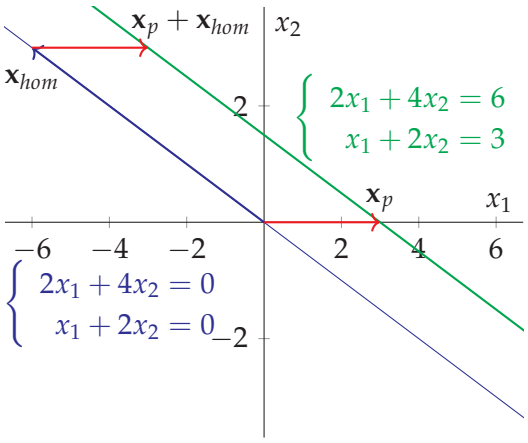


Figure 6.1: The solutions of an inhomogeneous system can be obtained by adding a particular solution $\mathbf{x}_p$ to the solutions $\mathbf{x}_{hom}$ of the corresponding homogeneous system.

Let us for the sake of completeness, give a small example of an inhomogeneous system of linear equations that has no solutions:

Example 6.1.6

Consider the following system of two linear equations in two variables over $\mathbb{R}$:

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 0 \end{cases}.$$

This system is inhomogeneous, since the right-hand side of the first equation is not a zero. This system has no solutions, since it is not possible that $x_1 + x_2$ is equal to 1 and 0 at the same time! Indeed if that would be possible, we could conclude that $0 = 1$, which would be a contradiction.

To make Theorems 6.1.1 and 6.1.2 constructive, we need to figure out a way to answer the following three questions:

(i) How do we describe all solutions to a homogeneous system of linear equations explicitly?

(ii) How do we decide if an inhomogeneous system of linear equations has a solution?

(iii) If it exists, how do we explicitly find a solution to an inhomogeneous system of linear equations?

Note that if we can answer these questions, Theorem 6.1.2 can be used to describe all solutions to an inhomogeneous system of linear equations that have at least one solution. In the next sections we will answer these questions.

6.2 Transforming a system of linear equations

In this section, we will come up with a procedure that transforms a given system of linear equations into a simpler one, without changing the solutions they have. In other words, we want to find a way to replace a possibly complicated looking system of linear equations with another, much simpler system of linear equations, but we want that the initial, possibly complicated, system has exactly the same solutions as the simpler one.

Before we start with that though, we will introduce a compact way to describe a system of linear equations using what are known as *matrices*. For now you can think of a matrix as a rectangular scheme containing elements from the field $\mathbb{F}$ one is working over. In a later chapter, we will have a more in depth discussion of matrices.

Definition 6.2.1

Given a linear system

$$\begin{cases} a_{11} \cdot x_1 & + & \cdots & + & a_{1n} \cdot x_n & = & b_1 \\ a_{21} \cdot x_1 & + & \cdots & + & a_{2n} \cdot x_n & = & b_2 \\ & & \vdots & & & & \vdots \\ a_{m1} \cdot x_1 & + & \cdots & + & a_{mn} \cdot x_n & = & b_m \end{cases}$$

we denote by

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

the *coefficient matrix* of the system of linear equations. The matrix

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{bmatrix}$$

is called the *augmented matrix* of the system of linear equations.

Example 6.2.1

Let us consider the system of linear equations as given in Example 6.1.5. The coefficient matrix of this system is given by

$$\begin{bmatrix} 2 & 5 & 0 & 1 \\ 3 & 0 & -1 & 0 \end{bmatrix},$$

while the augmented matrix of this system is

$$\left[\begin{array}{cccc|c} 2 & 5 & 0 & 1 & 0 \\ 3 & 0 & -1 & 0 & 6 \end{array}\right].$$

Sometimes one writes

$$\left[\begin{array}{cccc|c} 2 & 5 & 0 & 1 & 0 \\ 3 & 0 & -1 & 0 & 6 \end{array}\right]$$

for the augmented matrix to emphasize that the final 0 and 6 come from the right-hand side of the system of linear equations. This is just an esthetic choice though.

One says that a matrix has *rows* and *columns*. A row is a horizontal slice of a matrix, a column a vertical slice. For example, the matrix

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix}$$

has two rows: the first row is given by $[2\,6\,0\,1\,0]$, while the second row is given by $[4\,0\,-1\,0\,6]$. Similarly, it has five columns, namely

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 6 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$

A matrix is said to be an $m \times n$ matrix, if it has precisely m rows and precisely n columns. Hence the matrix we just considered is a 2×5 matrix. If we consider the matrices in Definition 6.2.1, we see that the coefficient matrix of a system of m linear equations in n variables is an $m \times n$ matrix. Similarly, its augmented matrix is an $m \times (n+1)$ matrix. Indeed, it has one more column than the coefficient matrix, containing the b_i from the right-hand sides of the linear equations.

Now let us return to our goal: transforming a system of linear equations over a field $\mathbb{F}$ into a simpler one, without changing the solution set. The idea is to gradually transform any given system over $\mathbb{F}$ into a much simpler system, at each step making sure that the set of solutions did not change. The operations that we will use to transform the systems will consist of three types:

1. Interchange two equations.
2. Multiply a given equation with a nonzero constant from $\mathbb{F}$.

3. Add a multiple of one equation to another.

Let us explain, what these three operations do in more detail. The first one takes two linear equations from a given system, say R_i and R_j, and interchanges them. This means that after the operation R_j occurs in position i and R_i in position j. We denote this operation by $R_i \leftrightarrow R_j$.

Example 6.2.2

Let us illustrate the interchange operation on the system from Example 6.1.5:

$$\begin{cases} 2 \cdot x_1 + 6 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 = 0 \\ 4 \cdot x_1 + 0 \cdot x_2 + (-1) \cdot x_3 + 0 \cdot x_4 = 6 \end{cases}.$$

In this case, we can perform the operation $R_1 \leftrightarrow R_2$ and obtain the system

$$\begin{cases} 4 \cdot x_1 + 0 \cdot x_2 + (-1) \cdot x_3 + 0 \cdot x_4 = 6 \\ 2 \cdot x_1 + 6 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 = 0 \end{cases}.$$

If we, which in fact is more convenient, work with the augmented matrix of this system, the effect of the operation $R_1 \leftrightarrow R_2$ is that the augmented matrix

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix}$$

is replaced by

$$\begin{bmatrix} 4 & 0 & -1 & 0 & 6 \\ 2 & 6 & 0 & 1 & 0 \end{bmatrix}.$$

Hence the operation $R_1 \leftrightarrow R_2$ simply interchanges the first and the second row of the augmented matrix. We will usually write this as follows:

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix} \underset{R_1 \leftrightarrow R_2}{\longrightarrow} \begin{bmatrix} 4 & 0 & -1 & 0 & 6 \\ 2 & 6 & 0 & 1 & 0 \end{bmatrix}.$$

The second operation we will use to simplify systems just multiplies one of the given linear equations with a *nonzero* constant $c \in \mathbb{F}$ (in other words $c \in \mathbb{F} \setminus \{0\}$). This simply means that one replaces the linear equation R_j, say given by $a_{j1}x_1 + \cdots + a_{jn}x_n = b_j$, with the linear equation $ca_{j1}x_1 + \cdots + ca_{jn}x_n = cb_j$ (which is for simplicity just denoted by $c \cdot R_j$). We denote this operation by $R_j \leftarrow c \cdot R_j$.

Example 6.2.3

Let us illustrate the operation $R_1 \leftarrow (1/2) \cdot R_1$ on the system from Example 6.1.5. This amounts to replacing the system

$$\begin{cases} 2 \cdot x_1 + 6 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 = 0 \\ 4 \cdot x_1 + 0 \cdot x_2 + (-1) \cdot x_3 + 0 \cdot x_4 = 6 \end{cases}$$

by

$$\begin{cases} 1\cdot x_1 + 3\cdot x_2 + 0\cdot x_3 + 1/2\cdot x_4 = 0 \\ 4\cdot x_1 + 0\cdot x_2 + (-1)\cdot x_3 + 0\cdot x_4 = 6 \end{cases}.$$

In matrix notation, we obtain

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix} \underset{R_1 \leftarrow (1/2)\cdot R_1}{\longrightarrow} \begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix}.$$

Hence the effect of the operation $R_1 \leftarrow (1/2)\cdot R_1$ on the augmented matrix is that all entries in the first row are multiplied with $1/2$. We have used the arrow $\longrightarrow$ to indicate one step when changing the matrix. Later on, we will gradually change the matrix and use the arrow $\longrightarrow$, each time an operation is used. Below the arrow, we write which operation is used (in this case $R_1 \leftarrow (1/2)\cdot R_1$).

Finally, the third operation, adding d times equation R_j to an equation R_i (where $i \neq j$ and $d \in \mathbb{F}$), simply means that the linear equation R_i given by $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$ is replaced by the equation $(a_{i1} + da_{j1})x_1 + \cdots + (a_{in} + da_{jn})x_n = b_i + db_j$. One can briefly state this by writing that the linear equation R_i is replaced by $R_i + d\cdot R_j$, or in other words as $R_i \leftarrow R_i + d\cdot R_j$.

Example 6.2.4

Again let us use the system from Example 6.1.5 to illustrate the effect of the operation $R_1 \leftarrow R_1 + 2\cdot R_2$. This amounts to replacing the system

$$\begin{cases} 2\cdot x_1 + 6\cdot x_2 + 0\cdot x_3 + 1\cdot x_4 = 0 \\ 4\cdot x_1 + 0\cdot x_2 + (-1)\cdot x_3 + 0\cdot x_4 = 6 \end{cases}$$

by

$$\begin{cases} 10\cdot x_1 + 6\cdot x_2 + (-2)\cdot x_3 + 1\cdot x_4 = 12 \\ 4\cdot x_1 + 0\cdot x_2 + (-1)\cdot x_3 + 0\cdot x_4 = 6 \end{cases}.$$

In matrix notation, we obtain

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix} \underset{R_1 \leftarrow R_1 + 2\cdot R_2}{\longrightarrow} \begin{bmatrix} 10 & 6 & -2 & 1 & 12 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix}.$$

Hence the effect of the operation $R_1 \leftarrow R_1 + 2\cdot R_2$ on the augmented matrix, is that the first row is replaced by the first row plus two times the second row.

As is clear from the examples, the effect of the three operations $R_i \leftrightarrow R_j$, $R_j \leftarrow c\cdot R_j$, and $R_i \leftarrow R_i + d\cdot R_j$ can be seen as easy operations on the rows of the augmented matrix of the system of linear equations we started with. For this reason, they are called *elementary row operations*. This is in fact also the reason why we used capital R in the labels $R_1, \ldots, R_m$ for the linear equations in our system: the R simply was inspired by the first letter in the word "row".

Now let us make sure that when using any of these elementary operations, the solution set of the new system is identical to that of the original system of linear equations. In fact, let us state this as a theorem:

Theorem 6.2.1

Let $R_1, \ldots, R_m$ be a system of m linear equations in n variables over a field $\mathbb{F}$. Further, let i and j be two distinct integers between 1 and m. Then any of the systems obtained by applying one of the operations $R_i \leftrightarrow R_j$, $R_j \leftarrow c \cdot R_j$, with $c \in \mathbb{F} \setminus \{0\}$ or $R_i \leftarrow R_i + d \cdot R_j$, with $d \in \mathbb{F}$, has the same set of solutions as the original system.

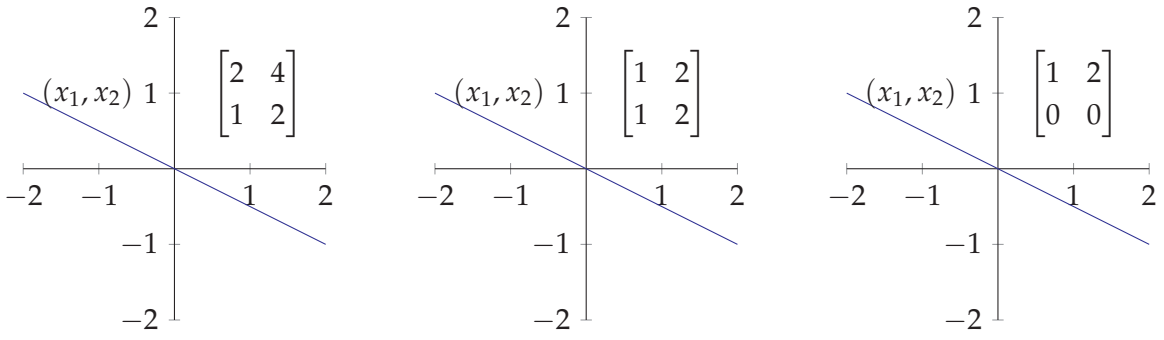
Proof. We only prove the theorem for the elementary operation $R_i \leftarrow R_i + d \cdot R_j$. The reader is encouraged to check that the theorem is also true for the remaining two elementary operations. We need to show that the set of solutions of the system of linear equations $R_1, \ldots, R_{i-1}, R_i, R_{i+1}, \ldots, R_m$ is the same as the set of solutions of the system given by $R_1, \ldots, R_{i-1}, R_i + d \cdot R_j, R_{i+1}, \ldots, R_m$. Let us denote the first set of solutions by S and the second set by T. We wish to show that $S = T$.

First of all, we claim that $S \subseteq T$. Therefore, let us choose arbitrary $(v_1, \ldots, v_n) \in S$. We want to show that $(v_1, \ldots, v_n) \in T$. In other words, assuming that $(v_1, \ldots, v_n) \in \mathbb{F}^n$ is a common solution to the linear equations $R_1, \ldots, R_m$, we need to show that it also is a common solution to the linear equations $R_1, \ldots, R_{i-1}, R_i + d \cdot R_j, R_{i+1}, \ldots, R_m$. But then we only need to show that $(v_1, \ldots, v_n)$ is a solution to $R_i + d \cdot R_j$. This is certainly true, since if $(v_1, \ldots, v_n)$ is a common solution to R_i and R_j, then it is also a solution to $R_i + d \cdot R_j$ for any constant $d \in \mathbb{F}$. Hence $(v_1, \ldots, v_n) \in T$. Since we chose $(v_1, \ldots, v_n) \in S$ arbitrarily, this implies that $S \subseteq T$.

Now we claim that $T \subseteq S$. We choose arbitrary $(v_1, \ldots, v_n) \in T$ and now want to show that $(v_1, \ldots, v_n) \in S$. This means that we may assume that $(v_1, \ldots, v_n) \in \mathbb{F}^n$ is a common solution to the linear equations $R_1, \ldots, R_{i-1}, R_i + d \cdot R_j, R_{i+1}, \ldots, R_m$. We need to show that $(v_1, \ldots, v_n)$ is a solution to R_i. However, this is true, since $R_i = (R_i + d \cdot R_j) - d \cdot R_j$. Hence $(v_1, \ldots, v_n) \in S$. Since we chose $(v_1, \ldots, v_n) \in T$ arbitrarily, this implies that $T \subseteq S$.

Now that we have shown that $S \subseteq T$ and $T \subseteq S$, Lemma 2.1.1 implies that $S = T$, which is what we wanted to show. □

Theorem 6.2.1 is illustrated in Figure 6.2. In this figure, the blue line indicates the set of solutions $(x_1, x_2) \in \mathbb{R}^2$ to the homogeneous system of equations given by $\begin{cases} 2x_1 + 4x_2 = 0 \\ x_1 + 2x_2 = 0 \end{cases}$, which has coefficient matrix $\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$. As this system is transformed to a new simpler system using elementary row operations, the solution set remains the same.



(a) The initial system (b) $R_1 \leftarrow (1/2)R_1$. (c) $R_2 \leftarrow R_1 + (-1)R_2$.

Figure 6.2: The solution set of a system of equations does not change when transforming the system using elementary row operations.

It turns out that with these three rather elementary operations in hand, we can find the set of solutions to any system of linear equations. Using one elementary row operation, may not simplify a system of linear equation so much, but the idea is that if we use several elementary row operations in succession, we can transform any given system into a much simpler one. In the next sections, we will see how, but for now, let us consider an example.

Example 6.2.5

Let us revisit Example 6.1.5. There we considered the following system of 2 equations in 4 variables over $\mathbb{R}$:

$$\begin{cases} 2 \cdot x_1 + 6 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 = 0 \\ 4 \cdot x_1 + 0 \cdot x_2 + (-1) \cdot x_3 + 0 \cdot x_4 = 6 \end{cases}.$$

Let us simplify this system, applying elementary row operations. As we have seen in Theorem 6.2.1, this does not change the solution set of the system. Since it is much more compact to work with the augmented matrix of the system, let us do that as well.

First, applying the transformation $R_1 \leftarrow (1/2) \cdot R_1$, we obtain the augmented matrix:

$$\begin{bmatrix} 2 & 6 & 0 & 1 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix} \xrightarrow[R_1 \leftarrow (1/2) \cdot R_1]{} \begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix}.$$

The point of this operation, was to get a one in the first entry of the first row. This makes it easy to eliminate the x_1 variable from the second equation. In other words, in the next step we want to create a zero in the first entry of the second row. We achieve this applying the elementary row operation $R_2 \leftarrow R_2 - 4 \cdot R_1$, since then we obtain

$$\begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 4 & 0 & -1 & 0 & 6 \end{bmatrix} \xrightarrow[R_2 \leftarrow R_2 - 4 \cdot R_1]{} \begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & -12 & -1 & -2 & 6 \end{bmatrix}.$$

Now we simplify further, by making the coefficient for x_2 in the second equation equal to one. In other words, we now want to make the second entry in the second row equal to one. To

achieve this, we apply $R_2 \leftarrow (-1/12) \cdot R_2$:

$$\begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & -12 & -1 & -2 & 6 \end{bmatrix} \xrightarrow[R_2 \leftarrow (-1/12) \cdot R_2]{} \begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & 1 & 1/12 & 2/12 & -6/12 \end{bmatrix}.$$

The fractions in the resulting matrix can actually be simplified a bit, so we could also have written:

$$\begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & -12 & -1 & -2 & 6 \end{bmatrix} \xrightarrow[R_2 \leftarrow (-1/12) \cdot R_2]{} \begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & 1 & 1/12 & 1/6 & -1/2 \end{bmatrix}.$$

The corresponding system is now nearly as simple as we can make it, but we can still use the second equation to get rid of the x_2 term in the first equation using $R_1 \leftarrow R_1 - 3 \cdot R_2$:

$$\begin{bmatrix} 1 & 3 & 0 & 1/2 & 0 \\ 0 & 1 & 1/12 & 1/6 & -1/2 \end{bmatrix} \xrightarrow[R_1 \leftarrow R_1 - 3 \cdot R_2]{} \begin{bmatrix} 1 & 0 & -1/4 & 0 & 3/2 \\ 0 & 1 & 1/12 & 1/6 & -1/2 \end{bmatrix}.$$

The corresponding system of linear equations is:

$$\begin{cases} x_1 + (-1/4) \cdot x_3 = 3/2 \\ x_2 + (1/12) \cdot x_3 + (1/6) \cdot x_4 = -1/2 \end{cases}.$$

It is important to remember that by Theorem 6.2.1, the set of solutions to this last system, is exactly the same as the set of solutions to the system we started with.

It is easy to find solutions $(v_1, v_2, v_3, v_4) \in \mathbb{R}^4$ to the last system: simply choose $v_3, v_4 \in \mathbb{R}$ as you want, then use the linear equations to solve for v_1 and v_2. For example, if we choose $v_3 = 0$ and $v_4 = 3$, then we find that $v_1 = (1/4)v_3 + 3/2 = 3/2$ and $v_2 = -(1/12)v_3 + (-1/6)v_4 - 1/2 = -1$. Hence $(3/2, -1, 0, 3)$ is a solution to the system. More, and in fact all, solutions can be obtained in this way: choose any value for v_3 and v_4 that you like and determine the corresponding v_1 and v_2 from the equations $v_1 = (1/4)v_3 + 3/2$ and $v_2 = -(1/12)v_3 + (-1/6)v_4 - 1/2$.

This example shows that it can help a great deal to simplify a given system of linear equation first, before trying to solve it.

6.3 The reduced row echelon form of a matrix

We have seen in Example 6.2.5 that using elementary row operations, can help to describe the solution set of a system of linear equations. What we will do now is to show that this approach always works. Rather than working with systems of linear equations, we will work with the coefficient and augmented matrix of the system. We have seen that if the system consists of m linear equations in n variables, then the coefficient matrix is an $m \times n$ matrix, while the augmented matrix is an $m \times (n+1)$ matrix. The entries in these matrices are from $\mathbb{F}$, the field we are working over. As mentioned before, we will typically work with either

$\mathbb{F} = \mathbb{R}$, the real numbers, or $\mathbb{F} = \mathbb{C}$, the complex numbers. The set of all $m \times n$ matrices with entries in $\mathbb{F}$ will be denoted by $\mathbb{F}^{m\times n}$. In formulas, we will typically use bold face letters, such as $\mathbf{A}, \mathbf{B}, \ldots$ for matrices.

We begin by defining a special kind of matrix:

Definition 6.3.1

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ an $m \times n$ matrix with entries in $\mathbb{F}$. One says that $\mathbf{A}$ is in *reduced row echelon form*, if all of the following are fulfilled.

(i) If a row of the matrix contains only zeros, it appears at the bottom of the matrix. Such rows are called zero rows.

(ii) The left-most non-zero entry in any non-zero row is equal to 1. This entry is called the *pivot* of the row.

(iii) Pivots of two non-zero rows of the matrix do not occur in the same column. Moreover, the pivot of the upper row is further to the left than the pivot of the lower row.

(iv) If a column of the matrix contains a pivot, then all other entries in that column are 0.

A matrix satisfying the first three items, but not necessarily the fourth item, is said to be in *row echelon form*.

Example 6.3.1

The 1×4 matrices $[0\ 0\ 0\ 0]$ and $[0\ 0\ 1\ 5]$ are both in reduced row echelon form. Also the 2×5 matrix

$$\begin{bmatrix} 1 & 0 & -1/4 & 0 & 3/2 \\ 0 & 1 & 1/12 & 1/6 & -1/2 \end{bmatrix}$$

which we obtained at the end of Example 6.2.5, is in reduced row echelon form.

An example of a 1×4 matrix that is not in reduced row echelon form is: $[0\ 0\ 2\ 0]$. Indeed, the left-most non-zero entry in the first (and only) row is not equal to 1. An example of a 3×4 matrix that is not in reduced row echelon form is:

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

This matrix is in row echelon form, but not in reduced row echelon form. The problem here is the third column. This column contains a pivot, namely the pivot of the second row, but apart from the pivot, this column contains another non-zero element (the 2).

The reason that reduced row echelon forms are so important for us is the following result:

Theorem 6.3.1

Let $\mathbf{A} \in \mathbb{F}^{m \times n}$ be a matrix. Then $\mathbf{A}$ can be brought into reduced row echelon form using elementary row operations.

Proof. We will give a sketch of the proof. The strategy is to first bring the matrix in row echelon form, and afterwards in reduced row echelon form. Let us therefore first show that we can use elementary row operations to bring the matrix $\mathbf{A}$ in row echelon form. To do this, we will use induction on m, the number of rows.

If $m = 1$ (the base case of the induction), then the only way $\mathbf{A}$ cannot be in row echelon form, is if the row contains a non-zero entry and the left-most non-zero entry, say c, is not equal to one. Then the operation $R_1 \leftarrow c^{-1} \cdot R_1$ will bring $\mathbf{A}$ in row echelon form.

For the induction step, suppose $m > 1$ is given and that the theorem is true for $(m-1) \times n$ matrices. If all entries in the matrix $\mathbf{A}$ are zero, it is already in row echelon form (and in fact also reduced row echelon form) and we are done. Therefore, let us now assume that the matrix $\mathbf{A}$ has at least one nonzero entry. We start by choosing the smallest possible j such that the j-th column of $\mathbf{A}$ contains a nonzero entry. In particular, if $j > 1$, then the first $j-1$ columns of $\mathbf{A}$ are all zero columns. After this, we choose the smallest possible i such that a_{ij}, the (i,j)-th entry of $\mathbf{A}$, is nonzero. Now we perform the operation $R_1 \leftrightarrow R_i$. The first row of the resulting matrix has a nonzero entry in its j-th position, say c, and zero entries in positions 1 up till $j-1$. Next, we perform the operation $R_1 \leftarrow c^{-1}R_1$, implying that now the j-th entry in the first row has become a 1. If not all elements below this 1 are zeros, we use elementary operations of the form $R_j \leftarrow R_j + dR_1$ for suitably chosen $d \in \mathbb{F}$ to transform the matrix further into a matrix, where there are only zeros below the pivot in row one. We have now transformed the matrix $\mathbf{A}$ into a matrix $\mathbf{B}$ of the form

$$\mathbf{B} = \begin{bmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * \\ 0 & \cdots & 0 & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{bmatrix}.$$

In this notation, the first part of the matrix $\mathbf{B}$ was given as

$$\begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}.$$

This reflects the fact that the first $j-1$ columns of $\mathbf{B}$ are zero. The notation is not meant to suggest that the first part of $\mathbf{B}$ contains at least two zero columns. Indeed, if $j = 2$, this part just consists of one zero column, since then $j - 1 = 1$. In the case that $j = 1$, the first column of the matrix $\mathbf{B}$ is actually not zero at all, but is the column whose first coordinate is 1 and otherwise contains zeroes.

Irrespective on the precise value of j, we now proceed by simply removing the first row of the matrix $\mathbf{B}$ and denote the $(m-1) \times n$ matrix that remains, by $\mathbf{C}$. Using the induction hypothesis, we can conclude that we can use elementary row operations to transform the matrix $\mathbf{C}$ into a matrix $\hat{\mathbf{C}}$ that is in row echelon form. Putting back the first row from $\mathbf{B}$, we find an $m \times n$ matrix, say $\hat{\mathbf{A}}$, that is in row echelon form.

This concludes the inductive proof that any matrix can be brought in row echelon form using elementary row operations. What remains to be done is to bring this matrix in reduced row echelon form. We know by definition of row echelon form that pivots of two non-zero rows of the matrix $\hat{\mathbf{A}}$ do not occur in the same column and moreover, that the pivot of the upper row is further to the left than the pivot of a lower row. Therefore, the entries below a pivot in the matrix $\hat{\mathbf{A}}$, are zero. However, the entries above a pivot in this matrix may not be zero. This can be achieved using elementary row operations of the form $R_i \leftarrow R_i + dR_j$, where row R_j contains a pivot and $i < j$. More precisely, we start using the row containing the right-most pivot to create zeros above this pivot and then work our way to the left, dealing with one pivot at the time. Once we have arrived at the left-most pivot and carried out the sketched procedure for that pivot as well, the obtained matrix will be in reduced row echelon form. □

As an example, we can simply look at Example 6.2.5. There we used elementary row operations to bring a matrix in reduced row echelon form. There are in principle many different ways to use elementary row operations to transform a given matrix $\mathbf{A}$ into reduced row echelon form. However, for a given matrix $\mathbf{A}$, it turns out that the outcome is always the same. Therefore we can talk about *the* reduced row echelon form of a matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$. In particular, the following definition is justified:

Definition 6.3.2

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m \times n}$ a matrix. Then the *rank* of $\mathbf{A}$, denoted by $\rho(\mathbf{A})$, is defined as the number of pivots in the reduced row echelon form of $\mathbf{A}$.

The proof of Theorem 6.3.1 is very algorithmic in nature and can indeed be made into an algorithm. Let us state the pseudo-code of an algorithm that computes a row echelon form of a matrix. Note how closely it follows the first part of the proof of Theorem 6.3.1. One could extend the algorithm and obtain pseudo-code of an algorithm that computes the reduced row echelon form of a matrix, but we will not do that.

Algorithm 9 for computing a row echelon form of a matrix

Input: Positive integers m, n and an $m \times n$ matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$
Output: $\text{ref}(\mathbf{A})$, the reduced row echelon form of $\mathbf{A}$

1: **if** $\mathbf{A} = \mathbf{0}$ **then**
2: $\quad \text{ref}(\mathbf{A}) \leftarrow \mathbf{0}$,
3: **if** $m = 1$ and $\mathbf{A} \neq \mathbf{0}$ **then**
4: $\quad j \leftarrow$ smallest column index such that $\mathbf{A}_{1j} \neq 0$
5: $\quad \text{ref}(\mathbf{A}) \leftarrow (\mathbf{A}_{1j})^{-1} \cdot \mathbf{A}$
6: **if** $m > 1$ and $\mathbf{A} \neq \mathbf{0}$ **then**
7: $\quad j \leftarrow$ least ℓ such that some row of $\mathbf{A}$ has non-zero ℓ-th entry
8: $\quad i \leftarrow$ least i such that the ith row of $\mathbf{A}$ has nonzero j-th entry
9: $\quad \mathbf{B} \leftarrow$ the matrix obtained from $\mathbf{A}$ by applying $R_1 \leftrightarrow R_i$
10: $\quad b \leftarrow$ the ith entry of the first row of $\mathbf{B}$
11: $\quad \mathbf{B} \leftarrow$ the matrix obtained from $\mathbf{B}$ by applying $R_1 \leftarrow b^{-1} \cdot R_1$
12: $\quad \mathbf{r} \leftarrow$ the 1-st row of $\mathbf{B}$
13: $\quad$ **for** $i = 2...m$ **do**
14: $\quad\quad b \leftarrow$ the first entry of the i-th row of $\mathbf{B}$
15: $\quad\quad \mathbf{B} \leftarrow$ the matrix obtained from $\mathbf{B}$ by applying $R_i \leftarrow R_i - bR_1$
16: $\quad \mathbf{C} \leftarrow$ the matrix obtained from $\mathbf{B}$ by deleting the first row
17: $\quad \mathbf{C} \leftarrow \text{ref}(\mathbf{C})$ (here the algorithm call itself recursively)
18: $\quad \text{ref}(\mathbf{A}) \leftarrow$ the matrix obtained by adding $\mathbf{r}$ on top of $\mathbf{C}$

In line 13 of the pseudo-code we have used what is known as a *for loop*. In this case, this just means that lines 14 and 15 are executed first for i equal to two, then for i equal to three, and so on up til the case where i is equal to m.

6.4 Computing all solutions to systems of linear equations

Up till now, we have usually written elements from $\mathbb{F}^n$ as n-tuples $(a_1, \ldots, a_n)$. It is quite common to identify $\mathbb{F}^n$ with $\mathbb{F}^{n \times 1}$, that is to say, to identify an n-tuple with an $n \times 1$ matrix. Such a matrix only contains one column. This means for example that:

$$(1, 2, 4, 7) \text{ is identified with } \begin{bmatrix} 1 \\ 2 \\ 4 \\ 7 \end{bmatrix}.$$

A small warning is in place. Even though, we will always identify $\mathbb{F}^n$ and $\mathbb{F}^{n \times 1}$, some books prefer to identify $\mathbb{F}^n$ and $\mathbb{F}^{1 \times n}$.

When performing elementary row operations, we have at times multiplied rows of a matrix with an element c from $\mathbb{F}$ or added one row to another. A similar operation can be performed

on columns in a matrix. In particular, it is customary to define

$$c \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} c \cdot a_1 \\ \vdots \\ c \cdot a_n \end{bmatrix} \text{ and } \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} a'_1 \\ \vdots \\ a'_n \end{bmatrix} = \begin{bmatrix} a_1 + a'_1 \\ \vdots \\ a_n + a'_n \end{bmatrix}.$$

This notation, combined with the theory of reduced row echelon matrices, will make it possible to determine whether or not a given system of linear equations has solutions, and if yes, to write all solutions down in a systematic way. Let us start with determining when a system has a solution.

Theorem 6.4.1

Let a system of m linear equations in n variables over a field $\mathbb{F}$ be given. Denote by $\mathbf{A}$ the coefficient matrix of the system and by $[\mathbf{A}|\mathbf{b}]$ its augmented matrix. Then the system has no solution if $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ do not have the same rank.

Proof. We know from Theorem 6.3.1, that there exists a sequence of elementary row operations that brings the matrix $\mathbf{A}$ in its row reduced echelon form, say $\hat{\mathbf{A}}$. Since the first n columns of the augmented matrix $[\mathbf{A}|\mathbf{b}]$ are identical with those of the coefficient matrix $\mathbf{A}$, applying exactly the same elementary row operations on $[\mathbf{A}|\mathbf{b}]$ yields a matrix, say $\mathbf{B}$, whose first n columns are identical with those of the reduced row echelon form of $\mathbf{A}$. Therefore we can write $\mathbf{B} = [\hat{\mathbf{A}}|\hat{\mathbf{b}}]$ for some $\hat{\mathbf{b}} \in \mathbb{F}^m$. Let us denote the bottom entry $\hat{\mathbf{b}}$ by $\hat{b}_m$. If the bottom row of $\hat{\mathbf{A}}$ contains a pivot, then the matrix $[\hat{\mathbf{A}}|\hat{\mathbf{b}}]$ is in reduced row echelon form. But then we see that the matrices $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ have the same rank, contrary to the assumption given in the theorem that $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ do not have the same rank. Therefore we may assume that the bottom row of $\hat{\mathbf{A}}$ does not contain a pivot, which simply means that this row is the zero row. If the last row of $\hat{\mathbf{A}}$ does not contain a pivot and $\hat{b}_m = 0$, then the matrix $[\hat{\mathbf{A}}|\hat{\mathbf{b}}]$ is in reduced row echelon form and we can conclude that $\rho(\mathbf{A}) = \rho([\mathbf{A}|\mathbf{b}])$, again leading to a contradiction. Therefore we may assume that the bottom row of $\hat{\mathbf{A}}$ does not contain a pivot and that $\hat{b}_m \neq 0$. But then the bottom row of the matrix $[\hat{\mathbf{A}}|\hat{\mathbf{b}}]$ corresponds to the equation $0 \cdot x_1 + \cdots 0 \cdot x_m = \hat{b}_m$. Since this equation has no solution, Theorem 6.2.1 implies that the system we started with has no solution either. □

Example 6.4.1

As in Example 6.1.6, consider the following system of two linear equations in two variables over $\mathbb{R}$:

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 0 \end{cases}.$$

We have already seen in Example 6.1.6 that this system has no solutions. Let us now try to confirm this using Theorem 6.4.1. The augmented matrix $[\mathbf{A}|\mathbf{b}]$ is given by

$$[\mathbf{A}|\mathbf{b}] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Applying the row operation $R_1 \leftrightarrow R_2$ followed by $R_2 \leftarrow R_2 - R_1$, we find the reduced row echelon form of the augmented matrix:

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

Hence $\rho([\mathbf{A}|\mathbf{b}]) = 2$. The reduced row echelon form of the coefficient matrix is the matrix

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

which can be obtained from $\mathbf{A}$ by applying the operation $R_2 \leftarrow R_2 - R_1$. Hence $\rho(\mathbf{A}) = 1$. Since $\rho(\mathbf{A}) \neq \rho([\mathbf{A}|\mathbf{b}])$, Theorem 6.4.1 implies that indeed the system we started with does not have a solution.

In case $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ do have the same rank, we can use the theory of reduced row echelon matrices, to describe a solution explicitly. Let us look at a concrete example first.

Example 6.4.2

Let us consider a system of three linear equations in four variables over $\mathbb{R}$, whose augmented matrix already is in reduced row echelon form:

$$\begin{cases} x_1 + 2 \cdot x_2 + 3 \cdot x_4 = 5 \\ x_3 + 4 \cdot x_4 = 6 \\ 0 = 0 \end{cases}.$$

We can see that in this case the coefficient matrix $\mathbf{A}$ and augmented matrix $[\mathbf{A}|\mathbf{b}]$ are

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \text{ respectively } [\mathbf{A}|\mathbf{b}] = \begin{bmatrix} 1 & 2 & 0 & 3 & 5 \\ 0 & 0 & 1 & 4 & 6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since both are already in reduced row echelon form, we can immediately determine the ranks of these matrices and conclude that $\rho(\mathbf{A}) = \rho([\mathbf{A}|\mathbf{b}]) = 2$. Theorem 6.4.1 does therefore not apply, and we cannot conclude anything about the existence of solutions yet. However, a solution is easily determined in the following way: first rewrite the equations in the following way:

$$\begin{cases} x_1 = 5 - 2 \cdot x_2 - 3 \cdot x_4 \\ x_3 = 6 - 4 \cdot x_4 \end{cases}.$$

Now we can choose $x_2 = v_2$ and $x_4 = v_4$ as we want for any $v_2, v_4 \in \mathbb{R}$ and then compute the

resulting values for x_1 and x_3. For example, choosing $v_2 = v_4 = 0$, we find the solution

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 6 \\ 0 \end{bmatrix}.$$

Exactly the same approach can be used in general to find a solution to a system of linear equations, provided that coefficient and augmented matrix have the same rank. The result is the following:

Theorem 6.4.2

Let a system of m linear equations in n variables over a field $\mathbb{F}$ be given. Denote by $\mathbf{A}$ the coefficient matrix of the system and by $[\mathbf{A}|\mathbf{b}]$ its augmented matrix and suppose that these matrices have the same rank ρ. Moreover, assume that the pivots of the reduced row echelon form of $\mathbf{A}$ are at the positions $(1, j_1), \ldots, (\rho, j_\rho)$, and that the top ρ entries of the last column of the reduced row echelon form of $[\mathbf{A}|\mathbf{b}]$ are given by $\hat{b}_1, \ldots, \hat{b}_\rho$. Then the m-tuple $(v_1, \ldots, v_n)$ defined as

$$v_j = \begin{cases} \hat{b}_\ell & \text{if } j = j_\ell \text{ for some } \ell = 1, \ldots, \rho, \\ 0 & \text{otherwise.} \end{cases}$$

is a possible solution to the system.

Proof. The idea of the proof is simply to generalize the approach used in Example 6.4.2. First of all, we use the equations corresponding to the rows of the reduced row echelon form of the augmented matrix $[\mathbf{A}|\mathbf{b}]$ to express the variables x_j with $j \in \{j_1, \ldots, j_\ell\}$ in terms of the remaining $n - \rho$ variables. Then putting all these remaining variables $x_j, j \notin \{j_1, \ldots, j_\ell\}$ equal to zero, we find that $x_j = \hat{b}_\ell$ for $j = j_\ell$ and $\ell = 1, \ldots, \rho$. Hence the n-tuple $(v_1, \ldots, v_n)$ is indeed a solution to the system whose augmented matrix is the reduced row echelon form of $[\mathbf{A}|\mathbf{b}]$. Now applying Theorem 6.2.1, we see that this n-tuple is also a solution to the system we started with. □

Theorem 6.4.2 does by no means state that the indicated solution is the only solution. Indeed, we know from Theorem 6.1.2 that there can be more. Recall that a solution to an inhomogeneous system of linear equations was called a particular solution. If the system of linear equations in inhomogeneous, Theorem 6.4.2 therefore gives such a particular solution, provided it exists.

Corollary 6.4.3

Let a system of m linear equations in n variables over a field $\mathbb{F}$ be given. Denote by $\mathbf{A}$ the coefficient matrix of the system and by $[\mathbf{A}|\mathbf{b}]$ its augmented matrix. Then the system has no solution if and only if $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ do not have the same rank.

Proof. The "if" part is precisely Theorem 6.4.1. In other words, we have already seen in Theorem 6.4.1 that if $\rho(\mathbf{A}) \neq \rho([\mathbf{A}|\mathbf{b}])$, then the system has no solutions. Conversely, if $\rho(\mathbf{A}) = \rho([\mathbf{A}|\mathbf{b}])$, then Theorem 6.4.2 implies that the system does have at least one solution. □

With Corollary 6.4.3 we can determine exactly if a given system of linear equations has a solution. Moreover, using Theorem 6.4.2, we can determine at least one solution if such solutions exist. Now recall that in Theorem 6.1.2, we have seen that in order to find all solutions of an inhomogeneous system of linear equations, it is enough to find all solutions of the corresponding homogeneous system of linear equations and one particular solution of the inhomogeneous system. Therefore, what is left to do, is to describe how one finds all solutions to a homogeneous system of linear equations. This is precisely the aim of the next theorem, but let us look at an example first to get the idea.

Example 6.4.3

Let us consider a system of three linear equations in four variables over $\mathbb{R}$, whose augmented matrix already is in reduced row echelon form:

$$\begin{cases} x_1 + 2 \cdot x_2 + 3 \cdot x_4 = 0 \\ x_3 + 4 \cdot x_4 = 0 \\ 0 = 0 \end{cases}.$$

This system is similar to the system of linear equation we studied in Example 6.4.2, but this time it is homogeneous. In particular, the coefficient matrices of the system above and the system from Example 6.4.2 are the same and as observed in Example 6.4.2, it is in reduced row echelon form.

It is not hard to find all solutions to the system. Since the coefficient matrix of the system is in reduced row echelon form with pivots in the first and third column, we can express x_1 and x_3 in terms of x_2 and x_4. More concretely, we can rewrite the equations as

$$\begin{cases} x_1 = -2 \cdot x_2 - 3 \cdot x_4 \\ x_3 = -4 \cdot x_4 \end{cases}.$$

Hence any solution $(v_1, v_2, v_3, v_4) \in \mathbb{R}^4$ to the system satisfies

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} -2 \cdot v_2 - 3 \cdot v_4 \\ v_2 \\ -4 \cdot v_4 \\ v_4 \end{bmatrix} = v_2 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + v_4 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}.$$

Therefore, we can think of $v_2, v_4 \in \mathbb{R}$ as parameters that we can choose arbitrarily, each choice giving us a solution to the system of linear equations we started with. Changing notation

from v_2 to t_1 and v_4 to t_2, we see that any solution to the system is of the form

$$t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \quad (t_1, t_2 \in \mathbb{R})$$

Conversely, since a direct check shows that

$$\begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}$$

are solution to the system, Theorem 6.1.1 implies that for any $t_1, t_2 \in \mathbb{R}$, the expression

$$t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}$$

is also a solution. Putting this together, we see that the solutions to the homogeneous system of linear equations we started are precisely those $(v_1, v_2, v_3, v_4) \in \mathbb{R}^4$ such that

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \quad (t_1, t_2 \in \mathbb{R}).$$

One calls such a description of the solutions, the *general solution* of the homogeneous system. The solution set to the homogeneous system of linear equations

$$\begin{cases} x_1 + 2 \cdot x_2 + 3 \cdot x_4 = 0 \\ x_3 + 4 \cdot x_4 = 0 \\ 0 = 0 \end{cases}$$

is precisely given by

$$\left\{ t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \mid t_1, t_2 \in \mathbb{R} \right\}.$$

In this example, we started out with a homogeneous system of linear equations whose

coefficient matrix was in reduced row echelon form. This was the reason that we could determine all solutions relatively fast. From the previous sections, we know however that even if we start with a more complicated system, we can always use elementary row operations to transform it in such a way that the resulting coefficient matrix is in reduced echelon form. Basically, Example 6.4.3 describes how to find all solutions, once the coefficient matrix of the system of linear equations is in reduced row echelon form. Exactly the same ideas work for any homogeneous system of linear equations: first simplify the system by bringing its coefficient matrix in reduced row echelon form, then follow the procedure exemplified in Example 6.4.3. It is possible to describe the outcome for the general case and for the sake of completeness we do so in the following theorem. However, when asked to solve a homogeneous system of linear equations in practice, it is often easier not to use this theorem, but instead to use a procedure similar to the one in Example 6.4.3 directly.

Theorem 6.4.4

Let a homogeneous system of m linear equation in n variables over a field $\mathbb{F}$ be given. Denote the coefficient matrix of this system by $\mathbf{A}$ and let $\hat{\mathbf{A}}$ denote the reduced row echelon form of $\mathbf{A}$. Further, suppose that $\hat{\mathbf{A}}$ has ρ pivots in columns $j_1, \ldots, j_\rho$, and denote by

$$\mathbf{c}_1 = \begin{bmatrix} c_{11} \\ \vdots \\ c_{m1} \end{bmatrix}, \ldots, \mathbf{c}_{n-\rho} = \begin{bmatrix} c_{1\,n-\rho} \\ \vdots \\ c_{m\,n-\rho} \end{bmatrix}$$

the $n-\rho$ columns of $\hat{\mathbf{A}}$ not containing a pivot. Finally, define

$$\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ \vdots \\ v_{n1} \end{bmatrix}, \ldots, \mathbf{v}_{n-\rho} = \begin{bmatrix} v_{1\,n-\rho} \\ \vdots \\ v_{n\,n-\rho} \end{bmatrix}$$

by

$$v_{ji} = \begin{cases} -c_{\ell i} & \text{if } j = j_\ell \text{ for some } \ell = 1, \ldots, \rho, \\ 1 & \text{if } \mathbf{c}_i \text{ is the } j\text{-th column in } \hat{\mathbf{A}}, \\ 0 & \text{otherwise.} \end{cases}$$

Then the solution set of the given homogeneous system of linear equations is given by

$$\left\{ t_1 \cdot \begin{bmatrix} v_{11} \\ \vdots \\ v_{n1} \end{bmatrix} + \cdots + t_{n-\rho} \cdot \begin{bmatrix} v_{1\,n-\rho} \\ \vdots \\ v_{n\,n-\rho} \end{bmatrix} \mid t_1, \ldots, t_{n-\rho} \in \mathbb{F} \right\}.$$

Proof. We will not prove this theorem, but only indicate the idea of the proof. First of all Theorem 6.2.1 is used to conclude that the homogeneous system with coefficient matrix $\mathbf{A}$ has exactly the same solution set as the homogeneous system with coefficient matrix $\hat{\mathbf{A}}$. Then

the same approach as in Example 6.4.3 is used to describe all solutions to the homogeneous system with coefficient matrix $\hat{\mathbf{A}}$. □

The expression

$$t_1 \cdot \begin{bmatrix} v_{11} \\ \vdots \\ v_{n1} \end{bmatrix} + \cdots + t_{n-\rho} \cdot \begin{bmatrix} v_{1\,n-\rho} \\ \vdots \\ v_{n\,n-\rho} \end{bmatrix} \quad (t_1, \ldots, t_{n-\rho} \in \mathbb{F})$$

is called the *general solution* of the homogeneous system with coefficient matrix $\mathbf{A}$. Looking back at Example 6.4.3, we see that the general solution of the homogeneous system of linear equations studied in that example was shown to be equal to

$$t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \quad (t_1, t_2 \in \mathbb{R}).$$

Corollary 6.4.5

Let a homogeneous system of m linear equation in n variables over a field $\mathbb{F}$ be given. Denote the coefficient matrix of this system by $\mathbf{A}$. Then the homogeneous system has only the all-zero tuple $(0, \ldots, 0) \in \mathbb{F}^n$ as solution if and only if $\rho(\mathbf{A}) = n$.

Proof. Theorem 6.4.4 implies that if the rank of $\mathbf{A}$ is less than n, then there exists a nonzero solution. Conversely, if the rank of $\mathbf{A}$ is equal to n, the number of parameters t_i in the description of the solution set in Theorem 6.4.4, is zero. This means that only the all-zero tuple $(0, \ldots, 0)$ is a solution. □

The status is now that we can determine all solutions to any homogeneous system of linear equations (which we called the general solution of the homogeneous system), can determine whether or not an inhomogeneous system has a solution, and find such a solution (which we called a particular solution) if it does. Hence using Theorem 6.1.2, we can in this case also determine a formula describing all solutions to an inhomogeneous system of linear equations: it is simply the sum of a particular solution and the general solution of the corresponding homogeneous system. This sum is called the *general solution* of the inhomogeneous system. Therefore we have answered in a constructive way all three questions posed at the end of Section 6.1.

Let us finish this section with an example, where we compute the general solution of an inhomogeneous system of linear equations.

Example 6.4.4

Let us return to the inhomogeneous system of linear equations considered in Example 6.4.2:

$$\left\{\begin{array}{rcrcrcl} x_1 & + & 2\cdot x_2 & + & & 3\cdot x_4 & = & 5 \\ & & & & x_3 & + 4\cdot x_4 & = & 6 \\ & & & & & 0 & = & 0 \end{array}\right. .$$

We have computed a particular solution in Example 6.4.2 and the general solution of the corresponding homogeneous system in Example 6.4.3. Using these previous calculations in combination with Theorem 6.1.2, we conclude that the general solution of the inhomogeneous system is given by:

$$\begin{bmatrix} 5 \\ 0 \\ 6 \\ 0 \end{bmatrix} + t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \quad (t_1, t_2 \in \mathbb{R}).$$

The solution set of the inhomogeneous system is therefore:

$$\left\{ \begin{bmatrix} 5 \\ 0 \\ 6 \\ 0 \end{bmatrix} + t_1 \cdot \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} -3 \\ 0 \\ -4 \\ 1 \end{bmatrix} \mid t_1, t_2 \in \mathbb{R} \right\}.$$

6.5 Uniqueness of the reduced row echelon form

Previously, we have stated that a given matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$ has a unique reduced row echelon form. Existence was shown in Theorem 6.3.1 and in this section we want to show uniqueness. This section can be skipped and is only meant for the reader who wants to see a proof of the uniqueness of the reduced row echelon form.

Theorem 6.5.1

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ a matrix. Suppose that $\mathbf{A}$ can be transformed using a sequence of elementary row operations to a matrix $\mathbf{B}_1$ in reduced row echelon form, but using another sequence of elementary row operations to a matrix $\mathbf{B}_2$ in reduced row echelon form. Then $\mathbf{B}_1 = \mathbf{B}_2$.

Proof. From Theorem 6.2.1, we know that the homogeneous systems of linear equations with coefficient matrices $\mathbf{A}$, $\mathbf{B}_1$, and $\mathbf{B}_2$ all have exactly the same solutions. The idea of the proof is to show that the homogeneous systems of linear equations with coefficient matrices $\mathbf{B}_1$

and $\mathbf{B}_2$ only can have the same solutions if $\mathbf{B}_1 = \mathbf{B}_2$. Moreover, we use induction on n, the number of columns.

Let us start with the induction basis. If $n = 1$, there are only two possible reduced row echelon forms: the $m \times 1$ matrices

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The first can only be a reduced row echelon form of $\mathbf{A}$, if $\mathbf{A}$ was the zero $m \times 1$ matrix to begin with. Performing any elementary row operation on the zero matrix, results in the zero matrix again. Hence if $\mathbf{B}_1$ or $\mathbf{B}_2$ is the zero matrix, then $\mathbf{A} = \mathbf{B}_1 = \mathbf{B}_2$, since they are all equal to the zero matrix. Now suppose that $\mathbf{B}_1$ or $\mathbf{B}_2$ is equal to the second possible $m \times 1$ reduced row echelon matrix. If $\mathbf{B}_1 \neq \mathbf{B}_2$, then at least one of them is equal to the only other $m \times 1$ reduced row echelon form matrix, namely the zero matrix. But we have just seen that this would imply that both $\mathbf{B}_1$ and $\mathbf{B}_2$ are equal to the zero matrix. This contradiction shows that if $\mathbf{B}_1$ or $\mathbf{B}_2$ is equal to the second $m \times 1$ reduced row echelon matrix, then $\mathbf{B}_1 = \mathbf{B}_2$.

We continue to the induction step. Assume $n > 1$ and that the theorem is true for $n - 1$. For any $m \times n$ matrix $\mathbf{A}$, let us denote by $\mathbf{A}|_{n-1}$, the $m \times (n-1)$ matrix one obtains by removing the final column of $\mathbf{A}$. The induction hypothesis implies that $\mathbf{A}|_{n-1}$ has a unique reduced row echelon form. Moreover, if $\mathbf{B}$ is an $m \times n$ matrix in reduced row echelon form, then also the matrix $\mathbf{B}|_{n-1}$ is in reduced row echelon form. This implies that if $\mathbf{B}_1$ and $\mathbf{B}_2$ are two possible reduced row echelon forms of $\mathbf{A}$, then the induction hypothesis implies that $\mathbf{B}_1|_{n-1} = \mathbf{B}_2|_{n-1}$. In other words: the first $n-1$ columns of $\mathbf{B}_1$ $\mathbf{B}_2$ are identical. Only the n-th (i.e., the last) columns may be distinct. Now denote by ρ the number of pivots occurring in $\mathbf{B}_1|_{n-1}$. If the n-th column of $\mathbf{B}_1$ contains a pivot, this column contains zeros only, except in the $(\rho+1)$-th position, where it contains a one. Hence any solution $(v_1, \ldots, v_n) \in \mathbb{F}^n$ to the homogeneous system of linear equations with coefficient matrix $\mathbf{B}_1$, satisfies $v_n = 0$. Conversely, using Theorem 6.4.4, if the n-th column of $\mathbf{B}_1$ does not contain a pivot, there exists a solution $(v_1, \ldots, v_n)$ such that $v_n = 1$. A similar reasoning applies to the last column of $\mathbf{B}_2$. Using Theorem 6.2.1, we can however conclude that the homogeneous systems of linear equations with coefficient matrices $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{A}$ all have exactly the same solution sets. It follows that either a pivot occurs in the n-th columns of both $\mathbf{B}_1$ and $\mathbf{B}_2$, or that no pivot occurs in the n-th columns of both $\mathbf{B}_1$ and $\mathbf{B}_2$. In the first case, we have already seen that the n-th columns are completely determined, implying that $\mathbf{B}_1 = \mathbf{B}_2$. In the second case, we can conclude that there is exactly one solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$ that has a zero in all variables corresponding to the columns not containing pivots, except in the n-th column, where it has a one. Using Theorem 6.4.4, one sees that the coefficients of this solution completely determine the n-th column of a reduced row echelon form of $\mathbf{A}$. We conclude that $\mathbf{B}_1 = \mathbf{B}_2$ also in the second case where no pivot occurs in the n-th columns of both $\mathbf{B}_1$ and $\mathbf{B}_2$. □

Chapter 7

Vectors and matrices

7.1 Vectors

As in the last chapter, we will denote by $\mathbb{F}$ a field. What we will explain works over any field, but the reader can just think of $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. When describing solutions to systems of linear equations, we already worked with $\mathbb{F}^n$, the set of all n-tuples with entries in $\mathbb{F}$. Also, we already explained that such an n-tuple is for convenience often identified with an $n \times 1$ matrix. This just means that:

$$(v_1, \ldots, v_n) \text{ can also be written as } \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}.$$

When an n-tuple is written as an $n \times 1$ matrix, we say that the n-tuple is written in *vector* form. Elements in $\mathbb{F}^n$ are therefore called *vectors* with n entries from $\mathbb{F}$. If all entries of such a vector are zero, we call that vector the *zero vector* of $\mathbb{F}^n$.

Remark 7.1.1

Elements in $\mathbb{F}^{n\times 1}$ are sometimes called *column vectors*, while likewise elements from $\mathbb{F}^{1\times n}$ are called *row vectors*.

We have already used in the previous chapter that there is a natural way to add two vectors $\mathbf{v}$ and $\mathbf{w}$ from $\mathbb{F}^n$, and also that one can multiply a vector from $\mathbb{F}^n$ with an element $c \in \mathbb{F}$, often called a *scalar* in this context, since multiplying a vector by a constant can be thought of as scaling the vector. More precisely, addition of vectors in $\mathbb{F}^n$ is defined as:

$$\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} + \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ \vdots \\ v_n + w_n \end{bmatrix}. \tag{7.1}$$

For $\mathbb{F} = \mathbb{R}$ and $n = 2$, addition of two vectors is illustrated in Figure 7.1.

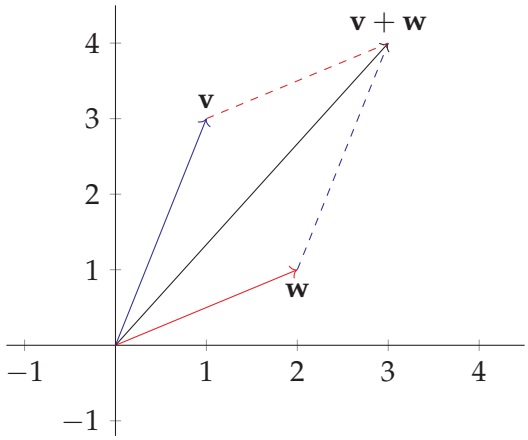


Figure 7.1: Addition of two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^2$.

