Note 2

Sets and functions

2.1 Sets

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

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

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

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

Saying that a is an element of A is expressed as: $a \in A$. Some authors prefer to write the set first and then the element, writing $A \ni a$ instead of $a \in A$. If an element a is not in the set A, one can use the negation from propositional logic and write $\neg(a \in A)$. It is also common though to write $a \notin A$ for the statement that a is not an element of the set A. If two elements are in the same set, say $a_1 \in A$ and $a_2 \in A$, it is common to write $a_1, a_2 \in A$.

|||| Example 2.1

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

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

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

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

|||| Example 2.2

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

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

|||| Example 2.3

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

|||| Example 2.4

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

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

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

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

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

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

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

and

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

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

It is also customary to define

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

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

and

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

Example 2.5

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

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

Lemma 2.6

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

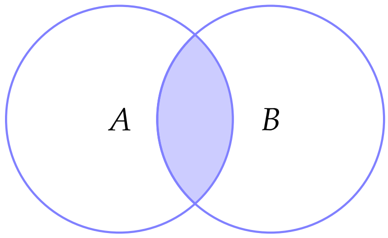
Proof. The statement $A = B$ for two sets A and B, is logically equivalent to the statement $a \in A \Leftrightarrow a \in B$ for all a. Using Equation (1-22), we can split the bi-implication up in two implications. Then we obtain the logically equivalent statement $(a \in A \Rightarrow a \in B) \quad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

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

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

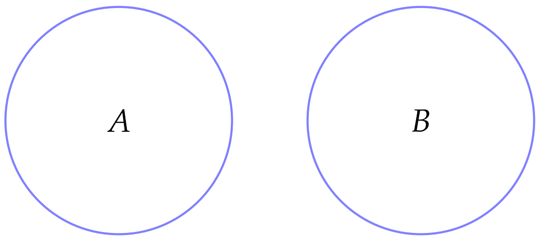
$$A \cap B = \{a \mid a \in A \wedge a \in B\}. \tag{2-1}$$

$A \cap B$



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

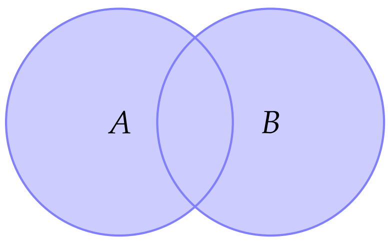
A and B disjoint



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

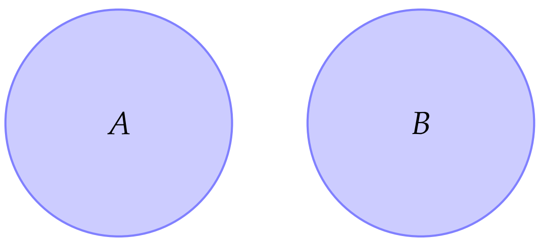
$$A \cup B = \{a \mid a \in A \vee a \in B\}. \tag{2-2}$$

$A \cup B$



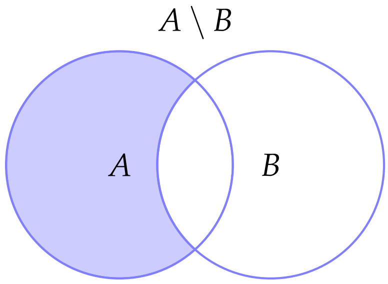
The union $A \cup B$ is called a *disjoint union* of A and B if $A \cap B = \emptyset$.

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



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

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



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

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

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

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

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

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

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

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

|||| Example 2.7

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

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

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

|||| Theorem 2.8

Let A, B and C be sets. Then

$$
\begin{aligned}
A \cap A &= A && \text{(2-4)}\\
A \cup A &= A && \text{(2-5)}\\
A \cup B &= B \cup A && \text{(2-6)}\\
A \cap B &= B \cap A && \text{(2-7)}\\
A \cup (B \cup C) &= (A \cup B) \cup C && \text{(2-8)}\\
A \cap (B \cap C) &= (A \cap B) \cap C && \text{(2-9)}\\
A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) && \text{(2-10)}\\
A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) && \text{(2-11)}
\end{aligned}
$$

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

|||| Example 2.9

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

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

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 2.10 that image and co-domain do not have to be equal. Common notations for the image of a function $f : A \to B$ are $f(A)$ or $\mathrm{image}(f)$. Let us consider some examples:

|||| Example 2.10

Let us again consider the function

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

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 2.6, it is enough to show that $f(\mathbb{R}) \subseteq \mathbb{R}_{\geq 0}$ and $\mathbb{R}_{\geq 0} \subseteq f(\mathbb{R})$.

