PROBABILITY 2021

JUHAN ARU

1

[1]All kinds of feedback, including smaller or bigger typos, is appreciated - juhan.aru@epfl.ch. In writing these notes I have consulted notes of I. Manolescu (Fribourg), Y. Velenik (Geneva), A. Eberle (Bonn) (all on their websites) and the book by R. Dalang & D. Conus published by EPFL press.

Section 0

Introduction

Probability theory provides a mathematical framework for studying random phenomena, i.e. everything that one cannot predict. We might not be able to predict because we don't have full information, or maybe because it's just not possible to predict. Maybe it is even a bit surprising to begin with that something precise and mathematical can be said about things we cannot predict, that interesting things can be said.

A bit of history

Currently probability theory is a rapidly developing branch of mathematics, with many applications. One could say that until 20th century probability was seen more as a part of applied mathematics, thereafter maybe more applicable mathematics and only with the last 20 years or so, and after 3 Fields medals, it has also been accepted as a branch of pure mathematics as well. Here are some questions people have asked in different periods, leaving aside very related questions that belong more to statistics:

Until 20th century, the main topic of probability were games of chance, lotteries, betting, but also questions about measurement errors started coming in:

- Should I accept the even chances for the bet that at least one six appears in 4 consecutive dice throws?
- How many lottery tickets should I buy to have even chance of winning the lottery?
- How long would it take to toss 5 consecutive heads with coin tosses?
- What can we describe the sum of small independent errors?

In fact the last question was properly answered only in the beginning of 20th century and is one of the most celebrated results of probability theory - the Central Limit Theorem. It says that under quite general conditions the sum of independent errors, when properly normalized converges to the Gaussian, also called the normal distribution. We will see this result in the course.

Over the 20th century, however topics in probability got much more diverse and rich. Here are some types of questions and models:

- Consider a rat in Manhattan that on each corner randomly chooses to go to left, right, back or forth. Will it ever return to the place he started? If there is another rat, and they are in love, and they want to find each other, how should they go about it?
- Relatedly, how to describe the diffusion of heat or a gas in terms of molecules? How does one single molecule behave, how does its trajectory look like?
- How to model flow of a gas or liquid through a porous medium, for example a gas mask or the earth?
- How to describe the fluctuations of a stock price over time?
- How quickly do diseases spread in a population? What parameters are important?

As you noticed, these questions can still be posed from a very non-mathematical perspective, but the mathematical models behind them are much richer than just a coin toss (which, I think, is already pretty interesting). We want to look into some of them.

Moreover, in 20th century probability theory also started playing a role in other parts of mathematics, through for example the so-called probabilistic method, often used to prove existence of certain objects:

- Dvoretzky's theorem: all high-dimensional convex bodies have low-dimensional ellipsoid sections.
- Existence of normal numbers for simultaneous basis: a number is said to be normal to base b, if the proportion of each digit in its expansion to base b is $1/b$, i.e in decimal expansion each digit $i = 0, 1, \ldots, 9$ appears with the same proportion. There is no concrete known number x for which this holds for $b = 2, 3$ simultaneously.

In the 21st century more new directions have entered due to interactions with computer science, for example ending in the Page-Rank search algorithm that Google uses.

At the same time also interactions with other domains of mathematics became stronger and probability started even sometimes influencing the development of some domains like complex analysis and dynamics. Here are some questions, where we still lack mathematical understanding:

- How to explain that certain structures like fractals, certain distributions like Gaussians, certain statistical symmetries like scale or rotation invariance appear in so many different contexts in nature?
- Why does deep learning work so well - e.g. why is it better than humans in GO? How far can one go?
- Are useful quantum computers theoretically possible?

The first questions is called universality. In fact the Central Limit Theorem can be seen as the basic example of it – it explains why the Gaussian distribution appears in many unrelated different contexts. You can find talks on universality by non-probabilists like T. Tao, by mathematical physicists like T. Spencer, and probabilists like W. Werner. I find it already inspiring that we can say anything mathematically meaningful about such a vague question. I find it's a question in the spirit of today's mathematics - we try to mathematically understand not only structures like pure symmetries, not only pure randomness like coin tosses, but a mixture of the two.

This course

Unfortunately, in this course we will not be able to address most of these exciting developments. We will be mainly dealing with setting up the basic mathematical framework, so that you have the basis for statistics, for applications in other fields and future courses in probability. We will also just try to get a glimpse of the probabilistic mathematical thinking, and there will be some intrinsically beautiful mathematical results.

The course will be roughly in three chapters:

(1) The basic framework of probability theory - here, we will properly set up the modern framework of probability theory, in other words see how one constructs a probabilistic model.

(2) Random variables - random variables are the central objects of probability theory, they are the random numbers, or other random objects that come up in our probabilistic model. We will see how to describe and study random variables, and meet several random variable that come up more frequently.
(3) Limit theorems - a special case of the Law of Large Numbers says that if you keep on tossing a fair coin, then the proportion of tails will get closer and closer to a half. We will be prove this result, but we will also prove a version of the Central Limit Theorem, discussed above.

We start, however, with an overview of some more elementary models for probability theory and discuss their limitations.

0.1 Some historical probability models and their limitations

In this section we shortly discuss some preliminary probability models.

Laplace model

For a few hundred years the following simple model (which we call Laplace or classical model) was used to study unpredictable situations, and to model the likelihood that a certain event happens in this situation.

- Gather together all possible outcomes $\Omega = \{\omega_1, \ldots, \omega_n\}$ and count the total number of possible outcomes $n_A := |\Omega|$ of the situation.
- Collect all the outcomes ω_i for which the desired event E happens, and count their number n_E.
- Set the probability of the event $p(E)$ to be the ratio $\frac{n_E}{n_A}$.

In other words, we can set up the following definition:

Definition 0.1 (Laplace/Classical model of probability)**.** *Laplace model of probability consists of a set of outcomes Ω and possible events, given by all subsets $E \subseteq \Omega$. The probability of each event is defined as $p(E) = \frac{|E|}{|\Omega|}$.*

In some sense, here we are really not defining any new mathematical structures - we are just giving a name to certain proportions.

For example if you want to model the event that two heads come up in two consecutive coin tosses you would do it as follows:

- We take $\Omega = \{HH, TT, HT, TH\}$,
- set $E = \{HH\}$
- and see that $p(E) = 1/4$ as $|\Omega| = 4$.

Many everyday or gambling situations can be described with this simple model.

Exercise 0.1. *Write down the Laplace model for calculating the probability of having two sixes in three throws of dice. What is this probability?*

This classical model has already some very nice properties, which we certainly want to keep for more general models.

Lemma 0.2 (Nice properties of the classical model)**.** *Consider the Laplace model on a set Ω. Let E, F be two events, i.e. two subsets of Ω.*

- *If the two events E, F cannot happen at the same time, i.e.then the probability of one of them happening $p(E \cup F) = p(E) + p(F)$.*
- *The complementary event of E, i.e. the event that E does not happen, has probability $1 - p(E)$.*

Both of these results follow directly from a definition. There are many other properties one could prove, e.g:

Exercise 0.2. *Consider the Laplace model on the set Ω and let E, F be any two events. Prove that $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.*

Using this, one can already also do basically all the calculations for lottery, betting, cards...as you see on the example sheet. But there is still one basic question - how come this ratio is of any use in telling you anything about the world, when actually you know that it doesn't predict what is happening?

The reason comes basically from the fact that if many of the events happen without influencing each other, then their proportion among all possible outcomes will converge to this notion of probability. Let us prove a weak version of this here:

Proposition 0.3 (Proportion of heads goes to 1/2)**.** *Consider the Laplace model for n coin consecutive fair coin tosses. Let $0 < \epsilon < 1/2$ be arbitrary and define the event E^n_ϵ to denote all sequences of n tosses where the proportion of heads is less than $1/2 - \epsilon$ or more than $1/2 + \epsilon$. Then for any $\epsilon > 0$, we have that $p(E^n_\epsilon) \to 0$ as $n \to \infty$.*

Behind this proposition is an implicit assumption: in the above Laplace model for n coin tosses exactly describes the situation where the n tosses do not influence each other, for all of them heads and tails are equally likely. In this situation any sequence of n fair coin tosses has probability exactly $1/2$.

To obtain the estimates needed in the proposition, we need an asymptotic of $n!$, i.e. a better expression about how it behaves as $n \to \infty$. This is called Stirling's formula. PS! The estimate on Binomial coefficient is not that easy!

Exercise 0.3 (Weak Stirling's formula (*))**.** *Prove that for some constants $c, C > 0$, we have that*

$$cn^n e^{-n} \leq n! \leq Cn^{n+1} e^{-n}.$$

(*) Deduce that there are $C, c > 0$, such that for all $\epsilon > 0$ small enough and all $n \in \mathbb{N}$ we have that

$$\binom{n}{\lceil n(1/2 - \epsilon) \rceil} \leq Cn^C 2^n \exp(-c\epsilon^2 n).$$

Armed with this, we are ready to prove the proposition.

Proof of proposition. Let $E^n_{\epsilon,<}$ and $E^n_{\epsilon,>}$ denote respectively the events that the proportion is less than $1/2 - \epsilon$, and that it is more than $1/2 + \epsilon$. As these events cannot happen at the same time, we have that $p(E) = p(E^n_{\epsilon,<}) + p(E^n_{\epsilon,>})$ and by symmetry it suffices to only show that $p(E^n_{\epsilon,<}) \to 0$ as $n \to \infty$. Moreover, as these events are increasing with ϵ, it suffices to prove the proposition for $\epsilon > 0$ small enough.

Now, the number of all possible sequences of n tosses is exactly 2^n as each toss has two options. On the other hand, the number of outcomes with k heads out of n tosses is given

by exactly $\binom{n}{k}$. So using Lemma 0.2 several times for disjoint events of exactly k tosses, we can write

$$p(E^n_{\epsilon,<}) \leq 2^{-n} \left(\sum_{k=0}^{\lceil n(1/2-\epsilon) \rceil} \binom{n}{k} \right).$$

A direct calculation convinces you that as long as $k < n/2$, we have that $\binom{n}{k-1} \leq \binom{n}{k}$. Thus we can further bound

$$p(E^n_{\epsilon,<}) \leq 2^{-n} n \binom{n}{\lceil n(1/2-\epsilon) \rceil}.$$

By Exercise 0.3, for all $\epsilon > 0$ small enough

$$\frac{\binom{n}{\lceil n(1/2-\epsilon) \rceil}}{2^n} \leq C' n^{C+1} \exp(-cn\epsilon^2)$$

and thus $p(E^n_{\epsilon,<}) \leq C' n \exp(-cn\epsilon^2)$, which goes to 0 as $n \to \infty$. □

Remark 0.4. *With the some strategy one could actually prove a somewhat stronger statement: for example that the probability of the event $\tilde{E}_n$ that the proportion of heads is outside of the interval $(1/2 - n^{-1/3}, 1/2 + n^{-1/3})$ goes to zero. This basically amounts to just setting $\epsilon = n^{-1/3}$ in the proof above.*

This is a special case of the Law of Large Numbers (LLN). We will prove LLN in much greater generality and with much less calculations, but only once we have developed some theory and only in the third section.

So we see that not only does Laplace model allow calculations, but it does tell you something about random phenomena - at least about reoccuring random phenomena. However, this model also has some drawbacks:

- In the Laplace model it is implicitly assumed that all outcomes of the situation are equally likely. What if this is not the case? For example, what if the coin is not fair, but after long number of tosses seems to give $1/\pi$ heads?
- Also, it is hard to work with more complicated situations, where you may have to look at an arbitrary large number of events like in the following exercise.

Exercise 0.4. *Suppose your event is: I will need no more than* 100 *tosses before getting three consecutive heads. Can you use the Laplace model? Can you use the Laplace model if your event is - I obtain three consecutive heads before three consecutive tails? But if you ask three consecutive heads before five consecutive tails? Can you use Laplace model for this?*

This is related to a more general worry: as soon as there are infinitely many possible outcomes, what should you do? Assuming that all of infinitely many outcomes are equally likely gives a contradiction, as their probabilities would still need to add up to one! What to do?

An intermediate model

The next probability model does not presuppose that all outcomes are equally likely and will allow also to handle an infinite number of outcomes:

Definition 0.5 (An intermediate probability model)**.** *We say that (Ω, p) is an intermediate probability model if Ω is a set (of outcomes) and $p : \Omega \to [0, 1]$ is a function such that*

- *The total probability is 1: $\sum_{\omega\in\Omega} p(\omega) = 1$* [2]*.*
- *The probabilities of disjoint subsets of Ω add up: $p(E \cup F) = p(E) + p(F)$ for all $E \cap F = \emptyset$.*

We identify an event E with a subset of Ω and set the probability $p(E) := \sum_{\omega\in E} p(\omega)$.

This intermediate model is set up so that we still keep the nice properties of the classical model that we saw above. Moreover, one can check that when $|\Omega| < \infty$ and we set all $p(\omega) = |\Omega|^{-1}$, we are back to the Laplace model. So it is really a generalization.

Before thinking about further mathematical properties of this model, let us think about using it for applications. One difficulty of applying this model to real situations is now the following question – how do we decide what should be the $p(\omega)$? Before we used a certain symmetry or exchangeability hypothesis on the set of outcomes, but if we don't have this, what could we do?

For example, here is a reasonable-sounding idea, based on the proportion above: in the case of the coin toss, i.e. two possibilities, we could just toss the coin it many times and set the proportion of heads to be the probability of heads in our model. That sounds meaningful. However, how many times should we toss it? If we toss it just once, we set the probability to be either 0 or 1? We will be able to give some sort of an idea of how many tosses would suffice in the last chapter of the course...but what should you do if you don't have a lot of data? Or if the model is much more complicated? Luckily for us, these complicated questions belong already more to the discipline of statistics...

So let us rather ask what is still mathematically missing in the intermediate model? Having a countable set is now not a problem. In fact, we will see that as long as Ω is a countable set, the intermediate model is equivalent to the modern framework of probability, introduced in the next section.

However, uncountable sample spaces enter naturally. For example, when you model for example an uniform random point on $[0, 1]$ then the space of outcomes is uncountable. Or similarly the space of infinite sequences of coin tosses is uncountable (why?) - such a space is needed when you consider for example the event that three consecutive heads occur before five consecutive tails, as it is not determined by any fixed number of coin tosses. Finally, many complicated discrete situations are easier to describe and study if one models them via continuous probabilities, like the Gaussian distribution where all values of $\mathbb{R}$ are possible. In fact, Gaussians enter through every door as we will see towards the end of the course.

And as soon as we have an uncountable Ω, say $\Omega = \mathbb{R}$ or $\Omega = [0, 1]$, things get more involved. Indeed, if you think about it, already sums over uncountable sets are pretty complicated (and not so well defined)! For example, there is just no function p satisfying the hypothesis of the definition and putting a positive mass on uncountable set of points of Ω:

Exercise 0.5. *Let Ω be any uncountable set. Consider a positive function $f : \Omega \to [0, 1]$. Then necessarily $\sum_{\omega\in\Omega} f(\omega) = \infty$.*

[2]Here, and elsewhere you might wonder what does this sum even mean if Ω is infinite. You can rigorously define it as the supremum of $\sum_{\omega\in\Omega'} f(\omega')$ over all finite subsets $\Omega' \subseteq \Omega$, if you wish, but in this Section nr 0 we don't yet worry about these things so much...

So how should we then model the uniform number on $[0,1]$? It intuitively feels that this notion exists, but we already discussed that putting equal probabilities on infinite sets doesn't work...Is there any way out?

Probability vs area: intermediate continuous model

There is one nice way out from the issues described above. Namely, the following hack was used up to 20th century: if we think of a raindrop falling on the segment $[0,1]$, then the probability that it falls into some set A should be exactly the area of this set! Thus to define continuous probability, at least on $[0,1]^n$ we could equate probability of a set with its area.

Now, this is very nice because we know that area is related to integrals - areas can be calculated! Thus we get an idea for defining a variety of probability distributions on $\mathbb{R}^n$ - for any Riemann-integrable function f with $\int_{\mathbb{R}^n} f(x)d^n x = 1$ we define the probability of being in A as $\int_A f(x)d^n x$, in case such a thing is defined. So in conclusion, we could also define an intermediate continuous probability model

Definition 0.6 (An intermediate continuous probability model)**.** *We say that $(\mathbb{R}^n, f)$ is an intermediate probability model if f is a non-negative Riemann-integrable function with total mass 1. We identify events with subsets A such that $\int_A f(x)d^n x$ is defined, and set their probability to be $p(A) := \int_A f(x)d^n x$.*

Such a model shares several nice properties both with the Laplace model or the intermediate model. So why do we call this again just an intermediate model, why is it not a satisfactory resolution? For all practical purposes, it is in fact already pretty good!

However, from a purely mathematical point of view there are some drawbacks:

- Firstly, it's just quite unsatisfactory to have two different notions of probability - one for discrete, one for the continuous setting! It would be much nicer to have one framework pretty much like topology offers a framework to talk about continuity both for real numbers and for continuous functions.
- Second, we would certainly also like to talk of random objects that are more complicated than $\mathbb{R}^n$ - for example random continuous functions that could describe say the shore line of Britan or mountainous landscapes or clouds. But what is the notion of area for such complicated spaces?

As we will see, both of those issues are resolved in the modern framework of probability theory.

Section 1

Basic notions

In this section we will build up the modern framework of probability, and see how it nicely unifies the attempts from the previous section. It is a bit abstract, but setting up this language allows us to move more swiftly and rigorously later on.

1.1 Basics of measure spaces and probability spaces

As in topology, a probability space will be a set together with a certain structure. We will start with a more general notion of a measure space.

1.1.1 Definition of a measure space

For a measure space the structure comes in two bits:

- first, a set of subsets closed under some operations, called this time a σ-algebra;
- and second, a function defined on these subsets, called a measure.

You can think of measure as of some generalization of area, and of the σ-algebra as of all subsets whose area can be measured.

Definition 1.1 (Measure space, Borel 1898, Lebesgue 1901-1903)**.** *A measure space is a triple $(\Omega, \mathcal{F}, \mu)$, where*

- *Ω is a set, called the sample space or the universe.*
- *$\mathcal{F}$ is a set of subsets of Ω, satisfying:*
  - *$\emptyset \in \mathcal{F}$;*
  - *if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$;*
  - *If $A_1, A_2, \dots \in \mathcal{F}$, then also $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ is called a σ-algebra and any $A \in \mathcal{F}$ is called a measurable set.
- *And finally, we have a function $\mu : \mathcal{F} \to [0, \infty]$ satisfying $\mu(\emptyset) = 0$ and countable additivity for disjoint sets: if $A_1, A_2, \dots \in \mathcal{F}$ are pairwise disjoint,*

$$\mu(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mu(A_n).$$

This function μ is called a measure. If $\mu(\Omega) < \infty$, we call μ a finite measure.

Let us consider an example of defining a measure on an arbitrary set Ω:

Definition 1.2 (Counting measure)**.** *On any set Ω one can define the counting measure μ_c: we set $\mathcal{F} := \mathcal{P}(\Omega)$, and $\mu_c(\{\omega\}) := 1$ for any $\omega \in \Omega$. Notice that if Ω is an infinite set, then $\mu_c(\Omega) = \infty$, so this is a measure, but not a finite measure.*

Here, we still used the power set $\mathcal{P}(\Omega)$ as the sigma-algebra, however the ability to restrict the measure only on a subcollection $\mathcal{F}$ is actually necessary. A way to think about it as follows: we think of measure as of a generalization of area, so $\mathcal{F}$ is the set of all subsets for which the notion of area exits. One way to think that there should be some restriction is by thinking of the following example: should you be able to calculate the area under any arbitrary (non-continuous!) function $f : [0, 1] \to \mathbb{R}$? Recall that for example in Riemann integration, that would be one way to give sense to an area, the function 1_E is not integrable

for every $E \subseteq [0,1]$, e.g. for $E = \mathbb{Q} \cap [0,1]$!

Also (similarly to the case of topology), it might not be intuitively clear why we should ask the σ-algebra to be closed exactly under countable unions and intersections of sets, or why we ask the measure to be countable additive. Why not finite, why not arbitrary? We will see some answers, but the main answer - as in the case of topology - is that this makes the framework function the best.

Before defining the probability space, let us see how to construct more measurable sets. Already the defining properties of the sigma-algebra $\mathcal{F}$ give us plenty of measurable sets. However, there are many more:

Lemma 1.3 (Constructing more measurable sets)**.** *Consider a set Ω with a σ-algebra $\mathcal{F}$.*

(1) Then also $\Omega \in \mathcal{F}$ and if $A, B \in \mathcal{F}$, then also $A \setminus B \in \mathcal{F}$.
(2) For any $n \geq 1$, if $A_1, \ldots, A_n \in \mathcal{F}$, then also $A_1 \cup \cdots \cup A_n \in \mathcal{F}$ and $A_1 \cap \cdots \cap A_n \in \mathcal{F}$.
(3) If $A_1, A_2, \ldots, \in \mathcal{F}$, then also $\bigcap_{n \geq 1} A_1 \in \mathcal{F}$.

Proof of Lemma 1.3. By de Morgan's laws for any sets $(A_i)_{i \in I}$, we have that

$$\bigcap_{i \in I} A_i = (\bigcup_{i \in I} A_i^c)^c.$$

Property (3) follows from this, as if $A_1, A_2, \cdots \in \mathcal{F}$, then by the definition of a σ-algebra also $A_1^c, A_2^c, \cdots \in \mathcal{F}$ and hence

$$(\bigcup_{i \geq 1} A_i^c)^c \in \mathcal{F}.$$

For (2), again by de Morgan laws, it suffices to show that $A_1 \cup \cdots \cup A_n \in \mathcal{F}$. But this follows from the definition of a σ-algebra, as $A_1 \cup \cdots \cup A_n = \bigcup_{i \geq 1} A_i$ with $A_k = \emptyset$ for $k \geq n+1$. Finally, for (1) we can just write $\Omega = \emptyset^c$. Moreover, writing $A \setminus B = A \cap B^c$, we conclude by using (2). $\square$

1.1.2 Definition of a probability space

We can now define a probability space - it is just a measure space of total measure 1. Although nowadays it is natural to see the concepts of a measure space and probability space side by side, realizing that measure theory is the right context for all probability theory took nearly 30 years! It was only the Russian mathematician Kolmogorov who realized that it encapsulates all the previous models and notions of probability in a satisfactory manner.

Definition 1.4 (Probability space, Kolmogorov 1933)**.** *A probability space is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with total mass 1, i.e. with $\mathbb{P}(\Omega) = 1$. In the case of a probability space we still call Ω the universe or the state space, the $\mathbb{P}$ the probability measure, the sets $E \in \mathcal{F}$ events and $\mathbb{P}(E)$ the probability of the event E.*

It is important to have a good mental picture of how these objects correspond to our description of the world. We think of $(\Omega, \mathcal{F}, \mathbb{P})$ as follows:

- Ω is the collection of all possible states of the situation, of all possible outcomes, very much like in the simple Laplace model.

- The new bit is the σ-algebra $\mathcal{F}$. It contains all events E whose happening we can observe. Notice that $\mathcal{F}$ is not necessarily equal to the space of all subsets of Ω. This means that pretty much like in case of area, we might not be able to observe or assign probabilities to all combinations of outcomes.
- Finally, the function $\mathbb{P} : \mathcal{F} \to [0, 1]$ assigns the probability of each event - this can be interpreted either as the frequency of the event over many independent trials as we saw in Section 0, or as a certain belief (we will come back to this later.) This is something we put into the model based on our assumptions.

This new framework is more general than the intermediate model (and thus Laplace model). Indeed, if Ω is countable, we just set $\mathcal{F} := \mathcal{P}(\Omega)$. Now if our intermediate model has a probability function $p : \Omega \to [0, 1]$ such that $\sum_{\omega\in\Omega} p(\omega) = 1$, we can just define $P(E) := \sum_{\omega\in E} p(\omega)$ and verify that all axioms of the probability space are indeed satisfied. For a concrete example, in the fair dice model $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} := \mathcal{P}(\Omega)$ and for any event E, we set $\mathbb{P}(E) := \frac{|E|}{6}$.

Now, the really new bit w.r.t. the previous models is the second bullet point - the notion of sigma-algebra. It becomes mathematically crucial in considering probability spaces where Ω is uncountable.

Conceptually, it is however nice already for discrete probability spaces as it helps to distinguish the level of information that one can observe.

For example, suppose we model the situation with two fair coins. To do this, we set $\Omega = \{(H, T), (H, H), (T, H), (T, T)\}$. Now, let us look at the role of different sigma-algebras:

- If we can observe the outcome of both tosses, then our sigma-algebra would be $\mathcal{P}(\Omega)$.
- However, suppose the only thing you can observe is the outcome of the first toss. Then we cannot differentiate whether the full outcome was (H, T) or (H, H), or similarly whether it was (T, H) or (T, T). We have thus no information about the second toss, and maybe also no way to assign to it some probabilities. To take this into account, we can without changing the sample space, change the sigma-algebra and set it to be $\mathcal{F} = \{\emptyset, \{(H, T), (H, H)\}, \{(T, H), (T, T)\}, \Omega\}$, where naturally the first of the sets corresponds to the first toss coming up heads, and the second to the first toss coming up tails.
- Similarly, maybe our friend only tells you whether the two tosses were the same or different. Then we cannot differentiate between (H, H) and (T, T), or between (H, T) or (T, H). We could model this situation by setting

$$\mathcal{F} = \{\emptyset, \{(H, H), (T, T)\}, \{(T, H), (H, T)\}, \Omega\}.$$

Often in fact such a situation happens in real life: we only obtain information about the world step by step, and thus if we want to keep on working on the same probability space, we can consider different filtrations $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \dots$ such that each next one contains more information.

Finally, let us end this subsection with a little remark saying that we are not generalizing all models we presented before. Indeed, at the end of Section 0 we discussed that a version of continuous probability could be defined using the Riemann integral. However, let us see that such a construction doesn't easily give rise to a probability space with the definition

above: consider $\Omega = [0,1]$ and let $\mathcal{F}$ be the subset of all sets A such that $\mathbf{1}_{\{x\in A\}}$ is Riemann-integrable. Then surprisingly $\mathcal{F}$ is not a sigma-algebra, as shown by the following exercise.

Exercise 1.1 (Riemann integral doesn't mix with measure)**.** *Show that for any finite set $A \subseteq [0,1]$ the function $\mathbf{1}_{\{x\in A\}}$ is Riemann-integrable. On the other hand show that $\mathbf{1}_{\{x\in\mathbb{Q}\}}$ is not Riemann-integrable (i.e. the lower and upper sums don't converge to the same number). Deduce that the set $\mathcal{F}$ of all subsets such that $\mathbf{1}_{\{x\in A\}}$ is Riemann-integrable is not a σ-algebra.*

So we will have to come up with something better for $[0,1]$! Moreover, we would certainly also want to give a sense to probability measures on arbitrary metric spaces. However, before we get into this, let us consider some basic properties of measure spaces and probability spaces.

1.1.3 Some general properties of measure spaces

Next, let us look at some basic properties of a measure defined on σ-algebras, not dissimilar to some laws we have proved on the Laplace model of probability.

Proposition 1.5 (Basic properties of a measure and a probability measure)**.** *Consider a measure space $(\Omega, \mathcal{F}, \mu)$. Let $A_1, A_2, \cdots \in \mathcal{F}$. Then*

(1) For any $n \geq 1$, and $A_1, \ldots, A_n$ disjoint, we have finite additivity

$$\mu(A_1) + \cdots + \mu(A_n) = \mu(A_1 \cup \cdots \cup A_n).$$

In particular if $A_1 \subseteq A_2$ then $\mu(A_1) \leq \mu(A_2)$.

(2) If for all $n \geq 1$, we have $A_n \subseteq A_{n+1}$, then as $n \to \infty$, it holds that $\mu(A_n) \to \mu(\bigcup_{k\geq 1} A_k)$.

(3) We have countable subadditivity (also called the union bound): $\mu(\bigcup_{n\geq 1} A_n) \leq \sum_{n\geq 1} \mu(A_n)$.

If in fact $\mu(\Omega) = 1$, and thus we have a probability space (and we set $\mathbb{P} := \mu$), we also have the following properties:

(4) For any $A \in \mathcal{F}$, we have that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(5) If for all $n \geq 1$, we have $A_n \supseteq A_{n+1}$, then as $n \to \infty$, it holds that $\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_{k\geq 1} A_k)$.

Notice that for two events A, B properties 1 and 4 correspond to properties we already saw for the Laplace model of probability.

Proof of Proposition 1.5. This is on exercise sheet. □

Exercise 1.2 (Counterexample for general measure spaces)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Find measurable sets $(A_n)_{n\geq 1} \in \mathcal{F}$ such that for $n \geq 1$ we have that $A_n \supseteq A_{n+1}$. Show that contrary to probability spaces, it does not necessarily hold that $\mu(A_n) \to \mu(\bigcap_{n\geq 1} A_n)$.*

In fact, another nice property of the Laplace model holds in the more general framework:

Lemma 1.6 (Inclusion and Exclusion)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, \ldots, A_n \in \mathcal{F}$. Then*

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{S\subset\{1,\ldots,n\},\, S\neq\emptyset} (-1)^{|S|+1} \mathbb{P}\Big(\bigcap_{i\in S} A_i\Big).$$

In particular, we have that $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$.

Proof. This proof is on the exercise sheet. □

Before concentrating more concretely on probability spaces, let us build up a little bit more vocabulary and concepts for working with measure spaces.

1.1.4 Measurable and measure preserving maps

In topological spaces continuous functions mix well with topology. In measure spaces functions that mix well with σ-algebra are called measurable maps.

Definition 1.7 (Measurable and measure-preserving maps)**.** *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be two measure spaces.*

- *We call a function $f : \Omega_1 \to \Omega_2$ measurable if the preimages of measurable sets are measurable, i.e. if $\forall F \in \mathcal{F}_2 \implies f^{-1}(F) \in \mathcal{F}_1$.*
- *Further, a measurable function such that $\forall F \in \mathcal{F}_2$ we have that $\mu_2(F) = \mu_1(f^{-1}(F))$ is called measure-preserving.*

Observe that the measure itself does not enter in the definition of a measurable map; the name measurable comes from the fact that the pair $(\Omega, \mathcal{F})$, where Ω is a set and $\mathcal{F}$ is a σ-algebra is often called a measurable space.

Intuitively, measurable maps preserve the entity of sets whose area can be measured (i.e. all events in prob. spaces), and measure-preserving maps preserve in addition the area as well (i.e. the probability in prob. spaces).

Similarly to topological spaces we will from now onwards always denote a measurable function as $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ to keep track of the σ-algebras involved. However the function f is still defined from Ω_1 to Ω_2.

As in topological spaces, measurability can be checked on a smaller subset of sets. This is an important fact that helps you verify measurability:

Lemma 1.8. *Suppose $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are two measurable spaces and $\mathcal{G}$ generates $\mathcal{F}_2$, in the sense that the smallest σ-algebra containing $\mathcal{G}$ is equal to $\mathcal{F}_2$. Prove that if $f^{-1}(G) \in \mathcal{F}_1$ for all $G \in \mathcal{G}$, then f is in fact a measurable function from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$.*

Proof. The proof is on the exercise sheet. □

As for sets and topological spaces, there is also a notion of equivalence for measure spaces - we want bijective measure-preserving bimeasurable maps:

Definition 1.9 (Isomorphic as measure spaces)**.** *We say that two measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ are isomorphic (as measure spaces) if there exists a measurable bijection $f : \Omega_1 \to \Omega_2$ such that f^{-1} is also measurable, and such that for all $E \in \mathcal{F}_1$, it holds that $\mu_1(E) = \mu_2(f(E))$.*

This equivalence will not come up too often. However, the situation that will come up more often is the fact that we want to define a probability measure on the image space of a function. For example, when we consider the model of two consequent fair dices, we might want to define a function that gives the sum of the two throws. This will naturally induce a probability measure on the space $\{2, 3, \ldots, 12\}$.

This is formalized by the idea of a push-forward measure:

Lemma 1.10 (Push-forward measure)**.** *Consider a measurable map f from $(\Omega_1, \mathcal{F}_1, \mu_1)$ to $(\Omega_2, \mathcal{F}_2)$. Then f induces a measure μ_2 on $(\Omega_2, \mathcal{F}_2)$ by $\mu_2(F) := \mu_1(f^{-1}(F))$. Moreover, then the map f from $(\Omega_1, \mathcal{F}_1, \mu_1)$ to $(\Omega_2, \mathcal{F}_2, \mu_2)$ is measure-preserving.*

Often this measure μ_2 is called the push-forward measure of μ_1. Notice when μ_1 is a probability measure, then so is μ_2 as then $\mu_2(\Omega_2) = \mu_1(\Omega_1) = 1$.

Proof. We need to just check that μ_2 is a measure. It clearly satisfies $\mu_2(\emptyset) = 0$. Further, notice that if $F_1, F_2, \ldots$ are disjoint, then so are $f^{-1}(F_1), f^{-1}(F_2), \ldots$. Thus countable additivity for μ_2 also follows from that of μ_1. □

In fact, this will be a very important tool to induce probability measures. For example, we will see that all natural probability measures on $\mathbb{R}$ can be constructed via suitable functions from probability measures on $[0, 1]$, or we often only care about functions from our probability space that take values in $\mathbb{R}^n$ and concentrate then on these induced probability measures.

For now, let us consider a very caricatural example to illustrate what is happening.

Example 1.11. *Consider the probability space of a fair dice:*

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\{1, 2, 3, 4, 5, 6\}, \mathcal{P}(\{1, 2, 3, 4, 5, 6\}), \mathbb{P})$$

where $\mathbb{P}(\{i\}) = 1/6$. *Consider also the measurable space corresponding to a coin toss*

$$(\Omega_2, \mathcal{F}_2) = (\{H, T\}, \mathcal{P}(\{H, T\})).$$

Now, define $f : \{1, 2, 3, 4, 5, 6\} \to \{H, T\}$ *that maps* $1, 2, 3$ *to* H *and* $4, 5, 6$ *to* T. *Then* f *is a measurable map from* $(\Omega, \mathcal{F})$ *to* $(\Omega, \mathcal{F}_2)$ *that intuitively gives us a way to encode a fair coin toss using a dice:* $\{1, 2, 3\}$ *corresponds to heads, and* $\{4, 5, 6\}$ *to tails.*

The lemma above tells us that via this map f *we can indeed induce a probability measure on this coin model, i.e. on* $(\{H, T\}, \mathcal{P}(\{H, T\}))$ *that exactly gives both options half a probability, i.e. encodes the fair coin.*

1.2 Probability spaces

Although most of what follows would work in the realm of measure spaces, let us now concentrate on probability spaces. We have listed several properties that a probability space should satisfy, let us see that one can even construct such probability spaces for situations of interest.

In particular, we will now look at different types of probability spaces in some more detail and see that whereas for discrete probability spaces (i.e. countable state space Ω) there is basically nothing new w.r.t. the simpler model we called the intermediate model, then for continuous probability spaces (i.e. uncountable state space Ω) some work is needed to even construct probability spaces with nice properties.