The product of a scalar c from $\mathbb{F}$ with a vector in $\mathbb{F}^n$ is defined as:

$$c \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} c \cdot v_1 \\ \vdots \\ c \cdot v_n \end{bmatrix}. \tag{7.2}$$

Instead of $(-1) \cdot \mathbf{v}$ one can also write $-\mathbf{v}$. Similarly, an expression such as $\mathbf{v} + (-1) \cdot \mathbf{w}$ is often written as $\mathbf{v} - \mathbf{w}$. It is also typical to omit the multiplication sign between scalar and vector. In other words, an expression like $c\mathbf{v}$ should just be read as $c \cdot \mathbf{v}$. For $\mathbb{F} = \mathbb{R}$ and $n = 2$ scaling of a vector is illustrated in Figure 7.2.

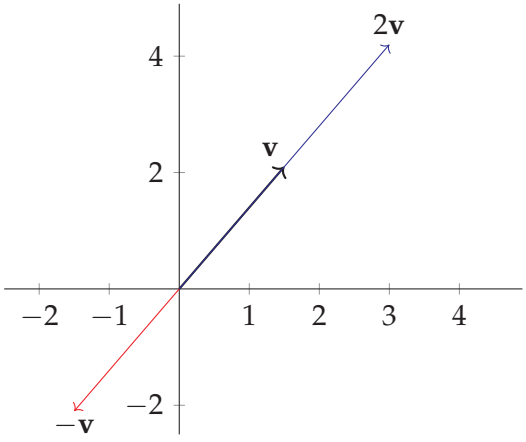


Figure 7.2: Scaling of a vector $\mathbf{v} \in \mathbb{R}^2$.

As in the case for matrices, we will often use boldface fonts for vectors and typically use letters such as $\mathbf{u}, \mathbf{v}, \mathbf{w}$. For future reference, we state the following theorem, which collects a number of properties of the vector addition and scalar multiplication:

Theorem 7.1.1

Let $\mathbb{F}$ be a field, $c, d \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{F}^n$. Then

(i) $(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$

(ii) $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$

(iii) $c \cdot (d \cdot \mathbf{u}) = (c \cdot d) \cdot \mathbf{u}$

(iv) $c \cdot (\mathbf{u}+\mathbf{v}) = c \cdot \mathbf{u} + c \cdot \mathbf{v}$

(v) $(c+d) \cdot \mathbf{u} = c \cdot \mathbf{u} + d \cdot \mathbf{u}$

We leave the proof of this theorem out.

Now that we have vectors at out disposal, we will be able to discuss further properties they have. We start with an example.

Example 7.1.1

Consider the vectors

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in \mathbb{R}^2.$$

(a) Compute $4 \cdot \mathbf{u} + 3 \cdot \mathbf{v}$.

(b) Find c and d such that $c \cdot \mathbf{u} + d \cdot \mathbf{v} = \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^2$.

Answer:

(a) Using the definition of scalar multiplication and vector addition, we find

$$4 \cdot \mathbf{u} + 3 \cdot \mathbf{v} = 4 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 3 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix} + \begin{bmatrix} 6 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 11 \end{bmatrix}.$$

(b) We have

$$c \cdot \mathbf{u} + d \cdot \mathbf{v} = c \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + d \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} c \\ 2c \end{bmatrix} + \begin{bmatrix} 2d \\ d \end{bmatrix} = \begin{bmatrix} c+2d \\ 2c+d \end{bmatrix}.$$

If we want the outcome to be the zero vector, this means that we need to solve the homogeneous system of linear equations:

$$\begin{cases} c+2d = 0 \\ 2c+d = 0 \end{cases}.$$

Now subtracting the first equation twice from the second equation, in other words performing the elementary row operation $R_2 \leftarrow R_2 - 2 \cdot R_1$, we obtain the system

$$\begin{cases} c + 2d = 0 \\ 0c - 3d = 0 \end{cases}.$$

We could continue and bring the system in reduced row echelon form, but it is already clear now that the only solution is $c = d = 0$.

An expression like $4 \cdot \mathbf{u} + 3 \cdot \mathbf{v}$ is called a *linear combination* of the vectors $\mathbf{u}$ and $\mathbf{v}$. More general, given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ and scalars $c_1, \ldots, c_n \in \mathbb{F}$, an expression of the form

$$c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$$

is called a linear combination of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. The second part of the example implies that apparently the only linear combination of the vectors $\mathbf{u}$ and $\mathbf{v}$ given there that is equal to the zero vector, is the linear combination $0 \cdot \mathbf{u} + 0 \cdot \mathbf{v}$. In general, a sequence of vectors can have this property. This is captured in the following:

Definition 7.1.1

A sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ is called *linearly independent* if and only if the equation $c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n = \mathbf{0}$ can only hold if $c_1 = \cdots = c_n = 0$.

If the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ is not linearly independent, one says that it is *linearly dependent*.

In other words, a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ is linearly independent if and only if the only linear combination of the vectors that is equal to the zero vector, occurs for $c_1 = \cdots = c_n = 0$. Using some logical expressions, linear independence of a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ can be phrased as follows:

$$\text{for all } c_1, \ldots, c_n \in \mathbb{F} \text{ one has: } c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n = \mathbf{0} \Rightarrow c_1 = \cdots = c_n = 0. \tag{7.3}$$

Similarly, linear dependence of the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ can be phrased in the following way:

$$\text{there exist } c_1, \ldots, c_n \in \mathbb{F} \text{ such that: } c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n = \mathbf{0} \ \wedge \ \text{not all } c_i \text{ are zero.} \tag{7.4}$$

Instead of saying that a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly (in)dependent, it is also quite common to simply say that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly (in)dependent. We will use this way of phrasing things quite often.

Example 7.1.2

The sequence of vectors consisting of

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \in \mathbb{R}^2$$

is linearly dependent. Indeed, since $\mathbf{v} = 2 \cdot \mathbf{u}$, we see that $(-2) \cdot \mathbf{u} + \mathbf{v} = \mathbf{0}$.

This example illustrates a more general principle: two vectors $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent if and only if one is a scalar multiple of the other. Indeed, if for example $\mathbf{u} = c \cdot \mathbf{v}$, then $1 \cdot \mathbf{u} + (-c) \cdot \mathbf{v} = \mathbf{0}$, showing that the vectors are linearly dependent. Similarly, if $\mathbf{v} = c \cdot \mathbf{u}$, then $(-c) \cdot \mathbf{u} + 1 \cdot \mathbf{v} = \mathbf{0}$, again showing that the vectors are linearly dependent. Conversely if the vectors are linearly dependent, there exist $c, d \in \mathbb{F}$, not both zero, such that $c \cdot \mathbf{u} + d \cdot \mathbf{v} = \mathbf{0}$. If $c \neq 0$, then we obtain that $\mathbf{u} = (-d/c) \cdot \mathbf{v}$ so that $\mathbf{v}$ is a scalar multiple of $\mathbf{u}$. If $d \neq 0$, we similarly obtain that $\mathbf{v} = (-c/d) \cdot \mathbf{u}$ showing that in that case $\mathbf{u}$ is a scalar multiple of $\mathbf{v}$. Hence intuitively, one can say that two vectors $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent if and only if there is a line through the origin containing both $\mathbf{u}$ and $\mathbf{v}$.

Example 7.1.3

The sequence of vectors consisting of

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in \mathbb{R}^2$$

is linearly independent. Indeed, we have seen in Example 7.1.1 that the equation $c \cdot \mathbf{u} + d \cdot \mathbf{v} = \mathbf{0}$ implies that $c = d = 0$.

This example suggests that the linear independence of a sequence of vectors can be investigated using the theory of systems of linear equations. This is indeed the case and the general result is the following:

Lemma 7.1.2

Let vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ be given and let $\mathbf{A} \in \mathbb{F}^{m \times n}$ be the $m \times n$ matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$, that is

$$\mathbf{A} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \ldots & \mathbf{v}_n \\ | & & | \end{bmatrix}.$$

The sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly independent if and only if the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$ only has the zero vector $\mathbf{0} \in \mathbb{F}^n$ as solution.

Proof. First suppose that the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly independent and let $(c_1, \ldots, c_n) \in \mathbb{F}^n$ be a solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$. This system can directly be rewritten as the equation $c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n = \mathbf{0}$. Using that we assumed that the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly independent, we see that $(c_1, \ldots, c_n) = (0, \ldots, 0)$.

Now conversely, assume that the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$ only has the zero vector $\mathbf{0} \in \mathbb{F}^n$ as solution. If $(c_1, \ldots, c_n) \in \mathbb{F}^n$ satisfies

$c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n = \mathbf{0}$, then we can immediately conclude that $(c_1, \ldots, c_n)$ is also a solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$. But then by assumption, we may conclude that $(c_1, \ldots, c_n) = (0, \ldots, 0)$. □

This lemma leads to a short characterisation of linear independence:

Theorem 7.1.3

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{F}^m$ be given and let $\mathbf{A} \in \mathbb{F}^{m \times n}$ be the matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$. The sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly independent if and only if the matrix $\mathbf{A}$ has rank n.

Proof. This follows from Corollary 6.4.5 and Lemma 7.1.2. □

Example 7.1.4

Consider the following three vectors in $\mathbb{C}^3$:

$$\mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 1+i \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 0 \\ 1+i \\ 0 \end{bmatrix}, \text{ and } \mathbf{w} = \begin{bmatrix} 1+i \\ -1+5i \\ 2i \end{bmatrix}.$$

(a) Are the vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ linearly independent?

(b) Are the vectors $\mathbf{u}$, $\mathbf{v}$ linearly independent?

(c) Is the vector $\mathbf{u}$ linearly independent?

Answer: The general strategy for this type of questions is to use Theorem 7.1.3. Recall that in order to compute the rank of a matrix, it is by Definition 6.3.2, the definition of the rank of a matrix, enough to compute its reduced row echelon form. Now let us answer the three questions, one at the time.

(a) Theorem 7.1.3 implies that to find the answer, we should determine the rank of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1+i \\ 0 & 1+i & -1+5i \\ 1+i & 0 & 2i \end{bmatrix}.$$

We have

$$\begin{bmatrix} 1 & 0 & 1+i \\ 0 & 1+i & -1+5i \\ 1+i & 0 & 2i \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 - (1+i) \cdot R_1} \begin{bmatrix} 1 & 0 & 1+i \\ 0 & 1+i & -1+5i \\ 0 & 0 & 0 \end{bmatrix}$$

$$\xrightarrow[R_2 \leftarrow (1+i)^{-1} \cdot R_2]{} \begin{bmatrix} 1 & 0 & 1+i \\ 0 & 1 & 2+3i \\ 0 & 0 & 0 \end{bmatrix}.$$

We can conclude that $\rho(\mathbf{A}) = 2$, which is less than three, the number of vectors we are considering. Hence the vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are linearly dependent.

(b) In this case, we should compute the rank of the matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1+i \\ 1+i & 0 \end{bmatrix}.$$

Using exactly the same elementary row operations as when solving the first questions, we find that the reduced row echelon form of $\mathbf{B}$ is the matrix

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

In particular, $\rho(\mathbf{B}) = 2$, which is equal to the number of vectors we are considering. Hence the vectors $\mathbf{u}$, $\mathbf{v}$ are linearly independent.

(c) If we only consider the vector $\mathbf{u}$, we need to determine the rank of the matrix

$$\mathbf{C} = \begin{bmatrix} 1 \\ 0 \\ 1+i \end{bmatrix}.$$

This matrix has rank one, since the one column this matrix has, is not the zero column. We can conclude that the sequence consisting of the vector $\mathbf{u}$ is linearly independent. In general, a sequence consisting of only one vector $\mathbf{u} \in \mathbb{F}^m$ is linearly independent if and only if $\mathbf{u} \neq \mathbf{0}$.

7.2 Matrices and vectors

When studying systems of linear equations, we introduced the notion of a matrix. A matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$ was introduced as a rectangular scheme containing $m \times n$ elements from a given field $\mathbb{F}$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}.$$

Sometimes one just writes $\mathbf{A} = [a_{ij}]_{1\leq i\leq m, 1\leq j\leq m}$ for brevity. When a matrix is given in this form, the element a_{ij}, sometimes also written as $a_{i,j}$, is the entry in row i and column j of the matrix $\mathbf{A}$. It is also common to denote this entry by $\mathbf{A}_{ij}$ or $\mathbf{A}_{i,j}$. The matrix $\mathbf{A}$ given above has m rows: $[a_{i1} \ \dots \ a_{in}]$ for $i = 1, \dots, m$ and n columns:

$$\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} \text{ for } j = 1, \dots, n.$$

We will call rows of a matrix *row vectors* and similarly columns of a matrix *column vectors*.

It turns out to be extremely useful to be able to multiply a matrix with a vector. We define the following:

Definition 7.2.1

Let $\mathbf{A} = (a_{ij})_{1\leq i\leq m, 1\leq j\leq m} \in \mathbb{F}^{m\times n}$ be a matrix and $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{F}^n$ a vector. Then we define $\mathbf{A} \cdot \mathbf{v} \in \mathbb{F}^m$ as follows:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} a_{11} \cdot v_1 + \dots a_{1n} \cdot v_n \\ \vdots \\ a_{m1} \cdot v_1 + \dots a_{mn} \cdot v_n \end{bmatrix}$$

Note that we can not multiply any matrix with any vector. Their sizes have to "fit": the number of columns of the matrix has to be the same as the number of entries in the vector. If this is not the case, the corresponding matrix-vector multiplication is not defined.

Example 7.2.1

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 7 \\ -1 \\ -2 \end{bmatrix}.$$

Compute $\mathbf{A} \cdot \mathbf{v}$.

Answer: Using Definition 7.2.1, we find that:

$$\mathbf{A} \cdot \mathbf{v} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \cdot 7 + 2 \cdot (-1) + 3 \cdot (-2) \\ 4 \cdot 7 + 5 \cdot (-1) + 6 \cdot (-2) \end{bmatrix} = \begin{bmatrix} -1 \\ 11 \end{bmatrix}.$$

Note that the matrix vector product occurs very naturally when considering a system of linear equations. A system of linear equations

$$\begin{cases} a_{11} \cdot x_1 & + & \cdots & + & a_{1n} \cdot x_n & = & b_1 \\ & & \vdots & & & & \vdots \\ a_{m1} \cdot x_1 & + & \cdots & + & a_{mn} \cdot x_n & = & b_m \end{cases}$$

can be expressed as

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}. \tag{7.5}$$

Now that we have defined a matrix vector product, one may wonder if more generally, matrices can be multiplied with each other as well. The answer turns out to be yes, provided again that their sizes fit. More precisely, we can do the following:

Definition 7.2.2

Let $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{B} \in \mathbb{F}^{n\times \ell}$. Suppose that the columns of $\mathbf{B}$ are given by $\mathbf{b}_1, \ldots, \mathbf{b}_\ell \in \mathbb{F}^n$, that is to say, suppose that

$$\mathbf{B} = \begin{bmatrix} | & & | \\ \mathbf{b}_1 & \ldots & \mathbf{b}_\ell \\ | & & | \end{bmatrix}.$$

Then we define

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} | & & | \\ \mathbf{A} \cdot \mathbf{b}_1 & \ldots & \mathbf{A} \cdot \mathbf{b}_\ell \\ | & & | \end{bmatrix}.$$

Note that the matrix product $\mathbf{A} \cdot \mathbf{B}$ is defined only if the number of columns of $\mathbf{A}$ is the same as the number of rows of $\mathbf{B}$. If these numbers match, then $\mathbf{A} \cdot \mathbf{B}$ is a matrix with m rows and ℓ columns. In other words, if $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{B} \in \mathbb{F}^{n\times \ell}$, then $\mathbf{A} \cdot \mathbf{B} \in \mathbb{F}^{m\times \ell}$.

Another way to look at the definition of the matrix product is to give a formula for the entries of the product $\mathbf{A} \cdot \mathbf{B}$ one at the time. Let us say, that we want to find a formula for the (i, j)-th entry of the product, $(\mathbf{A} \cdot \mathbf{B})_{i,j}$, that is to say, the entry in row i and column j. This amounts to determining the i-th entry of the product $\mathbf{A} \cdot \mathbf{b}_j$, where $\mathbf{b}_j$ is the j-th column of $\mathbf{B}$. This in turn is exactly the same as the outcome of multiplying the i-th row of the matrix $\mathbf{A}$ with the j-th column of the matrix $\mathbf{B}$. Since the i-th row of $\mathbf{A}$ can be written as equals $[a_{i1} \ \ldots \ a_{in}]$ and the j column of $\mathbf{B}$ as

$$\mathbf{b}_j = \begin{bmatrix} b_{1j} \\ \vdots \\ b_{mj} \end{bmatrix},$$

we see that

$$(\mathbf{A} \cdot \mathbf{B})_{i,j} = [a_{i1} \ \ldots \ a_{in}] \cdot \begin{bmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{bmatrix} = a_{i1} \cdot b_{1j} + \cdots + a_{in} \cdot b_{nj}.$$

Using the summation symbol from Section 5.3, we can rewrite this formula as follows:

$$(\mathbf{A} \cdot \mathbf{B})_{i,j} = \sum_{r=1}^{n} a_{ir} \cdot b_{rj}. \tag{7.6}$$

Example 7.2.2

In this example, let $\mathbb{F} = \mathbb{R}$ and write

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 7 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}.$$

Compute, if possible, the matrix products $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{B} \cdot \mathbf{A}$.

Answer: First consider the matrix product $\mathbf{A} \cdot \mathbf{B}$. Since $\mathbf{A} \in \mathbb{R}^{2\times 3}$ and $\mathbf{B} \in \mathbb{R}^{3\times 3}$, the product $\mathbf{A} \cdot \mathbf{B}$ is defined. We have already computed the product of $\mathbf{A}$ and the first column of $\mathbf{B}$ in Example 7.2.1, so we will not repeat those computations. Taking that into account, we obtain that:

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 7 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 1\cdot 0+2\cdot 1+3\cdot 0 & 1\cdot 0+2\cdot 0+3\cdot 1 \\ 11 & 4\cdot 0+5\cdot 1+6\cdot 0 & 4\cdot 0+5\cdot 0+6\cdot 1 \end{bmatrix}$$
$$= \begin{bmatrix} -1 & 2 & 3 \\ 11 & 5 & 6 \end{bmatrix}.$$

Now let us consider the matrix product $\mathbf{B} \cdot \mathbf{A}$. Since $\mathbf{B}$ has three columns and $\mathbf{A}$ has two rows, the matrix product $\mathbf{B} \cdot \mathbf{A}$ is not defined.

This example shows that in general $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$. In other words, matrix multiplication is not commutative. In fact, as we have just seen, it may even happen that one of the products is not defined. Even if both products are defined, the order of the matrices still matters and $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$ in general. Consider for example $\mathbf{A} = [1\,0]$ and $\mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Then

$$\mathbf{A} \cdot \mathbf{B} = [1\,0] \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1\cdot 0 + 0 \cdot 1 = 0 \text{ and } \mathbf{B} \cdot \mathbf{A} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot [1\,0] = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Let us define addition of matrices as well.

Definition 7.2.3

Let $\mathbf{A}, \mathbf{A}' \in \mathbb{F}^{m\times n}$ be given, say

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \text{ and } \mathbf{A}' = \begin{bmatrix} a'_{11} & \dots & a'_{1n} \\ \vdots & & \vdots \\ a'_{m1} & \dots & a'_{mn} \end{bmatrix}.$$

Then we define $\mathbf{A} + \mathbf{A}'$ as follows:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} a'_{11} & \dots & a'_{1n} \\ \vdots & & \vdots \\ a'_{m1} & \dots & a'_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + a'_{11} & \dots & a_{1n} + a'_{1n} \\ \vdots & & \vdots \\ a_{m1} + a'_{m1} & \dots & a_{mn} + a'_{mn} \end{bmatrix}.$$

Addition of matrices is only defined if they have the same sizes. On the level of entries, we can see that $(\mathbf{A} + \mathbf{A}')_{i,j} = a_{ij} + a'_{ij}$. Addition and multiplication of matrices satisfy many similar rules as summation and multiplication of real or complex numbers. We collect some in the following theorem. The main exception, as already mentioned before, is that matrix multiplication is not commutative in general.

Theorem 7.2.1

Let $\mathbb{F}$ be a field. Then

(i) $\mathbf{A} + \mathbf{A}' = \mathbf{A}' + \mathbf{A}$ for all $\mathbf{A}, \mathbf{A}' \in \mathbb{F}^{m \times n}$.

(ii) $(\mathbf{A} + \mathbf{A}') + \mathbf{A}'' = \mathbf{A} + (\mathbf{A}' + \mathbf{A}'')$ for all $\mathbf{A}, \mathbf{A}', \mathbf{A}'' \in \mathbb{F}^{m \times n}$.

(iii) $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$ for all $\mathbf{A} \in \mathbb{F}^{m \times n}$, $\mathbf{B} \in \mathbb{F}^{n \times \ell}$, and $\mathbf{C} \in \mathbb{F}^{\ell \times k}$.

(iv) $\mathbf{A} \cdot (\mathbf{B} + \mathbf{B}') = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{B}'$ for all $\mathbf{A} \in \mathbb{F}^{m \times n}$ and $\mathbf{B}, \mathbf{B}' \in \mathbb{F}^{n \times \ell}$.

(v) $(\mathbf{A} + \mathbf{A}') \cdot \mathbf{B} = \mathbf{A} \cdot \mathbf{B} + \mathbf{A}' \cdot \mathbf{B}$ for all $\mathbf{A}, \mathbf{A}' \in \mathbb{F}^{m \times n}$ and $\mathbf{B} \in \mathbb{F}^{n \times \ell}$.

Proof. We will prove the third item only and leave the other parts to the reader. Using Equation (7.6) for the product $(\mathbf{B} \cdot \mathbf{C})$, we obtain that $(\mathbf{B} \cdot \mathbf{C})_{s,j} = \sum_{r=1}^{\ell} b_{sr} \cdot c_{rj}$. Using this and Equation (7.6) for the product $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C})$ and rewriting the resulting expression, we see that:

$$\begin{aligned}
(\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}))_{i,j} &= \sum_{s=1}^{n} a_{is} \cdot (\mathbf{B} \cdot \mathbf{C})_{s,j} \\
&= \sum_{s=1}^{n} a_{is} \cdot \sum_{r=1}^{\ell} b_{sr} \cdot c_{rj} \\
&= \sum_{s=1}^{n} \sum_{r=1}^{\ell} a_{is} \cdot (b_{sr} \cdot c_{rj}) \\
&= \sum_{s=1}^{n} \sum_{r=1}^{\ell} (a_{is} \cdot b_{sr}) \cdot c_{rj} \\
&= \sum_{r=1}^{\ell} \sum_{s=1}^{n} (a_{is} \cdot b_{sr}) \cdot c_{rj} \\
&= \sum_{r=1}^{\ell} \left(\sum_{s=1}^{n} a_{is} \cdot b_{sr} \right) \cdot c_{rj}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{r=1}^{\ell} (\mathbf{B} \cdot \mathbf{C}))_{i,r} \cdot c_{rj} \\
&= ((\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C})_{i,j}.
\end{aligned}
$$

□

We finish this section by explaining two more operations on matrices. We have already seen that vectors can be multiplied with a scalar. The generalisation to matrices is immediate: for $c \in \mathbb{F}$ and

$$
\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{F}^{m \times n}, \text{ we define } c \cdot \mathbf{A} = \begin{bmatrix} c \cdot a_{11} & \cdots & c \cdot a_{1n} \\ \vdots & & \vdots \\ c \cdot a_{m1} & \cdots & c \cdot a_{mn} \end{bmatrix}. \tag{7.7}
$$

Finally, there is a way to reverse the roles of rows and columns in a matrix $\mathbf{A}$. This is simply done by taking the *transpose* of a matrix, which is denoted by $\mathbf{A}^T$. More precisely, given

$$
\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbb{F}^{m \times n}, \text{ we define } \mathbf{A}^T = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \dots & a_{mn} \end{bmatrix}. \tag{7.8}
$$

Example 7.2.3

Let the matrix

$$
\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \in \mathbb{R}^{2 \times 3}
$$

be given. Compute $\mathbf{A}^T$.

Answer:

We have

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

Note that if $\mathbf{A} \in \mathbb{F}^{m \times n}$, then $\mathbf{A}^T \in \mathbb{F}^{n \times m}$. On the level of entries, we simply have that the (i, j)-th entry of $\mathbf{A}^T$ is equal to the (j, i)-th entry of $\mathbf{A}$.

The transpose behaves well with respect to matrix additions and matrix products. More precisely, we have the following theorem.

Theorem 7.2.2

Let $\mathbb{F}$ be a field. Then

(i) $(\mathbf{A}^T)^T = \mathbf{A}$ for all $\mathbf{A} \in \mathbb{F}^{m\times n}$.

(ii) $(\mathbf{A} + \mathbf{A}')^T = \mathbf{A}^T + (\mathbf{A}')^T$ for all $\mathbf{A}, \mathbf{A}' \in \mathbb{F}^{m\times n}$.

(iii) $(\mathbf{A} \cdot \mathbf{B})^T = \mathbf{B}^T \cdot \mathbf{A}^T$ for all $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{B} \in \mathbb{F}^{n\times \ell}$.

Proof. We only show the first item. In general, the (i, j)-th entry of $\mathbf{B}^T$ is equal to the (j, i)-th entry of $\mathbf{B}$ for any matrix $\mathbf{B}$. Applying this first for the matrix $\mathbf{A}^T$, then for the matrix $\mathbf{A}$, we obtain that $((\mathbf{A}^T)^T)_{i,j} = (\mathbf{A}^T)_{j,i} = (\mathbf{A})_{i,j}$. This shows that the matrices $\mathbf{A}^T$ and $\mathbf{A}$ have exactly the same entries and hence that they are equal. □

It is important to remember the order of multiplication in item 3 before and after transposing. In some sense, transposing reverses the order of the terms in a product. There is a good reason for this. Given matrices $\mathbf{A} \in \mathbb{F}^{m\times n}$ and $\mathbf{B} \in \mathbb{F}^{n\times \ell}$, the product $\mathbf{A}^T \cdot \mathbf{B}^T$ is in general not even defined! Indeed, the number of columns in $\mathbf{A}^T$ is m, while the number of rows in $\mathbf{B}^T$ is ℓ. However, the product $\mathbf{B}^T \cdot \mathbf{A}^T$ makes perfect sense, since the number of columns in $\mathbf{B}^T$ is n, which is the same as the number of rows in $\mathbf{A}^T$. Though these observations do not prove item three from Theorem 7.2.2, they do explain why it is quite natural that the multiplication order is given as it is.

7.3 Square matrices

If the number of rows and columns of a matrix are the same, it is called a *square* matrix. In other words, a matrix $\mathbf{A}$ is a square matrix, if $\mathbf{A} \in \mathbb{F}^{n\times n}$ for some positive integer n. The entries in the (i, i)-th positions of a square matrix are called the *diagonal entries* of the matrix. For example the diagonal entries of the matrix

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

are the numbers 1, 5 and 9. All diagonal entries together form what is known as the *diagonal* of a square matrix.

Given n, the $n \times n$ matrix $\mathbf{I}_n$, called the *identity* matrix, is the matrix having 1's on its diagonal, and 0's everywhere else. So for $n = 4$, we have for example

$$\mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix is called the identity matrix because it has no effect on a vector when multiplied with that vector. More precisely, a direct calculation shows that $\mathbf{I}_n \cdot \mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in \mathbb{F}^n$. Hence the function $L : \mathbb{F}^n \to \mathbb{F}^n$ defined by $L(\mathbf{v}) = \mathbf{I}_n \cdot \mathbf{v}$ is just the identity function. With this matrix in place, the following definition makes sense:

Definition 7.3.1

A square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ is called *invertible* if there exists a matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$ such that

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n.$$

The matrix $\mathbf{B}$, if it exists, is called the inverse matrix of $\mathbf{A}$ and denoted by $\mathbf{A}^{-1}$.

Inverse matrices will appear in many situations later on, but already when solving some systems of linear equations, they can be handy. Suppose for example, that one wants to solve the system of linear equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, with a square coefficient matrix $\mathbf{A}$, vector of variables $\mathbf{x} = (x_1, \ldots, x_n)$ and righthand-side $\mathbf{b} = (b_1, \ldots, b_n)$. If the coefficient matrix $\mathbf{A}$ has an inverse, we can multiply from the left with $\mathbf{A}^{-1}$ and simplify: $\mathbf{A}^{-1} \cdot (\mathbf{A} \cdot \mathbf{x}) = (\mathbf{A}^{-1} \cdot \mathbf{A}) \cdot \mathbf{x} = \mathbf{I}_n \cdot \mathbf{x}$. But this means that the equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ implies that $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$. Conversely, if $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$, then by multiplying with $\mathbf{A}$ from the left, we obtain that $\mathbf{A} \cdot \mathbf{x} = \mathbf{A} \cdot (\mathbf{A}^{-1} \cdot \mathbf{b}) = (\mathbf{A} \cdot \mathbf{A}^{-1}) \cdot \mathbf{b} = \mathbf{I}_n \cdot \mathbf{b} = \mathbf{b}$. Hence we have shown that:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \text{ if and only if } \mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}, \text{ provided } \mathbf{A}^{-1} \text{ exists.} \tag{7.9}$$

This observation actually has a nice consequence about the rank of invertible matrices:

Lemma 7.3.1

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given and suppose that its inverse matrix exists. Then $\rho(\mathbf{A}) = n$.

Proof. Equation (7.9) implies that the homogeneous system of linear equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$ only has the solution $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{0} = \mathbf{0}$. But then by Corollary 6.4.5, the rank of $\mathbf{A}$ is equal to n. □

More is true, but we will return to that later. The question is now how to figure out when a matrix has an inverse and if it does, how to compute it. We will first find an algorithmic answer and after that describe a theoretical characterisation of invertible matrices.

What we will do first, is to find an algorithm that for a given $n \times n$ matrix $\mathbf{A}$, computes an $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ if it exists. Hence the outcome of the algorithm will either be that such a $\mathbf{B}$ does not exist, or it will return such a $\mathbf{B}$. Note that according to Definition 7.3.1, the inverse of $\mathbf{A}$, here denoted by $\mathbf{B}$, should satisfy $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ *and* $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$. Fortunately, it turns out that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ *implies* $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$, so that the algorithm we are about to describe indeed will compute the inverse matrix $\mathbf{B} = \mathbf{A}^{-1}$, provided it exists.

Let us denote the i-th column of the identity matrix $\mathbf{I}_n$ by $\mathbf{e}_i$ for $i = 1, \ldots, n$. So for example for $n = 4$, we have

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \text{ and } \mathbf{e}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

The idea of the algorithm to find inverse matrices is the following: we are trying to find a matrix $\mathbf{B} \in \mathbb{F}^{n \times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ for a given $\mathbf{A} \in \mathbb{F}^{n \times n}$. Now let us denote the columns of $\mathbf{B}$ as $\mathbf{b}_1, \ldots, \mathbf{b}_n$. The i-th column of $\mathbf{A} \cdot \mathbf{B}$ is by definition of the matrix product equal to $\mathbf{A} \cdot \mathbf{b}_i$, while the i-th column of $\mathbf{I}_n$ is equal to $\mathbf{e}_i$. Hence $\mathbf{A} \cdot \mathbf{b}_i = \mathbf{e}_i$ for all i between 1 and n. Conversely, if $\mathbf{A} \cdot \mathbf{b}_i = \mathbf{e}_i$ for all i between 1 and n, then the matrices $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{I}_n$ have the same columns, whence $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$. Therefore we see that

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n \text{ if and only if } \mathbf{A} \cdot \mathbf{b}_i = \mathbf{e}_i \text{ for all } i \text{ between 1 and } n.$$

Therefore, we can find $\mathbf{b}_i$, by solving the inhomogeneous system of linear equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$.

From the theory from the previous chapter, we see that to figure out if the system of equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$, it is enough to compute the reduced row echelon form of the augmented matrix $[\mathbf{A}|\mathbf{e}_i]$. If $\rho(\mathbf{A}) = \rho([\mathbf{A}|\mathbf{e}_i])$, then according to Corollary 6.4.3, there exists a solution and otherwise not. Hence precisely if for all i between 1 and n it holds that $\rho(\mathbf{A}) = \rho([\mathbf{A}|\mathbf{e}_i])$, we will be able to find a matrix $\mathbf{B} \in \mathbb{F}^{n \times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$.

Now, one could deal with the system of equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$ for one i at the time and in that way compute one column of the matrix $\mathbf{B}$ at the time, if it exists. However, the first part of the corresponding augmented matrices is always the same, namely $\mathbf{A}$. Therefore, it is faster to deal with all n systems at the same time by computing the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{e}_1|\mathbf{e}_2|\ldots|\mathbf{e}_n] = [\mathbf{A}|\mathbf{I}_n]$.

Hence the algorithm of how to determine if a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ has an inverse, and if yes how to compute it, is the following:

(i) Compute the reduced row echelon form of the $n \times 2n$ matrix $[\mathbf{A}|\mathbf{I}_n]$. This can be done using elementary row operations, just as we did in Section 6.3

(ii) If the resulting reduced row echelon form is not of the form $[\mathbf{I}_n|\mathbf{B}]$, conclude that $\mathbf{A}$ does not have an inverse.

(iii) If it is of the form $[\mathbf{I}_n|\mathbf{B}]$, conclude that $\mathbf{A}$ does have an inverse, namely $\mathbf{A}^{-1} = \mathbf{B}$.

To see how this works in practice, let us consider two examples.

Example 7.3.1

Let $\mathbb{F} = \mathbb{R}$ and

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Determine whether or not this matrix has an inverse and if yes, compute it.

Answer: First we determine the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{I}_2]$. We obtain:

$$\begin{aligned} [\mathbf{A}|\mathbf{I}_2] &= \begin{bmatrix} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \end{bmatrix} \xrightarrow[R_2 \leftarrow R_2 - 3 \cdot R_1]{} \begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & -2 & -3 & 1 \end{bmatrix} \\ &\xrightarrow[R_2 \leftarrow (-1/2) \cdot R_2]{} \begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & 1 & 3/2 & -1/2 \end{bmatrix} \xrightarrow[R_1 \leftarrow R_1 - 2 \cdot R_2]{} \begin{bmatrix} 1 & 0 & -2 & 1 \\ 0 & 1 & 3/2 & -1/2 \end{bmatrix}. \end{aligned}$$

Hence we conclude that $\mathbf{A}$ has an inverse, namely

$$\mathbf{A}^{-1} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix}.$$

Example 7.3.2

Let $\mathbb{F} = \mathbb{R}$ and

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

Determine whether or not this matrix has an inverse and if yes, compute it.

Answer: We start determining the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{I}_3]$. We obtain:

$$\begin{aligned} [\mathbf{A}|\mathbf{I}_3] &= \begin{bmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 5 & 7 & 9 & 0 & 0 & 1 \end{bmatrix} \\ &\xrightarrow[R_2 \leftarrow R_2 - 4 \cdot R_1]{} \begin{bmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 5 & 7 & 9 & 0 & 0 & 1 \end{bmatrix} \\ &\xrightarrow[R_3 \leftarrow R_3 - 5 \cdot R_1]{} \begin{bmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & -3 & -6 & -5 & 0 & 1 \end{bmatrix} \\ &\xrightarrow[R_3 \leftarrow R_3 - R_2]{} \begin{bmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \end{bmatrix}. \end{aligned}$$

Even though we have not found the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_3]$ yet, we already found an echelon form of it. The pivots can already be read off and are contained in the first, second, and fourth columns of the matrix. When proceeding to find the reduced row echelon form, the first three entries of the third row will remain zero. The reader is encouraged to compute the reduced echelon form and see that this indeed is true. Hence the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_3]$ will not be of the form $[\mathbf{I}_3|\mathbf{B}]$. We conclude that the matrix $\mathbf{A}$ does not have an inverse.

In principle, we now have an algorithm that can determine if a square matrix has an inverse and if yes, computes it. However, we have not shown that the algorithm is correct. In other words, if we follow the steps of the algorithm, will the outcome always be what it should be? First of all, we should make sure that if the reduced row echelon form of the $n \times 2n$ matrix $[\mathbf{A}|\mathbf{I}_n]$ is not of the form $[\mathbf{I}_n|\mathbf{B}]$, then $\mathbf{A}$ indeed has no inverse. And second of all, we should make sure that if a matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$ satisfies $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$, then also $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$, so that we indeed can conclude that $\mathbf{B}$ is the inverse of $\mathbf{A}$. We will address these issues in the rest of this section. It turns out that everything is as it should be and one can show that:

$$\begin{aligned} \mathbf{A}^{-1} \text{ exists} &\Leftrightarrow \text{ the reduced row echelon form of } [\mathbf{A}|\mathbf{I}_n] \text{ is of the form } [\mathbf{I}_n|\mathbf{B}] \\ &\Leftrightarrow \rho(\mathbf{A}) = n \text{ (that is: the rank of } \mathbf{A} \text{ is } n). \end{aligned} \tag{7.10}$$

A reader willing to accept this without proof can skip the remainder of this section, but for the other readers we will give a proof below.

Theorem 7.3.2

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a square matrix. Then the following statements are logically equivalent:

(i) The reduced row echelon form of the $n \times 2n$ matrix $[\mathbf{A}|\mathbf{I}_n]$ is of the form $[\mathbf{I}_n|\mathbf{B}]$ for some square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$.

(ii) There exists a square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$.

Proof. Let us assume that the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{I}_n]$ is of the form $[\mathbf{I}_n|\mathbf{B}]$ for some square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$. Let us denote by $\mathbf{b}_i$ the i-th column of the matrix $\mathbf{B}$. Then using the same elementary row operations to transform the matrix $[\mathbf{A}|\mathbf{I}_n]$ into the form $[\mathbf{I}_n|\mathbf{B}]$ can be used to transform the matrix $[\mathbf{A}|\mathbf{e}_i]$ into $[\mathbf{I}_n|\mathbf{b}_i]$. Since $[\mathbf{I}_n|\mathbf{b}_i]$ is in reduced row echelon form, we can conclude that the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{e}_i]$ is equal to $[\mathbf{I}_n|\mathbf{b}_i]$. This implies that $\mathbf{b}_i$ is a solution to the system of linear equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$. But then $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$. In particular $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ for some square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$, namely the matrix occurring in the right part of the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$.

Now conversely, suppose that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ for some square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$. Then for all i from 1 to n, the system of linear equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$ has a solution, namely the i-th column of the matrix $\mathbf{B}$. We claim that the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ only contains pivots in its first n columns. We will proof this claim using a proof by contradiction. Assume therefore

that the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ in fact has a pivot contained in a column with index $n+i$ for some $i > 0$. Then the reduced row echelon form of the matrix $[\mathbf{A}|\mathbf{e}_i]$ would contain a pivot in its $(n+1)$-th column. In particular, $\mathbf{A}$ and $[\mathbf{A}|\mathbf{e}_i]$ would not have the same rank. But then by Corollary 6.4.3, the system $\mathbf{A} \cdot \mathbf{x} = \mathbf{e}_i$ has no solution. Since we already observed that it does have a solution, we obtain a contradiction. This proves the claim that the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ only contains pivots in its first n columns. Next, we claim that the rank of $[\mathbf{A}|\mathbf{I}_n]$ is equal to n. To obtain a contradiction, suppose that the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ contains a zero row. Considering the second part of the matrix, $\mathbf{I}_n$, we can conclude that there exist a sequence of elementary row operations that can transform $\mathbf{I}_n$ into a matrix for a zero row. But $\mathbf{I}_n$ is a matrix with rank n, while an $n \times n$ matrix with a zero row can have rank at most $n-1$. This proves the second claim. Combining the two claims, we conclude that the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ contains a pivot in each of its first n columns. But then it is of the form $[\mathbf{I}_n|\mathbf{C}]$ for some square matrix $\mathbf{C} \in \mathbb{F}^{n\times n}$. □

Corollary 7.3.3

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Then there exists $\mathbf{B} \in \mathbb{F}^{n\times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ if and only if $\rho(\mathbf{A}) = n$.

Proof. If $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$ for some $\mathbf{B} \in \mathbb{F}^{n\times n}$, then by Theorem 7.3.2 the reduced row echelon form of the $n \times 2n$ matrix $[\mathbf{A}|\mathbf{I}_n]$ is of the form $[\mathbf{I}_n|\mathbf{C}]$ for some $\mathbf{C} \in \mathbb{F}^{n\times n}$. But then the reduced row echelon form of $\mathbf{A}$ itself is $\mathbf{I}_n$, implying that $\rho(\mathbf{A}) = n$.

Conversely, if $\rho(\mathbf{A}) = n$, the reduced row echelon form of $\mathbf{A}$ is equal to $\mathbf{I}_n$. But then the reduced row echelon form of $[\mathbf{A}|\mathbf{I}_n]$ is of the form $[\mathbf{I}_n|\mathbf{C}]$ for some square matrix $\mathbf{C} \in \mathbb{F}^{n\times n}$. By Theorem 7.3.2, we may conclude that there exists $\mathbf{B} \in \mathbb{F}^{n\times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$. □

Corollary 7.3.4

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a square matrix and suppose that there exists a square matrix $\mathbf{B} \in \mathbb{F}^{n\times n}$ such that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$. Then $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$ and therefore $\mathbf{B} = \mathbf{A}^{-1}$, the inverse of $\mathbf{A}$.

Proof. To conclude that $\mathbf{B}$ is the inverse of $\mathbf{A}$, we need to show that $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$. Since we are given that $\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n$, what is left to show, is that $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$.

Now note that $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{A}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{A} = \mathbf{I}_n \cdot \mathbf{A} = \mathbf{A} = \mathbf{A} \cdot \mathbf{I}_n$. Hence $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{A} - \mathbf{I}_n) = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{A}) - \mathbf{A} \cdot \mathbf{I}_n = \mathbf{0}$, where here $\mathbf{0}$ denotes the $n \times n$ zero matrix.

Note that the previous equation implies that any column of $\mathbf{B} \cdot \mathbf{A} - \mathbf{I}_n$ is a solution to the homogeneous system of equations $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$. On the other hand, the previous corollary implies that the matrix $\mathbf{A}$ has rank n. Hence, we know from Corollary 6.4.5 that the system $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$ only has the solution $\mathbf{x} = \mathbf{0}$. Hence all columns of $\mathbf{B} \cdot \mathbf{A} - \mathbf{I}_n$ are zero, implying that $\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n$. This is exactly what we needed to show. □

Corollary 7.3.5

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Then its inverse matrix exists if and only if $\rho(\mathbf{A}) = n$.

Proof. This follows from the previous two corollaries. □

Chapter 8

Determinants

8.1 Determinant of a square matrix

In this section, we will introduce the *determinant* of a square matrix. Determinants will be useful when investigating if a given matrix is invertible, but will also become very useful in later chapters. We start with a notational convention:

Definition 8.1.1

Let $\mathbf{A} = (a_{ij})_{1\leq i\leq n, 1\leq j\leq n} \in \mathbb{F}^{n\times n}$ be a given square matrix. Then we define the matrix $\mathbf{A}(i;j) \in \mathbb{F}^{(n-1)\times(n-1)}$ as:

$$\mathbf{A}(i;j) = \begin{bmatrix} a_{11} & \dots & a_{1\,j-1} & a_{1\,j+1} & \dots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1\,1} & \dots & a_{i-1\,j-1} & a_{i-1\,j+1} & \dots & a_{i-1\,n} \\ a_{i+1\,1} & \dots & a_{i+1\,j-1} & a_{i+1\,j+1} & \dots & a_{i+1\,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{n\,j-1} & a_{n\,j+1} & \dots & a_{nn} \end{bmatrix}.$$

In words: the matrix $\mathbf{A}(i;j)$ is obtained from $\mathbf{A}$ by deleting the i-th row and j-th column of $\mathbf{A}$. With this in place, we can define the determinant of a square matrix recursively as follows:

Definition 8.1.2

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a square matrix. Then we define

$$\det(\mathbf{A}) = \begin{cases} \mathbf{A} & \text{if } n = 1, \\ \sum_{i=1}^{n} (-1)^{i+1} \cdot a_{i1} \cdot \det(\mathbf{A}(i;1)) & \text{if } n \geq 2. \end{cases}$$

Instead of using the summation symbol, one may also write:

$$\det(\mathbf{A}) = a_{11} \cdot \det(\mathbf{A}(1;1)) - a_{21} \cdot \det(\mathbf{A}(2;1)) + \cdots + (-1)^{n+1} \cdot a_{n1} \cdot \det(\mathbf{A}(n;1)).$$

Example 8.1.1

Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

To compute the determinant of this matrix, we will use Definition 8.1.2. First of all, note that $\mathbf{A}(1;1) = a_{22}$ and $\mathbf{A}(2;1) = a_{12}$. Therefore

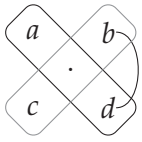
$$\det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11} \cdot \det(a_{22}) - a_{21} \cdot \det(a_{12}) = a_{11}a_{22} - a_{21}a_{12}.$$

Just using the letters a, b, c and d for the entries of the matrix, we can reformulate this as:

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc. \tag{8.1}$$

When given the task to compute the determinant of a 2×2 matrix, this equation is very practical. Figure 8.1 visualizes the equation for the determinant of a 2×2 matrix: to compute it, just multiply the two diagonal entries of the matrix and subtract the product of the two remaining entries.

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) =$$



$$= ad - cb.$$

Figure 8.1: Determinant of 2×2 matrix.

Example 8.1.2

As in Example 7.3.2, let $\mathbb{F} = \mathbb{R}$ and

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

Compute the determinant of $\mathbf{A}$.

Answer: First of all, note that

$$\mathbf{A}(1;1) = \begin{bmatrix} 5 & 6 \\ 7 & 9 \end{bmatrix}, \mathbf{A}(2;1) = \begin{bmatrix} 2 & 3 \\ 7 & 9 \end{bmatrix}, \text{ and } \mathbf{A}(3;1) = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}.$$

Hence using Definition 8.1.2, we obtain that

$$\det(\mathbf{A}) = 1 \cdot \det\left(\begin{bmatrix} 5 & 6 \\ 7 & 9 \end{bmatrix}\right) - 4 \cdot \det\left(\begin{bmatrix} 2 & 3 \\ 7 & 9 \end{bmatrix}\right) + 5 \cdot \det\left(\begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}\right).$$

Using Equation (8.1), we can quickly compute the determinants of 2×2 matrices. Then we obtain that

$$\det(\mathbf{A}) = 1 \cdot (45 - 42) - 4 \cdot (18 - 21) + 5 \cdot (12 - 15) = 3 + 12 - 15 = 0.$$

Later, we will have a few more techniques at our disposal for computing determinants of matrices, but for now we consider one particular class of matrices.

Definition 8.1.3

A matrix $\mathbf{A} = \mathbb{F}^{n \times n}$ is called a *diagonal matrix*, if there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$ such that

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda_{n-1} & 0 \\ 0 & \ldots & 0 & 0 & \lambda_n \end{bmatrix}.$$

We already mentioned previously that for square matrix $\mathbf{A} = (a_{ij})_{1 \leq i \leq n; 1 \leq j \leq n}$, the entries $a_{11}, \ldots, a_{nn}$ are called the diagonal entries of $\mathbf{A}$. This explains that name diagonal matrix in Definition 8.1.3: a diagonal matrix is a square matrix all of whose entries are zeroes, except possibly those on its diagonal. For example, the identity matrix $\mathbf{I}_n$ mentioned in the beginning of Section 7.3, is a diagonal matrix, with diagonal entries all equal to 1.

Proposition 8.1.1

Let $\mathbf{A} = \mathbb{F}^{n \times n}$ be a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$. Then

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_n.$$

In particular $\det(\mathbf{I}_n) = 1$.

Proof. We show this using induction on n. Indeed, if $n = 1$, then $\mathbf{A} = \lambda_1$ and Definition 8.1.2 implies that $\det(\mathbf{A}) = \lambda_1$. Now assume that $n \geq 2$ and that the proposition holds for diagonal

matrices in $\mathbb{F}^{(n-1)\times(n-1)}$. Using Definition 8.1.2, we then see that:

$$\begin{aligned}\det(\mathbf{A}) &= \lambda_1 \cdot \det(\mathbf{A}(1;1)) - 0 \cdot \det(\mathbf{A}(2;1)) + \cdots + (-1)^{n+1} \cdot 0 \cdot \det(\mathbf{A}(n;1)) \\ &= \lambda_1 \cdot \det(\mathbf{A}(1;1)) \\ &= \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_n,\end{aligned}$$

where in the last equality we used the induction hypothesis. The induction hypothesis applies, since $\mathbf{A}(1;1)$ is a diagonal matrix with diagonal entries $\lambda_2, \ldots, \lambda_n$. This completes the induction step. Using the induction principle, we conclude that the proposition is true. The particular case of the identity matrix now also follows, since then all diagonal entries are equal to one. □

We can in fact at this point already give a formula for the determinant of a larger class of matrices called upper triangular matrices:

Definition 8.1.4

A matrix $\mathbf{A} = \mathbb{F}^{n\times n}$ is called an *upper triangular matrix*, if there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$ and $a_{ij} \in \mathbb{F}$ for $1 \leq i < j \leq n$, such that

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & \lambda_2 & a_{23} & \ldots & a_{2n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda_{n-1} & a_{n-1\,n} \\ 0 & \ldots & 0 & 0 & \lambda_n \end{bmatrix}.$$

In words: an upper triangular matrix has all its nonzero entries above or on its diagonal. In particular, all entries below the diagonal of an upper triangular matrix are zero.

Theorem 8.1.2

Let $\mathbf{A} = \mathbb{F}^{n\times n}$ be an upper triangular matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$. Then

$$\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_n.$$

Proof. The proof is very similar as the proof of Proposition 8.1.1 and left to the reader. □

Another type of matrices, in the same spirit as upper triangular matrices, is the following:

Definition 8.1.5

A matrix $\mathbf{A} = \mathbb{F}^{n\times n}$ is called an *lower triangular matrix*, if there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$ and $a_{ij} \in \mathbb{F}$

for $1 \leq j < i \leq n$, such that

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ a_{21} & \lambda_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{n-11} & \dots & a_{n-1\,n-2} & \lambda_{n-1} & 0 \\ a_{n1} & \dots & a_{n\,n-2} & a_{n\,n-1} & \lambda_n \end{bmatrix}.$$

In words: a lower triangular matrix has all its nonzero entries below or on its diagonal. In particular, all entries above the diagonal of a lower triangular matrix are zero. Also here, we can find a formula for its determinant. Before showing that, we need a lemma that can be useful in its own right.

Lemma 8.1.3

If a square matrix in $\mathbb{F}^{n\times n}$ contains a zero row, its determinant is zero.

Proof. This can be shown using induction on n. Providing the details, is left to the reader. □

Theorem 8.1.4

Let $\mathbf{A} = \mathbb{F}^{n\times n}$ be a lower triangular matrix with diagonal entries $\lambda_1, \dots, \lambda_n$. Then

$$\det(\mathbf{A}) = \lambda_1 \cdot \dots \cdot \lambda_n.$$

Proof. We show this using induction on n. Indeed, if $n = 1$, then $\mathbf{A} = \lambda_1$ and Definition 8.1.2 implies that $\det(\mathbf{A}) = \lambda_1$. Now assume that $n \geq 2$ and that the proposition holds for lower triangular matrices in $\mathbb{F}^{(n-1)\times(n-1)}$. Now note that $\mathbf{A}(1;1)$ is a lower triangular matrix with diagonal entries $\lambda_2, \dots, \lambda_n$. Hence the induction hypothesis implies that $\det(\mathbf{A}(1;1)) = \lambda_2 \cdot \dots \cdot \lambda_n$. The matrices $\mathbf{A}(2;1), \dots, \mathbf{A}(n;1)$ all have the zero row as first row. The reason for this is that the first row of $\mathbf{A}$ only has a nonzero entry in its first position, but this position has been removed when constructing the matrices $\mathbf{A}(2;1), \dots, \mathbf{A}(n;1)$. By Lemma 8.1.3, we therefore have $\det(\mathbf{A}(2;1)) = 0, \dots, \det(\mathbf{A}(n;1)) = 0$.

Using Definition 8.1.2, we then see that:

$$\begin{aligned} \det(\mathbf{A}) &= \lambda_1 \cdot \det(\mathbf{A}(1;1)) - a_{21} \cdot \det(\mathbf{A}(2;1)) + \dots + (-1)^{n+1} \cdot a_{n1} \cdot \det(\mathbf{A}(n;1)) \\ &= \lambda_1 \cdot \det(\mathbf{A}(1;1)) - a_{21} \cdot 0 + \dots + (-1)^{n+1} \cdot a_{n1} \cdot 0 \\ &= \lambda_1 \cdot \det(\mathbf{A}(1;1)) \\ &= \lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_n, \end{aligned}$$

where in the last equality we used the induction hypothesis. This completes the induction step. Using the induction principle, we conclude that the theorem is true. □

8.2 Determinants and elementary row operations

Using Definition 8.1.2 is not always the fastest way to compute the determinant of a square matrix. When studying systems of linear equations, three types of elementary row operations could be used to simplify a given system immensely. Motivated by this, we now study the effect of these three types of elementary row operations on the value of a determinant. The easiest to deal with is an elementary row operation of the form $R_i \leftarrow c \cdot R_i$. We start by proving a more general result.

Theorem 8.2.1

Consider the following three matrices in $\mathbb{F}^{n\times n}$:

$$\mathbf{A} = \begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{a}_i & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix}, \mathbf{B} = \begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{b}_i & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix}, \text{ and } \mathbf{C} = \begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & c \cdot \mathbf{a}_i + \mathbf{b}_i & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix},$$

where $c \in \mathbb{F}$. Then $\det(\mathbf{C}) = c \cdot \det(\mathbf{A}) + \det(\mathbf{B})$.

Proof. We use induction on n. If $n = 1$, we have $\mathbf{A} = a$ for some $a \in \mathbb{F}$, $\mathbf{B} = b$ for some $b \in \mathbb{F}$ and $\mathbf{C} = c \cdot a + b$. Then according to Definition 8.1.2, we see that $\det(\mathbf{C}) = c \cdot a + b = c \cdot \det(\mathbf{A}) + \det(\mathbf{B})$.