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

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

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

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

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

|||| Example 2.11

Consider the functions

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

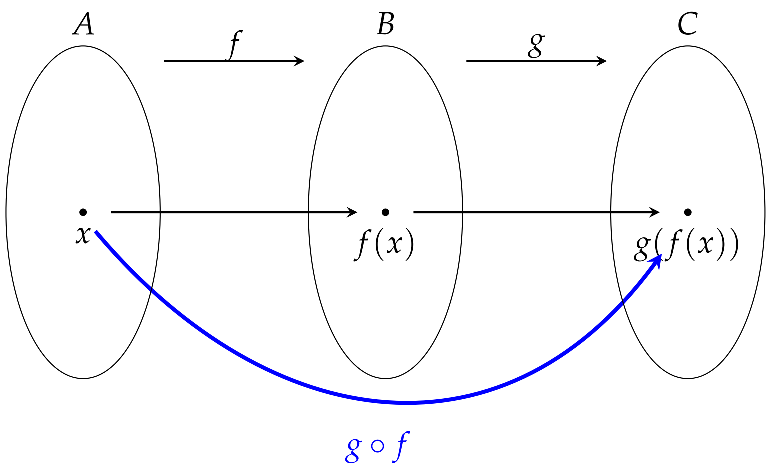


Figure 2.1: composition of the functions $f : A \to B$ and $g : B \to C$

> and
>
> $$\begin{array}{rcl} g : \{0,1\} & \to & \{0,1\} \\ a & \mapsto & a^2 \end{array}$$
>
> The functions f and g have the same domain and co-domain. Moreover, $f(0) = 0$, $f(1) = 1$, while $g(0) = 0^2 = 0$ and $g(1) = 1^2 = 1$. Hence $f = g$.
>
> This example shows that two functions may be the same even if they are described using different formulas.

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

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

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

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

Example 2.12

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

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

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

Lemma 2.13

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

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

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

while

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

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

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

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

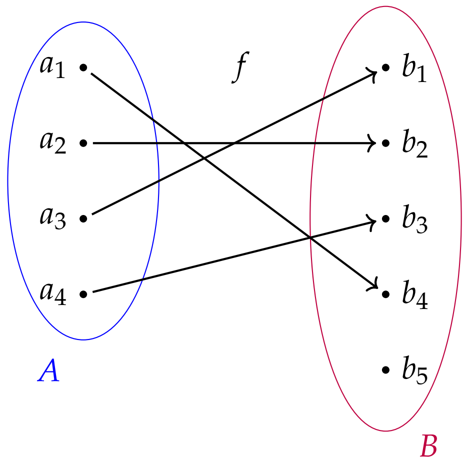


Figure 2.2: injective function $f : A \to B$

logical expressions, this means that:

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

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

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

This reformulation can be convenient in practice.

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

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

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

|||| Example 2.14

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

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

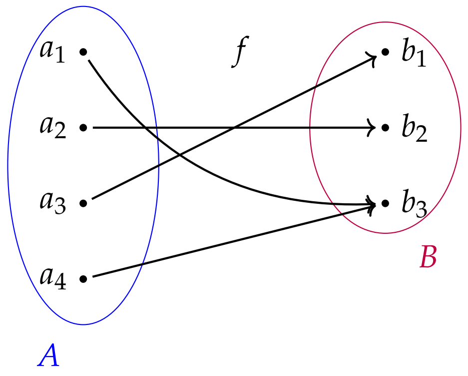


Figure 2.3: surjective function $f : A \to B$

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

|||| Example 2.15

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

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

|||| Definition 2.16

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

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

|||| Lemma 2.17

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

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

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

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

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

|||| Example 2.18

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

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

2.2.1 Computational aspects of functions

The way we have looked at a function $f : A \to B$, we completely ignored more practical aspects like: given some $a \in A$, how do you actually compute $f(b)$? For the general mathematical theory of functions, this is not an issue and the "inner workings" of the function f are then treated as a black box. However, for applications of the theory, it can be very important to know how to compute function values.

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

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

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

Step 3. Return the value from Step 2.

A bit more formally, we can rewrite this as:

Step 0. Denote by x the given input.

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

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

Step 3. Return z.

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

Algorithm 1 for $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$

Input: $x \in \mathbb{R}$
1: $y \leftarrow x \cdot x$
2: $z \leftarrow x \cdot y$
3: **return** z

Let us consider another example:

|||| Example 2.19

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

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

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

Algorithm 2 to compute $|x|$ for $x \in \mathbb{R}$

Input: $x \in \mathbb{R}$
if $x < 0$ **then**
    return $-x$
else
    return x

2.3 Examples of functions

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

|||| Lemma 2.20

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

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

or

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

Then f is an injective function.