1.2.1 Discrete probability spaces

Probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ with a countable sample space Ω are called discrete probability spaces. As already mentioned, if $|\Omega| < \infty$ and we set $\mathbb{P}(\{\omega\}) = |\Omega|^{-1}$, then we are back at the Laplace model - i.e. to the model of a coin or a fair dice.

It is also easy to see that we are back to the intermediate model in case when σ-algebra contains all subsets:

Lemma 1.12. *Let* Ω *be a countable set. Then the set of probability measures on* $(\Omega, \mathcal{P}(\Omega))$ *is in one to one correspondence with the set of functions* $p : \Omega \to [0, 1]$ *with* $\sum_{\omega \in \Omega} p(\omega) = 1$.

The proof is a rather boring affair:

Proof. First, given any probability measure $\mathbb{P}$ on $(\Omega, \mathcal{P}(\Omega))$, consider the function $p_{\mathbb{P}} : \Omega \to \mathbb{R}$ given by just $p_{\mathbb{P}}(\omega) = \mathbb{P}(\{\omega\})$. As $\mathbb{P}$ is a probability measure, in fact $p_{\mathbb{P}}$ takes values in $[0, 1]$. Further, by countable disjoint additivity

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = \mathbb{P}(\Omega) = 1.$$

In the other direction, given such a function p, define $\mathbb{P}_p : \mathcal{P}(\Omega) \to [0, 1]$ for every $E \subseteq \Omega$ by

$$\mathbb{P}_p(E) = \sum_{\omega \in E} p(\omega).$$

We know that this sum is well defined as p is non-negative and this sum is bounded from above by 1. It is then immediate to check that $\mathbb{P}_p$ satisfies all conditions for being a probability measure: from definition it is countable additive, and also $\mathbb{P}(\Omega) = 1$.

Finally, as the two maps $\mathbb{P} \to p_{\mathbb{P}}$ and $p \to \mathbb{P}_p$ are inverses of each other, we obtain the necessary bijection. □

But this of course doesn't meen automatically that by introducing the notion of $\sigma-$algebra, we didn't still somehow enlarge the possible probability spaces with a countable state space. In other words, is there maybe an extra level of generality induced by this σ-algebras in w.r.t the intermediate models? The next proposition says that this is not the case.

Proposition 1.13 (Discrete probability spaces = intermediate spaces)**.** *Let Ω be a countable set and consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. One can construct a probability space $(\Omega_2, \mathcal{P}(\Omega_2), \mathbb{P}_2)$ such that Ω_2 is countable and there is a surjective function $f : \Omega \to \Omega_2$ that is measurable, measure-preserving and more-over such that for every $F \in \mathcal{F}$ also $f(F)$ is measurable.*

Indeed, this proposition says that we can encode all the information that can be measured, observed and assigned probabilities to via a probability space $(\Omega_2, \mathcal{P}(\Omega_2), \mathbb{P}_2)$ and we saw just above that each such model is in 1-1 correspondence with what we called an intermediate model.

[$\star$ This proof is non-examinable $\star$]

Proof of Proposition 1.13. The idea is to partition Ω into indecomposable sets $F \in \mathcal{F}$, i.e. to write $\Omega = \bigcup_{i \in I} F_i$ such that F_i are disjoint and for any $F \in \mathcal{F}$ and any F_i, either $F \cap F_i = \emptyset$ or $F_i \subseteq F$. These F_i will correspond to elements or 'atoms' of Ω_2.

To do this, define for each $\omega \in \Omega$ the set $F_\omega = \bigcap_{F \in \mathcal{F}, \omega \in F} F$. We claim that $F_\omega \in \mathcal{F}$. This is not obvious as the intersection might be uncountable. Now, for any $\widehat{\omega} \notin F_\omega$, pick some $G_{\widehat{\omega}} \in \mathcal{F}$ with $\omega \in G_{\widehat{\omega}}$ but $\widehat{\omega} \notin G_{\widehat{\omega}}$. Notice that such a set must exist, as otherwise $\widehat{\omega} \in F_\omega$. Moreover, notice that $\widehat{\Omega} := \{\widehat{\omega} \notin F_\omega\}$ is countable. Thus $\widehat{F}_\omega := \bigcap_{\widehat{\omega} \in \widehat{\Omega}} G_{\widehat{\omega}} \in \mathcal{F}$. We claim that in fact $\widehat{F}_\omega = F_\omega$. As $\omega \in \widehat{F}_\omega$, by definition $F_\omega \subseteq \widehat{F}_\omega$. On the other hand also by definition $F_\omega^c \subseteq \widehat{F}_\omega^c$ and thus $F_\omega = \widehat{F}_\omega \in \mathcal{F}$.

We now claim that the sets F_ω partition Ω as explained above: first let $\omega, \widehat{\omega} \in \Omega$. We claim that either $F_{\widehat{\omega}} = F_\omega$ or they are disjoint. Suppose they are not disjoint. Then both $F_\omega \cap F_{\widehat{\omega}} \in \mathcal{F}$ and $F_\omega \backslash F_{\widehat{\omega}} \in \mathcal{F}$. But if $F_\omega \neq F_{\widehat{\omega}}$ then one of these sets contains ω and is strictly smaller than F_ω, contradicting the definition of F_ω. Now, consider any other $F \in \mathcal{F}$. Then either $F_\omega \cap F = \emptyset$, or there is some $\widehat{\omega} \in F_\omega$. The by definition $F_{\widehat{\omega}} \subseteq F$. But also as $F_{\widehat{\omega}} \cap F_\omega \neq \emptyset$ we have that $F_{\widehat{\omega}} = F_\omega$ and thus $F_\omega \subseteq F$.

Now, as Ω is countable, there are countably many sets F_ω. Thus we can enumerate them using a countable index set I as $(F_i)_{i\in I}$. We now define $f : \Omega \to I$ by $f(\omega) = i_\omega$, where $i_\omega \in I$ corresponds to the index of i such that $\omega \in F_i$. It is now easy to verify that f is measurable from $(\Omega, \mathcal{F})$ to $(I, \mathcal{P}(I))$. Thus we can induce a probability measure $\mathbb{P}_I$ on $(I, \mathcal{P}(I))$ as a push-forward of $\mathbb{P}$, i.e. via Lemma +, and obtain that f is in fact measure-preserving as a map from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(I, \mathcal{P}(I), \mathbb{P}_I)$. It remains to argue that every measurable set $F \in F$ map to a measurable set. But all subsets of I are measurable and thus this follows trivially. □

[⋆ End of the non-examinable part ⋆]

As, discussed the parameters of a discrete probability model (i.e. $p(\omega)$ for $\omega \in \Omega$) come via statistics from the real world, or by assumptions of equal probabilities like in the case of the Laplace model for finite Ω. Thus in this respect, finite and countably infinite spaces behave very similarly. One should, however, notice one difference - there are no probability measures on countably infinite sets that treat each element of the sample space as equally likely. Let us illustrate it in the case of $\Omega = \mathbb{Z}$, though a similar proof would work for any countably infinite Ω, when replacing shifts with general bijections.

Lemma 1.14. *There is no probability measure $\mathbb{P}$ on $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$ that is invariant under shifts, i.e. such that for any $A \in \mathcal{P}(\mathbb{Z}), n \in \mathbb{Z}$, we have that $\mathbb{P}(A + n) = \mathbb{P}(A)$* [3].

Proof. By shift-invariance $\mathbb{P}(\{k\}) = \mathbb{P}(\{0\})$ for any $k \in \mathbb{Z}$. By countable additivity

$$1 = \mathbb{P}(\mathbb{Z}) = \sum_{k\in\mathbb{Z}} \mathbb{P}(\{k\}) = \sum_{k\in\mathbb{Z}} \mathbb{P}(\{0\}),$$

which is either equal to 0 if $\mathbb{P}(\{0\}) = 0$, or equal to ∞ if $\mathbb{P}(\{0\}) > 0$, giving a contradiction. □

Notice that this in particular means that we cannot really conveniently talk about a random whole number, or about a random prime number - we would want all of them to have the same probability! Still, thinking of prime numbers as random numbers has been a very successful recent idea. For example, we refer to a beautiful theorem about arithmetic progressions in prime numbers, called the Green-Tao theorem.

Let us finish this small subsection on discrete spaces by discussing two example: simple random walk and uniform random graphs. These models and their generalizations will accompany us throughout the course and that have given rise to many beautiful theorems in combinatorics and probability theory.

Example 1.15 (Symmetric simple random walk)**.** *Let $n \in \mathbb{N}$ and let Ω be the set of all simple walks of n steps, i.e. $\mathbb{Z}$-valued vectors $(S_0, S_1, S_2, \ldots, S_n)$ such that $S_0 = 0$ and $|S_i - S_{i-1}| = 1$.*

Now set $\mathcal{F} = \mathcal{P}(\Omega)$ and define $\mathbb{P}$ such that $\mathbb{P}(\{\omega\}) = |\Omega|^{-1} = 2^{-n}$ for each $\omega \in \Omega$ (what does each ω here correspond to?). The corresponding probability model is called that of a symmetric simple random walk.

Example 1.16 (Uniform random graph)**.** *Let $n \in \mathbb{N}$. A simple graph is a set of vertices $V = \{v_1, \ldots, v_n\}$ together with an edge set E, that is some subset of $\{\{v_i, v_j\} : (v_i, v_j) \in$*

[3] Here, as customary, $A + n = \{a + n : a \in A\}$.

$V \times V, i \neq j\}$. You can imagine the graph as drawing all the n points $v_1, \dots, v_n$ on the plane and then drawing a line between v_i and v_j with iff $\{v_i, v_j\} \in E$.

The probability model for a uniform random graph is defined as follows: we let Ω be the set of all simple graphs, set $\mathcal{F} = \mathcal{P}(\Omega)$ and define $\mathbb{P}$ such that $\mathbb{P}(\{\omega\}) = |\Omega|^{-1}$ for each $\omega \in \Omega$ (what does each ω here correspond to?).

1.2.2 A cautionary tail for continuous probability spaces

Probability spaces where Ω is uncountable are called continuous probability spaces. The most typical examples are the space of sequences of coin tosses $\Omega = \{0,1\}^{\mathbb{N}}$, the unit interval $\Omega = [0,1]$ or $\Omega = \mathbb{R}$. It could also be $\Omega = \mathbb{R}^n$ or why not even $\Omega = \mathcal{C}_0([0,1])$, i.e. the set of continuous functions on $[0,1]$.

In the uncountable case, things get a bit more involved. Now, given any uncountable set Ω, one can still always define some probability measure on $(\Omega, \mathcal{P}(\Omega))$: for example we could just pick a single $\omega \in \Omega$ and set $\mathbb{P}(E) = 1$ if $\omega \in E$ and $\mathbb{P}(E) = 0$ otherwise (check this is a probability measure!). But in some sense this is not really looking at the whole set Ω - only one point is picked out. It comes out, however, that probability measures that somehow consider all points become much more difficult to define. For example, as the following example illustrates, shift invariance will no longer be defined on $\mathcal{P}(\Omega)$, but rather on smaller collections of subsets.

Indeed, it seems very reasonable that there should exist a uniform probability measure $\mathbb{P}$ on the circle $S^1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ that would be invariant under rotating the circle by any fixed angle. This seems like common sense! However, the following proposition says that this is impossible when we want to make all subsets of S^1 measurable, i.e. when we take $\mathcal{F} = \mathcal{P}(S^1)$:

Proposition 1.17. *There is no probability measure $\mathbb{P}$ on $(S^1, \mathcal{P}(S^1))$ that is invariant under shifts, i.e. such that for any $A \in \mathcal{P}(S^1), \alpha \in [0, 2\pi)$, we have that $\mathbb{P}(A + \alpha) = \mathbb{P}(A)$, where here we denote $A + \alpha$ the set obtained by rotating the circle by α radians.*

You should compare this to Lemma 1.14 and think why this is more interesting and more difficult.

[$\star$ This proof is non-examinable $\star$]

Proof. The idea is to decompose S^1 into a countable number of shifted copies of a set R and then to draw a contradiction like in Lemma 1.14.

Consider some irrational number $r \in [0,1]$ and the following operation $T : S^1 \to S^1$: we rotate the circle by $r2\pi$ radians. The inverse operation T^{-1} rotates it by $-r2\pi$ radians.

For any $x \in S^1$, consider set

$$S_x = \{\dots, T^{-2}(x), T^{-1}(x), x, T(x), T^2(x), \dots\}.$$

Notice that by the fact that r is irrational, we have that $T^k(x) \neq T^l(x)$ for all $k, l \in \mathbb{Z}$ and thus S_x is countably infinite: indeed, otherwise $T^{k-l}(x) = x$, but T^{k-l} is a rotation of $r(k-l)2\pi \notin 2\pi\mathbb{Z}$ radians and thus this is impossible.

We claim that the countably infinite sets S_x are either disjoint or coincide and that they partition S^1. First, notice that each $x \in S_x$, thus $\bigcup_{x \in S^1} S_x = S^1$. Hence it remains to show that if $S_x \cap S_y \neq \emptyset$, then $S_x = S_y$. So suppose that there is some $z \in S_x \cap S_y$.

Then by definition there is some $k_x, k_y \in \mathbb{Z}$ such that $T^{k_x}(x) = T^{k_y}(y) = z$. But then $x = T^{-k_x}(z) = T^{k_y - k_z}(y)$ and hence for any $l \in \mathbb{Z}$, $T^l(x) = T^{l+k_y-k_z}(y)$ and $S_x = S_y$.

By the Axiom of choice [4] we can pick one element s_x from each disjoint S_x and define R as the union of all such elements.

Now for $i \in \mathbb{Z}$, let $R_i = T^i(R)$. We claim that all R_i are disjoint. Indeed if $z \in R_i$ and $z \in R_j$, then there must exist $w, y \in R$ such that $T^i(w) = z = T^j(y)$ and in particular $T^{i-j}(w) = y$. Thus on the other hand w and y would need to belong to the same S_x, and on the other hand this is impossible as we saw that $T^k(x) \neq x$ for all $k \in \mathbb{Z}$. Moreover, $\bigcup_{i\in\mathbb{Z}} R_i = S^1$ as $\bigcup_{i\in\mathbb{Z}} R_i = \bigcup_{x\in S^1} S_x$.

Hence by countable additivity $1 = \mathbb{P}(S^1) = \sum_{i\in\mathbb{Z}} \mathbb{P}(R_i)$ and shift-invariance $\mathbb{P}(R_i) = \mathbb{P}(R)$ gives a contradiction as in the proof of Lemma 1.14. □

[⋆ End of the non-examinable part ⋆]

As the circle can be seen as the interval $[0, 1]$ pinned together at its endpoints, the same proposition says that there is no shift-invariant probability distribution on $[0, 1]$ that is defined on all subsets. This might seem like very bad news at first sight. However, it comes out that things can be mended, when one just restricts the collection of subsets $\mathcal{F}$. As mentioned, such a necessity can be compared to restricting the functions f for which one can define say a Riemann integral.

1.2.3 Borel σ-algebra on topological spaces

To define a probability measure on a set we need to define a σ-algebra - collection of subsets that we call events - and a probability measure that assigns each of these events a number, called probability. We saw that for uncountable spaces the set of all subsets might be too large. So what could we use?

- Suppose you want to talk of a random element of a topological space (X, τ). What should be the measurable sets be?

We should somehow use the notion of topology, and recall a topology is given by a collection of open sets. So it feels natural that we should be able to observe whether the random element is inside any given open set. So it is natural to define:

Definition 1.18 (Borel σ-algebra)**.** *Let (X, τ) be a topological space. The Borel σ-algebra $\mathcal{F}_\tau$ on X is defined to be the smallest σ-algebra that contains τ.*

This is well-defined because of the following lemma, whose proof you had on the exercise sheet, and which says that the intersection of σ-algebras is still a σ-algebra. Indeed, using this one can define the Borel sigma algebra $\mathcal{F}_\tau$ as the intersection of all σ-algebras containing τ.

Lemma 1.19 (Exo 1.3 in Dalang-Conus)**.** *Let Ω and I be two non-empty sets. Suppose that for each $i \in I$, $\mathcal{F}_i$ is a σ-algebra on Ω.*

- *Prove that $\mathcal{F} := \bigcap_{i\in I} \mathcal{F}_i$ is also a σ-algebra on Ω.*

[4] Recall that the Axiom of choice says the following: if you are giving any collection of non-empty sets $(X_i)_{i\in I}$, then their product is non-empty. In other words, you can define a function $f : I \to \bigcup_{i\in I} X_i$ such that for all $i \in I$, $f(i) \in X_i$.

- *Now, let $\mathcal{G}$ be any subset of $\mathcal{P}(\Omega)$. Then there exists a σ-algebra that contains $\mathcal{G}$ and that is contained in any other σ-algebra containing $\mathcal{G}$. This is called the σ-algebra generated by $\mathcal{G}$.*

Proof. On the exercise sheet. □

Notice that by definition all closed sets are then also measurable - for example the Borel σ-algebra on $\mathbb{R}$ contains all intervals (open, closed, half-open). Moreover, notice that depending on the topology the Borel σ-algebra can be much smaller than the power-set: for an example consider any space with the indiscrete topology. We will be mainly interested in the case of $(\mathbb{R}^n, \tau_E)$ and it comes out (but is not as easy to prove) that in this case the Borel σ-algebra is much smaller than the power set.

One of the reasons for using the Borel $\sigma-$algebra in probability[5] is the following nice proposition:

Proposition 1.20. *Consider two topological spaces (X_1, τ_1) and (X_2, τ_2). Prove that a continuous map $f : (X_1, \tau_1) \to (X_2, \tau_2)$ is at the same time also a measurable map from $(X_1, \mathcal{F}_{\tau_1})$ to $(X_2, \mathcal{F}_{\tau_2})$, where $\mathcal{F}_{\tau_1}$ and $\mathcal{F}_{\tau_2}$ denote the respective Borel σ-algebras.*

Proof. Let $f : (X_1, \tau_1) \to (X_2, \tau_2)$ be continuous. Then in particular for every $U \in \tau_2$ we have that $f^{-1}(U) \in \tau_1$. But then by the definition of the Borel σ-algebra, $f^{-1}(U) \in \mathcal{F}_{\tau_1}$. Moreover, by definition the open sets U generate the Borel σ-algebra $\mathcal{F}_{\tau_2}$. Thus we can use Exercise 1.8 to deduce that f is measurable. □

Now the collection of all open sets is still rather cumbersome to work with. However, in fact in the case of $(\mathbb{R}^n, \tau_E)$ one can find a much smaller collection of sets, basically just boxes, that also generate the Borel σ-algebra. This is quite similar to the fact that the Euclidean topology itself can be generated by the collection of open balls around rational points with rational radii.

Exercise 1.3. *The aim of this exercise is to prove that the Borel σ-algebra on $(\mathbb{R}^n, \tau_E)$, where τ_E is the Euclidean topology, is also*

(1) the smallest σ-algebra containing all boxes of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$;

(2) the smallest σ-algebra containing all half-boxes of the form $(-\infty, a_1] \times \cdots \times (-\infty, a_n]$.

To prove the first bullet point show that every open set $U \in \tau_E$ can be written as a countable union of boxes of the above form (hint: around each rational point in U consider a box that barely fits in U). Deduce the second bullet point from te first one.

Finally, as we will see in the next section, by restricting to Borel sigma-algebra we can finally talk about a uniform point on $[0, 1]$.

1.3 Probability measures on $\mathbb{R}^n$

This course will be mainly about studying probability measures on $\mathbb{R}$ and $\mathbb{R}^n$ - a random number will induce a probability measure on $\mathbb{R}$ and a $n-$tuple of random numbers a probability measure on $\mathbb{R}^n$. They come about also by studying abstract and complicated probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ via the notion of random variable that is introduced in the next

[5]Next semester you will see also something called the Lebesgue $\sigma-$algebra on $\mathbb{R}^n$ which is somewhat larger.

chapter. In short, random variables will be measurable functions from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{F}_{\tau_E})$ and help us describe the complicated universe: for example when we model the movement of a particle of dust, the whole state space is very complicated, it contains all the information on the movement, but we might be interested in the speed or the distance to its starting point and these would be real-valued random variables.

So this section is groundwork - for classifying and comparing random variables we need to understand probability measures on $(\mathbb{R}^n, \mathcal{F}_E)$, where by denote by $\mathcal{F}_E$ the Borel σ-algebra induced by the Euclidean topology, i.e. our usual $\sigma-$algebra from now onwards.

1.3.1 The uniform or Lebesgue measure

We start by the existence of the uniform probability measure on $[0,1]^n$ and the related shift-invariant measure of infinite mass on $\mathbb{R}^n$ - both sometimes called the Lebesgue measure and defined on the relevant Borel $\sigma-$algebra [6]. Interestingly proving their existence and uniqueness is really not that easy! Basically the reason is that it is just not simple to describe all Borel subsets and thus prescribe the probability measure, e.g. the function $\mathbb{P} : \mathcal{F}_E \to [0,1]$.

Thus the following result is out of the scope for this course, but will be proved in Analysis IV:

Theorem 1.21 (Existence and uniqueness of the Lebesgue measure on $\mathbb{R}^n$)**.** *Consider $(\mathbb{R}^n, \tau_E)$ with its Borel σ-algebra $\mathcal{F}_E$. Then there exists a unique measure μ on $(\mathbb{R}^n, \mathcal{F}_E)$ such that $\mu([a_1, b_1] \times \cdots \times [a_n, b_n]) = \Pi_{i=1}^n (b_i - a_i)$ for all vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ with real numbers $a_i < b_i$ for all $i < n$.*

Remark 1.22. *In fact, as you will see next semester the σ-algebra on which we can take the measure can be taken to be even larger - basically can also add all sets $S \subseteq \mathbb{R}^n$ such that there is some $B \in \mathcal{F}_E$ with $\mu(B) = 0$ and $S \subseteq B$. The resulting σ-algebra is called the Lebesgue σ-algebra. For probability, however, one usually works with the Borel σ-algebra.*

Notice that the Lebesgue measure is shift invariant:

Lemma 1.23. *Te Lebesgue measure on $(\mathbb{R}^n, \mathcal{F}_E)$ is shift invariant, i.e. for any $x \in \mathbb{R}^n$ and any $A \in \mathcal{F}_E$, we have that $\mu(A) = \mu(x + A)$, where $x + A = \{a + x : a \in A\}$.*

Proof. For every $r \in \mathbb{R}^n$, consider the function $f_r : \mathbb{R}^n \to \mathbb{R}^n$ defined by $f_r(x) := x + r$. As f_r is continuous, it is measurable by Proposition 1.20. Thus via Lemma 1.10 it induces a measure $\widetilde{\mu}(A)$ on $(\mathbb{R}^n, \mathcal{F}_E)$ by setting $\widetilde{\mu}(A) = \mu(f_r^{-1}(A))$.

The claim of the lemma is then equivalent to saying that $\widetilde{\mu}$ is the Lebesgue measure for every $r \in \mathbb{R}^n$. But notice that for every box $[a_1, b_1] \times \cdots \times [a_n, b_n]$ we have that

$$\widetilde{\mu}([a_1, b_1] \times \cdots \times [a_n, b_n]) = \mu([a_1 - r, b_1 - r] \times \cdots \times [a_n - r, b_n - r] = \Pi_{i=1}^n (b_i - a_i).$$

Thus by the uniqueness of Lebesgue measure we obtain that $\widetilde{\mathbb{P}}$ is also the Lebesgue measure as desired. □

In fact Lebesgue measure is the only shift invariant measure on $(\mathbb{R}^n, \mathcal{F}_E))$ up to a multiplicative constant.

[6] In fact they can be defined even on a slightly bigger $\sigma-$algebra that you will see next semester

Exercise 1.4 (Lebesgue measure is the only shift-invariant measure)**.** *Show that the Lebesgue measure is the only shift invariant measure μ on $(\mathbb{R}^n, \mathcal{F}_E)$ such that $\mu([0,1]^n) = 1$ and that every other shift invariant measure with $\mu([0,1]^n) \in (0,\infty)$ is given by a constant multiple of the Lebesgue measure.*

From the existence of Lebesgue measure on $\mathbb{R}^n$, we can also deduce the existence of what is called the uniform measure on $[0,1]^n$ by restriction. As its total mass is 1, this is really a probability measure on $[0,1]^n$ or if you wish on $\mathbb{R}^n$, in which case it just puts zero mass everywhere outside of the unit cube. As the arguments going inbetween are not of importance for us, let us just admit it too. [7]

Theorem 1.24 (Existence of Lebesgue measure on the unit cube, Lebsegue 1901 (admitted))**.** *There exists a unique probability measure $\mathbb{P}_U$ on $([0,1]^n, \mathcal{F}_E)$ such that $\mathbb{P}_U([0,x_1] \times \ldots [0,x_n]) = \Pi_{i=1}^n x_i$. Moreover such a $\mathbb{P}_U$ is shift-invariant: i.e. for any set $A \in \mathcal{F}_E$ and any $y \in [0,1]^n$ we have that $\mathbb{P}_U(A) = \mathbb{P}_U(A+y)$[8]. This is called the uniform measure or the Lebesgue measure on $[0,1]^n$.*

As first properties, notice that the uniform measure on $[0,1]$ (or $[0,1]^n$) doesn't put any mass on single points of $[0,1]$. This is really different from the countable situations where $\mathbb{P}$ was uniquely defined by its value on individual points of Ω!

Exercise 1.5. *Consider the Lebesgue measure $\mathbb{P}_U$ on $([0,1]^n, \mathcal{F}_E)$ as defined in the notes. Argue that for each $x \in [0,1]^n$ we have that $\{x\} \in \mathcal{F}_E$. Show that also $\mathbb{Q}^n \cap [0,1]^n \in \mathcal{F}_E$. What is $\mathbb{P}_U(\{x\})$? What is $\mathbb{P}_U(\mathbb{Q}^n \cap [0,1]^n)$?*

In particular, this means that for example (a,b) and $[a,b)$ and $[a,b]$ have the same Lebesgue measure.

Finally, let us verify that restricting to Borel-measurable sets was an actual restriction w.r.t. power set.

Exercise 1.6. *Show that not all subsets of $[0,1]$ are Borel-measurable. Can you find a description of a non-measurable subset? [hint: use Proposition 1.17]*

1.3.2 General probability measures on $\mathbb{R}$

We already saw one nice probability measure on $\mathbb{R}$ - the uniform measure on $[0,1]$. We now ask about general probability measures on $(\mathbb{R}, \mathcal{F}_E)$.

In fact the situation is really nice - as in the discrete case, we can identify all possible probability measures with a certain set of functions:

Definition 1.25 (Cumulative distribution function)**.** *We call a function $F : \mathbb{R} \to [0,1]$ a (cumulative) distribution function (abbreviated c.d.f.) if it satisfies the following conditions:*

(1) F is non-decreasing;
(2) $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$;
(3) F is right-continuous, i.e. for any $x \in \mathbb{R}$ and any sequence $(x_n)_{n \geq 1} \in [x, \infty)$ such that $x_n \to x$, we have that $F(x_n) \to F(x)$.

[7] Yet, for the curious - you can in fact deduce either of the theorems from each other. Can you figure out the arguments?

[8] here $A + y$ is considered modulo 1, i.e. in $n = 1$ for example $A + y = \{a + y \mod 1 : a \in A\}$.

The following key theorem says that cumulative distribution functions are in one-to-one correspondence with probability measures on $(\mathbb{R}, \mathcal{F}_E)$:

Theorem 1.26 (Classification of probability measures on $(\mathbb{R}, \mathcal{F}_E)$)**.** *Each probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_E)$ gives rise to a cumulative distribution function by defining $F(x) := \mathbb{P}((-\infty, x])$. Inversely, each cumulative distribution F gives rise to a unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_E)$ satisfying $\mathbb{P}((-\infty, x]) = F(x)$.*

For example here are three quite common probability measures describing very different situations:

- For a fair coin toss we can use a probability measure takes values 0 and 1 with probability 1/2. This probability measure on $(\mathbb{R}, \mathcal{F}_E)$ can be defined by setting $\mathbb{P}_C(F) = 0.5|F \cap \{0, 1\}|$ for every $F \in \mathcal{F}_E$. You can easily verify that this is a probability measure and then $F(x) = 0.5 1_{x \geq 0} + 0.5 1_{x \geq 1}$
- For the random stick-length we can use the uniform measure on $[0, 1]$. We defined it on $[0, 1]$ above, but we can extend it to $\mathbb{R}$ by setting $\widetilde{\mathbb{P}}_U(F) = \mathbb{P}_U(F \cap [0, 1])$ for every $F \in \mathcal{F}_E$. Again, it is an easy check that this indeed gives a probability measure. $\mathbb{P}_U$ corresponds to the function F defined by $F(x) = x 1_{x \in [0,1]} + 1_{x > 1}$.
- Finally, measurement errors are often described by what is called the Gaussian measure. The standard Gaussian measure on $\mathbb{R}$ by definition corresponds to the function F defined by $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{y^2}{2}) dy$.

This theorem above is pretty powerful! It says that every possible probability measure for a random number is described by just some monotonously growing right-continuous function. Notice that the c.d.f. F encodes a priori just the probability of each half interval $(-\infty, x]$, so actually $\mathbb{P} : \mathcal{F}_E \to [0, 1]$ is uniquely determined by just its values on sets $(-\infty, x]$.

We will prove this theorem in two steps: first we show that each probability measure gives rise to a c.d.f and show that conversely each c.d.f. gives rise to at least one probability measure with the above conditions; thereafter we argue that there is a unique such probability measure. To prove the first part we need a strengthening of Proposition 1.20 in the case of real numbers.

Exercise 1.7 (Monotonicity and measurability)**.** *Let $B \subseteq \mathbb{R}$ be an interval. Consider a non-decreasing (or non-increasing) function $f : B \to \mathbb{R}$. Then f is measurable from $(B, \mathcal{F}_E)$ to $(\mathbb{R}, \mathcal{F}_E)$.*

Proof of Theorem 1.26: part I. First, suppose that $\mathbb{P}$ is a probability measure on $(\mathbb{R}, \mathcal{F}_E)$ and let $F(x) = \mathbb{P}((-\infty, x])$. Then as $(-\infty, x] \subseteq (-\infty, y]$ for $x \leq y$, we have by (1) of Proposition 1.5 that F is non-decreasing.

Let us next check right-continuity of F. So let $(x_n)_{n \geq 1}$ be any sequence in $[x, \infty)$ converging to x. Then setting $A_n := \cap_{1 \leq k \leq n} (-\infty, x_k]$ we get that $\bigcap_{n \geq 1} A_n = (-\infty, x]$ and right-continuity follows from (5) of Proposition 1.5.

Now, if $(x_n)_{n \geq 1} \to -\infty$ we have that $\bigcap_{n \geq 1} (-\infty, x_n] = \emptyset$. Hence similarly to above (5) of Proposition 1.5 implies that $F(x_n) \to 0$. Finally, if $(x_n)_{n \geq 1} \to \infty$, we have $\bigcup_{n \geq 1} (-\infty, x_n] \to \mathbb{R}$ and thus by (2) of the same proposition again $F(x_n) \to 1$.

The other direction is more interesting. Suppose we are given a cumulative distribution function F. The idea is to now push the uniform measure on $((0, 1], \mathcal{F}_E)$ to $\mathbb{R}$ via a suitable

function f, defined using F. To do this define $f : (0,1] \to \mathbb{R}$ by

$$f(x) = \inf_{y \in \mathbb{R}} \{F(y) \geq x\}.$$

Then clearly f is non-decreasing and hence by Exercise 1.7 above measurable from $((0,1], \mathcal{F}_E)$ to $(\mathbb{R}, \mathcal{F}_E)$. Hence by Lemma 1.10 the uniform measure $\mathbb{P}_U$ induces a push-forward measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_E)$.

But now

$$\mathbb{P}((-\infty, x]) = \mathbb{P}_U((0, \sup_{z \in (0,1]} \{z < F(x)\})) = \mathbb{P}_U((0, F(x)]) = F(x)$$

and hence indeed F is the cumulative distribution function of $\mathbb{P}$.

This finishes the existence part of the theorem. □

To prove the uniqueness part we will try to understand a bit better how general Borel sets on $\mathbb{R}$ look like.

Remark 1.27 ($\star$ non-examinable $\star$)**.** *In fact, one can also deduce the uniqueness part in the theorem above from the uniqueness of uniform measure on $[0,1]$ by hand. Indeed, the function $F : \mathbb{R} \to [0,1]$ is non-decreasing and hence Borel-measurable. Notice that by definition any push-forward measure $\mathbb{P}_*$ on $([0,1], \mathcal{F}_E)$ under this map satisfies $\mathbb{P}_*([a,b]) = b - a$. By the uniqueness of uniform measure on $[0,1]$, i.e. Theorem 1.24, this must be the uniform measure. Now, the key observation (that is not as easy to prove, but you can try!) is that such a non-decreasing function F also maps any Borel measurable set $C \subseteq B$ to a Borel measurable set. But then by construction the measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_E)$ satisfies $\mathbb{P}(E) = \mathbb{P}_*(F(E))$ for any $E \in \mathcal{F}_E$ and thus it is uniquely determined.*

The following proposition says that given a probability measure on $(\mathbb{R}, \mathcal{F}_E)$ we can approximate each Borel set $B \in \mathcal{F}_E$ by disjoint (not necessarily open or closed, nor finite) intervals:

Lemma 1.28 (Approximation of Borel sets by disjoint intervals)**.** *Let $\mathbb{P}$ be a probability measure on $(\mathbb{R}, \mathcal{F}_E)$. Then for every $B \in \mathcal{F}_E$ and every $\epsilon > 0$, one can find a finite number of disjoint intervals or half-lines $I_1, \ldots, I_n$ such that $\mathbb{P}(B \Delta (I_1 \cup \ldots I_n)) < \epsilon$.*

The strategy of the proof used here is sometimes called Dynkin's argument:

Proof. Let $\mathcal{H}$ be the set of all subsets of $\mathbb{R}$ for which this property is true. Then $\mathcal{H}$ contains all intervals and half-lines and hence the smallest σ-algebra containing $\mathcal{H}$ is the Borel $\sigma-$algebra. Thus it suffices to show that $\mathcal{H}$ also is a σ-algebra as then by definition of the smallest σ-algebra containing a certain set, we must have $\mathcal{F}_E \subseteq \mathcal{H}$.

To show that $\mathcal{H}$ is a σ-algebra we verify the defining properties: the case of $\emptyset$ is clear, as we can just take half-lines $(-\infty, x]$ and we have seen that then $\mathbb{P}((-\infty, x]) \to 0$ as $x \to -\infty$. The case of complements is also clear, as firstly the complement of a finite disjoint union of intervals and half-lines is of the same form and secondly

$$A \Delta B = (A \backslash B) \cup (B \backslash A) = (A \cap B^c) \cup (B \cap A^c) = (B^c \backslash A^c) \cup (A^c \backslash B^c) = B^c \Delta A^c.$$

So we are left to show that if $H_1, H_2, \ldots$ are in $\mathcal{H}$, then so is $\bigcup_{i \geq 1} H_i$. To prove this, let us fix some $\epsilon > 0$. First, as $\mathbb{P}$ is a probability measure and $A_n = \bigcup_{i=1}^n H_i$ are increasing sets with $\bigcup_{n \geq 1} A_n = \bigcup_{i \geq 1} H_i$, we can choose $n \in \mathbb{N}$ such that $\mathbb{P}(\bigcup_{i \geq 1} H_i) \setminus A_n) < \epsilon/2$. Further,

for each H_i with $i = 1 \dots n$ we can pick disjoint intervals or half-lines $U_{i,1}, \dots, U_{i,m_i}$ such that $\mathbb{P}(H_i \Delta (U_{i,1} \cup \dots \cup U_{i,m_i})) < \frac{\epsilon}{2n}$.

