BASIC PROBABILITY THEORY 2024

JUHAN ARU

[1]

[1]Version of 2024. All kinds of feedback, including smaller or bigger typos, is appreciated - juhan.aru@epfl.ch. This is a third version of the notes. In writing previous version of 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

This course is about probability theory: the mathematical framework for formalising our questions about random phenomena, and their mathematical study.

When we want to describe a random phenomena in the real world, we build a mathematical model. This is itself an interesting process and a good model involves lots of well-chosen simplifications and righteous choices - e.g. to model a coin toss, we usually discard the possibility of it landing on the edge, or without further knowledge we consider the heads and tails equiprobable, although that may not be the case for example already because of different weight distributions. But this all is not the topic of this course.

In this course we will study the general mathematical framework and formulation of such models and then discuss the mathematical tools necessary and useful to study such models. Hopefully we also have some time to discuss some interesting models.

Section 1

Basic framework

In this chapter we discuss some basic but important notions of probability theory:

- Probability space
- Random variables
- Independence

1.1 Probability space

Our first aim is to motivate the notion of a probability space or a probabilistic model. To do this let us consider two examples:

(1) A random number with values in $\{1, 2, \ldots, 12\}$ e.g. something that comes from a lottery.
(2) Describing the weather in Lausanne the day after.

In describing these two random phenomena we will still use everyday vocabulary / intuitions. Thereafter we will give the mathematical definitions that will fix the vocabulary for the rest of the course.

(1) Random number. To describe a random number mathematically, we basically need three inputs:

- The set of all possible outcomes: in this case $\Omega = \{1, 2, 3, \ldots, 12\}$
- The collection of yes / no questions that we can answer about the actual outcome, i.e. this random number. For example:
  - Is this number equal to 3?
  - Is this number even?
  - Is this number smaller than 4?

  To each of these questions we put in correspondence the subset of outcomes that corresponds to the answer yes: $\{3\}$, $\{2, 4, 6, 8, 10, 12\}$ or $\{1, 2, 3\}$ respectively. We call each such subset an event.
- Finally, to each event $E \subseteq \Omega$ we want to assign a numerical value $\mathbb{P}(E) \in [0, 1]$ that we call the probability. This should correspond to the fraction of times an event happens if the random number is given to us many times, e.g. if the lottery is played many times. [2]

Here the set of possible outcomes was easy and directly given by the problem. Also it is natural to assume that each subset $E \subseteq \Omega$ is an event - or in other words that for each E we can ask the question: is the number in E? This means that the we can take the collection of events to correspond to all subsets of Ω.

Determining the probability really depends on what we want to model - e.g. if we are trying to model the lottery, we may assume that all numbers are equally likely and then we

[2] In fact, one uses probabilistic models also to model phenomena that only happens once. In that case probability measures somehow our degree of belief.

rediscover the model from high-school: we set $\mathbb{P}(E) = |E|/|\Omega|$. However, if we wanted to describe the sum of two dice, we would need to choose the numbers $\mathbb{P}(E)$ very differently! [3]

Now, if we want our model to correspond to the intuitive notion of probability and to predict the fraction of repeated experiments, then these choices are not quite free - we need to add some constraints. E.g. we cannot put in an arbitrary function $\mathbb{P}$: indeed, if we have two events $E_1 \subseteq E_2$ then we should have $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$ as every time E_1 happens, also E_2 happens. We should also have $\mathbb{P}(\Omega) = 1$ as something always happens and $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ if E and F are disjoint (why?). Of course not all these constraints are distinct - some might imply others and when giving the definition of a probability space below we will purify and choose only some conditions that will then mathematically imply all the others.

(2) Weather in Lausanne the day after. We would again want to make the three decisions, but here the task is already harder at the very first step. What should be the state space? A natural state space could probably be all possible microscopic states of the atmosphere up to 20km of height over Lausanne...but here we of course have many arbitrary choices - why 20 km, how wide should we look over Leman etc? And in any case, any natural state would be impossibly complicated!

Luckily, we do not actually need to worry about it - we only have to assign probabilities to all the events in the collection of events. And we have some freedom in choosing this collection events - it could be determined by our possibility to measure the states, e.g. we are able to measure the temperature up to some precision, or the density of CO_2 or water molecules to some precision and this determines some subsets of the state space.

However, as with the probability function, also for the collection of events there are some natural consistency conditions: we would assume that if one can observe if event E happened, we should be also able to measure if its complement E^c happened. Or if we are able to say if E happened or if F happened, we should be able to say if one of the two happened - i.e. $E \cup F$ should also be an event. And in fact it comes out that this is all we need!

Naturally, setting up probabilities for this model is also horribly complicated - there are no natural symmetry assumptions like the one we used for the uniform distribution. Also, even the best physicist in the world will not be able to describe the natural probability distribution of all microscopic states of the atmosphere, especially as it will heavily depend on what is happening just before! Thus, our only choice basically is to try to somehow use the combination of our knowledge about atmospheric processes together with our observations from history to set up some estimates for the model; and then naturally we will try to improve it with every next day. Luckily, this difficult task is not up to us but rather the office of meteo and the statisticians!