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

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

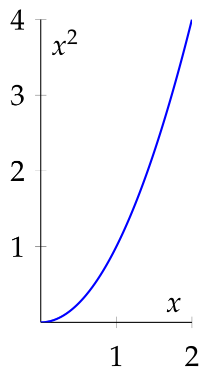
|||| Example 2.21

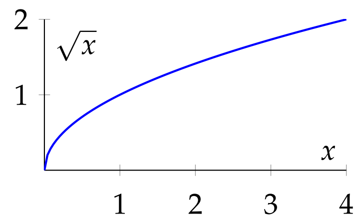
Consider the function $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = x^2$. We have already seen in Example 2.10 that the image of this function equals $\mathbb{R}_{\geq 0}$. In other words, $f(\mathbb{R}) = \mathbb{R}_{\geq 0}$. The function f is therefore not surjective. In fact, it is not injective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

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

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

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

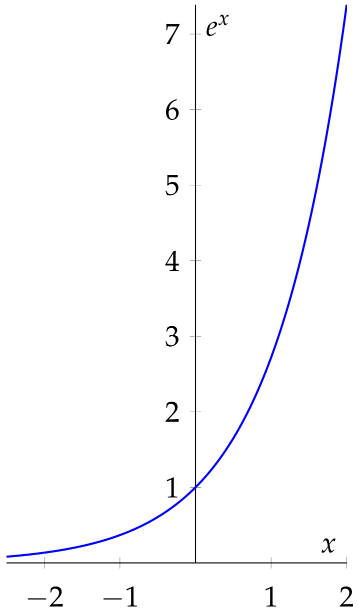


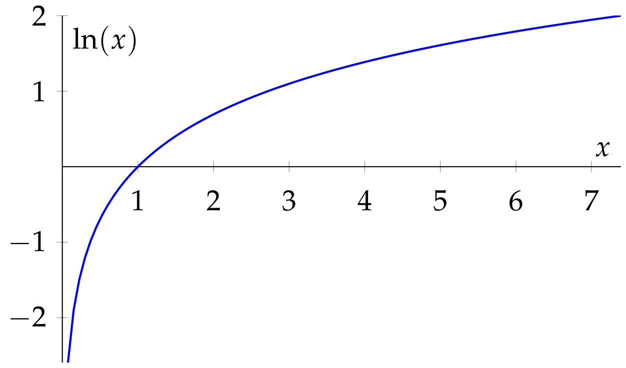


|||| Example 2.22

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

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

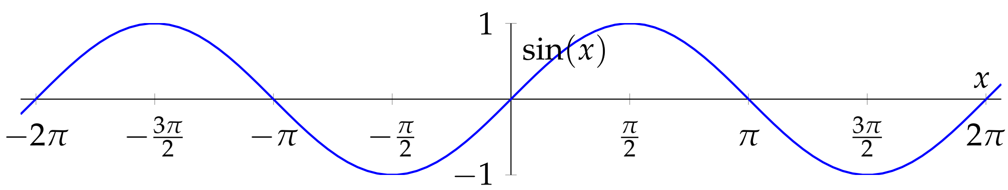




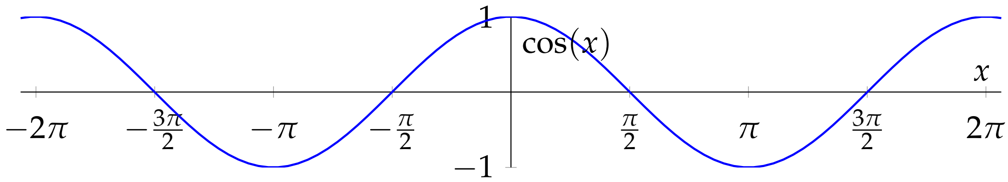
2.3.1 The trigonometric functions sin, cos and tan.