Now, notice that the union of all $U_{i,j}$ with $1 \leq i \leq n, 1 \leq j \leq m_i$ is still a finite union of disjoint intervals or half-lines $V_1, \dots, V_l$. Indeed, this is true for any pair of intervals or half-lines [9] and by induction follows for the union of any k intervals or half-lines. On the other hand we can write

$$\mathbb{P}(A_n \Delta (V_1 \cup \dots V_l)) = \mathbb{P}\left((\bigcup_{i=1}^{n} H_i) \Delta (\bigcup_{i=1\dots n, j=1\dots m_i} U_{i,j}) \right).$$

Using $(A \cup B)\Delta(C \cup D) \subset (A \Delta C) \cup (B \Delta D)$ repeatedly we obtain

$$\left(\bigcup_{i=1}^{n} H_i \right) \Delta \left(\bigcup_{i=1\dots n, j=1\dots m_i} U_{i,j} \right) \subseteq \bigcup_{i=1}^{n} H_i \Delta (\cup_{j=1}^{m_i} U_{i,j})$$

and hence

$$\mathbb{P}\left((\bigcup_{i=1}^{n} H_i) \Delta (\bigcup_{i=1\dots n, j=1\dots m_i} U_{i,j}) \right) \leq \mathbb{P}(\bigcup_{i=1}^{n} H_i \Delta (\cup_{j=1}^{m_i} U_{i,j})) \leq \sum_{i=1}^{n} \mathbb{P}(H_i \Delta (\cup_{j=1}^{m_i} U_{i,j})) < \epsilon/2.$$

Using further $(A \cup B)\Delta C \subset (A \Delta C) \cup B$ we finally have

$$\mathbb{P}\left((\bigcup_{i \geq 1} H_i) \Delta (V_1 \cup \dots V_l) \right) \leq \mathbb{P}\left((\bigcup_{i=1}^{n} H_i) \Delta (V_1 \cup \dots V_l) \right) + \mathbb{P}(\bigcup_{i \geq 1} H_i) \setminus A_n) < \epsilon$$

and the proposition follows. □

The uniqueness part in Theorem 1.26 now follows from this more general corollary:

Corollary 1.29. *Suppose that two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $(\mathbb{R}, \mathcal{F}_E)$ satisfy $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$. Then in fact $\mathbb{P}_1(F) = \mathbb{P}_2(F)$ for all $F \in \mathcal{F}_E$. Further, the same conclusion holds if $\mathbb{P}_1((x, y)) = \mathbb{P}_2((x, y))$ for all real numbers $x < y$.*

Proof. First, notice that from the condition it follows that in fact $\mathbb{P}_1(I) = \mathbb{P}_2(I)$ for all intervals I. Let us check this for open intervals: we have that

$$(x, y) = \cup_{n \geq 1} \left((-\infty, y - 1/n] \setminus (-\infty, x] \right).$$

Hence we can write

$$\mathbb{P}_1((x, y)) = \lim_{n \to \infty} \mathbb{P}_1(-\infty, y - 1/n] \setminus (-\infty, x]) = \lim_{n \to \infty} (\mathbb{P}_1((-\infty, y - 1/n]) - \mathbb{P}_1((-\infty, x])).$$

By the assumption $\mathbb{P}_1((-\infty, y - 1/n]) - \mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, y - 1/n]) - \mathbb{P}_2((-\infty, x])$ and we conclude that $\mathbb{P}_1((x, y)) = \mathbb{P}_2((x, y))$. Similar arguments show this for all intervals. In particular the final claim of the corollary follows from the first claim.

To prove the first claim, we would naively like to apply Lemma 1.28 to deduce the equality for all Borel sets by using the fact that for any disjoint intervals or half-lines $I_1, \dots, I_n$, the above implies that $\mathbb{P}_1(I_1 \cup \dots \cup I_n) = \mathbb{P}_2(I_1 \cup \dots \cup I_n)$. Now, the problem is that Lemma 1.28 might give for the two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ very different approximating intervals for a fixed Borel set B. So in fact we have to rather repeat the argument.

[9] I encourage you to prove this rigorously - in fact we will see soon a convenient way to think about this in topology.

Thus, we consider the collection $\widetilde{\mathcal{H}}$ such that for all sets $H \in \widetilde{\mathcal{H}}$ and for every $\epsilon > 0$, there exists a finite number of intervals or half-lines $I_1, \ldots, I_n$ such that $\mathbb{P}_i(H \Delta (I_1 \cup \ldots I_n)) < \epsilon$ for both $i = 1, 2$. By above we know that $\widetilde{\mathcal{H}}$ contains all intervals. Further exactly the same proof as above implies that this collection is stable under complements and countable unions and hence $\widetilde{\mathcal{H}}$ is a σ-algebra containing all intervals, and hence equals the Borel σ-algebra. The corollary now follows.

Indeed, for any Borel set B, we just pick finite a sequence $(U_i)_{i \geq 1}$ of sets such that each U_i is a finite union of disjoint intervals (or half-lines) and further $\mathbb{P}_1(B \Delta U_i) \to 0$ and $\mathbb{P}_2(B \Delta U_i) \to 0$ as $i \to \infty$. Then in particular

$$\mathbb{P}_1(B) = \lim_{i \to \infty} \mathbb{P}_1(U_i) = \lim_{i \to \infty} \mathbb{P}_2(U_i) = \mathbb{P}_2(B).$$

□

Notice that we already had to basically go through the very same argument twice. In fact, there is a nice abstract results that avoids such repetition called the Dynkin theorem, which we will admit without proof:

Theorem 1.30 (Dynkin's uniqueness of extension (admitted))**.** *Let Ω be any set and $\mathcal{F}$ a σ-algebra. Suppose that $\mathcal{H} \subseteq \mathcal{F}$ is stable under intersection, i.e. if $H_1, H_2 \in \mathcal{H}$, then also $H_1 \cap H_2 \in \mathcal{H}$, and moreover $\sigma(\mathcal{H}) = \mathcal{F}$. Then any two finite measures $\mathbb{P}_1$ and $\mathbb{P}_2$ that agree on $\mathcal{H}$, agree on the whole of $\mathcal{F}$.*

One might ask, doesn't such an approximation already also give the construction of the Lebesgue measure? The problem is that here we supposed that a probability measure already exists and then showed that there is some nice way to approximate the measure of every Borel set. However, when we wanted to define the probability measure, we should really find a consistent way to choose approximations for every Borel set, and further then check that all the axioms for a probability measure hold. This is more strenuous and out of the scope of this course, yet will be done in Analysis 4 for the Lebesgue measure on $\mathbb{R}^n$.

1.3.3 General probability measures on $\mathbb{R}^n$

For probability measures on $(\mathbb{R}^n, \mathcal{F}_E)$, a similar characterization holds as for the one-dimensional case. We first define the cumulative distribution for probability measures on $\mathbb{R}^n$:

Definition 1.31 (Joint cumulative distribution function)**.** *Any function $F : \mathbb{R}^n \to [0, 1]$ is called a joint cumulative distribution function (c.d.f.), if it satisfies the following conditions:*

(1) F is non-decreasing in each coordinate.
(2) $F(x_1, \ldots, x_n) \to 1$ when all of $x_i \to \infty$.
(3) $F(x_1, \ldots, x_n) \to 0$, when at least one of $x_i \to -\infty$.
(4) F is right-continuous, meaning that for any sequence $(x_1^m, \ldots, x_n^m)_{m \geq 1}$ converging to $(x_1, \ldots, x_n)$ such that for all $m \geq 1$ we have that $x_i^m \geq x_i$, it holds that $F(x_1^m, \ldots, x_n^m) \to F(x_1, \ldots, x_n)$.

Notice that for $n = 1$ we are back to the case of individual c.d.f. Moreover, if we send any $n - 1$ coordinates to infinity, then we also obtain the c.d.f. of the remaining coordinate:

$$F_{X_i}(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty).$$

The key result now says that c.d.f.s are in one to one correspondence to probability measures on $\mathbb{R}^n$.

Theorem 1.32 (Joint c.d.f.s characterise probability measures on $\mathbb{R}^n$ (admitted))**.** *Each probability measure $\mathbb{P}$ on $(\mathbb{R}^n, \mathcal{F}_E)$ gives rise to a joint cumulative distribution function by defining*

$$F_{\overline{X}}(x_1, \dots, x_n) := \mathbb{P}_{\overline{X}}((-\infty, x_1] \times \cdots \times (-\infty, x_n]).$$

Inversely, each joint cumulative distribution F gives rise to a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^n, \mathcal{F}_E)$ satisfying $\mathbb{P}((-\infty, x_1] \times \cdots \times (-\infty, x_n)) = F(x_1, \dots, x_n)$.

In fact, it is again not hard to prove that given a probability measures, F gives rise to a joint cumulative distribution function - the proof is really like in the 1D case. However, the opposite statement - showing that every joint c.d.f. gives rise to a unique probability measure can not be concluded as simply and hence we assume it. In fact uniqueness follows from Dynkin's theorem above, but for existence we are not able to use a similar trick as in the 1D case.

There is one very special case of joint c.d.f: given n one-dimensional c.d.f.-s $F_1, F_2, \dots, F_n$ we can define $F(x_1, \dots, x_n) := F_1(x_1) \dots F_n(x_n)$. One can check that this really is a joint c.d.f. on $\mathbb{R}^n$ and hence gives rise to a probability measure. The induced probability measure is quite special and is called the product measure.

1.4 Product measures on $\mathbb{R}^n$ and $\mathbb{R}^{\mathbb{N}}$

To finish this section we will look more closely at the product measure mentioned above. Constructing probability spaces by taking products of existing probability spaces is natural from a mathematical point of view - like product spaces in topology - but maybe even more importantly, it also relates to one of the most important concepts in probability, that of independence.

Notice that we need to do steps: first, constructing the product σ-algebra which is very analogous to the construction of a product topology and is rather intuitive. However, the presence of the measure adds a layer of extra difficulty - we also need to define a product measure on this product σ-algebra. We will do the first step in some generality, and for the second we mainly concentrate on our case of interest - the cases $\mathbb{R}^n$ and $\mathbb{R}^{\mathbb{N}}$.

1.4.1 Product σ-algebras

The definition of a product σ-algebra for countably many spaces is rather direct:

Definition 1.33 (Product σ-algebra)**.** *Let I be countable and $(\Omega_i, \mathcal{F}_i)$ with $i \in I$ be non-empty measurable spaces. We define the product σ-algebra $\mathcal{F}_\Pi$ on $\Pi_{i \in I} \Omega_i$ by taking the σ-algebra generated by $\{\Pi_{i \in I} F_i : F_i \in \mathcal{F}_i \forall i \in I\}$, where $F_i \in \mathcal{F}_i$ and $F_i \neq \Omega_i$ for only finitely many $i \in I$.*

Remark 1.34. *It comes out that contrary to the case of product topology, for countable products of measurable spaces, we could as well take the σ-algebra generated by $\{\Pi_{i \in I} F_i : F_i \in \mathcal{F}_i \forall i \in I\}$, with no constraints on the number of non-trivial sets. I leave it as an exercise for the interested.*

One can ask if this is the right choice of a $\sigma-$algebra. At least on easy examples the product σ-algebra seems to behave well:

Exercise 1.8. *Suppose I is finite, each Ω_i countable and $\mathcal{F}_i = \mathcal{P}(\Omega_i)$. Show that the product σ-algebra on $\Pi_{i\in I}\Omega_i$ is the equal to $\mathcal{P}(\Pi_{i\in I}\Omega_i)$. Is it still the case when I is not finite?*

Also, as in the case of product topology one can characterize the product σ-algebra as the smallest σ-algebra such that all projection maps are measurable. In this respect we record the following lemma, that really comes directly from the definition:

Lemma 1.35 (Projections are measurable)**.** *Let I be countable and $(\Omega_i, \mathcal{F}_i)$ with $i \in I$ be non-empty measurable spaces. Consider the product space $(\Pi_{i\in I}\Omega_i, \mathcal{F}_\Pi)$. Then for any $J \subseteq I$, the projection $\pi : (\Pi_{i\in I}\Omega_i, \mathcal{F}_\Pi) \to (\Pi_{i\in J}\Omega_j, \mathcal{F}_\Pi)$ is measurable, when we consider $\Pi_{i\in J}$ with its $\sigma-$algebra.*

Finally, if (X_i, τ_i) are topological spaces, we now have two ways to construct a σ-algebra on $\Pi_{i\in I}X_i$: either as the Borel σ-algebra of the product topology, or the product σ-algebra of the individual Borel σ-algebras. The following proposition says that it doesn't matter as long as spaces are nice enough.

Proposition 1.36. *Let I be countable and (X_i, τ_i) are topological spaces, each with a countable basis and an associated Borel σ-algebra $\mathcal{F}_{X_i}$. Then the Borel σ-algebra on $(\Pi_{i\in I}X_i, \tau_{\Pi_{i\in I}X_i})$ is equal to the product σ-algebra $\mathcal{F}_\Pi$ on $\Pi_{i\in I}X_i$.*

For us the important message is that the Borel σ-algebra on $\mathbb{R}^n$ is the product of the Borel σ-algebras on $\mathbb{R}$. The proof itself is not difficult, but not that interesting, and thus non-examinable.

[$\star$ This proof is non-examinable $\star$]

Proof. Let us denote by $\sigma(\tau_\Pi)$ the Borel σ-algebra of the product topology on $\Pi_{i\in I}$.

We start by showing that $\mathcal{F}_\Pi \subseteq \sigma(\tau_\Pi)$. To show that, it suffices to show that $\sigma(\tau_\Pi)$ contains every set in some generating set of $\mathcal{F}_\Pi$. So consider the generating set of Definition 1.33 and consider one element of this set - $F = \Pi_{i\in I}F_i$ with $F_i \in \mathcal{F}_{X_i}$. By definition of the product topology, for every open set $U_j \in \tau_j$, we have that $\Pi_{i\in I}V_i \in \tau_\Pi$ where $V_j = U_j$ and $V_i = X_i$ for $i \neq j$. Thus as $\sigma(\tau_\Pi)$ is a σ-algebra, we conclude that $E_j = \Pi_{i\in I}\widehat{F}_i \in \sigma(\tau_\Pi)$, where $\widehat{F}_j = F_j$ and $\widehat{F}_i = X_i$ for $i \neq j$. But we have that $\bigcap_{i\in I} E_j = F$ and thus $F \in \sigma(\tau_\Pi)$.

We now show that $\sigma(\tau_\Pi) \subseteq \mathcal{F}_\Pi$. It suffices to prove that $\tau_\Pi \subseteq \mathcal{F}_\Pi$. We first claim that

Claim 1.37. *τ_Π admits a countable basis τ_Π^B.*

Proof of claim: Choose for each (X_i, τ_i) a countable basis τ_i^B. Now we know from the topology course that the sets of the form $\Pi_{i\in I}V_i$ where $V_i \in \tau_i^B$ and $V_i \neq X_i$ only for finitely many indices $i \in I$ form a basis of the product topology τ_Π. As the number of finite subsets of a countable set is countable, each τ_i^B is countable and a finite product of countable sets is countable, we conclude that there are countably many elements in this basis. □

From this claim it follows that each open set in τ_Π can be written as a countable union of sets in τ_Π^B. Hence, if $\tau_\Pi^B \subseteq \mathcal{F}_\Pi$, it follows that $\tau_\Pi \subseteq \mathcal{F}_\Pi$ as well. But by definition of the Borel σ-algebra, if $U_i \in \tau_i^B$, then $U_i \in \mathcal{F}_{X_i}$. Hence by the definition of the product σ-algebra $\mathcal{F}_\Pi$, it follows that $\tau_\Pi^B \subseteq \mathcal{F}_\Pi$. □

[$\star$ End of the non-examinable part $\star$]

1.4.2 Product probability measures on product σ-algebras

One can define many different probability measures on the product σ-algebra - indeed, we characterized for example all possible probability measures on $\mathbb{R}^n$ with its Borel $\sigma-$algebra (which by the proposition above is the same as the product $\sigma-$algebra).

Among these probability measures, there is one that is quite special - the so called product probability measure. Inducing such a probability measure is in fact not a trivial thing. It is rather easy, however, in one concrete setting - for finite products of discrete probability spaces:

Exercise 1.9 (Finite products of discrete spaces)**.** *Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1), \ldots, (\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be discrete probability spaces. Show by a direct construction that there is a unique probability measure $\mathbb{P}_\Pi$ on the measurable space $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_\Pi)$ such that for every $F \in \mathcal{F}_\Pi$ of the form $F_1 \times \cdots \times F_n$ with $F_i \in \mathcal{F}_i$ for all $i = 1 \ldots n$, we have that $\mathbb{P}_\Pi(F) = \Pi_{i=1}^n \mathbb{P}_i(F_i)$.*

This last property is also the very definition of the product measure:

Definition 1.38 (Product measure)**.** *For $i \in \mathbb{N}$, let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ be probability spaces. We call a probability measure $\mathbb{P}_\Pi$ on $(\Pi_{i\in\mathbb{N}}\Omega_i, \mathcal{F}_\Pi)$ a product measure of the collection $((\Omega_i, \mathcal{F}_i, \mathbb{P}_i))_{i\geq 1}$ if for any finite subset $J \subset \mathbb{N}$ and any event E of the form $E = \Pi_{i\in\mathbb{N}} F_i$ with $F_i = \Omega_i$ for $i \notin J$ and $F_i = E_i \in \mathcal{F}_i$ for $i \in J$, we have that*

$$\mathbb{P}_\Pi(E) = \Pi_{i\in J}\mathbb{P}_i(E_i). \tag{1.1}$$

Remark 1.39. *Notice that because every finite subset J is included in some set of the form $\{1, \ldots, n\}$, one can equivalently ask the condition for all sets J of this form.*

We will soon see that the product measure describes the situation where all probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ describe independent experiments. Namely, suppose that you can define the product of $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ with product measure as defined above. Now, consider the events $E_1 = F_1 \times \Omega_2$ and $E_2 = \Omega_1 \times F_2$ with $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$. Then by definition of product measure

$$\mathbb{P}_\Pi(E_1 \cap E_2) = \mathbb{P}_1(F_1)\mathbb{P}_2(F_2) = \mathbb{P}_\Pi(E_1)\mathbb{P}_\Pi(E_2).$$

Such a multiplicativity property will be the very definition of independence between events E_j and E_i in the next section.

Let us first verify that we have already constructed product measures on $(\mathbb{R}^n, \mathcal{F}_E)$: indeed, as mentioned the candidates are the probability measures induced by c.d.f.-s of the form $F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)$. It is easy to see that the c.d.f on any product probability measure does have to be of the form - this just follows by applying Condition (1.1) to events of the form $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$. In the other direction something needs to be verified!

Exercise 1.10. *Let $F_1, \ldots, F_n$ be c.d.f. Prove that $F(x_1, \ldots, x_n) := F_1(x_1) \ldots F_n(x_n)$ gives rise to a joint c.d.f. Further, show that the probability measure on $(\mathbb{R}^n, \mathcal{F}_E)$ induced by the c.d.f. F gives rise to a product probability measure. [Hint: try out the lemma above approximation of Borel sets and induction]*

This settles the bill for finite products of copies of $\mathbb{R}$. But what about countable products? As discussed, these appear naturally when considering sequences of experiments, e.g. coin tosses.

The general case of countable product spaces is actually quite tricky, and is out of the scope of this course. However, in case of product space of $\mathbb{R}$ with its Borel $\sigma-$algebra, there is again a slick way to go about it. The key lemma is the following bi-measurable correspondence between $(\{0,1\}^{\mathbb{N}}, F_\Pi)$ and $([0,1], \mathcal{F}_E)$:

Lemma 1.40 (Dyadic correspondence)**.** *For each $x \in [0,1]$ consider its dyadic expansion $x = \sum_{i\geq 1} 2^{-i} x_i$, where we make the expansion unique by choosing it such that it doesn't end in a infinite sequence of 1-s. Then the map $f : [0,1] \to \{0,1\}^{\mathbb{N}}$ defined by $f(x) = (x_1, x_2, \dots)$ is injective and measurable from $([0,1], \mathcal{F}_E)$ to $(\{0,1\}^{\mathbb{N}}, \mathcal{F}_\Pi)$. Similarly, the map $g : \{0,1\}^{\mathbb{N}} \to [0,1]$ given by $(x_1, x_2, \dots) \to \sum_{i\geq 1} 2^{-i} x_i$ is surjective and measurable.*

Proof. Injectivity and surjectivity are clear. Measurability in both directions follows from the following points:

(1) $\mathcal{F}_\Pi$ is generated by the sets of the form $F_1 \times F_2 \times \cdots \times F_n \times \{0,1\} \times \{0,1\} \times \dots$ (from definition);
(2) $\mathcal{F}_E$ is generated by intervals of the form $[j2^{-n}, (j+1)2^{-n})$ over $j = 1 \dots 2^n$ and $n \geq 1$ (this is a small check);
(3) the sets of the form $F_1 \times F_2 \times \cdots \times F_n \times \{0,1\} \times \{0,1\} \times \dots$ are correspondence with finite unions of intervals of the type above via f or g.

To see the third point, notice that every set of the form $E = \Pi_{i\in I} F_i$ where $F_i = \{\omega_i\}$ for all $i \leq n$ and $F_i = \{0,1\}$ otherwise is in correspondence with an interval of length 2^{-n} of the form above. □

As a first consequence, we can already construct the product space for infinitely many fair coin tosses:

Proposition 1.41 (Space of infinite fair coin tosses)**.** *For each $i \geq 1$ let $\Omega_i = \{0,1\}$, $\mathcal{F}_i = \mathcal{P}(\Omega_i)$ and $\mathbb{P}_i(0) = \mathbb{P}_i(1) = 1/2$. Then there exists a product probability measure $\mathbb{P}_\Pi$ on $(\Pi_{i\geq 1}\Omega_i, \mathcal{F}_\Pi)$.*

Notice that in particular each sequence of n coin tosses has probability exactly 2^{-n}, i.e. like in the case of Laplace model for n equivalent coin tosses.

Proof. Consider the dyadic map $f : [0,1] \to \{0,1\}^{\mathbb{N}}$ from the lemma above. This lemma says that the map is measurable from $([0,1], \mathcal{F}_E)$ to $(\{0,1\}^{\mathbb{N}}, \mathcal{F}_\Pi)$. Thus, by Lemma 1.10, the uniform measure $\mathbb{P}_U$ on $([0,1], \mathcal{F}_E)$ induces a probability measure $\mathbb{P}_\Pi$ on $(\{0,1\}^{\mathbb{N}}, \mathcal{F}_\Pi)$.

It remains to see that this measure is indeed a product measure. Fix some $\omega \in \{0,1\}^{\mathbb{N}}$, i.e. $\omega_i \in \{0,1\}$ for each $i \geq 1$. Now, consider a finite subset $J = \{1, \dots, n\} \subseteq \mathbb{N}$ and set $F_i = \{\omega_i\}$ for all $i \in J$ and $F_i = \{0,1\}$ otherwise, and let $E = \Pi_{i\in I} F_i$. Now observe that $\mathbb{P}_U(f^{-1}(E)) = 2^{-n}$. But this is exactly equal to $\Pi_{i\in J}\mathbb{P}_i(F_i)$ and thus we indeed have a product measure. □

We now go on to prove the existence of general product measures on $\mathbb{R}^{\mathbb{N}}$.

Theorem 1.42 (Product probability measure on $\mathbb{R}^{\mathbb{N}}$)**.** *For $i \geq 1$, let $(\mathbb{R}, \mathcal{F}_E, \mathbb{P}_i)$ be probability measures. Then there exists a probability measure $\mathbb{P}_\Pi$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi)$ such that $\mathbb{P}_\Pi$ is a product measure of the collection $((\mathbb{R}, \mathcal{F}_E, \mathbb{P}_i))_{i\geq 1}$.*

The non-examinable proof of this theorem uses slickly the fact that $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$ are in bijection and the Dyadic lemma we proved in the opposite direction. Namely, from any

countable product of infinite fair coin tosses we can construct the uniform probability measure on $([0,1], \mathcal{F}_E)$. But now by looking at this countable product as a countable product of countable products, we can get a countable product of such uniform probability measures! In other words, we can construct a nice measure on $[0,1]^{\mathbb{N}}$. The general case then proceeds as in Theorem 1.26.

[$\star$ This proof is non-examinable $\star$]

Proof. **Step 1: uniform measure on $[0,1]^{\mathbb{N}}$.**

We start by constructing the product probability measure on $\mathbb{N}$ copies of $([0,1], \mathcal{F}_E, \mathbb{P}_U)$. In this respect, consider any bijection $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ (exo: find an explicit such map!).

This induces a map $G : \{0,1\}^{\mathbb{N}} \to [0,1]^{\mathbb{N}}$ as follows: each sequence $(a_i)_{i\in\mathbb{N}} \in \{0,1\}^{\mathbb{N}}$ is mapped to $(b_i)_{i\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$ by setting $b_i = \sum_{j\geq 1} a_{g(i,j)} 2^{-j}$, i.e. we define b_i via its dyadic expansion $0.a_{g(i,1)}a_{g(i,2)}\dots$.

We claim that this map is measurable from $(\{0,1\}^{\mathbb{N}}, \mathcal{F}_\Pi)$ to $([0,1]^{\mathbb{N}}, \mathcal{F}_\Pi)$, where as customary we abuse a bit the notation and denote by $\mathcal{F}_\Pi$ the relevant product σ-algebras (that are not the same in this case). Indeed, the product $\sigma-$algebra on $[0,1]^{\mathbb{N}}$ is generated by the events of the form $\Pi_{i\in\mathbb{N}} F_i$ with $F_i \in F_{[0,1]}$. By Lemma 1.40 together with Lemma 1.35 we know that each $G^{-1}([0,1] \times \dots \times F_i \times [0,1] \times \dots)$ is measurable in the product $\sigma-$algebra of $\{0,1\}^{\mathbb{N}}$. But $G^{-1}(\Pi_{i\in\mathbb{N}} F_i)$ is a countable intersection of such sets, and thus is also measurable.

It remains to check that we indeed have a product probability measure, as defined just above. First, let $n \in \mathbb{N}$ and for $i \leq n$ let F_i be of the form $[a_i, b_i)$ with a_i, b_i both of the form $k2^{-m}$ for some $k, m \in \mathbb{N}$. Then from the correspondence in Lemma 1.40 it follows that $G^{-1}(F_1 \times \dots \times F_n \times [0,1] \times \dots)$ is of the form $\Pi_{i\in\mathbb{N}} E_i$, where only finitely many E_i are different from $\{0,1\}$. From the fact that $\mathbb{P}_\Pi$ is a product measure, it then readily follows that

$$\widetilde{\mathbb{P}}_\Pi(F_1 \times \dots \times F_n \times [0,1] \times \dots) = \Pi_{i\leq n}(b_i - a_i).$$

To obtain the condition for product measure for all sets of the form $F_1 \times \dots \times F_n \times [0,1] \times \dots$, with $F_i \in \mathcal{F}_E$ we first notice that the condition also holds if F_1 is a disjoint union of the intervals of the above form, and by approximation it holds for any disjoint union of intervals or half-lines. We can then use Lemma 1.28 to conclude it for all $F_1 \in \mathcal{F}_E$. We then further apply induction to extend the applicability to all $F_i \in F_E$.

Step 2: the general case.

The general case is now rather easy. Namely, we can define $F : ([0,1]^{\mathbb{N}}, \mathcal{F}_\Pi) \to (\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi)$ by setting $F(x_1, x_2, \dots) = (f_1(x_1), f_2(x_2), \dots)$, where the functions f_i are defined as in Theorem 1.26 by

$$f_i(x) = \inf_{y\in\mathbb{R}} \{F_i(y) \geq x\}.$$

The rest follows similarly to Step 1.

□

[$\star$ End of non-examinable part $\star$]

Finally, it is important to notice that the measure we just constructed on $([0,1]^{\mathbb{N}}, \mathcal{F}_\Pi)$ interacts well with the Lebesgue measure on $[0,1]^n$:

Exercise 1.11. *For $i \geq 1$, let $(\mathbb{R}, \mathcal{F}_E, \mathbb{P}_i)$ be probability measures. Consider the product probability measure $\mathbb{P}_\Pi$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi)$ of the collection $((\mathbb{R}, \mathcal{F}_E, \mathbb{P}_i))_{i\geq 1}$.*

Further, for $n \in \mathbb{N}$, consider the projection $\pi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^n$ to first $n-$coordinates, i.e. the map $(x_1, x_2, \dots) \to (x_1, \dots, x_n)$. Show that the pushforward measure of $\mathbb{P}_\Pi$ induced on $(\mathbb{R}^n, \mathcal{F}_E)$ by this projection is characterized by the c.d.f.

$$F(x_1, \dots, x_n) = \Pi_{i=1}^n \mathbb{P}_i((-\infty, x_i]).$$

1.5 Conditional probability and independence

In this subsection we work solely with probability spaces and introduce a central notion of probability - that of independence. Recall that then the σ-algebra $\mathcal{F}$ is the collection of all events that can be observed, and for each such event $E \in \mathcal{F}$, we have defined a probability $\mathbb{P}(F) \in [0,1]$.

We saw in the case of Laplace model that probability has one interpretation as modelling the frequency of something happening in a repeated experiment, when each experiment 'does not influence' the others. We will now develop a mathematical meaning to this 'does not influence'. More generally, we will set up the vocabulary to talk about how the knowledge of about some random event, influences the probabilities we should assign to other events. Here, the other common interpretation of probability as a degree of belief enters very naturally.

1.5.1 Conditional probability

We have already considered (in the course and on the example sheets) many unpredictable situations where several events naturally occur either at the same time or consecutively: a sequence of coin tosses, random walks. Here, whether we observe some event at step n depends on what has happened before. Similarly, if you want to model the weather or the financial markets tomorrow, you better take into account what happened today. To talk about the change of probabilities when we observe something, we introduce the notion of conditional probability:

Definition 1.43 (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$. Then for any $F \in \mathcal{F}$, we define the conditional probability of the event F given E (i.e. given that the event E happens), by*

$$\mathbb{P}(F|E) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}.$$

Recall that $E \cap F$ is the event that both E and F happen. Hence, as the denominator is always given by $\mathbb{P}(E)$, the conditional probability given E is proportional to $\mathbb{P}(E \cap F)$ for any event F. Here is the justification for dividing by $\mathbb{P}(E)$:

Lemma 1.44. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$. Then $P(\cdot|E)$ defines a probability measure on $(\Omega, \mathcal{F})$.*

Proof. First, notice that $\mathbb{P}$ is indeed defined for every $F \in \mathcal{F}$. Next, $\mathbb{P}(\emptyset|E) = \mathbb{P}(\emptyset)/\mathbb{P}(E) = 0$ and $\mathbb{P}(\Omega|E) = \mathbb{P}(E)/\mathbb{P}(E) = 1$. So it remains to check countable additivity.

So let $F_1, F_2, \ldots \mathcal{F}$ be disjoint. Then also $E \cap F_1, E \cap F_2, \ldots$ are also disjoint. Hence

$$\mathbb{P}(\bigcup_{i\geq 1} F_i|E) = \frac{\mathbb{P}((\bigcup_{i\geq 1} F_i) \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(\bigcup_{i\geq 1}(F_i \cap E))}{\mathbb{P}(E)} = \sum_{i\geq 1} \frac{\mathbb{P}(F_i \cap E)}{\mathbb{P}(E)} = \sum_{i\geq 1} \mathbb{P}(F_1|E),$$

and countable additivity follows.

□

It should be remarked that conditional probability might be similar to the initial probability (we will see more about this very soon), but might also be drastically different. A somewhat silly but instructive example is the following: conditional probability of the event E^c, conditioned on E is always zero, no matter what the original probability was; similarly the conditional probability of E, conditioned on E is always 1.

Exercise 1.12. *The French, Swiss and German decide to elect the greatest mathematician of all time. The French propose Poincaré, the Swiss propose Euler and the German Gauss. Each country has one vote, and the candidate with most votes wins. In case of equal votes, the winner is chosen uniformly randomly. Now Mathematico, an organization that predicts elections, forecasts that*

- *the French will give their vote with probability* $1/2$ *to Poincaré and equally with probability* $1/4$ *to Euler or Gauss;*
- *the Swiss will give their vote with probability* $1/2$ *to Euler and equally with probability* $1/4$ *to Poincaré or Gauss;*
- *the German will give their vote with probability* $1/2$ *to Gauss and equally with probability* $1/4$ *to Poincaré or Euler.*

Moreover, Mathematico thinks that none of the countries cares about the opinion of the others.

Build a probabilistic model to be able to predict the winner. What assumptions are you making? In this model, what is the probability that Euler wins? What is the probability that Euler gets at least 2 *votes? Now, surprisingly it comes out that the Swiss have elected Gauss instead of Euler. How would you now estimate the probability that Euler still wins the election?*

One also has to be very careful about the exact conditioning, as similarly sounding conditionings can also have very different conditional probabilities.

Exercise 1.13. *Roger Federer is now* 70 *years old and still playing. He is a bit tired of running and has limited his strategy in his serve game: he either serves an ace with probability* $1/2$ *and obtains a point, or with the same probability makes a double fault and his opponent gains a point. The game has also been simplified and the player who first obtains 3 points wins. Build a probabilistic model (or several) to answer the following questions and answer them:*

- *What is the probability that Roger wins his serve game?*
- *What is the probability that Roger won his serve game, given that he hit at least two aces?*
- *What is the probability that he will win his serve game, given that he started by hitting two aces?*

Still, although conditional probabilities are often tricky, they are very important and useful. The following result is a generalization of the following intuitive result: if you know that exactly one of three events E_1, E_2, E_3 happens, then to understand the probability of any other event F, it suffices to understand the conditional probabilities of this event, conditioned on each of E_i, i.e. the probabilities $\mathbb{P}(F|E_i)$.

Proposition 1.45 (Law of total probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Further, let I be countable and $(E_i)_{i\in I}$ be disjoint events with positive probability and such that $\Omega \setminus \left(\bigcup_{i\in I} E_i\right)$ has zero probability. Then for any $F \in \mathcal{F}$, we can write*

$$\mathbb{P}(F) = \sum_{i\in I} \mathbb{P}(F|E_i)\mathbb{P}(E_i).$$

Proof. We can write F as a disjoint union

$$F = \left(F \cap (\bigcup_{i\in I} E_i)\right) \cup \left(F \cap (\Omega \setminus (\bigcup_{i\in I} E_i))\right)$$

and as $\mathbb{P}\left(F \cap (\Omega \setminus (\bigcup_{i\in I} E_i))\right) = 0$ by assumption, we see by additivity of $\mathbb{P}$ under disjoint unions that $\mathbb{P}(F) = \mathbb{P}\left(F \cap (\bigcup_{i\in I} E_i)\right)$.

