# Some Notes for the Direct Entrants Course

Michael Dreher, 2015/16, MACS, Heriot-Watt University Edinburgh

Recommended reading speed: at least four hours per chapter

# Chapter 1

# Limits and Continuity

## Purpose

Limits of sequences and functions are *the* main idea of *analysis*, and analysis is a certain branch of mathematics that justifies the computations which you did in calculus at school. Later we will get an understanding that analysis does a bit more.

## Limits of Sequences

**Example 1.** *We have* $\lim_{n\to\infty} \frac{n+1}{n} = 1$. *We can guess this by testing with* $n = 100$, $n = 1000$, $n = 10^6$ *and computing mentally.*

**Definition 1.1 (Only a first attempt).** *Let* $(a_0, a_1, a_2, \dots) = (a_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers.*
*We say that a real number* $A$ *is the limit of the sequence* $(a_n)_{n\in\mathbb{N}}$ *(and we write this as* $A = \lim_{n\to\infty} a_n$*) if the following holds:*
*We can make the distance* $|a_n - A|$ *as small as we want for all* $n$ *above some advantageously large chosen threshold.*

Obviously, $|a_n - A|$ is the distance of the real numbers $a_n$ and $A$. The last sentence (with the colours) is much too vague and does not follow scientific standards.

**Definition 1.2 (Now properly).** *Let* $(a_0, a_1, a_2, \dots) = (a_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers.*
*We say that a real number* $A$ *is the limit of the sequence* $(a_n)_{n\in\mathbb{N}}$ *(and we write this as* $A = \lim_{n\to\infty} a_n$*) if the following holds:*
*For each real positive* $\varepsilon$
*it exists a natural number* $N_0$ *(perhaps depending on* $\varepsilon$*)*
*such that for all* $n \geq N_0$
*it holds that* $|a_n - A| < \varepsilon$.
*We rewrite this much shorter as follows:*

$$\forall \varepsilon > 0\colon\ \exists N_0(\varepsilon) \in \mathbb{N}\colon\ \forall n \geq N_0(\varepsilon)\colon\ |a_n - A| < \varepsilon.$$

**Remark 1.3.** *The* $\forall$ *and* $\exists$ *are read as "for all" and "it exists". The colon ":" starts a new sub-sentence. We have various layers of sub-sentences inside each other, like the layers of an onion.*
*The interesting case for* $\varepsilon$ *is when* $\varepsilon$ *is small. Nobody cares about large* $\varepsilon$ *(say,* $\varepsilon \geq 1$*). The natural number* $N_0$ *is the threshold value mentioned in the naive definition.*
*When proving a statement* $\lim_{n\to\infty} a_n = A$*, your task is to present a recipe that computes* some $N_0$ *from a given* $\varepsilon$ *such that the colour line becomes true.*

**Example 2.** *We have guessed* $\lim_{n\to\infty} \frac{n+1}{n} = 1$*, but a proof is still missing. Here it comes.*
*Let a positive number* $\varepsilon$ *be given to us. Our job is to find* one *natural number* $N_0$ *(which is allowed to depend on* $\varepsilon$*) such that for all* $n \geq N_0$ *we have* $|\frac{n+1}{n} - 1| < \varepsilon$*. We think of* $N_0$ *as a threshold value: we don't care what* $a_n$ *does for* $n \leq N_0$*. But each* $n$ *above this threshold value* $N_0$ *possesses a certain property.*
*Clearly,* $|\frac{n+1}{n} - 1| = |\frac{1}{n}| = \frac{1}{n}$*, since* $n$ *is positive. A valid* $N_0$ *therefore is* $N_0(\varepsilon) = \frac{1}{\varepsilon} + 1$ *(rounded up to the next integer). Another valid value for* $N_0$ *is* $\frac{2}{\varepsilon} + 83$ *(rounded up to the next integer). It is not needed to the "the optimal* $N_0$*". Our job has only been to find* some $N_0$*.*

## Limits of Functions

Limits of functions can be defined similarly to limits of sequences, but unfortunately, there is a twist.
Consider the function

$$f(x) = \begin{cases} 7 & \text{if } x = 2 \\ 5 - x & \text{if } x \neq 2. \end{cases} \tag{$\clubsuit$}$$

We observe from its graph: if $x$ runs towards 2 (from the left or from the right), then $f(x)$ runs towards 3 (from the top or from the bottom), but $f(2)$ is something completely different, namely 7. We have $\lim_{x\to 2} f(x) = 3$, the definition of which follows now.

**Definition 1.4.** *Let* $f = f(x)$ *be a function from* $\mathbb{R}$ *into* $\mathbb{R}$*. We say that a real number* $A$ *is the limit of the function* $f$ *at the point* 2 *(and write this as* $A = \lim_{x\to 2} f(x)$*) if the following holds:*
*For each real positive* $\varepsilon$
*it exists a positive real number* $\delta$ *(perhaps depending on* $\varepsilon$*)*
*such that for all* $x$ *with* $|x - 2| < \delta$ *but* $x \neq 2$
*it holds that* $|f(x) - A| < \varepsilon$*.*
*In compressed notation:*

$$\forall \varepsilon > 0\colon\ \exists \delta > 0\colon\ \forall x \text{ with } 0 < |x - 2| < \delta\colon\ |f(x) - A| < \varepsilon.$$

The twist announced earlier is that we must forbid $x = 2$. If we allowed[1] $x = 2$ then our function $f$ from ($\clubsuit$) would not possess any limit at the point 2.

**Remark 1.5.** *The relevant* $\varepsilon$ *and* $\delta$ *are typically small. Nobody cares about large* $\varepsilon$ *or large* $\delta$*. When proving a statement of the form* $\lim_{x\to x_*} f(x) = A$*, our task is to present a recipe that calculates some positive* $\delta$ *from a given positive* $\varepsilon$*, such that the coloured statement becomes true.*

**Example 3.** *We have guessed* $\lim_{x\to 2} f(x) = 3$ *for* $f$ *from* ($\clubsuit$)*, and now we prove it.*
*Let a positive real number* $\varepsilon$ *be given to us. Our job is to find* some $\delta$ *(that is allowed to depend on* $\varepsilon$*) such that for all* $x \neq 2$ *that differ from* 2 *by less than* $\delta$ *(which means* $0 < |x - 2| < \delta$*) possess the property* $|f(x) - 3| < \varepsilon$*.*
*Provided* $x \neq 2$*, we have* $|f(x) - 3| = |(5 - x) - 3| = |2 - x| = |x - 2|$*, but now we may exploit that* $|x - 2| < \delta$*, and our desire is* $|f(x) - 3| < \varepsilon$*. The number* $\varepsilon$ *is given, and* $\delta$ *is wanted. After some thinking we find the recipe: choose* $\delta$ *as* $\delta := \varepsilon$*. We could choose* $\delta$ *also smaller than* $\varepsilon$*.*

Instead of defining $\lim_{x\to 2} f(x)$, we may define $\lim_{x\to x_*} f(x)$ for a real number $x_*$ (instead of 2) provided we have fixed that $x_*$ before. Just substitute $x_*$ for 2 in the definition.

[1] subjunctive mood of irreality

## Continuity of Functions

**Definition 1.6.** *We say that a function $f = f(x)$ is continuous at a point $x_*$ if each of the following conditions hold:*

- *$f$ is defined at $x_*$ (which means that $f(x_*)$ exists)*
- *$\lim_{x\to x_*} f(x)$ also exists*
- *both are equal (which means $f(x_*) = \lim_{x\to x_*} f(x)$).*

The function $f$ from (♣) is *not* continuous at $x_* = 2$, but continuous for all other $x_*$.
We present another description of continuity of a function $f$ at a point $x_*$:

**Lemma 1.7.** [2] *A function $f = f(x)$ is continuous at a point $x_*$ if and only if the following condition holds:*

$$\forall \varepsilon > 0\colon\ \exists \delta > 0\colon\ \forall x \text{ with } |x - x_*| < \delta\colon\ |f(x) - f(x_*)| < \varepsilon.$$

That is just a re-wording of the definition, and the proof is not very hard. Note that the twist is no longer needed. (WHY ?)
We have one more description of continuity of a function $f$ at a point $x_*$:

**Proposition 1.8.** [3]
*A function $f = f(x)$ is continuous at a point $x_*$ if and only if the following condition holds: For each sequence $(x_1, x_2, x_3, \dots)$ with $\lim_{n\to\infty} x_n = x_*$: the sequence $(y_1, y_2, y_3, \dots)$ of associated function values $y_n := f(x_n)$ has limit $f(x_*)$.*

This can be concisely written as

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right). \qquad (\heartsuit)$$

**Exercise 1.** *Show that the function $f$ from (♣) does* not *satisfy (♡). You may want to choose an $x$–sequence*

$$\left(2 + \frac{1}{2}, 2 + \frac{1}{3}, 2 + \frac{1}{4}, 2 + \frac{1}{5}, \dots\right)$$

*and another $x$–sequence*

$$(2, 2, 2, 2, \dots)$$

*and imagine a zip fastener.*

[2] A lemma is a little theorem

[3] A proposition is some result that is smaller than a theorem but bigger than a lemma

# Chapter 2

# On Taylor's Theorem

## Purpose

Taylor's theorem is useful because it allows to approximate complicated functions by easy functions, and it even enables us to estimate the error of approximation.
The mean value theorem of differentiation is a special case of Taylor's theorem.
Rolle's theorem is a theoretical tool, whose only purpose is to help us proving all the other theorems.
The extended mean value theorem is also a theoretical tool, which is crucial for proving Taylor's theorem and for proving l'Hospital's rule.
The pedagogical use of this note is to show how to write a mathematical text.

## Some other result

**Theorem 2.1** (**Continuous functions on compact sets**). *Let $[a,b]$ be an interval, and let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Then $f$ attains a smallest and biggest value over $[a,b]$. Expressed in symbols:*

$$\exists \overline{x} \in [a,b], \quad \exists \underline{x} \in [a,b]\colon \quad \forall x \in [a,b]\colon \quad f(\underline{x}) \le f(x) \le f(\overline{x}). \tag{$\spadesuit$}$$

We will need this theorem in the next sections.

## Differentiable functions on an interval

**Lemma 2.2.** *Let $f\colon [a,b] \to \mathbb{R}$ be differentiable on this interval, let $\underline{x}$ be an* ***interior*** *point of this interval, where $f$ attains a smallest value.*
*Then $f'(\underline{x}) = 0$.*

*Proof.* By definition, we have

$$f'(\underline{x}) = \lim_{x \to \underline{x}} \frac{f(x) - f(\underline{x})}{x - \underline{x}}.$$

Since $\underline{x}$ is an interior point, $x$ is able to approach $\underline{x}$ from the left, and to approach $\underline{x}$ from the right. Observe that

$$\frac{f(x) - f(\underline{x})}{x - \underline{x}} \quad \begin{cases} \le 0 & : x < \underline{x}, \\ \ge 0 & : x > \underline{x}, \end{cases}$$

Letting $x$ approach $\underline{x}$ from the left then tells us $f'(\underline{x}) \le 0$. And letting $x$ approach $\underline{x}$ from the right yields $f'(\underline{x}) \ge 0$. But the left-sided limit and the right-sided limit must coincide, which is only possible if

$$\lim_{x\to\underline{x}} \frac{f(x)-f(\underline{x})}{x-\underline{x}} = 0.$$

This was our goal. The proof is complete. □

**Remark 2.3.** *The requirement that $\underline{x}$ be an interior point of $[a,b]$ is essential. Think about the sine over $[0,\pi]$.*

**Remark 2.4.** *A similar lemma holds about an interior point $\overline{x}$ where $f$ attains a largest value.*

## Rolle's theorem and its conclusions

**Theorem 2.5 (Rolle's theorem).** *Let $f\colon[a,b]\to\mathbb{R}$ be differentiable, and let $f(a)=f(b)=0$. Then there is a point $\xi\in(a,b)$ with $f'(\xi)=0$.*
*Expressed in symbols:*

$$\exists\xi\in(a,b):\quad f'(\xi)=0.$$

*Proof.* By assumption, $f$ is differentiable. Therefore, $f$ is continuous.
Now we invoke the theorem about continuous functions on compact sets, and that theorem guarantees us the existence of a point $\overline{x}\in[a,b]$ and a point $\underline{x}\in[a,b]$ where $f$ attains a largest value and a smallest value[1].
We consider $\overline{x}$ first and $\underline{x}$ later (but only if needed).
Now we perform a proof by cases.

**Case A: $f(\overline{x})<0$:** this is impossible because we are permitted to choose $x=a$ in (♠), and then we arrive at the contradiction $0<0$.

**Case B: $f(\overline{x})>0$:** then $\overline{x}\neq a$, because $f(a)=0$. And $\overline{x}\neq b$, because $f(b)=0$. Consequently, $\overline{x}$ is an interior point of the interval $(a,b)$. Now we apply Remark 2.4 together with Lemma 2.2, resulting in $f'(\overline{x})=0$. We choose $\xi:=\overline{x}$.

**Case C: $f(\overline{x})=0$:** We refine our proof by cases.

**Case $\alpha$: $f(\underline{x})<0$:** then $\underline{x}\neq a$, because $f(a)=0$. And $\underline{x}\neq b$, because $f(b)=0$. Consequently, $\underline{x}$ is an interior point of the interval $(a,b)$. Now we apply Lemma 2.2, resulting in $f'(\underline{x})=0$. We choose $\xi:=\underline{x}$.

**Case $\beta$: $f(\underline{x})=0$:** then $f$ is taking the value zero everywhere on the interval $[a,b]$. We pick an arbitrary interior point and call it $\xi$.

**Case $\gamma$: $f(\underline{x})>0$:** this is impossible.

In each case, we have delivered a point $\xi$ with the desired properties. Other cases can not happen, the proof is complete. □

**Theorem 2.6 (Mean value theorem of differentiation).** *Let $f\colon[a,b]\to\mathbb{R}$ be differentiable. Then a point $\xi\in(a,b)$ exists with*

$$f'(\xi)=\frac{f(b)-f(a)}{b-a}.$$

*Proof.* We define an auxiliary function[2]

$$h(x):=f(x)-f(a)-\frac{f(b)-f(a)}{b-a}(x-a).$$

[1] Careful: it may happen that $\underline{x}$ or $\overline{x}$ are not interior points of $[a,b]$.

[2] it helps us, therefore we call it $h$

Two short calculations reveal $h(a) = 0$ and $h(b) = 0$. Now also the function $h$ is differentiable, because $f$ is. We apply Rolle's theorem to the function $h$, hence there is a point $\xi \in (a, b)$ with $h'(\xi) = 0$. However, we observe that

$$h'(x) = f'(x) - \frac{f(a) - f(b)}{b - a},$$

and substituting $\xi$ for $x$ here concludes the proof. □

**Theorem 2.7** (**Extended mean value theorem of differentiation**). *Let $f$ and $g$ be differentiable functions on an interval $[a, b]$, with $g'(x) \neq 0$ for all $x \in (a, b)$, and $g(b) - g(a) \neq 0$. Then there exists a point $\xi \in (a, b)$ with*

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

*Sketch of proof.* Apply Rolle's theorem to the auxiliary function

$$h(x) = f(x) - f(a) - \frac{f(b) - f(a)}{g(b) - g(a)} \cdot (g(x) - g(a)).$$

□

Now we come to the highlight of this note.