Remark 1.1. *Finally, before giving the mathematical definitions, let us stress again that all three components of the model - the sample space, the set of events and their probabilities - are inputs that we choose to build our model. When trying to model a real world phenomena we usually make simplifications for each of these choices. For example, for the coin toss we use only two outcomes: heads and tails, although theoretically edge is also possible. Also, we usually set probabilities to be a half, although that is not exactly true either.*

[3] See Exercise sheet 1.

1.2 Mathematical definition of a probability space

We are now ready to use our mathematical filter and give a mathematical definition of a probability space. In fact, we first use the mathematical purifier to come up with a definition in the restricted setting where Ω is a finite set, and then generalize it further.

Indeed, the discussions above lead us directly to:

Definition 1.2 (Elementary probability space, Kolmogorov 1933)**.** *An elementary probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- *Ω is a finite set, called the state or 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 $A_1, A_2, \in \mathcal{F}$, then also $A_1 \cup A_2 \in \mathcal{F}$.*

  $\mathcal{F}$ is called the collection of events and any $A \in \mathcal{F}$ is called an event.
- *And finally, we have a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying $\mathbb{P}(\Omega) = 1$ and additivity for disjoint sets: if $A_1, A_2 \in \mathcal{F}$ are pairwise disjoint, then*

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

  This function $\mathbb{P}$ is called the probability

Notice that some properties discussed above, like the fact that for events $E_1 \subseteq E_2$, we have $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$, follow directly from the definition.[4]

Now, most phenomena in the real world can be described by finite sets just because we are able to measure things only to a finite level of precision. However, like the notion of a continuous or differentiable function helps to simplify our mathematical descriptions of reality and thus improve our understanding, continuous probability spaces also make the mathematical descriptions neater, simpler and thereby also make it easier to understand and study the underlying random phenomena.

Some natural examples where infinite sample spaces come in: an uniform point on a line segment e.g. stemming from breaking a stick into several pieces; the position on the street where the first raindrop of the day falls; or the space of all infinite sequences of coin tosses. In all these cases the mathematically natural state space is even uncountable. Countably infinite state spaces can also come up: for example if we want to model the first moment that a repeated coin toss comes up heads, the value might be $1, 2, 3$ or with very very small probability also 10^{10}, so a natural state space would contain all natural numbers.

So let us state the general definition:

Definition 1.3 (Probability space, Kolmogorov 1933)**.** *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- *Ω is a set, called the state or 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 $A_1, A_2, \cdots \in \mathcal{F}$, then also $\bigcup_{n\geq 1} A_n \in \mathcal{F}$.*

  $\mathcal{F}$ is called the collection of events or a σ-algebra and any $A \in \mathcal{F}$ is called an event.

[4] See Exercise sheet 1.

- *And finally, we have a function* $\mathbb{P} : \mathcal{F} \to [0,1]$ *satisfying* $\mathbb{P}(\Omega) = 1$ *and additivity for disjoint sets: if* $A_1, A_2, \cdots \in \mathcal{F}$ *are pairwise disjoint,*

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

This function $\mathbb{P}$ *is called the probability*

Notice the only differences are 1) we do not assume Ω to be finite 2) we assume that the set of events is stable under countable unions 3) we assume also the additivity of the probability under countable unions.

Exercise 1.1. *Show that each elementary probability space is a probability space.*

In fact probability spaces are an example of a general notion of measure spaces - probability spaces are just measure spaces with total mass equal to 1.

Definition 1.4 (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* $A_1, A_2, \cdots \in \mathcal{F}$*, 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* $A_1, A_2, \cdots \in \mathcal{F}$ *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.*

Geometrically we interpret:

- Ω as our space of points
- $\mathcal{F}$ as the collection of subsets for which our notion of volume can be defined
- μ our notion of volume: it gives each measurable set its volume.

It is important to make this link to measure theory as many properties of probability spaces directly come from there. Yet it is also good to keep in mind that probability theory is not just measure theory - as M. Kac has put it well, 'Probability is measure theory with a soul' and we adhere to this philosophical remark.

Remark 1.5. *You should compare the definition of a probability space / measure space with the definition of a topological space: there also we use a collection of subsets with certain properties to attach structure to the set. A question you should ask is: why do we use exactly countable unions and intersections for the events, and not finite or arbitrary?*

1.3 Some basic properties of probability spaces

We start by a few small remarks about the definition of a probability space:

Remark 1.6. *It is worth considering why ask for countable stability of the σ-algebra or countable additivity of the probability measure. Whereas this is more a meta-mathematical question, it is good to keep it in mind throughout the course. Let us here just offer two simple observations.*

First, countable sums naturally come up when we take limits of finite sums. In fact, countable additivity can be seen to be equivalent to certain form of continuity for the probability measure (see below).

Second, allowing for arbitrary unions leads easily to power-sets, and sums of uncountably many positive terms cannot be finite (see the exercise sheet).

Exercise 1.2. *Show that the countable additivity in the axioms of a probability space can be replaced with finite additivity plus the following statement: for any decreasing sequence of events $E_1 \supseteq E_2 \supseteq E_3 \dots$ we have that $\mathbb{P}(\cap_{i=1}^n E_i) \to 0$ as $n \to \infty$.*
$\star$ Does this hold in a general measure space?

