Winter 2025/26
By Timo Freiesleben

Maths primer for mathematical philosophy students

This short course is tailored for students with little math background who want to enter mathematical philosophy as a field. It introduces basic topics in math:

- **1 Foundations:** 1.1 natural numbers, 1.2 basic proof principles, 1.3 rational/real/complex numbers, 1.4 sets
- **2 Calculus:** 2.1 functions, 2.2 standard functions, 2.3 derivatives, 2.4 integrals, 2.5 multi-dimensional case
- **3 Linear algebra:** 3.1 vector space, 3.2 matrices, 3.3 linear equations, 3.4 Eigenvalues and Eigenvectors
- **4 Probability theory:** 4.1 urn models, 4.2 probability space, 4.3 random variables, 4.4 conditional probability, 4.5 standard distributions and central theorems

No matter if you are interested in logic, general philosophy of science, decision theory, or philosophy of [Insert here math/physics/ statistics/machine learning/social sciences], you can always profit from these basics.

I focus on the topics that I think are useful for doing the masters at the MCMP but also for becoming a good mathematical philosopher more generally.

My general aims with the course are:

- teach you some basic mathematical concepts in a fun way
- introduce you to mathematical thinking
- enable you to solve simple exercises

Let's start our little journey!

1 Fundamentals of mathematics

1.1 Natural numbers, primes, and integers

Definition 1.1 (Natural numbers)**.** We call the numbers $\mathbb{N} := \{0, 1, 2, 3, 4, \ldots, \}$ the natural numbers.

The natural numbers can also be introduced very formally but we don't have enough time for this. Google for the *Peano axioms* to learn more.

Definition 1.2 (Basic laws of natural numbers)**.** On the natural numbers, we can define addition + and multiplication ·. Let a, b and c be three natural numbers, then they satisfy the following rules:

- Neutral element of addition called 0:
$$a + 0 = 0 + a = a$$
- Neutral element of multiplication called 1:
$$a \cdot 1 = 1 \cdot a = a$$
- Commutativity of addition:
$$a + b = b + a$$
- Commutativity of multiplication:
$$a \cdot b = b \cdot a$$
- Associativity of addition:
$$a + (b + c) = (a + b) + c$$
- Associativity of multiplication:
$$a \cdot (b \cdot c) = (a \cdot b) \cdot c$$
- Distributivity of addition and multiplication:
$$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$$

Definition 1.3 (Sum)**.** We can also add several numbers together. For this, we use the sum symbol $\sum$:

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n$$

with n and $a_1, \ldots, a_n$ numbers.

Definition 1.4 (Product)**.** We can also multiply several numbers together. For this, we use the product symbol $\prod$:

$$\prod_{i=1}^{n} a_i = a_1 \cdots a_n$$

with n and $a_1, \ldots, a_n$ numbers.

Definition 1.5 (Factorial)**.** Another relevant operation is the so-called *factorial* of a natural number n, it is defined via

$$n! = \prod_{i=1}^{n} i.$$

Example 1.6. Simple Examples:

- $\sum\limits_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$
- $\prod\limits_{i=1}^{5} i = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 5!$

There are some particularly cool natural numbers called *primes* like 2,3,5, or, and that one is a beauty, 23. They are numbers that can only be divided by 1 or themselves.

Definition 1.7 (Prime)**.** We call a natural number p bigger than 1 a prime if for every product $p = a \cdot b$ of natural numbers holds that $a = 1$ or $b = 1$.

Integers

The natural numbers have their limits. For example, we cannot even solve some simple equations with them. Look at:

$$7 + a = 5$$

This equation cannot be solved using natural numbers.

Definition 1.8 (Integers)**.** We call the numbers $\mathbb{Z} := \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots, \}$ the integers.

With the integers, we add a so-called inverse element with respect to the plus operation. All the basic properties of the natural numbers remain the same in the integers, but we get one extra:

Definition 1.9. Let a be an integer, then there exists an *additive inverse* $-a$, such that:

$$a + (-a) = (-a) + a = 0$$

In mathematics, the integers are a highly interesting object of study. Unlike the natural numbers, the integers have what it needs to be called a *mathematical ring.*

Definition 1.10. For two numbers a and b we say a divides b and write $a \mid b$ if and only if:

$$\text{There exists a number } c \text{ such that} \quad a \cdot c = b$$

1.2 Proving theorems

The idea of a proof lies at the center of mathematics. When we prove mathematical statements we call them theorems. A proof shows that the statement of a theorem follows from our mathematical axioms or other statements that we derived from these axioms.

There are several strategies for proving statements!

Direct proof

In a direct proof, you take the given assumptions and show how they lead to a desired conclusion. You show $A \Rightarrow B$ by assuming A and showing B.

Theorem 1.11. *Let a, b, c be integers, then if $a \mid b$ and $a \mid b + c$ then $a \mid c$.*

Proof. We know that $a \mid b$ and $a \mid b + c$. Therefore, by definition, we know that there exist integers d_1 and d_2, such that:

$$a \cdot d_1 = b \quad \text{and} \quad a \cdot d_2 = b + c$$

Thus, we can derive that:

$$\begin{aligned} c &= b + c - b \\ &= a \cdot d_2 - a \cdot d_1 \\ &\overset{Distr}{=} a \cdot \underbrace{(d_2 - d_1)}_{:=d_3} \end{aligned}$$

That means, we can write c as $c = a \cdot d_3$, which means $a \mid c$ □

Proof by contradiction

Constructivists hate that trick. But it is often a very elegant proof technique. You show A by showing that $\neg A$ leads to a contradiction $\perp$.

Theorem 1.12. *There exist infinitely many prime numbers.*

Proof. We prove this by contradiction. Assume there would be finitely many primes $p_1, \dots, p_n$ with n a natural number. Then, consider the number

$$k := \prod_{i=1}^{n} p_i + 1$$

We know that $k > p_i$ for all prime numbers. Therefore, k cannot be a prime number.
Now, we know two things:

- Because k is not prime, we know that there is a j such that

$$p_j \mid k$$

- By construction, we know that

$$p_j \mid k - 1 = \prod_{i=1}^{n} p_i$$

By the theorem we proved above, we know that $p_j \mid k - (k-1) = 1$. This is a contradiction of p_j being prime and greater than 1. ` □

Proof by contraposition

You show that $A \Rightarrow B$ by showing that $\neg B \Rightarrow \neg A$.

Definition 1.13. We call an integer a *even* if $2 \mid a$. We call it odd if $2 \not| a$.

Theorem 1.14. *If n^2 is odd, then also n is odd.*

Proof. Assume n is even. Then, we can write $n = 2 \cdot a$ for some integer a. Then, we know that

$$n^2 = (2 \cdot a)^2 = 4 \cdot a^2$$

Since $2 \mid 4$, we know that $2 \mid n^2$. □

Proof by example or by counterexample

This is tailored to either statements that claim that something exists or false statements that claim that something holds for all elements. For an existence claim, all the evidence needed is one example. To falsify a for-all statement, all evidence needed is a single counterexample.

Theorem 1.15. *There exists integers a, b and c such that*

$$a^2 + b^2 = c^2$$

Proof. Let $a = 3$, $b = 4$, and $c = 5$, then

$$3^2 + 4^2 = 9 + 16 = 25 = 5^2$$

□

Conjecture 1.16. *Every number of the form $2^n - 1$ is prime.*

Counterexample. Consider $n = 4$

$$2^4 - 1 = 15 = 3 \cdot 5$$

□

Proof by mathematical induction

Mathematical induction is one of the standard methods of how to prove statements of the following form:

$$\text{For all natural numbers } n \text{ with } n \geq k \text{ holds } A(n)$$

Where k is some natural number and A a predicate that takes one input.
There are different ways of how this can be done, which are all equivalent. Two very common versions look like this:

$$A(k) \text{ and for all } n \geq k : (A(n) \to A(n+1))$$

or

$$A(k) \text{ and for all } n \geq k : ((\forall k \leq l \leq n : A(l)) \to A(n+1).$$

The simple idea standing behind mathematical induction is the following.

1. Assume I can prove some statement for a particular natural number k.

2. Assume moreover I can prove that whenever our statement holds for some natural number n greater than k then it also holds for $n+1$.