Now assume that $n \geq 2$ and that the theorem holds for $(n-1) \times (n-1)$ matrices. We know from Definition 8.1.2 that

$$\det(\mathbf{C}) = \sum_{k=1}^{n} (-1)^{k+1} \cdot c_{k1} \cdot \det(\mathbf{C}(k;1)).$$

Let us denote by $\sum_{k=1;k\neq i}^{n} (-1)^{k+1} \cdot c_{k1} \cdot \det(\mathbf{C}(k;1))$ the summation one obtains by letting k range from 1 to n, except that now the value i is skipped. Then we can write

$$\det(\mathbf{C}) = \sum_{k=1;k\neq i}^{n} (-1)^{k+1} \cdot c_{k1} \cdot \det(\mathbf{C}(k;1)) + (-1)^{i+1} \cdot c_{i1} \cdot \det(\mathbf{C}(i;1)).$$

For all k different from i, the induction hypothesis implies that $\det(\mathbf{C}(k;1)) = c \cdot \det(\mathbf{A}(k;1)) + \det(\mathbf{B}(k;1))$. Further $\mathbf{C}(i;1) = \mathbf{A}(i;1) = \mathbf{B}(i;1)$, since the i-th row is the only row in which the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ differ. Now using that $c_{i1} = c \cdot a_{i1} + b_{i1}$ and $c_{k1} = a_{k1}$ if $k \neq i$, we see

that

$$\begin{aligned}
\det(\mathbf{C}) &= \sum_{k=1;k\neq i}^{n} (-1)^{k+1} \cdot a_{k1} \cdot \det(\mathbf{C}(k;1)) + \\
&\quad (-1)^{i+1} \cdot (a_{i1} + b_{i1}) \cdot \det(\mathbf{C}(i;1)) \\
&= \sum_{k=1;k\neq i}^{n} (-1)^{k+1} \cdot a_{k1} \cdot (c \cdot \det(\mathbf{A}(k;1)) + \det(\mathbf{B}(k;1))) \\
&\quad + (-1)^{i+1} \cdot c \cdot a_{i1} \cdot \det(\mathbf{A}(i;1)) + (-1)^{i+1} \cdot b_{i1} \cdot \det(\mathbf{B}(i;1)) \\
&= \sum_{k=1;k\neq i}^{n} c \cdot (-1)^{k+1} \cdot a_{k1} \cdot \det(\mathbf{A}(k;1)) + (-1)^{i+1} \cdot c \cdot a_{i1} \cdot \det(\mathbf{A}(i;1)) \\
&\quad + \sum_{k=1;k\neq i}^{n} (-1)^{k+1} \cdot b_{k1} \cdot \det(\mathbf{B}(k;1))) + (-1)^{i+1} \cdot b_{i1} \cdot \det(\mathbf{B}(i;1)) \\
&= c \cdot \det(\mathbf{A}) + \det(\mathbf{B}).
\end{aligned}$$

This concludes the induction step and hence the induction proof. □

Corollary 8.2.2

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given and suppose that $\mathbf{C}$ is obtained from $\mathbf{A}$ by applying the elementary row operation $R_i \leftarrow c \cdot R_i$ on $\mathbf{A}$, for some i and some $c \in \mathbb{F}$. Then $\det(\mathbf{C}) = c \cdot \det(\mathbf{A})$.

Proof. If we choose $\mathbf{b}_i = \mathbf{0}$ in Theorem 8.2.1, we find that $\det(\mathbf{C}) = c \cdot \det(\mathbf{A}) + \det(\mathbf{B})$, where $\mathbf{B}$ is a matrix whose i-th row is the zero row. Lemma 8.1.3, implies that $\det(\mathbf{B}) = 0$. Hence the corollary follows. □

Investigating the effect of the remaining two types of elementary row operation turns out to be more elaborate. What turns out to happen is the following:

$$\text{Applying } R_i \leftrightarrow R_j \text{ on a square matrix, changes the sign of the determinant.} \tag{8.2}$$

$$\text{Applying } R_i \leftarrow R_i + c \cdot R_j \text{ on a square matrix, does not affect the determinant.} \tag{8.3}$$

Example 8.2.1

As in Example 7.3.2, let $\mathbb{F} = \mathbb{R}$ and

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

Compute the determinant of $\mathbf{A}$ using elementary row operations.

Answer: From Example 7.3.2 we can read off that:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 5 & 7 & 9 \end{bmatrix} \xrightarrow[R_2 \leftarrow R_2 - 4 \cdot R_1]{} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 5 & 7 & 9 \end{bmatrix}$$
$$\xrightarrow[R_3 \leftarrow R_3 - 5 \cdot R_1]{} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & -3 & -6 \end{bmatrix} \xrightarrow[R_3 \leftarrow R_3 - R_2]{} \begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{bmatrix}.$$

Using Equation (8.3) three times, we may conclude that

$$\det(\mathbf{A}) = \det\left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{bmatrix}\right).$$

Now note that the matrix on the right-hand side is an upper triangular matrix. Hence using Theorem 8.1.2, we obtain that

$$\det(\mathbf{A}) = \det\left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{bmatrix}\right) = 1 \cdot (-3) \cdot 0 = 0.$$

In the rest of this section, we will prove the validity of Equations (8.2) and (8.3). A reader willing to accept their validity without proof can directly proceed to Section 8.3. A reader who wants to read the proof of Equations (8.2) and (8.3) is invited to do so, but on a first reading it may be best to read Section 8.3 first.

We start with two lemmas.

Lemma 8.2.3

Assume that $n \geq 2$ and let a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ be given. Further, denote by $\mathbf{B} \in \mathbb{F}^{n \times n}$ a matrix obtained from $\mathbf{A}$ by interchanging two consecutive rows of $\mathbf{A}$. Then $\det(\mathbf{B}) = -\det(\mathbf{A})$.

Proof. We prove this using induction on n.

If $n = 2$, we have

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

implying $\det(\mathbf{A}) = a_{11} \cdot a_{22} - a_{21} \cdot a_{12}$ and $\det(\mathbf{B}) = a_{21} \cdot a_{12} - a_{11} \cdot a_{22}$. Hence $\det(\mathbf{B}) = -\det(\mathbf{A})$.

Now let $n \geq 3$ and assume that the lemma holds for $n-1$. Let us denote the two rows of $\mathbf{A}$ that are interchanged by R_i and R_{i+1}. Then, we see that $\mathbf{A}(i;1) = \mathbf{B}(i+1;1)$ and $\mathbf{A}(i+1;1) = \mathbf{B}(i;1)$. Further for $k \neq i$ and $k \neq i+1$, we have that $\mathbf{B}(k;1)$ can be obtained from $\mathbf{A}(k;1)$ by interchanging two consecutive rows. Hence for such k, we have $\det(\mathbf{B}(i;1)) = -\det(\mathbf{A}(i;1))$ from the induction hypothesis. Putting all this together, we find:

$$\begin{aligned}
\det(\mathbf{B}) &= \sum_{k=1;k\neq i;k\neq i+1}^{n} (-1)^{k+1} \cdot a_{k1} \cdot \det(\mathbf{B}(k;1)) \\
&\quad + (-1)^{i+1} \cdot a_{i+11} \cdot \det(\mathbf{B}(i;1)) + (-1)^{i+2} \cdot a_{i1} \cdot \det(\mathbf{B}(i+1;1)) \\
&= -\sum_{k=1;k\neq i;k\neq i+1}^{n} (-1)^{k+1} \cdot a_{k1} \cdot \det(\mathbf{A}(k;1)) \\
&\quad + (-1)^{i+1} \cdot a_{i+11} \cdot \det(\mathbf{A}(i+1;1)) + (-1)^{i+2} \cdot a_{i1} \cdot \det(\mathbf{A}(i;1)) \\
&= -\sum_{k=1}^{n} (-1)^{k+1} \cdot a_{k1} \cdot \det(\mathbf{A}(k;1)) \\
&= -\det(\mathbf{A}).
\end{aligned}$$

This concludes the induction step and hence the proof. □

Lemma 8.2.4

Assume that $n \geq 2$ and let a square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Assume that two consecutive rows of $\mathbf{A}$ are identical. Then $\det(\mathbf{A}) = 0$.

Proof. This can be shown following the same strategy as in the proof of Lemma 8.2.3 □

The above lemma is just a special case of a more general result:

Proposition 8.2.5

Assume that $n \geq 2$ and let a square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Assume that two rows of $\mathbf{A}$ are identical. Then $\det(\mathbf{A}) = 0$.

Proof. If two consecutive rows of $\mathbf{A}$ are identical, Lemma 8.2.4 implies $\det(\mathbf{A}) = 0$. Therefore we are left with the case that two rows of $\mathbf{A}$ are identical, but that these are not consecutive. Now let us denote the two given identical rows of $\mathbf{A}$ by R_i and R_j, for some $i > j \geq 1$. We interchange rows R_i and R_{i-1}, thus moving the row R_i up in the matrix. The effect on the determinant is a sign change using Lemma 8.2.3. In the new matrix, the identical rows are now rows R_j and R_{i-1}. If these rows are consecutive, we stop interchanging rows, but otherwise, we move the lowest of the two identical rows up, one row at the time. Therefore, we end up with a matrix $\mathbf{B}$ with two consecutive rows. Moreover, using Lemma 8.2.3 each time we interchange two consecutive rows, we know that $\det(\mathbf{B}) = \pm\det(\mathbf{A})$. On the other hand, $\det(\mathbf{B}) = 0$ by Lemma 8.2.4. Hence we can conclude that $\det(\mathbf{A}) = 0$. □

We now have all the ingredients needed to show the effect of interchanging two rows on the determinant of a square matrix.

Theorem 8.2.6

Let a square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given and denote by $\mathbf{B} \in \mathbb{F}^{n\times n}$ a matrix obtained from $\mathbf{A}$ using an elementary operation of the form $R_i \leftrightarrow R_j$ for some integers $i < j$. Then $\det(\mathbf{B}) = -\det(\mathbf{A})$.

Proof. Let us write

$$\mathbf{A} = \begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{a}_i & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_{j-1} & - \\ - & \mathbf{a}_j & - \\ - & \mathbf{a}_{j+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{a}_i + \mathbf{a}_j & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_{j-1} & - \\ - & \mathbf{a}_j + \mathbf{a}_i & - \\ - & \mathbf{a}_{j+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix}.$$

Applying Theorem 8.2.1 on row i of $\mathbf{C}$, we see that

$$\det(\mathbf{C}) = \det\left(\begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{a}_i & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_{j-1} & - \\ - & \mathbf{a}_j + \mathbf{a}_i & - \\ - & \mathbf{a}_{j+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix}\right) + \det\left(\begin{bmatrix} - & \mathbf{a}_1 & - \\ & \vdots & \\ - & \mathbf{a}_{i-1} & - \\ - & \mathbf{a}_j & - \\ - & \mathbf{a}_{i+1} & - \\ & \vdots & \\ - & \mathbf{a}_{j-1} & - \\ - & \mathbf{a}_j + \mathbf{a}_i & - \\ - & \mathbf{a}_{j+1} & - \\ & \vdots & \\ - & \mathbf{a}_n & - \end{bmatrix}\right)$$

Now we apply Theorem 8.2.1 again, but this time for row j in the matrices occurring in the two determinants on the right-hand side of this equation and use Proposition 8.2.5 afterwards. Then we obtain that

$$\det(\mathbf{C}) = \det(\mathbf{A}) + \det(\mathbf{B}).$$

However, Proposition 8.2.5 implies that $\det(\mathbf{C}) = 0$, since rows i and j of $\mathbf{C}$ are identical. Hence $0 = \det(\mathbf{A}) + \det(\mathbf{B})$, which implies what we wanted to show. □

Now that we know the effect of the elementary row operations $R_i \leftarrow c \cdot R_i$ and $R_i \leftrightarrow R_j$ on the determinant, let us also see what happens with the determinant when using an elementary operations of the form $R_i \leftarrow R_i + c \cdot R_j$.

Theorem 8.2.7

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given and suppose that the matrix $\mathbf{B}$ is obtained from $\mathbf{A}$ by applying the elementary row operation $R_i \leftarrow R_i + c \cdot R_j$ on $\mathbf{A}$, for some distinct row indices i, j, and $c \in \mathbb{F}$. Then $\det(\mathbf{B}) = \det(\mathbf{A})$.

Proof. This follows from Theorem 8.2.1 and Proposition 8.2.5. □

8.3 Alternative descriptions of the determinant

In our description of a determinant of a square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$, the first column of $\mathbf{A}$ played a special role. After all, in the recursive definition, we multiply entries from the first column of $\mathbf{A}$ with the determinants of smaller matrices. These smaller matrices were obtained from $\mathbf{A}$ by deleting the first column and some row. For this reason, one sometimes says that one in Definition 8.1.2 computes the determinant by expanding it along the first column. More precisely, one often refers to this as the *expansion* or *Laplace expansion* of the determinant along the first column.

One can now ask if there is any reason why the first column is so special. The answer is: it is not! It is possible to compute determinants by expansion along other columns and in fact also by expansion along rows. More precisely, we have the following theorem:

Theorem 8.3.1

Let $n \geq 2$ and $\mathbf{A} \in \mathbb{F}^{n\times n}$ be a square matrix. Then for any j between 1 and n:

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} \cdot a_{ij} \cdot \det(\mathbf{A}(i;j)). \tag{8.4}$$

Moreover, for any i between 1 and n:

$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} \cdot a_{ij} \cdot \det(\mathbf{A}(i;j)). \tag{8.5}$$

Proof. We will not prove this theorem, but the interested reader can find some remarks at the end of this section explaining the main ideas behind the proof. □

Note that for $j = 1$, Equation (8.4) simply becomes the formula given for the determinant given in Definition 8.1.2. Equation (8.4) describes the Laplace expansion of the determinant along the j-th column, while Equation (8.5) describes the Laplace expansion of the determinant along the i-th row. These equations can also be expressed without using the summation sign in the following way:

$$\det(\mathbf{A}) = (-1)^{1+j} \cdot a_{1j} \cdot \det(\mathbf{A}(1;j)) + (-1)^{2+j} \cdot a_{2j} \cdot \det(\mathbf{A}(2;j)) + \cdots + (-1)^{n+j} \cdot a_{nj} \cdot \det(\mathbf{A}(n;j)).$$

and

$$\det(\mathbf{A}) = (-1)^{i+1} \cdot a_{i1} \cdot \det(\mathbf{A}(i;1)) + (-1)^{i+2} \cdot a_{i2} \cdot \det(\mathbf{A}(i;2)) + \cdots + (-1)^{i+n} \cdot a_{in} \cdot \det(\mathbf{A}(i;n)).$$

Example 8.3.1

As in Example 7.3.2, let $\mathbb{F} = \mathbb{R}$ and

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

Compute the determinant of $\mathbf{A}$ using Laplace expansion along the first row.

Answer: First of all, note that

$$\mathbf{A}(1;1) = \begin{bmatrix} 5 & 6 \\ 7 & 9 \end{bmatrix}, \mathbf{A}(1;2) = \begin{bmatrix} 4 & 6 \\ 5 & 9 \end{bmatrix}, \text{ and } \mathbf{A}(1;3) = \begin{bmatrix} 4 & 5 \\ 5 & 7 \end{bmatrix}.$$

Hence using Laplace expansion along the first row, we obtain that

$$\det(\mathbf{A}) = (-1)^{1+1} \cdot 1 \cdot \det\left(\begin{bmatrix} 5 & 6 \\ 7 & 9 \end{bmatrix}\right) + (-1)^{1+2} \cdot 2 \cdot \det\left(\begin{bmatrix} 4 & 6 \\ 5 & 9 \end{bmatrix}\right) + (-1)^{1+3} \cdot 3 \cdot \det\left(\begin{bmatrix} 4 & 5 \\ 5 & 7 \end{bmatrix}\right).$$

Using Equation (8.1), we can quickly compute the determinants of 2×2 matrices. Then we obtain that

$$\det(\mathbf{A}) = 1 \cdot (45 - 42) - 2 \cdot (36 - 30) + 3 \cdot (28 - 25) = 3 - 12 + 9 = 0.$$

Theorem 8.3.1 has a nice consequence involving transpose matrices.

Corollary 8.3.2

Let $\mathbf{A} \in \mathbb{F}^{n \times n}$ be given. Then $\det(\mathbf{A}) = \det(\mathbf{A}^T)$.

Proof. We use induction on n. If $n = 1$, $\mathbf{A} = \mathbf{A}^T$, so certainly $\det(\mathbf{A}) = \det(\mathbf{A}^T)$. Now assume $n \geq 2$ and that the corollary holds for $n-1$. Note that $\mathbf{A}(j;1)^T = \mathbf{A}^T(1;j)$, so that using the induction hypothesis, we may use that $\det(\mathbf{A}^T(1;j)) = \det(\mathbf{A}(j;1)^T) = \det(\mathbf{A}(j;1))$. Now using Laplace expansion of the determinant of $\mathbf{A}^T$ along the first row, we see that

$$\begin{aligned}\det(\mathbf{A}^T) &= \sum_{j=1}^{n}(-1)^{1+j}\cdot(\mathbf{A}^T)_{1j}\cdot\det(\mathbf{A}^T(1;j)) \\ &= \sum_{j=1}^{n}(-1)^{1+j}\cdot a_{j1}\cdot\det(\mathbf{A}(j;1)) \\ &= \det(\mathbf{A}),\end{aligned}$$

where in the last equality, we used Definition 8.1.2. This concludes the induction step and thereby the proof. □

Finally, one very important property of determinants that we want to mention here, is that determinants behave well with respect to matrix multiplication:

Theorem 8.3.3
Let $\mathbf{A}, \mathbf{B} \in \mathbb{F}^{n\times n}$ be given. Then $\det(\mathbf{A}\cdot\mathbf{B}) = \det(\mathbf{A})\cdot\det(\mathbf{B})$.

The interested reader can find a sketch of the proof at the end of this section, but this is not required reading. This theorem looks innocent, but has a number of consequences that all are quite important for us later on. We formulate them as a number of corollaries.

Corollary 8.3.4
Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Then $\mathbf{A}$ has an inverse if and only if $\det(\mathbf{A}) \neq 0$.

Proof. If $\mathbf{A}$ has an inverse $\mathbf{A}^{-1}$, then $\mathbf{A}\cdot\mathbf{A}^{-1} = \mathbf{I}_n$. Applying Theorem 8.3.3, we see that $\det(\mathbf{A})\cdot\det(\mathbf{A}^{-1}) = \det(\mathbf{I}_n) = 1$. For the last equality we used Proposition 8.1.1. But then $\det(\mathbf{A}) \neq 0$, since otherwise the product $\det(\mathbf{A})\cdot\det(\mathbf{A}^{-1})$ would be zero.

Conversely, assume that $\det(\mathbf{A}) \neq 0$. If we transform $\mathbf{A}$ using any sequence of elementary row operations to a matrix $\mathbf{B}$ in reduced row echelon form, then Corollary 8.2.2 and Theorems 8.2.6, 8.2.7 imply that $\det(\mathbf{B}) = d\cdot\det(\mathbf{A})$ for some nonzero constant $d \in \mathbb{F}$. Therefore $\det(\mathbf{B}) \neq 0$. This means in particular that $\mathbf{B}$ does not contain a zero row, since otherwise its determinant would be zero by Lemma 8.1.3. But then $\mathbf{B} = \mathbf{I}_n$, implying that $\mathbf{A}$ has rank n. As observed in Equation (7.10) and Corollary 7.3.5, this implies that $\mathbf{A}$ has an inverse. □

Corollary 8.3.5
Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Then the columns of $\mathbf{A}$ are linearly independent if and only if $\det(\mathbf{A}) \neq 0$.

Proof. This follows by combining Theorem 7.1.3, the previous corollary, and Corollary 7.3.5. □

Corollary 8.3.6

Let $\mathbf{A} \in \mathbb{F}^{n\times n}$ be given. Then $\det(\mathbf{A}) \neq 0$ if and only if the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$ only has the zero vector as solution.

Proof. This follows by combining Corollaries 6.4.5, 7.3.5, and 8.3.4. □

We will not prove Theorem 8.3.3 in detail, but the reader who would like to know more, can read the remainder of this section and get a good impression on why this theorem as well as Theorem 8.3.1 is true. The remainder of this section can be skipped on a first reading. If a reader is willing to accept the statements of Theorems 8.3.1 and 8.3.3 without proof, feel free to continue to the next chapter.

The key to understanding why Theorem 8.3.1 is true is the following:

Lemma 8.3.7

Let $f : \mathbb{F}^{n\times n} \to \mathbb{F}$ be a function that satisfies the following two conditions:

(i) $f(\mathbf{A}) = 0$ for all square matrices $\mathbf{A} \in \mathbb{F}^{n\times n}$ that have two identical rows.

(ii) For all matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ as given in Theorem 8.2.1, it holds that $f(\mathbf{C}) = c \cdot f(\mathbf{A}) + f(\mathbf{B})$.

Then $f(\mathbf{A}) = \det(\mathbf{A}) \cdot f(\mathbf{I}_n)$ for all $\mathbf{A} \in \mathbb{F}^{n\times n}$.

Proof. We only sketch the proof: the two conditions that f satisfies, are enough to deduce exactly how the value of f changes, when a matrix $\mathbf{A}$ is changed using an elementary row operation. In fact, many of the proofs in Section 8.2 can be reused. The two given conditions are also enough to deduce that $f(\mathbf{A}) = 0$ for all $\mathbf{A}$ that have a zero row. The outcome is then that f behaves exactly the same as the determinant under elementary row operations and that both f and the determinant take the value zero for matrices with a zero row.

Given any square matrix $\mathbf{A}$ and a sequence of elementary row operations that transform $\mathbf{A}$ into its reduced row echelon form, say $\mathbf{B}$, one can then compare the values of f and the determinant under these elementary row operations. The outcome is that $f(\mathbf{A}) = d \cdot f(\mathbf{B})$ for some constant $d \in \mathbb{F}$, but also $\det(\mathbf{A}) = d \cdot \det(\mathbf{B})$ for *the same* constant d. If $\mathbf{A}$ has rank strictly less than n, its reduced row echelon form $\mathbf{B}$ contains a zero row. But then $f(\mathbf{B}) = 0$ and $\det(\mathbf{B}) = 0$. If $\mathbf{A}$ has rank n, then $\mathbf{B} = \mathbf{I}_n$. Hence in this case $f(\mathbf{A}) = d \cdot f(\mathbf{I}_n)$, while $\det(\mathbf{A}) = d \cdot \det(\mathbf{I}_n) = d \cdot 1 = d$. In all cases, we see that $f(\mathbf{A}) = \det(\mathbf{A}) \cdot f(\mathbf{I}_n)$. □

Note that the determinant as we defined it in Definition 8.1.2 satisfies the two conditions from Lemma 8.3.7, see Proposition 8.2.5 and Theorem 8.2.1. To prove that Theorem 8.3.1 is valid, what one needs to do is to show that the function f one obtains by expanding a determinant along some row or some column, always has the properties mentioned in Lemma 8.3.7 and that $f(\mathbf{I}_n) = 1$. To a high extent, this can be done similarly to how we showed these things for the determinant defined in Definition 8.1.2.

Finally, let us give a sketch of the proof of Theorem 8.3.3:

Proof. To give a proof sketch of Theorem 8.3.3, we consider the function $f : \mathbb{F}^{n\times n} \to \mathbb{F}$ defined by $f(\mathbf{A}) = \det(\mathbf{A} \cdot \mathbf{B})$ for some arbitrarily chosen $\mathbf{B} \in \mathbb{F}^{n\times n}$. Using Proposition 8.2.5 and Theorem 8.2.1, one first shows that f satisfies the conditions in Lemma 8.3.7. One can then conclude that $f(\mathbf{A}) = \det(\mathbf{A}) \cdot f(\mathbf{I}_n)$ for all $\mathbf{A} \in \mathbb{F}^{n\times n}$. But then $\det(\mathbf{A} \cdot \mathbf{B}) = f(\mathbf{A}) = \det(\mathbf{A}) \cdot f(\mathbf{I}_n) = \det(\mathbf{A}) \cdot \det(\mathbf{B})$. In the last equality, we used that $\mathbf{I}_n \cdot \mathbf{B} = \mathbf{B}$. □

Chapter 9

Vector spaces

9.1 Definition and examples of vector spaces

In the previous chapters, we have worked with linear combinations of vectors from $\mathbb{F}^n$, where $\mathbb{F}$ is a field (typically $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$). We have seen that elements of $\mathbb{F}^n$ can be added and multiplied with scalars, that is to say, multiplied with elements from $\mathbb{F}$. It turns out to be a great advantage to take a more abstract point of view and describe several essential properties right from the start. One says that one gives these properties as axioms. This is similar in spirit to what we did when we defined a field. Also there, several properties of the real and complex numbers were put as axioms for such a field. In case of vectors and scalars, the result is the following:

Definition 9.1.1

A *vector space* over a field $\mathbb{F}$ is a set V of elements called *vectors*, together with two operations satisfying eight axioms. The first operation is called addition and denoted by $+$. It takes as input two elements $\mathbf{u}, \mathbf{v} \in V$ and returns a vector in V denoted by $\mathbf{u} + \mathbf{v}$. The second operation is called scalar multiplication and denoted by $\cdot$. It takes as input an element of $c \in \mathbb{F}$, in this context often called a *scalar*, and a vector $\mathbf{u} \in V$ and returns a vector in V denoted by $c \cdot \mathbf{u}$. The eight axioms that should be satisfied are:

(i) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$

(ii) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in V$

(iii) There exists a vector $\mathbf{0} \in V$ called the *zero vector*, such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$ for all $\mathbf{u} \in V$

(iv) For any $\mathbf{u} \in V$ there exists an element $-\mathbf{u} \in V$ called the additive inverse of $\mathbf{u}$, such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$

(v) $c \cdot (d \cdot \mathbf{u}) = (c \cdot d) \cdot \mathbf{u}$ for all $\mathbf{u} \in V$ and all $c, d \in \mathbb{F}$

(vi) $1 \cdot \mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$

(vii) $c \cdot (\mathbf{u} + \mathbf{v}) = c \cdot \mathbf{u} + c \cdot \mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in V$ and all $c \in \mathbb{F}$

(viii) $(c + d) \cdot \mathbf{u} = c \cdot \mathbf{u} + d \cdot \mathbf{u}$ for all $\mathbf{u} \in V$ and all $c, d \in \mathbb{F}$

Note that in item 5 in the formula $(c \cdot d) \cdot \mathbf{u}$, the first $\cdot$ (in $c \cdot d$) denotes multiplication in the field $\mathbb{F}$, while the second $\cdot$ denotes the scalar multiplication on the vector space V. Similarly in item 8, in the formula $(c + d) \cdot \mathbf{u} = c \cdot \mathbf{u} + d \cdot \mathbf{u}$, the first $+$ denotes addition in $\mathbb{F}$, while the second $+$ denotes addition in V.

Example 9.1.1

Let us take $V = \mathbb{F}^n$ together with the addition and scalar product we have defined before in Equations (7.1) and (7.2). This gives an example of a vector space. To verify this, one should check if the eight vector space axioms from Definition 9.1.1 are satisfied. Note that five of them were mentioned already in Theorem 7.1.1. The zero vector in the third axiom is simply the zero vector in $\mathbb{F}^n$, while the additive inverse of a vector required in axiom four is given as:

$$-\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} -v_1 \\ \vdots \\ -v_n \end{bmatrix}$$

This only leaves the sixth axiom, but

$$1 \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} 1 \cdot v_1 \\ \vdots \\ 1 \cdot v_n \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \text{ for all } \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \in \mathbb{F}^n.$$

We see that $\mathbb{F}^n$ is a vector space over the field $\mathbb{F}$.

A vector space over the field $\mathbb{R}$ is often called a *real vector space*. Similarly, a vector space over the field $\mathbb{C}$ is often called a *complex vector space*. We have in the previous chapters actually encountered examples of vector spaces already. Let us give a few.

Example 9.1.2

Consider the set $\mathbb{C}$ of complex numbers. If we take $\mathbb{F} = \mathbb{C}$ and $n = 1$ in Example 9.1.1, we obtain that we can see $\mathbb{C}$ as a vector space over itself. However, we can also see $\mathbb{C}$ as a vector space over the real numbers $\mathbb{R}$. Indeed, as $+$, we simply take addition of complex numbers. Since we can multiply any two complex numbers, we can certainly multiply a real number with a complex number. This gives us the needed scalar product. That all eight axioms from Definition 9.1.1 are satisfied, can be deduced from Theorems 3.2.2 and 3.2.3.

Example 9.1.3

Very similarly as in Example 9.1.2, one can view the set of real numbers $\mathbb{R}$ as a vector space over itself, but also as a vector space over the field of rational numbers $\mathbb{Q}$ (see Examples 2.1.4 and 6.1.1 for a description of the field $\mathbb{Q}$).

Example 9.1.4

Consider the set $\mathbb{F}^{m\times n}$ of $m \times n$ matrices with entries in a field $\mathbb{F}$. Using addition of matrices as defined in Definition 7.2.3 and scalar multiplication defined by:

$$c \cdot \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} c \cdot a_{11} & \cdots & c \cdot a_{1n} \\ \vdots & & \vdots \\ c \cdot a_{m1} & \cdots & c \cdot a_{mn} \end{bmatrix},$$

the first two items of Theorem 7.2.1 state that the first two vector field axioms are satisfied. The $m \times n$ matrix having zero entries only, plays the role of zero vector. All other axioms can be checked as well, but we leave this to the reader.

Example 9.1.5

Consider the set $\mathbb{C}[Z]$ of polynomials in the variable Z with coefficients in $\mathbb{C}$ as defined in Definition 4.1.1. On this set, we have as addition $+$, the usual addition of polynomials. Also, we can multiply any two polynomials, so we certainly can multiply a constant polynomial with another polynomial. This gives us a scalar product on $\mathbb{C}[Z]$. We will not do so here, but one can show that all eight axioms from Definition 9.1.1 are satisfied. Hence $\mathbb{C}[Z]$ is a vector space over $\mathbb{C}$.

Example 9.1.6

Consider the set F of all functions with domain $\mathbb{R}$ and codomain $\mathbb{R}$. If $f : \mathbb{R} \to \mathbb{R}$ and $r \in \mathbb{R}$ are given, one can define the function $r \cdot f : \mathbb{R} \to \mathbb{R}$ as $(r \cdot f)(a) = r \cdot f(a)$ for all $a \in \mathbb{R}$. This gives a scalar multiplication on F. Addition on F is defined in a similar way: if $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are given, the function $(f + g) : \mathbb{R} \to \mathbb{R}$ is defined as $(f + g)(a) = f(a) + g(a)$ for all $a \in \mathbb{R}$. One can verify that this gives F the structure of a vector space over $\mathbb{R}$. As zero vector, one takes the zero function: $\mathbf{0} : \mathbb{R} \to \mathbb{R}$, satisfying $a \mapsto 0$ for all $a \in \mathbb{R}$.

In all examples we have given above, it holds that the product of the scalar 0 with any vector is equal to the zero vector $\mathbf{0}$. However, none of the eight vector space axioms state that $0 \cdot \mathbf{u} = \mathbf{0}$ for all $\mathbf{u} \in V$. Fortunately, the eight vector space axioms are chosen well: one can deduce quite a lot from them, for example that the formula $0 \cdot \mathbf{u} = \mathbf{0}$ indeed is true for any vector space. We prove this and another intuitive formula in the following lemma:

Lemma 9.1.1

Let V be a vector space. Then

$$0 \cdot \mathbf{u} = \mathbf{0} \text{ for all } \mathbf{u} \in V \tag{9.1}$$

and

$$(-1)\cdot\mathbf{u} = -\mathbf{u} \text{ for all } \mathbf{u} \in V. \tag{9.2}$$

Proof. Using that $0 = 0+0$ and vector space axiom eight, we see that $0\cdot\mathbf{u} = (0+0)\cdot\mathbf{u} = 0\cdot\mathbf{u} + 0\cdot\mathbf{u}$. Adding $-(0\cdot\mathbf{u})$ on both sides and using vector space axioms four, one and three, we get

$$\begin{aligned}
\mathbf{0} &= 0\cdot\mathbf{u} + (-(0\cdot\mathbf{u})) \\
&= (0\cdot\mathbf{u} + 0\cdot\mathbf{u}) + (-(0\cdot\mathbf{u})) \\
&= 0\cdot\mathbf{u} + (0\cdot\mathbf{u} + (-(0\cdot\mathbf{u}))) \\
&= 0\cdot\mathbf{u} + \mathbf{0} \\
&= 0\cdot\mathbf{u}.
\end{aligned}$$

This shows the first part. The second part follows similarly. Since $0 = (1+(-1))$, we obtain that $0\cdot\mathbf{u} = (1+(-1))\cdot\mathbf{u} = 1\cdot\mathbf{u} + (-1)\cdot\mathbf{u}$. The left-hand side of this equation is equal to $\mathbf{0}$ by the first part of this lemma. Using this and vector space axiom six, we see that $\mathbf{0} = \mathbf{u} + (-1)\cdot\mathbf{u}$. Hence $(-1)\cdot\mathbf{u} = -\mathbf{u}$. □

9.2 Basis of a vector space

Very similar to what we did in Section 7.1 for vectors in $\mathbb{F}^m$, one can talk about a *linear combination* of vectors in the setting of general vector spaces. Explicitly, given a vector space V over a field $\mathbb{F}$, vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ and scalars $c_1, \ldots, c_n \in \mathbb{F}$, an expression of the form

$$c_1\cdot\mathbf{v}_1 + \cdots + c_n\cdot\mathbf{v}_n$$

is called a linear combination of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Likewise, the notion of linear (in)dependency of a finite sequence of vectors from Definition 7.1.1 generalizes directly to the setting of vector spaces:

Definition 9.2.1

Let V be a vector space over a field $\mathbb{F}$. A sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ is called *linearly independent* if and only if the equation $c_1\cdot\mathbf{v}_1 + \cdots c_n\cdot\mathbf{v}_n = \mathbf{0}$ with $c_1, \ldots, c_n \in \mathbb{F}$ only holds if $c_1 = \cdots = c_n = 0$.

If the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ is not linearly independent, one says that it is *linearly dependent*.

Basically, the only difference with Definition 7.1.1 is that $\mathbb{F}^m$ has been replaced with V. Also in the setting of general vector spaces, it is common to simply say that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly (in)dependent rather than saying that the sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly (in)dependent.

There is one complication concerning linear independence of vectors in general vector spaces. In Definition 9.2.1, we only consider *finitely many* vectors. It turns out that sometimes, we would like to be able to state that the vectors from a possibly infinite set are linearly independent. The following definition will allow us to do that:

Definition 9.2.2

Let V be a vector space over a field $\mathbb{F}$. The vectors in a set S of vectors are called *linearly independent* if and only if any finite sequence of distinct vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ from S is a linearly independent sequence of vectors.

If the vectors in S are not linearly independent, one says that they are *linearly dependent*.

Basically, in Definition 9.2.2, the number of vectors in the set S we consider may be infinite, but when determining if they are linearly independent, we only consider finitely many at the same time. Often we will work with finite sequences of vectors only, in which case Definition 9.2.1 can be used.

In Examples 7.1.2 and 7.1.3 we have already given examples of linearly dependent and linearly independent vectors in the vector space $\mathbb{R}^2$. Let us consider some more examples.

Example 9.2.1

In Example 9.1.2, we considered $\mathbb{C}$ as a vector space over $\mathbb{R}$. In this example, we give examples of linearly dependent and independent vectors. First of all, consider the elements 1 and i. To determine if these are linearly independent, we consider the equation $c_1 \cdot 1 + c_2 \cdot i = 0$, where $c_1, c_2 \in \mathbb{R}$. The reason we only allow c_1 and c_2 to be real numbers, is that we in this example consider $\mathbb{C}$ as a vector space over the field $\mathbb{R}$. Hence in Definition 9.2.1, we have $V = \mathbb{C}$ and $\mathbb{F} = \mathbb{R}$. In particular, the scalars only come from $\mathbb{R}$ by definition.

Returning to the equation $c_1 \cdot 1 + c_2 \cdot i = 0$, where $c_1, c_2 \in \mathbb{R}$, we see that the complex number $c_1 \cdot 1 + c_2 \cdot i$ is in rectangular form. Since two complex numbers are equal if and only if they have the same real and imaginary part, the equation $c_1 \cdot 1 + c_2 \cdot i = 0$ implies that $c_1 = 0$ and $c_2 = 0$. We conclude that the complex numbers 1 and i are linearly independent over $\mathbb{R}$.

Similarly, one can show that the complex numbers 2 and $1 + i$ are linearly independent. Indeed, suppose that $c_1 \cdot 2 + c_2 \cdot (1 + i) = 0$, for some $c_1, c_2 \in \mathbb{R}$. Considering real and imaginary part, we see that this implies that $2c_1 + c_2 = 0$ and $c_2 = 0$, whence $c_1 = c_2 = 0$.

As a final example, let us consider a sequence of three complex numbers, for example 2, $1 + i$ and $2 + 3i$. Since $-(1/2) \cdot 2 + 3 \cdot (1 + i) + (-1) \cdot (2 + 3i) = 0$, we see that the three complex numbers 2, $1 + i$, and $2 + 3i$ are linearly dependent over $\mathbb{R}$.

Example 9.2.2

In Example 9.1.4, we viewed the set of matrices $\mathbb{F}^{m \times n}$ as a vector space over $\mathbb{F}$. For any pair (i, j) satisfying $1 \leq i \leq m$ and $1 \leq j \leq n$, define the matrix $\mathbf{E}^{(i,j)} \in \mathbb{F}^{m \times n}$ to be the matrix having zero entries, except for the entry (i, j), which is equal to one. For $m = n = 2$, we have

for example

$$\mathbf{E}^{(1,1)} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{E}^{(1,2)} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \mathbf{E}^{(2,1)} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \text{ and } \mathbf{E}^{(2,2)} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Continuing with $m = n = 2$, we see that the matrices $\mathbf{E}^{(1,1)}, \mathbf{E}^{(1,2)}, \mathbf{E}^{(2,1)}, \mathbf{E}^{(2,2)}$ are linearly independent over $\mathbb{F}$. Indeed for any $c_1, c_2, c_3, c_4 \in \mathbb{F}$, one has

$$c_1 \cdot \mathbf{E}^{(1,1)} + c_2 \cdot \mathbf{E}^{(1,2)} + c_3 \cdot \mathbf{E}^{(2,1)} + c_4 \cdot \mathbf{E}^{(2,2)} = \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix}.$$

Hence $c_1 \cdot \mathbf{E}^{(1,1)} + c_2 \cdot \mathbf{E}^{(1,2)} + c_3 \cdot \mathbf{E}^{(2,1)} + c_4 \cdot \mathbf{E}^{(2,2)} = \mathbf{0}$ implies that $c_1 = c_2 = c_3 = c_4 = 0$.

For general m and n one can show similarly that the $m \times n$ matrices $\mathbf{E}^{(1,1)}, \ldots, \mathbf{E}^{(m,n)}$ are linearly independent over F.

Returning to $m = n = 2$, an example of a sequence of linearly dependent matrices is:

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

since

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

Example 9.2.3

Consider the complex vector space $\mathbb{C}[Z]$ from Example 9.1.5. Recall that 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. This implies in particular, that a polynomial $p(Z) = c_0 + c_1 Z + \cdots + c_n Z^n$ is equal to the zero polynomial if and only if $c_i = 0$ for all i. This shows that the set $\{1, Z, Z^2, \ldots\}$ is a set of linearly independent polynomials over $\mathbb{C}$.

All these examples show that the notion of linear independence carries over well to the setting of general vector spaces. With this in place, we come to a very important notion in the theory of vector spaces.

Definition 9.2.3

Let V be a vector space over a field $\mathbb{F}$. A set S of vectors in V is called a *basis* of V if the two following conditions are met:

(i) The vectors in S are linearly independent.

(ii) Any $\mathbf{v} \in V$ can be written as a linear combination of vectors in S.

An *ordered basis* $(\mathbf{v}_1, \mathbf{v}_2, \ldots)$ of V is a list of vectors in V, such that the set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ is a basis of V.

It turns out that any vector space has a basis and we will freely use this fact. A reader who has time and motivation for a bit of extra material about this is referred to Section 9.4, but this is not required reading. If a vector space has a finite basis, i.e., if the set S containing the basis vectors, is finite, we can enumerate the elements in S and write $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Then $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is a finite ordered basis of V. Hence any vector space with a finite basis has an ordered basis.

Before giving examples, let us give one lemma and one more definition.

Lemma 9.2.1

Let V be a vector space over a field $\mathbb{F}$ that has a finite ordered basis $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. Then any vector $\mathbf{v} \in V$ can be written in exactly one way as a linear combination of the basis vectors.

Proof. The second part of Definition 9.2.3 guarantees that any vector $\mathbf{v} \in V$ can be written as a linear combination of the basis vectors, say $\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n$ for certain $c_1, \ldots, c_n \in \mathbb{F}$. What we need to show, is that this is the only way to write $\mathbf{v}$ as a linear combination of the basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Suppose therefore that $\mathbf{v} = d_1 \cdot \mathbf{v}_1 + \cdots d_n \cdot \mathbf{v}_n$ for certain $d_1, \ldots, d_n \in \mathbb{F}$. We wish to show that $c_1 = d_1, \ldots, c_n = d_n$. First of all, we have

$$c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n = \mathbf{v} = d_1 \cdot \mathbf{v}_1 + \cdots d_n \cdot \mathbf{v}_n.$$

Therefore,

$$c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n - (d_1 \cdot \mathbf{v}_1 + \cdots d_n \cdot \mathbf{v}_n) = \mathbf{0},$$

which in turn implies that

$$(c_1 - d_1) \cdot \mathbf{v}_1 + \cdots (c_n - d_n) \cdot \mathbf{v}_n = \mathbf{0}.$$

However, since the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent (this follows from the first part of Definition 9.2.3), we see that $c_1 - d_1 = 0, \ldots, c_n - d_n = 0$. But then $c_1 = d_1, \ldots, c_n = d_n$, which is what we wanted to show. □

This Lemma 9.2.1 motivates the following definition:

Definition 9.2.4

Let V be a vector space over a field $\mathbb{F}$ that has a finite ordered basis $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. If for $\mathbf{v} \in V$, we have

$$\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n,$$

then we define

$$[\mathbf{v}]_\beta = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{F}^n$$

to be the *coordinate vector* of $\mathbf{v}$ with respect to the ordered basis β. One also says that $[\mathbf{v}]_\beta$ is the *β-coordinate vector* of $\mathbf{v}$.

The function sending a vector of V to its β-coordinate vector, has several nice properties. Two of them will be useful later on.

Lemma 9.2.2

Let V be a vector space over a field $\mathbb{F}$ that has a finite ordered basis β. Then we have:

$$[\mathbf{u}+\mathbf{v}]_\beta = [\mathbf{u}]_\beta + [\mathbf{v}]_\beta \text{ for all } \mathbf{u},\mathbf{v} \in V$$

and

$$[c\cdot\mathbf{v}]_\beta = c\cdot[\mathbf{v}]_\beta \text{ for all } c\in\mathbb{F} \text{ and } \mathbf{v}\in V.$$

Proof. We prove the first item only and leave the proof of the second one to the reader. Let us say that the ordered basis β is given by $\mathbf{v}_1,\dots,\mathbf{v}_n$. If $\mathbf{u} = c_1\cdot\mathbf{v}_1+\cdots c_n\cdot\mathbf{v}_n$ and $\mathbf{v} = d_1\cdot\mathbf{v}_1+\cdots d_n\cdot\mathbf{v}_n$, then $\mathbf{u}+\mathbf{v} = (c_1+d_1)\cdot\mathbf{v}_1+\cdots(c_n+d_n)\cdot\mathbf{v}_n$. Hence

$$[\mathbf{u}+\mathbf{v}]_\beta = \begin{bmatrix} c_1+d_1 \\ \vdots \\ c_n+d_n \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = [\mathbf{u}]_\beta + [\mathbf{v}]_\beta.$$

□

Now we will use this lemma to prove a theorem involving linear independence of vectors.

Theorem 9.2.3

Let V be a vector space over a field $\mathbb{F}$ that has a finite ordered basis β consisting of n vectors. Suppose we are given $\mathbf{u}_1,\dots,\mathbf{u}_\ell \in V$ and $c_1,\dots,c_\ell\in\mathbb{F}$. Then

$$c_1\cdot\mathbf{u}_1+\cdots+c_\ell\cdot\mathbf{u}_\ell = \mathbf{0} \text{ if and only if } c_1\cdot[\mathbf{u}_1]_\beta+\cdots+c_\ell[\cdot\mathbf{u}_\ell]_\beta = \mathbf{0}.$$

In particular, the vectors $\mathbf{u}_1,\dots,\mathbf{u}_\ell$ in V are linearly independent if and only if the vectors $[\mathbf{u}_1]_\beta,\dots,[\mathbf{u}_\ell]_\beta$ in $\mathbb{F}^n$ are linearly independent.

Proof. A vector $\mathbf{v}$ in V is the zero vector if and only if its β-coordinate vector is the zero vector. Hence $c_1\cdot\mathbf{u}_1+\cdots+c_\ell\cdot\mathbf{u}_\ell = \mathbf{0}$ if and only if $[c_1\cdot\mathbf{u}_1+\cdots+c_\ell\cdot\mathbf{u}_\ell]_\beta = \mathbf{0}$. Using Lemma 9.2.2 repeatedly, we can also deduce that $[c_1\cdot\mathbf{u}_1+\cdots+c_\ell\cdot\mathbf{u}_\ell]_\beta = c_1\cdot[\mathbf{u}_1]_\beta+\cdots+c_\ell\cdot[\mathbf{u}_\ell]_\beta$. Hence the first part of the theorem follows. The second part follows directly from the first part. □

This theorem basically reduces the question of linear (in)dependence of vectors in V to a question of linear (in)dependence of vectors in $\mathbb{F}^n$. However, for $\mathbb{F}^n$, we already have techniques at our disposal, notably Theorem 7.1.3.

Example 9.2.4

Let $\mathbb{F} = \mathbb{R}$ and $V = \mathbb{R}^2$. We claim that the vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

form an ordered basis β for $\mathbb{R}^2$. Indeed, these vectors are linearly independent (the reader is encouraged to check this), and any vector is a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2$, since

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + v_2 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

This means that in this case $[\mathbf{v}]_\beta = \mathbf{v}$.

Now let γ be the sequence of vectors

$$\mathbf{u} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } \mathbf{v} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in \mathbb{R}^2.$$

We have seen in Example 7.1.3 that these two vectors are linearly independent. Further, one can show that any vector in $\mathbb{R}^2$ can be written as a linear combination of $\mathbf{u}$ and $\mathbf{v}$. Indeed, given $v_1, v_2 \in \mathbb{R}$, the equation

$$c_1 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

gives rise to a system of two linear equations in the variables c_1, c_2. Solving this system, one can show that for any $v_1, v_2 \in \mathbb{R}$, we have

$$c_1 = -\frac{v_1}{3} + \frac{2v_2}{3} \quad \text{and} \quad c_2 = \frac{2v_1}{3} - \frac{v_2}{3}$$

so that

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \left(-\frac{v_1}{3} + \frac{2v_2}{3}\right) \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \left(\frac{2v_1}{3} - \frac{v_2}{3}\right) \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

This means that $\gamma = (\mathbf{u}, \mathbf{v})$ is an ordered basis of $\mathbb{R}^2$. Moreover, from the above we see that

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_\gamma = \begin{bmatrix} -v_1/3 + 2v_2/3 \\ 2v_1/3 - v_2/3 \end{bmatrix}.$$

This first part of Example 9.2.4 can be expanded further: as in Section 7.3, let us denote the i-th column of the identity matrix $\mathbf{I}_n \in \mathbb{F}^{n \times n}$ by $\mathbf{e}_i$ for $i = 1, \ldots, n$. In other words: the vector $\mathbf{e}_i$ has 1 as its i-th coordinate and zeroes everywhere else. These vectors form an ordered

basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of the vector space $\mathbb{F}^n$ called the *standard (ordered) basis*. For the sake of completeness, let us show that they form an ordered basis:

Proposition 9.2.4

The vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ form an ordered basis of the vector space $\mathbb{F}^n$ over $\mathbb{F}$.

Proof. According to Definition 9.2.3, we need to check two things:

(i) The vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are linearly independent.

(ii) Any vector in $\mathbb{F}^n$ can be written as a linear combination of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

The first item follows from the observation that

$$c_1 \cdot \mathbf{e}_1 + c_2 \cdot \mathbf{e}_2 + \cdots + c_n \cdot \mathbf{e}_n = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

Indeed, this equation implies that if a linear combination is equal to the zero vector in $\mathbb{F}^n$, then all scalars $c_1, \ldots, c_n$ are zero. The second item follows, since if $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}^n$ is given, then

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = v_1 \cdot \mathbf{e}_1 + v_2 \cdot \mathbf{e}_2 + \cdots + v_n \cdot \mathbf{e}_n.$$

□

Similarly as in Example 9.2.4, if β is the standard ordered basis of $\mathbb{F}^n$, then $[\mathbf{v}]_\beta = \mathbf{v}$ for all $\mathbf{v} \in \mathbb{F}^n$. Note though that just as in Example 9.2.4, the vector space $\mathbb{F}^n$ has many more possible ordered bases. Now let us continue with giving examples of bases of vector spaces.

Example 9.2.5

Continuing Examples 9.1.2 and 9.2.1, we know that the complex numbers 1 and i are linearly independent over $\mathbb{R}$. They form an ordered basis $(1, i)$, which we denote by β, since any complex number is a linear combination of 1 and i over the real numbers. More specifically, for any $a, b \in \mathbb{R}$, we have $a + bi = a \cdot 1 + b \cdot i$. Therefore, for $a, b \in \mathbb{R}$, we have

$$[a + bi]_\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2.$$

Hence $[a + bi]_\beta$ is equal to the rectangular coordinates of the complex number $a + bi$.

There are many more possible bases (and hence ordered bases) for $\mathbb{C}$ when viewed as vector space over $\mathbb{R}$. For example, $(2, 1 + i)$ is a possible ordered basis. Indeed, we have

already seen in Example 9.2.1 that the complex numbers 2 and $1+i$ are linearly independent over $\mathbb{R}$. Also any complex number can be written as a linear combination with coefficients in $\mathbb{R}$ of 2 and $1+i$. To see this, we need to check that for a given complex number $a+bi$, where $a, b \in \mathbb{R}$, the equation $a+bi = c_1 \cdot 2 + c_2 \cdot (1+i)$ has a solution $c_1, c_2 \in \mathbb{R}$. Considering real and imaginary parts, we see that $a = 2c_1 + c_2$ and $b = c_2$. Hence we have as solution $c_2 = b$ and $c_1 = (a-c_2)/2 = (a-b)/2$. Denoting the ordered basis $(2, 1+i)$ by γ, we have

$$[a+bi]_\gamma = \begin{bmatrix} (a-b)/2 \\ b \end{bmatrix} \in \mathbb{R}^2.$$

Example 9.2.6

Continuing Examples 9.1.4 and 9.2.2, we can find an ordered basis β of the vector space $\mathbb{F}^{m\times n}$ over $\mathbb{F}$. This ordered basis is $(\mathbf{E}^{(1,1)}, \dots, \mathbf{E}^{(m,n)})$. We have already seen that the matrices $\mathbf{E}^{(1,1)}, \dots, \mathbf{E}^{(m,n)}$ are linearly independent, while any matrix $\mathbf{A} = (a_{ij})_{1\le i\le m; 1\le j\le n}$ can be written as a linear combination of them, namely $\mathbf{A} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\mathbf{E}^{(i,j)}$.

Specifically for $m = n = 2$, the matrices $\mathbf{E}^{(1,1)}, \mathbf{E}^{(1,2)}, \mathbf{E}^{(2,1)}, \mathbf{E}^{(2,2)}$ form an ordered basis $\beta = (\mathbf{E}^{(1,1)}, \mathbf{E}^{(1,2)}, \mathbf{E}^{(2,1)}, \mathbf{E}^{(2,2)})$ and we have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}_\beta = \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \end{bmatrix}.$$

Example 9.2.7

In this example, we again consider the complex vector space $\mathbb{C}[Z]$ from Examples 9.1.5 and 9.2.3. From these examples, we already know that the set $\{1, Z, Z^2, \dots\}$ is a set of linearly independent polynomials over $\mathbb{C}$. However, by definition of polynomials, any polynomial is a linear combination over $\mathbb{C}$ of finitely many elements from this set. Therefore the set $\{1, Z, Z^2, \dots\}$ is in fact a basis of the complex vector space $\mathbb{C}[Z]$. This is an example of a vector space having an infinite basis.

It turns out that for a given vector space V over a field $\mathbb{F}$, the number of vectors in a basis of V is always the same. Later in this section, we will prove this in the special case where the number of vectors in a basis is finite. In general, the number of elements in a basis of V is called the *dimension* of the vector space V. A common notation for the dimension of a vector space V is: $\dim(V)$ or just $\dim V$. If one wants to make clear over which field $\mathbb{F}$ the vector space is defined, one writes $\dim_{\mathbb{F}}(V)$ or $\dim_{\mathbb{F}} V$. If the number of vectors in a basis is finite, one says that V has finite dimension, otherwise one says that V has infinite dimension, which can also be expressed in a formula as: $\dim V = \infty$.

Example 9.2.8

Let us compute the dimensions of various examples of vector spaces that we have encountered

so far. First of all from Example 9.2.4, we see that $\dim_{\mathbb{R}}(\mathbb{R}^2) = 2$. Much more generally one has $\dim_{\mathbb{F}}(\mathbb{F}^n) = n$, since a possible basis of $\mathbb{F}^n$ is formed by the n vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

A special case of the above is when $\mathbb{C}$ is viewed as a vector space over itself. Then it has dimension one: $\dim_{\mathbb{C}}(\mathbb{C}) = 1$ (a possible basis is formed by the complex number 1). However, if $\mathbb{C}$ is viewed as a vector space over $\mathbb{R}$, a basis is given by $\{1, i\}$ as we have seen in Example 9.2.5. Hence $\dim_{\mathbb{R}}(\mathbb{C}) = 2$.

The vector space of $m \times n$ matrices $\mathbb{F}^{m\times n}$ has a basis consisting of the mn matrices $\mathbf{E}^{(i,j)}$ with $1 \leq i \leq m$ and $1 \leq j \leq n$, as we have seen in Example 9.2.6. Hence $\dim_{\mathbb{F}}(\mathbb{F}^{m\times n}) = mn$.

We have seen in Example 9.2.3 that the complex vector space $\mathbb{C}[Z]$ has a basis with infinitely many elements, namely $\{1, Z, Z^2, \ldots\}$. Hence $\dim_{\mathbb{C}}(\mathbb{C}[Z]) = \infty$.

Theorem 9.2.5

If V has a finite basis consisting of n vectors, any other set of linearly independent vectors in V has at most n elements.

Proof. Let us denote the basis vectors by $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and denote the resulting ordered basis by β. We will prove the theorem by contradiction. Assume therefore that there exists a set of at least $n+1$ linearly independent vectors, say $\mathbf{w}_1, \ldots, \mathbf{w}_{n+1}$. Since β is an ordered basis, we can find scalars $a_{ij} \in \mathbb{F}$ such that

$$\mathbf{w}_j = a_{1j}\mathbf{v}_1 + \cdots + a_{nj}\mathbf{v}_n \text{ for } j = 1, \ldots, n+1.$$

Now let $\mathbf{A} = (a_{ij}) \in \mathbb{F}^{n\times(n+1)}$ be the matrix with entries a_{ij}. Note that the j-th column in $\mathbf{A}$ is equal to $[\mathbf{w}_j]_\beta$. Since $\mathbf{A}$ has n rows, its rank $\rho(\mathbf{A})$ is at most n. Since $\mathbf{A}$ has $n+1$ columns, this implies that $\rho(\mathbf{A}) < n+1$. Then by Corollary 6.4.5, we see that the homogeneous system with coefficient matrix $\mathbf{A}$ has nonzero solutions. Let $(c_1, \ldots, c_{n+1}) \in \mathbb{F}^{n+1}$ be such a nonzero solution. Then we have

$$c_1 \cdot \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} + \cdots + c_{n+1} \cdot \begin{bmatrix} a_{1\,n+1} \\ \vdots \\ a_{n\,n+1} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Now recall that the j-th column in $\mathbf{A}$ is equal to $[\mathbf{w}_j]_\beta$. This means that we have $c_1 \cdot [\mathbf{w}_1]_\beta + c_{n+1} \cdot [\mathbf{w}_{n+1}]_\beta = \mathbf{0}$. Since from Lemma 9.2.2 one can deduce that $[c_1 \cdot \mathbf{w}_1 + \cdots + c_{n+1} \cdot \mathbf{w}_{n+1}]_\beta = c_1 \cdot [\mathbf{w}_1]_\beta + c_{n+1} \cdot [\mathbf{w}_{n+1}]_\beta$, we may conclude that $[c_1 \cdot \mathbf{w}_1 + \cdots + c_{n+1} \cdot \mathbf{w}_{n+1}]_\beta = \mathbf{0}$. Hence $c_1 \cdot \mathbf{w}_1 + \cdots + c_{n+1} \cdot \mathbf{w}_{n+1} = \mathbf{0}$. Since $(c_1, \ldots, c_{n+1})$ was not the zero vector, we conclude that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{n+1}$ are not linearly independent after all. This contradiction shows that the assumption that there exists sets with at least $n+1$ linearly independent vectors was wrong. Hence the theorem is true. □

Corollary 9.2.6

If V has a finite basis consisting of n vectors, any other basis for V contains precisely n vectors as well.