**Theorem 2.8** (**Taylor's theorem**). *Let the function $f: [a, b] \to \mathbb{R}$ be $(N+1)$ times differentiable on the interval $[a, b]$. Then the following holds: for all $x$, $x_0 \in (a, b)$, there is a point $\xi$ between $x$ and $x_0$ such that*

$$f(x) = \sum_{n=0}^{N} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n + R_N(x; x_0), \qquad R_N(x; x_0) = \frac{1}{(N+1)!} f^{(N+1)}(\xi) \cdot (x - x_0)^{N+1}.$$

*Proof.* Keep $x$ and $x_0$ fixed, and let $t$ be a variable running in the interval $(a, b)$. We define two functions[3]:

$$\begin{aligned} F(t) &:= f(x) - f(t) - f'(t) \cdot (x - t) - \frac{1}{2!} f''(t) \cdot (x - t)^2 - \ldots - \frac{1}{N!} f^{(N)}(t) \cdot (x - t)^N, \\ G(t) &:= \frac{(x - t)^{N+1}}{(N+1)!}. \end{aligned}$$

We apply the extended version of the mean value theorem, for the pair of functions $F$&$G$, but now on the interval $(x_0, x)$ instead of the interval $(a, b)$:

$$\frac{F(x) - F(x_0)}{G(x) - G(x_0)} = \frac{F'(\xi)}{G'(\xi)},$$

with some $\xi$ between $x$ and $x_0$.
We calculate the four items on the LHS:

$$\begin{aligned} F(x) &= 0, \\ F(x_0) &= f(x) - \sum_{n=0}^{N} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n, \\ G(x) &= 0, \\ G(x_0) &= \frac{(x - x_0)^{N+1}}{(N+1)!}. \end{aligned}$$

[3] for this step, a large amount of phantasy, bordering on ingenuity, is required

Hence we obtain

$$f(x) - \sum_{n=0}^{N} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n = F(x_0) = \frac{F'(\xi)}{G'(\xi)} \cdot G(x_0) = \frac{F'(\xi)}{G'(\xi)} \cdot \frac{(x - x_0)^{N+1}}{(N+1)!}.$$

It remains to evaluate $F'(\xi)$ and $G'(\xi)$. When we calculate $F'(t)$, we see that many terms cancel and only one term remains:

$$F'(t) = -\frac{1}{N!} f^{(N+1)}(t) \cdot (x - t)^N.$$

And we quickly find

$$G'(t) = -(N+1)\frac{(x-t)^N}{(N+1)!} = -\frac{(x-t)^N}{N!}.$$

Plugging in we then get

$$f(x) - \sum_{n=0}^{N} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n = \frac{-\frac{1}{N!} f^{(N+1)}(\xi) \cdot (x - \xi)^N}{-\frac{(x-\xi)^N}{N!}} \cdot \frac{(x - x_0)^{N+1}}{(N+1)!},$$

which resolves into

$$f(x) - \sum_{n=0}^{N} \frac{1}{n!} f^{(n)}(x_0) \cdot (x - x_0)^n = \frac{1}{(N+1)!} f^{(N+1)}(\xi) \cdot (x - x_0)^{N+1}.$$

But this is exactly the desired formula for the remainder $R_N$, which was our claim. □

**Remark 2.9.** *In case you struggle finding $F'(t)$, don't worry. The calculations are indeed a bit longer and require concentrated attention. That is why we do them. Consider the case $N = 2$ first. Consider the case $N = 3$ next.*

**Remark 2.10.** *We may choose even $N = 0$, and then Taylor's theorem boils down to the mean value theorem of differentiation.*

## l'Hospital's rule

**Theorem 2.11.** *Let $f$ and $g$ be differentiable functions on an interval $[a, b]$ with the following properties:*

$$\lim_{x \to a} f(a) = 0, \qquad \lim_{x \to a} g(x) = 0,$$
$$\forall x \in (a, b): \quad g'(x) \neq 0.$$

*Suppose that the limit*

$$\lim_{x \to a} \frac{f'(x)}{g'(x)}$$

*exists. Then also the limit*

$$\lim_{x \to a} \frac{f(x)}{g(x)}$$

*exists, and both are equal.*

*Proof.* Take some $x \in (a, b)$. We apply the extended version of the mean value theorem to the pair $f\&g$, and we choose the interval $(a, x)$:

$$\frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f'(\xi)}{g'(\xi)}.$$

Here $\xi$ is a certain unknown point between $a$ and $x$. Observe that $f(a) = g(a) = 0$. Hence we have

$$\frac{f(x)}{g(x)} = \frac{f'(\xi)}{g'(\xi)}, \qquad \text{for some} \quad \xi \in (a, x).$$

Now let $x$ run towards $a$. Then $\xi$ must also run towards $a$, because $\xi$ is squeezed between $a$ and $x$. We assumed that the limit of the RHS exists (for $\xi$ going to $a$). Then also the limit of the LHS must exist. □

**Remark 2.12.** *A little warning: there are nasty situations where* $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ *does not exist, but* $\lim_{x\to a} \frac{f(x)}{g(x)}$ *does exist. Then l'Hospital's rule cannot be applied (and is useless).*

## Applications

- Apply Taylor's theorem to the function $f(x) = \sqrt[3]{1+x}$ for $x \approx x_0 = 0$, with the choice $N = 1$. Use this formula to find $\sqrt[3]{1750}$ without calculator. How many digits of your calculation are reliable ? The relation $12^3 = 1728$, known from Gulliver's travels, might be useful here.

- The distance (direct line) between Fort Augustus and Urquhart Castle is 25 kilometres, the earth radius is 6370 kilometres. Due to the earth being a ball, the surface of Loch Ness is bent there, and a mountain of water is accumulated between Fort Augustus and Urquhart Castle. Determine the height of this mountain without calculator, but with an error of less than 10%.

# Chapter 3

# On the Construction of Complex Numbers

## Purpose

We shall define complex numbers in a mathematically correct way. Perhaps you have already heard that complex numbers are things like $3 - 4\mathrm{i}$, with 3 being called the *real part*, $-4$ being called the *imaginary part*, and i being defined as $\sqrt{-1}$. Every real number (the usual numbers you learned in school) can be seen as a complex number, because you are permitted to write $\pi + 0 \cdot \mathrm{i}$ instead of $\pi$.
That way of introducing complex numbers is troublesome for various reasons:

- it is mathematically delicate,
- it is philosophically wrong.

Defining $\mathrm{i} := \sqrt{-1}$ is mathematically delicate because $-1$ is a complex number, and taking square roots of complex numbers is tricky (because each complex number (assuming it is not zero) has two square roots, and which of the two square roots of $-1$ shall be i ?), and we should be more careful.
Defining $\mathrm{i} := \sqrt{-1}$ is also philosophically wrong, because a doing a definition means giving birth to a scientific term. And this newborn scientific term descends from other scientific terms that are an older generation and have been defined previously. In particular, we know that the older scientific terms *do exist.* And here the philosophical mistake has been done: we pretend to know what $\sqrt{-1}$ is, and that it exists as a rigorously defined scientific term.
The purpose of this note is to show you how to define complex numbers in the mathematically correct way. Moreover, in the mathematical programme, you will (sooner or later) get acquainted with various algebraic structures like group, semigroup, ring, field, vector space, topological space. The definition of such an algebraic structure always involves the following ingredients:

- a set of objects (the numbers, for instance),
- a collection of operations that you can apply to the objects,
- several rules that hold for the operations.

In this note, you will learn in detail these three ingredients of the algebraic structure *field of complex numbers*, which prepares you for your later studies.

## Looking again at the real numbers

The three ingredients of the *field of real numbers* are

**the set of real numbers:** e.g. 2 or 3.7 or $\sqrt{2}$ or $\pi$, but not $\infty$,

**a collection of operations:** there are three of them:

- **equality:** you take two numbers, ask if they are equal, and the result is one of the two states "true" and "false". The notation is $a = b$.
- **addition:** you take two numbers, add the second to the first number, and the result is again a real number. The notation is $a + b$.
- **multiplication:** you take two numbers, multiply the first by the second number, and the result is again a real number. The notation is $a \cdot b$.

In some sense, we also have subtraction and division, but they are introduced in the rules part.

**a collection of rules:** their effect is twofold. On the one hand, the rules restrict what you are able/allowed to do. For instance, we are not able to solve the equation $0 \cdot x = 17$ for $x$. On the other hand, the rules empower you: because we *are* permitted to morph $2 \cdot (x + y)$ into $2 \cdot x + 2 \cdot y$. The rules are the following:

- **$+$ is commutative:** for all real $a$ and $b$, we have $a + b = b + a$,
- **$+$ is associative:** for all real $a$, $b$ and $c$, we have $(a + b) + c = a + (b + c)$,
- **adding is reversible:** for each real $a$ and each real $b$, the equation $a + x = b$ has exactly one solution $x$. If you want, you may write $x = b - a$.
- **$+$ has a neutral element:** there is exactly one special real number (called 0) such that, for each real $a$, we have $a + 0 = a$.
- **$\cdot$ is commutative:** for all real $a$ and $b$, we have $a \cdot b = b \cdot a$,
- **$\cdot$ is associative:** for all real $a$, $b$ and $c$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$,
- **multiplying is almost always reversible:** for each real $a$ that is not equal to 0, and for each real $b$, the equation $a \cdot x = b$ has exactly one solution $x$. If you want, you may write $x = b/a$.
- **multiplying and adding are connected:** for each real numbers $a$, $b$, $c$, we have the equality $a \cdot (b + c) = a \cdot b + a \cdot c$.

The recipe of constructing complex numbers rigorously is the following:

**defining the objects:** we define complex numbers and write them in a funny way, in order to not confuse them with the numbers we already know.

**defining the operations:** we define operations that are acting upon the complex numbers, and again we write them in a funny way: $\boxminus$, $\boxplus$, $\boxdot$, in order not to confuse them with their real cousins $=$, $+$, $\cdot$.

**checking the rules:** we prove that these three operations obey the same list of rules mentioned above (provided that we use the new notation everywhere).

**observation:** we find that a certain subset of the set of all complex numbers behaves exactly as the set of real numbers does.

**changing the notation:** the duck test is this one: If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck. We will have found a subset of the complex numbers that behaves like the set of real numbers. Now is a good time to stop being so pedantic about the notation, replace $\boxminus$, $\boxplus$, $\boxdot$ by $=$, $+$, $\cdot$ everywhere, and consider every real number as a complex number. Pedantically speaking, a real number is not a complex number, but it can be seen as a complex number, which is enough for all purposes.

## Defining complex numbers

**Definition 3.1** (**Complex number**). *Let $a$ and $b$ be real numbers. Then the ordered pair $(a,b)$ is called a* complex number. *We say that $a$ is the* real part *of $(a,b)$, and $b$ is the* imaginary part *of $(a,b)$. The set of all such pairs is denoted by $\mathbb{C}$.*

As a formula, this means

$$\mathbb{C} := \{(a,b)\colon \quad a \in \mathbb{R}, \quad b \in \mathbb{R}\}.$$

We say "ordered pair" because the order matters: $(3,4)$ is not the same as $(4,3)$.
We also introduce the notation

$$a = \Re(a,b), \qquad b = \Im(a,b).$$

## Defining operations

**Definition 3.2** (**Equality**). *Let $(a,b)$ and $(c,d)$ be complex numbers. We say that they are* equal *if and only if $a = c$ and $b = d$. We write this equality as $(a,b) \boxminus (c,d)$.*

As a formula, this means

$$(a,b) \boxminus (c,d) \quad :\Longleftrightarrow \quad a = c \text{ and } b = d.$$

**Definition 3.3** (**Addition**). *Let $(a,b)$ and $(c,d)$ be complex numbers. We define their* sum *to be that complex number $(a,b) \boxplus (c,d)$ which equals $(a+c, b+d)$.*

As a formula, this means

$$(a,b) \boxplus (c,d) :\boxminus (a+c, b+d).$$

**Definition 3.4** (**Multiplication**). *Let $(a,b)$ and $(c,d)$ be complex numbers. We define their* product *to be that complex number $(a,b) \boxdot (c,d)$ which equals $(a \cdot c - b \cdot d, a \cdot d + b \cdot c)$.*

As a formula, this means

$$(a,b) \boxdot (c,d) :\boxminus (a \cdot c - b \cdot d, a \cdot d + b \cdot c).$$

Well, this notation is indeed cumbersome. Soon we will switch to a shorter one.

## Checking the rules

By means of several mathematical proofs, we show that:

$\boxplus$ **is commutative**

$\boxplus$ **is associative**

**the complex adding** $\boxplus$ **can be reversed**

$\boxplus$ **has a neutral element:** this is the number $(0,0)$,

$\boxdot$ **is commutative**

$\boxdot$ **is associative**

**the complex multiplying** $\boxdot$ **can be almost always reversed**

**complex multiplying and complex adding are connected**

The proofs are straight forward. To present an example, we prove the associativity of the multiplication (the other proofs are easier and nice homeworks).

**Lemma 3.5.** *The operation* $\boxdot$ *is associative. This means:*

$$\forall (a,b) \in \mathbb{C}, \quad \forall (c,d) \in \mathbb{C}, \quad \forall (e,f) \in \mathbb{C} \quad : \quad \Big((a,b) \boxdot (c,d)\Big) \boxdot (e,f) \boxminus (a,b) \boxdot \Big((c,d) \boxdot (e,f)\Big).$$

*Proof.* We compute the LHS, which is

$$\begin{aligned}\Big((a,b) \boxdot (c,d)\Big) \boxdot (e,f) &\boxminus \Big(a\cdot c - b\cdot d, a\cdot d + b\cdot c\Big) \boxdot (e,f) \\ &\boxminus \Big((a\cdot c - b\cdot d)\cdot e - (a\cdot d + b\cdot c)\cdot f, (a\cdot c - b\cdot d)\cdot f + (a\cdot d + b\cdot c)\cdot e\Big). \qquad (\spadesuit)\end{aligned}$$

And we compute the RHS, which is

$$\begin{aligned}(a,b) \boxdot \Big((c,d) \boxdot (e,f)\Big) &\boxminus (a,b) \boxdot \Big(c\cdot e - d\cdot f, c\cdot f + d\cdot e\Big) \\ &\boxminus \Big(a\cdot (c\cdot e - d\cdot f) - b\cdot (c\cdot f + d\cdot e), a\cdot (c\cdot f + d\cdot e) + b\cdot (c\cdot e - d\cdot f)\Big).\end{aligned}$$

Exploiting the fact that the real operations $+$, $\cdot$ are commutative and associative, we quickly check that this is the same as $(\spadesuit)$. The proof is complete. □

## Observation

After this hard work, we relax and play with some complex numbers:

$$(2,3) \boxplus (4,1) \boxminus (6,4), \qquad (2,3) \boxdot (4,1) \boxminus (5,14).$$

How about some numbers, with imaginary part equal to zero ?

$$(2,0) \boxplus (4,0) \boxminus (2+4, 0+0) \boxminus (6,0), \qquad (2,0) \boxdot (4,0) \boxminus (2\cdot 4 - 0\cdot 0, 2\cdot 0 + 0\cdot 4) \boxminus (8,0).$$

That looks similar to the real identities $2+4=6$ and $2\cdot 4 = 8$.
Our conjecture is: the complex operations $\boxminus$, $\boxplus$, $\boxdot$ behave on the subset of all those complex numbers $(a,0)$ with imaginary part equal to zero exactly in the same way as the real operations $=$, $+$, $\cdot$ behave on the set of real numbers. We just should replace the complex number $(a,0)$ by the real number $a$, the complex equality sign $\boxminus$ by the real equality sign $=$, the complex addition symbol $\boxplus$ by the real addition symbol $+$, and the complex multiplication symbol $\boxdot$ by the real multiplication symbol $\cdot$.