Also we would like to remark another setting that explains well the usefulness of σ-algebras:

Remark 1.7. *Often 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 (which is helpful as then $\mathbb{P}$ will only need to be extended not redefined), we can consider a sequence of σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \dots$ called a filtration - each day we can ask some more yes/no questions because we already for example know what happened on the previous day and maybe also have learned something new. All possible information is contained in the power set $\mathcal{P}(\Omega)$.*

Probability spaces are usually classified in two types:

Definition 1.8 (Discrete and continuous probability spaces)**.** *Probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ with a countable sample space Ω are called discrete probability spaces and those with an uncountable sample space are called continuous probability spaces.*

In this course we will mainly work with discrete probability spaces, as they are technically easier to deal with. However, continuous probability spaces come up naturally and we won't be able to fully avoid them either.

Their technical difference can be summoned in the following proposition, whose non-examinable proof will be left for enthusiasts.

Proposition 1.9. *Let Ω be countable and $\mathcal{F}$ a $\sigma-$algebra on Ω. Then one can find disjoint events $E_1, E_2, \dots \in \mathcal{F}$ such that for every $E \in \mathcal{F}$ we can express $E = \cup_{i \in I_E} E_i$.*

Essentially, this says that for every discrete probability space it suffices to determine $\mathbb{P}(E_i)$ for a countable collection of disjoint sets E_i, and thereafter for every other set E we can use countable additivity to extend $\mathbb{P}$. Notice that this means it is first easy to check whether a given $\mathbb{P}$ satisfies all the axioms and even more importantly it is easy to check when two probability measures are equal.

For continuous probability spaces this does not necessarily hold - the useful σ-algebras are usually more complicated. To examplify why one doesn't want to necessarily use the power-set consider the following proposition, whose proof is in the appendix and relies on the axiom of choice:

Proposition 1.10. *There is no probability measure $\mathbb{P}$ on $([0,1], \mathcal{P}([0,1]))$ that is invariant under shifts, i.e. such that for any $A \in \mathcal{P}([0,1]), \alpha \in [0,1)$, we have that $\mathbb{P}(A + \alpha \mod 1) =$*

$\mathbb{P}(A)$, *where here we denote* $A + \alpha \mod 1 := \{a + \alpha \mod 1 : a \in A\}$, *the set obtained by shifting* A *by* α, *modulo* 1.

In fact, it comes out that the only way to remedy this situation is to make the relevant $\sigma-$algebra smaller. We would still want to be able to answer yes or no to questions like: is my random number equal to $\{x\}$ or is it in an interval (a, b)? Thanks to the fact that we have only countable additivity, this does not imply that our σ-algebra would need to be the power-set. And thanks to the properties of the $\sigma-$algebras, we can always construct at least some $\sigma-$algebra containing all our favourite sets - see the exercise sheet.

Let us now state some immediate consequences of the definitions about the $\sigma-$algebras and the probability measures:

Lemma 1.11 (Stability of the $\sigma - algebra$)**.** *Consider a set* Ω *with a* σ*-algebra* $\mathcal{F}$.

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

Proof of Lemma 1.11. 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 (1) 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 (3), 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 $k \geq n + 1$.

Point (2) is left as an exercise. □

In a similar vein, the basic conditions on the measure give rise to several natural properties:

Proposition 1.12 (Basic properties of a probability measure)**.** *Consider a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $A_1, A_2, \cdots \in \mathcal{F}$. *Then*

(1) For any $A \in \mathcal{F}$, *we have that* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(2) For any $n \geq 1$, *and* $A_1, \ldots, A_n$ *disjoint, we have finite additivity*

$$\mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_n).$$

In particular if $A_1 \subseteq A_2$ *then* $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.
(3) If for all $n \geq 1$, *we have* $A_n \subseteq A_{n+1}$, *then as* $n \to \infty$, *it holds that* $\mathbb{P}(A_n) \to \mathbb{P}(\bigcup_{k\geq 1} A_k)$.
(4) We have countable subadditivity (also called the union bound): $\mathbb{P}(\bigcup_{n\geq 1} A_n) \leq \sum_{n\geq 1} \mathbb{P}(A_n)$.
(5) If for all $n \geq 1$, *we have* $A_n \supseteq A_{n+1}$, *then as* $n \to \infty$, *it holds that* $\mathbb{P}(A_n) \to \mathbb{P}(\bigcap_{k\geq 1} A_k)$.

Proof. Properties 1, 4 and second part of 2 were included in the Exercise sheet 1. The first part of property 2 follows like in the lemma above by taking $A_{n+1} = A_{n+2} = \cdots = \emptyset$ and using countable additivity.

So let us prove property 3: Write $B_1 = A_1$ and for $n \geq 2$, $B_n = A_n / A_{n-1}$. Then B_n are disjoint, $\bigcup_{n=1}^{N} B_n = A_N$ and $\bigcup_{n\geq 1} B_n = \bigcup_{n \geq 1} A_n$.