Proof. Let S be a basis of V consisting of n vectors and T any other basis. Since the vectors T are linearly independent, Theorem 9.2.5 implies that the number of vectors in T is at most n. Let us denote by m, the number of vectors in T. What we have just shown is that $m \leq n$. Now applying Theorem 9.2.5 again, but now taking T as a basis, we can conclude that the number of elements in S is at most m, that is: $n \leq m$. Combining the inequalities $m \leq n$ and $n \leq m$, we conclude that $n = m$, which is what we wanted to show. □

This corollary justifies the definition of dimension of a vector space V as the number of basis vectors in the finite dimensional case: no matter which basis of V you pick, it will contain precisely the same number of vectors. As mentioned before, the basis vectors themselves typically will be different when comparing two possible bases. In fact, for finite dimensional vector spaces, we can characterize all possible bases:

Theorem 9.2.7

Let V be a vector space over a field $\mathbb{F}$ of dimension n. Then any set of n linearly independent vectors in V is a basis for V.

Proof. Let us denote the vectors in some basis of V as $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and let us write β for the corresponding ordered basis. Further, let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be n linearly independent vectors in V. To show that these form a basis, all we need to check is item 2 in Definition 9.2.3. That is to say, we need to show that any $\mathbf{v} \in V$ can be written as a linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_n$. First of all, since β is a basis, we can find $a_{ij} \in \mathbb{F}$ such that

$$\mathbf{w}_j = a_{1j} \cdot \mathbf{v}_1 + \cdots + a_{nj} \cdot \mathbf{v}_n \text{ for } j = 1, \ldots, n,$$

or equivalently using the summation symbol:

$$\mathbf{w}_j = \sum_{i=1}^{n} a_{ij} \cdot \mathbf{v}_i \text{ for } j = 1, \ldots, n. \tag{9.3}$$

Now let $\mathbf{A} = (a_{ij}) \in \mathbb{F}^{n \times n}$ be the matrix with entries a_{ij}. As in the proof of Theorem 9.2.5, note that the j-th column in $\mathbf{A}$ is equal to $[\mathbf{w}_j]_\beta$. We claim that these columns are linearly independent vectors in $\mathbb{F}^n$. To see why, suppose that $c_1 \cdot [\mathbf{w}_1]_\beta + c_n \cdot [\mathbf{w}_n]_\beta = \mathbf{0}$ for certain $c_1, \ldots, c_n \in \mathbb{F}$. Then $[c_1 \cdot \mathbf{w}_1 + \cdots + c_n \cdot \mathbf{w}_n]_\beta = \mathbf{0}$, implying that $c_1 \cdot \mathbf{w}_1 + \cdots + c_n \cdot \mathbf{w}_n = \mathbf{0}$. Using that the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are linearly independent, we conclude that $c_1 = 0, \ldots, c_n = 0$, which is what we wanted to show to prove our claim. Now using Theorem 7.1.3 and Corollary 7.3.5, we conclude that the matrix $\mathbf{A}$ has an inverse matrix $\mathbf{A}^{-1}$.

Now let us return to what we want to show: $\mathbf{v} \in V$ can be written as a linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Since $\mathbf{v}$ is a linear combination of the basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, it is enough to show that each of the basis vectors themselves can be written as a linear combination of $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Let us write $\mathbf{A}^{-1} = (c_{ij})_{1 \leq i \leq n; 1 \leq j \leq n}$. We claim that:

$$\mathbf{v}_j = c_{1j} \cdot \mathbf{w}_1 + \cdot + c_{nj} \cdot \mathbf{w}_n \text{ for } j = 1, \ldots, n.$$

Equivalently, using the summation symbol, we claim that:

$$\mathbf{v}_j = \sum_{k=1}^{n} c_{kj} \cdot \mathbf{w}_k \text{ for } k = 1, \ldots, n.$$

To show the claim, first we use Equation (9.3) to see that:

$$\begin{aligned}
\sum_{k=1}^{n} c_{kj} \cdot \mathbf{w}_k &= \sum_{k=1}^{n} c_{kj} \cdot \left(\sum_{i=1}^{n} a_{ik} \cdot \mathbf{v}_i \right) \\
&= \sum_{k=1}^{n} \sum_{i=1}^{n} c_{kj} \cdot a_{ik} \cdot \mathbf{v}_i \\
&= \sum_{k=1}^{n} \sum_{i=1}^{n} a_{ik} \cdot c_{kj} \cdot \mathbf{v}_i \\
&= \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ik} \cdot c_{kj} \cdot \mathbf{v}_i \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} a_{ik} \cdot c_{kj} \right) \cdot \mathbf{v}_i.
\end{aligned}$$

Now note that the expression $\sum_{k=1}^{n} a_{ik} \cdot c_{kj}$ is the (i,j)-th entry of the matrix product $\mathbf{A} \cdot \mathbf{A}^{-1}$. However, since $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_n$, we see that $\sum_{k=1}^{n} a_{ik} \cdot c_{kj} = 1$ if $i = j$ and $\sum_{k=1}^{n} a_{ik} \cdot c_{kj} = 0$ otherwise. Hence we can conclude that $\sum_{k=1}^{n} c_{kj} \cdot \mathbf{w}_k = \mathbf{v}_j$, which is exactly what we wanted to show. □

9.3 Subspaces of a vector space

Given a vector space V over some field $\mathbb{F}$, it can happen that a subset W of V is closed under the scalar multiplication and the vector addition as defined on V. The word "closed" is just a way of saying that if $\mathbf{v} \in W$ and $c \in \mathbb{F}$, then $c \cdot \mathbf{v} \in W$ and if $\mathbf{u}, \mathbf{v} \in W$, then $\mathbf{u} + \mathbf{v} \in W$. Since V is a vector space, we always have $c \cdot \mathbf{v} \in V$ and $\mathbf{u} + \mathbf{v} \in V$, but if W is closed under the scalar multiplication and addition, the vectors $c \cdot \mathbf{v}$ and $\mathbf{u} + \mathbf{v}$ end up in W again. Let us consider two examples of this:

Example 9.3.1

Let us consider the complex vector space $\mathbb{C}^2$ and consider the subset $W = \{(z, 2 \cdot z) \mid z \in \mathbb{C}\}$. Then adding two elements of W yields another element of W, since $(z, 2 \cdot z) + (w, 2 \cdot w) = (z + w, 2 \cdot (z + w))$ for all $z, w \in \mathbb{C}$. Also multiplying an element from W with a scalar $c \in \mathbb{C}$ yields an element of W, since $c \cdot (z, 2 \cdot z) = (c \cdot z, 2 \cdot (c \cdot z))$. In fact W is a vector space using this scalar multiplication and addition. For example, one has $(0,0) \in W$, since $(0,0) = (0, 2 \cdot 0)$. Also $-(z, 2 \cdot z) = ((-z), 2 \cdot (-z))$ for any $z \in \mathbb{C}$, which shows that if $\mathbf{v} \in W$, then also $-\mathbf{v} \in W$. The reader is encouraged to check the remaining axioms of a vector space. Note that $\dim_{\mathbb{C}}(W) = 1$ (a possible basis is given by $\{(1,2)\}$).

Example 9.3.2

Consider the vector space $\mathbb{R}^{2\times 2}$ of 2 by 2 matrices with coefficients in $\mathbb{R}$. As we have seen, this is a real vector space of dimension four. Now let D be the subset of $\mathbb{R}^{2\times 2}$ consisting of all diagonal matrices, that is:

$$D = \left\{ \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \right\}.$$

Then the set D is closed under scalar multiplication and matrix addition. What this means is that if $\mathbf{A}, \mathbf{B} \in D$ and $c \in \mathbb{F}$, then $c \cdot \mathbf{A} \in D$ and $\mathbf{A} + \mathbf{B} \in D$. Let us check this. If $\mathbf{A}$ has diagonal elements λ_1 and λ_2 and $\mathbf{B}$ has diagonal elements μ_1 and μ_2, then:

$$c \cdot \mathbf{A} = c \cdot \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} c \cdot \lambda_1 & 0 \\ 0 & c \cdot \lambda_2 \end{bmatrix} \in D,$$

and

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} + \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 + \mu_1 & 0 \\ 0 & \lambda_2 + \mu_2 \end{bmatrix} \in D$$

One can check that D is in fact a real vector space of dimension two: a possible ordered basis is $(\mathbf{E}^{(1,1)}, \mathbf{E}^{(2,2)})$.

To capture these type of examples, we have the following:

Definition 9.3.1

Let V be a vector space over a field $\mathbb{F}$. A *subspace* of V is a subset W of V that is a vector space over $\mathbb{F}$ under the scalar multiplication and vector addition defined on V.

In other words, if $W \subseteq V$ is closed under the scalar multiplication and vector addition that V has, W "inherits" these operations. If W with these operations satisfies all vector space axioms from Definition 9.1.1, it is called a subspace of W. Any vector space V has at least two subspace: V itself can be seen as a subspace, and also the subspace $\{\mathbf{0}\}$ containing only the zero vector of V. In general, V has many more subspaces. In all cases, however, one can say the following about the dimension of a subspace:

Lemma 9.3.1

Let V be a vector space over a field $\mathbb{F}$ of dimension n and W a subspace of V. Then $\dim W \leq n$.

Proof. We know that V has a basis with n vectors and that W has a basis with $\dim W$ vectors. The basis vectors of W form a sequence of $\dim W$ linearly independent vectors. Hence Theorem 9.2.5 implies that $\dim W \leq n$. □

Since V already satisfies all vector space axioms, it turns out not to be necessary to check them all when investigating if a subset W is a subspace. More precisely, we have the following lemma:

Lemma 9.3.2

Let V be a vector space over $\mathbb{F}$ and W a nonempty subset of V. Then W is a subspace of V if the following is satisfied:

$$\text{for all } \mathbf{u}, \mathbf{v} \in W \text{ and all } c \in \mathbb{F} \text{ it holds that } \mathbf{u} + c \cdot \mathbf{v} \in W. \tag{9.4}$$

Proof. First let us show that W is closed under the scalar multiplication and vector addition of V. First of all, since W is not empty, it contains at least one vector, say $\mathbf{w}$. Then choosing $\mathbf{u} = \mathbf{w}$ and $\mathbf{v} = \mathbf{w}$ in Equation (9.4), we can conclude that the vector $\mathbf{w} + (-1) \cdot \mathbf{w}$ is also in W. Using for example Equation (9.2), this implies that $\mathbf{0} \in W$. Now that we know this, we can apply Equation (9.4) again, but now with $\mathbf{u} = \mathbf{0}$ and $\mathbf{v} \in W$ chosen arbitrarily. We can hence conclude that for arbitrary $\mathbf{v} \in W$, also $c \cdot \mathbf{v}$ is in W. This shows that W is closed under scalar multiplication. Applying Equation (9.4) for arbitrary $\mathbf{u}, \mathbf{v} \in W$ and $c = 1$, we conclude that $\mathbf{u} + \mathbf{v}$ is in W. Hence W is closed under vector addition.

Now let us show that W is a vector space by considering the eight vector space axioms from Definition 9.1.1. Items 1, 2, 5, 6, 7, and 8 actually hold for all vectors in V and therefore certainly for all vectors in a subset of V. Therefore, all that remains to be checked is that items 3 and 4 are satisfied. Item 3 is fulfilled, since we already have shown that Equation (9.4) implies that $\mathbf{0} \in W$. As for item 4, if $\mathbf{v} \in W$, then $(-1) \cdot \mathbf{v} \in W$, since W is closed under scalar multiplication. But by Equation (9.2), $(-1) \cdot \mathbf{v} = -\mathbf{v}$, so that indeed the additive inverse $-\mathbf{v}$ is in W, for all $\mathbf{v}$ in W. □

Example 9.3.3

Using Lemma 9.3.2, it is not hard to show that the subsets W and D from Examples 9.3.1 and 9.3.2 are subspaces. The reader is encouraged to check that the condition in Equation (9.4) is satisfied for these examples.

Example 9.3.4

Let $C_\infty(\mathbb{R})$ be the set of all infinitely differentiable functions $f : \mathbb{R} \to \mathbb{R}$. It is out of scope of this text to define very precisely what an infinitely differentiable function is, but roughly speaking this means the following: if for all $t \in \mathbb{R}$ the limit $\lim_{a \to 0}(f(t+a) - f(t))/a$ exists, we can define the derivative of f, denoted by f', to be the function $f' : \mathbb{R} \to \mathbb{R}$ with $t \mapsto \lim_{a \to 0}(f(t+a) - f(t))/a$. An infinitely differentiable function $f : \mathbb{R} \to \mathbb{R}$ has the property that one can keep on differentiating it as often as one wants. In particular, not only its derivative f' exists, but also the derivative of f' (denoted by f'' or $f^{(2)}$), the derivative of f'' (denoted by f''' or $f^{(3)}$), and so on. More generally for any positive integer n, one denotes with $f^{(n)}$ the n-th derivative of f. More precisely, one recursively defines the n-th derivative as follows:

$$f^{(n)} = \begin{cases} f & \text{if } n = 0, \\ (f^{(n-1)})' & \text{if } n > 0. \end{cases}$$

The set $C_\infty(\mathbb{R})$ is a subspace of the vector space F from Example 9.1.6. This amounts to showing that if $f, g \in C_\infty(\mathbb{R})$ and $c \in \mathbb{R}$, then also $f + c \cdot g \in C_\infty(\mathbb{R})$. In fact one can show inductively that $(f + c \cdot g)^{(n)} = f^{(n)} + c \cdot g^{(n)}$ for any $n \in \mathbb{Z}_{\geq 0}$. In particular, $f + c \cdot g$ is infinitely differentiable, which is what we needed to show.

There is one specific way to construct a subspace, which we will get in to now.

Definition 9.3.2

Let V be a vector space over $\mathbb{F}$ and S a set of vectors from V. Then the *span* of S, denotes by $\text{Span}(S)$ is the set of all possible linear combinations of vectors from S. In particular, if $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then

$$\text{Span}(S) = \{c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n \,|\, c_1, \ldots, c_n \in \mathbb{F}\}.$$

It is customary to define $\text{Span}(\emptyset) = \{\mathbf{0}\}$. As a consequence one also says that the empty set $\emptyset$ is a basis for the vector space $\{\mathbf{0}\}$. One can verify that for any subset $S \subseteq V$, the set $\text{Span}(S)$ is in fact a subspace of V, using for example Lemma 9.3.2. If W is a given subspace of a vector space V and $W = \text{Span}(S)$, one says that the vectors in S span W. One also says in this situation that W is spanned by the vectors in S. The vectors in a basis of W will certainly span W, but in general a set of vectors spanning W need not be linearly independent.

Example 9.3.5

Consider the real vector space $V = \mathbb{R}^2$ and let

$$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Let us determine the span of $\mathbf{v}_1$, or in other words $\text{Span}(S)$ with $S = \{\mathbf{v}_1\}$. Using Definition 9.3.2 we find:

$$\begin{aligned} \text{Span}(\{\mathbf{v}_1\}) &= \{c_1 \cdot \mathbf{v}_1 \,|\, c_1 \in \mathbb{R}\} \\ &= \left\{ c_1 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} \,|\, c_1 \in \mathbb{R} \right\} \\ &= \left\{ \begin{bmatrix} c_1 \cdot 2 \\ c_1 \cdot 1 \end{bmatrix} \,|\, c_1 \in \mathbb{R} \right\} \\ &= \left\{ \begin{bmatrix} 2c_1 \\ c_1 \end{bmatrix} \,|\, c_1 \in \mathbb{R} \right\}. \end{aligned}$$

Graphically the situation is that the span of $\mathbf{v}_1$ consists of all vectors lying on the line through $\mathbf{v}_1$ (see Figure 9.1, where the span is indicated by a blue line).

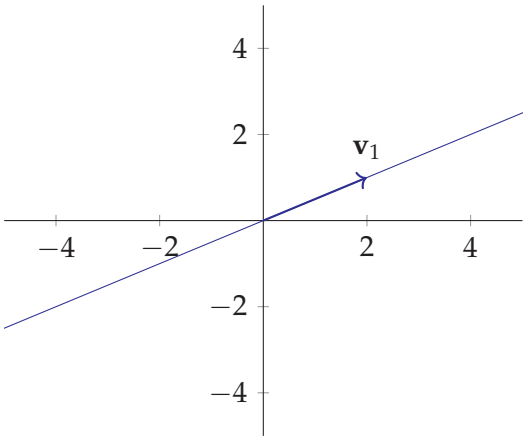


Figure 9.1: The span of one non-zero vector in $\mathbb{R}^2$.

Example 9.3.6

Consider the real vector space $V = \mathbb{R}^2$ and let

$$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Let us determine the span of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, or in other words $\text{Span}(S)$ with $S = \{\mathbf{v}_1, \mathbf{v}_2\}$. Using Definition 9.3.2 we find:

$$\begin{aligned} \text{Span}(\{\mathbf{v}_1, \mathbf{v}_2\}) &= \{c_1 \cdot \mathbf{v}_1 + c_2 \cdot \mathbf{v}_2 \,|\, c_1, c_2 \in \mathbb{R}\} \\ &= \left\{ c_1 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} \,|\, c_1, c_2 \in \mathbb{R} \right\} \\ &= \left\{ \begin{bmatrix} 2c_1 + c_2 \\ c_1 + 3c_2 \end{bmatrix} \,|\, c_1, c_2 \in \mathbb{R} \right\}. \end{aligned}$$

Graphically the situation is that the span of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ consists of the entire space $\mathbb{R}^2$. In Figure 9.1 we have indicated the area one obtains when plotting all vectors of the form $c_1 \cdot \mathbf{v}_1 + c_2 \cdot \mathbf{v}_2$ if c_1 and c_2 are chosen freely in the interval $[0, 1]$.

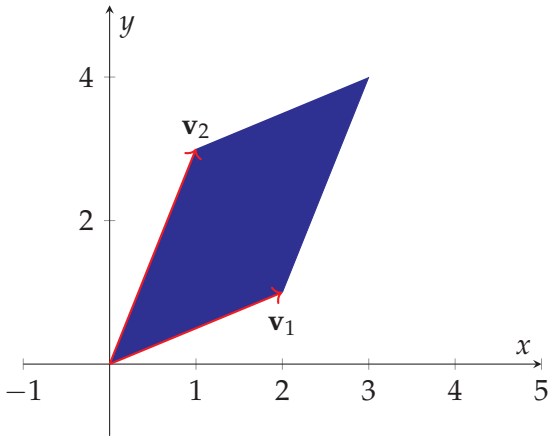


Figure 9.2: Area obtained by the linear combinations $c_1 \cdot \mathbf{v}_1 + c_2 \cdot \mathbf{v}_2$ where $c_1, c_2 \in [0, 1]$.

Example 9.3.7

Consider the real vector space $V = \mathbb{R}^3$ and let

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \text{ and } \mathbf{v}_3 = \begin{bmatrix} 0 \\ 3 \\ 6 \end{bmatrix}.$$

Question: Find a basis of the subspace W spanned by the three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

Answer:

A first, but unfortunately wrong, guess could be that the three vectors $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$ themselves form a basis. Certainly any vector in W can be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$. This is a direct consequence of the Definition 9.3.2 of the span. However, in order to be a basis, the three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ would have to be linearly independent as well. It turns out they are not. Using Theorem 7.1.3 this can be determined by calculating the reduced row echelon form of the 3×3 matrix $\mathbf{A}$ with columns $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$. We omit the details of this calculation, but instead encourage the reader to verify that this reduced row echelon form is:

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

This shows that the three vectors $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$ are linearly dependent, but at the same time that the first two of them are linearly independent (compare to Example 7.1.4, where a similar approach was used for three vectors in $\mathbb{C}^3$). We can conclude that $\mathbf{v}_3$ can be expressed as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. This in turns implies that the two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ span exactly the same subspace of $\mathbb{R}^3$ as the three vectors $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$. Hence $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis of W.

We have already fully answered the question, but suppose that we would like to see explicitly how to express $\mathbf{v}_3$ as a linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$. To do this, we need to find

a solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$ of the form $(c_1, c_2, 1)$. Looking at the reduced row echelon form of $\mathbf{A}$, we see that $(-4, 1, 1)$ is such a solution. Hence $(-4) \cdot \mathbf{v}_1 + 1 \cdot \mathbf{v}_2 + \mathbf{v}_3 = \mathbf{0}$, which implies that $\mathbf{v}_3 = 4 \cdot \mathbf{v}_1 - \mathbf{v}_2$.

Geometricallyy, what is going on in this example is that the three vectors $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$ all three lie in the same plane (namely the two-dimensional subspace W we found in the example). More generally, one can show that whenever three vectors in the real vectorspace $\mathbb{R}^n$ lie in a two-dimensional subspace of $\mathbb{R}^n$, then they are linearly dependent.

As we saw in the previous example, saying that a subspace is spanned by certain vectors, does not mean that these vectors are linearly independent. The procedure we used in Example 9.3.7 to find a basis can be generalized. Let us do that in the following theorem:

Theorem 9.3.3

Let a subspace W of the vector space $\mathbb{F}^n$ be spanned by vectors $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$. Further suppose that the reduced row echelon form of the matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ has pivots precisely in columns $j_1, \ldots, j_\rho$. Then $\{\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}\}$ is a basis of W.

Proof. First of all, let us denote by $\mathbf{A}$ the matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ and by $\mathbf{B}$ the reduced row echelon form of $\mathbf{A}$. By definition of the reduced row echelon form of a matrix, the columns of $\mathbf{B}$ with column indices $i_1, \ldots, i_\rho$ are the first ρ standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_\rho$. In particular, they are linearly independent. We claim that this implies that the columns of $\mathbf{B}$ with column indices $i_1, \ldots, i_\rho$ are also linearly independent. Indeed, if $c_{j_1} \cdot \mathbf{u}_{j_1} + \cdots + c_{j_\rho} \cdot \mathbf{u}_{j_\rho} = \mathbf{0}$, then the tuple $(v_1, \ldots, v_\ell) \in \mathbb{F}^\ell$ defined by $v_j = c_j$ if $j \in \{j_1, \ldots, j_\rho\}$ and $v_j = 0$ otherwise, is a solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$. However, we know that any such solution is also a solution to the homogeneous system with coefficient matrix $\mathbf{B}$. Since we already observed that the columns of $\mathbf{B}$ with column indices $i_1, \ldots, i_\rho$ are linearly independent, we conclude that necessarily $c_{j_1} = 0, \ldots, c_{j_\rho} = 0$. This shows that the vectors $\{\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}\}$ are linearly independent.

Now choose any column $\mathbf{u}_j$ of $\mathbf{A}$, where $j \notin \{j_1, \ldots, j_\rho\}$. Again by definition of the reduced row echelon form, the j-th column of $\mathbf{B}$ has zeroes for its last $n - \rho$ entries. Hence it can be expressed as a linear combination of $\mathbf{e}_1, \ldots, \mathbf{e}_\rho$, which are just columns $j_1, \ldots, j_\ell$ of $\mathbf{B}$. This means that the homogeneous system with coefficient matrix $\mathbf{B}$ has a solution $(v_1, \ldots, v_\ell)$ such that $v_j = 1$ and $v_k = 0$ for all $k \notin \{j, j_1, \ldots, j_\rho\}$. Now using that this is also a solution to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$, we find that the j-th column of $\mathbf{A}$ can be expressed as a linear combination of columns $j_1, \ldots, j_\rho$. This proves that the span of $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ is the same as the span of $\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}$.

Combining all the above, we conclude that $\{\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}\}$ is a basis of W. □

Looking back at Theorem 6.4.4, we see that in that theorem the solution set to a homogeneous system of linear equations was described exactly as the span of $n - \rho$ vectors. In this case

these vectors actually form a basis of the solution set and in particular they are linearly independent. Let us show this now.

Corollary 9.3.4

Let a homogeneous system of m linear equation in n variables over a field $\mathbb{F}$ be given. Denote the coefficient matrix of this system by $\mathbf{A}$ and assume that this matrix has rank ρ. The $n-\rho$ vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n-\rho}$ indicated in Theorem 6.4.4 form a basis of the solution set to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$.

Proof. Proof sketch: we use that same notation for the vectors $\mathbf{c}_i$ and the matrix $\hat{\mathbf{A}}$ as in Theorem 6.4.4. Looking back at the way the vector $\mathbf{v}_i$ was defined in Theorem 6.4.4, one can see that $\mathbf{v}_i$ has a 1 in the coordinate j, where j satisfies that $\mathbf{c}_i$ is the j-th column in $\hat{\mathbf{A}}$. Similarly, one sees that $\mathbf{v}_i$ has coefficients equal to 0 afterwards, since $\mathbf{c}_i$ contains zeroes only after its ith coefficient. Hence the matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_{n-\rho}$ is in row echelon form. This implies that the corresponding matrix in reduced row echelon form has pivots in each column. Theorem 9.3.3 then implies that $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n-\rho}\}$ is a basis. □

Corollary 9.3.5

Let V be a vector space over a field $\mathbb{F}$ of finite dimension n with ordered basis β and let $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ be vectors in V. Further suppose that the reduced row echelon form of the matrix with columns $[\mathbf{u}_1]_\beta, \ldots, [\mathbf{u}_\ell]_\beta$ has pivots precisely in columns $j_1, \ldots, j_\rho$. Then a basis of $\mathrm{Span}(\mathbf{u}_1, \ldots, \mathbf{u}_\ell)$ is given by $\{\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}\}$.

Proof. We give a sketch of the proof: first of all, we see from Theorem 9.3.3 that a basis of the subspace of $\mathbb{F}^n$ generated by $[\mathbf{u}_1]_\beta, \ldots, [\mathbf{u}_\ell]_\beta$ is given by $\{[\mathbf{u}_{j_1}]_\beta, \ldots, [\mathbf{u}_{j_\rho}]_\beta\}$. Now Theorem 9.2.3 can be used to see that $\{\mathbf{u}_{j_1}, \ldots, \mathbf{u}_{j_\rho}\}$ is a basis of $\mathrm{Span}(\mathbf{u}_1, \ldots, \mathbf{u}_\ell)$. □

9.4 Extra: why does any vector space have a basis?

This section is not required reading and can be skipped. It is meant as extra material for a student who has the time and motivation for it.

In the previous sections, we have simply used the fact that any vector space V has a basis. To prove this, we need to study the set $\mathcal{I}(V)$ consisting of all subsets of V whose elements are linearly independent vectors. For example $\emptyset \in \mathcal{I}(V)$, since the empty set contains no vectors and therefore cannot contain linearly dependent vectors. If $V \neq \{\mathbf{0}\}$ any subset of the form $\{\mathbf{v}\}$ is in $\mathcal{I}(V)$ as long as $\mathbf{v} \neq \mathbf{0}$. Intuitively, a basis B of V should be a set containing as many linearly independent vectors as possible. More precisely, this intuition would say that $B \in \mathcal{I}(V)$ and that no set of linearly independent vectors can contain B as a strict subset. This

second intuitive property can be reformulated by saying that if $C \in \mathcal{I}(V)$ and $B \subseteq C$, then $B = C$. Such a B is called a maximal element of $\mathcal{I}(V)$.

The above discussion is purely to get an intuitive idea, but the following theorem shows that there is merit in that discussion.

Theorem 9.4.1

Let B be a maximal element of $\mathcal{I}(V)$. Then B is a basis of V.

Proof. By definition of $\mathcal{I}(V)$, the vectors in B are linearly independent. What needs to be shown is that any vector in V can be written as a linear combination of vectors in B. Suppose that this is not the case. Then there exists $\mathbf{v} \in V$ such that any linear combination of vectors in B is distinct from $\mathbf{v}$. We claim that in this case, the set $B \cup \{\mathbf{v}\}$ consists of linearly independent vectors. To show this, suppose that

$$c_0 \cdot \mathbf{v} + c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n = \mathbf{0}, \tag{9.5}$$

for some $c_0, c_1, \ldots, c_n \in \mathbb{F}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n \in B$. If $c_0 = 0$, we immediately see that $c_1 = 0, \ldots, c_n = 0$, since the vectors in B are linearly independent. However, c_0 cannot be nonzero, since if it were, Equation (9.5) would imply that $\mathbf{v} = -c_0^{-1} \cdot c_1 \cdot \mathbf{v}_1 - \cdots - c_0^{-1} \cdot c_n \cdot \mathbf{v}_n$, contrary to the assumption that $\mathbf{v}$ cannot be written as a linear combination of vectors from B. Hence indeed, the set $B \cup \{\mathbf{v}\}$ consists of linearly independent vectors, just as claimed. Another way of saying this is that $B \cup \{\mathbf{v}\} \in \mathcal{I}(V)$, which in turn implies that B was not a maximal element of $\mathcal{I}(V)$, contrary to the assumption that it was. The contradiction shows that any vector in V can be written as a linear combination of vectors from B. Hence B is a basis. □

This theorem implies that in order to show that any vector space V has a basis, it is enough to show that the set $\mathcal{I}(V)$ always contains a maximal element. This is a direct consequence of a famous lemma called Zorn's lemma. Formulating and proving Zorn's lemma needs tools from foundational mathematics though that are out of scope of this text.

Chapter 10

Linear maps between vector spaces

Given two vector spaces V_1 and V_2, both over the same field $\mathbb{F}$, a linear map is a function from V_1 to V_2 that is compatible with scalar multiplication and vector addition. More precisely, we have the following:

Definition 10.0.1

Let V_1 and V_2 be vector spaces over a field $\mathbb{F}$. Then a *linear map* from V_1 to V_2 is a function $L : V_1 \to V_2$ such that:

(i) $L(\mathbf{u} + \mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in V_1$,

(ii) $L(c \cdot \mathbf{u}) = c \cdot L(\mathbf{u})$ for all $c \in \mathbb{F}$ and $\mathbf{u} \in V_1$.

A linear map is also called a *linear transformation*. Note that in the formula $L(\mathbf{u} + \mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$, the $+$ in $\mathbf{u} + \mathbf{v}$ denotes vector addition in V_1, while the $+$ in $L(\mathbf{u}) + L(\mathbf{v})$ denotes vector addition in V_2. Similarly, in the formula $L(c \cdot \mathbf{u}) = c \cdot L(\mathbf{u})$, the $\cdot$ in $c \cdot \mathbf{u}$ denotes the scalar multiplication in V_1, while in $c \cdot L(\mathbf{u})$, it denotes the scalar multiplication in V_2.

While in the previous chapter, we studied one vector space at the time, linear maps can connect different vector spaces with each other. Linear maps respect the vector space structure: choosing the scalar c equal to 0 and using Equation (9.1), one obtains for example

$$L(\mathbf{0}) = \mathbf{0}, \tag{10.1}$$

where the $\mathbf{0}$ on the left-hand side of the equation denotes the zero vector in V_1 and the one on the right denotes the zero vector in V_2. Similarly, choosing $c = -1$ and using Equation (9.2), one obtains that

$$L(-\mathbf{u}) = -L(\mathbf{u}). \tag{10.2}$$

Of course, there are many possible functions between two vector spaces and in general not many will be linear. Let us consider some examples.

Example 10.0.1

Consider the following function from $\mathbb{R}$ to $\mathbb{R}$. Which ones are linear maps?

(a) $f : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto x^2$,

(b) $g : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto 2x + 1$,

(c) $h : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto 2x$.

Answer:

(a) $f : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto x^2$. This is not a linear map. We have for example $f(1+1) = f(2) = 4$, but if f would have been a linear map, we should have had $f(1+1) = f(1) + f(1) = 1 + 1 = 2$.

(b) $g : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto 2x + 1$. This is not a linear map either, even though the graph of this function is a line. We have $g(0) = 1$, but if g would have been a linear map, we should have had $g(0) = 0$ by Equation (10.1).

(c) $h : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto 2x$. This is a linear map. For all $x, y \in \mathbb{R}$ we have $h(x+y) = 2(x+y) = 2x + 2y = h(x) + h(y)$ and for all $c \in \mathbb{R}$ and $x \in \mathbb{R}$, we have $c \cdot h(x) = c2x = 2cx = h(c \cdot x)$.

More general, linear maps from $\mathbb{R}$ to $\mathbb{R}$ are precisely those functions whose graph is a straight line passing through the origin. In other words, they are functions $L : \mathbb{R} \to \mathbb{R}$ such that $x \mapsto a \cdot x$ for some constant $a \in \mathbb{R}$. The reason is that if $L : \mathbb{R} \to \mathbb{R}$ is a linear map, then for all $x \in \mathbb{R}$, we have $L(x) = L(x \cdot 1) = x \cdot L(1)$. In the last equality, we used property 2 from Definition 10.0.1. Setting $a = L(1)$, we indeed obtain that $L(x) = a \cdot x$ for all $x \in \mathbb{R}$.

We will see that there is a strong connection between linear maps and matrices. For this reason, we start with studying linear maps coming from matrices and afterwards return to studying linear maps in a more general setting.

10.1 Linear maps using matrices

Let us start by defining a large class of linear maps.

Definition 10.1.1

Let $\mathbb{F}$ be a field. Given a matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$, define the function $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$ by defining $L_{\mathbf{A}}(\mathbf{v}) = \mathbf{A} \cdot \mathbf{v}$ for all $\mathbf{v} \in \mathbb{F}^n$.

It turns out that all functions $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$ defined above are linear.

Lemma 10.1.1
The function $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$ in Definition 10.1.1 is a linear map.

Proof. We need to check the two conditions from Definition 10.0.1. First of all

$$\begin{aligned} L_{\mathbf{A}}(\mathbf{u}+\mathbf{v}) &= \mathbf{A}\cdot(\mathbf{u}+\mathbf{v}) \\ &= \mathbf{A}\cdot\mathbf{u}+\mathbf{A}\cdot\mathbf{v} \\ &= L_{\mathbf{A}}(\mathbf{u})+L_{\mathbf{A}}(\mathbf{v}), \text{ for all } \mathbf{u},\mathbf{v}\in\mathbb{F}^n. \end{aligned}$$

Secondly:

$$L_{\mathbf{A}}(c\cdot\mathbf{u}) = \mathbf{A}\cdot(c\cdot\mathbf{u}) = c\cdot(\mathbf{A}\cdot\mathbf{u}) = c\cdot L_{\mathbf{A}}(\mathbf{u}) \text{ for all } c\in\mathbb{F} \text{ and } \mathbf{u}\in\mathbb{F}^n.$$

□

In Example 10.0.1, we saw that the function $h : \mathbb{R} \to \mathbb{R}$, $x \mapsto 2x$ was a linear map. It is actually a very special case of Definition 10.1.1: if we choose $n = m = 1$, $\mathbb{F} = \mathbb{R}$ and $\mathbf{A} = [2]$ in Definition 10.1.1, we find the function h. Instead of $\mathbf{A} = [2]$, we could also just have written $\mathbf{A} = 2$. Indeed, when writing down a 1×1 matrix, it is quite common to leave the brackets [ ] out.

Example 10.1.1
Let $\mathbb{F} = \mathbb{R}$ and choose

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Further define

$$\mathbf{u} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Question: With $\mathbf{A}$, $\mathbf{u}$ and $\mathbf{v}$ given as above. Compute $L_{\mathbf{A}}(\mathbf{u})$, $L_{\mathbf{A}}(\mathbf{v})$ and $L_{\mathbf{A}}((1/2)\mathbf{u}+\mathbf{v})$.

Answer: We have

$$L_{\mathbf{A}}(\mathbf{u}) = \mathbf{A}\cdot\mathbf{u} = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix}\cdot\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} (-1)\cdot 2+0\cdot 1 \\ 0\cdot 2+(1/2)\cdot 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 1/2 \end{bmatrix}$$

and similarly

$$L_{\mathbf{A}}(\mathbf{u}) = \mathbf{A}\cdot\mathbf{v} = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix}\cdot\begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} (-1)\cdot 1+0\cdot 3 \\ 0\cdot 1+(1/2)\cdot 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3/2 \end{bmatrix}.$$

We can in principle compute $L_{\mathbf{A}}((1/2)\mathbf{u}+\mathbf{v})$ in various ways. The most direct method is to calculate the vector $(1/2)\mathbf{u}+\mathbf{v}$ first , then afterwards calculate $\mathbf{A}\cdot((1/2)\mathbf{u}+\mathbf{v})$. Doing this,

we would find

$$(1/2)\mathbf{u}+\mathbf{v} = (1/2)\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7/2 \end{bmatrix}$$

and hence

$$L_{\mathbf{A}}((1/2)\mathbf{u}+\mathbf{v}) = L_{\mathbf{A}}\left(\begin{bmatrix} 2 \\ 7/2 \end{bmatrix}\right) = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 7/2 \end{bmatrix} = \begin{bmatrix} -2 \\ 7/4 \end{bmatrix}.$$

There is also another way to compute $L_{\mathbf{A}}((1/2)\mathbf{u}+\mathbf{v})$ using that $L_{\mathbf{A}}$ is a linear map and that we already had calculated $L_{\mathbf{A}}(\mathbf{u})$ and $L_{\mathbf{A}}(\mathbf{v})$:

$$\begin{aligned} L_{\mathbf{A}}((1/2)\mathbf{u}+\mathbf{v}) &= L_{\mathbf{A}}((1/2)\mathbf{u}) + L_{\mathbf{A}}(\mathbf{v}) = (1/2)L_{\mathbf{A}}(\mathbf{u}) + L_{\mathbf{A}}(\mathbf{v}) \\ &= (1/2)\begin{bmatrix} -2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} -1 \\ 3/2 \end{bmatrix} = \begin{bmatrix} -2 \\ 7/4 \end{bmatrix}. \end{aligned}$$

See Figure 10.1 for an illustration of the image of the vectors $\mathbf{u}$ and $\mathbf{v}$ as well as the image of vectors lying in a parallelogram two of whose sides are the given vectors $\mathbf{u}$ and $\mathbf{v}$.

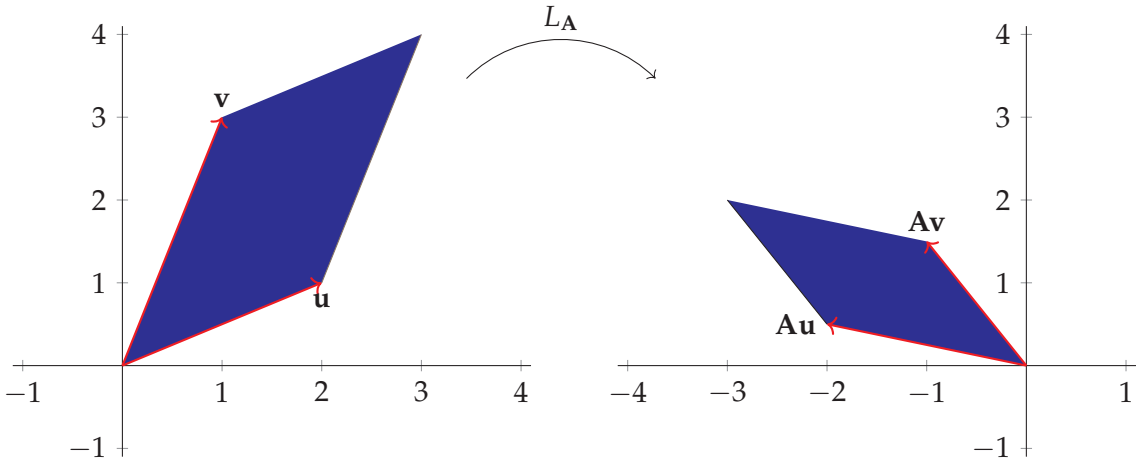


Figure 10.1: Example of a linear map given by a 2×2 matrix.

Example 10.1.2

As in the previous example let $\mathbb{F} = \mathbb{R}$ and define

$$\mathbf{u} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

This time, we choose the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

In this case we have

$$L_{\mathbf{A}}(\mathbf{u}) = \mathbf{A} \cdot \mathbf{v} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and

$$L_{\mathbf{A}}(\mathbf{u}) = \mathbf{A} \cdot \mathbf{v} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}.$$

An illustration is given in Figure 10.2. This time, the image of the blue area (the parallelogram) is just the line connecting $(-1, 1)$ and $(2, -2)$. It is therefore not visible in the figure, since it is hidden behind the drawing of the vectors $\mathbf{Au}$ and $\mathbf{Av}$.

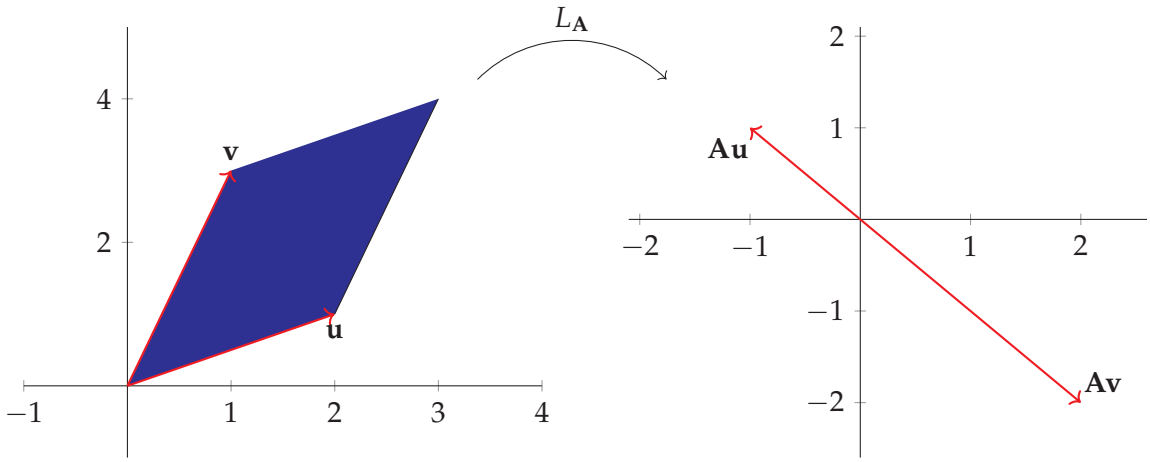


Figure 10.2: Another example of a linear map given by a 2×2 matrix.

Example 10.1.3

Let $\mathbb{F} = \mathbb{R}$ and choose

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 4}.$$

Then the corresponding linear map $L_{\mathbf{A}} : \mathbb{R}^4 \to \mathbb{R}^2$ works as follows:

$$L_{\mathbf{A}}\left(\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}\right) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} v_1 + v_2 + v_3 + v_4 \\ v_1 + 2v_2 + 3v_3 + 4v_4 \end{bmatrix}.$$

So for example

$$L_{\mathbf{A}}\left(\begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad L_{\mathbf{A}}\left(\begin{bmatrix} 0 \\ 2 \\ -1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \quad \text{and} \quad L_{\mathbf{A}}\left(\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

As Example 10.1.3, it is possible that a linear map $L_{\mathbf{A}}$ maps a vector to the zero vector. The set of such vectors has a special name:

Definition 10.1.2

Let $\mathbb{F}$ be a field. Given a matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$, the *kernel* of the matrix $\mathbf{A}$, denoted by $\ker \mathbf{A}$, is the following set of vectors:

$$\ker \mathbf{A} = \{\mathbf{v} \in \mathbb{F}^n \mid \mathbf{A} \cdot \mathbf{v} = \mathbf{0}\}.$$

Note that one can equivalently define the kernel of a matrix $\mathbf{A}$ to be all vectors from $\mathbb{F}^n$ that are mapped to the zero vector by the linear map $L_{\mathbf{A}}$. We can also think of the vectors in the kernel as precisely those vectors that are solutions to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$.

Remark 10.1.1

A remark about terminology is in place here. Some authors prefer to use the words *null space*, *right kernel* or *right null space* for what we have called the kernel of a matrix. The reason for adding the word "right" is that we have multiplied the matrix with a column vector from the right. One could also have considered the set of row vectors $\mathbf{u} \in \mathbb{F}^{1\times m}$ such that $\mathbf{u} \cdot \mathbf{A} = \mathbf{0}$. This set is called the *left kernel* of $\mathbf{A}$ or sometimes also the *left null space*.

One of the reasons that we introduced the notion of kernel of a matrix, is that it actually is a subspace. Let us show this in the following lemma.

Lemma 10.1.2

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ a matrix. Then the kernel of $\mathbf{A}$ is a subspace of $\mathbb{F}^n$.

Proof. First of all, note that $\mathbf{0} \in \ker \mathbf{A}$ so that $\ker \mathbf{A}$ is not the empty set. This means that if we set $W = \ker \mathbf{A}$, then the requirement that W is not empty in Lemma 9.3.2 is met.

Let $\mathbf{u}, \mathbf{v} \in \ker \mathbf{A}$ and $c \in \mathbb{F}$. Then

$$\mathbf{A} \cdot (\mathbf{u} + c \cdot \mathbf{v}) = \mathbf{A} \cdot \mathbf{u} + \mathbf{A} \cdot (c \cdot \mathbf{v}) = \mathbf{A} \cdot \mathbf{u} + c \cdot (\mathbf{A} \cdot \mathbf{v}) = \mathbf{0} + c \cdot \mathbf{0} = \mathbf{0}. \tag{10.3}$$

Here we used that $\mathbf{A} \cdot \mathbf{u} = \mathbf{0}$ and $\mathbf{A} \cdot \mathbf{v} = \mathbf{0}$, since $\mathbf{u}, \mathbf{v} \in \ker \mathbf{A}$. Equation (10.3) implies that $\mathbf{u} + c \cdot \mathbf{v} \in \ker \mathbf{A}$. Then Lemma 9.3.2 implies that $\ker \mathbf{A}$ is a subspace of $\mathbb{F}^n$. □

The dimension of $\ker \mathbf{A}$ is called the *nullity* of the matrix $\mathbf{A}$. It is denoted by $\text{null}\mathbf{A}$.

Example 10.1.4

Let $\mathbb{F} = \mathbb{R}$ and consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

Compute a basis for $\ker \mathbf{A}$ and compute the nullity of $\mathbf{A}$.

Answer: The kernel of $\mathbf{A}$ consists of all vectors $\mathbf{v} = (v_1, v_2, v_3, v_4) \in \mathbb{R}^4$ such that $\mathbf{A} \cdot \mathbf{v} = \mathbf{0}$. We have for example seen in Example 10.1.3 that the vector $(1, -1, -1, 1)$ is mapped to $(0, 0)$ by the linear map $L_{\mathbf{A}}$. Therefore $(1, -1, -1, 1) \in \ker \mathbf{A}$.

We can think of the vectors in the kernel as precisely those vectors that are solutions to the homogeneous system of two linear equations with coefficient matrix $\mathbf{A}$. To describe all these solutions, we follow the same procedure as explained in Example 6.4.3 and Theorem 6.4.4. Hence, we first bring the matrix $\mathbf{A}$ in reduced row echelon form:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \xrightarrow[R_2 \leftarrow R_2 - R_1]{} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \xrightarrow[R_1 \leftarrow R_1 - R_2]{} \begin{bmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \end{bmatrix}.$$

Now we can see that $\mathbf{v} = (v_1, v_2, v_3, v_4) \in \ker \mathbf{A}$ if and only if $v_1 - v_3 - 2v_4 = 0$ and $v_2 + 2v_3 + 3v_4 = 0$. Similarly as in Example 6.4.3 (or directly using Theorem 6.4.4), we see that

$$\ker \mathbf{A} = \left\{ t_1 \cdot \begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} 2 \\ -3 \\ 0 \\ 1 \end{bmatrix} \mid t_1, t_2 \in \mathbb{R} \right\}.$$

Hence the vectors

$$\begin{bmatrix} 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 \\ -3 \\ 0 \\ 1 \end{bmatrix}$$

span $\ker \mathbf{A}$. In fact, Corollary 9.3.4 tells us that these two vectors form a basis of $\ker \mathbf{A}$. Hence a basis for $\ker \mathbf{A}$ is given by

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

The nullity of the matrix $\mathbf{A}$ is by definition the dimension of the subspace $\ker \mathbf{A}$. Since we have just computed a basis of $\ker \mathbf{A}$ and this basis consists of two vectors, we conclude that the nullity of $\mathbf{A}$ is two. In other words: $\text{null}\mathbf{A} = 2$.

We have already observed that we can think of the vectors in the kernel as precisely those vectors that are solutions to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$. Using Corollary 9.3.4, we obtain the following result, which often is called the *rank-nullity theorem for matrices*.

Theorem 10.1.3

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ a matrix. Then

$$\rho(\mathbf{A}) + \text{null}(\mathbf{A}) = n,$$

where $\rho(\mathbf{A})$ denotes the rank of the matrix $\mathbf{A}$ and $\text{null}(\mathbf{A})$ its nullity.

Proof. Using Corollary 9.3.4, we see that the kernel of $\mathbf{A}$ has a basis containing precisely $n - \rho(\mathbf{A})$ many vectors. Hence $\text{null}(\mathbf{A}) = \dim\ker(\mathbf{A}) = n - \rho(\mathbf{A})$. This implies that $\rho(\mathbf{A}) + \text{null}(\mathbf{A}) = n$. □

We have seen in Lemma 10.1.2, that the kernel of a matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$ is a linear subspace of $\mathbb{F}^n$. In other words: $\ker \mathbf{A}$ is a linear subspace of the domain of the linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$. To a matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$ one can also associate a linear subspace of $\mathbb{F}^m$, the codomain of the linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$. We do this in the following definition.

Definition 10.1.3

Let $\mathbb{F}$ be a field. Given a matrix $\mathbf{A} \in \mathbb{F}^{m\times n}$, the *column space* of the matrix $\mathbf{A}$, denoted by $\text{colsp}\mathbf{A}$, is the subspace of $\mathbb{F}^m$ spanned by the columns of $\mathbf{A}$. The dimension of the column space of a matrix $\mathbf{A}$ is called the *column rank* of $\mathbf{A}$.

Lemma 10.1.4

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ a matrix. Then $\text{colsp}\mathbf{A}$, the column space of the matrix $\mathbf{A}$ is precisely the image of the linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$.

Proof. An element from the column space of a matrix $\mathbf{A}$ is by definition a linear combination of the columns of $\mathbf{A}$. On the other hand, an element of the image of the linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$ is of the form $\mathbf{A} \cdot \mathbf{v}$ for some vector $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}^n$. Using Definition 7.2.1, we can rewrite this as the linear combination of the columns of $\mathbf{A}$ as follows:

$$\mathbf{A} \cdot \mathbf{v} = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \ldots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = v_1 \cdot \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \cdots v_n \cdot \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix}.$$

Hence the image of the linear map $L_{\mathbf{A}}$ consists precisely of all linear combinations of the columns of $\mathbf{A}$. But this is precisely the column space of the matrix $\mathbf{A}$. □

Remark 10.1.2

Because of Lemma 10.1.4, the column space of a matrix $\mathbf{A}$ is sometimes also called the *range* or the *image* of $\mathbf{A}$.

We have previously introduced the rank of a matrix in Definition 6.3.2. The rank $\rho(\mathbf{A})$ of a matrix $\mathbf{A}$ as defined in Definition 6.3.2 is sometimes more properly called the *row rank* of the matrix $\mathbf{A}$, since one can show that the dimension of the vector space spanned by the rows of $\mathbf{A}$ is equal to $\rho(\mathbf{A})$. It turns out however, that for any matrix, its row rank and column rank are the same. Therefore, we will from now on simply call the column rank of a matrix $\mathbf{A}$, the rank of the matrix and denote it by $\rho(\mathbf{A})$, using the same notation as in Definition 6.3.2.

It is not obvious from Definitions 6.3.2 and 10.1.3 that row rank and column rank of a matrix are always the same. A reader willing to accept this can skip the remainder of this section, but for the interested reader, we give a short proof of why row rank and column rank are always the same.

Theorem 10.1.5

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{m\times n}$ a matrix. Then the row rank and the column rank of the matrix $\mathbf{A}$ are the same.

Proof. The row rank $\rho(\mathbf{A})$ of a matrix $\mathbf{A}$ is by definition equal to the number of pivots in the reduced row echelon form of $\mathbf{A}$. On the other hand, Theorem 9.3.3 implies that the number of vectors in a basis of the column space of $\mathbf{A}$ is also equal to this number of pivots. Hence the dimension of the column space of $\mathbf{A}$ is also equal to $\rho(\mathbf{A})$. □

Example 10.1.5

In this example we want to compute the rank and nullity of the matrix $\mathbf{A}$ from Example 10.1.1. In other words, we have

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1/2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

To compute the kernel of $\mathbf{A}$, one can in principle follow the same procedure as in Example 10.1.4, but since we are working with a square matrix in this example, we choose a slightly different approach using determinants.

First of all, $\det \mathbf{A} = (-1)\cdot 1/2 = -1/2$. In particular, $\det \mathbf{A} \neq 0$. Hence Corollary 8.3.6 implies that the kernel of the matrix $\mathbf{A}$ only contains the zero vector, that is $\ker \mathbf{A} = \{\mathbf{0}\}$. Hence in this example, $\text{nulll}(\mathbf{A}) = \dim\{\mathbf{0}\} = 0$.

The rank-nullity theorem (Theorem 10.1.3) then implies that the matrix $\mathbf{A}$ has rank $\rho(\mathbf{A}) = 2 - 0 = 2$. Hence $\rho(\mathbf{A}) = 2$. We could also have computed the rank using Corollary 8.3.5: since $\det \mathbf{A} \neq 0$, this corollary implies that the two columns of $\mathbf{A}$ are linearly independent (one could verify this directly as well). Hence $\mathbf{A}$ has column rank two, but then Theorem 10.1.5 implies that $\rho(\mathbf{A}) = 2$.

Example 10.1.6

In this example we want to compute the rank and nullity of the matrix $\mathbf{A}$ from Example 10.1.2.

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Note that $\det \mathbf{A} = 0$ this time. In particular the columns of $\mathbf{A}$ are not linearly independent by Corollary 8.3.5. Indeed, the two columns of $\mathbf{A}$ add up to the zero vector, confirming that they are are linearly dependent. We conclude that

$$\text{colsp}\mathbf{A} = \text{Span}\left\{\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right\} = \text{Span}\left\{\begin{bmatrix} -1 \\ 1 \end{bmatrix}\right\}.$$

Since the vector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ is not the zero vector, we conclude that the dimension of colsp$\mathbf{A}$ is one. In particular $\rho(\mathbf{A}) = 1$ (as in the previous example we use Theorem 10.1.5). Then the rank-nullity theorem implies that $\text{null}(\mathbf{A}) = 2 - 1 = 1$.

Note that Figure 10.2 intuitively suggests that the image of the linear map $L_{\mathbf{A}} : \mathbb{R}^2 \to \mathbb{R}^2$ is one-dimensional. By Lemma 10.1.4 the image of $L_{\mathbf{A}}$ is the same as the column space of $\mathbf{A}$, so that the intuition fits.

10.2 Linear maps between general vector spaces

In the previous section, we have focused on linear maps coming from matrices, but Definition 10.0.1 allows for much more general linear maps. It turns out that the notions of kernel and image also make sense in the general setting. Let us first consider some more examples.