## Changing the notation

We are using the cumbersome notation for a last time. Take a complex number $(a, b)$. Observe that $(0, b) \boxed{=} (b, 0) \boxdot (0, 1)$. Hence we have

$$(a, b) \boxed{=} (a, 0) \boxplus (0, b) \boxed{=} (a, b) \boxplus (b, 0) \boxdot (0, 1).$$

We already announced that we prefer to replace $(a, 0)$ by $a$, and, by analogy, $(b, 0)$ by $b$.
Let us make the agreement

$$\mathrm{i} := (0, 1).$$

Then the above line $(a, b) \boxed{=} (a, b) \boxplus (b, 0) \boxdot (0, 1)$ becomes

$$(a, b) \boxed{=} a \boxplus b \boxdot \mathrm{i},$$

and if we finally substitute the operation symbols, we get

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

Now what is $\mathrm{i} \cdot \mathrm{i}$ ? By its very definition, i is just an abbreviation of $(0, 1)$. Hence the answer to the question is found by this calculation:

$$\mathrm{i} \cdot \mathrm{i} = (0, 1) \boxdot (0, 1) \boxed{=} (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) \boxed{=} (-1, 0),$$

and we agreed to write $(-1, 0)$ shorter as $-1$. The answer is

$$\mathrm{i} \cdot \mathrm{i} = -1.$$

However, we also have $(-\mathrm{i}) \cdot (-\mathrm{i}) = -1$, by a very similar calculation in the cumbersome notation. Consequently, the equation $z^2 = -1$ has at least two solutions, namely i and $-$i. It might possess even more solutions, who knows. An attempt to “define” $\mathrm{i} := \sqrt{-1}$ then is clearly questionable, because we have no way to specify which of the (at least) two square root candidates is the one we mean.

The definitions of the operations in the new notation are easier to remember, perhaps:

$$\begin{aligned} a + b \cdot \mathrm{i} = c + d \cdot \mathrm{i} \quad &:\Longleftrightarrow \quad a = c \text{ and } b = d, \\ (a + b \cdot \mathrm{i}) + (c + d \cdot \mathrm{i}) &:= (a + c) + (b + d) \cdot \mathrm{i}, \\ (a + b \cdot \mathrm{i}) \cdot (c + d \cdot \mathrm{i}) &:= (a \cdot c - b \cdot d) + (a \cdot d + b \cdot c) \cdot \mathrm{i}. \end{aligned}$$

The assumption of the author is that you know already these three lines, and now you have learned where do they come from.

# Chapter 4

# On the Exponential Function

## Purpose

We (attempt to) explain the mysteries of the exponential functions and highlight their key properties. We continue our training in reading proofs and doing mathematics rigorously. We enjoy beautiful mathematics.

## Key Properties

Euler's number $e$ is $e = 2.7182818284 59\ldots$, it is irrational, and the exponential function is

$$\exp(x) = e^x, \qquad x \in \mathbb{R}.$$

Its key properties are mostly these here:

$$\begin{aligned} e^{a+b} &= e^a \cdot e^b, \qquad a, b \in \mathbb{R}, \\ \frac{\mathrm{d}}{\mathrm{d}\,x} e^x &= e^x, \qquad x \in \mathbb{R}, \\ e^{\ln x} &= x, \qquad \ln(e^x) = x, \qquad x \in \mathbb{R}_+. \end{aligned}$$

The last line means that the natural logarithm ln is the inverse function to the exponential function. And the natural logarithm is beautiful because it has an exceptionally nice derivative (which one ???).

In this note, we try to answer four questions which you might have found yourself already (every teacher appreciates curious students who raise questions):

- where does this funny number 2.718281828459... come from ? Why didn't we select a simpler number like 2 or 10 ?
- assuming our pocket calculator is broken, how can we calculate $e^x$ if $e$ is the funny number from above, and $x$ is similarly complicated ? Calculating $e^3$ by hand is easy: you simply multiply $2.71828 \cdot 2.71828 \cdot 2.71828$ using pen and paper (which is principally doable if we have some time), and you take more digits if you need more digits. But if $x$ is more complicated, like $x = 2.34527910\ldots$, what exactly is $e^x$ supposed to be ? Already its meaning is a mystery.
- Can we do it with complex numbers $x$, ideally without investing too much work ?
- Can we find some beauty here ?

## The real case

Suppose you have 100£ in a bank account, and the bank pays 1% interest. If you keep the money there from 1 January till 31 December untouched, you will obtain at the end of the year $100 \cdot (1 + 0.01) = 101$£ .

If you withdraw the money on 30 June, the bank will pay you half the interest, so you will get $100 \cdot (1 + \frac{0.01}{2}) = 100.50$£ , and then you directly pay back this amount into your account again. Then this amount of 100.50£ will earn an interest of 0.5% for the remaining half of the year, giving you a total amount of $100 \cdot (1 + \frac{0.01}{2})^2 = 101.0025$£ on 31 December.

If you withdraw the money of 100£ (plus the earned interests) after a third of a year, you will receive $100 \cdot (1 + \frac{0.01}{3})$£ , and then you pay back this amount into your account again, and then you withdraw everything after two thirds of the year (with the interests), and then you pay what you received back into your account. On 31 December, you will have an amount of $100 \cdot (1 + \frac{0.01}{3})^3 = 101.003337037$£ . This is making us even richer (by 0.08337037p) than the midyear split approach.

The question is: how rich can we become following this scheme ? In mathematical terms: what is the value of

$$100 \cdot \lim_{n\to\infty} \left(1 + \frac{0.01}{n}\right)^n \ ?$$

The answer is: $100 \cdot e^{0.01} = 101.0050167\ldots$. We will explain it soon.

The funny number $e = 2.718281828459\ldots$ is *Banker's Constant.*

Therefore, we are interested in the limit

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n, \qquad x \in \mathbb{R},$$

and we do not know yet whether this limit even exists.

**Lemma 4.1.** *For each $x \in \mathbb{R}$, the limit*

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$$

*does exist, and its value is equal to $\sum_{n=0}^{\infty} \frac{1}{n!} x^n$, where this summation with infinitely many items is defined like this:*

$$\sum_{n=0}^{\infty} \tfrac{1}{n!} x^n := \lim_{N\to\infty} \left( \sum_{n=0}^{N} \tfrac{1}{n!} x^n \right). \qquad (\clubsuit)$$

*Proof.* Let $x \in \mathbb{R}$ be given (hence fixed). We will prove elsewhere that the limit on the RHS of ($\clubsuit$) exists. So (for today) we presume that this limit exists and give it the name $A_*$. Hence we know the following:

$$\forall\ \varepsilon > 0: \quad \exists\ N_0(\varepsilon): \quad \forall\ N' \geq N_0(\varepsilon): \quad \left| \left( \sum_{n=0}^{N'} \frac{1}{n!} x^n \right) - A_* \right| < \varepsilon. \qquad (\spadesuit)$$

By the definition of the limit of the sequence, we are required to prove the following:

$$\forall\ \varepsilon > 0: \quad \exists\ N_1(\varepsilon): \quad \forall\ n \geq N_1(\varepsilon): \quad \left| \left(1 + \frac{x}{n}\right)^n - A_* \right| < \varepsilon \qquad (\heartsuit)$$

A positive number $\varepsilon$ is given to us. Our job is to deliver a number $N_1(\varepsilon)$ with the desired property mentioned in ($\heartsuit$). Using the binomial formula $(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$, where $\binom{n}{k} = \frac{n!}{(n-k)!k!}$, we calculate

$$\begin{aligned}\left(1 + \frac{x}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} 1^{n-k} \left(\frac{x}{n}\right)^k = \sum_{k=0}^{n} \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{k!} \cdot \frac{x^k}{n^k} \\ &= \sum_{k=0}^{n} \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{n \cdot n \cdot \ldots \cdot n} \cdot \frac{x^k}{k!}.\end{aligned}$$

The first fraction behind the summation symbol has $k$ factors in the numerator and $k$ factors in the denominator, so we can simplify this to

$$\left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{n} 1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\left(1-\frac{3}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}.$$

Let us pick a special $k$, say $k=5$, and consider a special item of the sum:

$$1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\left(1-\frac{3}{n}\right)\cdot\left(1-\frac{4}{n}\right)\cdot\frac{x^5}{5!},$$

and this term converges (assuming $n\to\infty$) to $1\cdot 1\cdot 1\cdot 1\cdot 1\cdot \frac{x^5}{5!}$. This looks promising, but we are not done yet: although we may write

$$\lim_{n\to\infty}\left(\left(1+\frac{x}{n}\right)^n\right) = \lim_{n\to\infty}\left(\sum_{k=0}^{n}\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\left(1-\frac{3}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}\right),$$

we are not allowed to swap the $\lim_{n\to\infty}$ and the $\sum_{k=0}^{n}$ (because then we are in a situation where the identifier $n$ is used in an outer layer of the formula, but defined only in an inner layer of the formula, which is logically absurd).

The trouble is that the items in the sum $\sum_{k=0}^{n}$ are getting more and more as $n$ increases.

For the number $\varepsilon$ fixed in the very beginning, the line (♠) gives us a number $N_0(\frac{\varepsilon}{3})$. Let $\tilde{N}$ be a number $\geq N_0(\varepsilon/3)$ which will be specified later, and let us assume $n\geq\tilde{N}$. We will now split two summations at the index $\tilde{N}$. Hence we write

$$\begin{aligned}\left|\left(1+\frac{x}{n}\right)^n - A_*\right| &= \left|\left(\sum_{k=0}^{n}\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}\right)-\sum_{m=0}^{\infty}\frac{x^m}{m!}\right| \\ &= \left|\left(\sum_{k=0}^{\tilde{N}}(\ldots)+\sum_{k=\tilde{N}+1}^{n}(\ldots)\right)-\sum_{m=0}^{\tilde{N}}\frac{x^m}{m!}-\sum_{m=\tilde{N}+1}^{\infty}\frac{x^m}{m!}\right| \\ &\leq \left|\sum_{k=0}^{\tilde{N}}\left(\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}-\frac{x^k}{k!}\right)\right| \qquad (\diamond) \\ &\qquad + \left|\sum_{k=\tilde{N}+1}^{n}\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}\right| \\ &\qquad + \left|\sum_{m=\tilde{N}+1}^{\infty}\frac{x^m}{m!}\right|.\end{aligned}$$

Our desire is to prove that this complicated RHS is smaller than $\frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3}$, for all $n\geq N_1(\varepsilon)$, and we still have some freedom in choosing $N_1(\varepsilon)$. We start with the last item on the RHS ($\diamond$), because it is the easiest one:

$$\left|\sum_{m=\tilde{N}+1}^{\infty}\frac{x^m}{m!}\right| = \left|A_* - \sum_{m=0}^{\tilde{N}}\frac{x^m}{m!}\right| < \frac{\varepsilon}{3},$$

where we have exploited (♠) and tacitly assumed $N_1(\varepsilon)\geq N_0(\varepsilon/3)$. This is a condition on $N_1(\varepsilon)$. Now we consider the middle item on the RHS ($\diamond$). The triangle inequality permits us to pull the modulus bars into the summation:

$$\begin{aligned}\left|\sum_{k=\tilde{N}+1}^{n}\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}\right| &\leq \sum_{k=\tilde{N}+1}^{n}\left|\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}\right|\\ &= \sum_{k=\tilde{N}+1}^{n}\left|1-\frac{1}{n}\right|\cdot\ldots\cdot\left|1-\frac{k-1}{n}\right|\cdot\left|\frac{x^k}{k!}\right|\\ &\leq \sum_{k=\tilde{N}+1}^{n} 1\cdot\ldots\cdot 1\cdot\frac{|x|^k}{k!}\\ &= \sum_{k=\tilde{N}+1}^{n}\frac{|x|^k}{k!}.\end{aligned}$$

In the same way as we presumed that the summation $\sum_{n=0}^{\infty}\frac{1}{n!}x^n$ exists (and has a certain value which we called $A_*$), we may presume that also the summation $\sum_{n=0}^{\infty}\frac{1}{n!}|x|^n$ exists (and has a certain value $A_{|*|}$. Expressed as a formula:

$$\forall\ \varepsilon>0\colon\quad \exists\ N_{|0|}(\varepsilon)\colon\quad \forall\ N'\geq N_{|0|}(\varepsilon)\colon\quad \left|\left(\sum_{n=0}^{N'}\frac{1}{n!}|x|^n\right)-A_{|*|}\right|<\varepsilon. \qquad (|\spadesuit|)$$

Then we can continue our calculation like this:

$$\sum_{k=\tilde{N}+1}^{n}\frac{|x|^k}{k!}\leq\sum_{k=\tilde{N}+1}^{\infty}\frac{|x|^k}{k!}=\left|A_{|*|}-\sum_{k=0}^{\tilde{N}}\frac{|x|^k}{k!}\right|<\frac{\varepsilon}{3},$$

provided that $\tilde{N}\geq N_{|0|}(\varepsilon/3)$. We have exploited ($|\spadesuit|$) here.
Now we come to the first item on the RHS ($\diamond$). We again use the triangle inequality to pull the modulus bars into the summation:

$$\begin{aligned}\left|\sum_{k=0}^{\tilde{N}}\left(\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}-\frac{x^k}{k!}\right)\right| &\leq \sum_{k=0}^{\tilde{N}}\left|\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\cdot\frac{x^k}{k!}-\frac{x^k}{k!}\right|\\ &= \sum_{k=0}^{\tilde{N}}\left|\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)-1\right|\cdot\frac{|x|^k}{k!}.\end{aligned}$$