Now rewrite $F \cap (\bigcup_{i\in I} E_i) = \bigcup_{i\in I}(F \cap E_i)$. Because $(E_i)_{i\in I}$ are disjoint, so are $(F \cap E_i)_{i\in I}$. Hence again by countable additivity for disjoint sets

$$\mathbb{P}(F) = \mathbb{P}\left(\bigcup_{i\in I}(F \cap E_i)\right) = \sum_{i\in I} \mathbb{P}(F \cap E_i).$$

Now, by definition $\mathbb{P}(F \cap E_i) = \mathbb{P}(F|E_i)\mathbb{P}(E_i)$ and the proposition follows.

□

1.5.2 Independence

Things simplify a lot when the probability of an event does not change, when conditioned on another event - i.e. when $\mathbb{P}(E|F) = \mathbb{P}(E)$. Such events are called independent. In fact the rigorous definition is slightly different:

Definition 1.46 (Independence for two events)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that two events E, F are independent if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$.*

Observe that when $\mathbb{P}(F) > 0$, then we get back to the intuitive statement of independence, i.e.that $\mathbb{P}(E|F) = \mathbb{P}(E)$. Indeed, if E and F are independent we can write

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(E)\mathbb{P}(F)}{\mathbb{P}(F)} = \mathbb{P}(E).$$

Here are some basic properties of independence:

Lemma 1.47 (Basic properties)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

- *If E is an event with $\mathbb{P}(E) = 1$ then it is independent of all other events.*
- *If E, F are independent, then also E^c and F are independent.*
- *Finally, if an event is independent of itself, then $\mathbb{P}(E) \in \{0, 1\}$.*

Proof. Let $E, F \in \mathcal{F}$. By inclusion-exclusion formula

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).$$

Now, if $\mathbb{P}(E) = 1$ then also $\mathbb{P}(E \cup F) \geq \mathbb{P}(E) = 1$ and hence this gives $\mathbb{P}(E \cap F) = \mathbb{P}(F) = \mathbb{P}(F)\mathbb{P}(E)$ and hence E and F are independent.

For the second property, we can write by law of total probability

$$\mathbb{P}(E^c \cap F) + \mathbb{P}(E \cap F) = \mathbb{P}(F).$$

By independence of E, F we have $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$ and thus it follows that

$$\mathbb{P}(E^c \cap F) = \mathbb{P}(F)(1 - \mathbb{P}(E)) = \mathbb{P}(F)\mathbb{P}(E^c)$$

as desired.

Finally, if E is independent of itself then $\mathbb{P}(E) = \mathbb{P}(E \cap E) = \mathbb{P}(E)^2$. Hence $\mathbb{P}(E)(1 - \mathbb{P}(E)) = 0$, implying that $\mathbb{P}(E) \in \{0, 1\}$. □

There are two different ways to generalize independence to several events:

- mutual independence
- and pairwise independence

The stronger and more important notion is that of mutual independence.

Definition 1.48 (Mutual independence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let I be an index set. Then the events $(E_i)_{i \in I}$ are called mutually independent if for any finite subsets $I_1 \subseteq I$ we have that*

$$\mathbb{P}\left(\bigcap_{i \in I_1} E_i\right) = \Pi_{i \in I_1} \mathbb{P}(E_i).$$

Further, for two sets of events $(E_i)_{i \in I}$ and $(F_j)_{j \in J}$ we say that they are mutually independent if for all $i \in I, j \in J$:

$$\mathbb{P}(E_i \cap F_j) = \mathbb{P}(E_i)\mathbb{P}(F_j).$$

Mutual independence is naturally linked to product measures. As we haven't discussed product measures on general spaces, let us restrict ourselves here to product measures of $(\mathbb{R}, \mathcal{F}_E, \mathbb{P})$.

Lemma 1.49. *Let $(\mathbb{R}, \mathcal{F}_E, \mathbb{P}_i)$ for $i \geq 1$ be probability measures and consider their product measure $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$. Then for every collection $(E_i)_{i \geq 1}$ with $E_i \in F_E$ we have that the events $F_i = \mathbb{R} \times \mathbb{R} \times \ldots E_i \times \mathbb{R} \times \ldots$ with E_i in the i-th coordinate are mutually independent.*

Proof. This follows directly from the definition of product measure. □

In particular, to model events that we expect to be mutually independent we also naturally go towards product measures. For example. To model a sequence of n independent fair coin tosses we take the product space of $(\{0, 1\}, \mathcal{P}(\{0, 1\}))$ with the probability measure that sets $\mathbb{P}(\{0\}) = \mathbb{P}(\{1\}) = 1/2$. You can check that the model you get is exactly the Laplace model on n indistinguishable fair coin tosses that we discussed in the beginning of the course.

Similarly, one can check that the uniform random graphs we considered in Example 1.16 can be actually modeled on a product space:

Exercise 1.14. *Consider the model of uniform random graphs in Example 1.16. Let $E_{i,j}$ be the event that the edge $\{i,j\}$ is present. Prove that the events $E_{i,j}$ are independent. Find the appropriate product space to model uniform random graphs.*

The assumption of mutual independence helps to also build more complicated probability models. For example, suppose you have a coin that is not fair, but comes up heads with probability $p \in (0,1)$. How would you assign probabilities to a sequence of n tosses? The assumption of all sequences being equally likely does not make sense any longer (e.g. think of the case when p is near 1, then certainly the sequence of all zeros and all ones cannot have the same probabilities).

However, the assumption of mutual independence and its relation to product measures help. Indeed, you would still take the product space of $(\{0,1\}, \mathcal{P}(\{0,1\}), \mathbb{P}_p)$ but would now consider $\mathbb{P}_p$ such that it gives 1 with probability p and 0 with probability $1-p$. You can check that then the probability of a specific sequence of n tosses with m heads and tails $n-m$ is by independence just $p^m(1-p)^{n-m}$. (Why do those sum up to one? Check!)

Sometimes one does not have the full mutual independence or at least does not know it holds, and just pairwise independence can be asserted. There are similar notions of $k-$wise independence.

Definition 1.50 (Pairwise independence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let I be an index set. Then the events $(E_i)_{i\in I}$ are called pairwise independent if for any $i \neq j \in I$ the events E_i and E_j are independent.*

It is important to notice that, whereas mutual independence clearly implies pairwise independence, the opposite is not true in general:

Exercise 1.15 (Pairwise independent but not mutually independent)**.** *Consider the probability space for two independent coin tosses. Let E_1 denote the event that the first coin comes up heads, E_2 the event that the second coin comes up heads and E_3 the event that both coin come up on the same side. Show that E_1, E_2, E_3 are pairwise independent but not mutually independent.*

Finally, the notion of independence works as expected also under the conditional probability measure:

Definition 1.51 (Conditional independence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let I be an index set. Then the events $(F_i)_{i\in I}$ are called conditionally independent given E if for any finite subsets $I_1 \subseteq I$ we have that*

$$\mathbb{P}\left(\bigcap_{i\in I_1} F_i | E\right) = \Pi_{i\in I_1} \mathbb{P}(F_i|E).$$

As with conditional probability, conditioning can also change the presence or absence of independence - as a silly extreme example again the event E on which you condition, becomes independent of everything. We will meet a more interesting example very soon.

Exercise 1.16. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and E_1, E_2, E_3 pairwise independent events with positive probability. Show that if E_1 and E_2 are conditionally independent, given E_3, then E_1, E_2, E_3 are mutually independent.*

1.5.3 Bayes' rule

Mostly one hears about conditional probabilities not through independence, but through the Bayes' rule:

Proposition 1.52 (Bayes' rule)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and E, F two events of positive probability. Then*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F)}$$

It's not only that the statement looks innocent, but also the proof is a one-liner - by definition of conditional probability, we can write

$$\mathbb{P}(E|F)\mathbb{P}(F) = \mathbb{P}(E \cap F) = \mathbb{P}(F|E)\mathbb{P}(E).$$

So why is this simple result so important and talked-about? Let us look at some examples. Thomas Bayes himself was looking at (a slightly more advanced version of) the following example: suppose that every week the same lottery takes place with the same rules. To begin with, you don't know what is the probability p of winning this lottery, you only know it is either $1/3$ or $2/3$.

But now, you have played n times and won m times - can you say whether anything about the winning probability? Clearly, the number of times you have won tells you something about this probability - if you win every single time, you would guess that this probability is rather $2/3$ than $1/3$; if you never win in 10000 rounds, you probably guess the opposite.

To analyse this situation more precisely, we want to construct a probability space containing both the information about the winning probability and the outcomes of each weekly lottery. The notion of conditional independence helps us in this construction - whereas the events of winning are not independent of each other if the value of p is unknown, they become independent, if you condition it being equal to $1/3$ or $2/3$. (Why?) Thus we can build our probability space as follows

- $\Omega = \{1/3, 2/3\} \times \{0, 1\}^n$, where the first co-ordinate denotes the unknown winning probability and the others the outcomes of n weekly lotteries by setting 1 if we win, and 0 if we lose.
- A priori all possible combinations could be observed, so you set $\mathcal{F} := \mathcal{P}(\Omega)$.
- Finally, how should we set the probabilities? As we know nothing about p, we should probably consider both possibilities of p equally likely. As mentioned, for any fixed choice of probability p, all the weekly lotteries are conditionally independent given p and win with probability p. Thus, conditoned on p, a sequence with m wins and $n-m$ losses would have probability $p^m(1-p)^{n-m}$, as in the case of coin tosses above.

Now, if we denote by F_i the event that $p = i/3$ and by E_m the event that we got m wins, then from our model we can calculate that $\mathbb{P}(E_m|F_i) = \binom{n}{m}(i/3)^m(1-i/3)^{n-m}$. Also, by assumption $\mathbb{P}(F_i) = 1/2$. Finally, to calculate $\mathbb{P}(E_m)$ we can use the law of total probability to get that $\mathbb{P}(E_m) = \sum_{i=1}^{2} \frac{1}{2}\binom{n}{m}(i/3)^m(1-i/3)^{n-m}$. Thus using Bayes formula we obtain an exact expression for $\mathbb{P}(F_i|E_m)$:

$$\mathbb{P}(F_i|E_m) = \frac{\frac{1}{2}\binom{n}{m}(i/3)^m(1-i/3)^{n-m}}{\sum_{i=1}^{2} \frac{1}{2}\binom{n}{m}(i/3)^m(1-i/3)^{n-m}}.$$

This is quite nice! And this explains the usefulness of Bayes' rule. Namely, very often we start modelling unknown situations from very little information, so to build up our probabilistic model we have to use some assumptions – like the assumptions of equal probability for each winning probability in this concrete case – and when we have more data, and more observations we can start updating our model to build a more accurate description of the situation.

Most often, one hears about Bayes' rule though in the realm of medicine. Let us give an example of this from late spring of a year that will not be remembered happily:

Exercise 1.17 (Bayes' rule and positive test results)**.** *In late spring 2020 several antibody tests to see whether your body has produced antibodies against SARS-CoV-2 and thus whether you could be immune to COVID at least that moment. Their preciseness was a good-sounding* 95%*, meaning that both false-positives (the test tells that you have antibodies when you actually don't) and false-negatives (the test tells that you don't have antibodies, but you actually do) would only appear in* 5% *of the tests taken. However, despite this good preciseness, caution was recommended in interpreting your result. Let's try to understand why:*

- *You hear someone claim that, when some tests positive they have* 95% *chance of actually having antibodies. Is this statement correct?*
- *Now, consider this additional information: in late spring 2020 it was estimated that* 5% *of the population have actually been in contact with SARS-CoV-2. Which probability space would you now build to estimate the probability that you have antibodies after a positive test? What is this probability? What if you take two independent tests on the same day and both come up positive?*
- *Suppose now that* 50% *of the population have been in contact with SARS-CoV-2. How does this change the result?*

Section 2

Random variables and random vectors

The notion of a random variable is central in probability theory and in describing the world using probability theory - they help us model the random quantities we observe.

Random variables are measurable functions

Mathematically, a $(\Omega_2, \mathcal{F}_2)$-valued random variable is just a measurable function $X : \Omega \to \Omega_2$ from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Omega_2, \mathcal{F}_2)$. Often one uses the notion of a random variable to only talk about $(\mathbb{R}, \mathcal{F}_E)$-valued random variables. We will follow this custom and call $(\mathbb{R}, \mathcal{F}_E)$-valued random variables just random variables. In case we consider the more general notion, we will talk explicitly of $(\Omega_2, \mathcal{F}_2)$-valued random variables.

Events as random variables

In fact, we have already seen some random variables: for example, given $(\Omega, \mathcal{F}, \mathbb{P})$, for every event $E \in \mathcal{F}$, we can define the indicator function $1_E : \Omega \to \mathbb{R}$ by $1_E(\omega) = 1_{\omega \in E}$. This indeed defines a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{F}_E)$ as the the preimages of $F \in \mathcal{F}_E$ under this map are either $\emptyset, E, E^c$ or Ω. The random variable 1_E thus encodes whether the event E happened or not.

Many more random variables

However, random variables go much further and allow us to talk not only about whether a random event happened or not, but also about what exactly happened. E.g. using random variables we can ask the exact number of dots on a dice or what will be the temperature tomorrow? How many people will vote for Trump? How many students will show up for the live ZOOM discussion? Or, in case of more general random variables - how does the trajectory of an errantly moving molecule look like? What is the shape of a random walk?

Random variables vs probability measures

Random variables are in fact very strongly connected to probability measures. Namely any probability measure $(\Omega, \mathcal{F}, \mathbb{P})$ gives rise to a $(\Omega, \mathcal{F})$-valued random variable by just defining the measurable map $X : \Omega \to \Omega$ as the identity map $X(\omega) = \omega$. Thus every probability space can be also seen as a random variable.

In the other direction, we have seen in Lemma 1.10 that any measurable map from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Omega_2, \mathcal{F}_2)$, i.e. a $(\Omega_2, \mathcal{F}_2)$-valued random variable always induces a probability $\mathbb{Q}$ measure on $(\Omega_2, \mathcal{F}_2)$ - for all $F \in \mathcal{F}_2$, we set $\mathbb{Q}(F) = \mathbb{P}(X^{-1}(F))$. Thus also every random variable gives rise to a probability measure on $(\Omega_2, \mathcal{F}_2)$.

Hence we can in some sense equate any $(\Omega_2, \mathcal{F}_2)$-random variable with just a probability measure on $(\Omega_2, \mathcal{F}_2)$.

Why random variables at all?

Given that random variables are just measurable functions and moreover the relation between probability measures and random variables above, one might ask, why do we need this concept at all?

Maybe indeed, mathematically, random variables are not something really new, at least not like the concept of a topological space is. However, they do simplify life and offer a new way of thinking:

- *Complicated situations are described by several random variables defined on the same space.* Whereas it is true that a single random variable can be as well just equated with a probability measure on its image space, usually we are studying complicated situations, like weather, and they are described by many random variables at the same time. In this case, it is much more convenient that all the unknown / all the randomness is encoded in this one space $(\Omega, \mathcal{F}, \mathbb{P})$ that denotes the universe, and random variables are quantities that we have access to, that we can measure.
- *Random variables allow for arithmetic operations.* Often, in addition to wanting to talk about numerical values of some experiments and observations, we also want to further manipulate random quantities. The concept of random variable is large enough to allow for that.
- *Both 'variable' and 'random' are good to keep in mind.* Also, in comparison to measurable functions, the words of 'variable' and 'random' fit better with our mental picture. Indeed, the idea of a 'variable' is useful to keep in mind – we think of a value that varies when the state $\omega \in \Omega$ varies. And although a random variable is in the end just a function. Similarly, the word 'random' also has its place – we think of the state $\omega \in \Omega$ in the domain space of this function as something unknown, as something we cannot predict and don't have access to, so as something 'random'.
- *We can forget the underlying probability space.* We will see that when looking at real-valued random variables, we can actually even just forget about the basic, possibly over-complicated space $(\Omega, \mathcal{F}, \mathbb{P})$ and start concentrating on what we really can measure and observe - the random variables.

This is now enough of chit-chat. Let us get to maths.

2.1 (Real) random variables

For concreteness, let us define again:

Definition 2.1 (Random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then any measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F}_E)$ is called a random variable. We call the probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_E)$ defined for all $E \in \mathcal{F}_E$ by*

$$\mathbb{P}_X(E) = \mathbb{P}(X^{-1}(E)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in E\})$$

the law or the distribution of the random variable X.

Notice that the fact that $\mathbb{P}_X$ is a probability measure follows from Lemma 1.10. For $E \in \mathcal{F}_E$ we will often use the notations

$$\mathbb{P}(X \in E) := \mathbb{P}(X^{-1}(E))$$

insisting that we think of X as a random quantity taking some values. We also denote the event $\{\omega \in \Omega : X(\omega) = k\}$ simply by $\{X = k\}$ or even by just $X = k$. By custom, we keep the capital letters X, Y, Z often for random variables - not to confuse with the same notation also often used for topological spaces!

Here are some concrete examples of probability spaces and random variables defined on them.

- *Indicator functions of events.* As explained in the introduction of this section, the easiest random variables arise when asking whether and event happened or not. So for example if we consider the probability space for a fair dice $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ the uniform measure on Ω, then for any $E \subseteq \Omega$, the indicator function 1_E is a random variable. Indeed, for any $F \in \mathcal{F}_E$, the preimage of F under 1_E is either equal to E, E^c, Ω or $\emptyset$ and by definition they are all measurable sets of Ω. We will return to such random variables soon and call them *Bernoulli random variables.*
- *The number of heads.* For $n \in \mathbb{N}$ consider the probability space $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathbb{P})$ where $\mathbb{P}$ is the probability measure that treats each sequence of coin tosses as equal. Let us show that

  $$X_1 = \text{total number of heads}$$

  is a random variable: indeed, we just need to show that X_1 is a measurable function from $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathbb{P})$ to $(\mathbb{R}, \mathcal{F}_E)$. But all subsets of the probability space are measurable, so the condition is automatically satisfied. This happens always when the σ-algebra on our initial probability space is the power-set – this should remind you of the fact that all functions from a top. space with the discrete topology are continuous. We will in a few lectures time introduce a general class of similar random variables called *Binomial random variables.*
- *Properties of a random graph.* Further, we could also consider the example of uniform random graphs on n vertices as in the Exercise sheet 1 or 3. Then again, we used the power-set as the σ-algebra on the set Ω of all possible graphs on n vertices. Thus both

  $$Y_1 = \text{the number of edges that are present}$$

  and

  $$Y_2 = \text{the number of connected components}$$

  are random variables. Notice that using these random variables we can much more freely talk about this random graph and about how it looks like.
- *Properties of a random walk.* As a final example, consider the model of random walks on n steps as on the Example sheet 2 – again, we can describe this model well using random variables. E.g.

  $$Z_1 = \text{maximal value of the walk}$$

  and

  $$Z_2 = \text{the number of times the walk visits zero}$$

are both random variables. This is again just because our probability space for random graphs was built using the power set as a σ-algebra and in that case all real valued functions $F : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F}_E)$ are measurable and hence random variables.

- *Standard normal random variable.* The random variable X, whose c.d.f. is given by the c.d.f. of the Gaussian measure will be called the standard normal or standard Gaussian random variable. We will come back to this very soon.

As you notice in all cases we are really interested in the image of the function X - the domain Ω is of little interest, we really care about which values in $\mathbb{R}$ are taken with which probability. This also motivates the main notion of equality in the world of random variables - the equality in law:

Definition 2.2 (Equality in law)**.** *Two random variables X, Y are said to be equal in law or equal in distribution, denoted $X \sim Y$ if for every $E \in \mathcal{F}_E$ we have that $\mathbb{P}_X(E) = \mathbb{P}_Y(E)$.*

Pictorially, this means the following. For a random variable that takes a finite number of values, you can always describe it using a histogram: you make a column for each possible value y of the random variable X, and then make the value of the column equal to $\mathbb{P}(X = y)$. In this respect equality in law just means that the two histograms are the same.

Notice that a priori even the underlying probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ could be different - we are only interested that they give rise to the same law on $(\mathbb{R}, \mathcal{F}_E)$. So in that sense the underlying probability space plays only an auxiliary role here. In particular, we can really start comparing different probabilistic phenomena in different context and mathematically saying that some random numbers explaining them look the same or look different.

We will later on see another notion of equality called almost sure equality, which requires the random variables to be defined on the same space.

2.1.1 The cumulative distribution function of a random variable

Our first aim is to get some understanding about which random variables exist and how to classify them. To do this, recall that we obtained in the first chapter a characterization of all probability measures on $(\mathbb{R}, \mathcal{F}_E)$ using cumulative distribution functions (Theorem 1.26).

But now, each random variable is described by the probability measure it induces on $(\mathbb{R}, \mathcal{F}_E)$. Thus this very same theorem implies directly that all random variables can be characterized using c.d.f.-s too: (Verify that you understand why this is just a rewording of what we have!)

Proposition 2.3 (Cum.dist. function of a random variable)**.** *For each random variable X (defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$), we have that $F_X(x) = \mathbb{P}(X \in (-\infty, x])$ defines a cumulative distribution function (c.d.f). Moreover, each cumulative distribution function gives rise to a unique law of a random variable.*

The final bit can be rephrased by saying that two random variables with the same cumulative distribution function are equal in law.

For example, what would be the c.d.f of the so called Bernoulli random variable X that takes value 1 with probability p and 0 with probability $1 - p$? All indicator functions of events correspond to such random variables with $\mathbb{P}(E) = p$. We would have $F_X(x) = (1 - p)1_{x \geq 0} + p1_{x \geq 1}$. More generally for a random variable that takes only finite number of values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, we would have $F_X(x) = \sum_{i=1 \ldots n} p_1 1_{x \geq x_i}$. (Why?)

Thus we see that F_X encodes the behaviour of X rather naturally. Let us now look at this relation between the cumulative distribution function F_X and the random variable X more closely. By $F(x^-)$ we denote the limit of $F(x_n)$ with $(x_n)_{n\geq 1} \to x$ from below.

Lemma 2.4 (C.d.f vs r.v.)**.** *Let X be a random variable on some probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and F_X its cumulative distribution function. Then for any $x < y \in \mathbb{R}$*

(1) $\mathbb{P}(X < x) = F(x-)$
(2) $\mathbb{P}(X > x) = 1 - F(x)$
(3) $\mathbb{P}(X \in (x, y)) = F(y-) - F(x)$.
(4) $\mathbb{P}(X = x) = F(x) - F(x-)$.

Proof. This is on exercise sheet. □

Thus we see that all jumps of F_X correspond to points where $\mathbb{P}_X(X = x) > 0$. But how many jumps are there?

Lemma 2.5. *A cumulative distribution function F_X of a random variable X has at most countably many jumps.*

Proof. Let S_n be the set of jumps that are larger than $1/n$ and $\widehat{S}_n$ any finite subset of S_n. Then $\widehat{S}_n$ is measurable and $1 \geq \mathbb{P}(X \in S_n) \geq |\widehat{S}_n| n^{-1}$. Thus it follows that $|\widehat{S}_n| \leq n$. As this holds for any finite subset of S_n, we deduce that $|S_n| \leq n$ and in particular S_n is finite.

Now the set of all jumps can be written as a union $\bigcup_{n\geq 1} S_n$. Hence as each S_n is finite and a countable union of finite sets is countable, we conclude. □

These jumps of a c.d.f. F_X are sometimes called atoms of the law of X. More precisely, we call $s \in \mathbb{R}$ an atom for the law of X if and only if $\mathbb{P}(X = s) > 0$.

In the extreme case F_X increases only via jumps, i.e. is piece-wise constant changing value at most countable times. Precisely:

Definition 2.6 (Piece-wise constant with at most countable jumps)**.** *We say that $f : \mathbb{R} \to [0, \infty)$ is piece-wise constant with countably many jumps iff there is some countable set S and some real numbers $c_s > 0$ for $s \in S$ such that $\sum_{s\in S} c_s < \infty$ and*

$$f(x) = \sum_{s\in S} c_s 1_{x\geq s}.$$

In the other extreme F_X could also be everywhere continuous. These motivate the following definitions:

Definition 2.7 (Discrete and continuous random variable)**.** *Let X be a random variable. If F_X is piece-wise constant changing value at most countable many times, we then call the X a discrete random variable. If F_X is continuous, we call X a continuous random variable.*

Another, maybe somewhat simpler equivalent description of discrete random variable is as follows:

Exercise 2.1. *Prove that a random variable X is discrete if and only if there is a countable set $S \subseteq \mathbb{R}$ such that for all $s \in S$ we have that $\mathbb{P}(X = s) > 0$ and $\mathbb{P}(X \in S) = 1$. We call S the support of the discrete random variable X.*

This also makes the vocabulary coherent with what we have seen in the first chapter - although a priori a discrete random variable X takes values on $\mathbb{R}$, we have seen that effectively it takes values only on the countable set S and thus $\mathbb{P}_X$ can be defined on a discrete probability space.

As the following proposition says, the c.d.f. of any random variable can be written as a convex combination of c.d.f-s of a discrete and continuous random variable.

Proposition 2.8. *Any cumulative distribution function F_X of a random variable X can be written uniquely as convex combination of cumulative distribution functions of a continuous random variable Y_1 and of a discrete random variable Y_2 - i.e. for some $a \in [0,1]$ we have that $F_X = aF_{Y_1} + (1-a)F_{Y_2}$.*

Proof. If X is either continuous or discrete, the existence of such writing is clear. So suppose that X is neither continuous nor discrete. Write S for the countable set of jumps of F_X. Define

$$\widehat{F}_{Y_1}(x) = \sum_{s \in S} 1_{x \geq s}(F_X(s) - F_X(s-)).$$

Then $\widehat{F}_{Y_2} := F_X - \widehat{F}_{Y_1}$ is continuous: indeed, by definition both F_X and $\widehat{F}_{Y_1}$ both right-continuous, and thus is also their difference. Moreover, both are continuous at any continuity point of F_X, i.e. when $x \notin S$. Finally, when $x \in S$, then again by definition of $\widehat{F}_{Y_1}$, we have that

$$F_X(s) - F_X(s-) = 1_{s \geq s}(F_X(s) - F_X(s-)) = \widehat{F}_{Y_1}(s) - \widehat{F}_{Y_1}(s-).$$

Now, as X is neither discrete nor continuous, we have that $0 < \widehat{F}_{Y_1}(\infty) < 1$ and $0 < \widehat{F}_{Y_2}(\infty) < 1$. Hence, we can define

$$F_{Y_1}(x) := \frac{\widehat{F}_{Y_1}(x)}{\widehat{F}_{Y_1}(\infty)}$$

and

$$F_{Y_2}(x) := \frac{\widehat{F}_{Y_2}(x)}{\widehat{F}_{Y_2}(\infty)}.$$

By definition both of those are non-decreasing, right-continuous with $F_{Y_i}(-\infty) = 0$ and $F_{Y_i}(\infty) = 1$ and hence are c.d.f-s for random variables. As F_{Y_1} increases only via jumps and F_{Y_2} is continuous, we have the desired writing with $a = \widehat{F}_{Y_1}(\infty)$ and $1 - a = \widehat{F}_{Y_2}(\infty)$.

To see the uniqueness of the decomposition, suppose that one can write

$$F_X = aF_{Y_1} + (1-a)F_{Y_2} = bF_{Z_1} + (1-b)F_{Z_2},$$

where both Y_1 and Z_1 are discrete and Y_2, Z_2 continuous random variables. Then $aF_{Y_1} - bF_{Z_1}$ has to be continuous, but also piecewise constant with countably many jumps. As $aF_{Y_1}(-\infty) - bF_{Z_1}(-\infty) = 0$, the only possibility is that it is constantly zero. As $F_{Y_1}(\infty) = 1 = F_{Z_1}(\infty)$, it follows that $a = b$ and $F_{Y_1} = F_{Z_1}$. Thus also $F_{Y_2} = F_{Z_2}$ and the proposition follows.

□

We will see how to interpret this result by saying any random variable can be seen as combination of a discrete and continuous random variable. However, to get there we first

have to develop some theory, e.g. the notion of independence for random variables. Let us start by just meeting some more random variables.

2.1.2 Discrete random variables

There are several families of laws of discrete random variables that come up again and again. As we will see, sometimes these laws also have very nice mathematical characterizations. For each of these cases we really determine only the law of the random variable: a probability law of $(\mathbb{R}, \mathcal{F}_E)$. To characterise it, we can either give a c.d.f. or determine $\mathbb{P}_X(F)$ for a sufficiently large set of F (e.g. all intervals). In the discrete case we saw that the random variable only takes a countable set of values, so we could just also determine this set of values S and determine $\mathbb{P}_X(X = s)$ for each $s \in S$.

Uniform random variable
Any random variable that takes values in a finite set $S = \{x_1, \ldots, x_n\}$, each with equal probability $1/n$ is called the uniform random variable on S. We call the law of this random variable the uniform law. Its c.d.f is given by simply $F_X(x) = n^{-1} \sum_{i=1}^{n} 1_{x \geq x_i}$.

Examples are - a fair dye, the outcome of roulette, taking the card from the top of a well-mixed pack of cards etc...Concretely, for a trivial example is that if we model a fair dye on $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(i) = 1/6$, then the random variable $X(\omega) = \omega \in \mathbb{R}$ gives rise to a uniform random variable (why?).

We use this family of random variables every time we have no a priori reason to prefer one outcome over the other. A fancy mathematical way of saying this would be to say that the uniform law is the only probability law on a finite set that is invariant under permutations of the points. We will also see on the example sheet that this is the so called maximum entropy probability distribution with values in a finite set S.

Bernoulli random variable
As mentioned already, a random variable that takes only values $\{0, 1\}$, taking value 1 with probability p is called a Bernoulli random variable of parameter p. It is named after the Swiss mathematician Bernoulli, who also thought that all sciences need mathematics, but mathematics doesn't need any. Leaving you to judge, let us see that these examples come up very often.

Namely, on every probability space $(\Omega, \mathcal{F}, \mathbb{P})$, every indicator function of an event, i.e. 1_E gives rise to a Bernoulli random variable and the parameter p is equal to the probability of the event. Indeed for any event E in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ the indicator function $1_E : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F})$ is measurable and hence a random variable. Moreover, it is $\{0, 1\}$ valued by definition and $\mathbb{P}(\{1_E = 1\}) = \mathbb{P}(E) = p$. Sometimes one talks about Bernoulli random variables more generally whenever there are two different outcomes, e.g. also when the values are $\{-1, 1\}$.

Binomial random variable
A random variable that takes values in the set $\{0, 1, \ldots, n\}$, and takes each value k with probability

$$p^k (1-p)^{n-k} \binom{n}{k}$$

is called a binomial random variable of parameters $n \in \mathbb{N}$ and $0 \leq p \leq 1$ (why do the probabilities sum to one?). We denote the law of such a binomial random variable by $Bin(n, p)$.

Notice that for $n = 1$, we have the Bernoulli random variable. We met the binomial random variable already in the beginning of the section, where we considered the number of heads for a sequences of n fair coin tosses, in a situation where each sequence has equal probability. We will see it also naturally comes up in models of random graphs, or models of random walks. The reason why it comes up so often is that it always describes the following situation - we have a sequence of independent indistinguishable events and we count the number of those who occur. Here is a precise statement:

Exercise 2.2 (Binomial r.v. is the number of occurring events)**.** *Suppose we have n mutually independent events $E_1, \ldots, E_k$ of probability p on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the random number of events that occurs: $X = \sum_{i=1}^{n} 1_{E_i}$. Prove that X is a random variable and has the law $Bin(n, p)$.*

For a concrete lively example, let's go back to the Erdos-Renyi random graph on n vertices from the example sheet, where each edge is independently included with probability p. We can then fix some vertex v and consider the random variable M_v giving the number of vertices adjacent to v, i.e. linked to v by an edge. The exercise above shows that this random variable has law $Bin(n-1, p)$.

Geometric random variable

A random variable that takes values in the set $\mathbb{N}$, each value k with probability $p(1-p)^{k-1}$ for some $0 < p \leq 1$ is called a geometric random variable of parameter p. We denote the law of a geometric random variable by $Geo(p)$. One should again check that this even defines a random variable, by seeing that the probabilities do sum to one.

A geometric random variable describes the following situation: we have independent events $E_1, E_2, \ldots$ each of success probability p and we are asking for the smallest index k such that the event E_k happens. For example, $Geo(1/2)$ describes the number of tosses needed to get a first heads. This will be made precise on the exercise sheet.

There is also a nice property that characterizes the geometric r.v.:

Lemma 2.9 (Geometric r.v. is the only memoryless random variable)**.** *We say that a random variable X with values in $\mathbb{N}$ is memoryless if for every $k, l \in \mathbb{N}$ we have that $\mathbb{P}_X(X > k + l | X > k) = \mathbb{P}_X(X > l)$. Every geometric random variable is memoryless, and in fact these are the only examples of memoryless random variables on $\mathbb{N}$.*

Proof. Let us start by proving that the geometric random variable satisfies the memoryless property. First, notice that if $\mathbb{P}(X = 1) = 1$, then X is a degenerate geometric random variable with $p = 1$. So we can suppose that we work in the case $\mathbb{P}(X > 1) > 0$.

Let us check that a geometric r.v. is memoryless. First, it is easy to check that for a geometric random variable X, we have that $\mathbb{P}(X > l) = (1-p)^l$ for some $p \in (0, 1]$. As by the definition of conditional probability

$$\mathbb{P}(X > k + l | X > k) = \frac{\mathbb{P}(X > k + l)}{\mathbb{P}(X > k)},$$

it follows that $\mathbb{P}(X > k + l | X > k) = (1-p)^{k+l-k} = \mathbb{P}(X > l)$ as desired.

Now, let us show that each random variable satisfying the memoryless property has the law of a geometric random variable. Again if $\mathbb{P}(1) = 1$, we are done. Otherwise we can write

$$\mathbb{P}(X > 1 + l | X > 1)\mathbb{P}(X > 1) = \mathbb{P}(X > 1 + l).$$

Thus for a memoryless random variable

$$\mathbb{P}(X > l)\mathbb{P}(X > 1) = \mathbb{P}(X > l + 1).$$

Thus inductively $\mathbb{P}(X > l) = \mathbb{P}(X > 1)^l$ and hence X is a geometric random variable of parameter $p = 1 - \mathbb{P}(X > 1)$. □

Poisson random variable

Poisson was a French mathematician who has famously said that the life is good for only two things - mathematics and teaching mathematics. His random variables come up quite often.

The Poisson random variable is a discrete random variable with values in $\{0\} \cup \mathbb{N}$ and taking the value k with probability

$$e^{-\lambda}\frac{\lambda^k}{k!}$$

for some $\lambda > 0$. We denote this distribution by $Poi(\lambda)$. Poisson random variables describe occurrences of rare events over some time period, where events happening in any two consecutive time periods are independent. For example, it has been used to model

- The number of visitors at a small off-road museum.
- More widely, the number of stars in a unit of the space.
- Or more darkly, it was used to also model the number of soldiers killed by horse kicks in the Prussian army.