Example 10.2.1

Let $\mathbb{F} = \mathbb{C}$ and consider the complex vector space $\mathbb{C}[Z]$ (see Example 9.1.5). Recall that $\mathbb{C}[Z]$ denotes the set of all polynomials with coefficients in $\mathbb{C}$. Now consider the map $D : \mathbb{C}[Z] \to \mathbb{C}[Z]$ defined by $D(a_0 + a_1 Z + a_2 Z^2 + \cdots + a_n Z^n) = a_1 + 2a_2 Z + \cdots + n a_n Z^{n-1}$. In words, the map D sends a polynomial $p(Z)$ to its derivative $p(Z)'$. One can show that D is a linear map.

Example 10.2.2

Let $V_1 = \mathbb{F}^{n\times n}$ and $V_2 = \mathbb{F}$. Given a square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$, the *trace*, denoted by $\text{Tr}(\mathbf{A})$, is defined as the sum of the elements on its diagonal. In other words:

$$\text{Tr}\left(\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}\right) = a_{11} + \cdots + a_{nn}.$$

Question: Is the map $\text{Tr} : \mathbb{F}^{n\times n} \to \mathbb{F}$, defined by $\mathbf{A} \mapsto \text{Tr}(\mathbf{A})$ a linear map?

Answer: To find out whether or not the trace map $\text{Tr} : \mathbb{F}^{n\times n} \to \mathbb{F}$ as defined above, is linear,

we check if all conditions in Definition 10.0.1 are satisfied. First of all, using the notation from Definition 10.0.1, we have $V_1 = \mathbb{F}^{n\times n}$ and $V_2 = \mathbb{F}$. We should first check that these are vector spaces over a field $\mathbb{F}$. Both are indeed vector spaces over $\mathbb{F}$: For V_1, see Example 9.1.4 with $m = n$ and for V_2, see Example 9.1.1 with $n = 1$.

Now we need to check if Tr satisfies the two conditions from Definition 10.0.1. Let us choose arbitrary $c \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{F}^{n\times n}$. Hence we can write

$$\mathbf{u} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix}.$$

Then

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & \dots & a_{1n}+b_{1n} \\ \vdots & & \vdots \\ a_{n1}+b_{n1} & \dots & a_{nn}+b_{nn} \end{bmatrix}$$

and

$$c \cdot \mathbf{u} = c \cdot \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} c\cdot a_{11} & \dots & c\cdot a_{1n} \\ \vdots & & \vdots \\ c\cdot a_{n1} & \dots & c\cdot a_{nn} \end{bmatrix}.$$

Hence

$$\mathrm{Tr}(\mathbf{u}+\mathbf{v}) = a_{11}+b_{11}+\cdots a_{nn}+b_{nn} = a_{11}+\cdots a_{nn}+b_{11}+\cdots b_{nn} = \mathrm{Tr}(\mathbf{u})+\mathrm{Tr}(\mathbf{v})$$

and

$$\mathrm{Tr}(c\cdot\mathbf{u}) = c\cdot a_{11}+\cdots c\cdot a_{nn} = c\cdot(a_{11}+\cdots a_{nn}) = c\cdot\mathrm{Tr}(\mathbf{u}).$$

We can conclude that $\mathrm{Tr} : \mathbb{F}^{n\times n} \to \mathbb{F}$, defined by $\mathbf{A} \mapsto \mathrm{Tr}(\mathbf{A})$ is a linear map.

Example 10.2.3

Let $\mathbb{F} = \mathbb{R}$ and consider the map $m_5 : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $m_5(v_1, v_2) = (5v_1, 5v_2)$. In other words, the effect of map m_5 on a vector is that it multiplies a vector with the scalar 5. Visually, this means that the direction of a vector is not changed, but its length becomes five times longer. One can show that this is a linear map of real vector spaces.

More generally, one can show that if $\mathbb{F}$ is a field and $c \in \mathbb{F}$ is a scalar, then the map $m_c : \mathbb{F}^n \to \mathbb{F}^n$ defined by $m_c(\mathbf{u}) = c \cdot \mathbf{u}$ is a linear map of vector spaces.

Example 10.2.4

For $\alpha \in \mathbb{R}$, consider the map $R_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $R_\alpha(v_1, v_2) = (\cos(\alpha)\cdot v_1 - \sin(\alpha)\cdot v_2, \sin(\alpha)\cdot v_1 + \cos(\alpha)\cdot v_2)$. Geometrically, the effect of R_α on $(v_1, v_2) \in \mathbb{R}^2$ is a rotation over an angle α against the clock, where the rotation has center in $(0, 0)$. For example, if $\alpha = \pi/2$,

then $R_{\pi/2}(v_1, v_2) = (-v_2, v_1)$. One can show that R_α is a linear map.

Example 10.2.5

We choose $\mathbb{F} = \mathbb{C}$. Let V_1 be the set of polynomials in $\mathbb{C}[Z]$ of degree at most three and similarly let V_2 be the set of polynomials in $\mathbb{C}[Z]$ of degree at most four. Both V_1 and V_2 are vector spaces over $\mathbb{C}$. A possible basis for V_1 is given by the set $\{1, Z, Z^2, Z^3\}$, while a basis for V_2 is $\{1, Z, Z^2, Z^3, Z^4\}$. Hence $\dim V_1 = 4$ and $\dim V_2 = 5$. Now define the map $L : V_1 \to V_2$ by $p(Z) \mapsto (i + 2Z) \cdot p(Z)$. Note that indeed for any $p(Z) \in V_1$, we have $(i + 2Z) \cdot p(Z) \in V_2$ using Equation (4.1). One can show that L is a linear map. Indeed, if $p_1(Z), p_2(Z) \in V_1$ and $c \in \mathbb{C}$ one has

$$\begin{aligned} L(p_1(Z) + p_2(Z)) &= (i + 2Z) \cdot (p_1(Z) + p_2(Z)) \\ &= (i + 2Z) \cdot p_1(Z) + (i + 2Z) \cdot p_2(Z) \\ &= L(p_1(Z)) + L(p_2(Z)) \end{aligned}$$

and

$$L(c \cdot p_1(Z)) = (i + 2Z) \cdot c \cdot p_1(Z) = c \cdot (i + 2Z) \cdot p_1(Z) = c \cdot L(p_1(Z)).$$

Example 10.2.6

As a final example of a linear map, we consider the map $\mathrm{ev} : \mathbb{C}[Z] \to \mathbb{C}^2$ defined by $p(Z) \mapsto (p(0), p(1))$. So for example $\mathrm{ev}(Z^2 + Z + 1) = (0^2 + 0 + 1, 1^2 + 1 + 1) = (1, 3)$. One can show that ev is a linear map.

We finish this section with some general properties of linear maps. First we consider the composition of two linear maps, see Section 2.2 for the definition of the composite of two functions.

Theorem 10.2.1

Let $\mathbb{F}$ be a field and V_1, V_2, V_3 vector spaces over $\mathbb{F}$. Further, suppose that $L_1 : V_1 \to V_2$ and $L_2 : V_2 \to V_3$ are linear maps. Then the composition $L_2 \circ L_1 : V_1 \to V_3$ is also a linear map.

Proof. Let us choose arbitrary $\mathbf{u}, \mathbf{v} \in V_1$ and $c \in \mathbb{F}$. Then using linearity of L_1 and L_2 as well as the definition of the composition of two functions, we obtain that

$$\begin{aligned} (L_2 \circ L_1)(\mathbf{u} + \mathbf{v}) &= L_2(L_1(\mathbf{u} + \mathbf{v})) \\ &= L_2(L_1(\mathbf{u}) + L_1(\mathbf{v})) \\ &= L_2(L_1(\mathbf{u})) + L_2(L_1(\mathbf{v})) \\ &= (L_2 \circ L_1)(\mathbf{u}) + (L_2 \circ L_1)(\mathbf{v}) \end{aligned}$$

and

$$(L_2 \circ L_1)(c \cdot \mathbf{u}) = L_2(L_1(c \cdot \mathbf{u})) = L_2(c \cdot L_1(\mathbf{u})) = c \cdot L_2(L_1(\mathbf{u})) = c \cdot (L_2 \circ L_1)(\mathbf{u}).$$

Hence by Definition 10.0.1, the map $L_2 \circ L_1 : V_1 \to V_3$ is a linear map. □

Since any function $f : A \to B$ has an image, namely the set $\text{image}(f) = \{f(a) \mid a \in A\}$, see Section 2.2, a linear map $L : V_1 \to V_2$ has an image as well. In view of Lemma 10.1.4, this generalizes the idea of the notion of the column space of a matrix to the setting of general linear maps. One can show that the image of a linear map $L : V_1 \to V_2$ is a subspace of V_2. The notion of a kernel can also directly be generalized.

Definition 10.2.1

Let $\mathbb{F}$ be a field and V_1 and V_2 vector spaces over $\mathbb{F}$. Given a linear map $L : V_1 \to V_2$, the *kernel* of the map L is:

$$\ker L = \{\mathbf{v} \in V_1 \mid L(\mathbf{v}) = \mathbf{0}\}.$$

Similarly as in the case of the kernel of a matrix, one can show that the kernel of a linear map $L : V_1 \to V_2$ is a subspace of V_1.

Example 10.2.7

Let us revisit Example 10.2.1. We considered the linear map $D : \mathbb{C}[Z] \to \mathbb{C}[Z]$, sending a polynomial $p(Z)$ to its derivative. The only polynomials whose derivative is 0 are constant polynomials, that is to say polynomials of the form $p(Z) = a_0$. Hence $\ker D = \{a_0 \mid a_0 \in \mathbb{C}\} = \mathbb{C}$. Note that $\{1\}$ is a basis of $\ker D$, so that we can conclude that $\dim \ker D = 1$.

With a view to a later application of the theory to differential equations, we consider another example involving derivatives.

Example 10.2.8

Let $C_\infty(\mathbb{R})$ be the vector space of all infinitely differentiable functions from $\mathbb{R}$ to $\mathbb{R}$, see Example 9.3.4. Let us consider the map $L : C_\infty(\mathbb{R}) \to C_\infty(\mathbb{R})$ where $f \mapsto f' - f$. As usual f' denotes the derivative of the function f. Since $f \in C_\infty(\mathbb{R})$, also f' is infinitely often differentiable, so that $f' \in C_\infty(\mathbb{R})$. One can show that L is a linear map. Using Definition 10.2.1, we see that $\ker L = \{f \in C_\infty(\mathbb{R}) \mid f' - f = 0\}$. In other words: the kernel of L consists of those functions $f \in C_\infty(\mathbb{R})$ such that the derivative of f is the same as f itself. In yet other words: the kernel of L consists exactly of all functions in $C_\infty(\mathbb{R})$ of the differential equation $f' = f$. An example of a function satisfying this differential equation is the exponential function $\exp : \mathbb{R} \to \mathbb{R}$ defined by $t \mapsto e^t$. Also all scalar multiples $f = c \cdot \exp$ with $c \in \mathbb{R}$, are solutions to the differential equation $f' = f$. It is in fact possible to show that there are no more solutions in $C_\infty(\mathbb{R})$. Hence $\ker L$ turns out to be a one-dimensional subspace of $C_\infty(\mathbb{R})$ with basis $\{\exp\}$.

Remark 10.2.1

The exponential function was also discussed in Example 2.3.2, but there its codomain was defined to be $\mathbb{R}_{\geq 0}$. Strictly speaking, the exponential function from Example 2.3.2 is therefore not the same function as the exponential function we used in this example. However, since both functions map any $x \in \mathbb{R}$ to exactly the same value, namely e^x, it is a bit overkill to use different notations for these functions. For this reason we have denoted both functions with exp.

Example 10.2.9

As a final example of the kernel of a linear map, we consider the map ev from Example 10.2.6. The map $\text{ev} : \mathbb{C}[Z] \to \mathbb{C}^2$ was defined by $p(Z) \mapsto (p(0), p(1))$. Hence we have

$$\ker \text{ev} = \{p(Z) \in \mathbb{C}[Z] \mid (p(0), p(1)) = (0,0)\} = \{p(Z) \in \mathbb{C}[Z] \mid p(0) = 0 \ \wedge \ p(1) = 0\}.$$

It is possible to describe the kernel of ev more specifically. Let us start by describing the set of polynomials $p(Z)$ satisfying $p(0) = 0$, that is to say, such that 0 is a root of $p(Z)$. Using Lemma 4.6.2, we conclude that

$$\{p(Z) \in \mathbb{C}[Z] \mid p(0) = 0\} = \{Z \cdot q(Z) \mid q(Z) \in \mathbb{C}[Z]\}.$$

Now if both $p(0) = 0$ and $p(1) = 0$, then we see that $p(Z) = Z \cdot q(Z)$ for some $q(Z) \in \mathbb{C}[Z]$, and $p(1) = 0$. But this is equivalent with saying that $p(Z) = Z \cdot q(Z)$ for some $q(Z) \in \mathbb{C}[Z]$ and $q(1) = 0$. Using Lemma 4.6.2 again, but now for $q(Z)$ and the root 1, we see that $q(Z) = (Z-1) \cdot s(Z)$ for some $s(Z) \in \mathbb{C}[Z]$. Hence we obtain $p(Z) \in \ker \text{ev}$ if and only if $p(Z) = Z \cdot (Z-1) \cdot s(X)$ for some $s(Z) \in \mathbb{C}[Z]$. We conclude that

$$\ker \text{ev} = \{p(Z) \in \mathbb{C}[Z] \mid p(0) = 0 \ \wedge \ p(1) = 0\} = \{Z \cdot (Z-1) \cdot s(Z) \mid s(Z) \in \mathbb{C}[Z]\}.$$

10.3 Linear maps between finite dimensional vector spaces

Let us assume that we are given a finite dimensional vector space V over a field $\mathbb{F}$, say $\dim V = n$. In such a setting, we can choose an ordered basis of V, say $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent vectors in V. As we have seen in Definition 9.2.4, for each $\mathbf{v} \in V$, we can produce a unique coordinate vector $[\mathbf{v}]_\beta \in \mathbb{F}^n$. This means that we can define a function $\phi_\beta : V \to \mathbb{F}^n$ by $\mathbf{v} \mapsto [\mathbf{v}]_\beta$. Now combining Lemma 9.2.2 and Definition 10.0.1, we can immediately conclude that the function ϕ_β is a linear map. Given a vector $(c_1, \ldots, c_n) \in \mathbb{F}^n$, it is simple to write down a vector of V having $(c_1, \ldots, c_n)$ as its coordinate vector (with respect to β). Indeed, the vector $\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$ is that vector and it is the only vector with coordinates $(c_1, \ldots, c_n)$ according to Lemma 9.2.1! What we in fact have found is the inverse function of ϕ_β. Let us put these statements in a lemma and give a complete proof.

Lemma 10.3.1

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of dimension n, and $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, an ordered basis of V. Then the function $\phi_\beta : V \to \mathbb{F}^n$ defined by $\mathbf{v} \mapsto [\mathbf{v}]_\beta$ is a linear map. Moreover, the function $\psi_\beta : \mathbb{F}^n \to V$ defined by $(c_1, \ldots, c_n) \mapsto c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$ is the inverse of ϕ_β and also a linear map.

Proof. We have already shown in the discussion before this lemma that $\phi_\beta : V \to \mathbb{F}^n$ is a linear map of vector spaces over $\mathbb{F}$. Now let us denote by $\psi_\beta : \mathbb{F}^n \to V$ the map defined

by $(c_1, \ldots, c_n) \mapsto c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$. We first show that ψ_β is the inverse function ϕ_β^{-1}. In order to check this, we need to show that $\psi_\beta \circ \phi_\beta(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v} \in V$ as well as that $\phi_\beta \circ \psi_\beta(c_1, \ldots, c_n) = (c_1, \ldots, c_n)$ for all $(c_1, \ldots, c_n) \in \mathbb{F}^n$. We have

$$(\psi_\beta \circ \phi_\beta)(\mathbf{v}) = \psi_\beta([\mathbf{v}]_\beta) = \mathbf{v}$$

and

$$(\phi_\beta \circ \psi_\beta)(c_1, \ldots, c_n) = \phi_\beta(c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n) = (c_1, \ldots, c_n).$$

It is left to the reader to check that ψ_β is a linear map. □

The reason the linear maps ϕ_β and ψ_β are so useful, is that they can be used to describe a general linear map more explicitly. More to the point, suppose that we are given a linear map $L : V_1 \to V_2$ as in Definition 10.0.1, but that we know that both V_1 and V_2 are finite dimensional vector spaces, say that $\dim V_1 = n$ and $\dim V_2 = m$. This means that we can choose an ordered basis of V_1, say $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent vectors in V_1. Similarly, we can choose an ordered basis of V_2, say $\gamma = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$, where $\mathbf{w}_1, \ldots, \mathbf{w}_m \in V_2$ are linearly independent vectors in V_2. Then instead of studying the abstract linear map $L : V_1 \to V_2$, we will study the function $\phi_\gamma \circ L \circ \psi_\beta : \mathbb{F}^n \to \mathbb{F}^m$. The effect is that the abstract vector spaces V_1 and V_2 have been replaced by the more down to earth vector spaces $\mathbb{F}^n$ and $\mathbb{F}^m$. Using Theorem 10.2.1 in combination with Lemma 10.3.1, we can also conclude that the function $\phi_\gamma \circ L \circ \psi_\beta$ actually is a linear map of vector spaces over $\mathbb{F}$, since it is the composite of linear maps.

We have in Section 10.1 seen that any matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$ gives rise to a linear map $L_\mathbf{A} : \mathbb{F}^n \to \mathbb{F}^m$, by defining $\mathbf{v} \mapsto \mathbf{A} \cdot \mathbf{v}$. In fact, any linear map from $\mathbb{F}^n$ to $\mathbb{F}^m$ is of this form. Let us show this now:

Lemma 10.3.2

Let $\mathbb{F}$ be a field and $\tilde{L} : \mathbb{F}^n \to \mathbb{F}^m$ a linear map. Then there exists exactly one matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$ such that $\tilde{L} = L_\mathbf{A}$. Moreover, if we denote by $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the standard basis vectors of $\mathbb{F}^n$, then $\mathbf{A}$ is the matrix whose columns consist of $\tilde{L}(\mathbf{e}_1), \ldots, \tilde{L}(\mathbf{e}_n)$.

Proof. If $\mathbf{v} = (c_1, \ldots, c_n) \in \mathbb{F}^n$, then $\mathbf{v} = c_1 \cdot \mathbf{e}_1 + \cdots + c_n \cdot \mathbf{e}_n$, since the i-th standard basis vector of $\mathbb{F}^n$ has a one in coordinate i and zeroes otherwise. Since $\tilde{L}$ is a linear map, we have $\tilde{L}(\mathbf{v}) = \tilde{L}(c_1 \cdot \mathbf{e}_1 + \cdots + c_n \cdot \mathbf{e}_n) = c_1 \cdot \tilde{L}(\mathbf{e}_1) + \cdots + c_n \cdot \tilde{L}(\mathbf{e}_n)$. Hence the matrix $\mathbf{A}$ with columns $\tilde{L}(\mathbf{e}_1), \ldots, \tilde{L}(\mathbf{e}_n)$ satisfies that $\mathbf{A} \cdot \mathbf{v} = c_1 \cdot \tilde{L}(\mathbf{e}_1) + \cdots + c_n \cdot \tilde{L}(\mathbf{e}_n) = \tilde{L}(\mathbf{v})$. This shows that $\tilde{L} = L_\mathbf{A}$.

What is left to show is that the matrix $\mathbf{A}$ is unique. Suppose that there exist another matrix $\mathbf{B} \in \mathbb{F}^{m \times n}$ such that $\tilde{L} = L_\mathbf{B}$. We want to show that $\mathbf{A} = \mathbf{B}$. If $\mathbf{A} \neq \mathbf{B}$, one can find a column, say column i, where the matrices $\mathbf{A}$ and $\mathbf{B}$ are distinct. Note that the ith column of $\mathbf{A}$ equals $\tilde{L}(\mathbf{e}_i)$ by construction of the matrix $\mathbf{A}$. On the other hand, $\tilde{L}(\mathbf{e}_i) = L_\mathbf{B}(\mathbf{e}_i) = \mathbf{B} \cdot \mathbf{e}_i$, which is precisely the ith column of $\mathbf{B}$. Apparently, the ith columns of $\mathbf{A}$ and $\mathbf{B}$ are both equal to $L(\mathbf{e}_i)$

and not distinct after all. This contradiction show that the assumption $\mathbf{A} \neq \mathbf{B}$ cannot be valid and therefore that $\mathbf{A} = \mathbf{B}$. □

Given a linear map $L : V_1 \to V_2$ we will apply this lemma to the associated linear map $\tilde{L} = \phi_\gamma \circ L \circ \psi_\beta : \mathbb{F}^n \to \mathbb{F}^m$. Let us before continuing with the general theory, first consider an example.

Example 10.3.1

We revisit Example 10.2.5. In that example V_1 was the vector space consisting of polynomials in $\mathbb{C}[Z]$ of degree at most three and V_2 the vector space of polynomials in $\mathbb{C}[Z]$ of degree at most four. Hence as ordered basis for V_1, we can choose $\beta = (1, Z, Z^2, Z^3)$, while a possible ordered basis for V_2 is given by $\gamma = (1, Z, Z^2, Z^3, Z^4)$. The linear map $L : V_1 \to V_2$ described in Example 10.2.5 mapped a polynomial $p(Z)$ to $(i + 2Z) \cdot p(Z)$.

Let us start by explaining what the linear map $\phi_\gamma : V_2 \to \mathbb{F}^5$ is in this case. An element in V_2 is a polynomial of degree at most four. Hence $\mathbf{v} \in V_2$ is a polynomial of the form $a_0 + a_1 Z + \cdots + a_4 Z^4$ with $a_0, a_1, \ldots, a_4 \in \mathbb{C}$, which is already written as a linear combination of the vectors in the ordered basis $(1, Z, \ldots, Z^4)$. Hence $\phi_\gamma(a_0 + a_1 Z + \cdots + a_4 Z^4) = (a_0, a_1, \ldots, a_4)$, or in vector notation:

$$\phi_\gamma(a_0 + a_1 Z + \cdots + a_4 Z^4) = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_4 \end{bmatrix}.$$

Similarly,

$$\psi_\beta\left(\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}\right) = b_0 + b_1 Z + b_2 Z^2 + b_3 Z^3.$$

We can describe the linear map L by figuring out what happens with the vectors in the chosen ordered basis β when L is applied. It is convenient to express the outcome as a linear combination of the vectors in the chosen ordered basis γ. We obtain:

$$L(1) = (i + 2Z) \cdot 1 = i + 2Z, \quad L(Z) = (i + 2Z) \cdot Z = iZ + 2Z^2,$$
$$L(Z^2) = (i + 2Z) \cdot Z^2 = iZ^2 + 2Z^3, \quad L(Z^3) = (i + 2Z) \cdot Z^3 = iZ^3 + 2Z^4.$$

Now let us compute the matrix $\mathbf{A}$ described in Lemma 10.3.2. We need to compute $\tilde{L}(\mathbf{e}_i)$ for $i = 1, \ldots, 4$, where $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4)$ is the standard ordered basis of $\mathbb{F}^4$ and $\tilde{L} = \phi_\gamma \circ L \circ \psi_\beta : \mathbb{F}^4 \to \mathbb{F}^5$. Then we find:

$$\tilde{L}(\mathbf{e}_1) = (\phi_\gamma \circ L \circ \psi_\beta)(\mathbf{e}_1) \quad = \quad (\phi_\gamma \circ L \circ \psi_\beta)\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$
\begin{aligned}
&= (\phi_\gamma \circ L)(1) \\
&= \phi_\gamma(i+2Z) \\
&= \begin{bmatrix} i \\ 2 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

and similarly

$$\tilde{L}(\mathbf{e}_2) = \begin{bmatrix} 0 \\ i \\ 2 \\ 0 \\ 0 \end{bmatrix}, \quad \tilde{L}(\mathbf{e}_3) = \begin{bmatrix} 0 \\ 0 \\ i \\ 2 \\ 0 \end{bmatrix}, \quad \text{and} \quad \tilde{L}(\mathbf{e}_4) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ i \\ 2 \end{bmatrix}.$$

Using Lemma 10.3.2, we see that $\tilde{L} = \phi_\gamma \circ L \circ \psi_\beta = L_{\mathbf{A}}$, where

$$\mathbf{A} = \begin{bmatrix} i & 0 & 0 & 0 \\ 2 & i & 0 & 0 \\ 0 & 2 & i & 0 \\ 0 & 0 & 2 & i \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Definition 10.3.1

Let $\mathbb{F}$ be a field and $L : V_1 \to V_2$ a linear map between two finite dimensional vector spaces, say $\dim V_1 = n$ and $\dim V_2 = m$. Let β be an ordered basis of V_1 and γ one of V_2. Then we denote with ${}_\gamma[L]_\beta \in \mathbb{F}^{m\times n}$ the matrix described in Lemma 10.3.2 when applied to the linear map $\tilde{L} = \phi_\gamma \circ L \circ \psi_\beta : \mathbb{F}^n \to \mathbb{F}^m$. We say that the matrix ${}_\gamma[L]_\beta$ is the *matrix representation* of L with respect to the ordered bases β and γ. One also calls ${}_\gamma[L]_\beta$ the *mapping matrix* of L with respect to the ordered bases β and γ.

To avoid unnecessary computations, let us describe the mapping matrix ${}_\gamma[L]_\beta$ more directly:

Lemma 10.3.3

Let $\mathbb{F}$ be a field and $L : V_1 \to V_2$ a linear map between two finite dimensional vector spaces, say $\dim V_1 = n$ and $\dim V_2 = m$. Let $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be an ordered basis of V_1 and γ one of V_2. Then the mapping matrix of L with respect to the ordered bases β and γ has $[L(\mathbf{v}_1)]_\gamma, \ldots, [L(\mathbf{v}_n)]_\gamma$ as columns. That is to say:

$${}_\gamma[L]_\beta = [[L(\mathbf{v}_1)]_\gamma \cdots [L(\mathbf{v}_n)]_\gamma].$$

Proof. Combining Definition 10.3.1 and Lemma 10.3.2, we see that ${}_\gamma[L]_\beta$ has columns $\tilde{L}(\mathbf{e}_1), \ldots, \tilde{L}(\mathbf{e}_n)$, where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are the standard basis vectors of $\mathbb{F}^n$ and $\tilde{L} = \phi_\gamma \circ L \circ \psi_\beta$. Now note that for all i between 1 and n, we have $\psi_\beta(\mathbf{e}_i) = \mathbf{v}_i$ using the definition of ψ_β given in Lemma 10.3.1. Further, $\phi_\gamma(\mathbf{w}) = [\mathbf{w}]_\gamma$ for all $\mathbf{w} \in V_2$ by definition of the map ϕ_β. Hence we see that for all i between 1 and n, we have

$$\tilde{L}(\mathbf{e}_i) = (\phi_\gamma \circ L \circ \psi_\beta)(\mathbf{e}_i) = (\phi_\gamma \circ L)(\mathbf{v}_i) = \phi_\gamma(L(\mathbf{v}_i)) = [L(\mathbf{v}_i)]_\gamma.$$

□

In Example 10.3.1, we already computed the matrix representation of a linear map (the matrix denoted by $\mathbf{A}$ in the example). Let us consider a few more examples.

Example 10.3.2

This example is a continuation of Example 10.2.4. There, for $\alpha \in \mathbb{R}$, we considered the linear map $R_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $R_\alpha(v_1, v_2) = (\cos(\alpha) \cdot v_1 - \sin(\alpha) \cdot v_2, \sin(\alpha) \cdot v_1 + \cos(\alpha) \cdot v_2)$. Choosing the standard ordered basis $\beta = \gamma = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$ for $\mathbb{R}^2$ both in case of the domain and the codomain of the linear map R_α, we obtain that

$${}_\gamma[R_\alpha]_\beta = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}. \tag{10.4}$$

The point of representing a linear map L with the matrix ${}_\gamma[L]_\beta$, is that the structure of the original linear map is "encoded" in this matrix. The following theorem makes this more precise.

Theorem 10.3.4

Let $\mathbb{F}$ be a field and V_1, V_2 and V_3 three finite dimensional vector spaces over $\mathbb{F}$. Further, let β, γ and δ be ordered bases of respectively V_1, V_2 and V_3. Then one has

(i) $[L(\mathbf{v})]_\gamma = {}_\gamma[L]_\beta \cdot [\mathbf{v}]_\beta$ for any linear map $L : V_1 \to V_2$ and any $\mathbf{v} \in V_1$.

(ii) ${}_\delta[M \circ L]_\beta = {}_\delta[M]_\gamma \cdot {}_\gamma[L]_\beta$ for any linear maps $L : V_1 \to V_2$ and $M : V_2 \to V_3$.

Proof. We first prove the first item. Let us write $\mathbf{A} = {}_\gamma[L]_\beta$ for convenience. We have seen that $\phi_\gamma \circ L \circ \psi_\beta = L_\mathbf{A}$, using the notation from Lemma 10.3.1. Hence $\phi_\gamma \circ L = L_\mathbf{A} \circ (\psi_\beta)^{-1} = L_\mathbf{A} \circ \phi_\beta$. But then for any $\mathbf{v} \in V_1$, we obtain that $(\phi_\gamma \circ L)(\mathbf{v}) = (L_\mathbf{A} \circ \phi_\beta)(\mathbf{v})$. Simplifying the left-hand and right-hand side, we find that

$$(\phi_\gamma \circ L)(\mathbf{v}) = \phi_\gamma(L(\mathbf{v})) = [L(\mathbf{v})]_\gamma$$

and

$$(L_\mathbf{A} \circ \phi_\beta)(\mathbf{v}) = L_\mathbf{A}(\phi_\beta(\mathbf{v})) = L_\mathbf{A}([\mathbf{v}]_\beta) = {}_\gamma[L]_\beta \cdot [\mathbf{v}]_\beta.$$

Hence $[L(\mathbf{v})]_\gamma = {}_\gamma[L]_\beta \cdot [\mathbf{v}]_\beta$, which is what we needed to show.

The proof of the second item is somewhat similar. We write $\mathbf{A} = {}_\gamma[L]_\beta$ and $\mathbf{B} = {}_\delta[M]_\gamma$ for convenience. We have $L_\mathbf{A} = \phi_\gamma \circ L \circ \psi_\beta$ and $L_\mathbf{B} = \phi_\delta \circ M \circ \psi_\gamma$, which implies that $L_\mathbf{B} \circ L_\mathbf{A} = \phi_\delta \circ M \circ \psi_\gamma \circ \phi_\gamma \circ L \circ \psi_\beta$. Now using that ψ_γ and ϕ_γ are each other's inverses, see Lemma 10.3.1, we obtain that $L_\mathbf{B} \circ L_\mathbf{A} = \phi_\delta \circ M \circ L \circ \psi_\beta$. Since on the one hand $L_\mathbf{B} \circ L_\mathbf{A} = L_{\mathbf{B}\cdot\mathbf{A}}$ and on the other hand $\phi_\delta \circ M \circ L \circ \psi_\beta = L_\mathbf{C}$ with $\mathbf{C} = {}_\delta[M \circ L]_\beta$, this implies that ${}_\delta[M \circ L]_\beta = \mathbf{B} \cdot \mathbf{A} = {}_\delta[M]_\gamma \cdot {}_\gamma[L]_\beta$. This is what we wanted to show. □

The first item in this theorem simply tells us that the matrix ${}_\gamma[L]_\beta$ contains all information we need to know to describe the linear map L: computing $L(\mathbf{v})$ and then computing the coordinate vector of the outcome with respect to the ordered basis γ of V_2, is exactly the same as multiplying the matrix ${}_\gamma[L]_\beta$ with the coordinate vector of $\mathbf{v}$ with respect to the ordered basis β of V_1. The second item says that composition of linear maps behaves nice with respect to matrix representations. Let us look at an example of this.

Example 10.3.3

We continue with Example 10.3.2. We have seen that if we choose β and γ to be the standard basis of $\mathbb{R}^2$, then ${}_\gamma[R_\alpha]_\beta$ is as in Equation (10.4). Recall that the map $R_\alpha : \mathbb{R}^2 \to \mathbb{R}^2$ itself, geometrically can be described as a rotation over an angle α against the clock with midpoint in the origin. In particular $R_{\pi/2}$ corresponds with a rotation over $\pi/2$ radians (90 degrees). This means that $R_{\pi/2} \circ R_{\pi/2} = R_\pi$, a rotation over π radians (180 degrees). In particular, this means that $R_\pi(v_1, v_2) = (-v_1, -v_2)$. Let us check the second item in Theorem 10.3.4 for $V_1 = V_2 = V_3 = \mathbb{R}^2$, $\beta = \gamma = \delta = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)$ and $L = M = R_{\pi/2}$. Then on the one hand we have

$$ {}_\delta[R_{\pi/2}]_\gamma = {}_\gamma[R_{\pi/2}]_\beta = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} $$

and therefore

$$ {}_\delta[R_{\pi/2}]_\gamma \cdot {}_\gamma[R_{\pi/2}]_\beta = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}. $$

On the other hand, using Equation (10.4) for $\alpha = \pi$, we see that

$$ {}_\delta[R_{\pi/2} \circ R_{\pi/2}]_\beta = {}_\delta[R_\pi]_\beta = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}. $$

We conclude that indeed ${}_\delta[R_{\pi/2} \circ R_{\pi/2}]_\beta = {}_\delta[R_{\pi/2}]_\gamma \cdot {}_\gamma[R_{\pi/2}]_\beta$, just as it should be.

If one would do the same computation for $M = R_{\alpha_1}$ and $L = R_{\alpha_2}$ and use that $R_{\alpha_1} \circ R_{\alpha_2} = R_{\alpha_1+\alpha_2}$, one obtains that

$$ \begin{bmatrix} \cos(\alpha_1) & -\sin(\alpha_1) \\ \sin(\alpha_1) & \cos(\alpha_1) \end{bmatrix} \cdot \begin{bmatrix} \cos(\alpha_2) & -\sin(\alpha_2) \\ \sin(\alpha_2) & \cos(\alpha_2) \end{bmatrix} = \begin{bmatrix} \cos(\alpha_1+\alpha_2) & -\sin(\alpha_1+\alpha_2) \\ \sin(\alpha_1+\alpha_2) & \cos(\alpha_1+\alpha_2) \end{bmatrix}. $$

This identity actually implies the addition formulas for the cosine and the sine that we used in the proof of Lemma 3.4.1:

$$\cos(\alpha_1 + \alpha_2) = \cos(\alpha_1)\cos(\alpha_2) - \sin(\alpha_1)\sin(\alpha_2)$$

and

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

Example 10.3.4

Let $\mathbb{F} = \mathbb{R}$ and $V_1 = \mathbb{R}^2$, $V_2 = \mathbb{R}^2$ and let

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

Denote by $L_{\mathbf{A}} : \mathbb{R}^2 \to \mathbb{R}^2$ the linear map defined by $\mathbf{v} \mapsto \mathbf{A} \cdot \mathbf{v}$. We have for example

$$L_{\mathbf{A}}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

and

$$L_{\mathbf{A}}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Question:

(a) Choosing the standard ordered bases $\beta = \gamma = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$ for V_1 and V_2, compute ${}_\gamma[L_{\mathbf{A}}]_\beta$.

(b) Choosing the ordered bases $\beta = \gamma = \left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)$ for V_1 and V_2, compute ${}_\gamma[L_{\mathbf{A}}]_\beta$.

Answer:

(a) Since γ is chosen to be the standard basis and

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = v_1 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + v_2 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

we see that $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}_\gamma = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ for all $v_1, v_2 \in \mathbb{R}$. Using Lemma 10.3.3, we see that

$${}_\gamma[L_{\mathbf{A}}]_\beta = [[L_{\mathbf{A}}(\mathbf{e}_1)]_\gamma \; [L_{\mathbf{A}}(\mathbf{e}_2)]_\gamma] = [L_{\mathbf{A}}(\mathbf{e}_1) \; L_{\mathbf{A}}(\mathbf{e}_2)] = [\mathbf{A} \cdot \mathbf{e}_1 \; \mathbf{A} \cdot \mathbf{e}_2] = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} = \mathbf{A}.$$

One can in fact see in a similar way that for any field $\mathbb{F}$ and any matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$, one has ${}_\gamma[L_{\mathbf{A}}]_\beta = \mathbf{A}$ if β and γ are the standard ordered bases of $\mathbb{F}^m$ and $\mathbb{F}^n$.

(b) Now we choose the ordered bases $\beta = \gamma = \left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$ for V_1 and V_2. Using Lemma 10.3.3, we see that

$$\begin{aligned} {}_\gamma[L_{\mathbf{A}}]_\beta &= \left[[L_{\mathbf{A}} \left(\begin{bmatrix} -1 \\ 2 \end{bmatrix} \right)]_\gamma \, [L_{\mathbf{A}}(\begin{bmatrix} 1 \\ 1 \end{bmatrix})]_\gamma \right] = \left[[\mathbf{A} \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix}]_\gamma \, [\mathbf{A} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}]_\gamma \right] \\ &= \left[\begin{bmatrix} 1 \\ -2 \end{bmatrix}_\gamma \, \begin{bmatrix} 2 \\ 2 \end{bmatrix}_\gamma \right]. \end{aligned}$$

Now in order to compute $[\mathbf{w}]_\gamma$ for $\mathbf{w} \in \mathbb{R}^2$, one needs in general to solve a linear system of equations. More precisely, let us write $\mathbf{w} = (w_1, w_2)$, then we want to find $c_1, c_2 \in \mathbb{R}^2$ such that

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = c_1 \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} + c_2 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Therefore we need to solve the system of linear equations in the indeterminates c_1 and c_2 given by:

$$\begin{bmatrix} -1 & 1 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

This can in principle be done using the theory of Chapter 6 or by multiplying the system on both sides of the equality sign with the matrix

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

However, in this case we are lucky, since we can see directly that

$$\begin{bmatrix} 1 \\ -2 \end{bmatrix} = (-1) \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} \quad \text{and hence} \quad \begin{bmatrix} 1 \\ -2 \end{bmatrix}_\gamma = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} 2 \\ 2 \end{bmatrix} = 2 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \text{implying} \quad \begin{bmatrix} 2 \\ 2 \end{bmatrix}_\gamma = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

We conclude that

$${}_\gamma[L_{\mathbf{A}}]_\beta = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}.$$

The result is a surprisingly nice looking matrix: it is a diagonal matrix (see Definition 8.1.3).

As a last item in this section, we consider matrices of the form ${}_\gamma[L]_\beta$ in case L is the identity map from a vector space V to itself: $\mathrm{id}_V : V \to V$, $\mathbf{v} \mapsto \mathbf{v}$. Here β and γ are two, possibly

distinct, ordered bases of V. From the first part of Theorem 10.3.4, we see that

$$ {}_\gamma[\mathrm{id}_V]_\beta \cdot [\mathbf{v}]_\beta = [\mathbf{v}]_\gamma \quad \text{for all } \mathbf{v} \in V. \tag{10.5} $$

In words Equation (10.5) states that if one multiplies the matrix ${}_\gamma[\mathrm{id}_V]_\beta$ with the β-coordinate vector of a vector $\mathbf{v}$ in V, the outcome is the γ-coordinate vector of $\mathbf{v}$. For this reason, the matrix ${}_\gamma[\mathrm{id}_V]_\beta$ is called a *change of coordinates matrix* also known as a *change of basis matrix* .

Example 10.3.5

Let $V = \{p(Z) \in \mathbb{C}[Z] \mid \deg p(Z) \leq 3\}$. Then $\beta = (1, Z, Z^2, Z^3)$ and $\gamma = (Z^3, Z^2, Z, 1)$ are two ordered bases of V.

Question: Compute the corresponding change of coordinates matrix ${}_\gamma[\mathrm{id}_V]_\beta$.

Answer: Using Lemma 10.3.3, what we need to do is to compute $[1]_\gamma, [Z]_\gamma, [Z^2]_\gamma$ and $[Z^3]_\gamma$. Since the only difference between β and γ is the order of the basis vectors this is not so hard to do. For example

$$ [1]_\gamma = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, $$

since 1 is the fourth basis vector of γ. Proceeding similarly for the other basis vectors, one obtains the desired change of coordinates matrix:

$$ {}_\gamma[\mathrm{id}_V]_\beta = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} $$

We finish this section with a few facts on change of coordinates matrices that will come in handy later.

Lemma 10.3.5

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of finite dimension n and β, γ and δ ordered bases of V. Then

(i) ${}_\delta[\mathrm{id}_V]_\gamma \cdot {}_\gamma[\mathrm{id}_V]_\beta = {}_\delta[\mathrm{id}_V]_\beta$,

(ii) ${}_\beta[\mathrm{id}_V]_\beta = \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, and

(iii) $({}_\gamma[\mathrm{id}_V]_\beta)^{-1} = {}_\beta[\mathrm{id}_V]_\gamma$.

Proof. The first item follows directly from the second item in Theorem 10.3.4. The second item is clear, since if the ordered basis β is not changed, the coordinates of a vector with

respect to β do not change either. For the third item, note that according to the first and second part of the theorem, we have ${}_\gamma[\mathrm{id}_V]_\beta \cdot {}_\beta[\mathrm{id}_V]_\gamma = {}_\gamma[\mathrm{id}_V]_\gamma = \mathbf{I}_n$ and similarly ${}_\beta[\mathrm{id}_V]_\gamma \cdot {}_\gamma[\mathrm{id}_V]_\beta = {}_\beta[\mathrm{id}_V]_\gamma = \mathbf{I}_n$. Hence $({}_\gamma[\mathrm{id}_V]_\beta)^{-1} = {}_\beta[\mathrm{id}_V]_\gamma$. □

10.4 Usages of the matrix representation of a linear map

Now that we have the ability to represent linear maps between finite dimensional vector spaces with a matrix, we will use this to describe in more detail how to compute the kernel and image of a linear map. We start with a more general description of solutions to equations involving a linear map.

Theorem 10.4.1

Let $\mathbb{F}$ be a field and $L : V_1 \to V_2$ a linear map between vector spaces over $\mathbb{F}$. Further, let a vector $\mathbf{w} \in V_2$ be given and denote by $S = \{\mathbf{v} \in V_1 \mid L(\mathbf{v}) = \mathbf{w}\}$. Then exactly one of the following two possibilities occurs:

(i) $S = \emptyset$. This is the case if and only if $\mathbf{w} \notin \mathrm{image} L$.

(ii) $S = \{\mathbf{v}_p + \mathbf{v} \mid \mathbf{v} \in \ker L\}$, where $\mathbf{v}_p \in V_1$ is a vector such that $L(\mathbf{v}_p) = \mathbf{w}$.

Proof. If $S = \emptyset$, then the equation $L(\mathbf{v}) = \mathbf{w}$ has no solutions. This is equivalent to the statement that no vector $\mathbf{v} \in V_1$ is mapped to $\mathbf{w}$. This in turn is the same as saying that $\mathbf{w}$ is not in the image of L.

If $S \neq \emptyset$, we may conclude that there exists a vector $\mathbf{v}_p \in V_1$ such that $L(\mathbf{v}_p) = \mathbf{w}$. If $\tilde{\mathbf{v}}$ is some vector, such that $L(\tilde{\mathbf{v}}) = \mathbf{w}$, then using linearity of L, we see that $L(\tilde{\mathbf{v}} - \mathbf{v}_p) = \mathbf{w} - \mathbf{w} = \mathbf{0}$. Hence $\tilde{\mathbf{v}} - \mathbf{v}_p \in \ker L$. Since $\tilde{\mathbf{v}} = \mathbf{v}_p + (\tilde{\mathbf{v}} - \mathbf{v}_p)$ and, as we already have seen $\tilde{\mathbf{v}} - \mathbf{v}_p \in \ker L$, this shows that $S \subseteq \{\mathbf{v}_p + \mathbf{v} \mid \mathbf{v} \in \ker L\}$. Conversely, if a vector is of the form $\mathbf{v}_p + \mathbf{v}$ for some $\mathbf{v} \in \ker L$, then $L(\mathbf{v}_p + \mathbf{v}) = L(\mathbf{v}_p) + L(\mathbf{v}) = \mathbf{w} + \mathbf{0} = \mathbf{w}$. This shows that $\{\mathbf{v}_p + \mathbf{v} \mid \mathbf{v} \in \ker L\} \subseteq S$. Combining both inclusions, we may conclude that $S = \{\mathbf{v}_p + \mathbf{v} \mid \mathbf{v} \in \ker L\}$. □

Hence the structure of the solution set of an equation of the form $L(\mathbf{v}) = \mathbf{w}$ is completely determined. The vector $\mathbf{v}_p$, if it exists, is called a *particular solution*. Notice how similar this is to Theorem 6.1.2. This is not a coincidence. After all, the solution set to a system of linear equations with augmented matrix $[\mathbf{A}|\mathbf{b}]$ is exactly the same as the solution set to the equation $L_\mathbf{A}(\mathbf{v}) = \mathbf{b}$. Moreover, $\ker L_\mathbf{A}$ is exactly the same as the solution set to the homogeneous system of linear equations with coefficient matrix $\mathbf{A}$. Hence, Theorem 6.1.2 is really just a special case of Theorem 10.4.1.

In case both V_1 and V_2 are finite dimensional vector spaces, we can computationally solve an equation of the form $L(\mathbf{v}) = \mathbf{w}$ by solving a suitable system of linear equations. We make this more precise in the following theorem.

Theorem 10.4.2

Let $\mathbb{F}$ be a field and $L : V_1 \to V_2$ a linear map between finite dimensional vector spaces over $\mathbb{F}$. Let $\beta = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be an ordered basis of V_1 and $\gamma = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ be an ordered basis of V_2. Then

$$\{\mathbf{v} \in V_1 \mid L(\mathbf{v}) = \mathbf{w}\} = \{c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n \mid \mathbf{c} = (c_1, \ldots, c_n) \text{ satisfies } {}_\gamma[L]_\beta \cdot \mathbf{c} = [\mathbf{w}]_\gamma\}.$$

In particular

$$\ker L = \{c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n \mid (c_1, \ldots, c_n) \in \ker {}_\gamma[L]_\beta\}.$$

Proof. Applying Lemma 10.3.1 to the vector space V_1 and the given ordered basis β, we see that the linear maps $\phi_\beta : V_1 \to \mathbb{F}^n$ defined by $\mathbf{v} \mapsto [\mathbf{v}]_\beta$ and $\psi_\beta : \mathbb{F}^n \to V_1$ defined by $(c_1, \ldots, c_n) \mapsto c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$ are inverses of each other.

Assume that $L(\mathbf{v}) = \mathbf{w}$, then $[L(\mathbf{v})]_\gamma = [\mathbf{w}]_\gamma$, which using the first item in Theorem 10.3.4 implies that ${}_\gamma[L]_\beta[\mathbf{v}]_\beta = [\mathbf{w}]_\gamma$. If we write $(c_1, \ldots, c_n) = [\mathbf{v}]_\beta$, then $\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n$. This shows that

$$\{\mathbf{v} \in V_1 \mid L(\mathbf{v}) = \mathbf{w}\} \subseteq \{c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n \mid \mathbf{c} = (c_1, \ldots, c_n) \text{ satisfies } {}_\gamma[L]_\beta \cdot \mathbf{c} = [\mathbf{w}]_\gamma\}.$$

Conversely, assume that $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{F}^n$ satisfying ${}_\gamma[L]_\beta \cdot \mathbf{c} = [\mathbf{w}]_\gamma$ is given. The vector $\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n$ has the property that $[\mathbf{v}]_\beta = \mathbf{c}$. Therefore ${}_\gamma[L]_\beta \cdot [\mathbf{v}]_\beta = [\mathbf{w}]_\gamma$. Using Theorem 10.3.4 again, we see that $[L(\mathbf{v})]_\gamma = [\mathbf{w}]_\gamma$. But then $L(\mathbf{v}) = \mathbf{w}$. This shows that

$$\{\mathbf{v} \in V_1 \mid L(\mathbf{v}) = \mathbf{w}\} \supseteq \{c_1 \cdot \mathbf{v}_1 + \cdots c_n \cdot \mathbf{v}_n \mid \mathbf{c} = (c_1, \ldots, c_n) \text{ satisfies } {}_\gamma[L]_\beta \cdot \mathbf{c} = [\mathbf{w}]_\gamma\}.$$

Combining the above two inclusions, we see that the first part of the theorem follows.

Choosing $\mathbf{w} = \mathbf{0}$, the statement on $\ker L$ follows. □

The point of this theorem is that in order to compute all solutions to the equation $L(\mathbf{v}) = \mathbf{w}$, it is enough to compute all solutions to the equation ${}_\gamma[L]_\beta \cdot \mathbf{c} = [\mathbf{w}]_\gamma$. The latter equation is a system of linear equation with augmented matrix $[{}_\gamma[L]_\beta | [\mathbf{w}]_\gamma]$, which we can solve using the techniques from Chapter 6. The fact that the kernel of a linear map can be computed using the matrix representation of that map, has a nice consequence known as the *rank-nullity theorem for linear maps*.

Corollary 10.4.3

Let $\mathbb{F}$ be a field and $L : V_1 \to V_2$ a linear map between finite dimensional vector spaces over

$\mathbb{F}$. Then

$$\dim(\ker L) + \dim(\text{image} L) = \dim V_1.$$

Proof. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ is basis of $\ker L$, then $\{[\mathbf{v}_1]_\beta, \ldots, [\mathbf{v}_d]_\beta\}$ is a basis of $\ker {}_\gamma[L]_\beta$ using Theorem 9.2.3. Hence $\dim \ker L = \dim \ker {}_\gamma[L]_\beta$. Moreover, $\dim \text{image} L = \dim \text{image}_\gamma[L]_\beta$ using Corollary 9.3.5. Then the result follows from the rank-nullity theorem for matrices (see Theorem 10.1.3). □

Example 10.4.1

This example is a variation of Example 10.2.9. In that example, we consider the map $\text{ev} : \mathbb{C}[Z] \to \mathbb{C}^2$ defined by $p(Z) \mapsto (p(0), p(1))$ and computed its kernel. Let $V_1 \subseteq \mathbb{C}[Z]$ be the subspace of $\mathbb{C}[Z]$ consisting of all polynomials of degree at most three. Then $\beta = (1, Z, Z^2, Z^3)$ is an ordered basis of V_1. For $\mathbb{C}^2$, we choose the standard ordered basis $(\mathbf{e}_1, \mathbf{e}_2)$. Now let us consider the linear map $L : V_1 \to \mathbb{C}^2$ defined by $L(p(Z)) = (p(0), p(1))$. In other words: we restrict the domain of ev to V_1, but otherwise do not change anything.

Questions: What is the kernel of the linear map L described above? What are all solutions to the equation $L(p(Z)) = (5, 8)$?

Answer:

We can compute $\ker L$ in several ways, but let use Theorem 10.4.2. To compute $\ker L$, we first compute the kernel of ${}_\gamma[L]_\beta$. We have $L(1) = (1,1) = 1 \cdot \mathbf{e}_1 + 1 \cdot \mathbf{e}_2$, $L(Z) = (0,1) = 1 \cdot \mathbf{e}_2$, $L(Z^2) = (0,1) = 1 \cdot \mathbf{e}_2$, and $L(Z^3) = (0,1) = 1 \cdot \mathbf{e}_2$. Hence

$${}_\gamma[L]_\beta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Computing the reduced row echelon form of this matrix in this case just amounts to subtracting the first row from the second row. One finds the matrix:

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

Hence using Theorem 6.4.4 and Corollary 9.3.4, we find that a basis of $\ker {}_\gamma[L]_\beta$ is given by the set

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

and hence a basis of $\ker L$ is given by the set $\{-Z + Z^2, -Z + Z^3.\}$ Hence

$$\ker L = \{t_1 \cdot (-Z + Z^2) + t_2 \cdot (-Z + Z^3) \mid t_1, t_2 \in \mathbb{C}\}.$$

To solve the final question about the solutions to the equation $L(p(Z)) = (5,8)$, we use Theorem 10.4.1. All we still need to do is to compute a particular solution. We could in principle again transform the equation into a system of linear equations. Doing this would give rise to a system of inhomogeneous linear equations with augmented matrix

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 5 \\ 1 & 1 & 1 & 1 & 8 \end{array}\right],$$

which has reduced row echelon form

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 5 \\ 0 & 1 & 1 & 1 & 3 \end{array}\right].$$

A particular solution (c_1, c_2, c_3, c_4) should satisfy $c_1 = 5$ and $c_2 + c_3 + c_4 = 3$. Therefore $(5,3,0,0)$ is a particular solution, which corresponds to the polynomial $f(Z) = 5 + 3Z$. Using Theorem 10.4.1, we conclude that all solutions to the equation $L(p(Z)) = (5,8)$ form the set

$$\{5 + 3Z + t_1 \cdot (-Z + Z^2) + t_2 \cdot (-Z + Z^3) \mid t_1, t_2 \in \mathbb{C}\}.$$

Just as an aside: another way to compute $\ker L$ is to use that we already have computed the kernel of $\mathrm{ev} : \mathbb{C}[Z] \to \mathbb{C}^2$ in Example 10.2.9. Then

$$\begin{aligned} \ker L &= \ker \mathrm{ev} \cap V_1 \\ &= \{Z \cdot (Z-1) \cdot s(Z) \mid s(Z) \in \mathbb{C}[Z]\} \cap V_1 \\ &= \{Z \cdot (Z-1) \cdot s(Z) \mid s(Z) \in \mathbb{C}[Z], \deg s(Z) \leq 1\}. \end{aligned}$$

Here we used that $Z \cdot (Z-1) \cdot s(Z) \in V_1$ precisely if $\deg(Z \cdot (Z-1) \cdot s(Z)) \leq 3$. Since $\deg(Z \cdot (Z-1) \cdot s(Z)) = 1 + 1 + \deg s(Z)$, we see that $Z \cdot (Z-1) \cdot s(Z) \in V_1$ precisely if $\deg s(Z) \leq 1$. It is left to the reader to check that this computation of $\ker L$ gives exactly the same result as before.

Chapter 11

The eigenvalue problem and diagonalization

Let us take a look at Example 10.3.4 again. In it, we considered the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

and the linear map $L_{\mathbf{A}} : \mathbb{R}^2 \to \mathbb{R}^2$ associated to it. We further saw that if we chose the same ordered basis $\beta = \left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$ for both the domain and the codomain of $L_{\mathbf{A}}$, then the resulting mapping matrix ${}_\beta[L_{\mathbf{A}}]_\beta$ of $L_{\mathbf{A}}$ was particularly nice:

$${}_\beta[L_{\mathbf{A}}]_\beta = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}.$$

In this chapter, we investigate to which extent this can be done for an arbitrary square matrix.

11.1 Eigenvalues and eigenvectors

We start out by studying linear maps $L : V \to V$. The difference with our previous studies of linear maps is that we now assume that the domain of L is the same as the codomain of L, namely the vector space V.

Definition 11.1.1

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ and $L : V \to V$ a linear map. Let $\mathbf{v} \in V$ be a nonzero

vector and $\lambda \in \mathbb{F}$ a scalar such that

$$L(\mathbf{v}) = \lambda \cdot \mathbf{v}.$$

Then the vector $\mathbf{v}$ is called an *eigenvector* of the linear map L with *eigenvalue* λ.

Note that by definition an eigenvector is always a nonzero vector. The reason for this is to avoid uninteresting solutions to the equation $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$. Indeed, if one chooses $\mathbf{v} = \mathbf{0}$ and any $\lambda \in \mathbb{F}$, then it will hold that $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$, since $L(\mathbf{0}) = \mathbf{0}$ and $\lambda \cdot \mathbf{0} = \mathbf{0}$. Further note that an eigenvalue always is an element from the field $\mathbb{F}$ over which V is a vector space. Intuitively, what an eigenvector of a linear operator L is, is a vector that is scaled when L operates on it. Indeed, we can think of $\lambda \cdot \mathbf{v}$ as a scaling of the vector $\mathbf{v}$ by a factor λ. See Figure 11.1 for an illustration.