And here it is easy to perform the limit $n\to\infty$, because the number of items in the sum is never more than $\tilde{N}+1$, and there is no trouble with logics.
Let $k$ be fixed. We easily check

$$\lim_{n\to\infty}\left|\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)-1\right|\cdot\frac{|x|^k}{k!}=0.$$

Expressed as a formula:

$$\forall\ \varepsilon>0\colon\quad \exists\ N_{0,k}(\varepsilon)\colon\quad \forall\ n\geq N_{0,k}(\varepsilon)\colon\quad \left|\left(1-\frac{1}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)-1\right|\cdot\frac{|x|^k}{k!}<\varepsilon. \qquad (\triangle_k)$$

For each $k\in\{0,1,\ldots,\tilde{N}\}$, we have its own statement ($\triangle_k$) written here. These are $\tilde{N}+1$ statements.
Now let us assume

$$N_1(\varepsilon)\geq N_{0,0}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right),\quad N_1(\varepsilon)\geq N_{0,1}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right),\quad \ldots,\quad N_1(\varepsilon)\geq N_{0,\tilde{N}}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right),$$

and $n \geq N_1(\varepsilon)$. For such $n$, we then have

$$\sum_{k=0}^{\tilde{N}} \left| \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right) \right| \cdot \frac{|x|^k}{k!} < \sum_{k=0}^{\tilde{N}} \frac{\varepsilon}{3(\tilde{N}+1)} = \frac{\varepsilon}{3},$$

hence we can continue from ($\diamond$) like this:

$$\begin{aligned} &\left| \sum_{k=0}^{\tilde{N}} \left( \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right) \cdot \frac{x^k}{k!} - \frac{x^k}{k!} \right) \right| \\ &\qquad + \left| \sum_{k=\tilde{N}+1}^{n} \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right) \cdot \frac{x^k}{k!} \right| \\ &\qquad + \left| \sum_{m=\tilde{N}+1}^{\infty} \frac{x^m}{m!} \right| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned} \qquad (\diamond)$$

The recipe for chosing $N_1(\varepsilon)$ is this:

- choose $\tilde{N} = \max\{N_0(\varepsilon/3), N_{|0|}(\varepsilon/3)\}$ with $N_0$ specified in ($\spadesuit$) and $N_{|0|}$ specified in ($|\spadesuit|$);
- with this $\tilde{N}$, define

  $$N_1(\varepsilon) \coloneqq \max\left\{\tilde{N}, \quad N_{0,0}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right), \quad N_{0,1}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right), \quad \ldots, \quad N_{0,\tilde{N}}\left(\frac{\varepsilon}{3(\tilde{N}+1)}\right)\right\},$$

  where $N_{0,k}$ is specified in ($\triangle_k$).

The proof is complete. □

After this long proof, we now relax a bit and define the exponential function on $\mathbb{R}$:

**Definition 4.2.** *For $x \in \mathbb{R}$, we define*

$$\exp_{\mathbb{R}}(x) \coloneqq \sum_{n=0}^{\infty} \frac{1}{n!} x^n.$$

And to get a deeper understanding of the behaviour of this function, we prove a key property:

**Proposition 4.3.** *For all $x \in \mathbb{R}$ and all $y \in \mathbb{R}$, we have*

$$\exp_{\mathbb{R}}(x+y) = \exp_{\mathbb{R}}(x) \cdot \exp_{\mathbb{R}}(y).$$

*Proof.* We apply the definition of $\exp_{\mathbb{R}}$ to the LHS and to the RHS. Hence we have

$$\begin{aligned} LHS &= \sum_{n=0}^{\infty} \frac{1}{n!}(x+y)^n = \sum_{n=0}^{\infty} \frac{1}{n!} \left( \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \right) = \sum_{n=0}^{\infty} \frac{1}{n!} \left( \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} x^k y^{n-k} \right) \\ &= \sum_{n=0}^{\infty} \left( \sum_{k=0}^{n} \frac{x^k}{k!} \cdot \frac{y^{n-k}}{(n-k)!} \right). \end{aligned}$$

On the other hand,

$$RHS = \left( \sum_{j=0}^{\infty} \frac{1}{j!} x^j \right) \cdot \left( \sum_{m=0}^{\infty} \frac{1}{m!} y^m \right).$$

We wish to transform the LHS into the RHS. A natural transformation consists in swapping the summation symbols. This step is non-trivial because infinitely many terms in the sum are involved, and you will learn a justification of this swapping elsewhere in your maths programme. In any case, because of $0 \le k \le n$, we have $n \ge k$, hence we obtain

$$\begin{aligned}
LHS &= \sum_{k=0}^{\infty}\left(\sum_{n=k}^{\infty}\frac{x^k}{k!}\cdot\frac{y^{n-k}}{(n-k)!}\right) && \Big|\ \text{rename } m := n-k \\
&= \sum_{k=0}^{\infty}\left(\sum_{m=0}^{\infty}\frac{x^k}{k!}\cdot\frac{y^m}{m!}\right) && \Big|\ \frac{x^k}{k!} \text{ has no } m\text{, can be taken out} \\
&= \sum_{k=0}^{\infty}\frac{x^k}{k!}\cdot\left(\sum_{m=0}^{\infty}\frac{y^m}{m!}\right) && \Big|\ \left(\sum_{m\cdots}^{\cdots}\ldots\right) \text{ has no } k\text{, can be taken out} \\
&= \left(\sum_{m=0}^{\infty}\frac{y^m}{m!}\right)\cdot\sum_{k=0}^{\infty}\frac{x^k}{k!},
\end{aligned}$$

which is the same as the RHS, ignoring the different names of the running indices. □

**Example 4.** *We have*

$$\begin{aligned}
\exp_{\mathbb{R}}(7) &= \exp_{\mathbb{R}}(6+1) = \exp_{\mathbb{R}}(6)\cdot\exp_{\mathbb{R}}(1) = \exp_{\mathbb{R}}(5+1)\cdot\exp_{\mathbb{R}}(1) \\
&= \exp_{\mathbb{R}}(5)\cdot(\exp_{\mathbb{R}}(1))^2 = \ldots \\
&= (\exp_{\mathbb{R}}(1))^7 .
\end{aligned}$$

*Similarly, we get* $\exp_{\mathbb{R}}(m) = (\exp_{\mathbb{R}}(1))^m$ *for* $m \in \mathbb{N}$.

**Definition 4.4.** *We define* EULER*'s number*

$$e := \exp_{\mathbb{R}}(1) = \sum_{n=0}^{\infty}\frac{1}{n!} = \frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\ldots = 2.718281828459\ldots.$$

Hence we have $\exp_{\mathbb{R}}(m) = e^m$ for every $m \in \mathbb{N}$.

**Definition 4.5.** *For* $x \in \mathbb{R}$, *we agree to write* $e^x$ *as a short form of* $\exp_{\mathbb{R}}(x)$.

In this sense, the key property $e^{a+b} = e^a \cdot e^b$ has been proved for all real $a$ and all real $b$.

**Proposition 4.6.** *The function* $\exp_{\mathbb{R}}$ *is differentiable everywhere on* $\mathbb{R}$, *and its derivative is*

$$\frac{\mathrm{d}}{\mathrm{d}\,x}\exp_{\mathbb{R}}(x) = \exp_{\mathbb{R}}(x).$$

*Non–Proof.* We do not have the tools yet to give a proof. □

**Remark 4.7.** *The Maclaurin formula* $f(x) = \sum_{n=0}^{\infty}\frac{1}{n!}f^{(n)}(0)\cdot x^n$ *applied to* $f = \exp_{\mathbb{R}}$ *gives us*

$$\exp_{\mathbb{R}}(x) = \sum_{n=0}^{\infty}\frac{1}{n!}x^n,$$

*because of*

$$\left(\frac{\mathrm{d}}{\mathrm{d}\,x}\right)^n \exp_{\mathbb{R}}(x) = \exp_{\mathbb{R}}(x)$$

*and* $\exp_{\mathbb{R}}(0) = 1$. *This Maclaurin formula does not surprise us.*

## The complex case

We wish to perform similar investigations in the complex case, and our goal is to preserve as many properties of the real exponential function as possible. Therefore, we take the freedom to copy the results and their proofs preferably verbatim:

**Lemma 4.8.** *For each $z \in \mathbb{C}$, the limit*

$$\lim_{n\to\infty}\left(1+\frac{z}{n}\right)^n$$

*does exist, and its value is equal to $\sum_{n=0}^{\infty}\frac{1}{n!}z^n$, where this summation with infinitely many items is defined like this:*

$$\sum_{n=0}^{\infty}\frac{1}{n!}z^n := \lim_{N\to\infty}\left(\sum_{n=0}^{N}\frac{1}{n!}z^n\right).$$

The proof is literally the same as before, because of the following reasons:

- in the real proof, we have used the four basic operations $+$, $-$, $\cdot$, $/$, and their rules ($+$ and $\cdot$ are commutative and associative, etc.). But the complex versions of these four basic operations obey exactly the same rules.
- in the real proof, we have used the modulus function (in the calculations as well as in the definition of $\lim_{n\to\infty}$). There is also a modulus function in $\mathbb{C}$, and they are defined like this

  $$|x|_{\mathbb{R}} := \begin{cases} x & : x \ge 0, \\ -x & : x < 0, \end{cases} \qquad |x+y\mathrm{i}|_{\mathbb{C}} := \sqrt{x^2+y^2}.$$

  So they are defined differently. But this does not matter, because we only have used the following rules of a modulus function:

  - if $z \in \mathbb{C}$ (or $\in \mathbb{R}$) then $|z| \ge 0$,
  - if $|z| = 0$, then $z = 0$,
  - if $z$, $w \in \mathbb{C}$ (or $\in \mathbb{R}$), then $|z\cdot w| = |z|\cdot|w|$ and $|z+w| \le |z|+|w|$.

**Definition 4.9.** *For $z \in \mathbb{C}$, we define*

$$\exp_{\mathbb{C}}(z) := \sum_{n=0}^{\infty}\frac{1}{n!}z^n.$$

**Proposition 4.10.** *For all $z \in \mathbb{C}$ and all $w \in \mathbb{C}$, we have*

$$\exp_{\mathbb{C}}(z+w) = \exp_{\mathbb{C}}(z)\cdot\exp_{\mathbb{C}}(w).$$

*Proof. Mutatis mutandis* the same as in the real case. □

Now we wish to understand the complex exponential function slightly better.
By the key property, we have for $z = x + y\mathrm{i}$ with real $x$ and $y$ that

$$\exp_{\mathbb{C}}(z) = \exp_{\mathbb{C}}(x)\cdot\exp_{\mathbb{C}}(y\mathrm{i}) = \exp_{\mathbb{R}}(x)\cdot\exp_{\mathbb{C}}(y\mathrm{i}) = e^x\cdot\exp_{\mathbb{C}}(y\mathrm{i}).$$

Next we use the Maclaurin–formula for $\exp_{\mathbb{C}}(y\mathrm{i})$:

$$\begin{aligned}
\exp_{\mathbb{C}}(y\mathrm{i}) &= \sum_{n=0}^{\infty} \frac{1}{n!}(y\mathrm{i})^n \\
&= 1 + \frac{1}{1!}(y\mathrm{i}) + \frac{1}{2!}(y\mathrm{i})^2 + \frac{1}{3!}(y\mathrm{i})^3 + \frac{1}{4!}(y\mathrm{i})^4 + \frac{1}{5!}(y\mathrm{i})^5 + \frac{1}{6!}(y\mathrm{i})^6 + \dots \\
&= 1 + y\mathrm{i} - \frac{1}{2!}y^2 - \frac{1}{3!}y^3\mathrm{i} + \frac{1}{4!}y^4 + \frac{1}{5!}y^5\mathrm{i} - \frac{1}{6!}y^6 \pm \dots \\
&= \left(1 - \frac{1}{2!}y^2 + \frac{1}{4!}y^4 - \frac{1}{6!}y^6 + \frac{1}{8!}y^8 \mp \dots\right) \\
&\qquad + \left(y - \frac{1}{3!}y^3 + \frac{1}{5!}y^5 - \frac{1}{7!}y^7 \pm \dots\right) \cdot \mathrm{i} \\
&= \cos(y) + \sin(y) \cdot \mathrm{i},
\end{aligned}$$

because the big parantheses contain just the Maclaurin–formulas for $\cos(y)$ and $\sin(y)$.
If we now make the agreement of writing

$$e^z := \exp_{\mathbb{C}}(z), \qquad z \in \mathbb{C},$$

then we have

$$e^{x+y\mathrm{i}} = e^x \cdot (\cos y + \mathrm{i} \sin y), \qquad x \in \mathbb{R}, \quad y \in \mathbb{R}.$$

This allows us to reduce the complex exponential function $\exp_{\mathbb{C}}$ to easy real functions which we know from school.

## Mathematics is beautiful.

In the above identity, we choose $x = 0$ and $y = \pi$. Then we have

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

which boils down to the most beautiful formula of mathematics:

$$e^{\pi\mathrm{i}} + 1 = 0.$$

The beauty comes down from the fact that the five most important mathematical constants $e$, $\pi$, i, 1, 0 are compressed in one line.

There is more beauty. We have $e^{z+w} = e^z \cdot e^w$ for all $z \in \mathbb{C}$ and all $w \in \mathbb{C}$. This is just the key property of the exponential function. Let us take $z = \mathrm{i}\varphi$ and $w = \mathrm{i}\psi$ here with real $\varphi$ and real $\psi$. Then we have

$$e^{\mathrm{i}(\varphi+\psi)} = e^{\mathrm{i}\varphi} \cdot e^{\mathrm{i}\psi},$$

and the LHS equals $\cos(\varphi + \psi) + \mathrm{i} \sin(\varphi + \psi)$.
However, the RHS equals

$$(\cos \varphi + \mathrm{i} \sin \varphi) \cdot (\cos \psi + \mathrm{i} \sin \psi) = \Big( \cos \varphi \cos \psi - \sin \varphi \sin \psi \Big) + \mathrm{i}\Big( \cos \varphi \sin \psi + \cos \psi \sin \varphi \Big).$$

We learn a beautiful connection: the angle sum identities $\sin(\varphi + \psi) = \dots$ and $\cos(\varphi + \psi) = \dots$ are direct conclusions from the key property of the exponential function.

# Chapter 5

# On Polynomials

## Purpose

We give a pseudo-proof of the fact that every polynomial of degree $n$ has at least one complex zero. We then think again about division (with remainder). This enables us to prove that every polynomial of degree $n$ has even $n$ complex zeros (if you count them according to their multiplicity). We discuss greatest common divisors (first for numbers, then for polynomials).