Thus by countable additivity

$$\mathbb{P}(\bigcup_{i\geq 1} A_i) = \mathbb{P}(\bigcup_{i \geq 1} B_i) = \sum_{i \geq 1} \mathbb{P}(B_i)$$

But $\mathbb{P}$ is non-negative, so

$$\sum_{i \geq 1} \mathbb{P}(B_i) = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(B_i)$$

By countable additivity again

$$\sum_{i=1}^{n} \mathbb{P}(B_i) = \mathbb{P}(\bigcup_{i=1}^{n} B_n) = \mathbb{P}(A_n)$$

and (2) follows.

□

1.4 Random variables

In fact when studying a random phenomena we certainly don't want to restrict ourselves to yes and no questions. For example, in our model of a random number among $\{1, 2, \ldots, 12\}$ the natural question is not 'Is this number equal to 5?' but rather 'What number is it?'. Similarly in our example of discussing the weather, it is more natural to ask 'What is the temperature?', 'How much rain will there be in the afternoon?'?

Such numerical observations about our random phenomena will be formalised under the name of random variables. In essence they give a number for each state and thus as such are just functions $X : \Omega \to \mathbb{R}$ from the state-space to real numbers. However, we may not want to include all such functions for consistency reasons. Indeed, we want to be able to ask yes / no questions about our random numbers, e.g. Is the random number equal to 3? Is the temperature more than 18? But again the answer yes / no corresponds to certain subsets of states in the universe and as such should be events in our model. Thus there is a link between the collection of events, and and the collection of functions that can act as random variables. Let us without further give the general definition:

Definition 1.13 (Random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We call a function $X : \Omega \to \mathbb{R}$ a random variable if for every interval (a, b) the set $X^{-1}((a, b)) := \{\omega \in \Omega : X(\omega) \in (a, b)\}$ is an event on the original probabiliuty space, i.e. belongs to $\mathcal{F}$.*

There is a simplification in the case of discrete probability spaces:

Lemma 1.14 (Random variables on discrete probability spaces)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a discrete probability space. Then $X : \Omega \to \mathbb{R}$ is a random variable if and only if for every $y \in \mathbb{R}$ we have that $X^{-1}(\{y\}) \in \mathcal{F}$.*

Proof. This can be verified carefully from the definitions and will be on the exercise sheet. □

For the structurally minded the definition of a random variable might look somewhat arbitrary. And indeed, I have been hiding one piece of information - the natural collection of events on $\mathbb{R}$ that we alluded to a little bit already in the previous subsection. We will directly state it on $\mathbb{R}^n$.

Definition 1.15 (Borel σ-algebra)**.** *The smallest σ-algebra on $\mathbb{R}^n$ that contains all open boxes of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$ is called the Borel σ-algebra. We denote it by $\mathcal{F}_B$*

Remark 1.16. *In fact this definition is even more general: given any topological space (X, τ), the smallest σ-algebra containing all open sets is called the Borel σ-algebra. You will see on the exercise sheet that this more general definition reduces to the previous one in the case of $\mathbb{R}^n$ with its Euclidean topology.*

Based on this an equivalent, possibly more structural definition of a random variable is as follows: a function $X : \Omega \to \mathbb{R}$ is a random variable if the preimage of every set in the Borel $\sigma-$algebra under X is an event. [5]

An important notion that comes with random variables is its law:

Lemma 1.17 (The law of a random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable.*

Then there is a probability measure $\mathbb{P}_X$ induced on $(\mathbb{R}, \mathcal{F}_B)$ by defining $\mathbb{P}_X(F) := \mathbb{P}(X^{-1}(F)$ for every $F \in \mathcal{F}_B$. This probability measure $\mathbb{P}_X$ is called the law (or distribution) of a random variable X.

This is a lemma and not a definition as it needs to be proved that indeed $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{F}_B)$.

Proof of Lemma. We need to verify the axioms on a probability measure for a probability space:

- We have $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(\Omega) = 1$
- Similarly $\mathbb{P}_X(F) = \mathbb{P}(X^{-1}(F)) \in [0, 1]$ for all $F \in \mathcal{F}_B$
- Finally it remains to check countable additivity: let $F_1, F_2, \ldots$ be disjoint sets in $\mathcal{F}_B$. Then

  $$\mathbb{P}_X(\bigcup_{i\geq 1} F_i) = \mathbb{P}(X^{-1}(\bigcup_{i\geq 1} F_i)) = \mathbb{P}(\bigcup_{i\geq 1} X^{-1}(F_i)) = \sum_{i\geq 1} \mathbb{P}(X^{-1}(F_i)) = \sum_{i\geq 1} \mathbb{P}_X(F_i).$$

  Here we used the definition in the first and last equality, the properties of preimages in the second equality and the fact that $X^{-1}(F_i)$ are disjoint together with countable additivity in the third equality.

□

In words we showed that each random variable X induces a probability measure on the real numbers by just forgetting about the whole context and just concentrating on the number we see. For example in the case of weather in Lausanne, the temperature will give us a random variable and by just looking at its value and nothing else we have just a random real-valued number. Or more simply, if if we throw two fair coins and count the nunmber of heads, their sum will be a random variable that takes values in the set $\{0, 1, 2\}$. Thus the notion of the law of random variable gives us a way to compare random quantities arising in very different contexts.