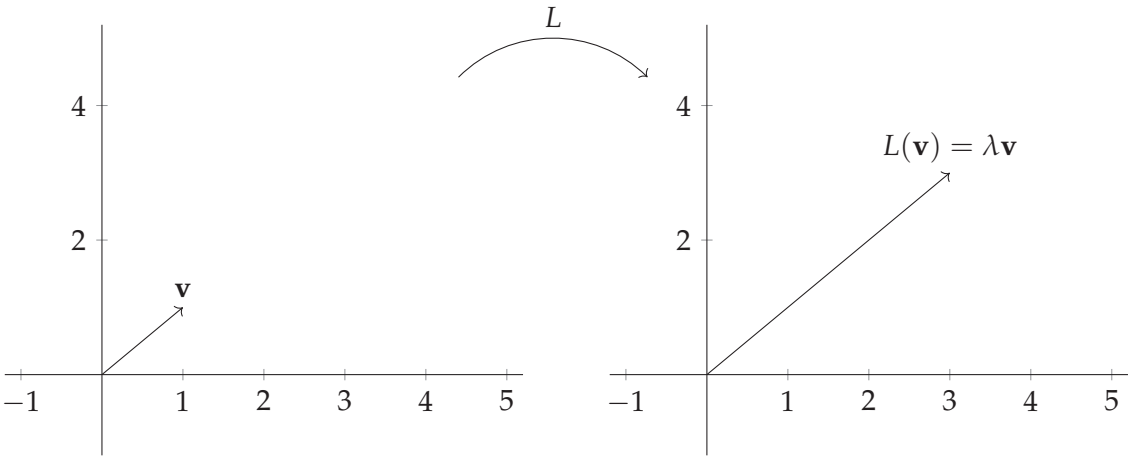


Figure 11.1: An eigenvector of a linear map L.

For matrices one can also talk about eigenvectors and eigenvalues:

Definition 11.1.2

Let $\mathbb{F}$ be a field, n a positive integer and $\mathbf{A} \in \mathbb{F}^{n \times n}$ a matrix. Let $\mathbf{v} \in \mathbb{F}^n$ be a nonzero vector and $\lambda \in \mathbb{F}$ a scalar such that

$$\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}.$$

Then the vector $\mathbf{v}$ is called an *eigenvector* of the matrix $\mathbf{A}$ with *eigenvalue* λ.

Note that this definition assumed that the matrix $\mathbf{A}$ is a square matrix. As we have seen, a matrix $\mathbf{A} \in \mathbb{F}^{m \times n}$ gives rise to a linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^m$. If $m = n$, we therefore see that a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ gives rise to a linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^n$. Note that $\mathbf{v} \in \mathbb{F}^n$ is an eigenvector of a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ if and only if $\mathbf{v} \in \mathbb{F}^n$ is an eigenvector of the linear map $L_{\mathbf{A}} : \mathbb{F}^n \to \mathbb{F}^n$. In that sense Definition 11.1.2 is just a special case of Definition 11.1.1. Also for matrices it holds that if the field is specified to be $\mathbb{F}$, then its eigenvalues are by definition elements of that field $\mathbb{F}$.

Let us consider some examples.

Example 11.1.1

Let $\mathbb{F} = \mathbb{R}$ and let us consider the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Question: Determine all eigenvalues of the matrix $\mathbf{A}$ as well as a corresponding eigenvector for each eigenvalue.

Answer: Assume that $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$ is an eigenvector with eigenvalue λ. The equation $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$ is equivalent to the two equations $-v_1 = \lambda v_1$ and $2v_2 = \lambda v_2$. These two equations can be rewritten as $(-1-\lambda)v_1 = 0$ and $(2-\lambda)v_2 = 0$, which in turn can be written in matrix form as follows:

$$\begin{bmatrix} -1-\lambda & 0 \\ 0 & 2-\lambda \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{11.1}$$

Now we distinguish three cases.

In the first case we assume that $-1-\lambda \neq 0$ and $2-\lambda \neq 0$. In other words: we assume that $\lambda \neq -1$ and $\lambda \neq 2$. In this case the diagonal elements of the matrix occurring in Equation (11.1) are both nonzero. Hence the only solution to Equation (11.1) is $(v_1, v_2) = (0,0)$. However, eigenvectors are by definition not equal to the zero vector, so we conclude that in this case there are no eigenvectors with eigenvalue λ.

In case two, we assume that $\lambda = -1$. In this case Equation (11.1) has solutions of the form $(v_1, 0)$, where $v_1 \in \mathbb{R}$ can be chosen freely. Hence $\lambda = -1$ is an eigenvalue of the given matrix $\mathbf{A}$. As eigenvector we can choose any vector of the form $\begin{bmatrix} v_1 \\ 0 \end{bmatrix}$ as long as $v_1 \neq 0$. For example $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is an eigenvector of the given matrix $\mathbf{A}$ with eigenvalue -1.

Finally, as the third and final case, we assume that $\lambda = 2$. In this case Equation (11.1) has solutions of the form $(0, v_2)$, where $v_2 \in \mathbb{R}$ can be chosen freely. Hence $\lambda = 2$ is an eigenvalue of the given matrix $\mathbf{A}$. As eigenvector we can choose any vector of the form $\begin{bmatrix} 0 \\ v_2 \end{bmatrix}$ as long as $v_2 \neq 0$. For example $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is an eigenvector of the given matrix $\mathbf{A}$ with eigenvalue 2.

Also in case V is an infinite dimensional vector space, the definition of eigenvectors and eigenvalues makes sense. We consider an example of this type.

Example 11.1.2

Let us consider the linear map $D : \mathbb{C}[Z] \to \mathbb{C}[Z]$ defined in Example 10.2.1. In particular, we

are working over the field $\mathbb{C}$, since in Example 10.2.1 we considered $\mathbb{C}[Z]$ as a complex vector space. The map D was defined by sending a polynomial to its derivative.

Question: What are the eigenvalues of D? Also for each eigenvalue, find a corresponding eigenvector.

Answer: We are looking for nonzero polynomials $p(Z)$ in $\mathbb{C}[Z]$ and scalars $\lambda \in \mathbb{C}$ such that $D(p(Z)) = \lambda \cdot p(Z)$. Let $p(Z) = a_0 + a_1 Z + a_2 Z^2 + \cdots + a_n Z^n$ be a nonzero polynomial. Since $D(a_0 + a_1 Z + a_2 Z^2 + \cdots + a_n Z^n) = a_1 + 2a_2 Z + \cdots + n a_n Z^{n-1}$, the degree of the polynomial $D(p(Z))$ will typically be one less than the degree of the polynomial $p(Z)$ itself. The only exception is if $p(Z) = a_0$, in which case $D(p(Z)) = 0$. Hence $D(p(Z)) = \lambda \cdot p(Z)$ can only hold for constant polynomials. If $p(Z)$ is a constant polynomial, then $p(Z) = a_0$ and $D(a_0) = 0 = 0 \cdot a_0$. This shows that 0 is the only eigenvalue that the linear map D has. Any polynomial $p(Z) = a_0$ with $a_0 \in \mathbb{C} \setminus \{0\}$ is an eigenvector of D with eigenvalue 0. The reason that the zero polynomial is not an eigenvector of D is that by definition eigenvectors must be nonzero. As this example shows, eigenvalues themselves can be zero.

The previous two examples, may suggest that a linear map always has at least one eigenvector, but this is not the case. Let us consider such an example.

Example 11.1.3

The rotation map $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ from Example 10.2.4 was defined by $R_{\pi/2}(v_1, v_2) = (-v_2, v_1)$.

Question: Does the linear map $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ have any eigenvectors?

Answer: Let us first give an intuitive answer and after that one using the definitions more directly. What the linear map $R_{\pi/2}$ does geometrically, is to take a vector as input and return as output the vector rotated over $\pi/2$ radians against the clock. If a nonzero vector would be an eigenvector, that would mean that rotation over $\pi/2$ radians would output a scaling of the input vector. This is intuitively not possible, so what we expect is that the linear map $R_{\pi/2}$ has no eigenvectors at all.

Let us now proceed to prove this using the definitions. If the map $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ does have eigenvectors, there exists $(v_1, v_2) \in \mathbb{R}^2 \setminus \{(0,0)\}$ and $\lambda \in \mathbb{R}$ such that $R_{\pi/2}(v_1, v_2) = \lambda \cdot (v_1, v_2)$. Equivalently $(-v_2, v_1) = (\lambda \cdot v_1, \lambda \cdot v_2)$, which in turn can be rewritten as the two equations $-\lambda v_1 - v_2 = 0$ and $v_1 - \lambda v_2 = 0$. Formulated in matrix form, we would get the matrix equation

$$\begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Adding λ times the second row to the first row, we obtain the equation $-(1 + \lambda^2)v_2 = 0$. But this implies that $v_2 = 0$, since $\lambda^2 + 1$ is not zero for any $\lambda \in \mathbb{R}$. Then using the second row, we also see that $v_1 = 0$. We conclude that $(v_1, v_2) = (0, 0)$, but eigenvectors were not allowed to be the zero vector. Hence the linear map $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ has no eigenvectors.

The procedure of determining the eigenvectors and eigenvalues in the previous examples was quite ad hoc. Fortunately, there is a procedure that always works in case V has a finite dimension. We will explain this procedure now, starting with eigenvalues of a square matrix.

Theorem 11.1.1

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{n\times n}$ a square matrix. Then $\lambda \in \mathbb{F}$ is an eigenvalue of $\mathbf{A}$ if and only if $\det(\mathbf{A} - \lambda \cdot \mathbf{I}_n) = 0$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

Proof. If $\lambda \in \mathbb{F}$ is an eigenvalue of the matrix $\mathbf{A}$, then there exists a nonzero vector $\mathbf{v} \in \mathbb{F}^n$ such that $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$. Since $\lambda \cdot \mathbf{v} = \lambda \cdot (\mathbf{I}_n \cdot \mathbf{v}) = (\lambda \cdot \mathbf{I}_n) \cdot \mathbf{v}$, we see that the equation $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$ can be rewritten as $\mathbf{A} \cdot \mathbf{v} = (\lambda \cdot \mathbf{I}_n)\mathbf{v}$, which in turn can be rewritten as $(\mathbf{A} - \lambda \cdot \mathbf{I}_n) \cdot \mathbf{v} = \mathbf{0}$. This shows that the homogeneous system of linear equations with coefficient matrix $\mathbf{A} - \lambda \cdot \mathbf{I}_n$ has a nonzero solution. Using Corollary 8.3.6 for the square matrix $\mathbf{A} - \lambda \cdot \mathbf{I}_n$, we conclude that $\det(\mathbf{A} - \lambda \cdot \mathbf{I}_n) = 0$.

Conversely, if $\det(\mathbf{A} - \lambda \cdot \mathbf{I}_n) = 0$, Corollary 8.3.6 implies that the homogeneous system of linear equations with coefficient matrix $\mathbf{A} - \lambda \cdot \mathbf{I}_n$ has a nonzero solution. Any such nonzero solution $\mathbf{v} \in \mathbb{F}^n$ then satisfies $(\mathbf{A} - \lambda \cdot \mathbf{I}_n) \cdot \mathbf{v} = \mathbf{0}$. This can be rewritten as $\mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}$. Hence $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ. □

For a given square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$, the expression $\det(\mathbf{A} - Z \cdot \mathbf{I}_n)$ is a polynomial in $\mathbb{F}[Z]$ of degree n. This polynomial is called the *characteristic polynomial* of $\mathbf{A}$. We will denote it by $p_{\mathbf{A}}(Z)$. The roots of this polynomial in the field $\mathbb{F}$ are exactly all the eigenvalues of the matrix $\mathbf{A}$.

Example 11.1.4

Theorem 11.1.1 makes it possible to describe all eigenvalues of a square matrix. For example, for the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

that we considered in Example 11.1.1, we have

$$p_{\mathbf{A}}(Z) = \det(\mathbf{A} - Z \cdot \mathbf{I}_2) = \det\left(\begin{bmatrix} -1-Z & 0 \\ 0 & 2-Z \end{bmatrix}\right) = (-1-Z)\cdot(2-Z) = (Z+1)\cdot(Z-2).$$

Therefore the roots of the characteristic polynomial $p_{\mathbf{A}}(Z)$ are precisely -1 and 2. This means that the eigenvalues of the matrix $\mathbf{A}$ are -1 and 2. Looking back at Example 11.1.1 we can make the answer given there a bit shorter, since the first case we considered is no longer needed. Indeed, in the first case, we considered all λ such that $\lambda \neq -1$ and $\lambda \neq 2$, but now we know already that there are no eigenvectors with such an eigenvalue λ.

Example 11.1.5

As a second example, let us consider the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

We have seen in Example 10.3.3 that his matrix represents the linear map $R_{\pi/2} : \mathbb{R}^2 \to \mathbb{R}^2$ if the standard ordered basis is chosen for $\mathbb{R}^2$. In this case

$$p_{\mathbf{A}}(Z) = \det(\mathbf{A} - Z \cdot \mathbf{I}_2) = \det\left(\begin{bmatrix} -Z & -1 \\ 1 & -Z \end{bmatrix}\right) = Z^2 + 1.$$

Since we are working over the real numbers $\mathbb{R}$ and the polynomial $Z^2 + 1$ has no roots in $\mathbb{R}$, we conclude that the matrix $\mathbf{A}$, when studied over $\mathbb{R}$, has no eigenvalues and hence no eigenvectors.

Now that we know how to find eigenvalues of a square matrix, it is natural to ask how to find eigenvectors. We will return to that question in the next section. In the remainder of this section, we will explain how to find eigenvalues of an arbitrary linear map $L : V \to V$ in case V is a finite dimensional vector space.

Theorem 11.1.2

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of dimension n and $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ an ordered basis of V. Then $\lambda \in \mathbb{F}$ is an eigenvalue of a linear map $L : V \to V$ if and only if $\det({}_\beta[L]_\beta - \lambda \cdot \mathbf{I}_n) = 0$.

Proof. If $\lambda \in \mathbb{F}$ is an eigenvalue of the linear map $L : V \to V$, then there exists a nonzero vector $\mathbf{v} \in V$ such that $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$. Hence by the first item in Theorem 10.3.4, we have ${}_\beta[L]_\beta \cdot [\mathbf{v}]_\beta = [L(\mathbf{v})]_\beta = [\lambda \cdot \mathbf{v}]_\beta$. Applying Lemma 9.2.2, we have $[\lambda \cdot \mathbf{v}]_\beta = \lambda \cdot [\mathbf{v}]_\beta$. Combining these two equations, we obtain that ${}_\beta[L]_\beta \cdot [(\mathbf{v})]_\beta = \lambda \cdot [\mathbf{v}]_\beta$. Hence $[\mathbf{v}]_\beta$ is an eigenvector of the matrix ${}_\beta[L]_\beta$ with eigenvalue λ.

Conversely, suppose that $\det({}_\beta[L]_\beta - \lambda \cdot \mathbf{I}_n) = 0$ for some $\lambda \in \mathbb{F}$. Then by Theorem 11.1.1, λ is an eigenvalue of the matrix ${}_\beta[L]_\beta$. Hence there exists a nonzero vector $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{F}^n$ that is an eigenvector of the matrix ${}_\beta[L]_\beta$ with eigenvalue λ. Now define $\mathbf{v} = c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n \in V$. Then $\mathbf{c} = [\mathbf{v}]_\beta$. Hence we have ${}_\beta[L]_\beta \cdot [(\mathbf{v})]_\beta = \lambda \cdot [\mathbf{v}]_\beta$, which implies $[L(\mathbf{v})]_\beta = \lambda \cdot [\mathbf{v}]_\beta = [\lambda \cdot \mathbf{v}]_\beta$. This implies that $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$. Hence λ is an eigenvalue of the linear map $L : V \to V$. □

This theorem shows that if V is a finite dimensional vector space, we can reduce the calculation of eigenvalues of a linear map $L : V \to V$ directly to the calculation of the eigenvalues of the square matrix ${}_\beta[L]_\beta$ representing the linear map. Here it does not matter at all, which ordered basis of V one chooses. For future use, let us nonetheless investigate the effect of choosing

another ordered basis on the matrix representing L. Here the change of coordinate matrices introduced in Equation (10.5) will play an important role.

Lemma 11.1.3

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of dimension n, and $L : V \to V$ a linear map. Further let β and γ be two ordered bases of V and denote by $\mathrm{id}_V : V \to V$ the identity map $\mathbf{v} \mapsto \mathbf{v}$. Then

$$ {}_\gamma[L]_\gamma = ({}_\beta[\mathrm{id}_V]_\gamma)^{-1} \cdot {}_\beta[L]_\beta \cdot {}_\beta[\mathrm{id}_V]_\gamma. $$

Proof. We know from Lemma 10.3.5 that $({}_\beta[\mathrm{id}_V]_\gamma)^{-1} = {}_\gamma[\mathrm{id}_V]_\beta$. Hence

$$ \begin{aligned} ({}_\beta[\mathrm{id}_V]_\gamma)^{-1} \cdot {}_\beta[L]_\beta \cdot {}_\beta[\mathrm{id}_V]_\gamma &= {}_\gamma[\mathrm{id}_V]_\beta \cdot {}_\beta[L]_\beta \cdot {}_\beta[\mathrm{id}_V]_\gamma \\ &= {}_\gamma[\mathrm{id}_V]_\beta \cdot {}_\beta[L \circ \mathrm{id}_V]_\gamma \\ &= {}_\gamma[\mathrm{id}_V \circ L \circ \mathrm{id}_V]_\gamma \\ &= {}_\gamma[L]_\gamma. \end{aligned} $$

In the second and third equality, we used the first item of Theorem 10.3.4. □

Two square matrices $\mathbf{A} \in \mathbb{F}^{n\times n}$ and $\mathbf{B} \in \mathbb{F}^{n\times n}$ are called *similar* if there exists an invertible matrix $\mathbf{Q} \in \mathbb{F}^{n\times n}$ such that $\mathbf{A} = \mathbf{Q}^{-1} \cdot \mathbf{B} \cdot \mathbf{Q}$. Hence Lemma 11.1.3 can be rephrased in words as follows: the effect of choosing a different ordered basis of V is that the matrix representing L is replaced by a similar matrix. It turns out that this lemma also explains why it does not matter which ordered basis one chooses when computing the eigenvalues of a linear map. In fact, we have the following:

Theorem 11.1.4

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of dimension n, and $L : V \to V$ a linear map. Further let β and γ be two ordered bases of V. Then the characteristic polynomials of ${}_\beta[L]_\beta$ and ${}_\gamma[L]_\gamma$ are identical.

Proof. For convenience, let us write $\mathbf{Q} = {}_\beta[\mathrm{id}_V]_\gamma$. Using Lemma 11.1.3, we see that:

$$ \begin{aligned} p_{{}_\gamma[L]_\gamma}(Z) &= \det({}_\gamma[L]_\gamma - Z \cdot \mathbf{I}_n) \\ &= \det(\mathbf{Q}^{-1} \cdot {}_\beta[L]_\beta \cdot \mathbf{Q} - Z \cdot \mathbf{I}_n) \\ &= \det(\mathbf{Q}^{-1} \cdot {}_\beta[L]_\beta \cdot \mathbf{Q} - Z \cdot \mathbf{Q}^{-1} \cdot \mathbf{Q}) \\ &= \det(\mathbf{Q}^{-1} \cdot {}_\beta[L]_\beta \cdot \mathbf{Q} - Z \cdot \mathbf{Q}^{-1} \cdot \mathbf{I}_n \cdot \mathbf{Q}) \\ &= \det(\mathbf{Q}^{-1} \cdot ({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \cdot \mathbf{Q}). \end{aligned} $$

At this point Theorem 8.3.3 comes in handy. Using this theorem, we can namely continue as follows:

$$\begin{aligned} p_{\gamma[L]_\gamma}(Z) &= \det(\mathbf{Q}^{-1} \cdot ({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \cdot \mathbf{Q}) \\ &= \det(\mathbf{Q}^{-1}) \cdot \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \cdot \det(\mathbf{Q}) \\ &= \det(\mathbf{Q}^{-1}) \cdot \det(\mathbf{Q}) \cdot \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \\ &= \det(\mathbf{Q})^{-1} \cdot \det(\mathbf{Q}) \cdot \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \\ &= 1 \cdot \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \\ &= \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \\ &= p_{\beta[L]_\beta}(Z). \end{aligned}$$

This is exactly what we wanted to show. □

Corollary 11.1.5

With the same notation as before, we have $\det {}_\beta[L]_\beta = \det {}_\gamma[L]_\gamma$.

Proof. This follows by putting $Z = 0$ in the characteristic polynomials $p_{\beta[L]_\beta}(Z)$ and $p_{\gamma[L]_\gamma}(Z)$. □

We can now define the characteristic polynomial of a linear map $L : V \to V$ as long as V is a finite dimensional vector space.

Definition 11.1.3

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of finite dimension n, and $L : V \to V$ a linear map. Then the *characteristic polynomial* is defined to be that polynomial $p_L(Z) = \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \in \mathbb{F}[Z]$, where β is some ordered basis of V.

The reason this definition makes sense, is that by Theorem 11.1.4, the choice of the ordered basis β does not matter: a different choice will not change the corresponding characteristic polynomial. In a similar way, based on Corollary 11.1.5, one can define the determinant of such a linear map: $\det L = \det {}_\beta[L]_\beta$.

Example 11.1.6

As an example, we will consider a linear map, similar to the linear map $D : \mathbb{C}[Z] \to \mathbb{C}[Z]$ from Example 10.2.1. However, since $\mathbb{C}[Z]$ is an infinitely dimensional vector space, we will modify the domain and codomain of the map a bit. More precisely, let V be the complex vector space of polynomials of degree at most three. Then we can define $\tilde{D} : V \to V$ as $p(Z) \mapsto p(Z)'$.

Question: What is the characteristic polynomial $p_{\tilde{D}}(\lambda)$ of the linear map $\tilde{D}$?

Answer: Let us choose the ordered basis $\beta = (1, Z, Z^2, Z^3)$ of V. Since $\tilde{D}(1) = 0, \tilde{D}(Z) =$

$1, \tilde{D}(Z^2) = 2Z$ and $\tilde{D}(Z^3) = 3Z^2$, we see that

$$ {}_\beta[\tilde{D}]_\beta = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and therefore} \quad {}_\beta[\tilde{D}]_\beta - Z \cdot \mathbf{I}_4 = \begin{bmatrix} -Z & 1 & 0 & 0 \\ 0 & -Z & 2 & 0 \\ 0 & 0 & -Z & 3 \\ 0 & 0 & 0 & -Z \end{bmatrix}. $$

We see that ${}_\beta[\tilde{D}]_\beta - Z \cdot \mathbf{I}_4$ is an upper triangular matrix (see Definition 8.1.4). This means that its determinant is simply the product of the elements on its diagonal, see Theorem 8.1.2. Therefore the characteristic polynomial of $\tilde{D}$ is $p_{\tilde{D}}(Z) = (-Z)^4 = Z^4$.

11.2 Eigenspaces

So far, we have focused mainly on how to find the eigenvalues of a matrix and a linear map. In this section, we will focus on finding all eigenvectors for a given eigenvalue.

Theorem 11.2.1

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of finite dimension n, and $L : V \to V$ a linear map. Suppose that $\lambda \in \mathbb{F}$ is an eigenvalue of L. Then the set

$$ E_\lambda = \{\mathbf{v} \in V \mid L(\mathbf{v}) = \lambda \cdot \mathbf{v}\} $$

is a subspace of V.

Proof. Let $\mathbf{u}, \mathbf{v} \in E_\lambda$ and $c \in \mathbb{F}$. According to Lemma 9.3.2, we can conclude that E_λ is a subspace of V, if we can show that $\mathbf{u} + c \cdot \mathbf{v} \in E_\lambda$. Now note that

$$ L(\mathbf{u} + c \cdot \mathbf{v}) = L(\mathbf{u}) + c \cdot L(\mathbf{v}) = \lambda \cdot \mathbf{u} + c \cdot \lambda \cdot \mathbf{v} = \lambda \cdot (\mathbf{u} + c \cdot \mathbf{v}). $$

Hence indeed $\mathbf{u} + c \cdot \mathbf{v} \in E_\lambda$, which is what we needed to show. □

For square matrices, this theorem has a direct consequence.

Corollary 11.2.2

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{n \times n}$ a square matrix. Suppose that $\lambda \in \mathbb{F}$ is an eigenvalue of $\mathbf{A}$. Then the set $E_\lambda = \{\mathbf{v} \in V \mid \mathbf{A} \cdot \mathbf{v} = \lambda \cdot \mathbf{v}\}$ is a subspace of $\mathbb{F}^n$.

Proof. This follows from Theorem 11.2.1 by applying it to the linear map $L_\mathbf{A} : \mathbb{F}^n \to \mathbb{F}^n$, $\mathbf{v} \mapsto \mathbf{A} \cdot \mathbf{v}$. □

For a given linear map $L : V \to V$ for a finite dimensional vector space V and an eigenvalue λ of L, the subspace E_λ is called the *eigenspace corresponding to the eigenvalue* λ of the linear map L. Similarly, for a given square matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$, the subspace E_λ is called the eigenspace corresponding to the eigenvalue λ of the matrix $\mathbf{A}$.

Now that we know that the set of all eigenvectors of a given eigenvalue λ together with the zero vector, forms a subspace E_λ, we can describe all eigenvectors for a given eigenvalue by giving a basis of this subspace E_λ. Fortunately, this turns out to be yet another application of the theory of systems of linear equations. First of all, we have:

Lemma 11.2.3

Let $L : V \to V$ be a linear map of vector spaces over a field $\mathbb{F}$ and assume that $\dim V = n$. Suppose that $\lambda \in \mathbb{F}$ is an eigenvalue of L. Then $E_\lambda = \ker(L - \lambda \cdot \mathrm{id}_V)$. Similarly, if $\mathbf{A} \in \mathbb{F}^{n\times n}$ is a matrix and $\lambda \in \mathbb{F}$ is an eigenvalue of $\mathbf{A}$, then $E_\lambda = \ker(\mathbf{A} - \lambda \cdot \mathbf{I}_n)$.

Proof. By definition, we have $\mathbf{v} \in E_\lambda$ if and only if $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$. Note that $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$ if and only if $(L - \lambda \cdot \mathrm{id}_n)(\mathbf{v}) = \mathbf{0}$, which in turn is equivalent to saying that $\mathbf{v} \in \ker(L - \lambda \cdot \mathbf{I}_n)$. The second part of the lemma involving the matrix $\mathbf{A}$ can be proved similarly. □

As we have observed before, computing vectors in the kernel of some matrix $\mathbf{B}$, is exactly the same as finding solutions to the homogeneous system of linear equations with coefficient matrix $\mathbf{B}$. Moreover, we already know how to compute a basis for the solution space of a homogeneous system of linear equations using Corollary 9.3.4 and Theorem 6.4.4. Hence, we do not need to develop new tools when computing a basis for the eigenspace E_λ of a matrix. Also when dealing with the similar problem for linear maps, we do not need any new tools: Theorem 10.4.2 implies that we can compute the kernel of a linear map $L : V \to V$ by computing the kernel of a matrix ${}_\beta[L]_\beta$ representing the linear map, where β is an ordered basis of V. This settles the computation of eigenvectors completely. Let us illustrate this in two examples.

Example 11.2.1

First, let us consider the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in \mathbb{C}^{2\times 2}.$$

We have encountered this matrix before in Example 11.1.5, but there is one important difference: in this example we work over the complex numbers $\mathbb{C}$. This was indicated by introducing the matrix as an element in $\mathbb{C}^{2\times 2}$, rather than as element of $\mathbb{R}^{2\times 2}$. First of all, we have, just as in Example 11.1.5, that

$$p_{\mathbf{A}}(Z) = \det(\mathbf{A} - Z \cdot \mathbf{I}_2) = \det\left(\begin{bmatrix} -Z & -1 \\ 1 & -Z \end{bmatrix}\right) = Z^2 + 1.$$

Since we are working over the field $\mathbb{C}$, the polynomial Z^2+1 has two roots namely i and $-i$.

Question: Find a basis for the eigenspace E_i.

Answer: We know from Lemma 11.2.3 that $E_i = \ker(\mathbf{A} - i \cdot \mathbf{I}_2)$. We have

$$\mathbf{A} - i \cdot \mathbf{I}_2 = \begin{bmatrix} -i & -1 \\ 1 & -i \end{bmatrix}.$$

To compute the kernel of this matrix, we bring it in reduced row echelon form:

$$\begin{bmatrix} -i & -1 \\ 1 & -i \end{bmatrix} \underset{R_2 \leftarrow R_2 - i \cdot R_1}{\longrightarrow} \begin{bmatrix} -i & -1 \\ 0 & 0 \end{bmatrix} \underset{R_1 \leftarrow i \cdot R_1}{\longrightarrow} \begin{bmatrix} 1 & -i \\ 0 & 0 \end{bmatrix}.$$

This means that $\mathbf{v} = (v_1, v_2) \in \ker(\mathbf{A} - i \cdot \mathbf{I}_2)$ if and only if $v_1 = i \cdot v_2$. Hence:

$$E_i = \ker(\mathbf{A} - i \cdot \mathbf{I}_2) = \left\{ c \cdot \begin{bmatrix} i \\ 1 \end{bmatrix} \mid c \in \mathbb{C} \right\}.$$

A basis of E_i is therefore given by

$$\left\{ \begin{bmatrix} i \\ 1 \end{bmatrix} \right\}.$$

This completely answers the question. In a similar way, one can show that a basis of E_{-i} is given by

$$\left\{ \begin{bmatrix} -i \\ 1 \end{bmatrix} \right\}.$$

Example 11.2.2

Let us revisit the linear map $\tilde{D} : V \to V$ introduced in Example 11.1.6. In that example V was the complex vector space of polynomials of degree at most three and $\tilde{D} : V \to V$ was defined by $p(Z) \mapsto p(Z)'$. We have already seen in Example 11.1.6 that $p_{\tilde{D}}(Z) = Z^4$. Hence $\tilde{D}$ has only one eigenvalue, namely 0.

Question: Compute a basis for the eigenspace E_0.

Answer: We know by Lemma 11.2.3 that $E_0 = \ker(\tilde{D} - 0 \cdot \mathrm{id}_V) = \ker \tilde{D}$. In order to compute a basis of $\ker \tilde{D}$, we first compute the kernel of a matrix ${}_\beta[\tilde{D}]_\beta \in \mathbb{C}^{4\times 4}$ representing $\tilde{D}$. Let us choose the ordered basis $\beta = (1, Z, Z^2, Z^3)$ of V. We have already seen in Example 11.1.6 that in that case:

$${}_\beta[\tilde{D}]_\beta = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

This matrix is already in echelon form and we can directly see that $(v_1, v_2, v_3, v_4) \in \ker_\beta[\tilde{D}]_\beta$ if and only if $v_2 = 0$ and $v_3 = 0$ and $v_4 = 0$. Therefore,

$$\ker_\beta[\tilde{D}]_\beta = \left\{ c \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \mid c \in \mathbb{C} \right\}.$$

We see that a basis for $\ker_\beta[\tilde{D}]_\beta$ is given by

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

The basis vector $(1, 0, 0, 0)$ corresponds to the polynomial $1 \cdot 1 + 0 \cdot Z + 0 \cdot Z^2 + 0 \cdot Z^3 = 1$. Hence using Theorem 10.4.2, we see that

$$E_0 = \ker \tilde{D} = \{c \cdot 1 \mid c \in \mathbb{C}\} = \mathbb{C}$$

and that a basis of E_0 is given by $\{1\}$.

Let us finish this section with a theoretical consideration about eigenvectors that will become very important later on. We start with a definition.

Definition 11.2.1

Let $\mathbb{F}$ be a field, V a finite dimensional vector space and $L : V \to V$ be a linear map. Suppose that $\lambda \in \mathbb{F}$ is an eigenvalue of L. Then we define the *algebraic multiplicity* $\text{am}(\lambda)$ of the eigenvalue λ to be the multiplicity of λ as root in the characteristic polynomial $p_L(Z)$ of L. Further, we define the *geometric multiplicity* $\text{gm}(\lambda)$ of the eigenvalue λ to be the dimension of E_λ.

Similarly for an eigenvalue $\lambda \in \mathbb{F}$ of a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$, we define $\text{am}(\lambda)$ to be the multiplicity of λ as root in the characteristic polynomial $p_{\mathbf{A}}(Z)$ of $\mathbf{A}$ and $\text{gm}(\lambda) = \dim E_\lambda$.

Example 11.2.3

In Example 11.2.1, the eigenvalue i is a root of multiplicity 1 in the characteristic polynomial $p_{\mathbf{A}}(Z) = Z^2 + 1$. Hence $\text{am}(i) = 1$. In that example, we also saw that E_i is a vector space of dimension one. Hence $\text{gm}(i) = 1$.

Example 11.2.4

In Example 11.2.2, the eigenvalue 0 is a root of multiplicity 4 in the characteristic polynomial $p_{\tilde{D}}(Z) = Z^4$. Hence $\text{am}(0) = 4$. In that example, we also saw that E_0 is a vector space of dimension one. Hence $\text{gm}(0) = 1$ in this case.

As the last example shows, the algebraic and the geometric multiplicity of an eigenvalue need not be the same. We do have the following theorem stating that $1 \leq \text{gm}(\lambda) \leq \text{am}(\lambda)$. A reader willing to accept this statement can continue to the next section.

Theorem 11.2.4

Let $\mathbb{F}$ be a field and $\lambda \in \mathbb{F}$ an eigenvalue of a linear map $L : V \to V$, with $\dim V = n < \infty$, or an eigenvalue of a square matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$. Then $1 \leq \text{gm}(\lambda) \leq \text{am}(\lambda) \leq n$.

Proof. First of all, if λ is an eigenvalue, there by definition exists at least one eigenvector. Hence $\text{gm}(\lambda) = \dim E_\lambda \geq 1$.

Now suppose that λ is an eigenvalue and let us write $s = \text{gm}(\lambda)$ for convenience. We will prove the theorem in case λ is an eigenvalue of a linear map $L : V \to V$ only, since the case of a matrix $\mathbf{A}$ follows by considering the linear map $L_\mathbf{A} : \mathbb{F}^n \to \mathbb{F}^n$. Since $\dim E_\lambda = \text{gm}(\lambda) = s$, any basis of E_λ contains precisely s vectors. Let us choose such a basis, say $\{\mathbf{v}_1, \ldots, \mathbf{v}_s\}$. Now choose vectors $\mathbf{v}_{s+1}, \ldots, \mathbf{v}_n \in V$ such that $\beta = \mathbf{v}_1, \ldots, \mathbf{v}_s, \mathbf{v}_{s+1}, \ldots, \mathbf{v}_n$ is an ordered basis of V. Since $L(\mathbf{v}_i) = \lambda \cdot \mathbf{v}_i$ for all i between 1 and s, we have

$$ {}_\beta[L]_\beta = \begin{bmatrix} \lambda \cdot \mathbf{I}_s & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}, $$

for some matrices $\mathbf{B} \in \mathbb{F}^{s \times (n-s)}$ and $\mathbf{D} \in \mathbb{F}^{(n-s) \times (n-s)}$ and where $\mathbf{0}$ denotes the $(n-s) \times s$ matrix all of whose coefficients are zero. Then

$$ {}_\beta[L]_\beta - Z \cdot \mathbf{I}_n = \begin{bmatrix} (\lambda - Z) \cdot \mathbf{I}_s & \mathbf{B} \\ \mathbf{0} & \mathbf{D} - Z \cdot \mathbf{I}_{n-s} \end{bmatrix} $$

and hence

$$ \begin{aligned} p_L(Z) &= \det({}_\beta[L]_\beta - Z \cdot \mathbf{I}_n) \\ &= \det\left(\begin{bmatrix} (\lambda - Z) \cdot \mathbf{I}_s & \mathbf{B} \\ \mathbf{0} & \mathbf{D} - Z \cdot \mathbf{I}_{n-s} \end{bmatrix} \right) \\ &= (\lambda - Z)^s \cdot \det(\mathbf{D} - Z \cdot \mathbf{I}_{n-s}). \end{aligned} $$

In the last equality, we used induction on s and developed the determinant in the first column to prove the induction basis as well as to perform the induction step. Now it is clear that the multiplicity of λ in $p_L(Z)$ is at least s. In other words: $\text{am}(\lambda) \geq s = \text{gm}(\lambda)$, which is exactly what we wanted to show. The final inequality $\text{am}(\lambda) \leq n$ follows, since $\text{am}(\lambda)$ is the multiplicity of the root λ in the polynomial $p_L(Z)$ and $\deg p_L(Z) = n$. □

11.3 Diagonalization

In this section, we describe when a linear map can be represented by a particularly nice matrix: a diagonal matrix. In other words: we will describe when a linear map has a diagonal

mapping matrix. To achieve this, we need to be able to choose a particularly nice ordered basis. Therefore we start with a lemma.

Lemma 11.3.1

Let $\mathbb{F}$ be a field, V a finite dimensional vector space over $\mathbb{F}$ and $L : V \to V$ a linear map. Further, suppose that $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$ are distinct eigenvalues of L and write $d_i = \text{gm}(\lambda_i)$ for $i = 1, \ldots, r$. If $(\mathbf{v}_1^{(i)}, \ldots, \mathbf{v}_{d_i}^{(i)})$ for $i = 1, \ldots, r$ are ordered bases of E_{λ_i}, then the vectors

$$\mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{d_1}^{(1)}, \ldots, \mathbf{v}_1^{(r)}, \ldots, \mathbf{v}_{d_r}^{(r)}$$

are linearly independent.

Proof. We will prove the lemma using induction on r.

If $r = 1$, there is nothing to prove, since we assume that $(\mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{d_1}^{(1)})$ is an ordered basis of E_{λ_1}. Then the vectors $\mathbf{v}_1^{(1)}, \ldots, \mathbf{v}_{d_1}^{(1)}$ are certainly linearly independent. Now let $r > 1$ and assume as induction hypothesis that the lemma is correct if there are $r - 1$ distinct eigenvalues. Suppose that

$$\sum_{i=1}^{r} \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \mathbf{0}, \tag{11.2}$$

for certain $\alpha_{i,j} \in \mathbb{F}$. We need to show that $\alpha_{i,j} = 0$ for all $i = 1, \ldots, r$ and $j = 1, \ldots, d_i$. Applying the linear map L to this equation and using that $L(\mathbf{v}_j^{(i)}) = \lambda_i \cdot \mathbf{v}_j^{(i)}$, we see that $\sum_{i=1}^{r} \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \lambda_i \cdot \mathbf{v}_j^{(i)} = \mathbf{0}$, which can be rewritten as

$$\sum_{i=1}^{r} \lambda_i \cdot \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \mathbf{0}. \tag{11.3}$$

Multiplying Equation (11.2) with λ_r and subtracting Equation (11.3) from the result, the term corresponding to $i = r$ cancels, while the result still equals $\mathbf{0}$. In other words:

$$\lambda_r \cdot \sum_{i=1}^{r-1} \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} - \sum_{i=1}^{r-1} \lambda_i \cdot \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \lambda_r \cdot \sum_{i=1}^{r} \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} - \sum_{i=1}^{r} \lambda_i \cdot \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \mathbf{0}.$$

Combining the first two sums into one, we obtain:

$$\sum_{i=1}^{r-1} \sum_{j=1}^{d_i} (\lambda_r - \lambda_i) \cdot \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \sum_{i=1}^{r-1} (\lambda_r - \lambda_i) \cdot \sum_{j=1}^{d_i} \alpha_{i,j} \cdot \mathbf{v}_j^{(i)} = \mathbf{0}.$$

Now we can apply the induction hypothesis and conclude that $(\lambda_r - \lambda_i) \cdot \alpha_{i,j} = 0$ for $i = 1, \ldots, r-1$ and $j = 1, \ldots, d_i$. Since all eigenvalues were assumed to be distinct, we see that $\lambda_r - \lambda_i \neq 0$ for all i between 1 and $r - 1$. Hence $\alpha_{i,j} = 0$ for $i = 1, \ldots, r-1$ and $j = 1, \ldots, d_i$. Substituting this in Equation (11.2), we obtain that $\sum_{j=1}^{d_r} \alpha_{r,j} \cdot \mathbf{v}_j^{(r)} = \mathbf{0}$, but then

we may also conclude that $\alpha_{r,j} = 0$ for $j = 1, \dots d_r$, since we assumed that $(\mathbf{v}_1^{(r)}, \dots, \mathbf{v}_{d_r}^{(r)})$ is an ordered basis of E_{λ_r}. This completes the induction step. Hence by the induction principle, we may conclude that the lemma holds for all r. □

As we have seen before, in order to be able to represent a linear map $L : V \to V$ by a matrix ${}_\beta[L]_\beta$, we need to choose an ordered basis β of V. The vectors in Lemma 11.3.1 are linearly independent, which is a good start, but may not span the entire space V. The next lemma clarifies when the eigenvectors span V.

Lemma 11.3.2

Let $\mathbb{F}$ be a field, V a vector space over $\mathbb{F}$ of dimension n, and $L : V \to V$ a linear map. Then the following two items are equivalent:

(i) The eigenvectors of L span V.

(ii) The characteristic polynomial of L is of the form

$$p_L(Z) = (-1)^n \cdot (Z - \lambda_1)^{m_1} \cdots (Z - \lambda_r)^{m_r}$$

for certain $\lambda_1, \dots, \lambda_r \in \mathbb{F}$ and positive integers $m_1, \dots, m_r$. Moreover, for each eigenvalue λ_i its algebraic and geometric multiplicity is the same: $\mathrm{am}(\lambda_i) = \mathrm{gm}(\lambda_i)$ for $i = 1, \dots, r$.

Proof. To show that the two items are logically equivalent, we first show (i) ⇒ (ii) and afterwards (ii) ⇒ (i)

(i) ⇒ (ii): assume that the eigenvectors of L span V. Then we can find a basis S of V consisting of eigenvectors only. Let $\lambda_1, \dots, \lambda_r \in \mathbb{F}$ be the eigenvalues of L and order the eigenvectors in S such that the eigenvectors with eigenvalue λ_1 come first, then those with eigenvalue λ_2, and so on, ending with the eigenvectors in S with eigenvalue λ_r. We then have constructed an ordered basis

$$\beta = (\mathbf{v}_1^{(1)}, \dots, \mathbf{v}_{n_1}^{(1)}, \dots, \mathbf{v}_1^{(r)}, \dots, \mathbf{v}_{n_r}^{(r)}),$$

where for $i = 1, \dots, r$, the vectors $\mathbf{v}_1^{(i)}, \dots, \mathbf{v}_{n_i}^{(i)}$ are the eigenvectors in S with eigenvalue λ_i.

Now on the one hand, we have $n_1 + n_2 + \cdots + n_r = n$, since the number of vectors in the ordered basis β is the same as the dimension of V. On the other hand, for all i, we have $n_i \leq \mathrm{gm}(\lambda_i)$, since $\mathbf{v}_1^{(i)}, \dots, \mathbf{v}_{n_i}^{(i)}$ are linearly independent vectors in E_{λ_i} and $\dim E_{\lambda_i} = \mathrm{gm}(\lambda_i)$.

Therefore, we have:

$$\begin{aligned} n &= n_1 + \cdots + n_r \\ &\leq \operatorname{gm}(\lambda_1) + \cdots + \operatorname{gm}(\lambda_r) \\ &\leq \operatorname{am}(\lambda_1) + \cdots + \operatorname{am}(\lambda_r) \\ &\leq \deg p_L(Z) \\ &= n. \end{aligned}$$

Since we both started and ended with n, all inequalities have to be equalities. This shows that $\operatorname{gm}(\lambda_i) = \operatorname{am}(\lambda_i)$ for all $i = 1, \ldots, r$ and that $p_L(Z)$ is of the form as stated in item 2.

(ii) $\Rightarrow$ (i): Now assume that $p_L(Z) = (-1)^n \cdot (Z - \lambda_1)^{m_1} \cdots (Z - \lambda_r)^{m_r}$ for certain distinct $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$ and positive integers $m_1, \ldots, m_r$ and that $\operatorname{am}(\lambda_i) = \operatorname{gm}(\lambda_i)$ for all $i = 1, \ldots, r$. Note that by definition, we have $m_i = \operatorname{am}(\lambda_i)$, which in turn implies that $m_i = \operatorname{gm}(\lambda_i)$, since we assume that $\operatorname{am}(\lambda_i) = \operatorname{gm}(\lambda_i)$ for all i. We conclude that $n = \deg p_L(Z) = \operatorname{gm}(\lambda_1) + \cdots + \operatorname{gm}(\lambda_r)$. On the other hand, by Lemma 11.3.1, we can find precisely $\operatorname{gm}(\lambda_1) + \cdots + \operatorname{gm}(\lambda_r)$ linearly independent eigenvectors of L. Combining these statements, we can conclude that we can find an ordered basis of V consisting of eigenvectors. In particular, the eigenvectors span V, which is what we wanted to show. □

Now we are ready to show the main result of this section.

Definition 11.3.1

Let a linear map $L : V \to V$ be given, where V is a finite dimensional vector space over a field $\mathbb{F}$. Then one says that L can be *diagonalized*, if there exists an ordered basis β of V such that the corresponding mapping matrix ${}_\beta[L]_\beta$ is a diagonal matrix. Likewise, if $\mathbf{A} \in \mathbb{F}^{n \times n}$ is a square matrix, then one says that $\mathbf{A}$ can be diagonalized, if $\mathbf{A}$ is similar to a diagonal matrix.

Theorem 11.3.3

Let V a finite dimensional vector space over a field $\mathbb{F}$. A linear map $L : V \to V$ can be diagonalized if and only if the characteristic polynomial of L is of the form $p_L(Z) = (-1)^n \cdot (Z - \lambda_1)^{m_1} \cdots (Z - \lambda_r)^{m_r}$ for certain $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$, and $\operatorname{am}(\lambda_i) = \operatorname{gm}(\lambda_i)$ for each eigenvalue λ_i.

Proof. Using Lemma 11.3.2, it is enough to show that L can be diagonalized if and only if its eigenvectors span V.

Therefore, first assume that L can be diagonalized. Then there exists an ordered basis β of V such that ${}_\beta[L]_\beta$ is a diagonal matrix. But this implies that each vector in β is an eigenvector. Hence V can be spanned by eigenvectors.

Conversely, assume that V can be spanned by eigenvectors. Then there exists an ordered basis $\beta = \mathbf{v}_1, \ldots, \mathbf{v}_n$ of V, containing eigenvectors only. The corresponding matrix ${}_\beta[L]_\beta$ is a diagonal matrix, with the eigenvalues of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ on its diagonal. □

Corollary 11.3.4

A matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ can be diagonalized if and only if the characteristic polynomial of $\mathbf{A}$ is of the form $p_\mathbf{A}(Z) = (-1)^n \cdot (Z - \lambda_1)^{m_1} \cdots (Z - \lambda_r)^{m_r}$ for certain $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$, and $\mathrm{am}(\lambda_i) = \mathrm{gm}(\lambda_i)$ for each eigenvalue λ_i.

Proof. This follows from Theorem 11.3.3 by applying it to the linear map $L_\mathbf{A} : \mathbb{F}^n \to \mathbb{F}^n$. □

Corollary 11.3.5

Let V be a finite dimensional complex vector space. A linear map $L : V \to V$, can be diagonalized if and only if $\mathrm{am}(\lambda_i) = \mathrm{gm}(\lambda_i)$ for each eigenvalue λ_i of L. Similarly, a complex matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is similar to a diagonal matrix if and only if $\mathrm{am}(\lambda_i) = \mathrm{gm}(\lambda_i)$ for each eigenvalue λ_i of $\mathbf{A}$.

Proof. If the field we work over is $\mathbb{C}$, it follows from Theorem 4.6.3 that the characteristic polynomial $p_L(Z)$ can be written as a product of its leading coefficient and terms of the form $Z - \lambda$. Hence this condition in Theorem 11.3.3 is always satisfied if $\mathbb{F} = \mathbb{C}$ and can therefore be removed. Theorem 11.3.3 then implies what we want. The proof of the corollary in the case of a complex matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is similar. □

Example 11.3.1

Consider the linear map $\tilde{D} : V \to V$ introduced in Example 11.1.6. In that example V was the complex vector space of polynomials of degree at most three and $\tilde{D} : V \to V$ was defined by $p(Z) \mapsto p(Z)'$.

Question: Can the linear map $\tilde{D}$ be diagonalized?

Answer: From Example 11.1.6, we see that $p_{\tilde{D}}(Z) = Z^4$. Using the notation of Theorem 11.3.3, we see that $r = 1$ and $\lambda_1 = 0$. Moreover, $\mathrm{am}(0) = 4$, since 0 is a root with multiplicity four of $p_{\tilde{D}}(Z)$. In Example 11.2.2, we have seen that E_0 is a one dimensional vector space with basis $\{1\}$. Hence $\mathrm{gm}(0) = \dim E_0 = 1$. Since $\mathrm{gm}(0) < \mathrm{am}(0)$ Theorem 11.3.3 implies that the linear map $\tilde{D}$ cannot be diagonalized.

Example 11.3.2

Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

that we also studied in Example 11.2.1 and that also occurred in Example 11.1.5.

Question 1: Can the matrix $\mathbf{A}$ be diagonalized when working over the real numbers $\mathbb{R}$? If yes, compute a matrix $\mathbf{Q} \in \mathbb{R}^{2\times 2}$ such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix.

Question 2: Can the matrix $\mathbf{A}$ be diagonalized when working over the complex numbers $\mathbb{C}$? If yes, compute a matrix $\mathbf{Q} \in \mathbb{C}^{2\times 2}$ such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix.

Answer to Question 1: We have computed in Example 11.1.5, that $p_{\mathbf{A}}(Z) = Z^2 + 1$. Since $Z^2 + 1$ has no real roots, it cannot be written in the form as required in Corollary 11.3.4. Therefore, the matrix $\mathbf{A}$ is not diagonalizable over $\mathbb{R}$.

Answer to Question 2: The characteristic polynomial $p_{\mathbf{A}}(Z) = Z^2 + 1$ has two complex roots, namely i and $-i$. Furthermore, we have $Z^2 + 1 = (Z - i) \cdot (Z + i)$. Hence $\text{am}(i) = 1$ and $\text{am}(-i) = 1$. Since by Theorem 11.2.4, we know that $1 \leq \text{gm}(\lambda) \leq \text{am}(\lambda)$ for any eigenvalue λ, we conclude that $\text{gm}(i) = \text{am}(i) = 1$ and $\text{gm}(-i) = \text{am}(-i) = 1$. Hence all in Corollary 11.3.4 are satisfied. We conclude that the given matrix $\mathbf{A}$ is diagonalizable over the complex numbers.

Now we explicitly compute an invertible matrix $\mathbf{Q} \in \mathbb{C}^{2\times 2}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix. Let us denote by ϵ the standard basis of $\mathbb{C}^2$. Then ${}_\epsilon[L_{\mathbf{A}}]_\epsilon = \mathbf{A}$. To diagonalize $\mathbf{A}$, we simply diagonalize the corresponding linear map $L_{\mathbf{A}}$. In order to do that, we need to find an ordered basis of $\mathbb{C}^2$ consisting of eigenvectors. In Example 11.2.1, we saw that:

$$E_i \quad \text{has basis} \quad \left\{ \begin{bmatrix} i \\ 1 \end{bmatrix} \right\} \qquad \text{and} \qquad E_{-i} \quad \text{has basis} \quad \left\{ \begin{bmatrix} -i \\ 1 \end{bmatrix} \right\}.$$

Hence $\beta = \left(\begin{bmatrix} i \\ 1 \end{bmatrix}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \right)$ is an ordered basis of $\mathbb{C}^2$ consisting of eigenvectors only. Using this ordered basis, we find that the mapping matrix ${}_\beta[L_{\mathbf{A}}]_\beta$ is a diagonal matrix with the eigenvalues of the vectors in the ordered basis β on its diagonal. Hence

$${}_\beta[L_{\mathbf{A}}]_\beta = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}.$$

To find the matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix, now observe that

$${}_\beta[L_{\mathbf{A}}]_\beta = {}_\beta[\text{id}_{\mathbb{C}^2}]_\epsilon \cdot {}_\epsilon[L_{\mathbf{A}}]_\epsilon \cdot {}_\epsilon[\text{id}_{\mathbb{C}^2}]_\beta = {}_\epsilon[\text{id}_{\mathbb{C}^2}]_\beta^{-1} \cdot \mathbf{A} \cdot {}_\epsilon[\text{id}_{\mathbb{C}^2}]_\beta.$$

Hence we can simply choose $\mathbf{Q} = {}_\epsilon[\text{id}_{\mathbb{C}^2}]_\beta$, the change of coordinate matrix from β-coordinates to ϵ-coordinates. This matrix contains the eigenvectors in β as columns. Hence

$$\mathbf{Q} = {}_\epsilon[\text{id}_{\mathbb{C}^2}]_\beta = \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$$

is the matrix we were looking for. Concretely, we have

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

11.4 Fibonacci numbers revisited

In Example 5.1.2, more precisely in Equation (5.3), we gave an example of a recursively defined sequence of numbers $F_1, F_2, F_3, \ldots$ called the Fibonacci numbers:

$$F_n = \begin{cases} 1 & \text{if } n = 1, \\ 1 & \text{if } n = 2, \\ F_{n-1} + F_{n-2} & \text{if } n \geq 3. \end{cases} \tag{11.4}$$

This recursion can also be expressed using matrices. Indeed, directly from Equation (11.4), one sees that

$$\begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} F_{n-1} \\ F_{n-2} \end{bmatrix} \quad \text{for all } n \geq 3.$$

This matrix form makes it possible to find a closed formula for the Fibonacci numbers. First of all, we have the following:

Lemma 11.4.1

For all $n \geq 2$ it holds that:

$$\begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^{n-2} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Proof. This can be shown using induction on n with base case $n = 2$. The details are left to the reader. □

To find a closed formula for F_n, it is enough to find a closed formula for powers of the matrix

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

We will diagonalize $\mathbf{P}$ to do this. The point is that if a matrix can be diagonalized, it is possible to find a closed formula for its powers:

Lemma 11.4.2

Let $\mathbb{F}$ be a field and $\mathbf{A} \in \mathbb{F}^{n \times n}$ a square matrix. Let $\mathbf{Q} \in \mathbb{F}^{n \times n}$ be an invertible matrix such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix $\mathbf{D}$ with the elements $d_1, \ldots, d_n$ on its diagonal. Then

$$\mathbf{A}^n = \mathbf{Q} \cdot \mathbf{D}^n \cdot \mathbf{Q}^{-1}.$$

Moreover, $\mathbf{D}^n$ is a diagonal matrix with the elements $d_1^n, \ldots, d_n^n$ on its diagonal.