## The Fundamental Theorem of Algebra

**Theorem 5.1** (**Fundamental theorem of algebra**). *Let $A(z) = \sum_{k=0}^{n} a_k z^k$ be a polynomial of degree n with complex coefficients. This means $a_n \neq 0$ and $a_k \in \mathbb{C}, \quad \forall k$.*
*Then this polynomial possesses at least one zero $z_0$:*

$$\exists\, z_0 \in \mathbb{C} \colon A(z_0) = 0.$$

*Pseudo–proof.* For didactical purposes, we only consider the polynomial

$$A(z) = z^4 - 2z^3 + 7z^2 + (3 + \mathrm{i}).$$

Choose a positive number $R = 1000$ and let $z \in \mathbb{C}$ run along the circle $\{z \in \mathbb{C} \colon |z| = R\}$ in the complex plane, in counter-clockwise direction, once. Then we can write

$$z = Re^{\mathrm{i}\varphi}, \qquad 0 \le \varphi < 2\pi,$$

and its fourth power then is $z^4 = 10^{12}e^{4\mathrm{i}\varphi}$, with $0 \le 4\varphi < 8\pi$, which means that $z^4$ runs along a circle with radius $10^{12}$ in counter-clockwise direction, four times.
If $z$ runs along the circle with radius $R$, where does $A(z)$ run ? To answer this question, we compare $A(z)$ and $z^4$:

$$\left|A(z) - z^4\right| = \left|-2z^3 + 7z^2 + 3 + \mathrm{i}\right| \le 2|z|^3 + 7|z|^2 + |3 + \mathrm{i}| \le 2 \cdot 10^9 + 7 \cdot 10^6 + 4 \le 10^{10},$$

which is much smaller (by a factor of at least 100) than $|z|^4 = 10^{12}$. We can say that $A(z) \approx z^4$ with a relative error of at most 1%. Therefore, we now consider an annular domain

$$\left\{w \in \mathbb{C} \colon 10^{12} - 10^{10} \le |w| \le 10^{12} + 10^{10}\right\},$$

and we know that $A(z)$ stays inside this narrow annulus, and $A(z)$ runs four times in counter-clockwise direction around the origin $w = 0$ of the complex plane (imagine a four-fold loop like a rubber ring).
Now we let $R$ decay continuously, starting from $R = 10^{12}$ until $R = \frac{1}{100}$, without doing jumps. What does the rubber ring do ? It seems clear that the rubber ring moves continuously without

doing jumps (here our proof turns into a pseudo-proof because we are far away from having precise definitions of the various funny words that we are using). At the end of this decay process, $R = \frac{1}{100}$, and then we have $A(z) \approx 3 + \mathrm{i}$ because of the following calculation:

$$|A(z) - (3 + \mathrm{i})| = \left|z^4 - 2z^3 + 7z^2\right| \le |z|^4 + 2|z|^3 + 7|z|^2 = \frac{1}{100^4} + \frac{2}{10^3} + \frac{7}{10^2} \le \frac{1}{2}.$$

Therefore, the final position of the rubber loop is contained in the disk

$$\left\{w \in \mathbb{C} \colon |w - (3 + \mathrm{i})| \le \tfrac{1}{2}\right\}.$$

Let us summarise: in the beginning, $R$ was $10^3$, and $A(z)$ was running along a rubber loop that goes four times around the origin. In the end, $R$ is $\frac{1}{100}$, and now $A(z)$ is running along a rubber loop that is completely contained in a disk that does not contain the origin. And inbetween, the rubber loop was moving continuously without jumps.
This is only possible if for some intermediate value of $R$, the origin lies on the rubber loop line. Then, for this special value of $R$, we have a number $z_0$ with $|z_0| = R$ such that $A(z_0) = 0$.
That was our goal. □

## Divisions with remainders

We think about the division $\frac{x^2+7}{x-2}$ and calculate it as we perhaps did in school:

$$\begin{array}{l} (x^2 \qquad\quad + 7) \quad : \quad (x - 2) = x + 2 \text{ with remainder } 11 \\ \underline{-(x^2 - 2x)} \\ \qquad\quad 2x + 7 \\ \qquad \underline{-(2x - 4)} \\ \qquad\qquad\quad 11 \end{array}$$

On the RHS, we first obtain the $x$, then the 2, then the remainder 11. The final formula then is

$$(x^2 + 7) = (x + 2) \cdot (x - 2) + 11.$$

We call this *division with remainder.*

**Lemma 5.2 (Division with remainder for polynomials).** *Let $A = \sum_{k=0}^{n} a_k z^k$ and $B(z) = \sum_{k=0}^{m} b_k z^k$ be polynomials with degree $n$ and $m$ respectively, where $m \ge 1$ and $a_n \ne 0$, $b_m \ne 0$. Then there are two polynomials $Q(z)$ and $R(z)$ satisfying the following two conditions:*

$$A(z) = Q(z) \cdot B(z) + R(z), \qquad \deg(R) < \deg(B).$$

*These two polynomials $Q$ and $R$ are* unique.

*Sketch of proof.* The uniqueness of $Q$ and $R$ is proved like this: suppose there are two other polynomials $\tilde{Q}$ and $\tilde{R}$ with the same properties. Then a subtraction gives

$$0 = \big(Q(z) - \tilde{Q}(z)\big) \cdot B(z) + \Big(R(z) - \tilde{R}(z)\Big).$$

Now $Q - \tilde{Q}$ is at either a non-zero constant, or it is a polynomial of degree $\ge 1$. Then the highest-order term of the product $(Q - \tilde{Q}) \cdot B$ has a degree of at least $m$, hence it cannot be cancelled by $R - \tilde{R}$ which as degree at most $m - 1$. Contradiction.
The existence of the polynomials $Q$ and $R$ is proved by mathematical induction on the degree of $A$. □

Now we go back to the polynomial $A(z) = \sum_{k=0}^{n} a_k z^k$. We know it has a zero $z_0 \in \mathbb{C}$. This means $A(z_0) = 0$. Choose $B(z) := z - z_0$. Then $\deg(B) = 1$. The Lemma on the division with remainder for polynomials then guarantees us the existence of two polynomials $Q(z)$ and $R(z)$ with

$$A(z) = Q(z) \cdot (z - z_0) + R(z), \qquad \deg(R) < 1.$$

Therefore, the degree of $R$ must be zero, hence $R(z)$ is a constant. Which constant ? To answer this, we substitute $z_0$ for $z$ and get

$$A(z_0) = Q(z_0) \cdot (z_0 - z_0) + R,$$

hence $R = 0$. The result then is $A(z) = Q(z) \cdot (z - z_0)$, where $Q$ is a certain polynomial with $\deg(Q) = n - 1$.

## The Fundamental Theorem of Algebra Reloaded

Let $A(z)$ be the above polynomial. We know that it possesses a zero which we are going to call $z_1$ instead of $z_0$, for reasons of beauty. Then we can write

$$A(z) = A_1(z) \cdot (z - z_1),$$

with some new polynomial $A_1$ (which was called $Q$ earlier). We know $\deg(A_1) = n - 1$. Now we apply the Fundamential Theorem of Algebra again, but now to $A_1$, which then has a zero which we call $z_2$, and this results in $A_1(z) = A_2(z) \cdot (z - z_2)$, for some new polynomial $A_2$ with $\deg(A_2) = n - 2$. And so on.

The final result then is

$$A(z) = a_n \cdot (z - z_1) \cdot (z - z_2) \cdot \ldots \cdot (z - z_n),$$

for certain $z_k \in \mathbb{C}$, and the factor $a_n$ comes from the highest term in $A(z) = \sum_{k=0}^{n} a_k z^k$.

We call this procedure *splitting a polynomial into linear factors*, because each factor $(z - z_k)$ is linear in $z$.

We observe that each polynomial of degree $n$ has $n$ complex zeros if you count them according to their multiplicity. For instance, $A(z) = (z - 7)^2 \cdot (z - 8)$ has the three zeros 7, 7, 8.

## Greatest common divisors for pairs of numbers

To find $\gcd(56, 12)$, we do repeated division with remainders:

$$\begin{aligned} 56 &= 4 \cdot 12 + 8, \\ 12 &= 1 \cdot 8 + 4, \\ 8 &= 2 \cdot 4 + 0. \end{aligned}$$

The last non-zero remainder is the desired greatest common divisor (which is Euclid's algorithm). Hence $\gcd(56, 12) = 4$. We can also express $\gcd(56, 12)$ using 56 and 12 working backwards:

$$\gcd(56, 12) = 4 = 12 - 1 \cdot 8 = 12 - 1 \cdot (56 - 4 \cdot 12) = (-1) \cdot 56 + 5 \cdot 12.$$

We formalize this:

**Lemma 5.3.** *Let $a$ and $b \in \mathbb{Z}$ be integers (not both being zero). Then there are integers $x$, $y \in \mathbb{Z}$ such that*

$$\gcd(a, b) = x \cdot a + y \cdot b.$$

A rigorous proof is given in the lecture on number theory (but basically, we have it already).

Another way of finding the gcd of two numbers uses their prime factor decompositions (assuming we are able to find the prime factors which can be a hard task).

## Greatest common divisors for pairs of polynomials

**Definition 5.4.** *Let $A(z)$ and $B(z)$ be polynomials. We say that $B(z)$ is a* divisor of $A(z)$ *if there is a polynomial $Q(z)$ such that*

$$A(z) = Q(z) \cdot B(z) \quad \forall\, z \in \mathbb{C}.$$

In this case, each zero of $B$ is also a zero of $A$.

**Definition 5.5.** *Let $A(z)$ and $B(z)$ be polynomials. We say that a polynomial $G(z)$ is a* greatest common divisor of $A$ and $B$ *if the following three conditions hold:*

- *$G$ is a divisor of $A$,*
- *$G$ is a divisor of $B$,*
- *every other polynomial that divides $A$ as well as $B$ is also a divisor of $G$.*

As an example, we mention that the polynomials $A(z) = z^4 - 1$ and $B(z) = z^3 - 1$ have the greatest common divisor $G(z) = 2z - 2$. But $7z - 7$ is also a greatest common divisor of $A$ and $B$.

Greatest common divisors for pairs of polynomials are not unique (you can always multiply them by a constant factor).

You can find greatest common divisors of pairs of polynomials using their zeroes (assuming we are able to find all their zeroes which can be a hard task). An easier way exploits Euclid's algorithm of repeated polynomial divisions with remainder.

You are strongly invited to figure out the details yourself by doing the homeworks.

# Chapter 6

# On the $\forall$ and $\exists$. And Series

## Purpose

We begin to love the symbols $\forall$ and $\exists$ because they help us to keep our thinking clear.
We apply our newly obtained skills to sequences and series.

## Distinguishing Assumptions and Claims

When doing financial accounting, we must distinguish the monetary numbers according to credit and debit.
When doing a calculation, we must distinguish the mathematical objects according to given objects and wanted objects.
When doing a proof, we must distinguish the statements according to assumption/presupposition and claim/assertion.
Whatever you do — when you mix these things up, your piece of work *will* go wrong.

## Simple Sentences with $\forall$ and $\exists$

Every human has a parent: let $\mathcal{H}$ denote the set of all humans (alive or already deceased). Then

$$\forall h \in \mathcal{H}\colon \exists \tilde{h} \in \mathcal{H}\colon \tilde{h} \text{ is parent of } h.$$

The expression $\exists \tilde{h}$ does not exclude that there is a second parent of $h$.
Some humans have children:

$$\exists h \in \mathcal{H}\colon \exists \hat{h} \in \mathcal{H}\colon h \text{ is parent of } \hat{h}.$$

For all two distinct points in the plane, there is exactly one straight line through these two points: let $\mathcal{P}$ be the set of points in the plane and $\mathcal{L}$ be the set of all straight lines in the same plane. Then

$$\forall p \in \mathcal{P}\colon \forall \tilde{p} \in \mathcal{P} \setminus \{p\}\colon \exists! \ell \in \mathcal{L}\colon (p \text{ is on } \ell) \text{ and } (\tilde{p} \text{ is on } \ell).$$

The expression $\mathcal{P} \setminus \{p\}$ means that we remove the point $p$ from the set $\mathcal{P}$, in order to enforce that $\tilde{p} \neq p$. This is the subtraction of two sets, and the result is again a set. The exclamation mark means that there is exactly one such line, not two or more. The purpose of the parantheses is mostly cosmetic and shall clarify that the word "and" combines two statements.
Each natural number has a prime factor decomposition: let $\mathbb{P}$ be the set of primes. Then

$$\forall n \in \mathbb{N}_+\colon \exists k \in \mathbb{N}_+\colon \exists p_1, \dots, p_k \in \mathbb{P}\colon \exists \alpha_1, \dots, \alpha_k \in \mathbb{N}_+\colon n = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot \ldots \cdot p_k^{\alpha_k}.$$

We did not assume that the $p_j$ are distinct (including such a requirement would complicate the line even more than it already is).

## The General Principle

Let $\mathcal{A}$ be any set and $T$ be a statement taking a variable $a$. Then a for-all-statement looks in the formally correct way like this:

$$\forall a \in \mathcal{A} : T(a),$$

and we read it like this:

For all members $a$ of the set $\mathcal{A}$ we have that the statement $T(a)$ is true.

In particular, we read the : after the $\forall$ as "we have that".

And an it-exists-statement looks in the formally correct way like this:

$$\exists a \in \mathcal{A} : T(a),$$

and we read it like this:

There is some member $a$ of the set $\mathcal{A}$ for which the statement $T(a)$ is true.

In particular, we read the : after the $\exists$ as "for which".

The symbols $\forall$ and $\exists$ are called *quantors* because they quantify for how many members of the set $\mathcal{A}$ the statement $T(a)$ is true: for all members, or for at least one of the members.
Some survival rules for young students:

- always put the $T$ part behind the $\forall/\exists$ part, never in front.
- do not forget the "$\in \mathcal{A}$", or abbreviate carefully. For instance, "$\forall \varepsilon > 0$" is an abbreviation of "$\forall \varepsilon \in \mathbb{R}_{>0}$", where $\mathbb{R}_{>}$ is meant as the set of all positive real numbers, and the authors hope that it is clear from the context that $\varepsilon$ is assumed to be real (and not from $\mathbb{N}$ or the set of prime numbers ...).