Then, I can conclude that it holds for all natural numbers n greater than k. Why is that? Easy! I know the statement is true for k since (1). But, if it holds for k it also has to hold for $k+1$ due to (2). Now, I know it holds for $k+1$, but then it also has to hold for $(k+1)+1$, and so on. At some point, we will reach every natural number greater than k by this procedure.

Theorem 1.17. *For all natural numbers, the following equation holds*

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Proof. Let $A(l) :\equiv \sum_{i=1}^{l} i = \frac{l(l+1)}{2}$.

- **Base Case:** Since the statement should hold for all natural numbers including zero our base case is to show that $A(0)$.

$$A(0) \Leftrightarrow \sum_{i=1}^{0} i = \frac{0(0+1)}{2} \Leftrightarrow 0 = 0$$

- **Induction Hypotheses:** for a fixed but arbitrary $n \in \mathbb{N}_0$ holds $A(n)$.

- **Inductive Step:**

$$\begin{aligned}\sum_{i=1}^{n+1} i &= (n+1) + \sum_{i=1}^{n} i \\ &\overset{I.H}{=} (n+1) + \frac{n(n+1)}{2} \\ &= 2\frac{(n+1)}{2} + \frac{n(n+1)}{2} \\ &= \frac{n(n+1)+2(n+1)}{2} \\ &= \frac{(n+1)(n+2)}{2}\end{aligned}$$

□

1.3 Rational, real, and complex numbers

Natural numbers and integers are great. I mean, really great! Just look at prime numbers and you see why they are amazing. But integers are also limited in what problems we can solve. Consider, the problem when you have one big pizza but four friends you want to share it with, how much does each of you get? Formally, this can be described by

$$1 = 5 \cdot a$$

where a describes the share of the pizza each one of you gets. The equation cannot be solved with an integer number a. But it can be solved with rational numbers.

Definition 1.18 (Rational numbers)**.** We call the numbers $\mathbb{Q} := \{\frac{a}{b}$ where a, b integers with $b \neq 0\}$ the rational numbers.

The rational numbers, satisfy all the properties we introduced above like commutativity, associativity, etc. But they add one additional property to our list – an inverse element to multiplication.

Definition 1.19. Let $\frac{a}{b}$ be a rational number with $a, b \neq 0$, then there exists a *multiplicative inverse* $\frac{b}{a}$, such that:

$$\frac{a}{b} \cdot \frac{b}{a} = \frac{b}{a} \cdot \frac{a}{b} = 1$$

The rational numbers are very powerful and mathematically interesting objects. They satisfy all properties of a mathematical *field*. Those are the basic structures on which we can define vector spaces.

Real numbers

But again, even very simple mathematical equations cannot be solved with rational numbers. There is one result that probably every one of you remembers from school mathematics – the theorem by Pythagoras. Pythagoras was obsessed with numbers and their properties. But some numbers made him horrified and unfortunately, they appeared in his most beloved field geometry.

The theorem by Pythagoras states that in a rectangular triangle, the square of the length of one side and the square of the length of the other side is equal to the square of the length of the hypotenuse. Short $a^2 + b^2 = c^2$. Let's look at what happens if $a = b = 1$. Then, $1^2 + 1^2 = 2 = a^2$.

Theorem 1.20. *There is no rational number a that satisfies $a^2 = 2$.*

Proof. We prove this via contradiction. Assume there exists integers b, c with $c \neq 0$ such that $(\frac{b}{c})^2 = 2$ and there is no integer d with $d \mid b$ and $d \mid c$.
Then, $b^2 = 2 \cdot c^2$. Thus, we know that b^2 is dividable by 2, and consequently, via the theorem (exercise in the afternoon) we can infer that $2 \mid b$. Thus, we know that b^* exists such that $b = 2 \cdot b^*$.
Now, we can write $2 \cdot c^2 = b^* 2 \cdot 4$. Equivalently, $c^2 = b^* \cdot 2$. Therefore, we know that c^2 is dividable by 2, and via the theorem (exercise in afternoon session) we can infer that c is dividable by 2.
This means both b and c are dividable by 2, which contradicts our assumption that they had no common divisor. □

Ok, but can we still find some a such that $a^2 = 2$? We again, have to extend the realm of our numbers. This time, to the so-called *reals*.
The reals extend our numbers to all possible decimal numbers. There are various ways to define the reals but all of them require more time and background. The most known approach is to define the reals as limits of Cauchy sequences but we have too little time to introduce them formally. The reals are everywhere and as we will show tomorrow, the number of reals is bigger than of the rational numbers.

Definition 1.21 (Real numbers)**.** We call the numbers $\mathbb{R} := \{\dots a_2 a_1 a_0, a_{-1} a_{-2} \dots \text{ with } a_i \in \{0, 1, \dots, 9\}\}$ the real numbers.

Famous examples of real numbers that are not rational are $\pi \approx 3.14$, $e \approx 2.71$, or $\phi \approx 1.62$. They can all be described as the limit of sequences of rational numbers.

Complex numbers

Are we done yet? Are these all relevant numbers? No, there is one further extension we can make – the complex numbers.
If you have a real number a, then you might have realized the following:

$$a^2 \geq 0$$

If you multiply two positive or two negative numbers with each other, you always end up with a positive number. But how can we then solve the following equation?

$$a^2 = -1$$

We cannot, at least if we have the real numbers. Some of you might even think that we shouldn't be able to solve such equations, it just feels wrong. But in math, we usually start to imagine we could and see if we run into contradictions. And luckily, we do not.

Definition 1.22 (Complex numbers)**.** We call the numbers $\mathbb{C} := \{a + b \cdot i \mid a, b \in \mathbb{R}\}$ the complex numbers. Here, i is the so-called *imaginary number* and is defined as $i^2 = -1$.

Finally, we are done. There are no further polynomial equations that cannot be solved. This is shown in the so-called *fundamental theorem of algebra*.

1.4 Sets

What are sets? Again, there is a very technical answer to this question in the form of the Zermelo-Fraenkel axioms. But we can also understand sets more intuitively:

Definition 1.23 (Set)**.** A set A describes a collection of objects. A set is fully identified by the objects it contains. If an object a is in A, we write $a \in A$ and say a is an element of A. If there are no elements in A, we call it the empty set and write $A = \emptyset$.

This intuitive definition is perfectly fine as a working definition. However, some weird objects are not allowed as sets, otherwise, we reach a contradiction. This is why the Zermelo-Fraenkel axioms are needed. One contradiction that can be derived has famously been shown by Bertrand Russell and is called Russell's Paradox.

Theorem 1.24 (Russell's Paradox)**.** *Consider the following set* $A := \{x \mid x \notin x\}$*. Then, we can derive the following contradiction:*

$$A \in A \Leftrightarrow A \notin A$$

Proof. Afternoon session! □

The popular version of this paradox goes as follows: Imagine there is a barber in the city of Seville, he shaves everyone in Seville who does not shave himself. Does the Barber shave himself?

Definition 1.25. Let A and B be two sets. We define the following relations:

- **Subset:**
$$A \subseteq B \Leftrightarrow \text{ for all } a \text{ holds } a \in A \Rightarrow a \in B$$
- **Superset:**
$$A \supseteq B \Leftrightarrow \text{ for all } b \text{ holds } b \in B \Rightarrow b \in A$$
- **Identity:**
$$A = B \Leftrightarrow A \subseteq B \text{ and } A \supseteq B$$
- **Powerset:**
$$\mathcal{P}(A) = 2^A := \{S \mid S \subseteq A\}$$

Definition 1.26. Let A and B be two sets. We define the following basic operations on them:

- **Union:**
$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$$
- **Intersection:**
$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$
- **Difference:**
$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$$
- **Complement:** Say $A \subseteq \Omega$ and Ω is your reference set.
$$A^C = \Omega \setminus A$$

Note that sets are also kind of strange. They are fully determined by their elements. Sets are the same, no matter how we order the elements in it. Moreover, you can list objects several times but it doesn't matter for the definition of the set:

$$\{1, 2, 3, 4, 5\} = \{3, 3, 3, 2, 5, 1, 4, 1\} \cup \emptyset$$

Exercises

Excercise 1.27. *Prove the following statements by mathematical induction:*

- *For all $n \in \mathbb{N}$ with $n \geq 1$:*
$$3 \mid n^3 + 2n$$
- *For all $n \in \mathbb{N}$ with $n \geq 1$:*
$$\sum_{i=1}^{n} (2i-1) = n^2$$
- *For all $n \in \mathbb{N}$ with $n \geq 4$:*
$$2^n \leq n! \leq n^n$$
- *For all $n \in \mathbb{N}$ with $n \geq 1$:*
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Excercise 1.28. *If n^2 is an even number, then also n is even.*