Definition 1.18 (Equality in law)**.** *Let X, Y be two random variables defined possibly on different probability spaces. We say that X and Y are equal in law or equal in distribution, denoted $X \sim Y$ if for every $E \in \mathcal{F}_B$ we have that $\mathbb{P}_X(E) = \mathbb{P}_Y(E)$.*

[5] In measure theory such functions would be called measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{F}_B)$; notice the similarity with the definition of continuous functions in your topology course.

We stress that when looking at the law of random variable the context gets forgotten - we only concentrate on the numerical value and the initial probability space $(\Omega, \mathcal{F}, \mathbb{P})$ only helps to determine $\mathbb{P}_X$ but plays no role thereafter. This means that we can nicely connect different random phenomena between each other. For example the indicator functions of all events that have probability p, independently on which probability space they have been defined, have the same law. Or more concretely, for example the following random variables have the same law:

- Number of heads in two independent tosses
- Number of prime factors when we choose uniformly a number among $\{1, 2, 3, 4\}$.

In some sense a large part of this course will be about studying and describing probability laws of random variables.

Section 2

Conditional probability and independence

In general, if we learn something new about our random phenomena, this knowledge influences and often changes our predictions for the rest of the model.

- For example in the case of a uniform random number between 1 and 12, if someone tells you that this number is even, then the probability of seeing 1 will suddenly be 0, but the probability of seeing 2 will rise from 1/12 to 1/6.
- In the case of weather in Lausanne, if someone tells us that it rains the whole day, then it is less likely to also be above 35 degrees.

The aim of this section is to set up the vocabulary to talk about how the knowledge about some event or random variable influences the probabilities we should assign to other events. This leads us to talk about conditional probabilities and to discuss the case where events do not influence each other, giving rise to an important notion of probability theory called independence.

2.1 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 or successive steps in a random walk, or different links or edges in a random graph. In all these cases, the fact that one event has happened could easily influence the others. For example, if you want to model the financial markets tomorrow, it seems rather advisable to take into account what happened today. To talk about the change of probabilities when we have observed something, we introduce the notion of conditional probability:

Definition 2.1 (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 2.2. *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 $(\Omega, \mathcal{F})$, called the conditional probability measure given E.*

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 of an event might sometimes be similar to the initial probability (we will see more about this very soon), but it 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.

Or for a more senseful exercise consider the following:

Exercise 2.1 (Random walk and conditional probabilities)**.** *Consider the simple random walk of length n.*

- *What is the probability that the walk ends up at the point n at time n? Now, suppose that the first step was -1. What is the probability that the walk ends up at the point n at time n now?*
- *Suppose that n is even. What is the probability that the walk ends up at the point 0 at time n? Now, suppose that the first step was -1. What is the probability that the walk ends up at the point 0 at time n now?*

One also has to be very careful about the exact conditioning, as two similarly sounding conditionings can induce very different conditional probabilities. In general, we need to know something extra about the relation of two events to know how the probability of one changes when conditioned on the other.

There are some cases where these relations and thus conditional probabilities are easy:

- When $E \subseteq F$, then the conditional probability of F given E is just 1.
- When $F \subseteq E^c$, then the conditional probability of F given E is just 0.
- The third case is when F and E are so called independent: in that case $\mathbb{P}(F|E) = \mathbb{P}(E)$ basically by definition (we will come back to that).

In general, there are not many tools to calculate conditional probabilities, but there is one very useful tool called the Bayes' formula or the Bayes' rule:

2.1.1 Bayes' rule

Proposition 2.3 (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

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

Still, it is a very nice observation that allows us not only to calculate, but also is behind the framework of Bayesian statistics / Bayesian thinking about probability.

Let us here analyse a simple example.

Example 2.4. *Consider the situation with three different coins: one has heads on both sides, one has tails on both sides, and one is a fair coin. Now someone picked using some procedure one of the three types of coins, told you that she tossed a coin and heads came up. Which coin did she toss?*

The relevant probability space that contains the three coins and three tosses is as follows. First, the state space is pairs the product space $\{C_h, C_t, C_f\} \times \{H, T\}$ *- the first coordinate describes the type of the coin, the second the result of the toss. As a* $\sigma-$*algebra we take the whole* $\sigma-$*algebra as we can ask both about what came up on top, and then which coin it was.*

We know that to define $\mathbb{P}$ *on a finite set with the power-set it suffices to define* $\mathbb{P}$ *for every element of the state-space. From the assumptions* $\mathbb{P}(\{C_h, T\}) = \mathbb{P}(\{C_t, H\}) = 0$ *and* $\mathbb{P}(\{C_f, T\}) = \mathbb{P}(\{C_f, H\})$*. If we further set* $p_f = \mathbb{P}(\{coin = C_f\})$*,* $p_h = \mathbb{P}(\{coin = C_h\})$*,* $p_t = \mathbb{P}(\{coin = C_t\})$ *it also has to hold that* $p_f + p_t + p_h = 1$*, leaving two free parameters altogether.*