One way we see the Poisson r.v. appearing is via a limit of the Binomial distribution if the success probability p scales like $1/n$:

Lemma 2.10 (Poisson random variable as the limit of Binomials)**.** *Consider the Binomial distribution $Bin(n, \lambda/n)$. Prove that as $n \to \infty$ it converges to $Poi(\lambda)$ in the sense that for every $k \in \{0\} \cup \mathbb{N}$, we have that*

$$\mathbb{P}(Bin(n, \lambda/n) = k) \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

Proof. By definition, for any fixed $n \in \mathbb{N}$ and $k \in \{0\} \cup \mathbb{N}$, we have

$$\mathbb{P}(Bin(n, \lambda/n) = k) = \binom{n}{k}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Using

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

we can write

$$\mathbb{P}(Bin(n, \lambda/n) = k) = \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1 - \frac{\lambda}{n}\right)^{-k}.$$

But now as $n \to \infty$

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

Moreover, for any fixed $t > 0$ also $\frac{n-t}{n} \to 1$ as $n \to \infty$ and hence

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$$

and

$$\left(1 - \frac{\lambda}{n}\right)^{-k} = \left(\frac{n-\lambda}{n}\right)^{-k} \to 1,$$

proving the lemma. □

To connect this to the occurrences of rare events described before, one could think as follows. Suppose we try to model the number of arrivals over time window $[0, 1]$, say one year in a distant location. We then cut a time-window $[0, 1]$ into n equal time-segments of length $1/n$ with n large, say into 365 days, so that we can suppose that at each time-segment, say each day, there is at most one arrival. In this case we can describe the arrival or non-arrival using $Ber(p)$ or 1_E for some event E. If we further suppose that all days are alike, we can take this parameter p to be the same for all time-segments of the same length, e.g. for all days. Moreover, if we suppose that an arrival in one time-segment does not influence arrivals in other time-intervals, we can assume that all events E corresponding to different time intervals are mutually independent. Hence the total number of arrivals is the number of independent events happening, when the event probability is p - we saw above that this gives a $Bin(n, p)$ random variable. But now, if you check carefully the proof above, you see that if p is not of the form λ/p for some $\lambda > 0$, then in fact the number of events will either go to infinity or go to zero - i.e. to have a non-trivial random variable in the limit $n \to \infty$, we are forced to set $p = \lambda/n$.

2.1.3 Independence of random variables

Recall that we called two events E, F defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ independent if $\mathbb{P}(E)\mathbb{P}(F) = \mathbb{P}(E \cap F)$. Moreover, we called two sets of events $(E_i)_{i\in I}$ and $(F_j)_{j\in J}$ independent when for all $i \in I, j \in J$ we have that $\mathbb{P}(E_i \cap F_j) = \mathbb{P}(E_i)\mathbb{P}(E_j)$. It is easy to generalize that to the notion of mutual independence of a collection of sets of events.

This is the notion that gives rise to the definition of independence of random variables: indeed, each random variable X is being characterized by all events $\{X \in E\}$ for Borel sets E, and thus mutual independence of random variables is defined as mutual independence of these sets of events. More precisely,

Definition 2.11 (Mutually independent random variables)**.** *Let I be some countable index set and $(X_i)_{i\in I}$ a family of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that these random variables are mutually independent if for every finite set $J \subseteq I$ and all Borel measurable sets $(E_j)_{j\in J}$ we have that*

$$\mathbb{P}(\bigcap_{j\in J}\{X_j \in E_j\}) = \Pi_{j\in J}\mathbb{P}(X_j \in E_j).$$

A silly-sounding but very reasonable question to ask is the following: does there even exist a probability space with finite or with countably many independent random variables?

It comes out that we can answer this question positively using our set-up work from before - using the construction of product spaces and the relation between product spaces and independence.

Proposition 2.12. *Consider random variables $(X_i)_{i\geq 1}$. Then we can find a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $(\widetilde{X}_i)_{i\geq 1}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

- *For all $i \geq 1$, $\widetilde{X}_i$ and has the law of X_i*
- *Moreover, the random variables $(\widetilde{X}_i)_{i\geq 1}$ are mutually independent.*

Proof. Denote by $(F_{X_i})_{i\geq 1}$ the c.d.f-s corresponding to the random variables $(X_i)_{i\geq 1}$ and denote the by $\mathbb{P}_{X_i}$ the corresponding probability measures on $(\mathbb{R}, \mathcal{F}_E)$. Then by Theorem 1.42, we can construct the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$ that is the product probability measure of the spaces $((\mathbb{R}, \mathcal{F}_E, \mathbb{P}_{X_i}))_{i\geq 1}$.

For each $j \geq 1$, define $\widetilde{X}_j : (\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi, \mathbb{P}_\Pi) \to (\mathbb{R}, \mathcal{F}_E)$ via the projection map, i.e. we set $\widetilde{X}_j(\bar{x}) = x_j$. Recall that by Lemma 1.35 these maps are measurable and thus $\widetilde{X}_j$ is a random variable.

Moreover, by definitions, for each $E \in \mathcal{F}_E$

$$\mathbb{P}_{\widetilde{X}_i}(E) = \mathbb{P}_\Pi(\widetilde{X}_i \in E) = \mathbb{P}_\Pi(\mathbb{R} \times \cdots \times E \times \mathbb{R} \times \cdots \times \mathbb{R}) = \mathbb{P}_{X_i}(E),$$

where in the second product the event E is at the i-th co-ordinate. Thus the random variables X_i and $\widetilde{X}_i$ have the same law for all $i \geq 1$.

Finally, we need to check that the random variables $(\widetilde{X}_i)_{i\in I}$ are mutually independent on the space $(\mathbb{R}^{\mathbb{N}}, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$ according to the definition above. To see this, consider any finite $J \subseteq I$. Then by definition of the product measure and equality in law $X_j \sim \widetilde{X}_j$ we obtain that

$$\mathbb{P}_\Pi(\bigcap_{j\in J}\{\widetilde{X}_j \in E_j\}) = \Pi_{j\in J}\mathbb{P}_{X_j}(E_j) = \Pi_{j\in J}\mathbb{P}_{\widetilde{X}_j}(E_j),$$

and we conclude. □

Again, it actually suffices to check independence already for a smaller collection of events. In the following lemma we see that it suffices to only show that every collection of finitely many random variables are mutually independent, and moreover that we can restrict to nicer sets to check independence:

Lemma 2.13 (Equivalent statement of independence)**.** *Consider random variables $X_1, X_2, \ldots$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $X_1, X_2, \ldots$ are mutually independent if and only if*

(1) for every $m \geq 2$, the random variables $X_1, \ldots, X_m$ are mutually independent;

(2) for every $m \geq 2$ and all pairs $(a_j, b_j)_{j=1\ldots m}$ with $a_j < b_j$ we have that

$$\mathbb{P}(\bigcap_{1\leq j\leq m}\{X_j \in (a_j, b_j]\}) = \Pi_{1\leq j\leq m}\mathbb{P}(X_j \in (a_j, b_j]);$$

(3) for every $m \geq 2$ and all pairs $a_j \in \mathbb{R}$ we have that

$$\mathbb{P}(\bigcap_{1\leq j\leq m}\{X_j \leq a_j\}) = \Pi_{1\leq j\leq m}\mathbb{P}(X_j \leq a_j).$$

Proof. The proof is on the exercise sheet. □

The notion of independent random variables is very important and widely used - often also just because otherwise it is very difficult to do any calculations! Often one talks about a sequence of i.i.d. random variables $X_1, X_2, \dots$ - this means that $(X_i)_{i\geq 1}$ are mutually independent (first 'i') and all have the same probability law, i.e. are identically distributed (the 'i.d.'). Let us bring it even out as a definition:

Definition 2.14 (Independent identically distributed random variables)**.** *Let $X_1, X_2, \dots$ be random variables defined on a common probability space. We call $X_1, X_2, \dots$ i.i.d., i.e. independent and identically distributed if they are mutually independent and all have the same probability distribution.*

Intuitively, this corresponds to repeating the very same random situation or experiment over and over again.

Independence also helps us for example rewrite some properties of discrete random variables. For example, we can say that the sum of n independent $\{0,1\}$-valued $Ber(p)$ random variables has the law of the $Bin(n,p)$ random variable - we already saw this, but we carefully worded it using only independence of events. It also helps to take a more thorough look at Poisson random variables and the related Poisson point processes:

Exercise 2.3 (Poisson random variables)**.** *Let $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$ be two independent random variables defined on the same probability space.*

- *Prove that then $X_1 + X_2$ is also a Poisson random variable with parameter $\lambda_1 + \lambda_2$.*
- *Let now $Y_1, Y_2, \dots$ be independent $Ber(p)$ random variables defined on the same probability space. Prove that $X := \sum_{i=1}^{X_1} Y_i$ also has the law of $Poi(p\lambda)$ and $X_1 - X$ has the law of $Poi((1-p)\lambda)$ and is independent of X.*

Now, we consider what is called a Poisson point process on $\mathbb{N}$: This is a collection of i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$ where each $X_i \sim Poi(\lambda)$. For example you can think that some Newtonian apples fall on each integer. What is the law of the total number of apples on a finite set $S \subseteq \mathbb{N}$? Now colour every apple independently red with probability p and green with probability $1-p$ - i.e. every apple is ripe with probability p. Prove that restricting to only ripe / green apples also gives a Poisson point process on $\mathbb{N}$ and that moreover these processes are independent.

Finally, let i_1 be the first index of $\mathbb{N}$, which contains at least one apples, let i_2 be the second index that contains at least one apple etc. What is the distribution of the vector $(i_1, i_2 - i_1, i_3 - i_2, \dots)$?

2.1.4 Continuous random variables

Recall that we called a random variable X continuous if F_X was continuous, i.e. without any jumps. From Lemma 2.4 it follows that $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$. Most often continuous random variables arise via what is called a density function and this is also how we will usually construct them.

Definition 2.15 (Continuous r.v. with density)**.** *Let X be a random variable and $f_X : \mathbb{R} \to \mathbb{R}$ be a non-negative integrable function with $\int_{\mathbb{R}} f_X(x)dx = 1$. Then we say that a r.v. X has*

density f_X if for every $x \in \mathbb{R}$

$$F_X(t) = \int_{-\infty}^{t} f_X(x)dx.$$

Remark 2.16. *We remark straight away that there are also continuous random variables without a density (see starred section of the exercises).*

You might have already heard - and if not you will hear from me, and more next semester - that there are several notion of an integral. In particular, next to the Riemann integral stands the Lebesgue integral. So what do we mean by integrable?

We have seen that Riemann integral does not go well with measure theory - for example the set $\mathbb{Q}$ is a Borel set in $\mathbb{R}$, however $1_{\mathbb{Q}}$ is not Riemann-integrable. So it would be much more convenient to use the notion called the Lebesgue integral that you meet fully in Analysis IV and partly later on in this course. However, for now, it is really no restriction for us if:

- *for the sake of precision we just consider Riemann integrals*

In fact, all examples of densities we will see are Riemann integrable, so this is not a real restriction. Moreover, none of the results change become untrue when you come back and change Riemann integrals for Lebesgue integrals - in fact, as you will see next semester, for any function f that is Riemann integrable, its Lebesgue integral and Riemann integral agree.

Let us now look at the definition more closely. First, it is important to check the definition even makes sense, i.e. that the F_X defined actually is a cumulative c.d.f.:

Exercise 2.4. *Consider a non-negative Riemann integrable function f_X with $\int_{\mathbb{R}} f_X(x)dx = 1$. Define $F_X(x) := \int_{-\infty}^{x} f_X(x)dx$.*

- *Prove that F_X is a cumulative distribution function.*
- *Prove that if two random variables have the same density function, they have the same law*
- *Prove that given F_X, there is at most one continuous f_X such that $F_X(t) := \int_{-\infty}^{t} f_X(x)dx$.*
- *Give examples to show that f_X is however not uniquely defined by F_X.*

Further, let us look at an interpretation. Using Lemma 2.4 and the remark above that $\mathbb{P}(X = x) = 0$ for every $a < b$, we can also write

$$\mathbb{P}(X \in (a, b)) = \mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)dx.$$

it is important to notice that f_X does not give you the probability of $\{X = x\}$ at each point - we already saw that for continuous random variables this probability is 0 for all $x \in \mathbb{R}$. However, taking $b = a + \epsilon$, we can still obtain an interpretation of f_X, explaining why it is called the density function. Indeed, if for example f_X is continuous, we can write

$$\mathbb{P}(X \in (a, a + \epsilon)) = \int_a^{a+\epsilon} f_X(x)dx = \epsilon f_X(a) + o(\epsilon),$$

and thus one can think of $\epsilon f_X(a)$ as of the probability in being in the interval $(a, a + \epsilon)$. In particular, notice that $\epsilon^{-1}\mathbb{P}(X \in (a, a + \epsilon)) \to f_X(a)$ as $\epsilon \to 0$. This is of course related to the Fundamental theorem of calculus, which in the case of continuous f_X tells us that $F'_X(x) = f_X(x)$.

Let us now look at some examples. From the exercise above we see that to describe a continuous random variable with density it suffices to give the density function: an integrable non-.negative function with total integral 1.

Uniform random variable on $[a, b]$
A random variable U with density $f_U(x) = \frac{1}{b-a} 1_{[a,b]}$ is called a uniform random variable on the interval $[a, b]$ and is denoted sometimes $U = U_{[a,b]}$. We have already met the uniform random variable on $[0, 1]$ - as expected its law $\mathbb{P}_U$ is equal to the uniform / Lebesgue measure on $[0, 1]$, considered as a probability measure on $\mathbb{R}$. It's c.d.f is given by $F_U(x) = 1_{0 \leq x} \min x, 1$. You can also think of it as the limit of discrete uniform random variables taking values in $\{i/n : i = 1 \dots n\}$ - we will make this precise on the example sheet in some form, and then come back to it again later in the course.

Exponential random variable
Let $\lambda > 0$. The random variable X with density $f_X(x) = \lambda e^{-\lambda x} 1_{x \geq 0}$ is called the exponential random variable of parameter λ, and its law is denoted sometimes $Exp(\lambda)$. (We will check on the exercise sheet that the total mass is 1). In this case you can think of the exponential random variable as a continuous friend of the geometric random variable, as it also satisfies the memoryless property:

Exercise 2.5 (Exponential r.v. is the only memoryless random variable)**.** *We say that continuous a random variable X satisfying $\mathbb{P}(X > 0) = 1$ is memoryless if for every $x, y > 0$ we have that $\mathbb{P}_X(X > x + y | X > y) = \mathbb{P}_X(X > x)$. Prove that the exponential random variable is memoryless. Moreover, prove that every continuous memoryless random variable has the law of the exponential random variable.*

As geometric random variables, exponential random variables too are related to waiting times, just the underlying process is no longer in discrete time (like a sequence of tosses) but continuous time (like waiting for the next call from a friend). We will be able to make some more precise statements later in the course.

Gamma random variable
Let $\lambda > 0$ and $t > 0$. Denote by $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ the Euler gamma function. The random variable X with density

$$f_U(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x} 1_{x \geq 0}$$

is called a Gamma random variable of parameters λ and t. Again it needs to be checked that the total mass really is 1.

Notice that if we take $t = 1$, we have the exponential variable of parameter λ. Moreover, if we add up independent exponential random variables, we again obtain a Gamma random variable. This will be on the example sheet.

Maybe the most frequent Gamma random variable is the case $\lambda = 1/2$ and $t = d/2$, when one talks of a chi-square distribution of d parameters. This distribution will be important in statistics, and this is the main reason for introducing it here...though we will see more of it coming up very soon!

Gaussian random variable

Maybe the most important example of a random variable is that of a normal or Gaussian random variable. Given two parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}$, we say that N has the law of a normal random variable of mean μ and variance σ^2, denoted $N \sim \mathcal{N}(\mu, \sigma^2)$ if its density is given by

$$f_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

We call the law $\mathcal{N}(0,1)$ the standard normal random variable, or the standard Gaussian. Normal laws come up everywhere because of the so called Central limit theorem. A weak version of it could be vaguely stated as follows:

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables such that X_i has the same law as $-X_i$ and moreover, each X_i is bounded in the sense that there is some $C > 0$ with $\mathbb{P}(X_i < C) = 1$. Let $S_n = \sum_{i=1}^n X_i$. Then in the limit $n \to \infty$ we have that $\frac{S_n}{\sqrt{n}}$ becomes a normal random variable: for every interval (a, b), we have that $\mathbb{P}(\frac{S_n}{\sqrt{n}} \in (a, b)) \to \mathbb{P}(N \in (a, b)$, where N is a Gaussian random variable.

For example in physics experiments often we rarely expect to get the 'exact' value, but rather it comes with an error. This error is assumed to be a sum of many independent smaller errors, and thus, unless there is some bias that has not been accounted for, the observed values will have a normal distribution around the actual value.

We will prove a version of this theorem towards the end of the course, after having developed more tools to work with random variables. There is a first version of this in the starred section of the exercises.

It is common to mention here that although the normal random variable is the most used one, its cumulative distribution function - that has earned its own notation Φ_{μ,σ^2} - given as always by

$$\Phi_{\mu,\sigma^2}(x) = \mathbb{P}(N \leq t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx$$

does not admit an explicit formula. So in the old days one had to really check a long table with values to see give a numerical answer for, say, $\mathbb{P}(N > 12)$ or $\mathbb{P}(|N| < 200)$. I suspect there might be more modern ways now...

2.1.5 More random variables

Like we have seen before in the course - when we want to create more objects, one way is to start applying some operations to already existing objects. Here, this means operations on random variables.

Recall, that we have already seen that any continuous function from (X, τ_X) to (Y, τ_Y) is measurable, when we endow both spaces with their Borel σ-algebras - this is Proposition 1.20. This, together with the fact that the composition of measurable functions is measurable (check!) implies directly:

Lemma 2.17. *Let X be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then for any continuous real function $\phi : \mathbb{R} \to \mathbb{R}$, we have that $\phi(X)$ is also a random variable that can be defined on the same probability space.*

It is natural to ask whether the two classes of random variables - discrete and continuous - are stable under this operation. It comes out that this is always the case for discrete random variables, but not for the continuous random variables.

Exercise 2.6 (Functions of a random variable)**.** *Let X be a discrete random variable and $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous real function. Prove that $\phi(X)$ is also a discrete random variable. Is the image of a continuous random variable necessarily a continuous random variable?*

Still, in case of continuous random variables X, when g is nice enough, we do know that $g(X)$ is also continuous and we can even determine its density:

Proposition 2.18 (Density of the image)**.** *Let X be a continuous random variable with continuous density f_X that is non-zero only inside U. Let U, V be any open connected subsets of $\mathbb{R}$ and $\phi : U \to V$ bijective and continuously differentiable with ϕ' non-zero everywhere. Then $\phi(X)$ is also a continuous random variable with a continuous density $f_{\phi(X)}$ given by 0 outside of V and inside of V by:*

$$f_{\phi(X)}(x) = \frac{1}{|\phi'(\phi^{-1}(x))|} f_X(\phi^{-1}(x))$$

Proof. As ϕ is bijective, it is either strictly increasing or decreasing (why?). We look at the case when ϕ is increasing, the other case being analogous:

Notice that because ϕ is bijective and increasing, we have that $\mathbb{P}(\phi(X) \leq x) = \mathbb{P}(X \leq \phi^{-1}(x))$. And thus $F_{\phi(X)}(x) = F_X(\phi^{-1}(x))$.

Now as by assumption both F_X and ϕ^{-1} are continuous differentiable, we can apply the chain rule for $x \in V$ to deduce that

$$F'_{\phi(X)}(x) = (\phi^{-1})'(x) F'_X(\phi^{-1}(x)) = \frac{1}{|\phi'(\phi^{-1}(x))|} f_X(\phi^{-1}(x)).$$

As $F_{\phi(X)}(x) = 0$ for $x \leq \inf\{y \in U\}$ and $F_{\phi(X)}(x) \geq 1$ for $x \geq \sup\{y \in U\}$ we obtain the claim. Thus we see that defining $f_{\phi(X)}$ as stated indeed gives the right c.d.f $\square$

Remark 2.19. *It might be more illustrative for you to actually also do the previous proof more by hand: we already saw that in case of continuous density for every $x \in X$ it holds that $\mathbb{P}(X \in (x, x+\epsilon)) = \epsilon f_X(x) + o(\epsilon)$ and thus $\epsilon^{-1}\mathbb{P}(X \in (x, x+\epsilon)) \to f_X(x)$ as $\epsilon \to 0$. Now, by bijectivity of ϕ, we have $\mathbb{P}(\phi(X) \in (x, x+\epsilon)) = \mathbb{P}(X \in (\phi^{-1}(x), \phi^{-1}(x+\epsilon)))$. Use this to deduce the above formula.*

2.2 Random vectors

We already saw in the notes and on the example sheet that often several random variables come up in the same probabilistic situation and are naturally defined on the same probability space. So far we were looking mainly at their individual laws, or the situation when they were independent. But this is not always the case. When one starts being interested in the joint behaviour of several random variables, one often thinks in terms of random vectors:

Definition 2.20 (Random vectors and marginal laws)**.** *Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $(X_1, X_2, \ldots, X_n)$ is a random vector if and only if each of $X_1, X_2, \ldots, X_n$ is a random variable. The law $\mathbb{P}_{X_i}$ of each r.v. X_i is called its marginal law.*

Marginal laws are just the individual laws of random variables X_i that appear as components of a random vector and that we have been discussing so far. We know how to describe those. Yet they don't encode the relation between the random variables.

For example consider on the one hand (X_1, X_2), where both X_1 and X_2 encode independent fair coin tosses. On the other hand, consider $(X_1, \widetilde{X}_2)$, where X_1 is a fair coin toss, but $\widetilde{X}_2$ is heads when X_1 is tails and $\widetilde{X}_2$ is tails if X_1 is heads. Then the marginal laws of the vector (X_1, X_2) and $(X_1, \widetilde{X}_2)$ are the same (why?), yet they clearly describe very different situations!

So how can we mathematically encode this relation between the random variables? In fact, to look at joint laws, it is actually natural to look at $(X_1, \ldots, X_n)$ as a $\mathbb{R}^n$-valued random variable:

Lemma 2.21 (Joint law of random vectors)**.** *Let $\overline{X} = (X_1, \ldots, X_n)$ be a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then $(X_1, \ldots, X_n)$ as a vector is a $(\mathbb{R}^n, \mathcal{F}_E)$-valued random variable. In particular it induces a probability measure $\mathbb{P}_{\overline{X}}$ on $(\mathbb{R}_n, \mathcal{F}_E)$ called the joint law of the vector $\overline{X}$.*

In the other direction, any $(\mathbb{R}^n, \mathcal{F}_E)$-valued random variable gives rise to a random vector by the definition above.

The question here is measurability: does measurability of each component as a function $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F}_E)$ guarantee the measurability of the function $(\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{F}_E)$ and vice-versa. Thus the lemmat follows directly from a general result: [10]:

Lemma 2.22. *Let $(\Omega, \mathcal{F})$ and $((\Omega_i, \mathcal{F}_i))_{1 \leq i \leq n}$ be measurable spaces. Then the map $f : (\Omega, \mathcal{F}) \to (\Pi_{1 \leq i \leq n} \Omega_i, \mathcal{F}_\Pi)$ is measurable if and only if for every $i = 1 \ldots n$ the map $f_i = p_i \circ f$ mapping $(\Omega, F) \to (\Omega_i, \mathcal{F}_i)$ is measurable (here p_i is the projection map to the i-th coordinate).*

Proof. We saw that for the product $\sigma-$algebra that every projection map p_i is measurable. Thus, as the composition of measurable maps is measurable (check!) we obtain one direction: if $f : (\Omega, \mathcal{F}) \to (\Pi_{1 \leq i \leq n} \Omega_i, \mathcal{F}_\Pi)$ is measurable, then so are the components f_i.

In the opposite direction, recall that it suffices to show $f^{-1}(E)$ is measurable for a set of events E that generates the product sigma algebra. For example, the events of the form $E_i = F_1 \times \cdots \times F_n$ with $F_j \in \mathcal{F}_j$ generate the product sigma algebra. But now, $f^{-1}(E_i) = \bigcap_{i=1}^n f_i^{-1}(F_i)$. By assumption $f_i^{-1}(F_i) \in \mathcal{F}$ and thus the measurability of f follows. □

This is very useful, as now in order to describe random vectors, we can use the knowledge we have about probability measures on $\mathbb{R}^n$ - we know that they are characterised by the so called joint cumulative distribution functions. Thus Theorem 1.32 directly implies the following proposition:

Proposition 2.23 (Joint c.d.f.s of random vectors)**.** *Let $\overline{X} := (X_1, \ldots, X_n)$ be a random vector defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$F_{\overline{X}}(x_1, \ldots, x_n) := \mathbb{P}_{\overline{X}}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

[10] Notice the similarity to the following statement from topology: if $f_i : (X, \tau_X) \to (Y_i, \tau_{Y_i})$ are continuous, then so is $f : (X, \tau_X) \to (Y_1 \times \cdots \times Y_n, \tau_\Pi)$ given by $f = (f_1, \ldots, f_n)$.

gives rise to a joint cumulative distribution function. Moreover, any joint c.d.f. gives rise to a unique joint law of a random vector.

Again, random vectors give us mainly a clearer way of looking at things. We can for example now rephrase the last point of Lemma 2.13 as follows:

Lemma 2.24 (Independence using joint c.d.f.)**.** *Consider a random vector $\overline{X} = (X_1, \dots, X_n)$ defined on some probability space. Then $X_1, \dots, X_n$ are mutually independent if and only if $F_{\overline{X}}(x_1, \dots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$ for all $\overline{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$.*

Moreover, we can start doing arithmetic operations using random variables:

Lemma 2.25. *Let $\Phi : (\mathbb{R}^n, \tau_E) \to (\mathbb{R}^m, \tau_E)$ be any continuous function and $\overline{X}$ a random vector in $\mathbb{R}^n$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\Phi(\overline{X})$ is a random vector in $\mathbb{R}^m$, defined on the same probability space.*

Proof. This is on the exercise sheet.

□

In particular we have the following direct corollary:

Corollary 2.26. *Let $\overline{X}$ be a random vector in $\mathbb{R}^n$ and $\overline{a}$ any fixed vector in $\mathbb{R}^n$. Then $\sum_{i=1}^n a_i X_i$ is a random variable. Also $\Pi_{i=1}^n X_i$ is a random variable.*

I encourage you to even prove by hand that the sum of two random variables X_1 and X_2 is a random variable!

As in the case of usual random variables, one can also talk about discrete and continuous random vectors - in both cases what we have in mind is that all components are either discrete or continuous. But there could well also mixed cases. Here is a concrete example of a discrete random vector:

Multinomial random vector. Recall that the Binomial random variable $Bin(n, p)$ models the number of heads out of n independent tosses of a coin that comes up heads with probability p. As n is equal to the sum of heads and tails, it actually models both the number of heads and the number of tails. But suppose you want to model the random vector $(n_1, n_2, \dots, n_6)$ that gives you respectively the numbers of 1-s, 2-s etc of n independent dice throws? This is modelled by the so called multinomial random variable of parameters n, 6 and $p_1 = \cdots = p_6 = 1/6$.

The probability law of the multinomial random vector $\overline{M} \sim Mul(n, m, \overline{p})$ with parameters $n, m, \overline{p}$ is defined by

$$\mathbb{P}_{\overline{M}}(\overline{M} = (k_1, \dots, k_m) = \frac{n!}{k_1! \cdots k_m!} p_1^{k_1} \cdots p_m^{k_m},$$

whenever $\sum_{i=1}^m k_i = n$ and by $\mathbb{P}_{\overline{M}}(\overline{M} = (k_1, \dots, k_m) = 0$ otherwise. Notice that the marginal law on any coordinate i is given by the Binomial law $Bin(n, p_i)$.

As explained above, the multinomial random vector appears in the following situation: we consider a discrete random variable X taking values $x_1, \dots, x_m$ with probabilities $p_1, \dots, p_m$. And let $X_1, X_2, \dots, X_n$ be i.i.d. copies of X defined on some common probability space. Now define the random vector $\overline{M} = (M_1, \dots, M_n)$ as $M_j = \sum_{i=1}^n 1_{X_i = x_j}$. Then it is simple to check that each M_j is a random variable (in fact you have already proved this!) and thus $\overline{M}$ is a random vector. You can also verify that this random vector has the multinomial law.

2.2.1 Random vectors with denisty

Let us now consider the very special case of continuous vectors with density. This will be also a good source for more interesting examples.

Definition 2.27 (Random vectors with density)**.** *Let $\overline{X} = (X_1, \ldots, X_n)$ be a random vector and let $f_{\overline{X}}$ be a non-negative Riemann-integrable function from $\mathbb{R}^n \to [0, \infty)$ with total integral equal to 1. Then we say that $f_{\overline{X}}$ is the joint density of $\overline{X}$ if and only for any box $(a_1, b_1] \times \ldots (a_n, b_n]$*

$$\mathbb{P}_{\bar{X}}(X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]) = \int_{(a_1,b_1]\times\cdots\times(-a_n,b_n]} f_{\overline{X}}(\bar{x})d\bar{x}. \tag{2.1}$$

Remark 2.28. *Again, given the Lebesgue integral the natural statement would be that for every Borel measurable set E:*

$$\mathbb{P}(\overline{X} \in E) = \int_E f_{\overline{X}}(\bar{x})d\bar{x}.$$

In the case of Riemann integral the notion of integral might just not be defined on all such E.

Similarly to the 1d case, we also have the interpretation of this density as representing the probability of being in an infinitesimal neighbourhood around a point $\bar{t} = (t_1, \ldots, t_n)$. Indeed, if $f_{\overline{X}}$ is continuous, then you can check that we have

$$\mathbb{P}_{\overline{X}}((X_1, \ldots, X_n) \in (t_1, \ldots, t_n) + [-\epsilon/2, \epsilon/2]^n) = f_{\overline{X}}(t_1, \ldots, t_n)\epsilon^n + o(\epsilon^n). \tag{2.2}$$

Further, we can let $a_i \to -\infty$, for every $(t_1, \ldots, t_n) \in \mathbb{R}^n$ set

$$F_{\bar{X}}(t_1, \ldots, t_n) := \int_{(-\infty,t_1]\times\cdots\times(-\infty,t_n]} f_{\overline{X}}(\bar{x})d\bar{x}$$

and verify that this indeed gives rise to a c.d.f. Hence as joint c.d.f. characterise the joint law of random variables, can define laws of random vectors via their density function.

Finally, from the results in your course in Analysis II it then follows that if E is a subset of $\mathbb{R}^n$ such that 1_E is Riemann-measurable, then in fact:

$$\mathbb{P}(\overline{X} \in E) = \int_{\mathbb{R}^n} 1_E f_{\overline{X}}(\bar{x})d\bar{x}.$$

Notice that by the Fubini theorem for multiple Riemann-integrable functions, if the random vector admits a density, then also do its components:

Lemma 2.29 (Marginal densities)**.** *Let $\overline{X} = (X_1, \ldots, X_n)$ be a random vector with density $f_{\overline{X}}$ such that for every $I_0 \subseteq \{1, \ldots, n\}$ the function $f_{I_0^c}(\overline{x}')$ obtained by fixing all the co-ordinates in I_0 is Riemmann-integrable. Then the marginal laws $\mathbb{P}_{I_0}$ obtained by projecting on the co-ordinates contained in I_0 admits a density given by integrating out all the components in $\{1, \ldots, n\} \setminus I_0$.*

Remark 2.30. *Here we ask the condition that fixing any set of coordinate gives a Riemann-integrable function. This might be tiresome to check, but it is for example always true when f is continuous, or when f is piece-wise continuous with finite number of jumps along any co-ordinate – we call the latter just piece-wise continuous.*

Here are some quick examples of random vectors:

Uniform random vector on $[a,b]^n$. Similarly to a uniform random point on an interval, we can talk of a uniform random point $\overline{U} = (U_1, \ldots, U_n)$ in a rectangular box. To do this, we just define the density:

$$f_{\overline{U}}(x_1, \ldots, x_n) = \frac{1}{|b-a|^n} 1_{\overline{x} \in [a,b]^n}.$$

Notice that in this case the marginal laws U_i are just uniform random variables on $[a,b]$. Can you see why the variables $(U_1, \ldots, U_n)$ are mutually independent?

Gaussian random vector. Maybe the most important example here is that of the Gaussian (also called a normal) random vector $\mathcal{N}(\overline{\mu}, C)$, where $\overline{\mu}$ is a vector in $\mathbb{R}^n$ and C positive definite symmetric $n \times n$ matrix. We will call $\overline{\mu}$ the mean of the Gaussian vector, and the matrix C the covariance matrix – we will get to the reasons for this vocabulary in a few lectures time. The density of the Gaussian random vector is given by:

$$f_{\overline{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(C)}} \exp(-\frac{1}{2}(\overline{x} - \overline{\mu})^T C^{-1}(\overline{x} - \overline{\mu})).$$

When $\overline{\mu} = 0$ and C is the $n \times n$ identity matrix I_n, we call the law $\mathcal{N}(0, I_n)$ the standard Gaussian in $\mathbb{R}^n$. As you will see on the exercise sheet, all other Gaussian vectors in $\mathbb{R}^n$ are given by just linear transformations of the standard Gaussian.

As you will see on the example sheet, marginal laws of Gaussians are again Gaussian vectors. Moreover, all linear transformations of Gaussians still give Gaussians. However, to check this we will need to develop a bit more theory, e.g. about how the density changes under co-ordinate changes. On the other hand, let us stress that given two Gaussian random variables X_1, X_2, the random vector (X_1, X_2) is not necessarily a Gaussian random vector.

We already saw that transformations of random vectors remain random vectors. As in the 1D case, in the case of random vectors with density, we can again also determine the density. In this respect recall that for a diffeomorphism $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ one defines the differential $D\Phi$ as the $n \times n$ matrix $(D\Phi)_{ij} = \frac{\partial \Phi_i}{\partial x_j}$. The Jacobian is defined as the determinant of this matrix.

Proposition 2.31 (Density of the image of a random vector)**.** *Let $\overline{X}$ be a continuous random vector in $\mathbb{R}^n$ with density continuous $f_{\overline{X}}$ and $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ bijective and continuously differentiable with an everywhere non-zero Jacobian $J_\Phi(\overline{x}) = \det D\Phi_{\overline{x}}$, i.e. a C^1-diffeomorphism. Then $\Phi(\overline{X})$ is also a continuous random vector in $\mathbb{R}^n$ with a density $f_{\Phi(X)}$ given by:*

$$f_{\Phi(\overline{X})}(\overline{x}) = \frac{1}{|J_\Phi(\Phi^{-1}(\overline{x}))|} f_{\overline{X}}(\Phi^{-1}(\overline{x})).$$

The proof is basically the same as in the one-dimensional case.

Proof. Let E be a box. Then $1_{\phi^{-1}(E)}$ is Riemann-integrable by results from Analysis II. By using the fact that Φ is bijective

$$\mathbb{P}(\phi(\overline{X}) \in E) = \mathbb{P}(\overline{X} \in \Phi^{-1}(E)).$$

As $\overline{X}$ has density we can thus write

$$\mathbb{P}(\overline{X} \in \Phi^{-1}(E)) = \int_{\mathbb{R}^n} 1_{\Phi^{-1}(E)} f_{\overline{X}}(\bar{x}) d\bar{x}.$$

Now, we can use the multidimensional change-of-coordinates theorem of Analysis II for the transformation Φ^{-1} to write

$$\int_{\mathbb{R}^n} 1_{\Phi^{-1}(E)} f_{\overline{X}}(\bar{x}) d\bar{x} = \int_{\mathbb{R}^n} 1_E f_{\overline{X}}(\Phi^{-1}(\bar{x})) |J_{\Phi^{-1}}(\bar{x})| d\bar{x}$$

As $|J_{\Phi^{-1}}(\bar{x})| = \frac{1}{|J_\Phi(\Phi^{-1}(\bar{x}))|}$ we conclude.