Excercise 1.29. *For all $n \in \mathbb{N}$ holds that $n(n+1)$ is even.*

Excercise 1.30. *Decide how many elements the following sets contain*

- $\{\emptyset\}$
- $\emptyset$
- $\{\emptyset, \emptyset\}$
- $\{1, \{1, 2\}\}$
- $\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}\}$
- $2^{\{1,2,3\}}$
- $\{1, 2\} \cup \{2, 3\}$

Excercise 1.31. *Describe the following sets:*

- $\mathbb{N} \setminus 2\mathbb{N}$
- $2^{\{1,2,3,4\}}$

Excercise 1.32. *Say we define the set A via the following predicate*

$$x \in A \quad \Leftrightarrow \quad Q(x)$$

with $Q(x) :\equiv x \notin x$. Show that

$$A \in A \quad \Leftrightarrow \quad A \notin A$$

2 Calculus

2.1 Functions

Definition 2.1 (Cartesian Product)**.** Let A and B be sets. Then, the cartesian product $A \times B := \{(a, b) \mid a \in A, b \in B\}$.

Definition 2.2 (Relation)**.** Let A and B be sets. We call R a relation if $R \subseteq A \times B$.

Intuitively, functions are just special relations. They map each x to some unique y.

Definition 2.3 (Function)**.** Let A and B be sets. We call $f : A \to B$ a function, if f is a relation with

i) For all $a \in A$ there exists $b \in B$ with $f(a) = b$.

ii) If $b, b' \in B$ with $f(a) = b$ and $f(a) = b'$, then $b = b'$.

Composition is like building a chain of functions, each mapping to the next space. Important: for composition to work, the co-domain of the inner function has to match the domain of the outer function.

Definition 2.4 (Composition)**.** Let A, B, C be sets and $f : A \to B$ and $g : B \to C$ be functions. We define the composition $g \circ f$ as follows:

$$\text{for } a \in A : \ (g \circ f)(a) := g(f(a))$$

This constitutes a function with $g \circ f : A \to C$.

The following distinctions are in the basics of every math course. If you understand what a function is, they are very intuitive. They check whether the conditions for a function are also satisfied in the opposite direction. We see this later when we define the inverse function.

Definition 2.5 (Injective, surjective, bijective)**.** Let $f : A \to B$ be a function. We call f

- injective if for all $a, a' \in A$ holds
$$f(a) = f(a') \Rightarrow a = a'$$
- surjective if for all $b \in B$ there exists an $a \in A$ such that
$$f(a) = b$$
- bijective if f is injective and surjective.

Definition 2.6 (Identity)**.** Let A be a set. We define the identity function as follows

$$id_A : A \to A \text{ with } id_A(a) = a$$

Theorem 2.7 (Inverse)**.** *If $f : A \to B$ is bijective, then there exists an inverse function $f^{-1} : B \to A$, such that*

- $f \circ f^{-1} = id_B$
- $f^{-1} \circ f = id_A$

You have already applied the definition of cardinality when you compared the number of elements in two finite sets. But the notion of cardinality introduced here is more general, it is also applicable to infinite sets. Two sets have the same cardinality if there is a one-to-one mapping between them.

Definition 2.8 (Cardinality)**.** Let A and B be sets. We say

- A has smaller or equal cardinality than B, if there exists an injective function $f : A \to B$.
- A has bigger or equal cardinality than B, if there exists a surjective function $f : A \to B$.
- A has the same cardinality as B, if there exists a bijective function $f : A \to B$.

We write $|A|$ to describe the cardinality of A.

Theorem 2.9. *Show that $\mathbb{N}$ and $\{2n \mid n \in \mathbb{N}\}$ have equal cardinality.*

Proof. Afternoon session! □

The following three theorems were probably the most central results in set theory. The first says that there are just as many rational numbers as natural numbers. The second says that there are more real numbers than natural numbers. And the third implies that there are infinitely many infinities. Quite a philosophical result I would say.

Theorem 2.10. *$\mathbb{N}$ and $\mathbb{Q}$ have equal cardinality.*

Theorem 2.11. *$\mathbb{N}$ and $\mathbb{R}$ have different cardinality.*

Theorem 2.12. *Let A be a set, then $|2^A| > |A|$.*

Proof. Afternoon session! □

2.2 Standard functions

The following functions you have encountered in school and probably in all contexts where math is applied.

Definition 2.13. Let $a_0, \dots, a_n \in \mathbb{R}, a, b \in \mathbb{R} n \in \mathbb{N}$, then the following are standard functions you will see often:

- Polynomials: $f_1(x) = \sum_{i=0}^{n} a_i x^i$
- Exponential: $f_2(x) = ae^{bx}$
- Logarithm: $f_3(x) = a\ ln(bx)$
- Sinus: $f_4(x) = a\ sin(x)$
- Cosinus: $f_5(x) = a\ cos(x)$

Theorem 2.14. *(Laws for Exponential)*
Let $x, y, z, \beta \in \mathbb{R}$ then:

- $x^y \cdot x^z = x^{y+z}$
- $x^z \cdot y^z = (x \cdot y)^z$
- $(x^y)^\beta = x^{\beta y}$
- $\frac{1}{x^y} = x^{-y}$

Theorem 2.15. *(Laws for Logarithm)*
Let $a, x, y \in \mathbb{R}^+, \beta, z \in \mathbb{R}$ then:

- $log_a(x \cdot y) = log_a(x) + log_a(y)$
- $log_a(\frac{x}{y}) = log_a(x) - log_a(y)$
- $\beta\ log_a(x) = log_a(x^\beta)$
- $x = a^z \Rightarrow z = log_a(x)$

Usually, we will have that $a = e$ is the Euler constant with $e \approx 2,7183$.

2.3 Derivatives

One essential idea in math is the idea of a limit. What happens with $f(x) = e^x$ if x goes to infinity? What happens if x goes to minus infinity? Are these even meaningful questions?
Worse, what happens with $g(x) = \frac{1}{x}$ if x goes to 0? Here it seems to matter from which side we approach 0. Mathematicians use the following notation to describe limits:

$$\underset{x \to \infty}{lim} \quad \text{or} \quad \underset{x \to a}{lim}$$

The central concept motivating the idea of these limits was the calculus of derivatives and integrals. In derivates, the idea is to compute the slope of a function in a certain point by looking at the slope of the secant line with a smaller and smaller distance to the original point.

Definition 2.16. Let $f : X \to Y$ be a function. We call f differentiable in $x \in X$ if the following limit exists

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

We call $f'(x)$ the derivative of f in x and define it via

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

Theorem 2.17. *The following are the derivatives of the standard functions above*

- *Polynomials:* $f_1'(x) = \sum_{i=0}^{n-1} (i+1) a_{i+1} x^i$
- *Exponential:* $f_2'(x) = bae^{bx}$
- *Logarithm:* $f_3'(x) = \frac{a}{x}$
- *Sinus:* $f_4'(x) = a\ cos(x)$
- *Cosinus:* $f_5'(x) = -a\ sin(x)$

Theorem 2.18. *(Rules for derivatives) Let $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ be functions that are differentiable in x and $\alpha \in \mathbb{R}$, then the following rules apply:*

- *(linearity)* $(u + v)'(x) = u'(x) + v'(x)$
- *(scalar)* $(\alpha u)'(x) = \alpha u'(x)$
- *(product rule)* $(u \cdot v)'(x) = u'(x)v(x) + u(x)v'(x)$
- *(chain rule)* $(u \circ v)'(x) = v'(x) \cdot u'(v(x))$
- *(quotient rule)* $(\frac{u}{v})'(x) = \frac{u'(x)v(x) - u(x)v'(x)}{v(x)^2}$

Theorem 2.19. *Let $f(x) := e^{g(x)}$ where g is a differentiable function. Then,*

$$f'(x) = g'(x) \cdot e^{g(x)}.$$

Proof. We make use of the chain rule. We define $u(z) := e^z$ and $v(x) = g(x)$. Then, $u'(z) = e^z$ and $v'(x) = g'(x)$. Thus, by the chain rule we get that $f'(x) = u'(v(x))v'(x) = e^{v(x)}v'(x) = g'(x)e^{g(x)}$ □

Example 2.20. Let $f(x) := x \cdot \log(x)$. We make use of the product rule. We define $u(x) := x$ and $v(x) := \log(x)$. We obtain that $u'(x) = 1$ and $v'(x) := \frac{1}{x}$. Thus $f'(x) = u'(x)v(x) + u(x)v'(x) = log(x) + 1$.