## How to Use Them. How to Prove Them

If we have such quantor statements on the assumption side, then we are permitted to use them (see the first line in the following table).
If we have such quantor statements on the claim side, then we are required to prove them (see the second line in the following table).

| The statement | $\forall a \in \mathcal{A}$: $T(a)$ | $\exists a \in \mathcal{A}$: $T(a)$ |
|---|---|---|
| is on the assumption side: | We choose our favourite member $a$, and then there is a guarantee that this $a$ satisfies the statement $T(a)$. Afterwards, we may choose another $a$. And then another. We use this statement as often as wanted. | We have the guarantee that $T(a)$ is true for at least one $a \in \mathcal{A}$. Sadly, we are not allowed to assume/hope that $a$ is a nice member. We cannot choose $a$, because somebody else will choose it. We use this statement once. |
| is on the claim side: | Whenever some other person gives us an $a$, we must prove that then $T(a)$ holds for that $a$. We cannot choose $a$, because somebody else has already chosen it. We do many proofs: one for each $a$. | We must find at least one member $a \in \mathcal{A}$ and prove $T(a)$ for this $a$. We are allowed to choose $a$ (among those for which $T(a)$ holds), and we pick an $a$ with an easy proof (since laziness is human). We do one proof. |

## How to Negate Them

The rule is: you write the negation operator in the left-most position. Then you shift it from left to right and turn $\forall$ into $\exists$ and conversely:

$$\begin{aligned} &\mathrm{not}\Big(\forall a \in \mathcal{A}\colon T(a)\Big) &&\iff &&\exists a \in \mathcal{A}\colon \mathrm{not}\Big(T(a)\Big) \\ &\mathrm{not}\Big(\exists a \in \mathcal{A}\colon T(a)\Big) &&\iff &&\forall a \in \mathcal{A}\colon \mathrm{not}\Big(T(a)\Big). \end{aligned}$$

The negation of "all people like marmite" is "there exists a person who does not like marmite". The negation of "there was one sunny day of our summer vacation" is "all the days of our summer vacation were non-sunny".
We know that $\lim_{n\to\infty} a_n = A^*$ means:

$$\forall \varepsilon \in \mathbb{R}_{>0}\colon \exists N_0 \in \mathbb{N}\colon \forall n \in \mathbb{N} \text{ with } n \geq N_0\colon |a_n - A^*| < \varepsilon.$$

The logical negation $\lim_{n\to\infty} a_n \neq A^*$ then means

$$\exists \varepsilon \in \mathbb{R}_{>0}\colon \forall N_0 \in \mathbb{N}\colon \exists n \in \mathbb{N} \text{ with } n \geq N_0\colon |a_n - A^*| \geq \varepsilon.$$

This line has been obtained in a completely automatic way:

- write the "not" in the left-most place,
- read from left to right, flip the first quantor which you find, and move the "not" behind the :
- lather, rinse, repeat

However, typically mathematicians write it in a different way. The line of thinking is that some member which is announced by an $\exists$ statement is perhaps something very special, something unique, something precious. And special members of society receive a medal which we write as a subscript$_0$. Therefore, the statement $\lim_{n\to\infty} a_n \neq A^*$ is formalized like this:

$$\exists \varepsilon_0 \in \mathbb{R}_{>0}\colon \forall N \in \mathbb{N}\colon \exists n_0 \in \mathbb{N} \text{ with } n_0 \geq N\colon |a_{n_0} - A^*| \geq \varepsilon_0.$$

This is logically totally equivalent to the previous one, but it follows more closely the mathematical traditions.

## Something about Series

Let $(a_1, a_2, \dots) = (a_n)_{n\in\mathbb{N}}$ be a sequence of complex numbers. For $N \in \{1, 2, \dots\}$, we define the partial sums

$$S_N := a_1 + a_2 + \ldots + a_N = \sum_{n=1}^{N} a_n.$$

If $\lim_{N\to\infty} S_N = S^*$ exists with $S^* \in \mathbb{C}$, then we say that the series $\sum_{n=1}^\infty a_n$ *converges with limit* $S^*$, and then the expression $\sum_{n=1}^\infty a_n$ means two things at the same time:

- the sequence $(S_1, S_2, S_3, \dots)$ of complex numbers,
- the limit $S^*$ of this sequence.

It is a bit unfortunate that this expression $\sum_{n=1}^\infty a_n$ possesses two meanings, but this is the mathematical habit.

**Lemma 6.1.** *If* $\sum_{n=1}^\infty a_n$ *converges, then* $\lim_{n\to\infty} a_n = 0$.

*Pre-Proof on scratch paper just to get a better understanding:* We have the statement

$$\exists S^* \in \mathbb{C}\colon\ \forall \varepsilon > 0\colon\ \exists N_{S,0}(\varepsilon) \in \mathbb{N}\colon\ \forall n \geq N_{S,0}(\varepsilon)\colon\ |S_n - S^*| < \varepsilon$$

on the assumption side, and we have the statement

$$\forall \varepsilon > 0\colon\ \exists N_{a,0}(\varepsilon) \in \mathbb{N}\colon\ \forall n \geq N_{a,0}(\varepsilon)\colon\ |a_n - 0| < \varepsilon$$

on the claim side.
In order to obtain a better feeling what the assumption means, we choose some special $\varepsilon$, namely $\varepsilon = 0.1$. We are allowed to do this because $\varepsilon$ comes with an $\forall$ on the assumption side. Hence we have the guarantee that the following is true:

$$\exists N_{S,0}(0.1) \in \mathbb{N}\colon\ \forall n \geq N_{S,0}(0.1)\colon\ |S_n - S^*| < 0.1.$$

We choose one more special $\varepsilon$, namely $\varepsilon = 0.01$, and we are sure that

$$\exists N_{S,0}(0.01) \in \mathbb{N}\colon\ \forall n \geq N_{S,0}(0.01)\colon\ |S_n - S^*| < 0.01.$$

So these $S_n$ are even closer to $S^*$ than before (which then should mean that $N_{S,0}(0.01) \gg N_{S,0}(0.1)$ because we should throw away the bad values of $n$ when making $\varepsilon$ ten times smaller). Now phantasy kicks in and recommends to compare $S_n$ and $S_{n-1}$:

$$S_n = a_1 + a_2 + \cdots + a_n, \qquad S_{n-1} = a_1 + a_2 + \cdots + a_{n-1},$$

and therefore $a_n = S_n - S_{n-1}$. However, if $n$ and $n-1$ are both at least $N_{S,0}(0.01)$, then $|S_n - S_{n-1}| < 0.02$, which means $|a_n| < 0.02$, which means $|a_n - 0| < 0.02$.
Our formal proof will be complete when we leave the special values 0.1 and 0.01 behind and return to general positive $\varepsilon$. □

*Proof.* We know

$$\exists S^* \in \mathbb{C}\colon\ \forall \varepsilon > 0\colon\ \exists N_{S,0}(\varepsilon) \in \mathbb{N}\colon\ \forall n \geq N_{S,0}(\varepsilon)\colon\ |S_n - S^*| < \varepsilon.$$

Let us be given a positive $\varepsilon$. For this $\varepsilon$, we define

$$N_{0,a}(\varepsilon) := N_{0,S}\left(\frac{\varepsilon}{2}\right) + 1.$$

Let $n \in \mathbb{N}$ be any number with $n \geq N_{0,a}(\varepsilon)$. Then we have

$$n \geq N_{0,S}\left(\frac{\varepsilon}{2}\right), \qquad n - 1 \geq N_{0,S}\left(\frac{\varepsilon}{2}\right),$$

and the assumption then implies (for this fixed $n$)

$$|S_n - S^*| < \frac{\varepsilon}{2}, \qquad |S_{n-1} - S^*| < \frac{\varepsilon}{2}.$$

Now we conclude (using the triangle inequality) that

$$|a_n - 0| = |a_n| = |S_n - S_{n-1}| = |(S_n - S^*) + (S^* - S_{n-1})| \leq |S_n - S^*| + |S^* - S_{n-1}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.$$

The proof is complete. □

# Chapter 7

# On Matrices

## Purpose

What **is** a matrix ? This is easily answered: you write certain numbers in a tabular form and put a pair of parentheses around. Next we could learn how to manipulate a matrix, with just a shallow understanding of what this is all about. I prefer another learning style.
A much better question is: What **does** a matrix ? We will give a geometrical answer and obtain a deeper understanding.

## Cakes. Their Ingredients. Their Prices

Imagine a cake shop that is baking their own cakes every day. Consider two types of different cakes (type 1, type 2). On a certain day, they are baking $c_1$ cakes of type 1 and $c_2$ cakes of type 2. We write this as

$$\vec{c} := \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \in \mathbb{R}^2.$$

Well, technically we have $c_1 \in \mathbb{N}$ and $c_2 \in \mathbb{N}$, hence $\vec{c} \in \mathbb{N}^2$, but each natural number is a real number.

Vectors are always written as columns.

The cakes have recipes like this:

| | flour | butter | sugar | eggs |
|---|---|---|---|---|
| type 1 | 0.4 kg | 0.2 kg | 0.25 kg | 3 |
| type 2 | 0.5 kg | 0.2 kg | 0.2 kg | 4 |

And we arrange them in a recipe matrix $R$:

$$R := \begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix}.$$

We calculate the matrix-vector-product $R\vec{c}$:

$$R\vec{c} = \begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} c_1 \cdot 0.4\,\mathrm{kg} + c_2 \cdot 0.5\,\mathrm{kg} \\ c_1 \cdot 0.2\,\mathrm{kg} + c_2 \cdot 0.2\,\mathrm{kg} \\ c_1 \cdot 0.25\,\mathrm{kg} + c_2 \cdot 0.2\,\mathrm{kg} \\ c_1 \cdot 3 + c_2 \cdot 4 \end{pmatrix}. \tag{$\spadesuit$}$$

This is a vector (column, as we agreed), with 4 entries, and it means the amount of ingredients taken out of the storage room on that particular baking day. This vector is a member of $\mathbb{R}^4$.

What **does** the matrix $R$ ?
It translates from the cake-amount-vector $\vec{c}$ to the ingredient-amount-vector $\vec{i} = R\vec{c}$.

$$\boxed{\text{A matrix with } q \text{ columns and } r \text{ rows induces a mapping from } \mathbb{R}^q \text{ into } \mathbb{R}^r.}$$

Now let us consider prices of the ingredients:

| | flour | butter | sugar | egg |
|---|---|---|---|---|
| price | $1.2\frac{£}{\mathrm{kg}}$ | $1.4\frac{£}{\mathrm{kg}}$ | $0.9\frac{£}{\mathrm{kg}}$ | $0.15£$ |

We arrange them in the price matrix:

$$P := \begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix}.$$

Now what is the product $P\vec{i}$ ?
The matrix $P$ has $q = 4$ columns and $r = 1$ rows, so it induces a mapping from $\mathbb{R}^4$ into $\mathbb{R}^1$, and the meaning of the product $P\vec{i}$ is the total financial value of the ingredients $\vec{i}$.
What is the product $PR$ ? Let us calculate:

$$PR = \begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.4\,\mathrm{kg} & 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} & 0.2\,\mathrm{kg} \\ 3 & 4 \end{pmatrix}.$$

Our calculations are too long for this line, so we proceed in steps. The first entry of $PR$ is

$$\begin{aligned} &\begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.4\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 0.25\,\mathrm{kg} \\ 3 \end{pmatrix} \\ &= \left(1.2\frac{£}{\mathrm{kg}} \cdot 0.4\,\mathrm{kg} + 1.4\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.9\frac{£}{\mathrm{kg}} \cdot 0.25\,\mathrm{kg} + 0.15£ \cdot 3\right) \\ &= \left(0.48£ + 0.28£ + 0.225£ + 0.45£\right) \\ &= 1.435£. \end{aligned}$$

This is the price of all the ingredients in one cake of type 1. Now we perform a dance of joy and happiness because all the kilogramm units have nicely cancelled, and all the four items which we added have the same unit £.
And the second entry of the product $PR$ is calculated here:

$$\begin{aligned} &\begin{pmatrix} 1.2\frac{£}{\mathrm{kg}} & 1.4\frac{£}{\mathrm{kg}} & 0.9\frac{£}{\mathrm{kg}} & 0.15£ \end{pmatrix} \cdot \begin{pmatrix} 0.5\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 0.2\,\mathrm{kg} \\ 4 \end{pmatrix} \\ &= \left(1.2\frac{£}{\mathrm{kg}} \cdot 0.5\,\mathrm{kg} + 1.4\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.9\frac{£}{\mathrm{kg}} \cdot 0.2\,\mathrm{kg} + 0.15£ \cdot 4\right) \\ &= \left(0.6£ + 0.28£ + 0.18£ + 0.6£\right) \\ &= 1.66£. \end{aligned}$$

Our final answer then is:

$$PR = \begin{pmatrix} 1.435£ & 1.66£ \end{pmatrix}.$$

This is a matrix with $q = 2$ columns and $r = 1$ rows, so it induces a mapping from $\mathbb{R}^2$ into $\mathbb{R}^1$, and this mapping $\vec{c} \mapsto (PR)\vec{c}$ means to calculate the total-ingredient-price $(PR)\vec{c}$ associated to the cake-amount-vector $\vec{c}$.

The following seems plausible:

$$(PR)\vec{c} = P(R\vec{c}).$$

The LHS means: we calculate the matrix-matrix-product $PR$ first, and then we multiply by the cake-amount-vector $\vec{c}$. The RHS means: we calculate $R\vec{c}$ first, obtain the ingredient-amount-vector $\vec{i}$, and then we calculate the matrix-vector-product $P\vec{i}$.

$$\boxed{\text{The matrix-matrix-product is associative.}}$$

$$\boxed{\text{If necessary, we read column vectors as matrices with just one column.}}$$

Be careful not to change the order of the letters: $P$, $R$, $\vec{c}$ must remain in this order.

$$\boxed{\text{The matrix-matrix-product is \textbf{not} commutative.}}$$

This means $PR \neq RP$, because the product $RP$ does not make any sense. Moreover, the formats of the matrices do not fit together (try it and you will see what I mean).
Now consider two matrices $A$ and $B$. Even if their formats fit together, typically $AB$ will not be the same as $BA$:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \\ AB &= \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ -2 & -2 \end{pmatrix}, \\ BA &= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We learn that a matrix-matrix-product can result in the zero-matrix, without any of the factors being a zero-matrix. That is surprising.

## Some Theory: What is a Vector Space

A *vector space* (also called *linear space*) is an algebraic structure. When we define an algebraic structure, we go three steps:

- list the members,
- list the actions which we can perform upon the members,
- list the rules that these actions have to obey.

The members of the vector space $\mathbb{R}^3$ can be understood as all points in $\mathbb{R}^3$, or as equivalence classes of arrows. There are two actions: adding two vectors produces a vector (in the usual parallelogram style). And multiplying a vector by a real number produces again a vector (for instance in the sense of stretching the arrow if the number is greater than 1). And the rules are:

$$\begin{array}{lcc} \forall\vec{u},\ \forall\vec{v} & : & \vec{u}+\vec{v} = \vec{v}+\vec{u}, \\ \forall\vec{u},\ \forall\vec{v},\ \forall\vec{w} & : & (\vec{u}+\vec{v})+\vec{w} = \vec{u}+(\vec{v}+\vec{w}), \\ \forall\vec{u},\ \forall\vec{v},\ \forall\lambda\in\mathbb{R} & : & \lambda\cdot(\vec{u}+\vec{v}) = \lambda\cdot\vec{u}+\lambda\cdot\vec{v}, \\ \forall\vec{u},\ \forall\lambda\in\mathbb{R},\ \forall\mu\in\mathbb{R} & : & (\lambda+\mu)\cdot\vec{u} = \lambda\cdot\vec{u}+\mu\cdot\vec{u}, \\ \forall\vec{u},\ \forall\lambda\in\mathbb{R},\ \forall\mu\in\mathbb{R} & : & (\lambda\cdot\mu)\cdot\vec{u} = \lambda\cdot(\mu\cdot\vec{u}), \\ \forall\vec{a},\ \forall\vec{b}\text{: } \exists!\vec{x} & : & \vec{a}+\vec{x} = \vec{b}, \\ \forall\vec{u} & : & 1\cdot\vec{u} = \vec{u}. \end{array}$$

## Some Theory: What is a Linear Mapping

**Adding vectors:** Consider again the recipe matrix $R$ and the cake-amount-vector $\vec{c}$. Suppose that the bake shop produces the cake-amount $\vec{c}_m$ on monday and the cake-amount $\vec{c}_t$ on tuesday. Then the cake shop takes $(R\vec{c}_m) + (R\vec{c}_t)$ of ingredients from the storage room.