Proof. With induction on n one can show that $(\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q})^n = \mathbf{Q}^{-1} \cdot \mathbf{A}^n \cdot \mathbf{Q}$ for all $n \geq 1$. Since $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q} = \mathbf{D}$, the result then follows. Also showing that $\mathbf{D}^n$ is a diagonal matrix with the elements $d_1^n, \ldots, d_n^n$ on its diagonal, can readily be shown by induction on n. □

The point of this lemma is that it makes the computation of powers of a matrix relatively easy if the matrix is diagonalizable. Now let us return to the matrix $\mathbf{P}$. The characteristic polynomial of $\mathbf{P}$ is

$$p_{\mathbf{P}}(Z) = \det\left(\begin{bmatrix} 1-Z & 1 \\ 1 & -Z \end{bmatrix}\right) = Z^2 - Z - 1.$$

Hence the eigenvalues of $\mathbf{P}$ are $\lambda_1 = \frac{1+\sqrt{5}}{2}$ and $\lambda_2 = \frac{1-\sqrt{5}}{2}$. This already means that the matrix $\mathbf{P}$ is diagonalizable. To find the desired change of coordinate matrix, we need to calculate a basis of the eigenspaces. To calculate a basis of the eigenspace E_{λ_1}, note that

$$\mathbf{P} - \lambda_1 \cdot \mathbf{I}_2 = \begin{bmatrix} \frac{1-\sqrt{5}}{2} & 1 \\ 1 & \frac{-1-\sqrt{5}}{2} \end{bmatrix} \begin{array}{c} \longrightarrow \\ R_2 \leftarrow R_2 - \lambda_2 \cdot R_1 \end{array} \begin{bmatrix} \frac{1-\sqrt{5}}{2} & 1 \\ 0 & 0 \end{bmatrix}$$

Hence we see that a basis of E_{λ_1} is given by

$$\left\{ \begin{bmatrix} -1 \\ \frac{1-\sqrt{5}}{2} \end{bmatrix} \right\}.$$

Similarly, one can show that a basis of E_{λ_2} is given by

$$\left\{ \begin{bmatrix} -1 \\ \frac{1+\sqrt{5}}{2} \end{bmatrix} \right\}.$$

Hence

$$\begin{bmatrix} \frac{1+\sqrt{5}}{2} & 0 \\ 0 & \frac{1-\sqrt{5}}{2} \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix},$$

which implies that

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix} \cdot \begin{bmatrix} \frac{1+\sqrt{5}}{2} & 0 \\ 0 & \frac{1-\sqrt{5}}{2} \end{bmatrix} \cdot \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix}^{-1}$$

Now applying Lemma 11.4.2 in order to compute powers of $\mathbf{P}$ and Lemma 11.4.1, we see that

$$\begin{aligned} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^{n-2} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix} \cdot \begin{bmatrix} \left(\frac{1+\sqrt{5}}{2}\right)^{n-2} & 0 \\ 0 & \left(\frac{1-\sqrt{5}}{2}\right)^{n-2} \end{bmatrix} \cdot \begin{bmatrix} -1 & -1 \\ \frac{1-\sqrt{5}}{2} & \frac{1+\sqrt{5}}{2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned}$$

After working out all the matrix products on the right-hand side, one obtains that

$$\begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2}\right)^n \\ \frac{1}{\sqrt{5}} \cdot \left(\frac{1+\sqrt{5}}{2}\right)^{n-1} - \frac{1}{\sqrt{5}} \cdot \left(\frac{1-\sqrt{5}}{2}\right)^{n-1} \end{bmatrix},$$

which explains where Equation (5.4) came from.

In this section we focused on the Fibonacci numbers, but very similar techniques can be used to find closed formulas for other recursively defined sequences of numbers, but we will not pursue this further here.

11.5 Extra: What if diagonalization is not possible?

This section is not required reading and can be skipped. It is meant as extra material for a student who has the time and motivation for it.

As we have seen in the previous section, diagonalization of a linear map $L : V \to V$ is not always possible. In this section, we discuss the well-known *Jordan normal form*. The key for diagonalization was to study the eigenspace $E_\lambda = \ker(L - \lambda \cdot \mathrm{id}_V)$ for a given eigenvalue λ. We defined $\mathrm{gm}(\lambda) = \dim E_\lambda$ and have seen that in order to be able to diagonalize a matrix or linear map, it was important that the condition $\mathrm{gm}(\lambda) = \mathrm{am}(\lambda)$ is met. It turns out that if $\mathrm{gm}(\lambda) < \mathrm{am}(\lambda)$, one needs to study the kernels of powers of the linear map $L - \lambda \cdot \mathrm{id}_V$. Here the i-th power f^i of a function $f : V \to V$ should be understood as the i-fold composite of f with itself (so $f^1 = f$, $f^2 = f \circ f$, etcetera).

Lemma 11.5.1

Let $\mathbb{F}$ be a field, V an n-dimensional vector space over $\mathbb{F}$ and $L : V \to V$ a linear map. Further assume that $\lambda \in \mathbb{F}$ is an eigenvalue of L. Then

$$\ker(L - \lambda \cdot \mathrm{id}_V) \subseteq \ker((L - \lambda \cdot \mathrm{id}_V)^2) \subseteq \ker((L - \lambda \cdot \mathrm{id}_V)^3) \subseteq \dots$$

Moreover,

(i) if $\ker((L - \lambda \cdot \mathrm{id}_V)^i) = \ker((L - \lambda \cdot \mathrm{id}_V)^{i+1})$ for some positive integer i, then $\ker((L - \lambda \cdot \mathrm{id}_V)^i) = \ker((L - \lambda \cdot \mathrm{id}_V)^m)$ for all $m \geq i$, and

(ii) $\ker((L - \lambda \cdot \mathrm{id}_V)^n) = \ker((L - \lambda \cdot \mathrm{id}_V)^{n+1})$.

Proof. It is clear that for all i, the kernel of $(L - \lambda \cdot \mathrm{id}_V)^{i+1}$ is a subspace of the kernel of $(L - \lambda \cdot \mathrm{id}_V)^i$. If equality holds for some i, then by the rank-nullity theorem for linear maps, see Corollary 10.4.3, we also obtain that the images of the linear maps $(L - \lambda \cdot \mathrm{id}_V)^{i+1}$ and

$(L - \lambda \cdot \mathrm{id}_V)^i$ are the same. But then also

$$\begin{aligned}
\mathrm{im}(L - \lambda \cdot \mathrm{id}_V)^{i+2} &= (L - \lambda \cdot \mathrm{id}_V)^{i+2}(V) \\
&= (L - \lambda \cdot \mathrm{id}_V)((L - \lambda \cdot \mathrm{id}_V)^{i+1}(V)) \\
&= (L - \lambda \cdot \mathrm{id}_V)((L - \lambda \cdot \mathrm{id}_V)^{i}(V)) \\
&= (L - \lambda \cdot \mathrm{id}_V)^{i+1}(V) \\
&= \mathrm{im}(L - \lambda \cdot \mathrm{id}_V)^{i+1} \\
&= \mathrm{im}(L - \lambda \cdot \mathrm{id}_V)^{i}.
\end{aligned}$$

But then, again using the rank-nullity theorem for linear maps, we see that the kernel of $(L - \lambda \cdot \mathrm{id}_V)^{i+2}$ is equal to the kernel of $(L - \lambda \cdot \mathrm{id}_V)^i$. Using induction on m, one can similarly show that for any $m \geq i$ the kernel of $(L - \lambda \cdot \mathrm{id}_V)^m$ is equal to the kernel of $(L - \lambda \cdot \mathrm{id}_V)^i$.

Now consider the sequence of subspaces:

$$\ker(L - \lambda \cdot \mathrm{id}_V) \subseteq \ker((L - \lambda \cdot \mathrm{id}_V)^2) \subseteq \ker((L - \lambda \cdot \mathrm{id}_V)^3) \subseteq \ldots$$

By the previous, we know that if equality holds at some point in the sequence, then equalities will hold from then on. Therefore there exists $e \geq 1$ such that

$$\ker(L - \lambda \cdot \mathrm{id}_V) \subsetneq \cdots \subsetneq \ker((L - \lambda \cdot \mathrm{id}_V)^e) = \ker((L - \lambda \cdot \mathrm{id}_V)^{e+1}) = \ldots$$

For every strict inclusion, the dimension of the subspace increases by at least one. Since $\dim V = n$ and $\dim \ker(L - \lambda \cdot \mathrm{id}_V) = \dim E_\lambda \geq 1$, this can occur at most n times. Hence $e \leq n$. In particular $\ker((L - \lambda \cdot \mathrm{id}_V)^n) = \ker((L - \lambda \cdot \mathrm{id}_V)^{n+1})$. □

Theorem 11.5.2

Let $\mathbb{F}$ be a field, V an n-dimensional vector space over $\mathbb{F}$ and $L : V \to V$ a linear map. Further assume that $\lambda \in \mathbb{F}$ is an eigenvalue of L. Further, write $U = \mathrm{im}(L - \lambda \cdot \mathrm{id}_V)^n$ and $W = \ker(L - \lambda \cdot \mathrm{id}_V)^n$. Then

(i) $L(U) \subseteq U$ and $L(W) \subseteq W$.

(ii) $\dim U + \dim W = \dim V$ and $U \cap W = \{\mathbf{0}\}$.

(iii) Any vector in V can be written as the sum of a vector in U and a vector in W.

Proof. We have seen in the third item of Lemma 11.5.1 that $W = \ker(L - \lambda \cdot \mathrm{id}_V)^n = \ker(L - \lambda \cdot \mathrm{id}_V)^{n+1}$. Now choose $\mathbf{w} \in W$. Then also $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{w}) \in W$, since $(L - \lambda \cdot \mathrm{id}_V)^n((L - \lambda \cdot \mathrm{id}_V)(\mathbf{w})) = (L - \lambda \cdot \mathrm{id}_V)((L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{w})) = (L - \lambda \cdot \mathrm{id}_V)(\mathbf{0}) = \mathbf{0}$. Hence $L(\mathbf{w}) - \lambda \cdot \mathbf{w} \in W$, which implies that $L(\mathbf{w}) \in W$. We may conclude that $L(W) \subseteq W$. Similarly, if $\mathbf{u} \in U$, then $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{u}) \in U$, since if $\mathbf{u} = (L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{v})$ for some $\mathbf{v} \in V$, then $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{u}) = (L - \lambda \cdot \mathrm{id}_V)((L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{v})) = (L - \lambda \cdot \mathrm{id}_V)^n((L - \lambda \cdot \mathrm{id}_V)(\mathbf{v})) \in U$. Hence $L(\mathbf{u}) - \lambda \cdot \mathbf{u} \in U$, which implies that $L(\mathbf{u}) \in U$. We may conclude that $L(U) \subseteq U$.

The rank-nullity theorem for linear maps applied to the linear map $(L - \lambda \cdot \mathrm{id}_V)^n : V \to V$ immediately implies that $\dim U + \dim W = \dim V$. Now, we prove that $U \cap W = \{\mathbf{0}\}$. Let $\mathbf{u} \in U \cap W$. We wish to show that $\mathbf{u} = \mathbf{0}$. First of all, since $\mathbf{u} \in U$, there exists $\mathbf{v} \in V$ such that $\mathbf{u} = (L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{v})$. Second, since $\mathbf{u} \in W$, we have $(L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{u}) = \mathbf{0}$. Combining these two, we see that

$$(L - \lambda \cdot \mathrm{id}_V)^{2n}(\mathbf{v}) = (L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{u}) = \mathbf{0}.$$

In other words, $\mathbf{v} \in \ker(L - \lambda \cdot \mathrm{id}_V)^{2n}$. However, Lemma 11.5.1 implies that $\ker(L - \lambda \cdot \mathrm{id}_V)^{2n} = \ker(L - \lambda \cdot \mathrm{id}_V)^n$ and hence $\mathbf{v} \in \ker(L - \lambda \cdot \mathrm{id}_V)^n$. But then $\mathbf{u} = (L - \lambda \cdot \mathrm{id}_V)^n(\mathbf{v}) = \mathbf{0}$, which is what we wanted to show.

Given an ordered basis $\beta_U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ of U and an ordered basis $\beta_W = (\mathbf{w}_1, \ldots, \mathbf{w}_s)$ of W, joining the two together yields an ordered basis $\beta = (\beta_U, \beta_W)$ of V. Indeed, the fact that $U \cap W = \{\mathbf{0}\}$ can be used to show that the vectors in β are linearly independent, while the identity $\dim U + \dim W = \dim V$ implies that β contains exactly n vectors. Now given an arbitrarily chosen $\mathbf{v} \in V$, we can write $\mathbf{v}$ in exactly one way as a linear combination of the $\mathbf{u}_i$ and the $\mathbf{w}_j$, say $\mathbf{v} = \sum_i \alpha_i \cdot \mathbf{u}_i + \sum_j \beta_j \cdot \mathbf{w}_j$. Now observing that $\sum_i \alpha_i \cdot \mathbf{u}_i \in U$ and $\sum_j \beta_j \cdot \mathbf{w}_j \in W$, the last item in the theorem follows. □

Corollary 11.5.3

Using the same notation as in Theorem 11.5.2, write $p_L(Z) = (\lambda - Z)^{\mathrm{am}(\lambda)} \cdot q(Z)$ for a suitably chosen $q(Z) \in \mathbb{F}[Z]$. Denote by $L|_W : W \to W$, respectively $L|_U : U \to U$, the linear maps obtained by restricting the domain and codomain of L to U, respectively W. Then $p_L(Z) = p_{L|_U}(Z) \cdot p_{L|_W}(Z)$ and λ is not a root of $p_{L|_U}(Z)$.

Proof. Given an ordered basis $\beta_U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ of U and an ordered basis $\beta_W = (\mathbf{w}_1, \ldots, \mathbf{w}_s)$ of W, we have already seen in the proof of Theorem 11.5.2, that $\beta = (\beta_U, \beta_W)$ is an ordered basis of V. Since we know that $L(U) \subseteq U$ and $L(W) \subseteq W$, the matrix ${}_\beta[L]_\beta \in \mathbb{F}^{n \times n}$ will have the form

$${}_\beta[L]_\beta = \begin{bmatrix} {}_{\beta_U}[L]_{\beta_U} & \mathbf{0} \\ \mathbf{0} & {}_{\beta_W}[L]_{\beta_W} \end{bmatrix}. \tag{11.5}$$

This implies that $p_L(Z) = p_{L|_U}(Z) \cdot p_{L|_W}(Z)$. Now observe that λ cannot be a root of $p_{L|_U}(Z)$. Indeed, if this would be the case, then there would exist a nonzero $\mathbf{u} \in U$ such that $L|_U(\mathbf{u}) = \lambda \cdot \mathbf{u}$. Since by definition of the linear map $L|_U$, we have $L|_U(\mathbf{u}) = L(\mathbf{u})$, this would imply that $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{u}) = \mathbf{0}$. But then $\mathbf{u} \in \ker(L - \lambda \cdot \mathrm{id}_V)$, implying that $\mathbf{u} \in \ker(L - \lambda \cdot \mathrm{id}_V)^n = W$. Since we have seen that $U \cap W = \{\mathbf{0}\}$, we would obtain that $\mathbf{u} = \mathbf{0}$, contrary to our assumption. □

Now let us return to what we are trying to achieve: to find a matrix representing L that is as simple as possible. Equation (11.5) is an important step on the way. Indeed, we have reduced the problem into two simpler ones: finding a simple matrix representing $L|_U$ and one

representing $L|_W$. Moreover, λ is not a root of the characteristic polynomial of $L|_U$, so to deal with the eigenvalue λ of L, we only have to continue with the study of the linear map $L|_W$. Let us first get an intuitive idea of what may be going on. If $\text{am}(\lambda) = \text{gm}(\lambda)$, we can find an ordered basis β_W of W consisting of eigenvectors for L only, all having λ as eigenvalue. Then

$$_{\beta_W}[L]_{\beta_W} = [\lambda \cdot \mathbf{I}_s],$$

where $s = \dim W$. What we did in the previous section is essentially to repeat this procedure for another eigenvalue and split the matrix representing $L|_U$ further up in smaller blocks. As long as the algebraic and geometric multiplicity of the eigenvalues is always the same, we end up with diagonalizing the entire matrix.

So what happens if $\text{am}(\lambda) > \text{gm}(\lambda)$? We can still find an ordered basis β_W of W containing the eigenvectors with eigenvalue λ, but we also need some more vectors in β_W that are not eigenvectors. To put this in a different way: if $\text{am}(\lambda) = \text{gm}(\lambda)$, then $W = E_\lambda$, so that $\ker(L - \lambda \cdot \text{id}_V) = \ker((L - \lambda \cdot \text{id}_V)^n)$. However, if $\text{am}(\lambda) > \text{gm}(\lambda)$, then W contains E_λ, but is not equal to it. Then apparently $\ker(L - \lambda \cdot \text{id}_V) \subsetneq \ker((L - \lambda \cdot \text{id}_V)^n)$. This implies in particular that $\ker(L - \lambda \cdot \text{id}_V) \subsetneq \ker((L - \lambda \cdot \text{id}_V)^2)$ using Lemma 11.5.1. If we choose a vector $\mathbf{w} \in \ker((L - \lambda \cdot \text{id}_V)^2) \setminus \ker(L - \lambda \cdot \text{id}_V)$, it has the nice property that $L(\mathbf{w}) - \lambda \cdot \mathbf{w} = (L - \lambda \cdot \text{id}_V)(\mathbf{w}) \in \ker(L - \lambda \cdot \text{id}_V) = E_\lambda$. Let us define $\mathbf{v} = L(\mathbf{w}) - \lambda \cdot \mathbf{w}$. Further, we see that: $L(\mathbf{v}) = \lambda \cdot \mathbf{v}$, since $\mathbf{v} \in E_\lambda$, and $L(\mathbf{w}) = \lambda \cdot \mathbf{w} + \mathbf{v}$, by the way we defined $\mathbf{v}$. So if we only consider the effect of L on the two-dimensional subspace of V spanned by $\mathbf{v}$ and $\mathbf{w}$, which has ordered basis $(\mathbf{v}, \mathbf{w})$, we can represent L by the matrix

$$\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$

This gives a first idea of what to expect more in general for a matrix representing $L|_W$. In particular, it motivates the following:

Definition 11.5.1

Let $\mathbb{F}$ be a field and $\lambda \in \mathbb{F}$. A *Jordan block* of size e is a matrix $\mathbf{J}_e(\lambda) \in \mathbb{F}^{e \times e}$ of the form

$$\mathbf{J}_e(\lambda) = \begin{bmatrix} \lambda & 1 & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \mathbf{0} & & & \lambda \end{bmatrix}.$$

Lemma 11.5.4

Let $L : V \to V$ be a linear map, $\dim V = n$ and λ an eigenvalue of L. Further, let $W = \ker((L - \lambda \cdot \text{id}_V)^n)$. Then there exists an ordered basis of W such that $L|_W : W \to W$,

the restriction of L to W has a mapping matrix $\mathbf{D}(\lambda)$ of the form

$$\mathbf{D}(\lambda) = \begin{bmatrix} \mathbf{J}_{e_1}(\lambda) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{J}_{e_s}(\lambda) \end{bmatrix}$$

for a certain positive integer s and certain positive integers $e_1, \ldots, e_s$.

Proof. It will be convenient to write $W_i = \ker((L - \lambda \cdot \mathrm{id}_V)^i)$ and $r_i = \dim W_i$. Note that $W_1 = E_\lambda$. Now let e be the largest exponent such that $W_{e-1} \subsetneq W_e$. Then $W_e = W$ and $r_1 < r_2 < \cdots < r_e = \dim W$. Let $\beta = (\mathbf{w}_1, \ldots, \mathbf{w}_{r_e})$ be an ordered basis of W with the additional property that $(\mathbf{w}_1, \ldots, \mathbf{w}_{r_i})$ is an ordered basis of W_i for all i.

We now gradually construct another ordered basis, say γ, of W. First of all, we add to γ the vectors $\mathbf{w}_i$ for any i between $r_{e-1}+1$ and r_e. By construction, for any i between $r_{e-1}+1$ and r_e the vector $\mathbf{w}_i$ lies in W_e, but not in W_{e-1}. Now, consider the vectors $\mathbf{w}_{i,j} = (L - \lambda \cdot \mathrm{id}_V)^j(\mathbf{w}_i)$, for $j = 0, \ldots, e-1$. Note that the vector $\mathbf{w}_{i,j}$ lies in W_{e-j}, but not in W_{e-j-1}. Also note that $\mathbf{w}_i = \mathbf{w}_{i,0}$ for any i between $r_{e-1}+1$ and r_e.

First we claim that vectors $\mathbf{w}_{i,e-1}$ are linearly independent. Indeed if $\sum_{i=r_{e-1}+1}^{r_e} \alpha_i \cdot \mathbf{w}_{i,e-1} = 0$, then $\sum_{i=r_{e-1}+1}^{r_e} \alpha_i \cdot \mathbf{w}_i \in W_{e-1}$ and hence in the span of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{r_{e-1}}$. But since the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{r_e}$ are linearly independent, this implies that $\alpha_i = 0$ for all $i = r_{e-1}+1, \ldots, r_e$. Next we claim that the vectors $\mathbf{w}_{i,j}$, with $i = r_{e-1}+1, \ldots, r_e$ and $j = 0, \ldots, e-1$ are linearly independent. If $\sum_i \sum_{j=0}^{e-1} \alpha_{i,j} \cdot \mathbf{w}_{i,j} = 0$, then applying $(L - \lambda \cdot \mathrm{id}_V)^{e-1}$ yields the equation $\sum_i \alpha_{i,0} \cdot \mathbf{w}_{i,e-1} = 0$. Hence $\alpha_{i,0} = 0$ for all i. Now applying lower and lower powers of $(L - \lambda \cdot \mathrm{id}_V)$ to the equation $\sum_i \sum_{j=0}^{e-1} \alpha_{i,j} \cdot \mathbf{w}_{i,j} = 0$, one obtains inductively that $\alpha_{i,j} = 0$ for all i and j.

Therefore, it makes sense to include all the vectors $\mathbf{w}_{i,j}$ in γ. More precisely, we now set $\gamma = (\mathbf{w}_{r_{e-1}+1,e-1}, \ldots, \mathbf{w}_{r_{e-1}+1,0}, \ldots, \mathbf{w}_{r_e,e-1}, \ldots, \mathbf{w}_{r_e,0})$. We have $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{w}_{i,j}) = \mathbf{w}_{i,j+1}$ for $j = 0, \ldots, e-2$ and $(L - \lambda \cdot \mathrm{id}_V)(\mathbf{w}_{i,e-1}) = \mathbf{0}$. This implies that

$$L(\mathbf{w}_{i,j}) = \lambda \cdot \mathbf{w}_{i,j} + \mathbf{w}_{i,j+1} \quad \text{for } j = 0, \ldots, e-2 \text{ and} \quad L(\mathbf{w}_{i,e-1}) = \lambda \cdot \mathbf{w}_{i,e-1}.$$

We have now in fact shown that the restriction of L to the subspace spanned by the $\mathbf{w}_{i,j}$ can be represented by a block diagonal matrix with $r_e - r_{e-1}$ many matrices $\mathbf{J}_e(\lambda)$ on its diagonal.

For any j between 0 and $e-1$, the vectors $\mathbf{w}_{i,j}$, with i varying from $r_{e-1}+1$ to r_e, span a subspace of W_{e-j}. If for all j, this subspace is equal to W_{e-j}, then γ is an ordered basis of W, giving rise to a matrix representing W in Jordan normal form, as we have seen. Otherwise, let $\tilde{j}$ be the smallest value of j such that this subspace is not all of W_{e-j} and define $\tilde{e} = e - \tilde{j}$. Then let $\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_d \in W_{\tilde{e}}$ be vectors such that

$$(\mathbf{w}_1, \ldots, \mathbf{w}_{r_{\tilde{e}-1}}, \tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_d, \mathbf{w}_{r_{e-1}+1,\tilde{j}}, \ldots, \mathbf{w}_{r_e,\tilde{j}})$$

is an ordered basis of $W_{\tilde{e}}$. Now we proceed similarly as in the start, defining vectors $\tilde{\mathbf{w}}_{i,j} = (L - \lambda \cdot \mathrm{id}_V)^j(\tilde{\mathbf{w}}_i)$, which we add to γ and which will give rise to Jordan blocks of size $\tilde{e}$ in the matrix representing L.

Continuing in this way, we end up with an ordered basis of γ giving rise to a matrix in Jordan normal form that represents the restriction of L to W. □

Theorem 11.5.5

Let $\mathbb{F}$ be a field, V a finite dimensional vector space, and $L : V \to V$. Suppose that there exist distinct $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$ and positive integers $m_1, \ldots, m_r$ such that $p_L(Z) = (-1)^n \cdot (Z - \lambda_1)^{m_1} \cdots (Z - \lambda_r)^{m_r}$. Then the linear map can be represented by a matrix of the form

$$\begin{bmatrix} \mathbf{D}(\lambda_1) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}(\lambda_r) \end{bmatrix},$$

where each matrix $\mathbf{D}(\lambda_i) \in \mathbb{F}^{m_i \times m_i}$ is of the form as in Lemma 11.5.4.

Proof. We prove the theorem with induction of the number of eigenvalues. We use the same notation for U and W as in Theorem 11.5.2. If $r = 1$, $W = V$ and hence $L|_W = L$. Hence the result follows from Lemma 11.5.4. Now let $r > 1$. Let $\lambda = \lambda_r \in \mathbb{F}$ be an eigenvalue of L. From Lemma 11.5.4, we conclude that we can choose an ordered basis for W such that $L|_W$ is represented by a block diagonal matrix with Jordan blocks $\mathbf{J}_{e_i}(\lambda_r)$ on its diagonal. Further, from Corollary 11.5.3, λ_r is not an eigenvalue of $L|_U$, while the characteristic polynomial of $L|_U$ is a divisor of $P_L(Z)$. Hence the induction hypothesis applies. □

The matrix given in Theorem 11.5.5 is said to be the *Jordan normal form* of the matrix $\mathbf{A}$.

Corollary 11.5.6

Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a complex matrix. Then $\mathbf{A}$ is similar to a matrix in Jordan normal form.

Proof. If we work over the complex numbers, it follows from Theorem 4.6.3 that $p_{\mathbf{A}}(Z)$ can be written as a product of its leading coefficient, which is $(-1)^n$, and terms of the form $Z - \lambda$. Hence Theorem 11.5.5 applies. □

Chapter 12

Systems of linear ordinary differential equations of order one with constant coefficients

In this chapter we will investigate some families of differential equations. Differential equations are used to model processes occurring in nature. They occur in almost every area of applied exact sciences, like (quantum) mechanics, (bio)chemistry, dynamics of biological systems, construction engineering, the study of electrical components and circuits, and many more. The theory of differential equations is vast and we will in this book only take a first look at some special cases. Before starting with that, let us fix a few conventions and notations that we will use in the remainder of this chapter.

As we have seen, in general a function $f : A \to B$ is a map between two sets. In this chapter, we will always assume that the domain of the function is the set of real numbers $\mathbb{R}$. If the codomain B is equal to $\mathbb{R}$, we call such a function a real-valued function.

If $B = \mathbb{C}$, we call such a function a complex-valued function. Real- and complex-valued functions occur in many places in mathematics, especially in analysis. The techniques and tools from linear algebra that we have discussed so far in previous chapters, can be used in analysis as well. More precisely, we will see how tools from linear algebra can be used to solve specific types of differential equations. Without being too formal, one can think of a differential equation as a way to find real-valued or complex-valued functions with additional properties involving the derivatives of that function. We will assume that the reader is familiar with the derivative of a real-valued function. Given a function $f : \mathbb{R} \to \mathbb{R}$, we denote by f', the derivative of f, provided it exists. The function $f' : \mathbb{R} \to \mathbb{R}$ is again a real-valued function and as such one can attempt to compute the derivative of f'. If it exists, it is typically denoted by f'' or by $f^{(2)}$. Similarly, one can recursively define for $n \geq 3$, the function

$f^{(n)} : \mathbb{R} \to \mathbb{R}$ to be the derivative of $f^{(n-1)}$, provided it exists. We have seen this notation in Example 9.3.4 as well, where we introduced the real vector space $C_\infty(\mathbb{R})$ consisting of all infinitely differentiable (real-valued) functions. It is customary to write $f^{(0)} = f$ and $f^{(1)} = f'$. In the theory of real- and complex-valued functions, it is quite common to write down a function as $f(t)$, rather than writing $f : \mathbb{R} \to \mathbb{R}$ (for real-valued functions) or $f : \mathbb{R} \to \mathbb{C}$ (for complex-valued functions). In the remainder of this section we will also often do this.

We can now explain in broad terms what we mean by an n-th order ordinary differential equation (abbreviated: *ODE*).

Definition 12.0.1

Let n be a natural number. An n-th order ODE is an equation of the form

$$F(f^{(n)}(t), \ldots, f'(t), f(t), t) = 0,$$

where F is a function taking $n + 2$ variables as input.

A solution of such an ODE is then a real-valued function $f(t)$ such that

$$F(f^{(n)}(t), \ldots, f'(t), f(t), t) = 0$$

for all $t \in \mathbb{R}$. As a first small example: the function $f(t) = e^t$ is a solution to the ODE $f'(t) - f(t) = 0$, because it holds that $(e^t)' = e^t$. We will see more examples later on.

There are many variations and more refined definitions. For example in some cases, one only needs that $F(f^{(n)}(t), \ldots, f'(t), f(t), t) = 0$ for all t in a subset of $\mathbb{R}$. However, all we need at this point is an intuitive understanding of what an ODE is and therefore we will not go into more depth here.

A special kind of ODE where tools from linear algebra can be used is called a linear ODE. We define what that is now:

Definition 12.0.2

A *linear* ODE is an equation of the form $L(f(t)) = q(t)$, where $q(t) \in C_\infty(\mathbb{R})$ is a function and $L : C_\infty(\mathbb{R}) \to C_\infty(\mathbb{R})$ is a linear map. If $q(t)$ is the zero function, one calls the linear ODE a *homogeneous* linear ODE. Otherwise it is called an *inhomogeneous* linear ODE.

Also here variations exists: one does not necessarily need the solutions to be infinitely differentiable nor to be defined on the whole of $\mathbb{R}$. However, as before, we will not go into such details and emphasize the role of linear algebra.

Considering the linear map from Example 10.2.8, which was defined as $L(f(t)) = f(t) - f'(t)$, we see that the linear ODE $L(f(t)) = q(t)$ simply amounts to the differential equation $f'(t) - f(t) = q(t)$. The point of studying linear ODEs in connection with linear algebra is that Theorem 10.4.1 applies in this situation and describes the structure of the set of solutions. Let us consider a couple more examples.

Example 12.0.1

Are the following ODEs linear or not linear? If the ODE is linear, is it homogeneous or inhomogeneous?

(a) $f''(t) + 2f'(t) + f(t) = \cos(t)$

(b) $e^t \cdot f'(t) + \cos(t) \cdot f(t) = 0$

(c) $(f'(t))^2 + f(t) = 0$

Answer:

(a) The given ODE is linear with $L(f(t)) = f''(t) + 2f'(t) + f(t)$ and $q(t) = \cos(t)$. Since $q(t)$ is not the zero function, it is an inhomogeneous linear ODE.

(b) The given ODE is linear with $L(f(t)) = e^t \cdot f'(t) + \cos(t) \cdot f(t)$ and $q(t) = 0$. Since $q(t)$ is the zero function, it is a homogeneous linear ODE.

(c) The given ODE is not linear because of the term $(f'(t))^2$.

One is often primarily interested in real-valued functions as solutions to an ODE, but sometimes it is convenient to look for complex-valued solutions as well. For us the main reason will be to use such complex-valued solutions to find real-valued solutions of an ODE. Let us therefore explain how to compute the derivative of complex-valued functions. Given a complex-valued function $f : \mathbb{R} \to \mathbb{C}$, one can for any $t \in \mathbb{R}$, write $f(t) = f_1(t) + if_2(t)$, where $f_1(t) = \mathrm{Re}(f(t))$ is the real part of $f(t)$ and $f_2(t) = \mathrm{Im}(f(t))$ is the imaginary part of $f(t)$. In this way, any complex-valued function $f : \mathbb{R} \to \mathbb{C}$, gives rise to two real valued-functions $\mathrm{Re}(f) : \mathbb{R} \to \mathbb{R}$ defined as $t \mapsto \mathrm{Re}(f(t))$ and $\mathrm{Im}(f) : \mathbb{R} \to \mathbb{R}$ defined as $t \mapsto \mathrm{Im}(f(t))$. Conversely, given two real-valued functions $f_1 : \mathbb{R} \to \mathbb{R}$ and $f_2 : \mathbb{R} \to \mathbb{R}$, we can define a complex-valued function $f = f_1 + i \cdot f_2$ as $t \mapsto f_1(t) + i \cdot f_2(t)$. If the derivatives of f_1 and f_2 exist, we will define the derivative of f to be the function $f' = f_1' + i \cdot f_2'$. Similarly, we can define for any nonnegative integer n, the n-th derivative $f^{(n)} = f_1^{(n)} + i \cdot f_2^{(n)}$, provided that both $f_1^{(n)}$ and $f_2^{(n)}$ exist. With these conventions in place, we can therefore also talk about complex-valued functions as solutions of an ODE. We will see examples of such solutions later on.

After this brief introductory sketch of what a linear ODE is, let us look at some particular cases and examples in the following sections.

12.1 Linear first-order ODEs

According to Definition 12.0.1, a first-order ODE gives a relation between a function $f(t)$ and its derivative $f'(t)$. For example $f'(t) = f(t)$ is a first-order ODE, but also a more complicated

expression like

$$\sin(f(t)f'(t)) = f'(t)^2 + e^t$$

is a first-order ODE. To bring these examples in the form of Definition 12.0.1, we just rewrite the expressions and make the righthand side zero. For example, the first expression can be written as $f'(t) - f(t) = 0$, while the second example can be written as $\sin(f(t)f'(t)) - f'(t)^2 - e^t = 0$. Let us consider a couple of examples of first-order ODEs.

Example 12.1.1

Investigate whether or not the function $f(t) = e^{2t}$ is a solution to one of the following ODEs:

(a) $f'(t) - 2f(t) = 0$

(b) $f'(t)^2 - 4f(t) = 0$

(c) $\ln(f'(t)) - \ln(f(t)) = \ln(2)$

Answer:

(a) Using the chain rule we find that $f'(t) = (e^{2t})' = e^{2t}(2t)' = e^{2t}2 = 2e^{2t}$. Therefore it holds that

$$f'(t) - 2f(t) = 2e^{2t} - 2e^{2t} = 0.$$

We can therefore conclude that the function $f(t) = e^{2t}$ is a solution to the ODE $f'(t) = 2f(t)$.

(b) We have seen that $f'(t) = 2e^{2t}$. Therefore it holds that

$$f'(t)^2 - 4f(t) = (2e^{2t})^2 - 4e^{2t} = 4(e^{2t})^2 - 4e^{2t} = 4e^{4t} - 4e^{2t} \neq 0.$$

Therefore the function $f(t) = e^{2t}$ is not a solution to the ODE $f'(t)^2 - 4f(t) = 0$.

(c) If $f(t) = e^{2t}$, we find that

$$\ln(f'(t)) - \ln(f(t)) = \ln(2e^{2t}) - \ln(e^{2t}) = \ln(2) + \ln(e^{2t}) - \ln(e^{2t}) = \ln(2),$$

so the function $f(t) = e^{2t}$ is a solution to the ODE $\ln(f'(t)) - \ln(f(t)) = \ln(2)$.

Let us take a look again at the ODE $f'(t) = f(t)$. We mentioned before that the function $f(t) = e^t$ is a solution to this ODE. However, it is not the only one. For example the functions $f(t) = 2e^t$ and $f(t) = -5e^t$ both also satisfy that $f'(t) = f(t)$. In fact any function of the form $f(t) = c \cdot e^t$, with $c \in \mathbb{R}$ a constant, is a solution to the ODE $f'(t) = f(t)$.

One can show that in fact any solution to the ODE $f'(t) = f(t)$ is of the form $f(t) = c \cdot e^t$. Such a description of all possible solutions to an ODE is called its *general solution*. The term general solution was used in a similar way when describing solutions to systems of linear equations. Using this terminology we can say that the general solution to the ODE $f'(t) = f(t)$ is given by $f(t) = c \cdot e^t$, with $c \in \mathbb{R}$.

It can be difficult to find an explicit expression for the general solution to an ODE. However, for some classes of ODEs, it is possible. We will now look at one such class. An ODE of the form

$$f'(t) = a(t)f(t) + q(t), \tag{12.1}$$

with $a(t)$ and $q(t)$ functions in the variable t, is called a *linear first-order ODE*. The function $q(t)$ is also called the *forcing function* of this ODE. Note that an ODE as in Equation (12.1) is indeed linear according to Definition 12.0.2: simply choose the linear map defined by $L(f(t)) = f'(t) - a(t)f(t)$ and $q(t)$ as given, then the equation $L(f(t)) = q(t)$ simply states that $f'(t) - a(t)f(t) = q(t)$, which is equivalent to the equation $f'(t) = a(t)f(t) + q(t)$.

For example the ODE $f'(t) = f(t)$ is a linear first-order ODE. More precisely, by choosing $a : \mathbb{R} \to \mathbb{R}$ to be the function defined by $t \mapsto 1$ and $q : \mathbb{R} \to \mathbb{R}$ to be the function defined by $t \mapsto 0$, Equation (12.1) simplifies to the equation $f'(t) = f(t)$.

Following Definition 12.0.2, the ODE from Equation (12.1) is called *homogeneous* if the forcing function $q(t)$ is the zero function and *inhomogeneous* otherwise.

It turns out that one can give a formula for the general solution to a linear first-order ODE. In this formula we will need a bit of notation. We will by $P(t)$ denote a primitive function (also known as an *antiderivative*) of the function $a(t)$, that is to say, a function satisfying $P'(t) = a(t)$. We will assume in the remainder of this section that the function $a(t)$ in fact has such a primitive function. We will also need to assume that the function $e^{P(t)}q(t)$ has a primitive function. One can show that these assumptions are true if for example both function $a(t)$ and $q(t)$ are differentiable. If this is the case, we have the following result.

Theorem 12.1.1

The general solution to the ODE $f'(t) = a(t)f(t) + q(t)$ is given by

$$f(t) = e^{P(t)} \int e^{-P(t)}q(t)dt.$$

Proof. Recall that $P'(t) = a(t)$. Using first the product rule and then the chain rule, we find that

$$\left(e^{-P(t)}f(t)\right)' = \left(e^{-P(t)}\right)' f(t) + e^{-P(t)}f'(t) = -e^{-P(t)}a(t)f(t) + e^{-P(t)}f'(t).$$

Therefore the following holds:

$$\begin{aligned} f'(t) = a(t)f(t) + q(t) &\Leftrightarrow e^{-P(t)}f'(t) - e^{-P(t)}a(t)f(t) = e^{-P(t)}q(t) \\ &\Leftrightarrow \left(e^{-P(t)}f(t)\right)' = e^{-P(t)}q(t) \\ &\Leftrightarrow e^{-P(t)}f(t) = \int e^{-P(t)}q(t)dt \\ &\Leftrightarrow f(t) = e^{P(t)} \int e^{-P(t)}q(t)dt. \end{aligned}$$

□

When computing the integral in Theorem 12.1.1, one should not forget the integration constant, since this constant is needed when finding the general solution. Let us look at some examples.

Example 12.1.2

Compute the general solution to the following ODEs:

(a) $f'(t) = f(t)$

(b) $f'(t) = -\sin(t)f(t) + \sin(t)$

(c) $f'(t) = -t^{-1}f(t) + 1$, with $t > 0$

Answer:

(a) Rewriting $f'(t) = f(t)$ as $f'(t) - f(t) = 0$, we see that we can apply Theorem 12.1.1, using $a(t) = 1$ and $q(t) = 0$. A primitive function of $a(t) = 1$ is given by for example $P(t) = t$. Then we get that the general solution is given by

$$f(t) = e^t \int e^{-t} 0 \, dt = e^t \int 0 \, dt = e^t c = ce^t.$$

This agrees with the general solution we found before for this ODE.

(b) We can use Theorem 12.1.1 with $a(t) = -\sin(t)$ and $q(t) = \sin(t)$. We can choose $P(t) = \cos(t)$ and we therefore find that the desired general solution is given by

$$f(t) = e^{\cos(t)} \int e^{-\cos(t)} \sin(t) dt = e^{\cos(t)} \left(e^{-\cos(t)} + c \right) = 1 + ce^{\cos(t)}.$$

(c) Theorem 12.1.1 applies with $a(t) = -t^{-1} = -1/t$ and $q(t) = 1$. Since $t > 0$, this means that we can choose $P(t) = -\ln(t)$. The general solution to the ODE $f'(t) = -t^{-1}f(t) + 1$ then becomes

$$f(t) = e^{-\ln(t)} \int e^{\ln(t)} dt = (1/e^{\ln(t)}) \int t dt = \frac{1}{t} \left(\frac{1}{2} t^2 + c \right) = \frac{t}{2} + \frac{c}{t}.$$

One important special case of Theorem 12.1.1 is when the function a is a constant function, say $a(t) = a_0$ for all t. In this case, Theorem 12.1.1 simplifies to the following statement.

Corollary 12.1.2

Let $a_0 \in \mathbb{R}$ and $q(t)$ be a real-valued, differentiable function. Then the ODE $f'(t) = a_0 f(t) + q(t)$ has the general solution $f(t) = e^{a_0 t} \int e^{-a_0 t} q(t) dt$. More concretely, if $Q(t)$ is a primitive function of $e^{-a_0 t} q(t)$, then the general solution can be written as $f(t) = c \cdot e^{a_0 t} + e^{a_0 t} Q(t)$, where $c \in \mathbb{R}$ is arbitrary.

As said before, ODEs are used to model processes occurring in nature. The general solution of an ODE describes all possible behaviors of the process. In order to find out which one of the possibilities is the right one in a particular situation, one needs more information, that one

usually can obtain by performing measurements. One possibility is to describe the behaviour of the function f for a specific value of the variable t. One could imagine that one measures the exact state of the process at the beginning of an experiment. Mathematically speaking, what we will do is to pose an *initial value condition*, that is to say, a condition on a function $f(t)$ of the form $f(t_0) = y_0$, .

Definition 12.1.1

Given a real-valued function $f(t)$ and real numbers t_0 and y_0 such that $f(t_0) = y_0$. Then the function $f(t)$ is said to satisfy the *initial value condition* $f(t_0) = y_0$.

It turns out that in many interesting applications, a function $f : \mathbb{R} \to \mathbb{R}$ is completely determined if it satisfies both a first-order ODE and an initial value condition. We give a description of the situation for general ODEs.

Definition 12.1.2

Let $f(t)$ be a real-valued function satisfying:

(i) An n-th order ODE $F(f^{(n)}(t), \dots, f'(t), f(t), t) = 0$.

(ii) The initial value conditions $f(t_0) = y_0, f'(t_0) = y_1, \dots, f^{(n-1)}(t_0) = y_{n-1}$, for given $t_0 \in \mathbb{R}$ and values $y_0, y_1, \dots, y_{n-1} \in \mathbb{R}$.

The two conditions together are called an *initial value problem*. The function $f(t)$ is said to be a solution to the initial value problem.

For a first-order ODE $F(f'(t), f(t), t) = 0$, this amount to saying that $f(t)$ is a solution to the initial value problem if it satisfies

(i) $F(f'(t), f(t), t) = 0$ and

(ii) $f(t_0) = y_0$, for given $t_0 \in \mathbb{R}$ and a value y_0.

Hence Definition 12.1.1 is just a special case of Definition 12.1.2 when $n = 1$.

The strategy of solving an initial value problem often follows the same pattern. First compute the general solution to the given ODE. This general solution should contain some parameters such as c. Then use the initial value condition to determine c. The resulting function is the desired solution. Let us look at two examples in the case of first-order ODEs.

Example 12.1.3

Solve the following initial value problems. That is to say, compute the function $f(t)$ satisfying:

(a) The ODE $f'(t) = f(t)$ and the initial value condition $f(0) = 7$.

(b) The ODE $f'(t) + \sin(t)f(t) = \sin(t)$ and the initial value condition $f(\pi) = 2$.

Answer:
Note that we already have computed the general solution to the given two ODEs in Example 12.1.2. Now let us look at each initial value problem separately.

(a) We have already seen that the general solution to $f'(t) = f(t)$ is given by $f(t) = ce^t$. The trick is to evaluate $f(t)$ in 0 an compare the result with the initial value condition. We get that $f(0) = c$, but according to the initial value condition we should have $f(0) = 7$. This means that $c = 7$. Now that we know c, we find that the desired function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$f(t) = 7e^t.$$

(b) The general solution is in this case given by $f(t) = 1 + ce^{\cos(t)}$. Using the initial value condition, we find that $2 = f(\pi) = 1 + ce^{\cos(\pi)} = 1 + ce^{-1}$. This means that $ce^{-1} = 1$ and therefore $c = e$. Hence, the desired function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$f(t) = 1 + e \cdot e^{\cos(t)} = 1 + e^{1+\cos(t)}.$$

Before starting to consider more general ODEs, let us establish one nice property of the complex exponential function. We know that the derivative of the real-valued function $f(t) = e^{\lambda t}$ is simply $f'(t) = \lambda e^{\lambda t}$ for any $\lambda \in \mathbb{R}$. It turns out that this is also true for the complex exponential function:

Lemma 12.1.3
Let $\lambda \in \mathbb{C}$ and consider the complex-valued function $f : \mathbb{R} \to \mathbb{C}$ defined as $f(x) = e^{\lambda t}$. Then $\mathrm{Re}(f) = e^{\mathrm{Re}(\lambda)t}\cos(\mathrm{Im}(\lambda)t)$, $\mathrm{Im}(f) = e^{\mathrm{Re}(\lambda)t}\sin(\mathrm{Im}(\lambda)t)$ and $f'(t) = \lambda e^{\lambda t}$.

Proof. Let us write $\lambda = \lambda_1 + i\lambda_2$ in rectangular form. Then for any $t \in \mathbb{R}$, we have

$$\begin{aligned} e^{\lambda t} &= e^{\lambda_1 t + i\cdot\lambda_2 t} \\ &= e^{\lambda_1 t} \cdot e^{i\cdot\lambda_2 t} \\ &= e^{\lambda_1 t} \cdot (\cos(\lambda_2 t) + i \cdot \sin(\lambda_2 t)) \\ &= e^{\lambda_1 t}\cos(\lambda_2 t) + i \cdot e^{\lambda_1 t}\sin(\lambda_2 t)). \end{aligned}$$

This shows that the real part of the expression $f(t) = e^{\lambda t}$ is given by $\mathrm{Re}(f(t)) = e^{\lambda_1 t}\cos(\lambda_2 t)$, while its imaginary part is given by $\mathrm{Im}(f(t)) = e^{\lambda_1 t}\sin(\lambda_2 t)$. Now we set $f'(t) = (\mathrm{Re}(f(t)))' + i \cdot (\mathrm{Im}(f(t)))'$. Using the product and chain rule to compute $\mathrm{Re}(f(t))'$ and $\mathrm{Im}(f(t))'$, we get

$$\begin{aligned} f'(t) &= \mathrm{Re}(f(t))' + i \cdot \mathrm{Im}(f(t))' \\ &= (e^{\lambda_1 t}\cos(\lambda_2 t))' + i \cdot (e^{\lambda_1 t}\sin(\lambda_2 t))' \\ &= (e^{\lambda_1 t}\lambda_1\cos(\lambda_2 t) + e^{\lambda_1 t}(-\sin(\lambda_2 t))\lambda_2) + i \cdot (e^{\lambda_1 t}\lambda_1\sin(\lambda_2 t) + e^{\lambda_1 t}\cos(\lambda_2 t)\lambda_2) \\ &= (\lambda_1 + i\lambda_2)e^{\lambda_1 t}\cos(\lambda_2 t) + (-\lambda_2 + i\lambda_1)e^{\lambda_1 t}\sin(\lambda_2 t) \\ &= (\lambda_1 + i\lambda_2)e^{\lambda_1 t}(\cos(\lambda_2 t) + i\sin(\lambda_2 t)) \end{aligned}$$

$$
\begin{aligned}
&= (\lambda_1 + i\lambda_2)e^{\lambda_1 t}e^{i\lambda_2 t} \\
&= \lambda e^{\lambda t}.
\end{aligned}
$$

□

This lemma will be extremely useful when finding solutions to certain types of ODEs later on.

12.2 Systems of linear first-order ODEs with constant coefficients

In the previous section, we considered linear, first-order ODEs. Now, we consider a system of such ODEs, but we will only consider the case where all the functions occurring as coefficients are constant. After this, in the next section, we will show that some higher order ODEs can be solved using the theory from this section.

Definition 12.2.1

Let $n > 0$ be an integer, $q_1(t), \ldots, q_n(t)$ real-valued differentiable functions and $\mathbf{A} \in \mathbb{R}^{n\times n}$ a matrix. Then a system of linear, first-order ODEs is an equation of the form

$$
\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} + \begin{bmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_n(t) \end{bmatrix} \tag{12.2}
$$

The matrix $\mathbf{A}$ is called the *coefficient matrix* of the system, while the functions $q_1(t), \ldots, q_n(t)$ are called the *forcing functions* of the system. If all forcing functions $q_1(t), \ldots, q_n(t)$ are equal to the zero function, the system of ODEs is called *homogeneous*, otherwise it is called *inhomogeneous*. A solution to an inhomogeneous system of linear, first-order ODEs is called a *particular solution*.

Example 12.2.1

Given is the following system of linear, first-order ODEs:

$$
\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix}. \tag{12.3}
$$

(a) Is the given system of ODEs in Equation (12.3) homogeneous or inhomogeneous?

(b) Is $(f_1(t), f_2(t)) = (e^{2t}, 0)$ a solution to Equation (12.3)?

(c) Is $(f_1(t), f_2(t)) = (-e^t, 0)$ a solution to Equation (12.3)?

Answer:

(a) The system of ODEs in Equation (12.3) is inhomogeneous. Even though the forcing function $q_2(t)$ is the zero function, the function $q_1(t)$ is not. For a homogeneous system, all forcing functions should be the zero function.

(b) If $(f_1(t), f_2(t)) = (e^{2t}, 0)$, then

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} (e^{2t})' \\ 0 \end{bmatrix} = \begin{bmatrix} 2e^{2t} \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \cdot e^{2t} + 1 \cdot 0 \\ 0 \cdot e^{2t} + 2 \cdot 0 \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix} = \begin{bmatrix} 2e^{2t} + e^t \\ 0 \end{bmatrix}.$$

Therefore $(f_1(t), f_2(t)) = (e^{2t}, 0)$ is not a solution to Equation (12.3).

(c) If $(f_1(t), f_2(t)) = (-e^t, 0)$, then

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} (-e^t)' \\ 0 \end{bmatrix} = \begin{bmatrix} -e^t \\ 0 \end{bmatrix}$$

and

$$\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \cdot (-e^t) + 1 \cdot 0 \\ 0 \cdot (-e^t) + 2 \cdot 0 \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix} = \begin{bmatrix} -e^t \\ 0 \end{bmatrix}.$$

Therefore $(f_1(t), f_2(t)) = (-e^t, 0)$ is a solution to Equation (12.3). By definition, it is in fact a particular solution to Equation (12.3).

Now, a bit similarly to what we did for systems of linear equations, we begin by describing the structure of the solutions of systems of linear, first-order ODEs.

Theorem 12.2.1

Let an inhomogeneous system of ODEs as in Equation (12.2) be given and suppose that $(g_1(t), g_2(t), \ldots, g_n(t))$ is a particular solution of this system. Then any other solution $(\tilde{g}_1(t), \tilde{g}_2(t), \ldots, \tilde{g}_n(t))$ to Equation (12.2) is of the form

$$\begin{bmatrix} \tilde{g}_1(t) \\ \tilde{g}_2(t) \\ \vdots \\ \tilde{g}_n(t) \end{bmatrix} = \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_n(t) \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix},$$

where $(f_1(t), f_2(t), \ldots, f_n(t))$ is a solution to the homogeneous system of ODEs corresponding

to Equation (12.2):

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}. \tag{12.4}$$

Proof. Suppose that $(\tilde{g}_1(t), \tilde{g}_2(t), \ldots, \tilde{g}_n(t))$ is an arbitrary solution to Equation (12.2), then a direct computation shows that $(\tilde{g}_1(t) - g_1(t), \tilde{g}_2(t) - g_2(t), \ldots, \tilde{g}_n(t) - g_n(t))$ satisfies Equation (12.4). If we then define $f_i(t) = \tilde{g}_i(t) - g_i(t)$ for $i = 1, \ldots, n$, we see that $(\tilde{g}_1(t), \tilde{g}_2(t), \ldots, \tilde{g}_n(t))$ can be written as stated in the theorem.

Conversely, if $(f_1(t), f_2(t), \ldots, f_n(t))$ is a solution to the homogeneous system from Equation (12.4), then a direct calculation shows that $(g_1(t) + f_1(t), g_2(t) + f_2(t), \ldots, g_n(t) + f_n(t))$ is a solution to the inhomogeneous system from Equation (12.2). □

Algorithmically, this means that in order to solve an inhomogeneous system of ODEs as in Equation (12.2), we need to find a particular solution of it and then all solutions to the corresponding homogeneous system of ODEs given in Equation (12.4). Conceptually, one can understand Theorem 12.2.1 in a different way. Let $C_\infty(\mathbb{R})$ be the vector space from Example 9.3.4. It consists of all functions with domain and codomain $\mathbb{R}$ that can be differentiated arbitrarily often. Now for a given matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, consider the map $L_\mathbf{A} : C_\infty(\mathbb{R})^n \to C_\infty(\mathbb{R})^n$ defined by

$$L_\mathbf{A}\left(\begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}\right) = \begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} - \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}. \tag{12.5}$$

One can show that $L_\mathbf{A}$ is a linear map of real vector spaces. The kernel of this map is exactly the solution set the homogeneous system of ODEs in Equation (12.4). This observation is a generalization of what we already have seen in Example 10.2.8. A particular solution is then nothing but a vector $\mathbf{v}_p = (g_1(t), g_2(t), \ldots, g_n(t)) \in C_\infty(\mathbb{R})^n$ such that $L_\mathbf{A}(\mathbf{v}_p) = (q_1(t), \ldots, q_n(t))$. Therefore, Theorem 12.2.1 is nothing but a special case of the second item in Theorem 10.4.1. As an aside, since the kernel of any linear map is a subspace, we can conclude that the solution set to a homogeneous system of linear, first-order ODEs (with constant coefficients) is in fact a vector space over the real numbers, since it is the kernel of the linear map $L_\mathbf{A}$. A very useful fact, that we will not prove here, is that this vector space has finite dimension, namely n. This is useful to know, since it means that to describe all solutions to the system in Equation (12.4), it is enough to find a basis, that is to say, n linearly independent solutions. We will use this freely later on. What we will primarily focus on in the remainder of this section is how to find such a basis. The notion of general solution we

already encountered in Section 12.1 for linear, first-order ODEs can now be generalized as follows:

Definition 12.2.2

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be given. The *general solution* of the homogeneous ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}$$

is an expression of the form

$$c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n\,, \quad c_1, \ldots, c_n \in \mathbb{R},$$

where $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ is an ordered basis of the kernel of the linear map $L_{\mathbf{A}} : C_\infty(\mathbb{R})^n \to C_\infty(\mathbb{R})^n$ defined in Equation (12.5). If $q_1(t), \ldots, q_n(t)$ are forcing functions (not all zero) and $\mathbf{v}_p = (g_1(t), \ldots, g_n(t)) \in C_\infty(\mathbb{R})^n$ a particular solution of the inhomogeneous system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} + \begin{bmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_n(t) \end{bmatrix},$$

then the general solution of the inhomogeneous system is an expression of the form

$$\mathbf{v}_p + c_1 \cdot \mathbf{v}_1 + \cdots + c_n \cdot \mathbf{v}_n\,, \quad c_1, \ldots, c_n \in \mathbb{R}.$$

A first important trick is to use the theory of eigenvalues and eigenvectors of the matrix $\mathbf{A}$, as we will see in the next lemma.