□

Remark 2.32. *With slight modifications the same proof also works if $\Phi : U \to V$ is a C^1 diffeomorphism from a sufficiently nice open set U to sufficiently nice open set V, and $\mathbb{P}(\overline{X} \in U) = 1$. E.g. it suffices to assume both 1_U and 1_V are Riemann-integrable.*

Let us stress again that these statements would be nicer, more natural and more general if we had the notion of Lebsegue integral - we have already seen that the Riemann integral and Borel $\sigma-$algebra are not an ideal couple!

A nice application of this is determining the density of a sum of i.i.d. random variables:

Corollary 2.33. *Let X_1, X_2 be two independent continuous random variables with continuous densities f_{X_1} and f_{X_2}. Then their sum is a continuous random variable with density given by $f_{X_1+X_2}(y) = \int_{\mathbb{R}} f_{X_1}(x) f_{X_2}(y-x) dx$, i.e. by the convolution of the two densities.*

This definition of the density might look asymmetric, but you should check that it is not.

Proof. We use Proposition 2.31 with $\Phi(x, y) = (x, x + y)$. Indeed, this is an invertible linear map and thus a C^1 diffeomorphism from $\mathbb{R}^2 \to \mathbb{R}^2$. Moreover, its Jacobian $J = 1$. Thus by Proposition 2.31 the density of the vector $\Phi(X, Y)$ at s, t is given by:

$$f_{X_1, X_1+X_2}(x, y) = f_{X_1, X_2}(x, y - x).$$

But now X_1, X_2 are independent and hence we can further write this as $f_{X_1}(x) f_{X_2}(y - x)$. Finally, we notice that the law of $X_1 + X_2$ is the marginal law of $\Phi(X, Y)$ in the second coordinate. So we can use Lemma 2.29 to calculate this marginal density and obtain the desired formula. □

Let us look at a cute example:

- Consider two independent standard Gaussian random variables X_1, X_2. Then also $\frac{X_1+X_2}{\sqrt{2}}$ is a standard Gaussian random variable. Indeed, by the corollary above the density of $X_1 + X_2$ is given by $\frac{1}{2\pi} \int_{\mathbb{R}} e^{-x^2/2} e^{-(y-x)^2/2} dx$, which we can rewrite as

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-(x-y/2)^2} e^{-y^2/4} dx = \frac{e^{-y^2/4}}{\sqrt{4\pi}} \int_{\mathbb{R}} \frac{1}{\sqrt{\pi}} e^{-(x-y/2)^2} dx.$$

  But the last integral is just the total mass of a Gaussian $\mathcal{N}(y/2, 1/2)$ and thus equal to 1. Hence we recognize that $X_1 + X_2$ is a Gaussian $\mathcal{N}(0, 2)$. It is an easy check that then $\frac{X_1+X_2}{\sqrt{2}}$ is a standard Gaussian.

The joint density gives us moreover a new condition for checking mutual independence:

Exercise 2.7 (Independence using densities)**.** *Consider a random vector $\overline{X} = (X_1, \ldots, X_n)$ defined on some probability space. Suppose that $\overline{X} = (X_1, \ldots, X_n)$ admits a continuous density and all X_i admit a continuous density. Prove that $X_1, \ldots, X_n$ are mutually independent if and only if $f_{\overline{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$. What happens if the densities are piece-wise continuous with finitely many jumps?*

Deduce that for the uniform random vector $\overline{U} = (U_1, \ldots, U_n)$ on $[a, b]^n$ the components $U_1, \ldots, U_n$ are mutually independent. Moreover, deduce that if (X, Y) is a Gaussian random vector $\mathcal{N}(\overline{\mu}, C)$, then X and Y are independent Gaussians if and only if $C(1, 2) = 0$.

Remark 2.34. *In fact, the statement holds in more generality, however one needs care. Indeed, we saw that density functions are not uniquely defined - for example changing the value at a point does not change the density function. So a natural statement is actually asking for the equality only on some very large set, but we don't really have tools to deal with this setting at the moment. So for now, you can just assume that whenever the density of $f_{\overline{X}}$ is given by the product of f_{x_i} for all but countable number of points, we have independence; and on the other hand, if there is independence the joint density functions is equal to the product of the densities in terms that all integrals over boxes agree.*

2.2.2 Conditional laws

Given a random vector $(X_1, \ldots, X_n)$, we talked about the joint law that describes the probability measure induced on $\mathbb{R}^n$. We also discussed marginal laws, that give the individual laws of each component or a vector of components.

We now add to this list the conditional laws. Recall that given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and any event $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$, one could define the conditional probability measure on $(\Omega, \mathcal{F})$ by setting $\mathbb{P}(F|E) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)}$ for each $F \in \mathcal{F}$.

Given two random variables X_1, X_2 we will be interested in knowing the conditional law of X_1, given the value of X_2 – so we are just calculating conditional probability measures, with events E of the type $X_2 = x$. I will state the definition in a larger context and then come back to a simpler example.

Definition 2.35 (Conditional law for discrete random variables)**.** *Let $X_1, X_2, \ldots, X_n$ be discrete random variables on a common probability space. Write $\{1, \ldots, n\}$ as a union of two disjoint subsets I_0 and I_1. Now consider some fixed vector $(x_i)_{i \in I_1}$ with $\mathbb{P}((X_i = x_i)_{i \in I_1}) > 0$. Then the conditional law of $(X_i)_{i \in I_0}$ given $(X_i = x_i)_{i \in I_1}$ is given by*

$$\mathbb{P}((X_i = y_i)_{i \in I_0} | (X_i = x_i)_{i \in I_1}) := \frac{\mathbb{P}((X_i = y_i)_{i \in I_0} \cap (X_i = x_i)_{i \in I_1})}{\mathbb{P}((X_i = x_i)_{i \in I_1})}.$$

Let us write this out in the case of $n = 2$: then, assuming that $\mathbb{P}(X_2 = x_2) > 0$ the conditional law of X_1 given $X_2 = x_2$, is - as expected - described by giving for each x in the support of X_1, the conditional probability

$$\mathbb{P}(X_1 = x | X_2 = x_2) := \frac{\mathbb{P}(\{X_1 = x\} \cap \{X_2 = x_2\})}{\mathbb{P}(X_2 = x_2)}.$$

Now continuous random variables take any value with zero probability, so this wouldn't work directly. And as you will see on the exercise sheet, conditioning on events of zero probability is tricky. So we cannot just blindly reuse the definition of the conditional probabilities. Yet, for variables with a nice density one can give sense to conditional laws via densities.

As the general version might be a bit harder to parse, let us start from a simple version

Definition 2.36 (Conditional law for continuous random variables with density (simple))**.** *Let $\overline{X} = (X_1, X_2)$ be random vector with a continuous joint density. Let y be such that the marginal density of X_2 is positive: $f_{X_2}(y) > 0$. Then the conditional law of X_1, given $X_2 = y$ is defined to be the continuous r.v. with the following density:*

$$f_{X_1|X_2=y}(x) := \frac{f_{X_1,X_2}(x,y)}{f_{X_2}(y)}.$$

It requires a check that the conditional density is indeed a density, but I leave this to you. As a philosophy - although densities are not like probabilities, one can sometimes use them in similar roles. Let me now state a general version of the definition, where one can condition on a part of the vector.

Definition 2.37 (Conditional law for continuous random variables with density (general))**.** *Let $\overline{X} = (X_1, X_2, \dots, X_n)$ be random vector with a continuous joint density. Write $\{1, \dots, n\}$ as a union of two disjoint subsets I_0 and I_1 and write $\overline{X}_{I_0}$ and $\overline{X}_{I_1}$ for the corresponding random vectors. Now consider some fixed vector $\overline{x}$ such that the marginal density at $\overline{x}_{I_1}$ is positive, i.e. $f_{\overline{X}_{I_1}}(\overline{x}_{I_1}) > 0$. Then the conditional density of $\overline{X}_{I_0}$ given $\overline{X}_{I_1} = \overline{x}_{I_1}$ is defined by*

$$f_{\overline{X}_{I_0}|\overline{X}_{I_1}=\overline{x}_{I_1}}(\overline{x}_{I_0}) := \frac{f_{\overline{X}}(\overline{x})}{f_{\overline{X}_{I_1}}(\overline{x}_{I_1})}.$$

As above, it is an easy check that this does actually define a density. As with conditional probabilities in general, conditional laws are usually notoriously difficult to calculate and might be very different from the initial law.

However, there is one case, where things are nice again - this is Gaussian vectors. Although this holds in a large generality and could even be proved with the methods we already have, we restrict ourselves here to the 2-dimensional case. We will come back to the general case, once we have some more elegant and efficient tools at hand.

Lemma 2.38 (Conditional laws for Gaussians in 2D)**.** *Let (X, Y) be a Gaussian random vector $\mathcal{N}(\mu, C)$. Then the conditional law of Y, given $X = x$ for any $x \in \mathbb{R}$ is also Gaussian, similarly if we switch the roles of X, Y.*

Proof. We will assume that $\overline{\mu} = 0$. It is a small check then to deduce the general case.

The trick for calculating the conditional laws is exactly the same as for marginal laws, that you saw on the example sheet: we complete the square to recognize a Gaussian distribution. To do this write $\widehat{C} = C^{-1}$ and proceed as follows. First, from the Exercise sheet we know that

$$f_X(x) = \frac{\sqrt{|C_{22}|}}{\sqrt{2\pi}\sqrt{\det C}} \exp(-\frac{1}{2}(\widehat{C}_{11} - \frac{\widehat{C}_{12}^2}{\widehat{C}_{22}})x^2).$$

Hence, from the definition of the conditional density it follows that:

$$f_{Y|X=x}(y) = \frac{\frac{1}{2\pi\sqrt{\det C}} \exp(-\frac{1}{2}(\widehat{C}_{11}x^2 + \widehat{C}_{22}y^2 + 2\widehat{C}_{12}xy))}{\frac{\sqrt{|C_{22}|}}{\sqrt{2\pi}\sqrt{\det C}} \exp(-\frac{1}{2}(\widehat{C}_{11} - \frac{\widehat{C}_{12}^2}{\widehat{C}_{22}})x^2)}.$$

We rewrite

$$\widehat{C}_{11}x^2 + \widehat{C}_{22}y^2 + 2\widehat{C}_{12}xy = \left(\widehat{C}_{22}^{1/2}y + \frac{\widehat{C}_{12}}{\widehat{C}_{22}^{1/2}}x\right)^2 + (\widehat{C}_{11} - \frac{\widehat{C}_{12}^2}{\widehat{C}_{22}})x^2.$$

Thus cancelling out terms gives

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi|C_{22}|}} \exp\left(-\frac{1}{2}(\widehat{C}_{22}^{1/2}y + \frac{\widehat{C}_{12}}{\widehat{C}_{22}^{1/2}}x)^2\right)$$

and we recognize a Gaussian distribution with $\sigma^2 = |C_{22}|$ and $\mu = -\frac{\widehat{C}_{12}}{\widehat{C}_{22}^{1/2}}x$. In other words, the conditional law is shifted version of the marginal law.

□

Section 3

Mathematical expectation

We will continue working with random variables and start looking at several different characteristics or properties of their law, based on the concept of mathematical expectation. Mathematical expectation, or just 'expectation', or 'expected value', or 'mean' is a fancy name for taking the average in context of probability measures. Its introduction in the early times of probability was roughly motivated by a very simple question:

- Suppose you are offered the following deal - a dice is thrown and you get as many francs as many dots come up on the top of the dice; but you have to pay n francs independently of the result in return. How many francs should you agree to pay?

Whereas what is really the 'right' answer still depends on some further conditions and assumptions. However, the following vaguely stated mathematical result gives some insight into the problem (and was used in these old times of gambling!):

- Let $X_1, X_2, \dots$ be independent random dice throws. Let $S_n = \sum_{i=1}^{n} X_i$. Then in the limit $n \to \infty$ we have that $\frac{S_n}{n}$ converges to $\frac{1+2+3+4+5+6}{6} = 3.5$.

This result is a specific case of the so called law of large numbers, and it tells you that the average gain from one dice throw is 3.5. So would this mean that you should offer anything below 3.5 francs? While pondering on this worldly problem, let us dig into the mathematical theory.

3.1 Expected value of a discrete random variable

We start with the discrete case to lay clear foundations. The continuous case can be seen as an extension of this:

Definition 3.1 (Expected value of a discrete random variable)**.** *Let X be a discrete random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with support S. We say that X admits an expected value or that X is integrable if $\sum_{x \in S} |x| \mathbb{P}(X = x) < \infty$.*

For an integrable random variable X, the expected value of X, denoted $\mathbb{E}(X)$ is defined as

$$\mathbb{E}(X) = \sum_{x \in S} x \mathbb{P}(X = x).$$

Remark 3.2. *Observe the following*

- *The condition for integrability is there of absolute summability - otherwise the order in the sum would matter, and there would be no unique answer to the expectation. We have that X is integrable if $|X|$ is.*
- *The expectation only depends on the law $\mathbb{P}_X$ of the random variable and not the probability space on the background.*
- *Discrete random variables with finite support are always integrable.*

Before proving some properties that make the expected value extremely useful, let us look at some examples:

Deterministic random variable

If a random variable X takes some value $x \in \mathbb{R}$ with probability 1, then its expectation is

also clearly equal to x

Bernoulli random variable
Let E be an event on a probability space, and consider the random variable 1_E. As its support is finite, it is integrable. From the definition of expectation, we directly have that $\mathbb{E}(1_E) = \mathbb{P}(E)$. Thus in particular if X is a $Ber(p)$ random variable, then its expectation is just $\mathbb{E}(X) = p$.

Uniform random variable
Consider the uniform random variable U_n on $\{1, 2, \dots, n\}$. Again as it takes only finitely many values, it is integrable. Its expected value is

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

Poisson random variable
Consider the Poisson random variable P of parameter $\lambda > 0$. The support of a Poisson random variable is not finite and thus one needs to verify that it is integrable. But in fact, the same computation also gives the expectation:

$$\mathbb{E}(P) = \sum_{n\geq 0} n\mathbb{P}(P = n) = \sum_{n\geq 1} n\frac{e^{-\lambda}\lambda^n}{n!} = \lambda e^{-\lambda}\sum_{m\geq 0}\frac{\lambda^m}{m!} = \lambda.$$

Hence, even if a random variable can take arbitrary large values, its expectation can be finite. This is, however, not always the case. For example

- Consider a random variable X such that it takes value 2^n with probability 2^{-n}. Then clearly $\mathbb{E}(X) = \infty$ and X is not integrable.

If a random variable is non-negative, then its expected value doesn't exist only if it is too large, i.e. is infinite. Sometimes one still defines expected value for any positive random variable, just saying that $\mathbb{E}(X) = \infty$, in case it is infinite.

You will see more examples on the exercise sheet:

Exercise 3.1 (Expectations of discrete random variables)**.** *Prove that the expected value of a Binomial random variable $Bin(n, p)$ is equal to np. Prove also that the expected value of a geometric random variable of parameter p is equal to $1/p$.*

For now, let us verify some nice conditions of the expectation. We will use the following notation: if X, Y are random variables, we write $X \geq Y$ to say that the event $X \geq Y$ happens with probability 1.

Proposition 3.3. *Let X, Y be two integrable discrete random variables defined on the same probability space. Then the expected value satisfies the following properties:*

- *It is linear: we have that $\mathbb{E}(\lambda X) = \lambda\mathbb{E}(X)$ for all $\lambda \in \mathbb{R}$. Further, $X+Y$ is integrable and $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.*
- *If $X \geq 0$ i.e. $\mathbb{P}(X \geq 0) = 1$, then $\mathbb{E}(X) \geq 0$,*
- *If $X \geq Y$ i.e. $\mathbb{P}(X \geq Y) = 1$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$. Deduce that if $\mathbb{P}(c \leq X \leq C) = 1$, then $c \leq \mathbb{E}(X) \leq C$.*
- *We have that $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.*

Proof. The fact that $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ follows directly from the definition. Let us next prove that $X + Y$ is integrable and $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$. Denote by S_X, S_Y the supports of X and Y respectively. Denote by S_{X+Y} the support of $X + Y$. Notice that

$$\mathbb{P}(X + Y = s) = \sum_{x \in S_X} \sum_{y \in S_Y} \mathbb{P}(X = x, Y = y) 1_{x+y=s}$$

Thus we can write

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X + Y = s) = \sum_{s \in S_{X+Y}} \sum_{x \in S_X} \sum_{y \in S_Y} |x + y| \mathbb{P}(X = x, Y = y) 1_{x+y=s}.$$

By triangle inequality we can bound $|x + y| \leq |x| + |y|$ and thus obtain

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X + Y = s) \leq \sum_{s \in S_{X+Y}} \sum_{x \in S_X} \sum_{y \in S_Y} (|x| + |y|) \mathbb{P}(X = x, Y = y) 1_{x+y=s}. \tag{3.1}$$

Now, observe that for fixed x and y either $\mathbb{P}(X = x, Y = y) = 0$ or $x + y \in S_{X+Y}$ and we have that

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x, Y = y) \sum_{s \in S_{X+Y}} 1_{x+y=s}.$$

Moreover, for fixed x by the law of total probability we have that

$$\sum_{y \in S_Y} \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x).$$

Thus as everything in Equation (3.1) is positive, we can now switch the order of summation, and to recognize the RHS as a sum of

$$\sum_{x \in S_X} \sum_{y \in S_Y} \sum_{s \in S_{X+Y}} |x| \mathbb{P}(X = x, Y = y) 1_{x+y=s} = \sum_{x \in S_X} |x| \mathbb{P}(X = x)$$

and

$$\sum_{y \in S_Y} \sum_{x \in S_X} \sum_{s \in S_{X+Y}} |y| \mathbb{P}(X = x, Y = y) 1_{x+y=s} = \sum_{y \in S_Y} |y| \mathbb{P}(Y = y).$$

Hence we bound

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X + Y = s) \leq \sum_{x \in S_X} |x| \mathbb{P}(X = x) + \sum_{y \in S_Y} |y| \mathbb{P}(Y = y)$$

and deduce integrability. Thereafter, the same way of separating sums also gives that $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

For the second bullet point, we notice that if $X \geq 0$ with full probability, then for every $s \in S_X$, we have that $s \geq 0$. Thus it follows from definition of expectation that $\mathbb{E}(X) \geq 0$.

For the third bullet point, notice that by the condition $X - Y \geq 0$. Thus $X - Y \geq 0$ with full probability, and hence by the second bullet point $\mathbb{E}(X - Y) \geq 0$. The first bullet point then gives that $\mathbb{E}(X) \geq \mathbb{E}(Y)$. Plugging in $Y = c$ in this inequality, and noticing that $\mathbb{E}c = c$, gives $\mathbb{E}(X) \geq c$. The other inequality follows similarly.

Finally, for the fourth bullet point notice that $-\mathbb{E}(X) = \mathbb{E}(-X)$ by the first point. Hence it suffices to show that $\mathbb{E}(X) \leq \mathbb{E}|X|$. But this just follows from the definition, as $\mathbb{P}(X = x)$

is always positive for $x \in S_X$ and hence

$$\mathbb{E}(X) = \sum_{x \in S_X} x\mathbb{P}(X = x) \leq \sum_{x \in S_X} |x|\mathbb{P}(X = x) = \mathbb{E}(|X|),$$

where in the last equality we use that $\mathbb{P}(|X| = |x|) = \mathbb{P}(X = x) + \mathbb{P}(X = -x)$ and the fact that $|x| \in |S_X|$ if and only if either $x \in S_X$ or $-x \in S_X$. □

3.2 Expected value of an arbitrary random variable

The idea for defining the expectation of a general random variable X is to approximate it by discrete random variables. More precisely, given X, we define the discretizations of X as:

$$X_n = 2^{-n}\lfloor 2^n X \rfloor = \sum_{k \in \mathbb{Z}} k2^{-n} 1_{X \in [k2^{-n}, (k+1)2^{-n})}.$$

Notice that X_n is indeed a discrete random variable - it is a non-decreasing function of X, so it is a random variable, and it takes only countably many values, thus it is discrete. The following exercise says that these discretizations really approximate the initial random variable very well.

Exercise 3.2 (Discretizations are nice)**.** *Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. and $(X_n)_{n \geq 1}$ be the discretizations $X_n = 2^{-n}\lfloor 2^n X \rfloor = \sum_{k \in \mathbb{Z}} k2^{-n} 1_{X \in [k2^{-n}, (k+1)2^{-n})}$.*

Prove that for every $\omega \in \Omega$, we have that $X_n(\omega) \leq X(\omega) \leq X_n(\omega) + 2^{-n}$ and thus the sequence $(X_n(\omega))_{n \geq 1}$ converges to $X(\omega)$.

We can now use the definition of the expectation $\mathbb{E}(X)$ for discrete random variables X to define expected value of an arbitrary random variable:

Proposition 3.4 (Expected value of a random variable)**.** *Let X be a random variable defined on some probability space. If $\mathbb{E}(|X_1|) < \infty$, then $\mathbb{E}(|X_n|) < C$ for some constant C and we call X integrable. The expected value of X is then defined as*

$$\mathbb{E}(X) = \lim_{n \to \infty} \mathbb{E}(X_n).$$

Remark 3.5. *Notice that X is integrable if and only if $|X|$ is integrable. It is important to verify that a random variable is integrable before calculating the expectation. We will see below that for example bounded random variables are automatically integrable.*

Remark 3.6. *Also, observe again that the expectation only depends on the law of X and not on the underlying probability space (why?).*

The idea for proving this proposition is just to show that the sequence $\mathbb{E}(X_n)$ is Cauchy.

Proof. Notice that from the Exercise 3.2 above we see that $X_1 - 1 \leq X_n \leq X_1 + 1$ and hence $|X_n| \leq |X_1| + 1$. Thus if $\mathbb{E}(|X_1|) < C - 1$, then from Proposition 3.3 it follows that $\mathbb{E}(|X_n|) < C$ for all $n \geq 1$. It follows that X_n is integrable for every $n \geq 1$ and hence $\mathbb{E}(X_n)$ well-defined.

We now claim that $\mathbb{E}(X_n)$ is a Cauchy sequence. So consider $m \geq n$. Then from Proposition 3.3 it follows that

$$|\mathbb{E}(X_n) - \mathbb{E}(X_m)| = |\mathbb{E}(X_n - X_m)| \leq \mathbb{E}(|X_n - X_m|).$$

But we can bound $|X_n - X_m| \leq 2^{-n}$ using Exercise 3.2. Hence $|\mathbb{E}(X_n) - \mathbb{E}(X_m)| \leq \mathbb{E}(2^{-n}) = 2^{-n}$. It follows that the sequence $(\mathbb{E}(X_n))_{n\geq 1}$ is Cauchy and thus converges to a unique limit as $n \to \infty$. □

An easy but important sanity check is that this definition indeed agrees with the previous definition for discrete random variables, i.e. that the Definition 3.1 of $E(X)$ and the definition of $E(X)$ by Proposition 3.4 agree for any discrete random variable X. Further, one can also check that all the properties that hold for the expectation of the discrete random variable, also hold for the expectation in general:

Proposition 3.7. *Let X, Y be two integrable random variables defined on the same probability space. Then the expected value satisfies the following properties:*

- *It is linear: we have that $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ for all $\lambda \in \mathbb{R}$. Further, $X+Y$ is integrable and $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.*
- *If $X \geq 0$ i.e. $\mathbb{P}(X \geq 0) = 1$, then $\mathbb{E}(X) \geq 0$,*
- *If $X \geq Y$ i.e. $\mathbb{P}(X \geq Y) = 1$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$. Deduce that if $\mathbb{P}(c \leq X \leq C) = 1$, then $c \leq \mathbb{E}(X) \leq C$.*
- *We have that $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.*

Proof. All these points follow from Proposition 3.3 via discretizations and Exercise 3.2. This is a somewhat tedious verification that I leave for you.

For example, as for all n , we have that $X_n + 2^{-n} \geq X$, then $X \geq 0$ means that $X_n \geq -2^{-n}$. It follows from Proposition 3.7 that $\mathbb{E}(X_n) \geq -2^{-n}$, implying that for every $\epsilon > 0$, for all n large enough $\mathbb{E}(X_n) \geq -\epsilon$ and hence $\mathbb{E}(X) \geq 0$. □

Let us now see that in the case of random variables with density, we can use Riemann integration and the density to calculate expectation.

Proposition 3.8 (Expected value for r.v. with density)**.** *Let X be a random variable with density f_X. Then X is integrable iff $\int_{\mathbb{R}} |x| f_X(x) dx < \infty$ and we have*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x) dx.$$

Proof. Consider the discretizations $X_n = 2^{-n} \lfloor 2^n X \rfloor$. Notice that

$$\mathbb{P}(X_n \in [k2^{-n}, (k+1)2^{-n})) = \int_{k2^{-n}}^{(k+1)2^{-n}} f_X(x) dx$$

and hence

$$\mathbb{E}(|X_1|) = \sum_{k \geq 0} k2^{-1} \left(\int_{k2^{-1}}^{(k+1)2^{-1}} f_X(x) dx + \int_{-k2^{-1}}^{(-k+1)2^{-1}} f_X(x) dx \right).$$

Now, if $|x| \in [k2^{-1}, (k+1)2^{-1})$ then $k2^{-1} \leq |x| \leq k2^{-1} + 2^{-1}$. Using the fact that $\int_{\mathbb{R}} f_X(x) dx = 1$ and that f_X is non-negative, we conclude that

$$-1 + \int_{\mathbb{R}} |x| f_X(x) dx \leq \mathbb{E}(|X_1|) \leq 1 + \int_{\mathbb{R}} |x| f_X(x) dx.$$

Thus X is integrable iff $\int_{\mathbb{R}} |x| f_X(x) dx < \infty$.

Next, as

$$\mathbb{E}(X_n) = \sum_{k\in\mathbb{Z}} k2^{-n} \int_{k2^{-n}}^{(k+1)2^{-n}} f_X(x)dx,$$

we see similarly to above that also

$$\mathbb{E}(X_n) \leq \int_{\mathbb{R}} x f_X(x)dx \leq \mathbb{E}(X_n) + 2^{-n}.$$

But $\mathbb{E}(X_n) \to \mathbb{E}(X)$ as $n \to \infty$, and hence the proposition now follows by taking $n \to \infty$. $\square$

Let us calculate densities for some known random variables:

Uniform random variable on $[a, b]$
Consider a uniform random variable U on $[a, b]$. Recall its density is given by $f_U(x) = (b-a)^{-1}1_{x\in[a,b]}$. First notice that U is bounded and hence integrable. Thus we calculate:

$$\mathbb{E}(U) = (b-a)^{-1} \int_{\mathbb{R}} x1_{x\in[a,b]}dx = (b-a)^{-1} \int_a^b xdx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Gaussian random variable
Consider a standard normal random variable $N \sim \mathcal{N}(0, 1)$. We first note that

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |x| \exp(-\frac{x^2}{2})dx = \frac{2}{\sqrt{2\pi}} \int_0^\infty x \exp(-\frac{x^2}{2})dx = \frac{2}{\sqrt{2\pi}} < \infty.$$

Thus N is integrable. We further notice that

$$\mathbb{E}(N) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x \exp(-\frac{x^2}{2})dx = \mathbb{E}(-N),$$

as the density of $-N$ is the same as that of N. Hence Proposition 3.7 implies that $\mathbb{E}(N) = 0$.

Now, consider a general Gaussian random variable $N_{\mu,\sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$. Recall that we can write $N_{\mu,\sigma^2} \sim \sigma N + \mu$ and hence N_{μ,σ^2} is integrable. Further, we can use Proposition 3.7 one more time to deduce that $\mathbb{E}N_{\mu,\sigma^2} = \sigma\mathbb{E}(N) + \mu = \mu$. This is the reason why μ is called the mean of the Gaussian random variable.

Again, further examples are on the exercise sheet.

3.3 Expected value of a function of a random variable

It comes out that the expected value, even if just a number, is very useful tool to describe a random variable. Often we might not be interested in the expectation of some given random variables, but of certain functions of these random variables. For example, given a r.v. X we might be interested in $\mathbb{E}\left((X - \mathbb{E}X)^2\right)$, or given X, Y, we might be interested in $\mathbb{E}XY$. In fact, as we will see, if we know $\mathbb{E}g(X)$ for sufficiently many functions g, then this determines the random variable itself!

To start, let us look at the following proposition telling us that sometimes there is a nice way to calculate expectations of functions of a r.v.:

Proposition 3.9. *Let $\overline{X} = (X_1, \ldots, X_n)$ be a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and ϕ a measurable function from $(\mathbb{R}^n, \mathcal{F}_E)$ to $(\mathbb{R}, \mathcal{F}_E)$, so that $\phi(\overline{X})$ is a random variable.*

- *If all $X_1, \dots, X_n$ are discrete and $\phi(\overline{X})$ is integrable, then*
$$\mathbb{E}(\phi(\overline{X})) = \sum_{\overline{x} \in S_{\overline{X}}} \phi(\overline{x}) \mathbb{P}(\overline{X} = \overline{x}),$$
where $S_{\overline{X}} \subseteq \mathbb{R}^n$ is the support of the random vector $\overline{X}$, i.e. the set of $\overline{s} = (s_1, \dots, s_n) \in \mathbb{R}^n$ such that $\mathbb{P}(\overline{X} = \overline{s}) > 0$ for all $\overline{x} \in S_{\overline{X}}$ and $\mathbb{P}(\overline{X} \in S_{\overline{X}}) = 1$.
- *If $\overline{X}$ is a random vector with density, $\phi(X)$ an integrable random variable and ϕ sufficiently nice - meaning that $\phi^{-1}([a,b))$ is Riemann measurable for any interval $[a,b)$ - then*
$$\mathbb{E}(\phi(\overline{X})) = \int_{\mathbb{R}^n} \phi(\overline{x}) f_{\overline{X}}(\overline{x}) d\bar{x}.$$

The condition 'sufficiently nice' is of course not quite natural. This is yet again due to the fact that Riemann integration and measurability in the sense of Borel (or Lebesgue) do not play together in full harmony. After Analysis IV next semester, you should be able to revisit many of these results and restate them in more natural ways, if interested of course. Still, notice that the condition holds for many natural functions like x^n or $\exp(x)$.

Proof. The discrete case is on the exercise sheet.

To prove the second case, we use discretizations - we set $\phi_n(\overline{x}) = 2^{-n} \lfloor \phi(\overline{x}) 2^n \rfloor$. Then - given integrability - we have that
$$\mathbb{E}(\phi_n(\overline{X})) = \sum_{k \in \mathbb{Z}} k 2^{-n} \mathbb{P}(\phi_n(\overline{X}) = k 2^{-n}).$$
Now, given that $\phi^{-1}([a,b))$ are Riemann-measurable, we can write
$$k 2^{-n} \mathbb{P}(\phi_n(\overline{X}) = k 2^{-n}) = \int_{\mathbb{R}^n} 1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))} k 2^{-n} f_{\overline{X}}(\overline{x}) d\bar{x}.$$
Again by absolute summability [11] we can switch the order of sum and integration to get
$$\mathbb{E}(\phi_n(\overline{X})) = \int_{\mathbb{R}^n} f_{\overline{X}}(\overline{x}) \sum_{k \in \mathbb{Z}} 1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))} k 2^{-n} d\bar{x}.$$
As above, for any fixed $\overline{x}$, we have that $1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))}$ is equal to 1 for only one value of k and thus from the definition of ϕ_n, we obtain
$$\sum_{k \in \mathbb{Z}} 1_{\overline{x} \in \phi^{-1}([k2^{-n}, (k+1)2^{-n}))} k 2^{-n} = \phi_n(\overline{x}).$$
Hence
$$\mathbb{E}(\phi_n(\overline{X})) = \int_{\mathbb{R}^n} \phi_n(\overline{x}) f_{\overline{X}}(\overline{x}) d\bar{x}.$$
We can now conclude similarly to Proposition 3.8. □

Looking at expectations of functions of a random variable turns out to be a powerful thing:

[11] More precisely, we are using there that if either $\sum_{n \geq 1} \int_{\mathbb{R}} |f_n(x)| dx < \infty$ or $\int_{\mathbb{R}} \sum_{n \geq 1} |f_n(x)| dx < \infty$, then $\int_{\mathbb{R}} \sum_{n \geq 1} f_n(x) dx = \sum_{n \geq 1} \int_{\mathbb{R}} f_n(x) dx$. You have met the analogous result for swapping two sums $\sum_{k \geq 1} \sum_{n \geq 1}$, and the proof is basically the same.

Proposition 3.10. *Let X, Y be two random variables. Then X and Y are equal in law if and only if for all bounded continuous functions $g : \mathbb{R} \to \mathbb{R}$ we have that $\mathbb{E}g(X) = \mathbb{E}g(Y)$.*

Proof. If X and Y have the same law, then also do $g(X)$ and $g(Y)$ for any continuous and bounded g. Hence, as bounded functions are integrable and the expectation only depends on the law of the r.v., we indeed have that $\mathbb{E}g(X) = \mathbb{E}g(Y)$.

In the other our aim is to show that $\forall t \in \mathbb{R}$, $F_X(t) = F_Y(t)$. To do this recall that $F_X(t) = \mathbb{P}(X \leq t) = \mathbb{E}(1_{x \leq t})$, so our aim will be to consider continuous approximations $g_{t,n}$ of the indicator function $1_{x \leq t}$, defined as follows. Fix some $t \in \mathbb{R}$ and set $g_{t,n}(x) = 1$ if $x \leq t$, we set $g_{t,n}(x) = 0$ if $x \geq t + 2^{-n}$ and we set $g_{t,n}(x) = 1 - 2^n(x - t)$ inside the interval $(t, t + 2^{-n})$.

Then, one the one hand

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{E}(1_{x \leq t}) \leq \mathbb{E}(g_{t,n}(X))$$

and on the other hand

$$\mathbb{E}(g_{t,n}(X)) \leq \mathbb{E}(1_{x \leq t+2^{-n}}) = \mathbb{P}(X \leq t + 2^{-n}) = F_X(t + 2^{-n}).$$

Thus by right-continuity of $F_X(t)$ we see that $\mathbb{E}(g_{t,n}(X))$ converges to $F_X(t)$ as $n \to \infty$. But similarly also $\mathbb{E}(g_{t,n}(Y))$ converges to $F_Y(t)$ as $n \to \infty$. As by assumption $\mathbb{E}(g_{t,n}(X)) = \mathbb{E}(g_{t,n}(Y))$, we can conclude the proposition. □

Also independence can be restated in an elegant way using expectations - X, Y are independent if the expectation factorizes for all continuous functions!