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

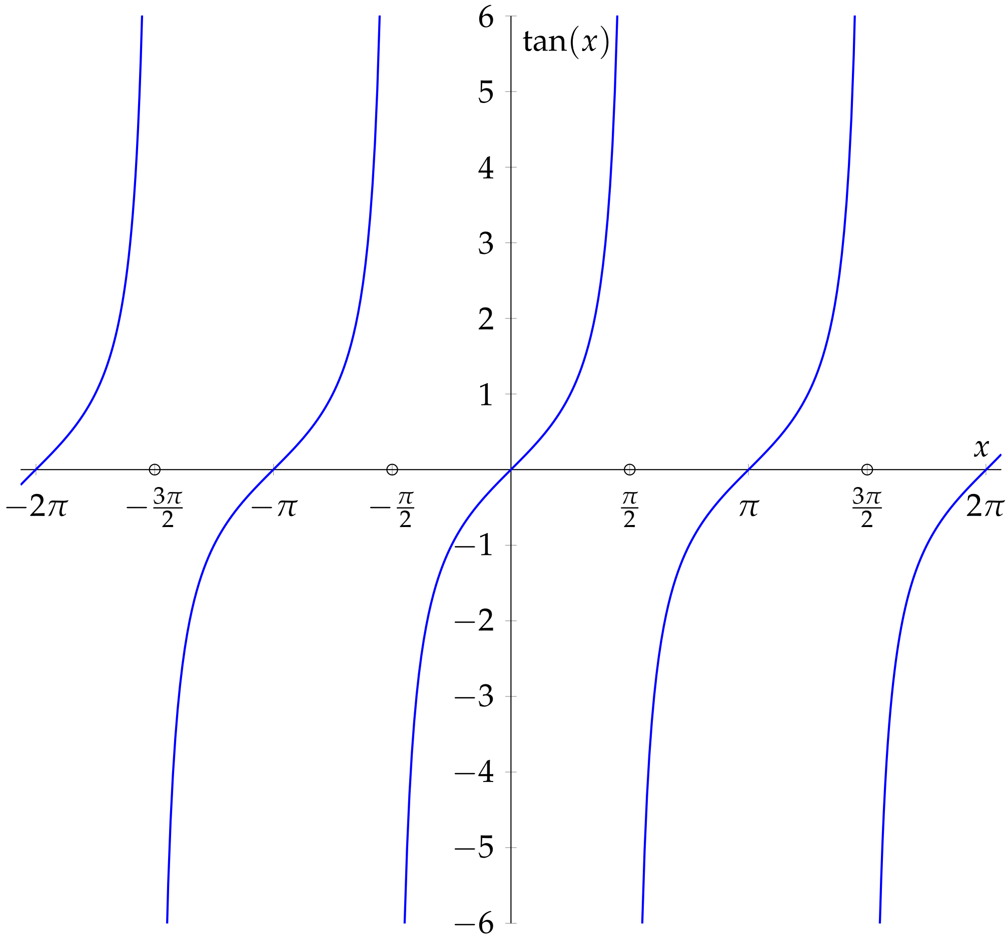
First of all, the sine function is usually denoted by sin, but let us in light of our definition of functions specify which domain and co-domain it has. First of all, we define the sine function $\sin : \mathbb{R} \to [-1, 1]$ to be the function such that $x \mapsto \sin(x)$. The image of sin is $[-1,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



Similarly, we define $\cos : \mathbb{R} \to [-1, 1]$. Again, the co-domain is chosen to be the closed interval $[-1, 1]$, which means that the function cos will be surjective. It is not injective though, since for example $\cos(-\pi/2) = \cos(\pi/2) = 0$. The graph of the cosine function is:



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

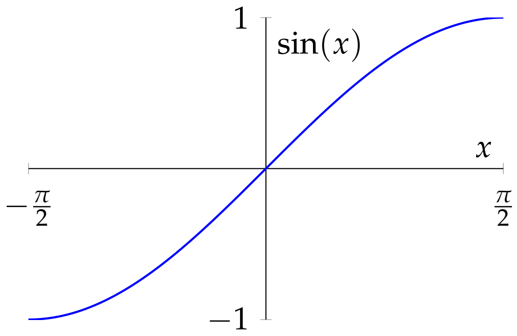


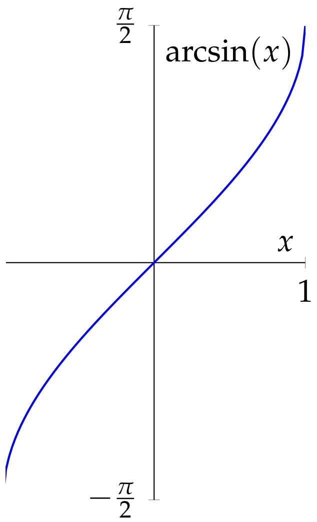
The small circles on the x-axis indicate the values of x for which the tangent function is not defined. The tangent function is surjective, since its image is $\mathbb{R}$. Just as the sine and cosine functions, it is not injective. We have for example $\tan(0) = 0$, but also $\tan(\pi) = 0$.