Let us now calculate the probabilities that we were interested in. Clearly,

$$\mathbb{P}(\{coin = C_t\}|\{toss = H\}) = 0$$

as the coin with two tails sides could not have produced heads. For the other combinations it is easiest to use Bayes' formula to calculate

$$\mathbb{P}(\{coin = c_h\}|\{toss = H\}) = \frac{\mathbb{P}(\{toss = H\}|\{coin = C_h\})\mathbb{P}(\{coin = C_h\})}{\mathbb{P}(\{toss = H\})} = \frac{\mathbb{P}(\{coin = C_h\})}{\mathbb{P}(\{toss = H\})}$$

and

$$\mathbb{P}(\{coin = c_f\}|\{toss = H\}) = \frac{\mathbb{P}(\{toss = H\}|\{coin = C_f\})\mathbb{P}(\{coin = C_f\})}{\mathbb{P}(\{toss = H\})} = \frac{\mathbb{P}(\{coin = C_f\})}{2\mathbb{P}(\{toss = H\})}.$$

Thus we see that

$$\frac{\mathbb{P}(\{coin = c_h\}|\{toss = H\})}{\mathbb{P}(\{coin = c_f\}|\{toss = H\})} = \frac{2\mathbb{P}(\{coin = c_h\})}{\mathbb{P}(\{coin = c_f\})} = 2p_h/p_f$$

and given that

$$\mathbb{P}(\{coin = c_h\}|\{toss = H\}) + \mathbb{P}(\{coin = c_f\}|\{toss = H\}) = 1$$

we conclude our estimates

$$\mathbb{P}(\{coin = c_f\}|\{toss = H\}) = \frac{p_f}{p_f + 2p_h}$$

and

$$\mathbb{P}(\{coin = c_h\}|\{toss = H\}) = \frac{2p_h}{p_f + 2p_h}.$$

What can we conclude? The first thing is maybe that without having any knowledge of how likely each coin was to begin with, we cannot say much about the final answer, as it contains that information! What we assume about the initial probability of each coin matters a lot: if we estimate that the coin with two heads was very unlikely compared to the fair coin, say $p_h = 0.000001p_f$*, then after seeing heads our estimate gives* $\mathbb{P}(\{coin = c_f\}|\{toss = H\}) = 0.999999$*. If however we have no reason to believe that any one coin was more likely to be taken than any other, for example because the person tossing the coin just picked it randomly among the three possibilities, then we have* $p_f = p_h = p_t = 1/3$ *and our formula gives* $\mathbb{P}(\{coin = c_f\}|\{toss = H\}) = 1/3$ *and* $\mathbb{P}(\{coin = c_h\}|\{toss = H\}) = 2/3$*.*

However, an important point is that independently of the initial probabilities, we can say how the probabilities or rather the rations of probabilities changed - our guess that it was the coin was heads/heads went up two times w.r.t. to the fair coin. An in fact, as you will see on the exercise sheet if we could follow more tosses we would become more and more knowledgeable which coin it was, independently of our possibly bad initial estimate. This is also the idea behind Bayesian approach to probability models - we may not know all the parameters to begin with, but we can then just fill them with guesses and as we observe more and more about the world, we can a posteriori improve on these guesses and make our models better.

2.1.2 Law of total probability

Although conditional probabilities are often tricky, they are necessary to deal with and even useful. For example, they help to decompose the probability space. Indeed, 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 2.5 (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 $\Omega = \bigcup_{i\in I} E_i$. 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. As $\Omega = \bigcup_{i\in I} E_i$ we have $\mathbb{P}(F) = \mathbb{P}\left(F \cap \left(\bigcup_{i\in I} E_i\right)\right)$.

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.

□

Remark 2.6. *In fact pretty much the same proof works if E_i don't cover the full space, but we only know that $\mathbb{P}(\Omega \setminus (\bigcup_i E_i)) = 0$. This generalisation is left as an exercise.*

2.2 Independence of events

Conditional probabilities are of course not at all difficult when the probability of an event does not change under conditioning - i.e. when $\mathbb{P}(E|F) = \mathbb{P}(E)$. Such pairs of events are called independent. In fact the rigorous definition is slightly different:

Definition 2.7 (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

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

We have chosen the other definition, as then we automatically also include the case where possibly $\mathbb{P}(F) = 0$.

Example 2.8. *Consider our model of a uniform random number among* $\{1, 2, 3, \ldots, 12\}$ *and the events* $E_1 := \{$*the number is equal to* $1\}$, $E_2 := \{$*the number is divisible by* $2\}$, $E_3 := \{$*the number is divisible by* $3\}$. *Which of these are independent?*

From a direct calculation, we have $\mathbb{P}(E_1) = 1/12$, $\mathbb{P}(E_2) = 1/2$ *and* $\mathbb{P}(E_3) = 1/3$. *But also we can directly calculate that* $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \cap E_3) = 0$ *and* $\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(\{$*the number is divisible by* $6\} = 1/6$. *We conclude that* E_2, E_3 *are independent, but* E_1 *and* E_2 *are not, neither are* E_1, E_3.