Proposition 3.11. *Let X, Y be two random variables. Then*

- *If for all $g : \mathbb{R} \to \mathbb{R}, h : \mathbb{R} \to \mathbb{R}$ continuous and bounded we have that*

$$\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}g(X)\mathbb{E}h(Y), \tag{3.2}$$

  then X and Y are independent.
- *On the other hand, if X and Y are independent, then for all measurable functions $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ and $h(Y)$ are integrable the Equation (3.2) holds.*

Proof. From Lemma 2.24 we know that to prove X, Y are independent, it suffices to prove that for all $s, t \in \mathbb{R}$ we have that $F_{(X,Y)}(s, t) = F_X(s)F_X(t)$. Further, recall that $F_{(X,Y)}(s, t) = \mathbb{E}1_{X \leq s, Y \leq t} = \mathbb{E}1_{X \leq s}1_{Y \leq t}$. We follow the strategy of Proposition 3.10. Indeed, consider the same continuous functions $g_{t,n}(x)$ satisfying $1_{x \leq t} \leq g_{t,n}(x) \leq 1_{x \leq t+2^{-n}}$.

Using the expression of $F_{(X,Y)}$ above, definition of $g_{t,n}$ and properties of expectation be can bound

$$\mathbb{E}g_{s-2^{-n},n}(X)g_{t-2^{-n},n}(Y) \leq F_{(X,Y)}(s, t) \leq \mathbb{E}g_{s,n}(X)g_{t,n}(Y).$$

By assumption

$$\mathbb{E}g_{s-2^{-n},n}(X)g_{t-2^{-n},n}(Y) = \mathbb{E}g_{s-2^{-n},n}(X)\mathbb{E}g_{t-2^{-n},n}(Y)$$

and similarly $\mathbb{E}g_{s,n}(X)g_{t,n}(Y) = \mathbb{E}g_{s,n}(X)\mathbb{E}g_{t,n}(Y)$. As $\mathbb{E}g_{s-2^{-n},n}(X)$ and $\mathbb{E}g_{s,n}(X)$ both converge to $F_X(s)$ and similarly $\mathbb{E}g_{t-2^{-n},n}(Y)$ and $\mathbb{E}g_{t,n}(Y)$ both converge to $F_X(t)$, we conclude.

For the other direction, we first observe the following (this will be on the exercise sheet):

Exercise 3.3. *Prove that if X, Y are independent random variables, then so are $g(X), h(Y)$.*

Given this, the second point follows when we show that for any integrable random variables X, Y we have that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. We first deal with the case of discrete random variables, and then pass to the limit using approximations

The discrete case
Denote the supports by S_X, S_Y and write

$$\mathbb{E}(X)\mathbb{E}(Y) = \left(\sum_{x\in S_X} x\mathbb{P}(X=x)\right)\left(\sum_{y\in S_Y} y\mathbb{P}(Y=y)\right) = \sum_{x\in S_X}\sum_{y\in S_Y} xy\mathbb{P}(X=x)\mathbb{P}(Y=y).$$

Now, for any random variables X, Y and every fixed $x \in S_X, y \in S_Y$ we have the identity

$$\mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x, Y=y)\sum_{s\in S_{XY}} 1_{xy=s}.$$

Further, by independence of X, Y we have $\mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x)\mathbb{P}(Y=y)$. Thus we can write

$$\sum_{x\in S_X}\sum_{y\in S_Y} xy\mathbb{P}(X=x, Y=y) = \sum_{x\in S_X}\sum_{y\in S_Y} xy\mathbb{P}(X=x, Y=y)\sum_{s\in S_{XY}} 1_{xy=s}.$$

By integrability of X, Y, this triple-series is absolutely summable, and thus we can change the order of sums and observe $xy1_{xy=s} = s1_{xy=s}$ to get

$$\sum_{s\in S_{XY}}\sum_{x\in S_X}\sum_{y\in S_Y} s1_{xy=s}\mathbb{P}(X=x, Y=y).$$

Finally, we observe that

$$\sum_{x\in S_X}\sum_{y\in S_Y} 1_{xy=s}\mathbb{P}(X=x, Y=y) = \mathbb{P}(XY=s)$$

which implies the claim for discrete r.v. Observe that this very same change of summation also gives the integrability of XY.

The general case
The general case proceeds via approximation and is left as an exercise.

□

Corollary 3.12. *Let us spell out a corollary of the proof: if X and Y are independent and integrable, then also XY is integrable.*

3.4 Variance and covariance

Next to the mean value or expectation, a key parameter or characteristic of a random variable is its variance (and its standard deviation, which is just the square-root of the variance).

Definition 3.13 (Variance of a random variable)**.** *Let X be an integrable random variable. Then if $\mathbb{E}(|X|^2) < \infty$, we say that X has a finite second moment and define its variance*

$$Var(X) := \mathbb{E}((X - \mathbb{E}X)^2) \geq 0.$$

Standard deviation is defined as $\sigma(X) := \sqrt{VarX}$.

Notice that indeed $(X - \mathbb{E}X)^2$ is integrable when $|X|^2$ is, as we can write $(X - \mathbb{E}X)^2 \leq 2|X|^2 + 2(\mathbb{E}X)^2$. A useful tool for calculating variance is to notice that by opening the square

$$\mathrm{Var}(X) = \mathbb{E}\left((X - \mathbb{E}X)^2\right) = \mathbb{E}(X^2) - 2\mathbb{E}(X\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.$$

So let us calculate some variances using this:

- The variance of a Bernoulli random variable $X \sim Ber(p)$ is $\mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$. Why is this reasonable?
- Similarly, using the same formula we can calculate the variance of an exponential random variable $X \sim Exp(\lambda)$. Indeed, as x^2 satisfies the conditions of Proposition 3.9, we can write
$$\mathbb{E}X^2 = \lambda \int_0^\infty x^2 \exp(-\lambda x) dx.$$
We now calculate by doing twice integration by parts
$$\lambda \int_0^\infty x^2 \exp(-\lambda x) dx = 2 \int_0^\infty x \exp(-\lambda x) dx = 2\lambda^{-1}\mathbb{E}X = 2\lambda^{-2}.$$
Hence $\mathrm{Var}(X) = \lambda^{-2}$.

Variance tells us how much the random variable fluctuates or deviates around its mean, as is illustrated for example by the following lemma:

Lemma 3.14 (Chebyshev's inequality)**.** *Let X be an integrable random variable with finite variance. Then $\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{Var(X)}{t^2}$.*

Proof. This follows directly from the Markov's inequality $\mathbb{P}(Y > t) \leq \frac{\mathbb{E}Y}{t}$ that we proved for non-negative integrable random variables Y on the previous exercise sheet. Indeed, we just apply Markov's inequality to $Y = (X - \mathbb{E}X)^2$ to get that

$$\mathbb{P}(|X - \mathbb{E}X| > t) = \mathbb{P}((X - \mathbb{E}X)^2 > t^2) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

□

In fact, variance also gives us a new view on expectation itself as the minimizer of certain error: if X is an integrable random variable of finite variance, then the real number a that minimizes the so called mean squared error: $\mathbb{E}(X - a)^2$ is given by $a = \mathbb{E}X$! Again, you will find this on the example sheet.

3.4.1 Covariance and correlation

As discussed, one is often is interested how two random variables are related to each other. We already saw the notion of independence - random variables are independent if they don't influence each other at all. In the other extreme there is the case where they are equal, i.e. $\mathbb{P}(X = Y) = 1$ in which case we say $X = Y$ almost surely. Both of those are very strong notions. A weaker measure of how two random variables are related, and a way to in some sense measure the level of dependence is described by notions of covariance and correlation.

Definition 3.15 (Covariance and correlation)**.** *Suppose that X, Y are two integrable random variables of finite variance defined on the same probability space. The covariance of X and Y, denoted $Cov(X, Y)$ is then defined as*

$$Cov(X, Y) = Cov(Y, X) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y.$$

If neither of X, Y is almost surely a constant, then the correlation $\rho(X, Y)$ is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}.$$

A first question might be why is even covariance well-defined? I.e. why is $\mathbb{E}(XY)$ finite when X, Y have finite variance? This follows from the Cauchy-Schwarz inequality, which I believe you have already seen in some form. You will find an non-eximinable proof at the end of the section.

Theorem 3.16 (Cauchy-Schwarz inequality)**.** *Let X, Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that X^2, Y^2 are integrable. Then $|XY|$ is also integrable, and moreover*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Moreover, the equality holds if and only if $|X| = \lambda |Y|$ almost surely for some $\lambda > 0$.

Notice that in particular it also follows that

$$\mathbb{E}(XY) \leq |\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

The relevant cases of equality can be also worked out.

Using this inequality, we see that not only are covariance and correlation well defined, but also we can see that having full correlation means that the random variables are almost surely equal.

Exercise 3.4 (Covariance and dependence)**.** *Let X, Y be two random variables of finite positive variance defined on the same probability space.*

- *Show that the correlation $\rho(X, Y) \in [-1, 1]$. When is it equal to 1, when is it equal to -1, how to interpret this?*
- *Show that if X, Y are independent, integrable with finite variance, then their covariance is zero.*
- *Show that if X, Y are integrable with finite variance, then*

  $$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

  and deduce that if X, Y are also independent, then $Var(X + Y) = Var(X) + Var(Y)$.
- *Finally, find random variables X, Y with zero covariance that are not independent.*

Given a random vector, it is often useful to define the covariance between each pair of components.

Definition 3.17 (Covariance matrix)**.** *Let $\overline{X} = (X_1, \ldots, X_n)$ be a random vector such that all components have finite variance. Then the covariance matrix $\Sigma_{i,j}$ is defined as*

$$\Sigma_{i,j} = Cov(X_i, X_j).$$

In fact, we have already met a covariance matrix! indeed, for a Gaussian random vector $\mathcal{N}(\overline{\mu}, C)$, the matrix positive-definite symmetric matrix C is the covariance matrix and $\overline{\mu} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$:

Exercise 3.5 (Independence and Gaussians)**.** *Prove that for a Gaussian random vector $\bar{X} \sim \mathcal{N}(\overline{\mu}, C)$, the matrix C is the covariance matrix and $\overline{\mu} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$. Show that in the case of a Gaussian random vector, if $Cov(X_i, X_j) = 0$, then X_i and X_j are independent.*

Observe that this in particular means that a Gaussian vector is determined only by its mean and covariance, which is a very nice indeed!

3.5 Moments of a random variable

We have seen that $\mathbb{E}(X)$ and $\mathbb{E}((X-\mathbb{E}X)^2)$ contain valuable information about a random variable X. Moreover, we saw that if we look at $\mathbb{E}g(X)$ for all bounded continuous g, then this determines the law of X completely. But this is already quite a lot of information! An intermediate task would be to ask $\mathbb{E}X^n$ for all $n \geq 1$. Does knowing this determine the random variable?

Definition 3.18 (Moments of a r.v.)**.** *Let X be a random variable and $n \in \mathbb{N}$. If $\mathbb{E}|X|^n < \infty$, we say that X admits a n-th moment. We call $\mathbb{E}X^n$ the n-th moment of X.*

To understand the relation between different moments, let's recall the Jensen's inequality. A function $\phi : \mathbb{R} \to \mathbb{R}$ is called convex if for all x, y and all $\lambda \in [0,1]$ we have that

$$\phi(\lambda x + (1-\lambda)y) \leq \lambda\phi(x) + (1-\lambda)\phi(y).$$

We call $\lambda x + (1-\lambda)y$ a convex combination of x, y. Using this vocabulary, Jensen's inequality can be reworded as saying that the image under ϕ of a convex combination of two points is always smaller than the convex combination of the images of the two points under ϕ. (What does it mean geometrically?)

Finally, recall that a convex function is continuous and thus if X is a random variable, then so is $\phi(X)$. We can now state Jensen's inequality:

Theorem 3.19 (Jensen's inequality)**.** *Let X be an integrable random variable and ϕ a convex function such that $\phi(X)$ is also integrable. Then*

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

Notice the similarity to the defining property of convexity: $\mathbb{E}X$ can be thought of as a convex combination of the possible values of X. Thus, for example if X takes only two values x, y with probabilities λ and $1-\lambda$ then Jensen's inequality is just a reformulation of the defining property of convexity.

I expect you have seen and will see many different proofs of this nice inequality. Still there is one in the appendix on this section for completeness.

We now have the following simple lemma, saying that the existence of higher moments implies the existence of lower moments too:

Lemma 3.20. *Let X be a random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P}$ that admits a n-th moment. Then it also admits a m-th moment for all $m \leq n$ and moreover $\mathbb{E}|X|^n \leq (\mathbb{E}(|X|^m))^{n/m}$.*

Proof. Let $m \leq n$. Let us first notice that if $|X|^n$ is integrable, then also is $|X|^m$ with $m \leq n$. Indeed, we can bound

$$|X(\omega)|^m \leq \max(|X(\omega)|^n, 1) \leq |X(\omega)|^n + 1$$

and thus integrability of $|X|^m$ follows from that of $|X|^n$.

Now, for $n \geq m$, consider $\phi(x) = |x|^{n/m}$. This is a convex function. Hence, as both $|X|^m$ and $|X|^n = \phi(|X|^m)$ are integrable, we can apply Jensen's inequality to ϕ and $|X|^m$ and obtain

$$\mathbb{E}|X|^n = \mathbb{E}(\phi(|X|^m)) \geq \phi(\mathbb{E}|X|^m) = (\mathbb{E}(|X|^m))^{n/m},$$

concluding the proof. □

In particular, the former Lemma says that if the second moment of X exists, then both X is integrable and of finite variance. Many random variables you will see in statistics or numerics will have finite variance, so it's useful to have a good condition for that. You will see on the example sheet that the converse is not true, there will be examples of integrable random variables with infinite variance and so on.

The existence of moments has a direct influence on how the tails of the random variable behave. Indeed, by Markov's inequality if $\mathbb{E}|X|^n < \infty$, we know that

$$\mathbb{P}(X > t) \leq \mathbb{P}(|X|^n > t^n) \leq \frac{\mathbb{E}|X|^n}{t^n},$$

i.e. the tail behaves like $O(t^{-n})$. In case of finite variance we only knew that the tail behaves like $O(t^{-2})$ for example. Or in simple words - having higher moments that very big values are taking with smaller probability.

Let us now come to the interesting question - do the moments uniquely determine the distribution? This is true in quite large generality, but not always. We will here prove a partial result:

Proposition 3.21. *Let X, Y be two almost surely bounded random variables, i.e. r.v. such that almost surely $X \in [-A, A]$ and $Y \in [-A, A]$ for some $A > 0$. Suppose further that $\mathbb{E}X^n = \mathbb{E}Y^n$ for every $n \in \mathbb{N}$. Then X and Y have the same law.*

Before embarking on the proof, observe that trivially for bounded random variables all moments do exist - namely, if X is bounded then every $|X|^n$ is bounded too. The proof we give relies on the following theorem of independent interest:

Theorem 3.22 (Stone-Weierstrass)**.** *Let f be a continuous function on some interval $I = [-A, A]$. Then f can be uniformly approximated by polynomials: i.e. there is a sequence of polynomials $(P_n)_{n\geq 1}$ such that $(P_n)_{n\geq 1}$ converges to f in $(C(I, \mathbb{R}), d_\infty)$, where as usual $d_\infty(f, g) = \sup_{x\in I} |f(x) - g(x)|$.*

Most likely, you will see the proof of this theorem in several courses from several points of view. As it is a beautiful result, it is well worth mentioning it several times. In fact, we will also provide a short probabilistic, but non-examinable proof at the end of the subsection. Let us first see how it implies the proposition.

Proof of Proposition 3.21. The proposition follows rather easily from Stone-Weierstrass theorem. Indeed, by the assumption and by linearity of expectation, we see that $\mathbb{E}P(X) = \mathbb{E}P(Y)$ for each polynomial P.

Our aim is to use Proposition 3.10, i.e. to prove that $\mathbb{E}\widehat{g}(X) = \mathbb{E}\widehat{g}(Y)$ for all continuous bounded $\widehat{g}$. Notice that any such $\widehat{g}$ gives rise to a continuous function $g : [-A, A] \to \mathbb{R}$, by restriction. Moreover as $X, Y \in [-A, A]$ almost surely, we see that $\mathbb{E}\widehat{g}(X) = \mathbb{E}g(X)$ and hence it suffices to argue that $\mathbb{E}g(X) = \mathbb{E}g(Y)$ for continuous functions on $[-A, A]$.

Given such a function g, by the Stone-Weierstrass theorem for every $\epsilon > 0$, there is some polynomial P_ϵ such that $d_\infty(g, P_\epsilon) < \epsilon$. As $\mathbb{E}P_\epsilon(X) = \mathbb{E}P_\epsilon(Y)$, we can write

$$|\mathbb{E}g(X) - \mathbb{E}g(Y)| = |\mathbb{E}g(X) - \mathbb{E}P_\epsilon(X) + \mathbb{E}P_\epsilon(Y) - \mathbb{E}g(y)|,$$

and bound this from above using by triangle inequality by

$$|\mathbb{E}\left(g(X) - P_\epsilon(X)\right)| + |\mathbb{E}\left(g(Y) - P_\epsilon(Y)\right)|.$$

Further,

$$|\mathbb{E}\left(g(X) - P_\epsilon(X)\right)| \leq \mathbb{E}|g(X) - P_\epsilon(X)| < \epsilon.$$

But now as $X \in [-A, A]$ almost surely, and $|g(x) - P_\epsilon(x)| < \epsilon$ for $x \in [-A, A]$, we see that $|g(X) - P_\epsilon(X)| < \epsilon$ almost surely, and hence by Proposition 3.7 we deduce that $\mathbb{E}|g(X) - P_\epsilon(X)| < \epsilon$.

Hence we conclude that $|\mathbb{E}g(X) - \mathbb{E}g(Y)| \leq 2\epsilon$ and as $\epsilon > 0$ was arbitrary we conclude that $\mathbb{E}g(X) = \mathbb{E}g(Y)$. As g was arbitrary, the proposition now follows from Proposition 3.10. □

So what could go wrong in general?

First, of course all moments might not exist and then only the few existing moments might not characterize the distribution. For example, if you define discrete random variables X_1 and X_2 with supports $\mathbb{Z} \setminus \{0\}$ and $2\mathbb{Z} \setminus \{0\}$ respectively by setting $\mathbb{P}(X_1 = k) = ck^{-3}$ and $\mathbb{P}(X_2 = 2k) = ck^{-3}$ with $c = \frac{1}{2\sum_{k\geq 1} k^{-3}}$, then X_1, X_2 are integrable with zero mean by symmetry. However neither admits a second moment (see Exercise sheet) and they are also not equal in law as their supports are different.

Second, even if all moments exist, they might grow too quickly to characterize the distribution:

Exercise 3.6 (Moment problem)**.** *Let X be a standard normal random variable. Prove that $W = \exp(X)$ admits all moments and calculate these moments. Let $a > 0$, and consider a discrete random variable Y_a with support*

$$S_a = \{ae^m : m \in \mathbb{Z}\}$$

and defined by

$$\mathbb{P}(Y_a = ae^m) = \frac{1}{Z}a^{-m}e^{-m^2/2}$$

with $Z = \sum_{m\in\mathbb{Z}} a^{-m}e^{-m^2/2}$ (why is it finite?). Show that Y_a admits all moments and that moreover for every $n \in \mathbb{N}$, $\mathbb{E}W^n = \mathbb{E}exp(Xn) = \mathbb{E}Y_a^n$.

3.5.1 Moment generating function

We considered moments of random variables and saw that they might give a useful countable collection of numbers that fully characterizes the underlying random variable. But what if instead of moments we look at some other family of functions $g(X)$ and their expectations? It comes out that a very useful family is directly related to moments: we consider $\mathbb{E}e^{tX}$ for all $t \in \mathbb{R}$ such that e^{tX} is integrable.

Definition 3.23 (Moment generating function)**.** *If X is a random variable such that $\exp(tX)$ is integrable for some interval $I = (-c, c)$ around 0. We say that X admits a moment-generating function (MGF) in a neighbourhood around 0 and denote $M_X(t) = \mathbb{E}\exp(tX)$ for $t \in I$.*

The name comes from the fact that when $M_X(t)$ exists in a small interval, we can write

$$M_X(t) = \mathbb{E}(\exp(tX)) = \mathbb{E}(\sum_{n\geq 1} \frac{t^n X^n}{n!}).$$

Checking that you can exchange the summation and the expectation (On the Exercise sheet), one obtains

$$M_X(t) = \sum_{n\geq 1} \frac{t^n}{n!}\mathbb{E}X^n.$$

In particular, from here it is not hard to deduce that if we look at $M_X(t)$ as a function of t, then in fact moments $\frac{d^n}{dt^n}M_X(t)$ evaluated at $t = 0$ just gives the n-th moment. We will skip this calculation that is not examinable.

It comes out that MGF-s also characterize the distribution. We state this result and you are free to use it, though the proof is out of the scope of this course:

Theorem 3.24 (MGF determines the distribution (admitted))**.** *Let X, Y be random variables such that $M_X(t)$ and $M_Y(t)$ exist in some open interval around 0, and moreover $M_X(t) = M_Y(t)$ in this interval. Then X and Y have the same law.*

In fact moment generating functions and this concrete theorem for MGFs also nicely generalize to random vectors:

Theorem 3.25 (MGF for random vectors (admitted))**.** *Let $\overline{X}$ be a random vector taking values in $\mathbb{R}^n$ such that $\mathbb{E}e^{\langle \bar{t}, \overline{x}\rangle} < \infty$ for $\bar{t}$ in some open neighbourhood of 0.*[12] *We then call $M_{\overline{X}}(\bar{t}) = \mathbb{E}e^{\langle \bar{t}, \overline{x}\rangle}$ the moment generating function of $\overline{X}$. Again, if MGFs of two random vectors $\overline{X}$ and $\overline{Y}$ are equal in some neighbourhood around 0, then $\overline{X}$ and $\overline{Y}$ have the same law.*

These two results are extremely useful. First, as an application MGF-s can be used to determine independence:

Lemma 3.26 (Independence and MGF)**.** *Let X, Y be random variables such that there exists an open interval $I \subset \mathbb{R}$ containing zero such that $M_X(t)$ and $M_Y(t)$ exist for all $t \in I$. Then X, Y are independent iff for each $t, s \in I$, $M_X(t)M_Y(s) = M_{(X,Y)}((t, s))$.*

Proof. Firstly, if X, Y are independent then the condition follows directly from Proposition 3.11. Indeed, for each $t, s \in I$ we can take $g(x) = \exp(tx)$ and $h(y) = \exp(sy)$. Then $M_X(t) = \mathbb{E}g(X)$ and $M_Y(s) = \mathbb{E}h(Y)$ and by assumption both are integrable. Hence that proposition implies that $M_X(t)M_Y(s) = \mathbb{E}\exp(tX + sY) = M_{(X,Y)}(t, s)$.

The other direction is a direct application of Theorem 3.25: indeed, let (X, Y) be a pair of random variables such that for each $t, s \in I$, $M_X(t)M_Y(s) = M_{(X,Y)}((t, s))$. Further, let $(\tilde{X}, \tilde{Y})$ be a pair of independent random variables such that $\tilde{X}$ has the law of X and $\tilde{Y}$ has the law of Y. In particular then $M_X(t) = M_{\tilde{X}}(t)$ and $M_Y(s) = M_{\tilde{Y}}(s)$for all $t, s \in I$.

Now, by the first part $M_{\tilde{X}}(t)M_{\tilde{X}}(s) = M_{(\tilde{X},\tilde{Y})}((t, s))$. We conclude that $M_{(X,Y)}((t, s)) = M_{(\tilde{X},\tilde{Y})}((t, s))$ and deduce from Theorem 3.25 that (X, Y) and $(\tilde{X}, \tilde{Y})$ have the same joint law. In particular X and Y are independent. □

Second, it really makes some things much easier, in particular calculations with Gaussians:

[12]Here $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^n$

Exercise 3.7. *Prove $\overline{X}$ is a Gaussian vector with mean $\overline{\mu}$ and covariance C if and only if $M_{\overline{X}}(\overline{t}) = \exp(\langle \overline{t}, \overline{\mu}\rangle + \frac{1}{2}\langle \overline{t}, C\overline{t}\rangle)$. Deduce that*

- *If X is a standard Gaussian on $\mathbb{R}^n$, then so is OX for every orthogonal $n \times n$ matrix.*
- *The Gaussian vector with mean $\overline{\mu}$ and covariance C on $\mathbb{R}^n$ can be written as $A\overline{Y} + \overline{\mu}$, where $\overline{Y}$ is the standard Gaussian on $\mathbb{R}^n$ and $C = \sqrt{AA^T}$ (You may assume such a matrix A exists, but you have seen it in linear algebra!)*

Thus having an MGF can really simplify and reduce calculations. The drawback of moment generating functions is that they do not always exist.

Exercise 3.8. *Consider the log-normal random variable, i.e. $Z = \exp(X)$ where X is a standard Gaussian. Prove that there is no open interval around 0 such that $M_t(Z)$ exists in this interval.*

This can be mended by considering what is called the characteristic function, defined by $c_X(t) = \mathbb{E}e^{itX}$. [13] The characteristic function always exists for all $t \in \mathbb{R}$ as $\exp(itX)$ is bounded (the integral is taken separately in the imaginary and real component)! Moreover, it uniquely characterizes the law of the random variable. But this will already topic of a future course...

3.6 ⋆ Proofs of some auxiliary results (non-examinable) ⋆

[⋆ non-examinable section begins ⋆]

In this non-examinable section we present proofs of some auxiliary results. I do recommend the probabilistic proof of the Stone-Weierstrass theorem, it is a gem!

First let us prove the Cauchy-Schwarz inequality:

Proof of Cauchy-Schwarz inequality. Define $\widehat{Y}, \widehat{X}$ as $\widehat{Y} = \frac{Y}{\sqrt{\mathbb{E}(Y^2)}}$ and $\widehat{X} = \frac{X}{\sqrt{\mathbb{E}(X^2)}}$. This is possible as X^2, Y^2 are integrable. Notice that by definition then $\mathbb{E}(\widehat{Y}^2) = \mathbb{E}(\widehat{X}^2) = 1$. Moreover, the Cauchy-Schwarz inequality is then equivalent to

$$\mathbb{E}(|\widehat{X}\widehat{Y}|) \leq 1. \tag{3.3}$$

But now for every $\omega \in \Omega$, we have that $|\widehat{X}(\omega)\widehat{Y}(\omega)| \leq \frac{1}{2}(\widehat{X}^2(\omega) + \widehat{Y}^2(\omega))$. Thus we see that $|XY|$ is integrable and by properties of expectation

$$\mathbb{E}(|\widehat{X}\widehat{Y}|) \leq \frac{1}{2}\mathbb{E}(\widehat{X}^2 + \widehat{Y}^2) = 1,$$

and the inequality 3.3 follows.

The equality holds if and only if $|\widehat{X}\widehat{Y}| = \frac{1}{2}(\widehat{X}^2 + \widehat{Y}^2)$ almost surely, which in turn holds if and only if $|\widehat{X}| = |\widehat{Y}|$ almost surely. As $\widehat{Y}, \widehat{X}$ are normalized versions of X, Y, this is turn holds if $|X| = \lambda|Y|$ almost surely for some $\lambda > 0$. □

Next, it is time to prove Jensen's inequality. We will do it using the following chracterization of convex functions:

- $\phi : \mathbb{R} \to \mathbb{R}$ is convex if and only if for every $x \in \mathbb{R}$, there is some $c = c(x) \in \mathbb{R}$ so that for every $y \in \mathbb{R}$, we have that $\phi(x + y) \geq \phi(x) + c_x y$.

[13] In fact, in case of random variables with density, it corresponds to the Fourier transform of the density (why?)

Proof of Jensen's inequality. Consider $x = \mathbb{E}X$ and $y = X - \mathbb{E}X$. Then injecting this in the formulation of convexity just above, we obtain

$$\phi(X) \geq \phi(\mathbb{E}X) + c(X - \mathbb{E}X)$$

almost surely. Taking now expectation, and using the fact that $\mathbb{E}(X - \mathbb{E}X)) = 0$, we deduce

$$\mathbb{E}\phi(X) \geq \phi(\mathbb{E}X)$$

as claimed. □

And finally the cute probabilistic proof of the Stone-Weierstrass theorem:

Proof of Theorem 3.22. By translation and scaling, it is simple to see that it suffices to prove the theorem for the case $I = [0, 1]$ and f continuous on $[0, 1]$. Now for every $x \in [0, 1], n \in \mathbb{N}$ let $X_{n,x}$ be a Binomial random variable of parameters (n, x) We define $P_n(x) = \mathbb{E}f(X_{n,x}/n)$. By Proposition 3.9 we then have

$$P_n(x) = \sum_{k=0}^{n} f(k/n)\binom{n}{k}x^k(1-x)^{n-k},$$

and hence $P_n(x)$ is a polynomial of order n in x.

We claim that $P_n(x)$ converges to f uniformly. First, notice that as f is continuous on $[0, 1]$ it is bounded by some M, and uniformly continuous - i.e. for every $\epsilon > 0$, there is some $\delta_\epsilon > 0$ so that if $|x - y| < \delta_\epsilon$, then $|f(x) - f(y)| < \epsilon$.

Now, write

$$|P_n(x) - f(x)| = |\mathbb{E}(f(X_{n,x}/n) - \mathbb{E}f(x)| \leq \mathbb{E}|f(X_{n,x}/n) - f(x)|.$$

The crux is something we have already seen: in fact $X_{n,x}$ is very close to its expectation xn for n large. Indeed, we by Chebyshev's inequality and the fact that $\mathrm{Var}(X_{n,x}) = nx(1 - x)$

$$\mathbb{P}(|X_{n,x}/n - x| > t/n) = \mathbb{P}(|X_{n,x} - nx| > t) \leq \frac{\mathrm{Var}X_{n,x}}{t^2} = \frac{nx(1-x)}{t^2}.$$

In particular, if we choose $t = n^{2/3}$, then $\mathbb{P}(|X_{n,x}/n - x| > n^{-1/3}) < n^{-1/3}$.

To use this fact we write:

$$\mathbb{E}|f(X_{n,x}/n) - f(x)| = \mathbb{E}\left(|f(X_{n,x}/n) - f(x)|1_{|X_{n,x}/n-x|>n^{-1/3}}\right) + \mathbb{E}\left(|f(X_{n,x}/n) - f(x)|1_{|X_{n,x}/n-x|<n^{-1/3}}\right).$$

Then as $|f(x)| < M$ for $x \in [-A, A]$, we can bound the first term by

$$M\mathbb{E}1_{|X_{n,x}/n-x|>n^{-1/3}} = M\mathbb{P}(|X_{n,x}/n - x| > n^{-1/3}) < Mn^{-1/3}.$$

Fix some $\epsilon > 0$ and choose n large enough so that $n^{-1/3} < \delta_\epsilon$. We can bound the second term by

$$\mathbb{E}\epsilon 1_{|X_{n,x}/n-x|<n^{-1/3}} \leq \epsilon.$$

Hence if we also require that $n^{-1/3} < \epsilon$, we obtain altogether

$$\mathbb{E}|f(X_{n,x}/n) - f(x)| < Mn^{-1/3} + \epsilon \leq (M + 1)\epsilon.$$

As this is uniform in x and holds for arbitrary ϵ, the theorem follows. □

[⋆ non-examinable section ends ⋆]

Section 4

Limit theorems

In this section, we will look at infinite sequences of events and infinite sequences of random variables. Some questions we will be interested in:

- When can we be sure that at least one of the events $A_1, A_2, \dots$ happens? For example, under what conditions can you guarantee that you will eventually win with a lottery or get a 6 in the exam? Or suppose, you start a random walk in Manhatten - at every corner you choose uniformly one of four directions. Will you ever get back to your hotel?
- Under what criteria do only finitely many of the events $A_1, A_2, \dots$ fail? For example, under what criteria do we know that a infectious disease that is spreading will only last for a finite time?
- When can we say something about the limit of the sequence of random variables $X_1, X_2, \dots$? We have already seen some vague statements in the lines that $Bin(n, \lambda/n)$ converge to Poisson or $Bin(n, 1/2)$ when normalized converges to the Gaussian. How to make such statements mathematically precise, especially and how to treat these situations in general?
- What about the limit of $\mathbb{E}X_1, \mathbb{E}X_2, \dots$ if the underlying random variables converge?

We will see how such questions come up naturally, find some cases where they become tractable and even easy. As often in mathematics, looking at limiting situations makes things more tractable. For example, somtimes to gain understanding of complex random systems, e.g. like complex networks, it is useful to see what happens if we let the size of the network go to infinite. Can we talk of some infinite network?

4.1 Infinite collections of events and random variables

Let us start by formalizing some of the limiting notions in the context of events. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of events $E_1, E_2, \dots$ that could for example be repetitions of the same random situation, like repetitive coin tosses. Recall that E_i is an event means that $E_i \subseteq \Omega$ and $E_i \in \mathcal{F}$. Each ω gives a random state of the universe, and $\omega \in E_i$ if the event E_i happens for this particular state.

Now, we say that

- First, we could ask whether at least one event of the sequence E_n happens. By definition, $\{\omega \in \Omega : \omega \in E_i \text{ for some } i\} = \bigcup_{n \geq 1} E_n$. Sometimes one says that 'E_i happens eventually'. An example would be the following example from an earlier example sheet: tossing independent coins, we eventually obtain heads with full probability (this also follows from the lemma just below). Notice that there is some sequence of tosses that gives no heads - the sequence $TTTTT\dots$, however as it has 0 probability, it does not matter.
- Second, we might ask whether the events E_i happen infinitely often. It requires a check to see that

$$\{\omega \in \Omega : \omega \in E_i \text{ for infinitely many } i\} = \bigcap_{m \geq 1} \bigcup_{n \geq m} E_n.$$

This event is also sometimes denoted by $\limsup_{n\geq 1} E_n$. In the case of coin tossing, each E_i could mean that the i-th toss comes up heads, and we have seen that in the case of independent coins, indeed E_i would happen infinitely often with full probability.

- Finally, we might ask whether all but finitely many E_i happen. One can again see (on the exercise sheet), that

$$\{\omega \in \Omega : \omega \in E_i \text{ for all but finitely many } i\} = \bigcup_{m\geq 1} \bigcap_{n\geq m} E_n.$$

This event is also denoted by $\liminf_{n\geq 1} E_n$. An example situation would be as follows: you start with 10 CHF, and as long as you have some money left, you bet with the European central bank (that can always print more money when needed!) on whether independent coin tosses are head or tails. The winner gets 1 CHF, and the loser loses 1 CHF. It's a mathematical fact that after almost surely, after finitely many bets you are left with 0 CHF. So if we denote by E_i the event after i bets you are bankrupt, this event fails only finitely many times.

Here are some useful criteria to study such events. First, a very naive criterion:

Lemma 4.1. *Let $E_1, E_2, \ldots$ be independent events of probability p_i. Then $\mathbb{P}(\bigcup_{i\geq 1} E_i) = 1$ if and only if $\Pi_{i=1}^n (1-p_i) \to 0$ as $n \to \infty$.*

Proof. This is on the exercise sheet. □

For example, if each event happens with the same probability p, then $\Pi_{i=1}^n p_i = p^n$, which clearly goes to zero. So even if you toss a coin that comes up heads with probability 0.00001, you will eventually see heads.