Definition 2.21 (Maxima and minima)**.** Let $f : D \to \mathbb{R};\ x \mapsto f(x)$ be a differentiable function with $D \subseteq \mathbb{R}$. We call $x_0 \in D$

- a local minimum of f if[1] $f'(x_0) = 0 \wedge f''(x_0) > 0$
- a local maximum of f if $f'(x_0) = 0 \wedge f''(x_0) < 0$
- a global minimum of f iff $\forall x \in D :\ f(x_0) \leq f(x)$
- a global maximum of f iff $\forall x \in D :\ f(x_0) \geq f(x)$

To see one example where gradients are useful, check out the gradient descent algorithm for training neural networks.

[1] Note that $f'(x_0) = 0$ is only a necessary condition and $f''(x_0) > 0$ is only a sufficient condition. However, usually, nothing more will be relevant for you. To see that the latter is not necessary consider $f(x) = x^4$ which has a minimum at $x = 0$ however $f''(0) = 0$.

2.4 Integrals

Integrals measure the area under curves. They are computed by looking at the areas of rectangles under the curve with smaller and smaller widths.

Definition 2.22 (Riemann integral)**.** Let f be a continuous function. Then, we define the integral between a and b with $a < b$ as

$$\int_a^b f(x)\ dx := \lim_{n \to \infty} \frac{b-a}{n} \sum_{i=0}^{n} f(a + \frac{i(b-a)}{n})$$

Theorem 2.23. *(Rules for integrals)*
Let $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ be continuous functions, $\alpha \in \mathbb{R}$ and $a < b < c$, then the following rules apply:

- *(linearity)* $\int_a^b (u+v)(x)\ dx = \int_a^b u(x)\ dx + \int_a^b v(x)\ dx$
- *(scalar)* $\int_a^b (\alpha u)(x)\ dx = \alpha \int_a^b u(x)\ dx$
- *(additivity)* $\int_a^c u(x)\ dx = \int_a^b u(x)\ dx + \int_b^c u(x)\ dx$

Definition 2.24 (Primitive integral)**.** Let f be a continuous function. Then, there exists a function F such that for all a and b with $a < b$ as

$$\int_a^b f(x)\ dx := F(b) - F(a)$$

It could be that there are many functions that satisfy this relationship. However, the following theorem shows that they are all almost the same.

Theorem 2.25 (Primitive integral)**.** *The primitive integral is unique up to a constant. For a given function f and corresponding primitive functions F, F^*, there exists a constant c such that*

$$F = F^* + c$$

The following is the most central result of every calculus class and it is indeed beautiful. It builds a bridge between the differential and integral calculus as basically inverse operations.

Theorem 2.26 (The fundamental theorem of calculus)**.** *Let f be a differentiable function and $a < b$, then*

$$\int_a^b f'(x)\ dx := f(b) - f(a)$$

Since differentiating and integrating functions are basically inverse operations, there is a counterpart to the product rule and the chain rule. We will only look at the counterpart to the product rule, which is partial integration. The counterpart to the chain rule is called the *substitution rule* but we won't cover that.

Theorem 2.27. *(Partial integration)*
Let $u : \mathbb{R} \to \mathbb{R}$ and $v : \mathbb{R} \to \mathbb{R}$ be differentiable functions $a < b$, then the following holds:

$$\int_a^b u(x)v'(x)\ dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)\ dx$$

Not every function can be Riemann integrated. A famous counterexample is the Dirichlet function

$$1_{\mathbb{Q}}(x) := \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{else.} \end{cases}$$

The Dirichlet function is not Riemann integrable but Lebesque integrable, which is a generalization of the Riemann integral but far less intuitive.

Theorem 2.28. *The following are the primitive integrals of the standard functions above*

- *Polynomials:* $F_1(x) = c + \sum\limits_{i=0}^{n} \frac{a_i}{i+1} x^{i+1}$
- *Exponential:* $F_2(x) = c + \frac{a}{b} e^{bx}$
- *Logarithm:* $F_3(x) = c + (x\ ln(x) - x)$
- *Sinus:* $F_4(x) = -a\ cos(x)$
- *Cosinus:* $F_5(x) = a\ sin(x)$

2.5 Multi-dimensional derivatives and integrals

Everything we do with functions $f : \mathbb{R} \to \mathbb{R}$ in one dimension, can also be done with multi-dimensional functions $g : \mathbb{R}^m \to \mathbb{R}$.
The trick is to look at g as if it were m different functions, where for each g_i, all other variables $j \neq i$ are constants.

Definition 2.29 (Gradients)**.** Let $g : \mathbb{R}^m \to \mathbb{R}$ be a function, we define the gradient of g as

$$\text{grad } g := (\frac{df}{dx_1}, \dots, \frac{df}{dx_m}) = \nabla g$$

We can also define a multi-dimensional integral but the theory behind that is more difficult to grasp. Thus, I only provide you with one example. If you want to learn more, check out the theorem by Fubini.

Example 2.30 (Multi-dimensional integral)**.** Let $g(x, y) := x^2 + y$ and $D := [0, 1] \times [0, 2]$, then

$$\int_D g(x, y)\ d(x, y) = \int_0^1 (\int_0^2 g(x, y))\ dy\ dx = \frac{8}{3}$$

Exercises

Excercise 2.31. *Show that* $\mathbb{N}$ *and* $\{2n \mid n \in \mathbb{N}\}$ *have equal cardinality.*

Excercise 2.32. *Assume the product rule and the chain rule and derive the quotient rule.*

Excercise 2.33. *Compute derivatives of the following functions*

- $f(x) = 3x^2 - 6x$
- $f(x) = x^3 - 3x^2$
- $f(x) = 3\ sin(x)$
- $f(x) = 3\ sin(2x)$
- $f(x) = e^x - x$
- $f(x) = e^x x$
- $f(x) = ln(g(x))$ *for some function* g.
- $f(x) = \frac{e^x}{x}$
- $f(x) = e^{-3x}x$

Excercise 2.34. *Compute maxima and minima of the following functions. Are these global or only local extrema?*

- $f(x) = 3x^2 - 6x$
- $f(x) = 3\ sin(x)$
- $f(x) = e^x - x$

Excercise 2.35 (Hard: Cantors Theorem)**.** *Let* A *be a set, then show that*

$$|A| \leq |2^A|$$

Hint: Assume there is a surjective mapping f *from* A *to its powerset and consider*

$$S := \{a \in A \mid a \notin f(a)\}$$

Excercise 2.36. *Let* $f(x) = x^3$*. Show that for all* $a \in \mathbb{R}$ *holds*

$$\int_{-a}^{a} f(x)\ dx = 0$$

Excercise 2.37. *Compute the primitive integrals for the following functions:*

- $f(x) = 2x^2 + 3$
- $f(x) = e^{-x}$
- $f(x) = |x| := \begin{cases} x & if\ x \geq 0 \\ -x & else \end{cases}$
- $f(x) = \frac{1}{x}$
- $f(x) = sin(3x)$

Excercise 2.38. *Use integration by parts to derive the following integrals*

- $\int_1^e ln(x)\ dx \quad$ *hint:* $ln(x) = ln(x) \cdot 1$
- $\int_0^\pi x\ cos(x)\ dx$

3 Linear algebra

Vectors and matrices are the central concepts in linear algebra. Vectors are points in multidimensional space. Matrices allow to define linear mappings between two vector spaces.