But the cake shop could have baked $\vec{c}_m + \vec{c}_t$ cakes on one day together and hence taken $R(\vec{c}_m + \vec{c}_t)$ out of the storage room. Which is the same amount of ingredients. Hence

$$(R\vec{c}_m) + (R\vec{c}_t) = R(\vec{c}_m + \vec{c}_t).$$

The addition on the LHS is a vectorial addition in $\mathbb{R}^4$, and the addition on the RHS is a vectorial addition in $\mathbb{R}^2$. Both are done in parallelogram style.

**Multiplying vectors by numbers:** Baking a triple amount of cakes needs a triple amount of ingredients:

$$R(3 \cdot \vec{c}) = 3 \cdot (R\vec{c}).$$

The multiplication by 3 on the LHS is a number-times-vector multiplication in $\mathbb{R}^2$. The multiplication by 3 on the RHS is a number-times-vector multiplication in $\mathbb{R}^4$.

We recall that $R$ induces a mapping from $\mathbb{R}^2$ into $\mathbb{R}^4$. These are vector spaces (also called *linear spaces*), and the typical actions in each of them are *vector plus vector gives vector* and *number times vector gives vector*. These two actions are called *linear operations.*

A mapping between linear spaces is *linear* when it is compatible with the 2 linear operations.

**Definition 7.1 (Linear mapping).** *A mapping $f\colon \mathbb{R}^q \to \mathbb{R}^r$ is called* linear *if the following two statements are true:*

$$\begin{aligned} \forall \vec{u} \in \mathbb{R}^q,\ \forall \vec{v} \in \mathbb{R}^q \quad &: \quad f(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v}), \\ \forall \vec{u} \in \mathbb{R}^q,\ \forall \lambda \in \mathbb{R} \quad &: \quad f(\lambda \cdot \vec{u}) = \lambda \cdot f(\vec{u}). \end{aligned}$$

Matrices and mappings are related:

- each matrix $A \in \mathbb{R}^{r\times q}$ induces a linear mapping $f_A$ from $\mathbb{R}^q$ into $\mathbb{R}^r$. The vector $\vec{u} \in \mathbb{R}^q$ is mapped to the result $f_A(\vec{u}) := A\vec{u}$.
- for each linear mapping $f$ from $\mathbb{R}^q$ into $\mathbb{R}^r$, there is exactly one matrix $A \in \mathbb{R}^{r\times q}$ with the property that $f(\vec{u}) = A\vec{u}$ for all $\vec{u} \in \mathbb{R}^q$.

## How to Find a Mapping From a Matrix

Easy (just multiply matrix times vector).

## How to Find a Matrix From a Mapping

Harder, but not by much. We have the linear mapping $f$ and want to find the matrix $A$.
Or in the cake example: there is another cake shop, and we as outsiders want to determine their recipe matrix. This can be done like this: let them bake one cake of type 1 and record[1] how the storage changes. This gives the first column of $R$. Let them bake one cake of type 2 and record how the storage changes. This gives the second column of $R$.
Baking just one cake of type 1 corresponds to the cake-amount-vector $\binom{1}{0}$.
Baking just one cake of type 2 corresponds to the cake-amount-vector $\binom{0}{1}$.

[1] yes, this analogy is broken a bit

A bit more schematic: $f$ maps from $\mathbb{R}^2$ into $\mathbb{R}^4$. The first canonical basis vector of $\mathbb{R}^2$ is $\vec{e}_1 := \binom{1}{0}$. The second canonical basis vactor of $\mathbb{R}^2$ is $\vec{e}_2 := \binom{0}{1}$. Then the first column of $R$ is $f(\vec{e}_1)$, and the second column of $R$ is $f(\vec{e}_2)$.

Now more general: let $f$ be a linear map from $\mathbb{R}^q$ into $\mathbb{R}^r$. Consider the canonical basis vectors $\vec{e}_1$, $\vec{e}_2$, ..., $\vec{e}_q$ of $\mathbb{R}^q$. They look like this: $\vec{e}_j$ is a column with $q$ entries, almost all of them are zero, except a number one at the position $j$. Then the corresponding matrix $A$ for $f$ looks like this: $A$ has $q$ columns, and the $j$-th column of $A$ is $f(\vec{e}_j)$.

$$\boxed{\text{The columns of the associated matrix are the images of the canonical basis vectors.}}$$

The deeper reason becomes visible when we rewrite (♠):

$$R\vec{c} = \begin{pmatrix} 0.4\,\text{kg} & 0.5\,\text{kg} \\ 0.2\,\text{kg} & 0.2\,\text{kg} \\ 0.25\,\text{kg} & 0.2\,\text{kg} \\ 3 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} c_1 \cdot 0.4\,\text{kg} + c_2 \cdot 0.5\,\text{kg} \\ c_1 \cdot 0.2\,\text{kg} + c_2 \cdot 0.2\,\text{kg} \\ c_1 \cdot 0.25\,\text{kg} + c_2 \cdot 0.2\,\text{kg} \\ c_1 \cdot 3 + c_2 \cdot 4 \end{pmatrix} = c_1 \cdot \begin{pmatrix} 0.4\,\text{kg} \\ 0.2\,\text{kg} \\ 0.25\,\text{kg} \\ 3 \end{pmatrix} + c_2 \cdot \begin{pmatrix} 0.5\,\text{kg} \\ 0.2\,\text{kg} \\ 0.2\,\text{kg} \\ 4 \end{pmatrix}.$$

Now $\vec{c} = \vec{e}_1$ means $c_1 = 1$ and $c_2 = 0$, but $\vec{c} = \vec{e}_2$ means $c_1 = 0$ and $c_2 = 1$.

Let us generalize this a bit: the matrix-vector-product

$$A\vec{x} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \dots & a_{rq} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix}$$

means that we take the first column of $A$ multiplied by $x_1$, the second column of $A$ multiplied by $x_2$, the third column of $A$ multiplied by $x_3$, and so on, the $q$-th column of $A$ multiplied by $x_q$, and then we make the summation over these $q$ vectors (each having $r$ entries).

## Example: How to Find a Rotation Matrix

Consider in $\mathbb{R}^2$ the rotation about the origin in counter-clockwise sense by 37°. This is a mapping from $\mathbb{R}^2$ into $\mathbb{R}^2$ because each point is sent to a point. We should check that this map is linear. How to find its corresponding matrix ?

We know how to do it: the first canonical basis vector is $\binom{1}{0}$, and a little bit of trigonometry reveals that its image is $\binom{\cos 37^\circ}{\sin 37^\circ}$. The second canonical basis vector is $\binom{0}{1}$, and its image is $\binom{-\sin 37^\circ}{\cos 37^\circ}$. Hence we obtain the mapping matrix

$$A = \begin{pmatrix} \cos 37^\circ & -\sin 37^\circ \\ \sin 37^\circ & \cos 37^\circ \end{pmatrix}.$$

This was not hard. You just need some training.

**Exercise 2.** *Consider the line through the origin which has an angle* $38^\circ$ *to the horizontal axis. What is the matrix that corresponds to the reflection across this line ?*

**Exercise 3.** *Imagine the sun in winter as it is only* $30^\circ$ *above the horizon. Each point in the atmosphere (imagine a bird) casts a shadow on the ground. We consider a two-dimensional analogue and ask for the matrix associated to this shadow map. This means: the sunlight comes from the left top part of the paper, as a parallel pencil of rays (* not *looking like a cone), with an angle of* $30^\circ$ *to the horizontal axis* $y = 0$*. Determine the mapping matrix (two rows, two columns).*

## Some Theory: What is a Linear Combination, Linear Dependence, Linear Independence

We start with a family[2] $(\vec{u}_1, \vec{u}_2, \dots, \vec{u}_m)$ of $m$ vectors, each of them living in the space $\mathbb{R}^n$. When you combine them *in a linear manner*, you obtain a linear combination. This means that only the two typical operations of a linear space are allowed, which are *vector plus vector gives vector* and *number times vector gives vector.*

**Definition 7.2** (**Linear combination**). *We say that a vector $\vec{v} \in \mathbb{R}^n$ is a* linear combination *of the vectors $\vec{u}_1, \dots, \vec{u}_m \in \mathbb{R}^n$ if real numbers $\lambda_1, \dots, \lambda_m$ exist with*

$$\vec{v} = \lambda_1 \cdot \vec{u}_1 + \lambda_2 \cdot \vec{u}_2 + \dots + \lambda_m \cdot \vec{u}_m.$$

Let us write this in another way. The entries of the $\vec{u}_j$ are christened following this scheme:

$$\vec{u}_1 = \begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{pmatrix}, \qquad \vec{u}_2 = \begin{pmatrix} u_{12} \\ u_{22} \\ \vdots \\ u_{n2} \end{pmatrix}, \qquad \dots, \qquad \vec{u}_m = \begin{pmatrix} u_{1m} \\ u_{2m} \\ \vdots \\ u_{nm} \end{pmatrix}.$$

Then the equation $\vec{v} = \lambda_1 \cdot \vec{u}_1 + \lambda_2 \cdot \vec{u}_2 + \dots + \lambda_m \cdot \vec{u}_m$ turns into

$$\vec{v} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nm} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix}.$$

$$\boxed{\text{Every matrix-vector-product is a linear combination of the columns of the matrix.}}$$

Now keep the vectors $\vec{u}_1, \dots, \vec{u}_m$ fixed and let the real numbers $\lambda_1, \dots, \lambda_m$ run through all of $\mathbb{R}$. All the linear combinations which you then get are the *linear span.*

**Definition 7.3** (**Linear span**). *The* linear span *of the family $(\vec{u}_1, \dots, \vec{u}_m)$ is this subset of $\mathbb{R}^n$:*

$$\operatorname{span}\Big(\vec{u}_1, \dots, \vec{u}_m\Big) := \left\{ \sum_{j=1}^{m} \lambda_j \cdot \vec{u}_j \colon \lambda_1 \in \mathbb{R}, \dots, \lambda_m \in \mathbb{R} \right\}.$$

**Definition 7.4** (**Linearly dependent**). *We say that a family $(\vec{u}_1, \dots, \vec{u}_m)$ of $\mathbb{R}^n$ is* linearly dependent *if one of the members can be written as a linear combination of the other ones.*

As an example, we take

$$\vec{u}_1 = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}, \qquad \vec{u}_2 = \begin{pmatrix} 3 \\ -4 \\ 2 \end{pmatrix}, \qquad \vec{u}_3 = \begin{pmatrix} 7 \\ 14 \\ 35 \end{pmatrix}. \tag{$\clubsuit$}$$

Then the family $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$ is linearly dependent because we can write $\vec{u}_1 = 0 \cdot \vec{u}_2 + \frac{1}{7} \cdot \vec{u}_3$. We can also write $\vec{u}_3 = 0 \cdot \vec{u}_2 + 7 \cdot \vec{u}_1$. It is impossible to express $\vec{u}_2$ as linear combination of $\vec{u}_1$ and $\vec{u}_3$, but this does not harm the linear dependence of the family $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$. We have three vectors here, but their linear span is a two-dimensional plane that lies in $\mathbb{R}^3$.

**Definition 7.5** (**Linearly independent**). *We say that a family $(\vec{u}_1, \dots, \vec{u}_m)$ of $\mathbb{R}^n$ is* linearly independent *when it is not linearly dependent.*

Another way of stating the linear independence of the family $(\vec{u}_1, \dots, \vec{u}_m)$: whenever the equation

$$\lambda_1 \cdot \vec{u}_1 + \lambda_2 \cdot \vec{u}_2 + \dots + \lambda_m \cdot \vec{u}_m = \vec{0}$$

is true, then the real numbers $\lambda_1, \lambda_2, \dots, \lambda_m$ must all be zero.

[2] The difference between *family* and *set* is that a family may contain an element twice or even more often, which is forbidden for a set.

## Image Spaces, Rank, and Null Spaces of Matrices

Consider a matrix $A$ with $m$ columns and $n$ rows. If $\vec{x} \in \mathbb{R}^m$ runs through all of $\mathbb{R}^m$, then $A\vec{x}$ runs through some set which we call *image space of $A$*.

**Definition 7.6 (Image space).** *For $A \in \mathbb{R}^{n\times m}$, we define*

$$\operatorname{img}(A) := \left\{ A\vec{x} \colon \vec{x} \in \mathbb{R}^m \right\}$$

*and call it the* image space of $A$.

This is the linear span of the columns of $A$, and it is a sub-vector-space of the vector space $\mathbb{R}^n$.

**Definition 7.7 (Rank of a matrix).** *For $A \in \mathbb{R}^{n\times m}$, we define*

$$\operatorname{rank}(A) := \dim\Big(\operatorname{img}(A)\Big),$$

*the dimension of the image space of $A$, and we call it the* rank of $A$.

This is the maximal number of linearly independent columns of $A$.

**Exercise 4.** *We take the vectors $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$ from (♣) and build a matrix from them:*

$$A := \begin{pmatrix} 1 & 3 & 7 \\ 2 & -4 & 14 \\ 5 & 2 & 35 \end{pmatrix}. \tag{$\diamond$}$$

*Check that the family $(\vec{u}_1, \vec{u}_2)$ is linearly independent. Check that the family $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$ is linearly dependent. Now determine* $\operatorname{rank}(A)$.

For this $A$, we may also ask for the maximal number of linearly independent *rows*. It equals two, because the first two rows are linearly independent (check this !), but all three rows are not:

$$\begin{aligned} &\tfrac{24}{10} \cdot \begin{pmatrix} 1 & 3 & 7 \end{pmatrix} \\ +&\tfrac{13}{10} \cdot \begin{pmatrix} 2 & -4 & 14 \end{pmatrix} \\ =&\begin{pmatrix} \frac{50}{10} & \frac{20}{10} & \frac{350}{10} \end{pmatrix} \end{aligned}$$

The following theorem gives the theoretical reason for this surprise:

**Theorem 7.8.** *For each matrix, the maximal number of linearly independent rows equals the maximal number of linearly independent columns.*

This enables us to calculate the rank of a matrix by transforming it into the row echelon form.