A verey useful criteria for verifying that some even cannot happen but finitely many times is given by the first Borel-Cantelli lemma:

Lemma 4.2 (Borel-Cantelli I)**.** *Let $E_1, E_2, \ldots$ be any sequence of events on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, then almost surely only finitely many events E_i happen, i.e. $\mathbb{P}(\bigcap_{m\geq 1} \bigcup_{n\geq m} E_n) = 0$.*

Notice that we are not assuming anything on the dependence or independence of the events E_i! Also, this lemma does not say that there is some fixed number 1000 of events that happen. Indeed, exactly how many events can happen and exactly which events happen depends on $\omega \in \Omega$.

For example, consider a sequence of unfair coins with probability of heads for the n-th coin given by n^{-2}. If E_n denotes the event of obtaining heads on the n-th toss, then $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$. Thus, by the lemma, we see that almost surely one obtains only finitely many heads in an infinite sequence of coin tosses. However, notice that whether you obtain 10 or even 100 heads depends on the exact sequence of tosses, i.e. on the 'randomness' encoded by the state $\omega \in \Omega$.

Proof. Fix some $\epsilon > 0$. As $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, we can find some $n_0 \in \mathbb{N}$ such that $\sum_{n\geq n_0} \mathbb{P}(E_n) < \epsilon$. But now as $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$,

$$\mathbb{P}(\bigcap_{m\geq 1} \bigcup_{n\geq m} E_n) \leq \mathbb{P}(\bigcup_{n\geq n_0} E_n) \leq \sum_{n\geq n_0} \mathbb{P}(E_n) < \epsilon,$$

where in the last inequality we use the union bound. As ϵ was arbitrary, the claim follows. $\square$

The short proof might make you suspicious if it is of any use. But think for example of the following. Assume that we have $X_1, X_2, \dots$ be a sequence of random variables on the same probability space, each with law $Geo(1/2)$ but such that we know nothing about the dependence structure. What can we say about the maximum of n first random variables?

Using Borel-Cantelli I, we can easily get some nice information:

Exercise 4.1. *Assume that we have $X_1, X_2, \dots$ be a sequence of random variables on the same probability space, each with law $Geo(1/2)$. Let $E_n = \{\max_{i=1}^n X_i > \sqrt{n}\}$. Show that almost surely only finitely many of $E_1, E_2, \dots$ happen, i.e. $\mathbb{P}(\bigcap_{n\geq 1}\bigcup_{i\geq n} E_i) = 0$. Deduce that there exists some random variable $C : \Omega \to \mathbb{R}$ that takes a.s. non-negative values and such that $\mathbb{P}(\max_{i=1}^n X_i(\omega) < C(\omega)\sqrt{n}) = 1$.*

This is partly complemented by the second Borel-Cantelli lemma, which gives a condition for infinitely many events to happen. Notice that here we again ask for independent events.

Lemma 4.3 (Borel-Cantelli II)**.** *Let $E_1, E_2, \dots$ be a sequence of independent events on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $\sum_{n\geq 1}\mathbb{P}(E_n) = \infty$. Then almost surely infinitely many events E_i happen, i.e. $\mathbb{P}(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n) = 1$.*

Proof. We have that

$$\mathbb{P}(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n) = 1 - \mathbb{P}(\bigcup_{m\geq 1}\bigcap_{n\geq m} E_n^c)$$

and hence it suffices to show that $\mathbb{P}(\bigcup_{m\geq 1}\bigcap_{n\geq m} E_n^c) = 0$. By the union bound

$$\mathbb{P}(\bigcup_{m\geq 1}\bigcap_{n\geq m} E_n^c) \leq \sum_{m\geq 1}\mathbb{P}(\bigcap_{n\geq m} E_n^c).$$

Further, as E_i are independent, so are E_i^c, and hence

$$\mathbb{P}(\bigcap_{n\geq m} E_n^c) = \Pi_{n\geq m}\mathbb{P}(E_n^c) = \Pi_{n\geq m}(1 - \mathbb{P}(E_n)).$$

Now using the inequality $1 - x \leq e^{-x}$ for $x \in [0, 1]$, we can bound the RHS further by $\exp(-\sum_{n\geq m}\mathbb{P}(E_n))$. But the sum in the exponential equals ∞ by the assumption. Thus $\mathbb{P}(\bigcap_{n\geq m} E_n^c) = 0$, hence $\mathbb{P}(\bigcup_{m\geq 1}\bigcap_{n\geq m} E_n^c) = 0$ and we conclude.

$\square$

As already exemplified by the proof, the criteria of independence is indeed necessary:

Exercise 4.2. *Find events $E_1, E_2, \dots$ on the same probability space such that $\sum_{n\geq 1}\mathbb{P}(E_n) = \infty$, but $\mathbb{P}(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n) = 0$. Also, find events $E_1, E_2, \dots$ such that $\mathbb{P}(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n)$ happens with probability p.*

These lemmas look very innocent, but actually have nice applications (we will see some later). First, a simple corollary says that independent events either happen infinitely often with probability 1 or 0 - this is quite remarkable, as a priori one might think that it could happen with any probability, like in the exercise above. So we see how the 'simple-looking' assumption of independence can really sway things:

Corollary 4.4. *Let $E_1, E_2, \dots$ be mutually independent events on a common probability space. Then $\mathbb{P}(\bigcap_{m\geq 1}\bigcup_{n\geq m} E_n) \in \{0,1\}$, i.e. E_i happens infinitely often either with probability 0 or 1.*

Proof. This follows directly from the Borel-Cantelli lemmas, as either $\sum_{n\geq 1}\mathbb{P}(E_n) < \infty$ or $\sum_{n\geq 1}\mathbb{P}(E_n) = \infty$. □

In fact, this is a special case of a more general Kolmogorov 0-1 law, that we only meet in the non-examinable section this year.

Things are similar, but a bit more exciting when we switch from events to sequences of random variables $X_1, X_2, \dots$. Again, firstly the question is what we can even ask about an infinite sequence of random variables - not all functionals might be measurable!

For example some questions that we might be interested in are:

- Is same value k attained by the sequence of random variables?
- Are all but finitely many of X_i positive?
- Is the sequence of random variables bounded in absolute value?
- Does it converge?

For the first one measurability is clear, as we can write it as the union $\bigcup_{n\geq 1}\{X_i = k\}$, similarly for the second one. For the third one, already some thought might be required: the event that the sequence of random variables is bounded in absolute value by $M \in \mathbb{N}$ is given by $E_M := \bigcap_{n\geq 1}\{|X_i| \leq M\}$. But we want to allow different bounds for different sequences. So we have to take also a union over M to get $\bigcup_{M\in\mathbb{N}} E_M$, which again shows that the question makes fully sense.

4.2 Convergence of random variables

We now get to the heart of this section which is not only asking whether sums or sequences of random variables converge or not, but trying to understand what do they converge to. So our model situation will be something as follows: $X_1, X_2, \dots$ are some random variables and we ask if X_n converges in some sense and to what it might converge. In fact, there are several notions of convergence: almost sure convergence, convergence in probability and convergence in law. They apply in different situations and describe different things.

4.2.1 Almost sure convergence

Maybe the most natural notion is that of almost sure convergence. For this notion, the setting is as follows: we have some random variables $X_1, X_2, \dots$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we just ask about the event $\{\omega \in \Omega : X_n(\omega) \text{ converges}\}$. For example, again with coin tossing you might toss coin a hundred times and take the average, and then a thousand times and take the average. Do these averages converge? The definition is as follows.

Definition 4.5 (Almost sure convergence)**.** *Let $X_1, X_2, \dots$ be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If for some random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ we have that $\mathbb{P}(\{\omega \in \Omega : (X_n(\omega)_{n\geq 1} \to X(\omega)\}) = 1$, then we say that the sequence $(X_n)_{n\geq 1}$ converges almost surely to X.*

Your first question should be again, why is this event in the definition even measurable! The exercise sheet will help you out.

Remark 4.6 ($\star$ non-examinable $\star$)**.** *In the spirit of the first half of the course, you might further ask - given the joint laws of any $(X_{i_1}, \dots, X_{i_n})$ for any finite subset $\{i_1, \dots, i_n\}$ of $\mathbb{N}$, can we even define a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_1, X_2, \dots$ are random variables defined on this space and satisfy the given joint laws? We have argued that this is possible in case $X_1, X_2, \dots$ are mutually independent by the construction of a product measure. This can be generalized to hold for more general sequences, as long as certain consistency conditions hold for the finite-dimensional joint laws. The relevant theorem is called Kolmogorov Extension Theorem. However, we will restrict ourselves to sequences of independent random variables, and thus will not go any deeper into this.*

4.2.2 Convergence in law

The most common notion and maybe the most important one is however 'convergence in law'. Convergence in law describes the convergence of distributions, if you wish - geometrically the convergence of histograms. For example, you could think of the following situation - your aim of life is to learn to toss a perfect random coin. In the beginning, you don't throw strong enough and there is a bias for the coin to do only one revolution and come on top with the side that was downwards. So you model your throw with $Ber(p)$ random variable with $p \neq 1/2$. As you practice more and more, you get better and finally your coin tosses are really nearly perfect $Ber(1/2)$ random variables. At different stages of your development you have different distributions, that you can model on different probability spaces. Over time these probability distributions start looking more and more like $Ber(1/2)$ in sense that their probability laws converge.

Definition 4.7 (Convergence in law)**.** *We say that a sequence of random variables $X_1, X_2, \dots$ converges in law (also: converges in distribution) to a random variable X if $F_{X_n}(t) \to F_X(t)$ for every t that is a continuity point of F_X, i.e. that is such that $\mathbb{P}(X = t) = 0$.*

Notice that we don't ask $X_1, X_2, \dots$ to be defined on the same probability space! This is not necessary, as we are in any case only looking at their laws $\mathbb{P}_{X_i}$, that are uniquely characterized by F_{X_i}.

It might be strange that we don't ask for convergence at all points $t \in \mathbb{R}$. The reason is the following: consider deterministic random variables X_n taking value $1/n$. Then we would intuitively want to say that X_n converge to the deterministic random variable X that takes value 0 almost surely. However, notice that $F_{X_i}(0) = 0$ for all $n \in \mathbb{N}$, but $F_X(0) = 1$. Thus if we asked for convergence for all t, the random variables X_n would not converge to 0...however, with the definition given above, they nicely do!

Still, notice that if the limiting random variable is continuous, we really do ask the pointwise convergence of c.d.f. at all points.

To better understand the notion of convergence in law, it might be useful to think of an equivalent criteria. In fact there are many equivalent criteria!

Proposition 4.8. *Let $X_1, X_2, \dots$ be a sequence of random variables. They converge to a random variable X in law if and only if for every $a < b$ with $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$ we have that $\mathbb{P}(X_n \in (a, b)) \to \mathbb{P}(X \in (a, b))$*

Proof. If $(X_n)_{n\geq 1}$ converge in law to X then by definition $F_{X_n}(t) \to F_X(t)$ for any continuity point t of $F_X(t)$. In particular, if $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$, then the points a, b are such continuity points. We can write

$$\mathbb{P}(X \in (a,b)) = F_X(b) - F_X(a) = \lim_{n\to\infty} (F_{X_n}(b) - F_{X_n}(a)).$$

But now $\mathbb{P}(X_n \in (a,b)) = (F_{X_n}(b^-) - F_{X_n}(a))$. It suffices to now see that $\lim_{n\to\infty} F_{X_n}(b^-) = \lim_{n\to\infty} F_{X_n}(b)$. But this follows from the fact that b is a continuity point as for every $\epsilon > 0$ we have that

$$F_{X_n}(b-\epsilon) \leq F_{X_n}(b^-) \leq F_{X_n}(b)$$

and if $b - \epsilon$ is also a continuity point, we deduce

$$F_X(b-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(b^-) \leq \limsup_{n\to\infty} F_{X_n}(b^-) \leq F_X(b),$$

which letting $\epsilon \to 0$ gives the desired equality.

In the other direction, we want to prove that for each t with $\mathbb{P}(X = b) = 0$, we have that $\mathbb{P}(X_n < b) \to \mathbb{P}(X < b)$. Now, we know that for any $a < b$ with $\mathbb{P}(a = 0)$, we have $\mathbb{P}(X_n \in (a,b)) \to \mathbb{P}(X \in (a,b))$. As there are only countably many a with $\mathbb{P}(X = a) > 0$, we can choose $a \to -\infty$ and conclude that $\mathbb{P}(X_n < b) \geq \mathbb{P}(X_n \in (a,b)) \to_{n\to\infty} \mathbb{P}(X \in (a,b))$. As $\mathbb{P}(X \in (a,b)) \to \mathbb{P}(X < b)$ as $a \to -\infty$, we deduce that $\liminf_{n\to\infty} \mathbb{P}(X_n < b) \geq \mathbb{P}(X < b)$. Similarly one can see that $\liminf_{n\to\infty} \mathbb{P}(X_n > b) \geq \mathbb{P}(X > b)$. But now

$$1 \geq \liminf_{n\to\infty} (\mathbb{P}(X_n < b) + \mathbb{P}(X_n > b)) \geq \liminf_{n\to\infty} \mathbb{P}(X < b) + \liminf_{n\to\infty} \mathbb{P}(X > b) \geq \mathbb{P}(X < b) + \mathbb{P}(X > b).$$

As $\mathbb{P}(X < b) + \mathbb{P}(X > b) = 1$, we see that in fact the inequalities have to be equalities and thus we conclude. □

Remark 4.9. *In fact the same proof gives a seemingly weaker but actually equivalent condition: we ask that for all $a < b$, it holds that $\liminf_{n\geq 1} \mathbb{P}(X_n \in (a,b) \geq \mathbb{P}(X \in (a,b)$. I leave it to you to check.*

4.2.3 Comparison of different modes of convergence

Almost sure convergence is a strictly stronger notion than convergence in law, even if the random variables are defined on the same probability space. First, that convergence in law does not imply almost sure convergence is illustrated by the following example

- Let $X_1, X_2, \ldots$ be i.i.d $Ber(1/2)$ random variables defined on the same probability space. Then clearly $(X_n)_{n\geq 1}$ converges in law to a $Ber(1/2)$ random variable as for every $n \geq 1$, we have that $X_n \sim Ber(1/2)$. Yet we claim that X_n does not converge almost surely. This can be seen in many ways, for example we have that in the case of $Ber(1/2)$ random variables

  $$\{\omega : (X_n(\omega))_{n\geq 1} \text{ converges}\} = \{\omega : X_n(\omega) = X_m(\omega) \text{ for all } m, n \text{ large enough}\}.$$

  I leave it to you to argue that these events are measurable (see also the exercise sheet). Now, define $E_n = \{\omega : X_k(\omega) \text{ is constant for } k \in [2^n, 2^{n+1}]\}$. If X_n converges, then at the very least it has to be constant on infinitely many of these intervals, thus

  $$\mathbb{P}((X_n)_{n\geq 1} \text{ converges}) \leq \mathbb{P}(\text{infinitely many } E_n \text{ happen}).$$

However, $\mathbb{P}(E_n) = \frac{2}{2^{2^n}}$ and thus $\sum_{n\geq 1}\mathbb{P}(E_n) < \infty$. In particular by Borel-Cantelli I we see that almost surely only finitely many of the events E_n happen end hencet not only we don't have almost sure convergence, instead

$$\mathbb{P}(\{\omega \in \Omega : (X_n(\omega))_{n\geq 1} \text{ does not converge}) = 1.$$

We now prove the other direction:

Proposition 4.10 (Almost sure convergence implies convergence in law)**.** *Let $X_1, X_2, \dots$ be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then if $(X_n)_{n\geq 1}$ converge almost surely, they also converge in law.*

Proof. The proof is based on the following claim:

Claim 4.11. *Suppose $X_1, X_2, \dots$ converge almost surely to X. Then for every $\epsilon > 0$, we have that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.*

Before proving the claim, let us see how it implies the proposition. Let x be a continuity point for F_X. Then both

$$F_X(x) = \lim_{m\to\infty} F_X(x - 1/m) = \lim_{m\to\infty} F_X(x + 1/m).$$

By the claim for every $m \in \mathbb{N}$, for n large enough it holds that $\mathbb{P}(|X_n - X| > 1/m) < 1/m$.

Notice further that

$$\{X_n \leq x\} \cap (X > x + 1/m)\} \subseteq \{|X - X_n| > 1/m\}.$$

Thus writing

$$F_{X_n}(x) = P(X_n \leq x) = \mathbb{P}((X_n \leq x) \cap (X \leq x + 1/m)) + \mathbb{P}((X_n \leq x) \cap (X > x + 1/m))$$

we can bound

$$F_{X_n}(x) \leq F_X(x + 1/m) + \mathbb{P}(|X - X| > 1/m) < F_X(x + 1/m) + 1/m.$$

Using a similar inequality for the other direction, we obtain that for every $m \in \mathbb{N}$, for all n large enough.

$$F_X(x - 1/m) - 1/m < F_{X_n}(x) < F_X(x + 1/m) + 1/m.$$

Taking first $n \to \infty$ and then $m \to \infty$, we obtain that $\lim_{n\geq 1} F_{X_n}(x) = F_X(x)$ and thus deduce the convergence in law of X_n to X.

It remains to prove the claim.

Proof of Claim. Fix some $\epsilon > 0$. Then

$$\{(X_n)_{n\geq 1} \to X\} \subseteq \{|X_n - X| < \epsilon \text{ for all large enough } n\} = \bigcup_{m\geq 1} E_m.$$

[14] with $E_m = \{\forall n \geq m : |X_n - X| < \epsilon\}$. Notice that these events are nested, i.e. $E_m \subseteq E_{m+1}$, as there are less conditions imposed by the latter. As $\mathbb{P}(\{(X_n)_{n\geq 1} \to X\}) = 1$ we get that

$$1 = \mathbb{P}(\bigcup_{m\geq 1} E_m) = \lim_{m\to\infty} \mathbb{P}(E_m).$$

But now $\mathbb{P}(|X_n - X| > \epsilon) \leq 1 - \mathbb{P}(E_n)$ and thus the claim follows. □

[14] In case you have trouble seeing what's happening, I recommend writing out everything using ω, e.g. $\{\omega : (X_n(\omega))_{n\geq 1} \to X(\omega)\} \subseteq \{\omega : |X_n(\omega) - X(\omega)| < \epsilon \text{ for all } n \geq n(\omega)\}$ etc.

□

In fact, in the claim above we introduced another notion of convergence that is often used: convergence in probability.

Definition 4.12 (Convergence in probability)**.** *One says that a sequence of random variables $X_1, X_2, \dots$ defined on the same probability space converge to X in probability if and only if for every $\epsilon > 0$ we have that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.*

The proof above then gives us the following implications:

- Convergence in probability implies convergence in law.
- Almost sure convergence implies convergence in probability.

We already saw that convergence in law doesn't imply almost sure convergence, but in fact stronger converses are true:

Exercise 4.3. *By considering the sequence of i.i.d. $Ber(1/2)$ random variables, or otherwise, prove that convergence in law does not imply convergence in probability.*

Now, let X_n be a random variable taking value 0 with probability $1 - 1/n$ and value 1 with probability $1/n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Prove that $(X_n)_{n\geq 1}$ converges to 0 in probability. Further, show that if X_n are mutually independent, then they do not converge to 0 almost surely. Does this remain true when X_n are not mutually independent?

There are in fact even further notions of convergence, but we will leave them to your further courses. You might already ask though, why should we care about so many different notions? The difference between almost sure convergence and convergence in law is maybe more intuitive and was already explained above. To recall, in the case of almost sure convergence we really look at the convergence of a sequence of numbers for each $\omega \in \Omega$; in the case of convergence in law, we look at the convergence of the respective probability laws, via e.g. their c.d.f-s. In the latter case the random variables don't need to defined on the same probability space. But why do we need this third notion of convergence in probability?

First, we saw it enter rather naturally when comparing almost sure convergence and convergence in law. Second, almost sure convergence is often a too strong notion, as illustrated in the exercise above. And third, convergence in probability is often much easier to work with than almost sure convergence, as one can work with events for fixed $n \in \mathbb{N}$. Finally, convergence in probability gives naturally rise to a very useful metric structure on random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, where there is no topology on the space of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that convergence in this topology would correspond to almost sure convergence! (See the non-examinable section of the exercise sheet.) So maybe in fact convergence in probability is natural and not the a.s. convergence? We will come back to this shortly, but of course this is only a meta-mathematical question, so let us for now push forward with actual mathematics.

4.3 Weak and Strong law of large numbers

Let us start by stating both theorems. Roughly, they both say that if you repeat the same random experiment independently n times to obtain i.i.d random variables $X_1, X_2, \dots, X_n$ then as $n \to \infty$ the average of X_i converges to the expectation of X_1. This is quite remarkable that the distribution of the variables does not play any larger role in this limit - only the

integrability and the expectation matter. Both of these theorems are related to so called ergodic theorems, which roughly link the temporal (here n) and spatial (here $\mathbb{E}$) averages.

Theorem 4.13 (Weak law of large numbers (WLLN)). *Let $X_1, X_2, \dots$ be i.i.d. integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then as $n \to \infty$, we have that*

$$\mathbb{P}(|\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}X_1| > \epsilon) \to 0,$$

i.e. the sequence $S_n = \frac{\sum_{i=1}^n X_i}{n}$ converges in probability to $\mathbb{E}X_1$.

The stronger version is as follows:

Theorem 4.14 (Strong law of large numbers (SLLN)). *Let $X_1, X_2, \dots$ be i.i.d. integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then we have that*

$$\mathbb{P}(\frac{\sum_{i=1}^n X_i}{n} \text{ converges to } \mathbb{E}X_1) = 1,$$

i.e. the sequence $S_n = \frac{\sum_{i=1}^n X_i}{n}$ converges almost surely to $\mathbb{E}X_1$.

As almost sure convergence implies convergence in probability, we see that the second result is indeed stronger. What is the difference of these two theorems?

The weak law says that if you do independent experiments $X_1, X_2, \dots$ and look at the average outcome of the first n of them with n large, then the random variable you obtain is very close to the constant $\mathbb{E}X_1$. Indeed, for evert $\epsilon > 0$, if you do sufficiently many experiments then the probability that this random average differs from $\mathbb{E}X_1$ by more than ϵ is less than, say, 0.00001. WLLN doesn't however say how the consecutive averages behave for a fixed sequence of outcomes.

The strong law on the other hand says exactly that almost surely for any sequence of outcomes, if you look at the average of the first n outcomes and then increase n, these averages converge to $\mathbb{E}X_1$. SLLN doesn't look only at snaphots for fixed n, but describes for every sequence the evolution of averages.

In both cases, both the integrability and independence are important. You will think about the role of integrability on the example sheet; for necessity of some independence you can consider the case $X_1 = X_2 \dots$. Then the average of $X_1, \dots, X_n$ is just equal to X_1 and has no reason to converge to a constant. In general, LLN also holds under some weak dependence, but this is out of scope here.

So why do we state the weak law at all? The reason is that it is considerably easier to prove! In fact, although we prove both theorems under weaker hypothesis then stated, the full case of the WLLN could be proved with not much more effort, whereas proving the sharp version of SLLN is already not that easy.

Proof of WLLN for i.i.d. random variables with bounded variance. Suppose that $\mathbb{E}X_1^2 < C$. In this case $\mathbb{E}(|S_n - \mathbb{E}X_1|^2)$ is well defined and we can write

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = \sum_{i,j \le n} n^{-2} \mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1).\right]$$

But $X_1, X_2, \dots$ are mutually independent and $\mathbb{E}X_j = \mathbb{E}X_1$. Thus we see that if $i \neq j$, then $E\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1).\right] = 0$. Hence

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = n^{-2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = n^{-1}C \to 0$$

as $n \to \infty$. By Chebyschev inequality we have that

$$\mathbb{P}(|S_n - \mathbb{E}X_1| > \epsilon) \leq \epsilon^{-1}n^{-1}C \to 0$$

and and WLLN for random variables with bounded variance follows.

□

Notice that we didn't really use independence here - just the fact that $Cov(X_i, X_j) = 0$ for all i, j! Moreover, we also didn't use that the variables were i.i.d., we just used that for all $i \geq 1$, we have that $\mathbb{E}X_i^2 < C$ - i.e. the variances are uniformly bounded. We prove SLLN under even stronger hypothesis. Notice how the proofs start similarly, but that there is an extra step in the end.

Proof of SLLN for i.i.d. random variables with $\mathbb{E}X_i^4 < C$. Suppose that for some $C > 0$, we have $\mathbb{E}X_i^4 < C$. By increasing the value of C (but not the number of notations!) we can assume that for this C also $\mathbb{E}(X_i - \mathbb{E}X_i)^4 < C$ for some $C > 0$ (why?). In this case $\mathbb{E}(|S_n - \mathbb{E}X_1|^4)$ is well defined and we can write

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) = \sum_{i,j,k,l \leq n} n^{-4}\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)(X_k - \mathbb{E}X_1)(X_l - \mathbb{E}X_1)\right].$$

Notice that if one index appears only once (e.g. we have $i = 1$, $j = k = l = 2$), then as in the proof of WLLN

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)(X_k - \mathbb{E}X_1)(X_l - \mathbb{E}X_1).\right] = 0$$

because of independence and the fact that $\mathbb{E}X_1 = \mathbb{E}X_i$. Hence

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) = n^{-4}\sum_{i,j \leq n} \mathbb{E}\left[(X_i - \mathbb{E}X_1)^2(X_j - \mathbb{E}X_1)^2\right].$$

By Cauchy-Schwarz,

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)^2(X_j - \mathbb{E}X_1)^2\right] \leq \mathbb{E}\left[(X_i - \mathbb{E}X_1)^4\right] \leq C.$$

Thus

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) \leq Cn^{-2}$$

and by Markov's inequality

$$\mathbb{P}(|S_n - \mathbb{E}X_1| > n^{-1/8}) = \mathbb{P}(|S_n - \mathbb{E}X_1|^4 > n^{-1/2}) \leq \frac{\mathbb{E}|S_n - \mathbb{E}X_1|^4}{n^{-1/2}} \leq Cn^{-3/2}.$$

Thus when we define $E_n = \{|S_n - \mathbb{E}X_1| > n^{-1/8}\}$, then $\sum_{n \geq 1}\mathbb{P}(E_n) < \infty$. Hence by Borel-Cantelli lemma applied to the evens E_n, we see that almost surely only finitely many of them occur. But this means that almost surely, $\{|S_n - \mathbb{E}X_1| \leq n^{-1/8}\}$ for all but finitely many n, implying that S_n converges to $\mathbb{E}X_1$ almost surely. □

Remark 4.15. *Again, notice that in this proof we don't use the fact that X_i are identically distributed, we only use that $\mathbb{E}X_i^4 < C$. You should ask yourself: why did we need in this proof the 4-th moment, and in WLLN only the 2-nd moment?*

These two theorems are the basis for the so called frequentist approach to probability. Indeed, we have the following immediate corollary (recall how annoying it was to prove it on the first example sheet!)

Corollary 4.16. *Let $E_1, E_2, \dots$ be independent events with $\mathbb{P}(E_i) = p$. Then $\frac{\#\{(E_i)_{i\leq n} \text{ that occur}\}}{n}$ converges almost surely to p.*

Proof. This follows directly from SLLN by noticing that $1_{E_1}, 1_{E_2}, \dots$ are i.i.d integrable random variables of expectation p. □

So for example, if you have a coin with unknown probability p of obtaining heads. Then to determine p, you start tossing the coin, and look at the average number of heads you get in n trials, and then SLLN says that with probability one these averages converge to p! It's an interesting question to see 'how fast' it converges to p, i.e. how precisely you might know p after, say, 25 or 100 throws...Although answering this question will be outside of the scope of this course, it is in certain settings related to the Central limit theorem, that describes the fluctuations of the average around its mean and is described in the next section.

4.4 Central limit theorem

The final result of the course is the Central Limit Theorem (CLT).

Theorem 4.17 (Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables of finite variance σ^2 defined on the same probability space. Then $n^{-1/2}\sum_{i=1}^n (X_i - \mathbb{E}X_i)$ converges in law to $N(0, \sigma^2)$.*

This is a remarkable result, saying that if we add up independent random variables of finite variance we always end up with the same distribution - the Gaussian distribution! This is the reason why at least heuristically measurement errors in physics look like Gaussians - they are sums of small independent contributions, or why Gaussians come up when looking at distributions of say heights in a population. This phenomenon that individual properties of the random variables X_i only influence the limiting law by a few parameters - the expectation, variance - is sometimes called universality.

In the CLT both the assumption of finite variance and independence are crucial: you will see an example about moment conditions on the exercise sheet. To see that without independence CLT could fail consider for example the case of $X_1 = X_2 = \dots$. Then $n^{-1/2}\sum_{i=1}^n X_i = n^{1/2}X_1$ which certainly does not converge and has no reason to be a Gaussian. Whereas the condition of independence can be relaxed somewhat, there has to be a fair amount independence to guarantee that the effect of each X_i on the sum is negligible!

We can now for example deduce very easily the following result, which has come up as a technical exercise in a non-examinable section of the exercise sheet:

Corollary 4.18. *Let X_n be a $Bin(n, p)$ random variable. Then $\frac{X_n - np}{\sqrt{n}}$ converges in law to a Gaussian of variance $\sigma^2 = p(1-p)$.*

Proof. We can write $X_n - np = \sum_{i=1}^n (Y_i - \mathbb{E}Y_i)$, where Y_i are i.i.d. $Ber(p)$ random variables. Then by the CLT, we have that $\frac{X_n - np}{\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)}{\sqrt{n}}$ converges to a Gaussian of variance $\mathrm{Var}(Y_i) = p(1-p)$.

□

We will again prove CLT under further hypothesis, in particular we assume $\mathbb{E}|X_i|^3 < \infty$. There are many different proofs of this theorem, all explaining different facets of the theorem. The one we follow is based on the following idea:

- The sum of Gaussians is always a Gaussian. Moreover, if $X_1, X_2, \dots$ are i.i.d. standard Gaussians, then $n^{-1/2}\sum_{i=1}^n X_i$ has again the same law! (Check!) Now, given general variables Y_i, we will just try to swap them one by one for Gaussian random variables of the same mean and variance. We always make an error, but if we can control the cumulative error, then we are done. This is exactly what we will do!

This key step is encapsulated in the following proposition, that we again prove under further hypothesis:

Proposition 4.19 (Lindeberg Exchange Principle)**.** *Let $X_1, X_2, \dots$ be i.i.d. zero mean unit variance random variables and with $\mathbb{E}|X_i|^3 < \infty$. Let further Y be a standard Gaussian. Define $S_n := n^{-1/2}\sum_{i=1}^n X_i$. Then for every $f : \mathbb{R} \to \mathbb{R}$ smooth with uniformly bounded derivatives up to third order, we have that $|\mathbb{E}f(S_n) - \mathbb{E}f(Y)| \to 0$ as $n \to \infty$.*

Before proving the proposition, let us see how to deduce CLT from this proposition. The idea is as follows: we saw already that knowing $\mathbb{E}g(X)$ for all continuous bounded g determines the distribution of X. In fact, this would be also true if we only assumed it to hold for smooth g! Moreover, convergence in law can be also deduced from knowing the convergence of $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for all g that are smooth and bounded, and have further conditions on derivatives. The idea is similar to Proposition 3.10 - we approximate indicator functions $1_{X<x}$ via smooth functions and thus obtain the convergence the c.d.f at all continuity points.

Lemma 4.20. *Suppose that $X, X_1, X_2, \dots$ are random variables. If for all smooth bounded g with uniformly bounded derivatives up to 3rd order we have $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ as $n \to \infty$, then X_n converge in law to X.*

Proof. This is on the exercise sheet. □

Proof of CLT:. Given random variables X_i of variance σ^2, we have that $\widehat{X}_i := \frac{X_i - \mathbb{E}X_i}{\sigma}$ are zero mean and unit variance. Thus we can apply Proposition 4.19 and Lemma 4.20 to deduce that $n^{-1/2}\sum_{i=1}^n \widehat{X}_i$ converges to a standard Gaussian. But now multiplying everything by σ gives the CLT. □

It remains to prove the proposition.

Proof of Lindeberg Exchange Principle: Let Y and $Y_1, Y_2 \dots$ be i.i.d. standard Gaussians. For $k \geq 1$, write

$$S_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k}^n Y_i}{n^{1/2}}.$$

Notice that $S_{n,n+1} = S_n$ and $S_{n,1} = n^{-1/2}\sum_{i=1}^n Y_i \sim N(0,1)$. Thus we can write

$$f(S_n) - f(Y) = \sum_{k=1}^n f(S_{n,k+1}) - f(S_{n,k}). \tag{4.1}$$

Our aim will be to control each individual summand. To do this write further

$$S^0_{n,k} := \frac{\sum_{i=1}^{k-1} X_i + \sum_{i=k+1}^n Y_i}{n^{1/2}},$$

where we have omitted the k-th term altogether.

By third-order Taylor's approximation we can write a.s.

$$f(S_{n,k+1}) = f(S^0_{n,k}) + \frac{X_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{X_k^2}{2n} f''(S^0_{n,k}) + \frac{X_k^3}{6n^{3/2}} f'''(x_1),$$

with x_1 between $S_{n,k+1}$ and $S^0_{n,k}$ and similarly

$$f(S_{n,k}) = f(S^0_{n,k}) + \frac{Y_k}{n^{1/2}} f'(S^0_{n,k}) + \frac{Y_k^2}{2n} f''(S^0_{n,k}) + \frac{X_k^3}{6n^{3/2}} f'''(x_2).$$

Taking expectations, as X_k is independent of $S^0_{n,k}$, we see that

$$\mathbb{E} f(S_{n,k+1}) = \mathbb{E} f(S^0_{n,k}) + \mathbb{E}\frac{X_k}{n^{1/2}} \mathbb{E}(S^0_{n,k}) + \mathbb{E}\frac{X_k^2}{2n} \mathbb{E} f''(S^0_{n,k}) + \mathbb{E}\left(\frac{X_k^3}{6n^{3/2}} f'''(x_1)\right).$$

Using further that X_k has mean zero, unit variance and $\mathbb{E}|X_k|^3 < \infty$, we obtain that

$$\mathbb{E} f(S_{n,k+1}) = \mathbb{E} f(S^0_{n,k}) + \frac{1}{2n} \mathbb{E} f''(S^0_{n,k}) + E_r,$$

with $|E_r| \leq \mathbb{E}\left(\frac{|X_k|^3}{6n^{3/2}} |f'''(x_1)|\right) = O(n^{-3/2})$ as by assumptions on f, we have that $|f'''(x)| < C$ and $\mathbb{E}|X_k|^3 < \infty$. Similarly, as also Y_k is independent of $S^0_{n,k}$, we obtain that

$$\mathbb{E} f(S_{n,k}) = \mathbb{E} f(S^0_{n,k}) + \frac{1}{2n} \mathbb{E} f''(S^0_{n,k}) + \widehat{E}_r,$$

with $|\widehat{E}_r| = O(n^{-3/2})$. Thus $|\mathbb{E} f(S_{n,k+1}) - \mathbb{E} f(S_{n,k})| = O(n^{-3/2})$. By the triangle inequality we obtain

$$|\mathbb{E}\left(f(S_n) - f(Y)\right)| \leq \sum_{k=1}^{n} |\mathbb{E} f(S_{n,k+1}) - \mathbb{E} f(S_{n,k})| = O(n^{-1/2})$$

and the proposition follows. □

I wish there was more...but that's all!