3.1 Vector space

Definition 3.1 (vector space)**.** Let $\mathbb{R}$ be the real numbers and V be a set. Let furthermore $+ : V \times V \to V$ and $\cdot : \mathbb{R} \times V \to V$ be functions. We call V a vector space over $\mathbb{R}$ if for all $u, v, w \in V$ and $\alpha, \beta \in \mathbb{R}$ holds:

- $(u + v) + w = u + (v + w)$
- There exists a $0_V \in V$, such that $0_V + v = v + 0_V = v$
- There exists $-v \in V$ such that $v - v = 0_V$
- $u + v = v + u$
- $\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$
- $(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$
- $(\alpha\beta) \cdot v = \alpha \cdot (\beta \cdot v)$
- $1 \cdot v = v$

This is a very general definition but for our purposes, it is enough to understand one specific vector space, namely $\mathbb{R}^n$ for $n \in \mathbb{N}$.

Theorem 3.2. *We can see $\mathbb{R}^n$ as a vector space with the following two operations. Let $x, x' \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, then we define*

- *Addition:* $x + x' := (x_1 + x'_1, \ldots, x_n + x'_n)$
- *Multiplication:* $\alpha x := (\alpha x_1, \ldots, \alpha x_n)$

Definition 3.3 (length)**.** For a given vector $x \in \mathbb{R}^n$, we define the length of the vector via:

$$|x| := \sqrt{x_1^2 + \cdots + x_n^2}$$

Definition 3.4 (scalar product)**.** For two vectors $x, y \in \mathbb{R}^n$, we define their scalar product as:

$$\langle x, y \rangle := x_1 y_1 + \cdots x_n y_n$$

It is easy to see that $\sqrt{\langle x, x, \rangle} = |x|$.

Theorem 3.5 (Geometric interpretation of scalar product)**.** *For two vectors $x, y \in \mathbb{R}^n$ holds*

$$cos(\alpha) = \frac{\langle x, y \rangle}{|x||y|}$$

where α describes the angle between x and y. Thus, two vectors are orthogonal to each other if and only if

$$\langle x, y \rangle = 0$$

Theorem 3.6 (Properties of scalar product)**.** *For two vectors $x, y, z \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ holds*

- *Symmetry:* $\langle x, y \rangle = \langle y, x \rangle$
- *Bilinearity:* $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

The following concepts might be a bit difficult at first. This is why, we only look at the two-dimensional case. However, for all of these definitions, there are generalizations.

Definition 3.7 (Linear combination)**.** For vectors $x_1, \dots, x_k \in \mathbb{R}^n$, we can define their linear combinations as

$$\operatorname{span}(x_1, \dots, x_k) = \{\sum_{i=1}^{k} \alpha_i x_i \mid \alpha_1, \dots, \alpha_k \in \mathbb{R}\}$$

Definition 3.8 (Linear independence)**.** We call vectors $x_1, \dots, x_k \in \mathbb{R}^n$ linear independent, if for all $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ holds

$$\sum_{i=1}^{k} \alpha_i x_i = 0 \Rightarrow \alpha_1 = \dots = \alpha_k = 0$$

Definition 3.9 (Generating system)**.** Vectors $x_1, \dots, x_k \in \mathbb{R}^n$ are called a generating system of $\mathbb{R}^n$ if for all $z \in \mathbb{R}^n$, there exist $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ such that

$$\sum_{i=1}^{k} \alpha_i x_i = z.$$

Definition 3.10 (Basis)**.** We call vectors $x_1, \dots . x_k \in \mathbb{R}^n$ a basis of $\mathbb{R}^n$ if

- $x_1, \dots, x_k$ are linearly independent
- $x_1, \dots, x_k$ are a generating system

One can show that all bases of a vector space have the same number of vectors. This provides a way to define the dimension of a vector space, it is defined by the number of vectors that form a basis. For example, in $\mathbb{R}^n$, the standard basis with $e_1, \dots, e_n$ has n many vectors. Thus, the dimension of $\mathbb{R}^n$ is n.

3.2 Matrices

Matrices provide a very convenient notation to describe linear functions between different vector spaces. What is a linear functions?
A function $f : \mathbb{R}^n \to \mathbb{R}^m$ between two vector spaces is called linear if for all $v, v' \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ holds

$$f(\alpha v + v') = \alpha f(v) + f(v').$$

One can show that every linear function can be described via a matrix multiplication. You can think of a matrix like of a table with rows and columns. In a given row-column combination, there is a scalar value as an entry.

Definition 3.11 (Matrix)**.** A matrix A is an element of $A \in \mathbb{R}^{m \times n}$. m describes the number of rows and n the number of columns. The matrix looks as follows

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Definition 3.12. Let $A, B \in \mathbb{R}^{m \times n}$ be two matrices and $\alpha \in \mathbb{R}$ a scalar. We can define the following two operations:

$$A + B = \begin{pmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} & \cdots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} & \cdots & a_{2,n} + b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} + b_{m,1} & a_{m,2} + b_{m,2} & \cdots & a_{m,n} + b_{m,n} \end{pmatrix}$$

and

$$\alpha A = \begin{pmatrix} \alpha\, a_{1,1} & \alpha\, a_{1,2} & \cdots & \alpha\, a_{1,n} \\ \alpha\, a_{2,1} & \alpha\, a_{2,2} & \cdots & \alpha\, a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha\, a_{m,1} & \alpha\, a_{m,2} & \cdots & \alpha\, a_{m,n} \end{pmatrix}$$

We can also multiply two different matrices and also multiply a matrix and a vector. Important is that they have fitting dimensions. The number of columns in the first matrix must match the number of rows in the second matrix for the operations to be defined.

Definition 3.13 (Matrix multiplication)**.** Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ be matrices and $x \in \mathbb{R}^n$ be a vector. Then we can define two more operations:

$$AB = \begin{pmatrix} \sum\limits_{k=1}^{n} a_{1,k}b_{k,1} & \sum\limits_{k=1}^{n} a_{1,k}b_{k,2} & \cdots & \sum\limits_{k=1}^{n} a_{1,k}b_{k,p} \\ \sum\limits_{k=1}^{n} a_{2,k}b_{k,1} & \sum\limits_{k=1}^{n} a_{2,k}b_{k,2} & \cdots & \sum\limits_{k=1}^{n} a_{2,k}b_{k,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum\limits_{k=1}^{n} a_{m,k}b_{k,1} & \sum\limits_{k=1}^{n} a_{m,k}b_{k,2} & \cdots & \sum\limits_{k=1}^{n} a_{m,k}b_{k,p} \end{pmatrix}$$

and

$$Ax = \begin{pmatrix} \sum\limits_{k=1}^{n} a_{1,k}x_k \\ \sum\limits_{k=1}^{n} a_{2,k}x_k \\ \vdots \\ \sum\limits_{k=1}^{n} a_{m,k}x_k \end{pmatrix}$$

Note that generally matrix multiplication is not commutative. Even if the dimensions fit and $A, B \in \mathbb{R}^{n\times n}$, still generally

$$AB \neq BA$$

Your task in the afternoon will be to find examples of such matrices.

3.3 Linear equations

Why are matrices such an important concept? Namely, because they provide an intuitive way to describe linear equations. Look at the following example:

Example 3.14. Father and son are together 34. Their age difference is 26. How old are they?

There are different ways to approach this question.

Solving by hand

Intuitively, you pack this information into equations. Let x=father and y=son, then we can formalize the problem as

$$x + y = 34 \text{ and } x - y = 26.$$

To solve the equation we isolate $x = y + 26$ and then substitute it for x in the other equation. Thus,

$$\begin{aligned} x + y = 34 &\Leftrightarrow y + 26 + y = 34 \\ &\Leftrightarrow 2y = 8 \\ &\Leftrightarrow y = 4 \end{aligned}$$

Gaußian elimination

Alternatively, we can see the linear equation as the result of a matrix-vector multiplication.

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad v = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 34 \\ 26 \end{pmatrix}$$

Then, we can write our equation as:

$$Av = b$$

Because we know that x and y are still unknown, we can ignore them for now. Instead, we write:

$$\left(\begin{array}{cc|c} 1 & 1 & 34 \\ 1 & -1 & 26 \end{array}\right)$$

The so-called *Gaussian elimination* provides a simple algorithm to solve such equations by three operations

- Multiply a row with a scalar $\alpha \neq 0$.
- Swap to rows i and j.
- Add one row multiplied by a scalar α to another row.