Lemma 12.2.2

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix and suppose that $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\lambda \in \mathbb{R}$. Then the vector of functions

$$\begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} = \begin{bmatrix} v_1 e^{\lambda t} \\ v_2 e^{\lambda t} \\ \vdots \\ v_n e^{\lambda t} \end{bmatrix}$$

satisfies the homogeneous system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

Proof. On the one hand, we have

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \begin{bmatrix} v_1(e^{\lambda t})' \\ v_2(e^{\lambda t})' \\ \vdots \\ v_n(e^{\lambda t})' \end{bmatrix} = \begin{bmatrix} v_1\lambda e^{\lambda t} \\ v_2\lambda e^{\lambda t} \\ \vdots \\ v_n\lambda e^{\lambda t} \end{bmatrix} = \lambda \begin{bmatrix} v_1 e^{\lambda t} \\ v_2 e^{\lambda t} \\ \vdots \\ v_n e^{\lambda t} \end{bmatrix} = \lambda \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

On the other hand, we find

$$\mathbf{A} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} v_1 e^{\lambda t} \\ v_2 e^{\lambda t} \\ \vdots \\ v_n e^{\lambda t} \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \cdot e^{\lambda t} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \cdot e^{\lambda t} = \lambda \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

□

Example 12.2.2

Let

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

Find a solution to the homogeneous system of linear, first-order ODEs with coefficient matrix $\mathbf{A}$.

Answer:

We are asked to find a solution to the following system of ODEs:

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}. \tag{12.6}$$

With Lemma 12.2.2 in mind, we start by finding an eigenvalue and eigenvector of the given matrix $\mathbf{A}$. The characteristic polynomial of $\mathbf{A}$ is:

$$p_{\mathbf{A}}(Z) = \det(\mathbf{A} - \lambda \mathbf{I}_2) = \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 0 & 2-\lambda \end{bmatrix} \right) = (2-\lambda)^2 = (\lambda - 2)^2.$$

Hence 2 is the the only eigenvalue the matrix $\mathbf{A}$ has. To find an eigenvector of $\mathbf{A}$ with eigenvalue 2, we need to compute a nonzero vector from the kernel of the matrix $\mathbf{A} - 2\mathbf{I}_2$.

In principle, we should then first find the reduced row echelon form of $\mathbf{A} - 2\mathbf{I}_2$, but in this particular case it is in reduced row echelon form already:

$$\mathbf{A} - 2\mathbf{I}_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

We conclude that $\ker(\mathbf{A} - 2\mathbf{I}_2)$ is a one-dimensional vector space with basis given by for example

$$\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}.$$

Now Lemma 12.2.2 implies that

$$\begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} = \begin{bmatrix} 1e^{2t} \\ 0e^{2t} \end{bmatrix} = \begin{bmatrix} e^{2t} \\ 0 \end{bmatrix}$$

is a solution to Equation (12.6).

Using eigenvectors as in Lemma 12.2.2 is enough to find the general solution of Equation (12.4) in case the matrix $\mathbf{A}$ can be diagonalized. Recall that this is equivalent to stating that there exists an invertible matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix. More precisely, we have seen in the previous chapter that the columns of the matrix $\mathbf{Q}$ are linearly independent eigenvectors of the matrix $\mathbf{A}$. If these columns have eigenvalues $\lambda_1, \ldots, \lambda_n$, then we have in fact $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q} = \Lambda$, where Λ is the diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ on its diagonal.

Theorem 12.2.3

Assume that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a diagonalizable matrix. More precisely, let $\mathbf{Q}$ be an invertible matrix such that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is the diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ on its diagonal. Then the homogeneous system in Equation (12.4) has the general solution

$$\begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} = \mathbf{Q} \cdot \begin{bmatrix} c_1 \cdot e^{\lambda_1 t} \\ c_2 \cdot e^{\lambda_2 t} \\ \vdots \\ c_n \cdot e^{\lambda_n t} \end{bmatrix}, \quad c_1, \ldots, c_n \in \mathbb{R}.$$

Proof. First note that the system in Equation (12.4) is equivalent to the system

$$\mathbf{Q}^{-1} \cdot \begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q} \cdot \mathbf{Q}^{-1} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

Defining

$$\begin{bmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \\ \vdots \\ \tilde{f}_n(t) \end{bmatrix} = \mathbf{Q}^{-1} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}, \tag{12.7}$$

we see that $(f_1(t), \ldots, f_n(t))$ is a solution to system in Equation (12.4) if and only if

$$\begin{bmatrix} \tilde{f}_1'(t) \\ \tilde{f}_2'(t) \\ \vdots \\ \tilde{f}_n'(t) \end{bmatrix} = \mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q} \cdot \begin{bmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \\ \vdots \\ \tilde{f}_n(t) \end{bmatrix}. \tag{12.8}$$

However, since the matrix $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$, the system in Equation (12.8) simply amounts to the system of differential equations given by $\tilde{f}_1'(t) = \lambda_1 \cdot \tilde{f}_1(t)$, $\tilde{f}_2'(t) = \lambda_2 \cdot \tilde{f}_2(t), \ldots, \tilde{f}_n'(t) = \lambda_n \cdot \tilde{f}_n(t)$. Each of these differential equations is a homogeneous linear ODE of order one and can therefore be solved individually using Theorem 12.1.1. The result is that $\tilde{f}_i(t) = c_i \cdot e^{\lambda_i t}$ for $i = 1, \ldots, n$ and $c_i \in \mathbb{R}$. In other words

$$\begin{bmatrix} \tilde{f}_1(t) \\ \tilde{f}_2(t) \\ \vdots \\ \tilde{f}_n(t) \end{bmatrix} = \begin{bmatrix} c_1 \cdot e^{\lambda_1 t} \\ c_2 \cdot e^{\lambda_2 t} \\ \vdots \\ c_n \cdot e^{\lambda_n t} \end{bmatrix}.$$

But then using Equation (12.7) the statement of the theorem follows. □

A direct consequence of the theorem is the following alternative description of the general solution.

Corollary 12.2.4

Assume that $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a diagonalizable matrix. More precisely, let $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be an ordered basis of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$. Then the homogeneous system in Equation (12.4) has the general solution

$$c_1 \cdot \mathbf{v}_1 e^{\lambda_1 t} + \cdots + c_n \cdot \mathbf{v}_n e^{\lambda_n t}, \quad c_1, \ldots, c_n \in \mathbb{R}.$$

Proof. Let $\mathbf{Q}$ be the matrix with columns consisting of the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then $\mathbf{Q}$ is an invertible matrix satisfying that $\mathbf{Q}^{-1} \cdot \mathbf{A} \cdot \mathbf{Q}$ is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ on its diagonal. Applying Theorem 12.2.3, the corollary follows. □

Note that in Theorem 12.2.3 and Corollary 12.2.4 it is allowed that some eigenvalues appear several times. In other words: we allow the case where the algebraic multiplicity of some

eigenvalues is greater than one. However, since we assume that the matrix $\mathbf{A}$ is diagonalizable the algebraic and geometric multiplicity of any eigenvalue need to be equal. Hence the theorem will not be applicable if some eigenvalue of $\mathbf{A}$ has a smaller geometric multiplicity than algebraic multiplicity.

Example 12.2.3

Let

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

Then $p_{\mathbf{A}}(Z) = (Z-2)^2 \cdot (Z^2-1) = (Z-2)^2 \cdot (Z-1) \cdot (Z+1)$. Hence $\mathbf{A}$ has three eigenvalues 2, 1 and -1 with algebraic multiplicities 2, 1 and 1 respectively. One can show that bases of the eigenspaces E_2, E_1 and E_{-1} are given by

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right\}, \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\} \text{ and } \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \right\}.$$

In particular, the geometric and algebraic multiplicity is the same for each eigenvalue. Using Corollary 12.2.4, we see that the general solution to the system of linear, first-order ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ f_3'(t) \\ f_4'(t) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ f_4(t) \end{bmatrix}$$

is given by

$$\begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ f_4(t) \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} e^{t} + c_4 \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} e^{-t} = \begin{bmatrix} c_1 e^{2t} \\ c_2 e^{2t} \\ c_3 e^{t} + c_4 e^{-t} \\ c_3 e^{t} - c_4 e^{-t} \end{bmatrix},$$

where $c_1, c_2, c_3, c_4 \in \mathbb{R}$.

Before moving on to the case where the matrix $\mathbf{A}$ cannot be diagonalized, let us mention that also for a system of differential equations as we studied, one can pose initial value conditions similar as we did in Definition 12.1.1.

Definition 12.2.3

Given real-valued functions $f_1(t), \ldots, f_n(t)$ a real numbers t_0 and real numbers $y_1, \ldots, y_n$ such

that $f_i(t_0) = y_i$ for $i = 1, \ldots, n$. Then the functions $f_1(t), \ldots, f_n(t)$ are said to satisfy the *initial value conditions* $f_i(t_0) = y_i$ for $i = 1, \ldots, n$.

It turns out that system in Equation (12.4) has exactly one solution satisfying initial value conditions as in Definition 12.2.3. Determining that solution can be done by first computing the general solution of the system in Equation (12.4), then determining the value of the constants $c_1, \ldots, c_n$ occurring in the general solution such that the initial value conditions are satisfied. We give a small example.

Example 12.2.4

Let us consider the same system of differential equations as in Example 12.2.3. There we saw that the general solution was given by

$$\begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ f_4(t) \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} e^{2t} + c_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} e^{t} + c_4 \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} e^{-t} = \begin{bmatrix} c_1 e^{2t} \\ c_2 e^{2t} \\ c_3 e^{t} + c_4 e^{-t} \\ c_3 e^{t} - c_4 e^{-t} \end{bmatrix},$$

where $c_1, c_2, c_3, c_4 \in \mathbb{R}$.

Question: determine the solution satisfying the initial value conditions $f_1(0) = 1$, $f_2(0) = 2$, $f_3(0) = 3$ and $f_4(0) = 4$.

Answer: Putting $t = 0$ in the general solution and using the given initial value conditions, we find that the constants c_1, c_2, c_3 and c_4 need to satisfy

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 + c_4 \\ c_3 - c_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

This implies that

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 + c_4 \\ c_3 - c_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 7/2 \\ -1/2 \end{bmatrix}.$$

Hence the solution we are looking for is:

$$\begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ f_4(t) \end{bmatrix} = \begin{bmatrix} e^{2t} \\ 2e^{2t} \\ (7e^{t} - e^{-t})/2 \\ (7e^{t} + e^{-t})/2 \end{bmatrix}.$$

Now let us return to studying the system in Equation (12.4). The requirement in Theorem 12.2.3 that there exists a basis of eigenvectors can fail. This happens for example when the characteristic polynomial $p_{\mathbf{A}}(Z)$ cannot be written as a product of polynomials of degree one. In other words $p_{\mathbf{A}}(Z)$ could have complex, non-real roots. The following theorem extends Theorem 12.2.3 for such cases.

Theorem 12.2.5

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a matrix and let $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ be an ordered basis of $\mathbb{C}^n$ consisting of eigenvectors of $\mathbf{A}$ corresponding to (possibly non-real) eigenvalues $\lambda_1, \ldots, \lambda_n$. Then over the complex numbers the homogeneous system in Equation (12.4) has the general solution

$$c_1 \cdot \mathbf{v}_1 e^{\lambda_1 t} + \cdots + c_n \cdot \mathbf{v}_n e^{\lambda_n t}, \quad c_1, \ldots, c_n \in \mathbb{C}.$$

Proof. We leave out the details of the proof, but the proof is practically identical to that of Theorem 12.2.3 and Corollary 12.2.4. The only difference is that we now work over the complex numbers. Note that Lemma 12.1.3 guarantees that $(e^{\lambda t})' = \lambda e^{\lambda t}$ also for $\lambda \in \mathbb{C}$. □

Now suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$, but that its characteristic polynomial $p_{\mathbf{A}}(Z)$ has complex roots. We could view $\mathbf{A}$ as a matrix in $\mathbb{C}^{n\times n}$ and apply Theorem 12.2.5 to obtain the general solution. The problem with this, is that we now found a general solution of complex-valued solutions to Equation 12.4. One often is interested in the general solution of the real-valued solutions instead. Fortunately, this can be achieved with a few tricks. The main trick is that since $p_{\mathbf{A}}(Z)$ has coefficients in $\mathbb{R}$ if $\mathbf{A} \in \mathbb{R}^{n\times n}$, non-real roots occur in pairs: if $\mu \in \mathbb{C} \setminus \mathbb{R}$ is a root, then also $\overline{\mu} \in \mathbb{C}$ is a root, where $\overline{\mu}$ denotes the complex conjugated of λ (see Lemma 4.3.3). In particular, the roots of $p_{\mathbf{A}}(Z)$ can be arranged in the form $\lambda_1, \ldots, \lambda_r$ for the real roots and $\mu_1, \ldots, \mu_s, \overline{\mu}_1, \ldots, \overline{\mu}_s$ for the complex, nonreal roots. Then $n = r + 2s$, where we simply repeat a root m times if it occurs with multiplicity m. Let us illustrate this with an example.

Example 12.2.5

Suppose that $p_{\mathbf{A}}(Z) = (Z-1)\cdot(Z-2)^3 \cdot (Z^2+1)^2$ for some matrix $\mathbf{A} \in \mathbb{R}^{7\times 7}$. Then the roots of this polynomial are 1, 2 with multiplicity 3 and $i, -i$, both with multiplicity 2. There are two real roots, namely 1 and 2, but if we consider these roots with their multiplicity, we should repeat the root 2 thrice. Hence $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 2$ and $\lambda_4 = 2$. There are two complex, nonreal roots i and $-i$, which both should be repeated twice. Hence we have $\mu_1 = i$, $\mu_2 = i$, whence $\overline{\mu}_1 = -i$ and $\overline{\mu}_2 = -i$. Hence in this setting, we have $r = 3$ and $s = 2$.

To describe the general solution of Equation (12.4) in case $p_{\mathbf{A}}(Z)$ has nonreal roots, it will be convenient to define the complex conjugate of a vector $\mathbf{w} \in \mathbb{C}^n$: if $\mathbf{w} = (w_1, \ldots, w_n)$, then $\overline{\mathbf{w}} = (\overline{w}_1, \ldots, \overline{w}_n)$. The point is that if $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{A}\cdot\mathbf{w} = \mu \cdot \mathbf{w}$ for some $\mathbf{w} \in \mathbb{C}^n$ and $\mu \in \mathbb{C} \setminus \mathbb{R}$, then taking the complex conjugate (and using that the entries of $\mathbf{A}$ are real numbers), we see that $\mathbf{A}\cdot\overline{\mathbf{w}} = \overline{\mu}\cdot\overline{\mathbf{w}}$. With this in mind, Theorem 12.2.5 implies the following.

Corollary 12.2.6

Suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$ and that the roots of its characteristic polynomial $p_{\mathbf{A}}(Z)$ are arranged with multiplicity as $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ and $\mu_1, \ldots, \mu_s, \overline{\mu}_1, \ldots, \overline{\mu}_s$, where $\mu_1, \ldots, \mu_s \in \mathbb{C} \setminus \mathbb{R}$. Now suppose that there exist vectors $\mathbf{v}_i \in \mathbb{R}^n$ for $i = 1, \ldots, s$ and $\mathbf{w}_j \in \mathbb{C}^n$ for $j = 1, \ldots, s$ such that:

(i) $\mathbf{A} \cdot \mathbf{v}_i = \lambda_i \cdot \mathbf{v}_i$ for $i = 1, \ldots, r$,

(ii) $\mathbf{A} \cdot \mathbf{w}_j = \mu_j \cdot \mathbf{w}_j$ for $j = 1, \ldots, s$,

(iii) the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{w}_1, \ldots, \mathbf{w}_s, \overline{\mathbf{w}}_1, \ldots, \overline{\mathbf{w}}_s$ form an ordered basis of $\mathbb{C}^n$.

Then the homogeneous system in Equation (12.4) has the general solution

$$c_1 \cdot \mathbf{v}_1 e^{\lambda_1 t} + \cdots + c_r \cdot \mathbf{v}_r e^{\lambda_r t} + c_{r+1} \cdot \mathrm{Re}(\mathbf{w}_1 e^{\mu_1 t}) + \cdots + c_{r+s} \cdot \mathrm{Re}(\mathbf{w}_s e^{\mu_s t}) + c_{r+s+1} \cdot \mathrm{Im}(\mathbf{w}_1 e^{\mu_1 t}) + \cdots + c_n \cdot \mathrm{Im}(\mathbf{w}_s e^{\mu_s t}), \quad c_1, \ldots, c_n \in \mathbb{R}.$$

Proof. When viewed as a matrix over $\mathbb{C}$, the eigenvalues of $\mathbf{A}$ are given by

$$\lambda_1, \ldots, \lambda_r, \mu_1, \ldots, \mu_s, \overline{\mu}_1, \ldots, \overline{\mu}_s.$$

Hence Theorem 12.2.5 implies that

$$\mathbf{v}_1 e^{\lambda_1 t}, \ldots, \mathbf{v}_r e^{\lambda_r t}, \mathbf{w}_1 e^{\mu_1 t}, \ldots, \mathbf{w}_s e^{\mu_s t}, \overline{\mathbf{w}}_1 e^{\overline{\mu}_1 t}, \ldots, \overline{\mathbf{w}}_s e^{\overline{\mu}_s t}$$

form a basis of the set of solutions of Equation (12.4) when working over $\mathbb{C}$. To find a basis of this set of solutions when working over $\mathbb{R}$, we modify this basis. First of all, the solutions $\mathbf{v}_1 e^{\lambda_1 t}, \ldots, \mathbf{v}_r e^{\lambda_r t}$ are already real-valued functions, so no modification is needed for these. Given a pair of complex-valued solutions $\mathbf{w}_j e^{\mu_j t}$ and $\overline{\mathbf{w}}_j e^{\overline{\mu}_j t}$ for some j, we can replace this pair by the pair

$$\frac{\mathbf{w}_j e^{\mu_j t} + \overline{\mathbf{w}}_j e^{\overline{\mu}_j t}}{2} = \mathrm{Re}(\mathbf{w}_j e^{\mu_j t}) \quad \text{and} \quad \frac{\mathbf{w}_j e^{\mu_j t} - \overline{\mathbf{w}}_j e^{\overline{\mu}_j t}}{2i} = \mathrm{Im}(\mathbf{w}_j e^{\mu_j t}).$$

Since $\mathrm{Re}(\mathbf{w}_j e^{\mu_j t})$ and $\mathrm{Im}(\mathbf{w}_j e^{\mu_j t})$ describe real-valued functions, we therefore obtain a basis of all real-valued solutions of Equation (12.4) from the n solutions

$$\mathbf{v}_1 e^{\lambda_1 t}, \ldots, \mathbf{v}_r e^{\lambda_r t}, \mathrm{Re}(\mathbf{w}_1 e^{\mu_1 t}), \ldots, \mathrm{Re}(\mathbf{w}_s e^{\mu_s t}), \mathrm{Im}(\mathbf{w}_1 e^{\mu_1 t}), \ldots, \mathrm{Im}(\mathbf{w}_s e^{\mu_s t}).$$

□

The first item in the corollary simply means that the vector $\mathbf{v}_i$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_i. The second item means that if we would work over the field of complex numbers $\mathbb{C}$, instead of $\mathbb{R}$, then $\mathbf{w}_j$ would be an eigenvector with eigenvalue μ_j. In that case $\overline{\mathbf{w}}_j$ can be shown to be an eigenvector of $\mathbf{A}$ with eigenvalue $\overline{\mu}_j$. Finally, the third item means that there exists a basis of $\mathbb{C}^n$ consisting of eigenvectors of $\mathbf{A}$, when viewed as a matrix in $\mathbb{C}^{n\times n}$.

Hence the three items together can also be reformulated as: when viewed as a matrix in $\mathbb{C}^{n\times n}$, the matrix $\mathbf{A}$ is diagonalizable.

Corollary 12.2.6 may look complicated at first sight, but it is very practical in concrete cases. Let us therefore consider an example.

Example 12.2.6

Let

$$\mathbf{A} = \begin{bmatrix} 0 & 13 \\ -1 & 4 \end{bmatrix}.$$

The aim in this example is to show how to obtain the general solution of the homogeneous system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 13 \\ -1 & 4 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}. \tag{12.9}$$

To be more precise, we want to find the general solution consisting of real-valued functions.

First of all, we compute that

$$p_{\mathbf{A}}(Z) = \det(\mathbf{A} - Z\mathbf{I}_2) = \det\left(\begin{bmatrix} -Z & 13 \\ -1 & 4-Z \end{bmatrix}\right) = (-Z)\cdot(4-Z) - 13\cdot(-1) = Z^2 - 4Z + 13.$$

This polynomial has roots $2+3i$ and $2-3i$ (see Theorem 4.2.1). Since the roots are nonreal, let us work over the complex numbers for now. First we compute a complex eigenvector for the nonreal root $2+3i$. We do this by finding the reduced row echelon form of the matrix $\mathbf{A} - (2+3i)\mathbf{I}_2$:

$$\begin{aligned} \mathbf{A} - (2+3i)\mathbf{I}_2 = \begin{bmatrix} -2-3i & 13 \\ -1 & 2-3i \end{bmatrix} &\longrightarrow \begin{bmatrix} -1 & 2-3i \\ -2-3i & 13 \end{bmatrix} \\ R_1 \leftrightarrow R_2 \quad &\longrightarrow \begin{bmatrix} 1 & -2+3i \\ -2-3i & 13 \end{bmatrix} \\ R_1 \leftarrow -R_1 \quad &\longrightarrow \begin{bmatrix} 1 & -2+3i \\ 0 & 0 \end{bmatrix}. \\ R_2 \leftarrow R_2 + (2+3i)R_1 \quad & \end{aligned}$$

Now we see that E_{2+3i}, that is to say the kernel of $\mathbf{A} - (2+3i)\mathbf{I}_2$ when viewed as a matrix in $\mathbb{C}^{2\times 2}$, is equal to $\{(v_1, v_2) \in \mathbb{C}^2 \mid v_1 = (2-3i)v_2\}$. Hence a basis of E_{2+3i} is for example given by

$$\left\{\begin{bmatrix} 2-3i \\ 1 \end{bmatrix}\right\}.$$

Similarly, one shows that a possible basis of E_{2-3i} is

$$\left\{\begin{bmatrix} 2+3i \\ 1 \end{bmatrix}\right\},$$

but we do not actually need this second basis. Now following the recipee described in Corollary 12.2.6, we first compute

$$\begin{aligned}\begin{bmatrix} 2-3i \\ 1 \end{bmatrix} e^{(2+3i)t} &= \begin{bmatrix} 2-3i \\ 1 \end{bmatrix} e^{2t}(\cos(3t)+i\sin(3t)) \\ &= \begin{bmatrix} (2-3i)e^{2t}(\cos(3t)+i\sin(3t)) \\ e^{2t}(\cos(3t)+i\sin(3t)) \end{bmatrix} \\ &= \begin{bmatrix} 2e^{2t}\cos(3t)+3e^{2t}\sin(3t)+i(2e^{2t}\sin(3t)-3e^{2t}\cos(3t)) \\ e^{2t}\cos(3t)+ie^{2t}\sin(3t) \end{bmatrix}\end{aligned}$$

Hence

$$\operatorname{Re}\left(\begin{bmatrix} 2-3i \\ 1 \end{bmatrix} e^{(2+3i)t}\right) = \begin{bmatrix} 2e^{2t}\cos(3t)+3e^{2t}\sin(3t) \\ e^{2t}\cos(3t) \end{bmatrix}$$

and

$$\operatorname{Im}\left(\begin{bmatrix} 2-3i \\ 1 \end{bmatrix} e^{(2+3i)t}\right) = \begin{bmatrix} 2e^{2t}\sin(3t)-3e^{2t}\cos(3t) \\ e^{2t}\sin(3t) \end{bmatrix}.$$

By Corollary 12.2.6, we can conclude that the general solution of the system in Equation (12.9) is given by

$$\begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} = c_1 \cdot \begin{bmatrix} 2e^{2t}\cos(3t)+3e^{2t}\sin(3t) \\ e^{2t}\cos(3t) \end{bmatrix} + c_2 \cdot \begin{bmatrix} 2e^{2t}\sin(3t)-3e^{2t}\cos(3t) \\ e^{2t}\sin(3t) \end{bmatrix},$$

where $c_1, c_2 \in \mathbb{R}$.

We have now given the general solution in case the matrix $\mathbf{A}$ is diagonalizable over $\mathbb{R}$ (Theorem 12.2.3) or over $\mathbb{C}$ (Corollary 12.2.6). If the matrix is not diagonalizable, not even over $\mathbb{C}$, a formula for the general solution is known, but this is out of scope of these notes. We will show an example though for a particular case.

Example 12.2.7

Let

$$\mathbf{A} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \text{ with } \lambda \in \mathbb{R}.$$

This matrix has λ as eigenvalue with algebraic multiplicity two and geometric multiplicity one. Hence Theorem 12.2.3 does not apply, since E_λ is only one-dimensional with basis for example formed by the vector $(1,0)$.

We wish to determine the general solution to the system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}. \tag{12.10}$$

In other words, we have the two ODEs $f_1'(t) = \lambda \cdot f_1(t) + f_2(t)$ and $f_2'(t) = \lambda \cdot f_2(t)$. One solution is found by putting $f_2(t) = 0$, the zero function, and $f_1(t) = e^{\lambda t}$. In other words: the vector of functions $(e^{\lambda t}, 0)$ is a solution to the system in Equation (12.10). Another solution can be found by choosing $f_2(t) = e^{\lambda t}$. Then $f_1(t)$ needs to satisfy the linear inhomogeneous ODE $f_1'(t) = \lambda \cdot f_1(t) + e^{\lambda t}$. Using Corollary 12.1.2, we see that $f_1(t) = e^{\lambda t} \int e^{-\lambda t} e^{\lambda t} dt = e^{\lambda t} t + c \cdot e^{\lambda t}$, where $c \in \mathbb{R}$. Choosing $c = 0$, we see that $(f_1(t), f_2(t)) = (te^{\lambda t}, e^{\lambda t})$ is also a solution to the system in Equation (12.10). Since we now have found two linearly independent solutions, we can conclude that the general solution of the system in Equation (12.10) is given by

$$\begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} = c_1 \cdot \begin{bmatrix} e^{\lambda t} \\ 0 \end{bmatrix} + c_2 \cdot \begin{bmatrix} te^{\lambda t} \\ e^{\lambda t} \end{bmatrix}, \quad c_1, c_2 \in \mathbb{R}.$$

12.3 Relating systems of linear, first-order ODEs with linear second order ODEs

As an application of the previous section, we briefly consider a very special type of second order ODEs:

Definition 12.3.1

Let $a_0, a_1 \in \mathbb{R}$ be constants and $q : \mathbb{R} \to \mathbb{R}$ a function. Then a linear, second order ODE with constant coefficients is an ODE of the form

$$f''(t) + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t). \tag{12.11}$$

The function $q(t)$ is called the *forcing function* of the ODE. If the forcing function $q(t)$ is the zero function, the ODE is called *homogeneous*, otherwise it is called *inhomogeneous*.

As mentioned in Definition 12.1.2, one often poses initial value conditions of the form $f(t_0) = y_0, f'(t_0) = y_1$ for a given $t_0 \in \mathbb{R}$ and values $y_0, y_1 \in \mathbb{R}$. One can show that if $q(t)$ is a differentiable function, then the ODE in Equation (12.11) has exactly one solution satisfying a given initial value condition. For ODEs as in Equation (12.11), a way to find this solution is to first determine its general solution. We will explain how to do this in this section using the theory from the previous section. The following theorem holds the key:

Theorem 12.3.1

Let a function $f : \mathbb{R} \to \mathbb{R}$ be given. If f is a solution to the ODE

$$f''(t) + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t), \tag{12.12}$$

then the vector of functions $(f(t), f'(t))$ is a solution to the system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ q(t) \end{bmatrix}. \tag{12.13}$$

Conversely, if $(f_1(t), f_2(t))$ is a solution to the system of ODEs in Equation (12.13), then $f_1(t)$ is a solution to the ODE in Equation (12.12).

Proof. This is left to the reader. □

Example 12.3.1

A function $f(t)$ is a solution to the linear, second-order ODE $f''(t) + 5f'(t) + 6f(t) = 0$ if and only if the vector of functions $(f(t), f'(t))$ is a solution to the system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -6 & -5 \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}.$$

Theorem 12.3.1 has the following nice consequence about the structure of the solution to a linear, second order ODE.

Corollary 12.3.2

Let an inhomogeneous, linear, second order ODE

$$f''(t) + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t)$$

be given and suppose that $f_p(t)$ is a particular solution of this differential equation. Then any other solution $f(t)$ is of the form $f_p(t) + f_h(t)$, where $f_h(t)$ is a solution to the corresponding homogeneous ODE

$$f''(t) + a_1 \cdot f'(t) + a_0 \cdot f(t) = 0. \tag{12.14}$$

Proof. This follows by combining Theorems 12.2.1 and 12.3.1. □

In order to use Theorem 12.3.1 to find the solution of a linear, second order ODE, one needs to study the eigenvalues and eigenvectors of the matrix occurring in the theorem. Studying its characteristic polynomial is therefore an important first step. We have

$$\begin{aligned} \det\left(\begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} - Z\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) &= \det\left(\begin{bmatrix} -Z & 1 \\ -a_0 & -a_1 - Z \end{bmatrix}\right) \\ &= (-Z)\cdot(-a_1 - Z) - 1 \cdot (-a_0) = Z^2 + a_1 Z + a_0. \end{aligned}$$

This motivates the following definition:

Definition 12.3.2

The polynomial

$$Z^2 + a_1 Z + a_0$$

is called the *characteristic polynomial* of the ODE

$$f''(t) + a_1 \cdot f'(t) + a_0 \cdot f(t) = 0.$$

With this definition in place, solving a homogeneous, linear, second order ODE can be done completely explicitly. There are three cases to distinguish, depending on whether this polynomial has two distinct real roots, two complex conjugated, nonreal roots, or one real root with multiplicity two (see Theorem 4.2.1).

Case 1: The polynomial $Z^2 + a_1 Z + a_0$ has two distinct real roots. If $Z^2 + a_1 Z + a_0$ has two distinct real roots, this means that its discriminant $D = a_1^2 - 4a_0$ is positive and that the real roots are $\lambda_1 = \dfrac{-a_1 + \sqrt{D}}{2}$ and $\lambda_2 = \dfrac{-a_1 + \sqrt{D}}{2}$. We could now use Theorem 12.3.1 and Theorem 12.2.3 to find the general solution to the ODE in Equation (12.14), but a direct approach is faster. The point is though that after the theory about systems of ODEs, we expect that the general solution will involve the functions $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$. Indeed, we simply claim that both $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ are solutions to the ODE in Equation (12.14). For example, we see that

$$\begin{aligned}(e^{\lambda_1 t})'' + a_1(e^{\lambda_1 t})' + a_0 e^{\lambda_1 t} &= \lambda_1^2 e^{\lambda_1 t} + a_1 \lambda_1 e^{\lambda_1 t} + a_0 e^{\lambda_1 t} \\ &= (\lambda_1^2 + a_1\lambda_1 + a_0)e^{\lambda_1 t} \\ &= 0,\end{aligned}$$

where in the last equality, we used that λ_1 is a root of the polynomial $Z^2 + a_1 Z + a_0$. Very similarly, one shows that the function $e^{\lambda_2 t}$ also is a solution. If $D = a_1^2 - 4a_0 > 0$, the general solution to the ODE in Equation (12.14) will therefore be:

$$c_1 \cdot e^{\lambda_1 t} + c_2 \cdot e^{\lambda_2 t} = c_1 \cdot e^{\left(\frac{-a_1+\sqrt{D}}{2}\right)t} + c_2 \cdot e^{\left(\frac{-a_1-\sqrt{D}}{2}\right)t}, \quad c_1, c_2 \in \mathbb{R}. \tag{12.15}$$

Case 2: The polynomial $Z^2 + a_1 Z + a_0$ has two nonreal roots. In this case the discriminant $D = a_1^2 - 4a_0$ is negative and the roots of $Z^2 + a_1 Z + a_0$ are $\lambda_1 = \dfrac{-a_1 + i\sqrt{|D|}}{2}$ and $\lambda_2 = \dfrac{-a_1 - i\sqrt{|D|}}{2}$. Very similarly as in the previous case, one can show, this time using Lemma 12.1.3, that both $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ are complex-valued solutions to the ODE in Equation (12.14). To find real-valued solutions, we simply take the real and imaginary parts of one of these solutions, inspired by what we did in Corollary 12.2.6. We have

$$\mathrm{Re}(e^{\lambda_1 t}) = \mathrm{Re}(e^{(\frac{-a_1+i\sqrt{|D|}}{2})t}) = e^{(\frac{-a_1}{2})t}\cos\left(\frac{\sqrt{|D|}}{2}t\right)$$

and similarly

$$\mathrm{Im}(e^{\lambda_1 t}) = \mathrm{Im}(e^{(\frac{-a_1+i\sqrt{|D|}}{2})t}) = e^{(\frac{-a_1}{2})t}\sin\left(\frac{\sqrt{|D|}}{2}t\right).$$

If $D = a_1^2 - 4a_0 < 0$, the general solution to the ODE in Equation (12.14) will therefore be:

$$c_1 \cdot e^{(\frac{-a_1}{2})t}\cos\left(\frac{\sqrt{|D|}}{2}t\right) + c_2 \cdot e^{(\frac{-a_1}{2})t}\sin\left(\frac{\sqrt{|D|}}{2}t\right), \quad c_1, c_2 \in \mathbb{R}. \tag{12.16}$$

Case 3: The polynomial $Z^2 + a_1 Z + a_0$ has one real root with multiplicity two. In this case the discriminant $D = a_1^2 - 4a_0$ is zero and the double root is given by $\lambda = -a_1/2$. As in the previous cases, one can show directly that $e^{\lambda t}$ is a solution to the ODE in Equation (12.14), but what is missing is a second solution. Again we can get inspiration from what happened for systems of linear ODEs. In Example 12.2.7, we were in the situation that the algebraic multiplicity of an eigenvalue was two, but its geometric multiplicity was one. We are in a similar situation here. Indeed, if $D = 0$, then the matrix $\begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix}$ occurring in Theorem 12.3.1 has eigenvalue λ with algebraic multiplicity two, but one can show that its geometric multiplicity is only one. Since in Example 12.2.7, the function $te^{\lambda t}$ appeared, it is natural to try if this function is a solution to the ODE in Equation (12.14). This is indeed the case:

$$\begin{aligned}
(te^{\lambda t})'' + a_1(te^{\lambda t})' + a_0 te^{\lambda t} &= (e^{\lambda t} + t\lambda e^{\lambda t})' + a_1(e^{\lambda t} + t\lambda e^{\lambda t}) + a_0 te^{\lambda t} \\
&= (\lambda e^{\lambda t} + \lambda e^{\lambda t} + t\lambda^2 e^{\lambda t}) + a_1(e^{\lambda t} + t\lambda e^{\lambda t}) + a_0 te^{\lambda t} \\
&= (\lambda^2 + a_1\lambda + a_0)te^{\lambda t} + (2\lambda + a_1)e^{\lambda t} \\
&= (2\lambda + a_1)e^{\lambda t} \\
&= 0,
\end{aligned}$$

where in the last two equalities we used that $\lambda^2 + a_1\lambda + a_0 = 0$ and $\lambda = -a_1/2$. We conclude the following. If $D = a_1^2 - 4a_0 = 0$, the general solution to the ODE in Equation (12.14) is:

$$c_1 \cdot e^{\lambda t} + c_2 \cdot te^{\lambda t} = c_1 \cdot e^{(\frac{-a_1}{2})t} + c_2 \cdot t \cdot e^{(\frac{-a_1}{2})t}, \quad c_1, c_2 \in \mathbb{R}. \tag{12.17}$$

We finish the section with considering several examples.

Example 12.3.2

Compute the general solution to the differential equation $f''(t) - 5f'(t) + 6f(t) = 0$.

Answer: The characteristic polynomial of the differential equation is $Z^2 - 5Z + 6$. This polynomial has discriminant 1 and therefore has two disctinct real roots. Computing these roots in the usual way, one finds that they are 2 and 3.

Using Equation (12.15), we then find the following general solution

$$f(t) = c_1 e^{2t} + c_2 e^{3t}, \quad c_1, c_2 \in \mathbb{R}.$$

Example 12.3.3

Compute the general solution to the differential equation $f''(t) - 4f'(t) + 4f(t) = 0$.

Answer: The characteristic polynomial of the differential equation is $Z^2 - 4Z + 4$, which has discriminant zero. More precisely, it has 2 as a root with multiplicity two. Equation (12.17) then implies that the general solution we are looking for is given by:

$$f(t) = c_1 e^{2t} + c_2 t e^{2t}, \quad c_1, c_2 \in \mathbb{R}.$$

Example 12.3.4

Compute the general solution to the differential equation $f''(t) - 4f'(t) + 13f(t) = 0$.

Answer: In this case, the characteristic polynomial of the differential equation is $Z^2 - 4Z + 13$, which has a negative discriminant, namely $D = (-4)^2 - 4 \cdot 13 = -36$. Hence the characteristic polynomial has two non-real roots, which turn out to be $2 + 3i$ and $2 - 3i$. According to Equation (12.16) the wanted general solution is:

$$f(t) = c_1 e^{2t} \cos(3t) + c_2 e^{2t} \sin(3t), \quad c_1, c_2 \in \mathbb{R}.$$

Finally, we give examples of inhomogeneous, linear, second-order ODEs.

Example 12.3.5

Compute the general solution to the following differential equations:

1. $f''(t) - 5f'(t) + 6f(t) = t$. It is given that there exists a particular solution of the form $f(t) = at + b$ with $a, b \in \mathbb{R}$.
2. $f''(t) - 4f'(t) + 4f(t) = e^t$. It is given that $f(t) = e^t$ is a solution.
3. $f''(t) - 4f'(t) + 13f(t) = 1$. It is given that there exists a solution of the form $f(t) = a$ with $a \in \mathbb{R}$.

Answer:

Using Corollary 12.4.2 and the previous examples, it is enough to find a particular solution to each of the differential equations.

1. Let us try to find a particular solution of the form $f(t) = at + b$, with $a, b \in \mathbb{R}$. Inserting this in the differential equation, we see that $0 - 4a + 4(at + b) = t$. Hence $4a = 1$ and $-4a + 4b = 0$. We see that $f(t) = t/4 + 1/4$ is a particular solution. Using Example 12.3.2 and Corollary 12.4.2, we conclude that the general solution is given by:

$$f(t) = \frac{t}{4} + \frac{1}{4} + c_1 e^{2t} + c_2 e^{3t}, \quad c_1, c_2 \in \mathbb{R}.$$

2. Since we are given a particular solution, we can find the general solution directly from Example 12.3.3 using Corollary 12.4.2. The result is:

$$f(t) = e^t + c_1 e^{2t} + c_2 t e^{2t}, \quad c_1, c_2 \in \mathbb{R}.$$

3. First we find a particular solution of the form $f(t) = a$. Inserting this in the differential equations, we see that $0 - 4 \cdot 0 + 13a = 1$ and therefore $f(t) = 1/13$ is a particular solution. Now similarly as before, combining this particular solution and the general solution for the corresponding homogeneous ODE given in Example 12.3.4, we find the desired general solution to the given inhomogeneous equation:

$$f(t) = \frac{1}{13} + c_1 e^{2t}\cos(3t) + c_2 e^{2t}\sin(3t), \quad c_1, c_2 \in \mathbb{R}.$$

12.4 Relating systems of linear, first-order ODEs with linear higher order ODEs

The same ideas that were used in the previous section can be used for certain n-th order ODEs as well. For completeness sake, we show how to do this in this section. This is section is not required reading and only meant as an extra section for those that want to know a little bit more. We start out the same way as for linear, second order ODEs.

Definition 12.4.1

Let n be a natural number, $a_0, \ldots, a_{n-1} \in \mathbb{R}$ constants and $q : \mathbb{R} \to \mathbb{R}$ a function. Then a linear, n-th order ODE with constant coefficients is an ODE of the form

$$f^{(n)}(t) + a_{n-1} \cdot f^{(n-1)}(t) + \cdots + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t). \tag{12.18}$$

The function $q(t)$ is called the *forcing function* of the ODE. If the forcing function $q(t)$ is the zero function, the ODE is called *homogeneous*, otherwise it is called *inhomogeneous*.

As mentioned in Definition 12.1.2, one often poses initial value conditions of the form $f(t_0) = y_0, f'(t_0) = y_1, \ldots, f^{(n-1)}(t_0) = y_{n-1}$, for a given $t_0 \in \mathbb{R}$ and values $y_0, y_1, \ldots, y_{n-1} \in \mathbb{R}$. One can show that if $q(t)$ is a differentiable function, then the ODE in Equation (12.18) has exactly one solution satisfying a given initial value condition. For ODEs as in Equation (12.18), a way to find this solution is to first determine its general solution. We will explain how to do this in this section.

The main trick is to relate a solution of a linear, n-th order ODE with constant coefficients with a solution of an appropriately chosen system of linear, first-order ODEs.

Theorem 12.4.1

Let a function $f : \mathbb{R} \to \mathbb{R}$ be given. If f is a solution to the ODE

$$f^{(n)}(t) + a_{n-1} \cdot f^{(n-1)}(t) + \cdots + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t), \tag{12.19}$$

then the vector of functions $(f(t), f'(t), \ldots, f^{(n-1)}(t))$ is a solution to the system of ODEs

$$\begin{bmatrix} f_1'(t) \\ f_2'(t) \\ \vdots \\ f_n'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix} \cdot \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ q(t) \end{bmatrix}. \tag{12.20}$$

Conversely, if $(f_1(t), \ldots, f_n(t))$ is a solution to the system of ODEs in Equation (12.20), then $f_1(t)$ is a solution to the ODE in Equation (12.19).

Proof. This is left to the reader. □

Theorem 12.4.1 implies that when investigating the ODE in Equation (12.18), we can use all theory we have developed in the previous section. For example, we can conclude the following. b

Corollary 12.4.2

Let an inhomogeneous, linear, n-th order ODE

$$f^{(n)}(t) + a_{n-1} \cdot f^{(n-1)}(t) + \cdots + a_1 \cdot f'(t) + a_0 \cdot f(t) = q(t)$$

be given and suppose that $f_p(t)$ is a particular solution of this differential equation. Then any other solution $f(t)$ is of the form $f_p(t) + f_h(t)$, where $f_h(t)$ is a solution to the corresponding homogeneous ODE

$$f^{(n)}(t) + a_{n-1} \cdot f^{(n-1)}(t) + \cdots + a_1 \cdot f'(t) + a_0 \cdot f(t) = 0. \tag{12.21}$$

Proof. This follows by combining Theorems 12.2.1 and 12.4.1. □

As for systems of n linear, first-order ODEs, one can show that the solution set of a homogeneous, linear, n-th order ODE forms a vector space of dimension n. Therefore, to describe the general solution, one needs to find n linearly independent solutions. Similarly as in the case of systems of linear, first-order ODEs, a first step towards computing the general solution of a linear, n-th order ODE, is to find the general solution of the corresponding homogeneous ODE. If we would use Theorem 12.4.1, the first step would be to compute the characteristic polynomial of matrices of the form occurring in Theorem 12.4.1. Fortunately, there is a practical formula for the characteristic polynomials of such matrices. It even works over any field $\mathbb{F}$.

Lemma 12.4.3

Let $\mathbb{F}$ be a field, $n \geq 2$ an integer and $a_0, \ldots, a_{n-1} \in \mathbb{F}$. Then the characteristic polynomial of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ -a_0 & \cdots & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}$$

is equal to

$$p_{\mathbf{A}}(Z) = (-1)^n \cdot (Z^n + a_{n-1}Z^{n-1} + \cdots + a_1 Z + a_0).$$

Proof. We prove this by induction on n for $n \neq 2$. If $n = 2$, we can directly see that

$$p_{\mathbf{A}}(Z) = \det\left(\begin{bmatrix} -Z & 1 \\ -a_0 & -a_1 - Z \end{bmatrix}\right) = (-Z) \cdot (-a_1 - Z) - 1 \cdot (-a_0) = Z^2 + a_1 Z + a_0.$$

Now assume that $n > 2$ and that the result is true for $n - 1$. Developing the determinant of $\mathbf{A} - Z\mathbf{I}_n$ in the first column, we see that:

$$\det(\mathbf{A} - Z\mathbf{I}_n) = -Z \cdot \det\left(\begin{bmatrix} -Z & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -Z & 1 & 0 \\ 0 & \cdots & 0 & -Z & 1 \\ -a_1 & \cdots & \cdots & -a_{n-2} & -a_{n-1} - Z \end{bmatrix}\right)$$
$$+ (-1)^n \cdot (-a_0) \cdot \det\left(\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -Z & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -Z & 1 & 0 \\ 0 & \cdots & 0 & -Z & 1 \end{bmatrix}\right).$$

Using the induction hypothesis on the first determinant after the equality and Theorem 8.1.4 for the second determinant, we see that

$$\begin{aligned} \det(\mathbf{A} - Z\mathbf{I}_n) &= (-Z) \cdot (-1)^{n-1} \cdot (Z^{n-1+} a_{n-1}Z^{n-2} + \cdots + a_1) + (-1)^{n-1} \cdot (-a_0) \cdot 1 \\ &= (-1)^n \cdot (Z^n + a_{n-1}Z^{n-1} + \ldots f a_1 Z + a_0). \end{aligned}$$

This concludes the induction step. Hence the lemma is true for any integer $n \geq 2$. □

The matrix in Lemma 12.4.3 is called the *companion matrix* of the polynomial $Z^n + a_{n-1}Z^{n-1} + \cdots + a_0$. Lemma 12.4.3 implies that when solving the linear, n-th order ODE in Equation

(12.18), then the first thing one needs to do is to find the roots of the polynomial $Z^n + a_{n-1}Z^{n-1} + \dots f a_1 Z + a_0$. The polynomial

$$Z^n + a_{n-1}Z^{n-1} + \cdots + a_1 Z + a_0$$

is often called the *characteristic polynomial* of the ODE

$$f^{(n)}(t) + a_{n-1} \cdot f^{(n-1)}(t) + \cdots + a_1 \cdot f'(t) + a_0 \cdot f(t) = 0.$$

At this point, we could continue to develop the theory of linear, n-th order ODEs, but we will not do this in these notes.

Appendix A

Appendices

A.1 The unit circle

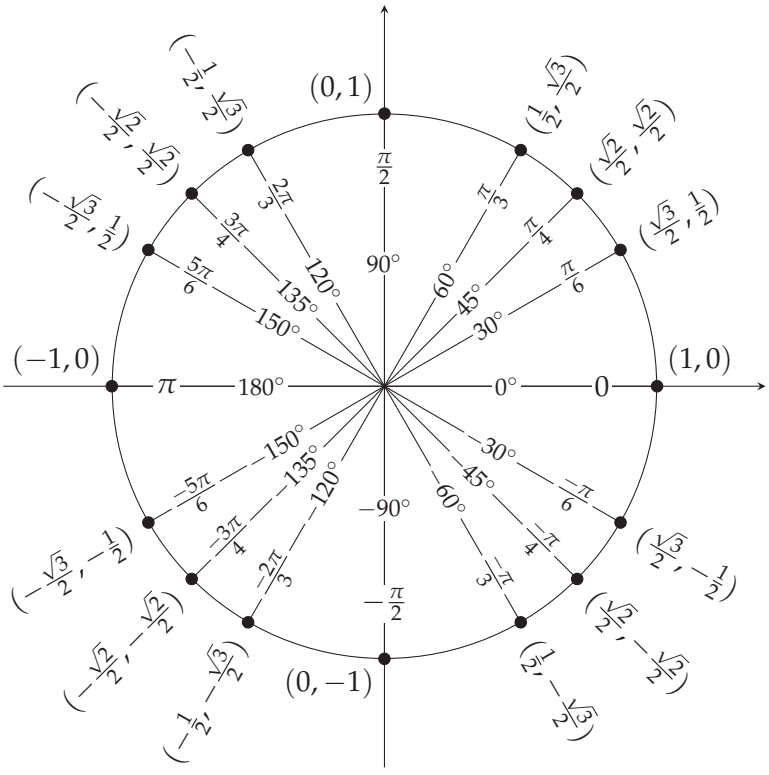


Figure A.1: The unit circle

A.2 Some rules for differentiation

Differentiation of a sum:

$$(f(t)+g(t))' = f'(t)+g'(t)$$

Differentiation of a product (product rule):

$$(f(t)\cdot g(t))' = f'(t)\cdot g(t)+f(t)\cdot g'(t)$$

Differentiation of composed function (chain rule):

$$(f(g(t)))' = f'(g(t))\cdot g'(t)$$

Differentiation of a quotient (quotient rule):

$$\left(\frac{f(t)}{g(t)}\right)' = \frac{f'(t)\cdot g(t)-f(t)\cdot g'(t)}{(g(t))^2}$$

Derivative of some standard functions (a denotes a real constant, in the case of $(a^t)'$ it is assumed that a is positive; n is an integer):

$(a)' = 0$	$(at)' = a$	$(t^n)' = nt^{n-1}$
$(e^t)' = e^t$	$(a^t)' = \ln(a)a^t$	$(\ln(t))' = 1/t$
$(\sin(t))' = \cos(t)$	$(\cos(t))' = -\sin(t)$	$(\tan(t))' = \frac{1}{\cos(t)^2} = 1+\tan(t)^2$

A.3 A small dictionary for mathematical terms

English to Danish

English	Danish
absolute value	absolutværdi, modulus (i tilfældet af komplekse tal)
argument (of a complex number)	argument (af et komplekst tal)
augmented matrix	totalmatrix (til et lineært ligningssystem)
associative	associativ
biimplication	biimplikation
bijective, bijection	bijektiv, bijektion
binomial equation, the	den binome ligning
chain rule	kædereglen
co-domain	dispositionsmængde
coefficient matrix	koefficientmatrix (til et lineært ligningssystem)
column (of a matrix)	søjle
column vector	søjlevektor
commutative	kommutativ
complex conjugate	kompleks konjugerede
complex number	komplekst tal
complex plane, the	den komplekse talplan
composite, composition	sammensat, sammensætning
continuous function	kontinuert funktion
coordinate vector	koordinatvektor
definite integral	bestemt integral
degree	grad
derivative of a function	afledte funktion
diagonal	diagonal
diagonal matrix	diagonalmatrix
disjoint	disjunkt
domain (of a function)	definitionsmængde
double root	dobbeltrod
(row) echelon form	trappeform
equation	ligning
even number	lige tal
equation	ligning
expansion of the determinant	udvikling af determinanten (efter en række eller søjle)

English	Danish
field	legeme
fraction	brøktal
general solution	fuldstændige løsning
identity matrix	identitetsmatrix
image (of a function)	billede, værdimængde
implication	implikation
indeterminate	ubekendte
injective	injektiv
integer	heltal
intersection of sets	fællesmængde
inverse element	inverst element
Laplace expansion of the determinant	udvikling af determinanten (efter en række eller søjle)
linear combination	linearkombination
linearly (in)dependent	lineært (u)afhængig
limit	grænsværdi
logically equivalent	logisk ækvivalent
logical proposition	logisk udsagn
lower triangular matrix	nedre trekantsmatrix
matrix, matrices	matrix, matricer
modulus	modulus, absolutværdi
monic	monisk
natural number	naturligt tal
odd number	ulige tal
particular solution	partikulær løsning
partial integration	delvis integration, partiel integration
polar coordinates	polære koordinater
polynomial	polynom, polynomium
preimage	urbillede
product rule	produktreglen
propositional logic	udsagnslogik
primitive function	stamfunktion
principal value of the argument	hovedargument (af et komplekst tal)
quotient rule	kvotientreglen

English	Danish
rank	rang (af en matrix)
real line, the	den reelle tallinje
real number	reelt tal
reduced row echelon form	reduceret trappeform (af en matrix)
root of a polynomial	rod i et polynom
row (of a matrix)	række
row echelon form	trappeform
row vector	rækkevektor
scalar	skalar
second order equation	andenordens ligning
set	mængde
span	udspænding
square matrix	kvadratisk matrix
square root	kvadratrod
surjective	surjektiv
transpose	transponerede (matrix)
truth table	sandhedstabel
union (of sets)	foreningsmængde
unit circle	enhedscirklen
upper triangular matrix	øvre trekantsmatrix
vector space	vektorrum
vertex (vertices)	hjørne(r), vinkelspids(er)
zero vector	nulvektoren

Dansk til Engelsk

Dansk	Engelsk
absolutværdi	absolute value, modulus (in case of complex numbers)
afledte funktion	derivative of a function
andenordens ligning	second order equation
antiderivative	stamfunktion
argument (af et komplekst tal)	argument (of a complex number)
associativ	associative
bestemt integral	definite integral
biimplikation	biimplication
bijektiv, bijektion	bijective, bijection
billede (af en funktion)	image
binome ligning, den	the binomial equation
brøktal	fraction
definitionsmængde	domain (of a function)
delvis integration	partial integration, integration by parts
diagonal	diagonal
diagonalmatrix	diagonal matrix
disjunkt	disjoint
dispositionsmængde	co-domain
dobbeltrod	double root, root of multiplicity two
enhedscirklen	unit cirlce
fuldstændige løsning	general solution
foreningsmængde	union (of sets)
fællesmængde	intersection (of sets)
grad	degree
grænsværdi	limit
heltal	integer
hjørne(r)	vertex (vertices)
hovedargument (af et komplekst tal)	principal value of the argument (of a complex number)
identitetsmatrix	identity matrix
implikation	implication
injektiv	injective
inverst element	inverse element
koefficientmatrix	coefficient matrix (of a system of linear equations)
kommutativ	commutative
komplekst tal	complex number
komplekse talplan, det	the complex plane

Dansk	Engelsk
kompleks konjugerede	complex conjugate
kontinuert funktion	continuous function
koordinatvektor	coordinate vector
kvadratisk matrix	square matrix
kvadratrod	square root
kvotientreglen	quotient rule
kædereglen	chain rule
legeme	field
lige tal	even number
ligning	equation
linearkombination	linear combination
lineært (u)afhængig	linearly (in)dependent
logisk udsagn	logical proposition
logisk ækvivalent	logically equivalent
løsning til en ligning	solution for an equation
matrix, matricer	matrix, matrices
modulus	modulus
monisk	monic
mængde	set
naturligt tal	natural number
nedre trekantsmatrix	lower triangular matrix
nulvektor	zero vector
numerisk værdi	absolute value, modulus (in case of a complex number)
partikulær løsning	particular solution
partiel integration	partial integration, integration by parts
polynom, polynomium	polynomial
polære koordinater	polar coordinates
produktreglen	product rule
rang	rank (of a matrix)
reduceret trappeform	reduced row echelon form
reelt tal	real number
reelle tallinje, den	the real line
rod i et polynomium	root of a polynomial
række (af en matrix)	row
rækkevektor	row vector
sammensat, sammensætning	composite, composition
sandhedstabel	truth table
skalar	scalar

Dansk	**Engelsk**
stamfunktion	antiderivative, primitive function
surjektiv	surjective
søjle (af en matrix)	column
søjlevektor	column vector
totalmatrix	augmented matrix (of a system of linear equations)
trappeform	row echelon form
trappematrix	matrix in echelon form
transponerede	transpose (of a matrix)
ubekendte	indeterminate
udsagnslogik	propositional logic
udspænding	span
udvikling af determinanten	expansion of the determinant (along a row or column)
ulige tal	odd number
urbillede	preimage
vektorrum	vector space
vinkelspids, vinkelspidser	vertex, vertices
værdimængde (af en funktion)	image
øvre trekantsmatrix	upper triangular matrix

Index