Already in this examples we actually had three events and one could also ask if there is some sort of notion of joint independence that generalises to more events. And indeed there are two different ways to generalize independence to several events:

- mutual or joint independence
- and pairwise independence

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

Definition 2.9 (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*

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

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 too.

Definition 2.10 (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 2.2 (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, one can also talk about independence of collections of events. This will be important when we try to generalize the notion of independence from events to random variables

Definition 2.11 (Mutual independence of collections of events)**.** *Consider two collections events* $(E_i)_{i \in I}$ *and* $(F_j)_{j \in J}$ *all defined on the same probability space. We say that they are independent if for all* $i \in I, j \in J$*:*

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

In case of several different collections of events $(E_{j,i})_{i\in I_j}$ for $j = 1\ldots$, we say that these collections are mutually independent if for any finite subset $J_1 \subseteq J$ and any events E_{j,i_j} with $j \in J_1$, it holds that

$$\mathbb{P}\left(\bigcap_{j\in J_1} E_{j,i_j}\right) = \Pi_{j\in J_1}\mathbb{P}(E_{j,i_j}).$$

Equivalently, we ask any subset of events E_{j,i_j} from different collection to be mutually independent.

Before going to the independence of random variables, here are some basic properties of independence for events:

Lemma 2.12 (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. In particular every event with $\mathbb{P}(E) = 0$ is independent of all other events.*
- *Finally, if an event is independent of itself, then $\mathbb{P}(E) \in \{0, 1\}$.*

Proof. This is on the example sheet.

□

2.3 Independence of random variables

We now formalise the notion of independence for random quantities, i.e. random variables. Recall that (the law of) a random variable X is characterized by all events $\{X \in (a, b)\}$ for intervals (a, b). The mutual independence of random variables is then defined as mutual independence of these sets of events. More precisely,

Definition 2.13 (Mutually independent random variables)**.** *Let I be an 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 set of intervals $((a_j, b_j))_{j\in J}$ we have that*

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

Remark 2.14. *The more structurally sound definition would use instead as the collection all Borel sets $E_j \in F_{\mathcal{B}}$. However, that is impractical, and in fact turns out (via some non-trivial measure theory) to be equivalent to the condition above.*

There are naturally more equivalent conditions. For example, a useful one as we see later is the following:

Exercise 2.3. *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 for every $m \geq 2$ and all pairs $a_j \in \mathbb{R}$ we have that*

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

Further, we again have a very nice and simple condition for random variables defined on discrete probability spaces.

Lemma 2.15 (Independence on the discrete probability space)**.** *Let $X_1, \dots, X_n$ be defined on a discrete probability space. Then $X_1, \dots, X_n$ are mutually independent if and only if for every $s_1 \ \dots, s_n \in \mathbb{R}$, we have that*

$$\mathbb{P}(\bigcap_{i=1}^{n}\{X_i = s_i\}) = \Pi_{i=1}^{n}\mathbb{P}(X_i = s_i).$$

The same holds more generally if $X_1, \dots, X_n$ are defined on any probability space but each take only a discrete number of values with full probability, i.e. for each of them there is some countable set S_i such that $\mathbb{P}(X_i \in S_i) = 1$. [6]

Proof. This is left as an exercise. □

As a sanity check it is now simple to see that the indicator events E, F of two events are independent if and only if E, F are independent as events: indeed $\mathbb{P}(\{1_E = x\}\{1_F = y\})$ is equal to

$$1_{x=1}1_{y=1}\mathbb{P}(E)\mathbb{P}(F) + 1_{x=1}1_{y=0}\mathbb{P}(E)\mathbb{P}(F^c) + 1_{x=0}1_{y=1}\mathbb{P}(E^c)\mathbb{P}(F) + 1_{x=0}1_{y=0}\mathbb{P}(E^c)\mathbb{P}(F^c)$$

which in turn can be rewritten as

$$(1_{x=1}\mathbb{P}(E) + 1_{x=0}\mathbb{P}(E^c))(1_{x=1}\mathbb{P}(F) + 1_{x=0}\mathbb{P}(F^c)) = \mathbb{P}(\{1_E = x\})\mathbb{P}(\{1_F = x\}.$$

Exercise 2.4 (Simple symmetric random walk)**.** *Prove that for a simple random walk of length n all the increments of the walk, i.e. $\Delta_i = S_i - S_{i-1}$ for $i = 1 \dots n$, are mutually independent random variables.*

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!

Remark 2.16 (i.i.d. random variables)**.** *Often one talks about collection of i.i.d. random variables $(X_j)_{j\in J}$ - this means that $(X_j)_{j\in J}$ are mutually independent (first 'i') and all have the same probability law, i.e. are identically distributed (the 'i.d.'). Intuitively, this corresponds to repeating the very same random situation or experiment over and over again.*

Now, we started the course by constructing probability spaces and then defining random variables on it. However, there are natural cases where one would like to go in the opposite direction - we know from observation or experience that we would like to study a bunch of independent random variables and our question is how to construct a probability space where they live? This might sound somewhat silly, but in fact mathematically it is not an easy question! We will partly deal with this question in the next subsection.

2.4 Independence and product probability spaces

Whereas independence is a probabilistic concept, it comes out that it is related also to a structure in measure spaces.