**Exercise 5.** *Determine the rank of the matrix from Exercise 2, and the rank of the matrix from Exercise 3.*

**Exercise 6.** *Let $A \in \mathbb{R}^{n\times m}$ be any matrix with $m$ columns and $n$ rows, and let $\vec{b} \in \mathbb{R}^n$ by any vector. Let us consider the problem of finding all $\vec{x} \in \mathbb{R}^m$ that solve $A\vec{x} = \vec{b}$. We consider the matrix $A$, and the extended matrix $(A|\vec{b})$ which is obtained when we write $\vec{b}$ next to $A$.*
*Give a reason why* $\operatorname{rank}(A) \le \operatorname{rank}(A|\vec{b})$. *Explain why the system $A\vec{x} = \vec{b}$ is unsolvable if* $\operatorname{rank}(A) \neq \operatorname{rank}(A|\vec{b})$.

Next we discuss all those $\vec{x}$ for which $A\vec{x} = \vec{0}$.

**Exercise 7.** *Consider $A$ from ($\diamond$). Determine all those $\vec{x} \in \mathbb{R}^3$ with $A\vec{x} = \vec{0}$. Where have you seen these numbers before ?*

**Definition 7.9** (**Null space, kernel**). *Let $A \in \mathbb{R}^{n\times m}$ be a matrix. Then the following subset of $\mathbb{R}^m$ is called the* null space of *$A$:*

$$\ker(A) := \left\{\vec{x} \in \mathbb{R}^m \colon A\vec{x} = \vec{0}\right\}.$$

*It also has the name* kernel of *$A$.*

This is a sub-vector-space of $\mathbb{R}^m$.

**Definition 7.10** (**Nullity**). *For $A \in \mathbb{R}^{n\times n}$, we define*

$$\operatorname{nul}(A) := \dim\Big(\ker(A)\Big),$$

*the dimension of the null space of $A$, and we call it the* nullity of *$A$.*

**Exercise 8.** *Determine the nullity of the matrix from Exercise 2, and the nullity of the matrix from Exercise 3, and the nullity of the matrix $A$ of $(\diamond)$. How do they relate to the ranks ?*

Now we have prepared the highlight of this note:

**Theorem 7.11** (**Rank–Nullity Theorem**). *For each $A \in \mathbb{R}^{n\times m}$, inducing a linear mapping from $\mathbb{R}^m$ into $\mathbb{R}^n$, we have*

$$m = \operatorname{nul}(A) + \operatorname{rank}(A).$$

One advantage of this is that we can draw a lot of information about a matrix from its row echelon form.

# Chapter 8

# The Beauty of Mathematics

## Introduction

Mathematics is beautiful, because

- it gives us deeper insight into structures ("the truth behind something"),
- it requires phantasie and creativity,
- it surprises,
- it connects seemingly unrelated topics.

Certainly there are more reasons, but we focus on these four here.

## A Question

We define the Fibonacci numbers $f_0$, $f_1$, $f_2$, ..., via the recursion

$$f_0 := 1, \qquad f_1 := 1, \qquad f_{n+2} := f_{n+1} + f_n \quad (\forall\, n \in \mathbb{N}_0).$$

Is there a non-recursive formula for $f_n$, which enables us to compute $f_{1273}$ without computing "all the earlier ones" ?

## One Answer

We start with some **seemingly unrelated topic** and consider, for a complex variable $z$, the power series

$$F(z) := \sum_{n=0}^{\infty} f_n z^n, \quad z \in \mathbb{C}.$$

This is a definition, and after each definition it is reasonable to ask whether that newly defined object exists at all. So, for which $z$ does the series on the RHS converge ? To this end, we benefit from a little result:

**Lemma 8.1.** *For each $n$, the Fibonacci number $f_n$ is a natural number with $1 \le f_n \le 2^n$.*

A proof can be done by mathematical induction. Then the root test applies and we deduce that the above series converges for all $z \in \mathbb{C}$ with $|z| < \frac{1}{2}$. In the sequel, we always assume $|z| < \frac{1}{2}$.

Now the **Formula Fairy** speaks to us and recommends to have a look at $(1-z-z^2)F(z)$:

$$\begin{aligned}(1-z-z^2)F(z) &= f_0 + f_1 z + f_2 z^2 + f_3 z^3 + f_4 z^4 + \dots \\ &\qquad - f_0 z - f_1 z^2 - f_2 z^3 - f_3 z^4 - \dots \\ &\qquad\qquad - f_0 z^2 - f_1 z^3 - f_2 z^4 - \dots \\ &= f_0 + (f_1 - f_0) z \\ &= 1,\end{aligned}$$

hence we find one more formula for $F(z)$, namely

$$F(z) = \frac{-1}{z^2+z-1}, \qquad \forall z \in \mathbb{C} \quad \text{with} \quad |z| < \frac{1}{2}.$$

This means that we have two different formulas for $F(z)$, and a very powerful principle of maths is to have two representations for the same object[1]. And now we find a third representation of $F$. To this end, we recall that poles of a function are always interesting, and this motivates a partial fraction decomposition: the equation $z^2+z-1=0$ has the solutions

$$z_1 := -\frac{1}{2} + \frac{\sqrt{5}}{2}, \qquad z_2 := -\frac{1}{2} - \frac{\sqrt{5}}{2},$$

and then the ansatz

$$F(z) \overset{!}{=} \frac{A_1}{z-z_1} + \frac{A_2}{z-z_2}$$

gives $A_1 = \frac{-1}{\sqrt{5}}$ and $A_2 = \frac{1}{\sqrt{5}}$.
What we are doing here is **connecting seemingly unrelated topics**, and since this is fun, we proceed to connect with another seemingly unrelated topic, now the geometric series:

$$\begin{aligned}F(z) &= \frac{A_1}{z-z_1} + \frac{A_2}{z-z_2} \\ &= -\frac{A_1}{z_1}\cdot\frac{1}{1-\frac{z}{z_1}} - \frac{A_2}{z_2}\cdot\frac{1}{1-\frac{z}{z_2}} \\ &= -\frac{A_1}{z_1}\sum_{n=0}^{\infty}\left(\frac{z}{z_1}\right)^{n+1} - \frac{A_2}{z_2}\sum_{n=0}^{\infty}\left(\frac{z}{z_2}\right)^{n+1} \\ &= \sum_{n=0}^{\infty}\left(\frac{-A_1}{z_1^{n+1}} + \frac{-A_2}{z_2^{n+1}}\right) z^n.\end{aligned}$$

Now we refer to some theoretical result (which must be proved !) that comparing coefficients is allowed even for power series, and this then gives

$$f_n = \frac{-A_1}{z_1^{n+1}} + \frac{-A_2}{z_2^{n+1}}.$$

We consider fractions with a variable in the denominator ugly, so we **connect** to Vieta's theorem: since $z_1$ and $z_2$ are solutions to $z^2+z-1=0$, we have $z_1+z_2=-1$ and $z_1 z_2 = -1$, hence $\frac{1}{z_j} = -z_{3-j}$ for $j=1,2$, and therefore

$$\begin{aligned}f_n &= -\left(A_1(-z_2)^{n+1} + A_2(-z_1)^{n+1}\right) \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right), \qquad n \in \mathbb{N}_0.\end{aligned}$$

This is known as BINET's Formula and it is an answer to our question.

**Exercise 9.** *Compute $f_0$, $f_1$, $f_2$, $f_3$ from this formula (as a check).*

[1] For instance, the scalar product $\vec{a}\cdot\vec{b}$ can be written in two ways: $\vec{a}\cdot\vec{b} = |\vec{a}|\cdot|\vec{b}|\cdot\cos(\angle(\vec{a},\vec{b}))$ and $\vec{a}\cdot\vec{b} = \sum_k a_k b_k$, and this is the whole point of the scalar product: to be able to compute angles from the coordinates only.

## Every Answer Generates a New Question

Binet's formula **surprises** us: the LHS is a natural number, but the RHS contains fractions, and even more disturbing are the various $\sqrt{5}$ terms on the RHS. So we have a new question:

Why do the $\sqrt{5}$ on the RHS cancel, for every $n$ ?

## Answer to the Second Question

We consider an algebraic **structure** which is called a *field*. This is a set of things (which we typically call *numbers*) and the four operations $+$, $-$, $\cdot$, $\div$ that behave in the expected way, and these four operations do not leave the set.

So, $\mathbb{N}$ is not a field (because $-$ and $\div$ leave $\mathbb{N}$), $\mathbb{Z}$ is also not a field, but $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields.

Another example is

$$\mathbb{Q}[\sqrt{5}] := \left\{a + b\sqrt{5} \colon a \in \mathbb{Q}, \quad b \in \mathbb{Q}\right\}.$$

It is clear that you can add, subtract, multiply $a+b\sqrt{5}$ and $c+d\sqrt{5}$ (where $a$, $b$, $c$, $d$ are rational), and the result will then be again in $\mathbb{Q}[\sqrt{5}]$. It is a bit more work to consider the division:

$$\frac{a+b\sqrt{5}}{c+d\sqrt{5}} = \frac{a+b\sqrt{5}}{c+d\sqrt{5}} \cdot \frac{c-d\sqrt{5}}{c-d\sqrt{5}} = \frac{ac-5bd}{c^2-5d^2} + \frac{bc-ad}{c^2-5d^2}\sqrt{5},$$

and therefore also $\div$ stays inside $\mathbb{Q}[\sqrt{5}]$, which is consequently a field.

This brings back old memories: we have seen something similar when we built $\mathbb{C}$. So we **connect** $\mathbb{Q}[\sqrt{5}]$ to $\mathbb{C}$, which we could write as $\mathbb{C} \overset{☺}{=} \mathbb{R}[\sqrt{-1}]$. To this end, we define a conjugation in $\mathbb{Q}[\sqrt{5}]$:

$$\overline{a+b\sqrt{5}} := a - b\sqrt{5}, \qquad a, b \in \mathbb{Q}.$$

**Lemma 8.2.** *The conjugation in* $\mathbb{Q}[\sqrt{5}]$ *commutes with the four operations* $+$, $-$, $\cdot$, $\div$ *and with taking powers in* $\mathbb{Q}[\sqrt{5}]$.

*Sketch of Proof.* Basically the same as in $\mathbb{C}$. □

**Lemma 8.3.** *If* $a_k z^k + a_{k-1} z^{k-1} + \ldots + a_1 z + a_0$ *is a polynomial in the variable* $z \in \mathbb{Q}[\sqrt{5}]$, *and if the coefficients* $a_k, \ldots, a_0$ *are rational, then the following holds: if* $z \in \mathbb{Q}[\sqrt{5}]$ *is a zero, then also* $\overline{z}$.

*Sketch of Proof.* Basically the same as the proof of a similar result in $\mathbb{C}$: if all the $a_\ell$ are real and $z \in \mathbb{C}$ is a zero, then also $\overline{z}$. □

Now the polynomial $z^2 + z - 1$ qualifies for this lemma, and therefore $z_1$ and $z_2$ should be conjugates to each other. Indeed, they are, as a quick check reveals: $\overline{z_1} = z_2$.

Since conjugation commutes with taking powers, we have

$$\overline{z_2^{n+1}} = (\overline{z_2})^{n+1} = z_1^{n+1},$$

hence we can conclude: if $(\frac{1+\sqrt{5}}{2})^{n+1} = a + b\sqrt{5}$, then $(\frac{1-\sqrt{5}}{2})^{n+1} = a - b\sqrt{5}$, hence $(\frac{1+\sqrt{5}}{2})^{n+1} - (\frac{1-\sqrt{5}}{2})^{n+1} = 2b\sqrt{5}$, and therefore $f_n = 2b$ which is rational.

We have found a **structural** reason why the $\sqrt{5}$ in Binet's formula cancel !

Hence we know that the $f_n$ must be rational. We have no reason though why they are integers.

Finally, we compare $\mathbb{Q}[\sqrt{5}]$ and $\mathbb{R}[\sqrt{-1}] = \mathbb{C}$:

| $\mathbb{Q}[\sqrt{5}]$ | $\mathbb{R}[\sqrt{-1}] = \mathbb{C}$ |
|---|---|
| First we augment $\mathbb{Q}$ with $\sqrt{5}$, | First we augment $\mathbb{R}$ with i, |
| then we fill the resulting set up with as many numbers as needed, until the operations stay inside (but not more numbers !). | then we fill the resulting set up with as many numbers as needed, until the operations stay inside (but not more numbers !). |
| Note that $\sqrt{5}$ is a solution to $x^2 - 5 = 0$, | Note that $\mathrm{i} = \sqrt{-1}$ is a solution to $x^2 + 1 = 0$, |
| but $x^2 - 5 = 0$ is unsolvable in $\mathbb{Q}$. | but $x^2 + 1 = 0$ is unsolvable in $\mathbb{R}$. |

These ideas are taken to a much higher **structural** level in the lectures on Higher Algebra, in particular Galois Theory.

## We Connect to Matrix Theory and Get One More Answer

By means of **Phantasy**, we find that the recursion formula $f_{n+2} = f_{n+1} + f_n$ can be written like this:

$$f_{n+2} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix},$$

understood as a product of two matrices on the RHS. We prefer matrices to be quadratic, and therefore we include the triviality $f_{n+1} = f_{n+1}$ into the above equation as second line, giving us

$$\begin{pmatrix} f_{n+2} \\ f_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix}.$$

If we introduce the notation $A := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, then we have, for instance,

$$\begin{pmatrix} f_5 \\ f_4 \end{pmatrix} = A \begin{pmatrix} f_4 \\ f_3 \end{pmatrix} = A^2 \begin{pmatrix} f_3 \\ f_2 \end{pmatrix} = A^3 \begin{pmatrix} f_2 \\ f_1 \end{pmatrix} = A^4 \begin{pmatrix} f_1 \\ f_0 \end{pmatrix} = A^4 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and similarly we have

$$\begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad n \in \mathbb{N}_0.$$

Now we only have to figure out how to compute $A^n$ with little effort.
The eigenvalues $\lambda_1$ and $\lambda_2$ of $A$ are solutions to $\det(A - \lambda I) = 0$, which means $-\lambda(1 - \lambda) - 1 = 0$, or $\lambda^2 - \lambda - 1 = 0$, hence

$$\lambda_1 = \frac{1 + \sqrt{5}}{2}, \qquad \lambda_2 = \frac{1 - \sqrt{5}}{2}.$$

Associated eigenvectors are

$$\text{for } \lambda_1\text{:} \quad \vec{u_1} = \begin{pmatrix} \lambda_1 \\ 1 \end{pmatrix}, \qquad \text{for } \lambda_2\text{:} \quad \vec{u_2} = \begin{pmatrix} \lambda_2 \\ 1 \end{pmatrix},$$

and then we build matrices

$$\Lambda := \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad S := \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix},$$

and then the two equations $A\vec{u_j} = \lambda_j \vec{u_j}$ turn into $AS = S\Lambda$, hence $A = S\Lambda S^{-1}$, and therefore

$$A^n = S\Lambda^n S^{-1} = \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix}^{-1}.$$

This then boils down again to Binet's formula. Please fill in the details yourself.