Your goal with these three operations should be to obtain a matrix in diagonal form:

$$\left(\begin{array}{ccccc|c} 1 & * & * & \cdots & * & * \\ 0 & 1 & * & \cdots & * & * \\ 0 & 0 & 1 & \cdots & * & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & * \end{array}\right)$$

Let's do this for our example, we start with our matrix above

$$\left(\begin{array}{cc|c} 1 & 1 & 34 \\ 1 & -1 & 26 \end{array}\right)$$

Then, we add (-1) times row one to row two and obtain

$$\left(\begin{array}{cc|c} 1 & 1 & 34 \\ 0 & -2 & -8 \end{array}\right)$$

Then, we multiply the second row with $-1/2$ and obtain

$$\left(\begin{array}{cc|c} 1 & 1 & 34 \\ 0 & 1 & 4 \end{array}\right)$$

Thus, we can derive

$$0 \cdot x + 1 \cdot y = 4$$

Thus, $y = 4$. By inserting this in the equation above, we get

$$1 \cdot x + 1 \cdot 4 = 34$$

And therefore $x = 30$.

While this looks like extra work, this approach is very efficient, especially if you solve linear equations with more variables.

Inverse

You may wonder if there is a simpler way of solving this equation by taking the inverse.

$$Av = b \Rightarrow v = A^{-1}b.$$

The problem is that this inverse does not always exist and it is usually hard to compute. But for two-dimensional matrices, there exists a simple formula for the inverse. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The inverse of A (if it exists) is given by

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Where the determinant of A is defined as

$$\det(A) = ad - bc$$

We see that the inverse is not defined if the determinant is zero. This is true more generally for arbitrary matrices. The determinant is a key concept in linear algebra but we will not introduce it here more deeply because its general definition is complex.

Let's compute the inverse for our example above:

$$\det(A) = 1 \cdot (-1) - 1 \cdot 1 = -2$$

Consequently,

$$A^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Hence,

$$\begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 34 \\ 26 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} -60 \\ -8 \end{pmatrix} = \begin{pmatrix} 30 \\ 4 \end{pmatrix}$$

3.4 Eigenvalues and Eigenvectors

For all matrices, there exist special vectors that are not really transformed by the matrix but only stretched. These vectors are called *Eigenvectors.* They characterize your matrix.

Definition 3.15 (Eigenvector and Eigenvalue)**.** Let A be a matrix and x be a vector. We call x and *Eigenvector* of A if there exists a scalar $\lambda \in \mathbb{R}$ such that

$$Ax = \lambda x.$$

We call the scalars λ for which such x exist the *Eigenvalues* of A.

There is a lot of theory on how to compute Eigenvalues and Eigenvectors, I only want to provide you with an intuition here. The Eigenvalues can be determined by the characteristic polynomial, using the following

$$Ax = \lambda x \Leftrightarrow Ax = \lambda Ex \Leftrightarrow Ax - \lambda Ex = 0 \Leftrightarrow (A - \lambda E)x = 0 \Leftrightarrow \det(A - \lambda E) \neq 0$$

Interesting properties of Eigenvalues and Eigenvectors

There are various theorems showing properties of Eigenvectors and Eigenvalues. I want to at least highlight some of them:

- The number of Eigenvalues is smaller or equal to the number of dimensions of the matrix A.
- There is often an orthogonal basis of Eigenvectors for vector spaces, which offer a more natural description of the system in relation to the considered matrix.
- Every matrix A of full rank where all eigenvalues differ, can be diagonalized, this means there exist a diagnoal matrix D

  $$D = Q^{-1}AQ$$

  for some invertible matrix Q.

Exercises

Excercise 3.16. *Give an example of two matrices $A, B \in \mathbb{R}^{2\times 2}$ such that*

$$AB \neq BA$$

Excercise 3.17. *If possible, multiply the following matrices/vectors:*

i) $\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

ii) $\begin{pmatrix} 0 & 1 \\ 2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

iii) $\begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$

iv) $\begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

v) $\begin{pmatrix} 2 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 & 2 \end{pmatrix}$

Excercise 3.18. *Show that $v_1, \dots, v_l \in \mathbb{R}^n$ is a generating system of $\mathbb{R}^n$ if and only if $e_1, \dots, e_n \in span(v_1 \dots, v_l)$.*

Excercise 3.19. *Compute the following:*

- $\langle \begin{pmatrix} 7 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 7 \end{pmatrix} \rangle$
- $\langle \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rangle$

Excercise 3.20. *Show that if $v_1, \dots, v_l \in \mathbb{R}^n$ is linear independent, then for all $v \in span(v_1, \dots, v_n)$, there exist unique $\alpha_1, \dots, \alpha_l$, such that*

$$v = \alpha_1 v_1 + \cdots + \alpha_l v_l$$

Excercise 3.21. *If possible compute the inverse matrix for*

- $\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}$
- $\begin{pmatrix} 1 & 3 \\ -2 & -6 \end{pmatrix}$

Excercise 3.22. *You go to the fruit market and buy 10 bananas and 2 watermelons, you pay 11 Euro. Your friend goes to the same supermarket (prices unchanged) and buys 3 bananas and 1 watermelon for 4.5 Euro. How much do bananas and watermelons cost?*

Excercise 3.23. *Solve the following set of linear equations*

- $x + y + z = 6$
- $-2x + 3y - z = -3$
- $3x - 5y - 2z = 7$

Excercise 3.24. *Solve if possible*

1. *Consider the following three equations*
   - $x + y = 5$
   - $2x - y = 1$
   - $-x + 2y = 6$
2. *Consider the following two equations*
   - $7x - y = 8$
   - $-14 + 2y = -16$

4 Probability theory

4.1 Urn models

Urns are the traditional model to introduce probabilities. But we can just as well take card games.

Imagine we have a classical poker set with 52 distinct cards (no jokers). You are allowed to draw 5 different cards. How many ways are there to draw five cards?

Drawing with order and with replacement

Say that the ordering of the cards matters. Also, whenever you draw a card, you have to put it back and the deck is shuffled again. Well, the first card could be one of 52 cards, just like the second, the third, the fourth, and the fifth. The set that describes all options is

$$S := \{(c_1, \dots, c_5) \mid c_1, \dots, c_5 \in \{1, \dots, 52\}\}$$

The size of this set is therefore 52^5.

More generally, there are

$$n^k$$

ways to draw k cards with replacement in order from a deck of n cards.

Drawing with order and without replacement

Say that the ordering of the cards matters but this time, the cards are drawn cards are not returned to the card deck. Well, the first card could be one of 52 cards, for the second draw 51 possible cards are left, for the third 50, and so on. The set that describes all options is

$$S := \{(c_1, \dots, c_5) \mid c_1, \dots, c_5 \in \{1, \dots, 52\} \text{ and } c_i \neq c_j \text{ for } i \neq j\}$$

The size of this set is therefore $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$.

More generally, there are

$$\frac{n!}{(n-k)!}$$

ways to draw k cards without replacement in order from a deck of n cards.

Drawing without order and without replacement

Say that the ordering of the cards does not matter nor are the drawn cards returned to the card deck. All possible draws can be described by the following set

$$S := \{\{c_1, \dots, c_5\} \mid c_1, \dots, c_5 \in \{1, \dots, 52\} \text{ and } c_i \neq c_j \text{ for } i \neq j\}$$

How big is this set? The best way to approach this is to look at the case where order matters and look at how many of the 5 drawn cards are equivalent. Remember, there are $\frac{n!}{(n-k)!}$ ways for drawing in order without replacement.

For a given such draw, there are $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ equivalent draws. The first card could be at five other positions and we still had the same deck, the second card at four, the third card at three, the fourth card at two, and the last card has already been switched back and forth by all the other cards. So, we obtain $(52 \cdot 51 \cdot 50 \cdot 49 \cdot 48)/(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)$ possible draws.

More generally, there are

$$\binom{n}{k} := \frac{n!}{(n-k)!\,k!}$$

ways to draw k cards without replacement without order from a deck of n cards.

Drawing without order and with replacement

There are $\binom{n+k-1}{k}$ ways to draw k cards with replacement without order from a deck of n cards.