2.3.2 The inverse trigonometric functions

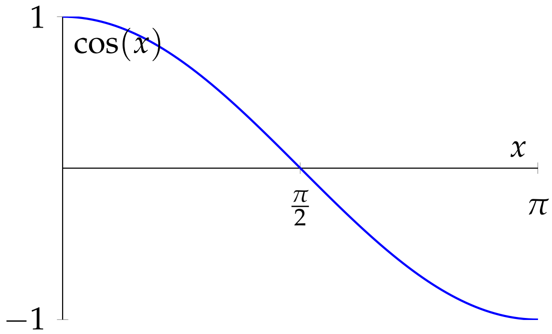
Since none of the trigonometric functions sin, cos and tan discussed in the previous subsection are bijections, we cannot find inverses for these functions. However, just as in Example 2.21, we can modify the domain of these functions and obtain functions that do have an inverse. These inverses are known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

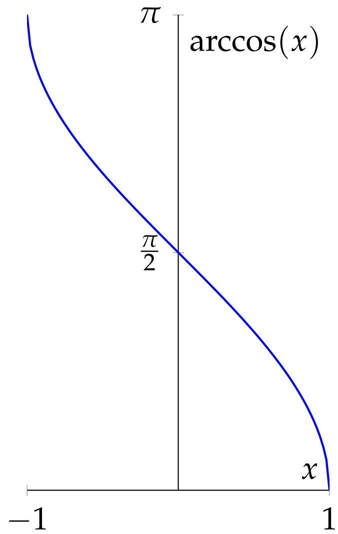
First of all, if the domain of the function $\sin : \mathbb{R} \to [-1, 1]$ is restricted to the closed interval $[-\pi/2, \pi/2]$, one obtains a function $f : [-\pi/2, \pi/2] \to [-1, 1]$ defined by $f(x) = \sin(x)$. The function f is a bijective function, since the graph of the sine function is strictly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



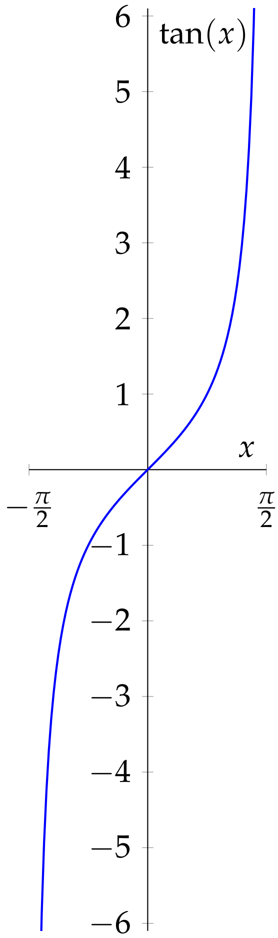


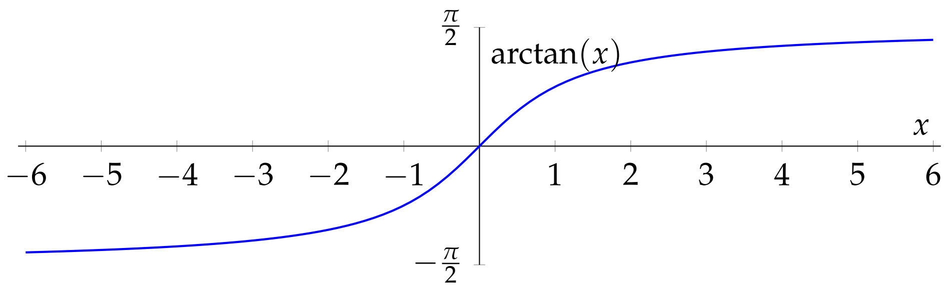
In a very similar way, we can define the *arccosine* function. First we restrict the domain of the usual cosine function to the closed interval $[0, \pi]$. The resulting function $g : [0, \pi] \to [-1, 1]$, where $g(x) = \cos(x)$, is strictly decreasing as well as surjective and thus bijective. The inverse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





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





|||| Example 2.23

Let us determine some values of the inverse trigonometric functions. Since $\sin(0) = 0$, we have $\arcsin(0) = 0$. However, even though $\sin(\pi) = 0$, we do not have $\arcsin(0) = \pi$. Indeed a function cannot take two distinct values for the same input! The issue is that arcsin is the inverse of the sine function with domain restricted to $[-\pi/2, \pi/2]$. Therefore $\sin(x) = y$ only implies $\arcsin(y) = x$ as long as $x \in [-\pi/2, \pi/2]$. For example, since $\sin(\pi/4) = \sqrt{2}/2$, we have $\arcsin(\sqrt{2}/2) = \pi/4$.

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

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