4.2 Probability space

Here we introduce the basics of probability theory. Mostly, we will assume that the sample space Ω is finite, i.e. we will always assume that $|\Omega| < \infty$.[2]

Definition 4.1 (Sample space)**.** A sample space is a set of possible outcomes. In philosophical contexts, this is often a set of possible worlds. This set is usually denoted Ω. We demand that $\Omega \neq \emptyset$.

Example 4.2. Let's say we have two dice and we want to describe all their possible outcomes. We can define

$$\Omega := \{(d_1, d_2) \mid d_1, d_2 \in \{1, \ldots, 6\}\}$$

Definition 4.3 (sigma-Algebra)**.** For us the σ-Algebra is just $\mathcal{F} := \mathcal{P}(\Omega) = 2^\Omega$ in every case. $\mathcal{P}(\Omega)$ denotes the powerset of Ω, which is the set of all subsets of Ω meaning $\mathcal{P}(\Omega) := \{A \mid A \subseteq \Omega\}$. The elements of $\mathcal{P}(\Omega)$ are usually called events.

Example 4.4. The σ Algebra of our Ω contains trivial elements like $\emptyset$ and Ω but also more interesting ones like $\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$, the set that contains all doubles.

Definition 4.5 (Probability measure)**.** A Probability measure P is a function

$$P : \mathcal{P}(\Omega) \to [0, \infty],\ A \mapsto P(A)$$

that satisfies the following two conditions:

- $P(\Omega) = 1$ and
- if $A, B \subseteq \Omega$ with $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Example 4.6. If we assume that the coins are fair, we can define the probability measure that gives each outcome the same probability, that is for all $\omega \in \Omega$:

$$P(w) = \frac{1}{36}$$

Definition 4.7 (Probability space)**.** A probability space is a triple $(\Omega, P, \mathcal{P}(\Omega))$ where Ω is a sample space, P is a probability measure on Ω, and $\mathcal{P}(\Omega)$ the powerset of Ω.

Theorem 4.8. *Let $A, B \subseteq \Omega$. The following are interesting properties of probability measures:*

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) \leq 1$
- $P(A) = \sum_{\omega \in A} P(\{\omega\})$
- *If* $\bigcup_{i=1}^{n} A_i = \Omega$ *and* $\forall i, j \in \{1, \ldots, n\}$ *with* $i \neq j$ *holds* $A_i \cap A_j = \emptyset$ *then for any set* $B \subseteq \Omega$ *holds* $P(B) = \sum_{i=1}^{n} P(A_i \cap B)$

4.3 Random variables

Definition 4.9 (Random Variable)**.** A random variable X is a (measureable)[3] function $X : \Omega \to \mathbb{R}$, $\omega \mapsto X(\omega)$. We define for any $x \in \mathbb{R}$ $[X = x] := X^{-1}(x) \subseteq \Omega$[4] or more generally $[X = A] = X^{-1}(A) \subseteq \Omega$ for $A \subseteq \mathbb{R}$. Thus,

$$P(X = x) = P(X^{-1}(x)).$$

[2] As always things get much more difficult but also interesting if we drop this assumption. However, this would demand a lot more work. If you are interested you should first acquire some basics in measure theory. The σ-Algebra we are working with will in infinite cases usually not be the powerset of Ω. (The coolest thing in measurement theory related to that: The Banach-Tarski Paradoxon.)

[3] Again, in your case any function is measureable. The requirement is $X^{-1}(A) \in \mathcal{F}$ for all A in the Borel-σ-Algebra of $\mathbb{R}$.

[4] This is only well defined if X is a measurable function, so you can see how the puzzle fits together.

Example 4.10. Intuitively, random variables are just a convenient way to talk about the relevant events. For example, we can define the following two random variables

$$X_1 : \Omega \to \mathbb{R}, X_1(d_1, d_2) = d_1$$

and

$$X_2 : \Omega \to \mathbb{R}, X_2(d_1, d_2) = d_2.$$

So the random variables X_1 and X_2 just map possible outcomes to the respective values of die 1 and die 2 in the outcome.

Now we can talk about events efficiently. What is the event where die 1 shows 5 and die 2 shows 3? We can express this as:

$$[X_1 = 5, X_2 = 3] = X_1^{-1}(5) \cap X_2^{-1}(3)$$

Definition 4.11 (Joint probability mass function)**.** Let $(\Omega, P, 2^\Omega)$ be a probability space where $|\Omega| < \infty$, 2^Ω denotes the powerset of Ω, and let $X : \Omega \to \mathbb{R}$ be a random variable. The *distribution* P_X of X is defined by

$$P_X(A') := P(\{\omega \in \Omega : X(\omega) \in A'\}) \text{ for all } A' \subseteq \mathbb{R}$$

The so-called *probability mass function* $f_X : \mathbb{R} \to [0, 1]$ of X is used among other things to visualize distributions. It is straightforwardly defined by

$$f_X : x \mapsto P(X = x).$$

For n random variables $X_1, \ldots, X_n$ and real numbers $x_1, \ldots, x_n \in \mathbb{R}$ with $n \in \mathbb{N}$ the joint probability mass function is defined as:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) := P(X_1 = x_1 \text{ and } \cdots \text{ and } X_n = x_n)$$

Example 4.12. The probability mass function of X_1 is just

$$f_{X_1}(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, \ldots, 6\} \\ 0 & \text{else} \end{cases}$$

Definition 4.13 (Identically distributed)**.** Let X, Y be random variables. We say that X and Y are distributed identical if $\forall a \in \mathbb{R}$ holds:

$$f_X(a) = f_Y(a)$$

Notice that this does not imply that the random variables are identical. They just assign the same probabilities to particular values, which could be to completely different events. Also, on zero measure sets (events that "almost never" happen) the values could be completely off.

Example 4.14. The random variables X_1 and X_2 describing the two dice are identically distributed because

$$f_{X_1}(x) = f_{X_2}(x)$$

Definition 4.15 (Expected value)**.** Let X be a random variable as defined above. Then,

$$\mathbb{E}[X] := \sum_{x \in \mathbb{R}} P(X = x) \cdot x.$$

This is only one way to state the expected value. As one can prove the following is equivalent:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} P(\{\omega\}) \cdot X(\omega).$$

Notice that we will regularly use $P(\omega)$ instead of $P(\{\omega\})$, which is only an abbreviation but does not change the fact that P is only defined for sets of outcomes.
Intuitively the expected value is something like a weighted average of the outcomes.

Example 4.16. The expected value of die X_1 can be described by

$$\mathbb{E}(X) = \frac{1}{6} \sum_{i=1}^{6} i = 3.5$$

Theorem 4.17 (Linearity of expected value)**.** *Let X, Y be random variables and $a, b \in \mathbb{R}$. Important properties of the Expected Value:*

- $\mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}[X] + b$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Definition 4.18 (Variance and covariance)**.** Let X, Y be random variables. The variance of X is defined as:

$$var(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

The variance tells you about how far the outcomes of the random variable are spread from the average value. Moreover, we define the covariance of X, Y as follows:

$$cov(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Covariance is a lot harder to interpret. I would say it mainly shows the linear relationships between two random variables. Notice, that if the Covariance is zero we say that the two random variables are uncorrelated. This does not mean that they are independent. Independence on the other side implies a covariance of zero.

Theorem 4.19. *Variance and covariance have the following properties:*

- $var(X) = cov(X, X)$.
- $var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Example 4.20. We can describe the variance of X_1 by Variance

$$\mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sum_{i=1}^{6} i^2 P(X = i) - (\frac{7}{2})^2 = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) - (\frac{7}{2})^2 \approx 2.92.$$

We can moreover describe the covariance between X_1 and X_2 by

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] \overset{indep}{=} \mathbb{E}[X_1]\mathbb{E}[X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2] = 0.$$

4.4 Conditional probabilities

Definition 4.21 (Conditional probabilities)**.** Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. For any $A, B \subseteq \Omega$ with $P(B) > 0$ we can define the probability of A given that B as follows:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Definition 4.22. Independence
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Let $A, B \subseteq \Omega$. We call A independent of B if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

One can derive by this that if $P(B) > 0$ then

$$A, B \text{ are independent iff } P(A \mid B) = P(A).$$

This has a lot of intuitive appeal! It says that event A is independent of event B if knowing that B happened does not tell us anything about whether event A happens.

Definition 4.23 (Independence of random variables)**.** Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Let X and Y be two random variables. We call X independent of Y if and only if

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \ \ \forall x, y \in \mathbb{R}$$

Very often we do not specify the sampling space exactly and instead, we start with the random variables and the probability distributions of these random variables. There is a theorem showing that there exists a sample space on which these random variables are well-defined.

Theorem 4.24 (General product rule)**.** *Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. For any $A_1, \dots, A_n \subseteq \Omega$ with $\bigcap\limits_{i=1}^{n-1} A_i \neq \emptyset$ with $n \in \mathbb{N}$ holds the following:*

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i \mid \bigcap_{j=1}^{i-1} A_j).$$

Note that we use the convention that $\bigcap\limits_{j=1}^{0} A_j := \Omega$

Theorem 4.25 (Bayes Theorem)**.** *Let $\bigcup\limits_{i=1}^{n} H_i = H$ for some set H in Ω s.t. $\forall i, j \in \{1, \dots, n\}$ with $i \neq j$ holds $H_i \cap H_j = \emptyset$. Let moreover $E \subseteq \Omega$. Then, the following holds:*

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{\sum\limits_{i=1}^{n} P(H_i)P(E \mid H_i)}$$

Definition 4.26 (Conditional expectation)**.** Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Moreover, let X, Y be random variables and $y \in Image(Y)$ with $P(Y = y) > 0$. Then, we can define the conditional expectation of X given $Y = y$ as follows:

$$\mathbb{E}[X \mid Y = y] := \sum_{x \in \mathbb{R}} P(X = x \mid Y = y)x$$

Generally, we can define the function

$$\mathbb{E}[X \mid Y] : A \to \mathbb{R};\ y \mapsto \mathbb{E}[X \mid Y = y]$$

Where $A = \{y \in Image(Y) \mid P(Y = y) > 0\}$. If we define this more carefully[5] we can even get a random variable over all $\mathbb{R}$.
Intuitively the conditional expectation of X on $Y = y$ expresses the expected value of the random variable X given that the random variable Y happened to be y.

Example 4.27. Let's define a third random variable, namely the sum of die 1 and die 2 by $X_3 := X_1 + X_2$. Then, we can compute the conditional expectation of X_3 given X_1 by

$$\mathbb{E}[X_3 \mid X_1] : \{1, \dots, 6\} \to \mathbb{R},\ i \mapsto \mathbb{E}[X_3 \mid X_1 = i] = i + 3.5.$$

Definition 4.28 (Conditional independence)**.** Let X, Y, and Z be random variables. Then, we define the conditional independence of X of Y given Z as follows:

$$X \perp\!\!\!\perp Y \mid Z \text{ if and only if } P(X = x, Y = y \mid Z = z) = P(X = x \mid Z) \cdot P(Y = y \mid Z = z)$$

for all $x, y, z \in \mathbb{R}$.

4.5 Common distributions and central results

Example 4.29. Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space and X be a random variable. The following are some common and interesting distributions.

- We call P uniformly distributed on Ω ($P \sim Unif(\Omega)$) iff $P(\omega) = \frac{1}{|\Omega|} \quad \forall \omega \in \Omega$.
- We call P_X (often also X) Bernoulli distributed ($X \sim Ber(p)$) with $p \in [0, 1]$ iff P(X=1)=p=1-P(X=0).
- We call P_X (often also X) Binomially distributed ($X \sim Bin(p, n)$) with $p \in [0, 1], n \in \mathbb{N}$ iff $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$.
- We call P_X (often also X) Geometrically distributed ($X \sim Geo(p)$) with $p \in [0, 1]$ iff $P(X = k) = p(1 - p)^{k-1}$.

[5] In maths this is defined not constructively but by a condition on the function. Consult the wiki article on conditional expectation to learn more about it.

A standard Bernoulli trial with parameter p would be a coin flip with probability p showing heads. A Binomially distributed variable with parameters p and n would be the sum of n many independent Bernoulli trials with parameter p. A geometrically distributed variable with parameter p could be interpreted as assigning to each k the probability that after k Bernoulli trials it turns out heads for the first time.

Theorem 4.30 (Law of large numbers)**.** *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables (i.i.d) with $E[X_1^2] < \infty$ then for $\overline{X}_n := \sum\limits_{i=1}^{n} \frac{X_i}{n}$ holds:*

$$\lim_{n\to\infty} \overline{X}_n = E[X_1] \textit{ almost certainly.}$$

Infinite sample spaces

Often Ω is not finite. Think of a dart board. The dart could land in arbitrarily fine-grained ways on the dart board. Or, if you are me, anywhere next to the dart board. What to do in such scenarios?

Luckily, we can generalize our notion of a mass function to a so-called probability density function. With density functions, each point on the dartboard can have a probability of zero but an area of the dartboard can have a positive probability. This works by so-called density functions that can be integrated over specific areas.

Example 4.31. The following are classical examples of distributions where we need density functions:

- We call X uniformly distributed on the interval $[a, b]$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{else} \end{cases}$$

- We call X exponentially distributed $X \sim exp(\alpha)$ with parameter α if

$$f_X(x) = \alpha\, e^{-\alpha x}$$

- We call X normal distributed $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ^2 if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Theorem 4.32 (Central limit theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2 (where $0 < \sigma^2 < \infty$).*
Define the sample mean as:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The Central Limit Theorem states that as $n \to \infty$, the distribution of the standardized sample mean approaches a standard normal distribution. Specifically:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\textit{in distribution}} N(0, 1)$$

In words, as the sample size increases, the distribution of the sample mean $\overline{X}_n$ approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, regardless of the original distribution of the X_i's, provided the X_i's have a finite variance.

This explains the central importance of the normal distribution. It is a good proxy for all processes that can be described as the sum of identical and independent processes.

Exercises

Excercise 4.33. *Say you have three coins and you want to toss them. Provide:*

1. *A sample space for this setting.*
2. *A probability measure such that the outcomes are uniformly distributed.*
3. *Describe the three coins each as an individual random variable. Assign* 0 *for heads and* 1 *for tails.*
4. *Describe the event that exactly two coins land tails. Use both the notation with random variables and also describe the event by its concrete outcomes.*
5. *Define a random variable* S *that counts the number of tails in the three coin flips. Describe the mass function of this random variable.*
6. *What is the expected value of* S*? What is its variance?*
7. *Describe the probability of coin 3 landing tails if you know that two of the coins landed tails.*
8. *Show that the random variables describing coin 1 and coin 2 are independent.*
9. *Compute the conditional expectation of* S *given you know that coin 1 has landed tails. Describe the conditional expectation of* S *as a function of coin 1.*
10. *Which distribution fits the description of a single coin? Which distribution fits the sum of several coins?*

Excercise 4.34. *Show that*

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Excercise 4.35. *Show that for* $\alpha > 0$ *holds*

$$\int_0^\infty \alpha e^{-\alpha x}\, dx = 1$$

Excercise 4.36. *Say there is a test that checks whether someone took cannabis or not. If a person is taking cannabis, the test is to 90% correct. If a person is not taking cannabis, the test is to 80% correct. Assume that the prevalence of cannabis consumption in the population is* 5%.
Imagine a random person walking on the street is tested positive. What is the probability that she is taking cannabis?

Excercise 4.37. *Let* $(\Omega, P, \mathcal{P}(\Omega))$ *be a probability space. Show that for any* $A_1, \dots, A_n \subseteq \Omega$ *with* $\bigcap\limits_{i=1}^{n-1} A_i \neq \emptyset$ *with* $n \in \mathbb{N}$ *holds the following:*

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i \mid \bigcap_{j=1}^{i-1} A_j).$$

Note that we use the convention that $\bigcap\limits_{j=1}^{0} A_j := \Omega$

Excercise 4.38. *Prove the following basic results only using the Kolmogorov Axioms. For all* $A, B \subseteq \Omega$ *holds*

i $P(\emptyset) = 0$

ii If $A \subset B$*, then* $P(A) \leq P(B)$.

iii $P(A) \leq 1$

iv $P(A^C) = 1 - P(A)$

v $P(A) = \sum\limits_{\omega \in A} P(\{\omega\})$ *given* $|A| < \infty$

vi $P(A \cup B) = P(A) + P(B) - P(A \cap B)$