Let us consider an example to see this.

Example 2.17 (The space for n fair coin tosses)**.** *We have seen that the probability space for n fair coin tosses can be modelled by taking the state space Ω to be the set of all n-tuples*

[6] Such random variables are called discrete random variables, as we will see soon.

$\{x_1, \ldots, x_n\}$ of length n with each $x_i \in \{H, T\}$, then taking $\mathcal{F}$ to be the power set and finally setting the probability of each singleton, i.e. each n-tuple, to be 2^{-n}.

Now, let us look at this as follows:

- *Each n-tuples is just an element of the product space $\{H, T\} \times \cdots \{H, T\}$, so we can use as Ω the product space. Let's denote also by $\Omega_0 = \{H, T\}$ the state spaces for the coordinates.*
- *The power-set is the smallest $\sigma-$algebra containing all sets of the form $E_1 \times \cdots \times E_n$ with each E_i in the power-set of a single coordinate $\{H, T\}$*
- *The uniform probability measure on Ω satisfies by definition*

$$\mathbb{P}(E_1 \times \cdots E_n) = \mathbb{P}_0(E_1) \cdots \mathbb{P}_0(E_n),$$

  where $\mathbb{P}_0$ is the uniform probability measure on the space of a single toss.
- *Finally the fact that the tosses are independent comes down to the following: all events $F_1, \ldots, F_n$ of the form $F_i = \Omega_0 \times \Omega_0 \times \ldots E_i \times \cdots \times \Omega_0$ with $E_i \in \mathcal{F}_i$ are mutually independent: indeed for $i \neq j$ we have for example*

$$\mathbb{P}(\Omega_0 \times \Omega_0 \times \ldots E_i \times \cdots \times \Omega_0 \cap \Omega_0 \times \Omega_0 \times \ldots E_j \times \cdots \times \Omega_0) = \mathbb{P}(\Omega_0 \times \Omega_0 \times E_i \times \Omega_0 \cdots \times E_j \times \Omega_0 \cdots \times \Omega_0)$$

  which by above equals $\mathbb{P}_0(E_i) \times \mathbb{P}_0(E_j)$ which again by above is equal to the product of $\mathbb{P}(\Omega_0 \times \Omega_0 \times \ldots E_i \times \cdots \times \Omega_0)$ and $\mathbb{P}(\Omega_0 \times \Omega_0 \times \ldots E_j \times \cdots \times \Omega_0$.

So we see that in some sense the product structure goes in hand with independence. And indeed, this is the general rule - mutual independence of random variables is naturally linked to products of probability spaces.

Let us follows this through mathematically, by first discussing product spaces in general and then looking at the construction of probability spaces for independent random variables.

2.4.1 Construction of product spaces

So let us have a brief look at the construction of product spaces. Consider probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ for $i = 1, 2 \ldots$. Then to construct the product probability space we need a product $\sigma-$algebra and a product measure.

(1) The product $\sigma-$algebra $\mathcal{F}_\Pi$ is simple and natural: it is the smallest $\sigma-$algebra containing all $E_{i_1} \times \cdots \times E_{i_n}$ with $E_{i_j} \in \mathcal{F}_{i_j}$ for all $j = 1 \ldots n$ and $\{i_j\}_{j=1 \ldots n}$ a finite subset of $\mathbb{N}$.

(2) The product probability measure $\mathbb{P}_\Pi$ of $\mathbb{P}_1, \mathbb{P}_2, \ldots$ on $(\Pi_{i \geq 1} \Omega_i, \mathcal{F}_\Pi)$ also sounds simple: it is the only probability measure such that

$$\mathbb{P}(E_{i_1} \times \cdots \times E_{i_n}) = \Pi_{j=1}^n \mathbb{P}_i(E_{i_j})$$

for all $E_{i_1} \times \cdot \times E_{i_n}$ with $E_{i_j} \in \mathcal{F}_{i_j}$ for $j = 1 \ldots n$. However, its construction and uniqueness even in the case of finite products is technical for general probability spaces and out of the scope of this course.

Thus we will state the following theorem without proof, which you will see in the measure theory or the third year probability course:

Theorem 2.18 (Product measure // admitted)**.** *For $i \in \mathbb{N}$, let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ be probability spaces. Then there exists a unique probability measure $\mathbb{P}_\Pi$ on $(\Pi_{i \in \mathbb{N}} \Omega_i, \mathcal{F}_\Pi)$ such that 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

$$\mathbb{P}_\Pi(E) = \Pi_{i\in J}\mathbb{P}_i(E_i). \tag{2.1}$$

We call such a measure the product measure of the collection $((\Omega_i, \mathcal{F}_i, \mathbb{P}_i))_{i\geq 1}$.

It is rather easy to see the existence and uniqueness in the case of a finite number of discrete probability spaces, so let us do that. Below, we state it in the case where the $\sigma-$algebras are equal to the power set, but as discussed before (see Proposition 1.9, this essentially encompasses the case of general $\sigma-$algebras on discrete spaces.

Lemma 2.19 (Discrete product spaces)**.** *Let $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$ for $i = 1 \dots n$ be discrete probability spaces. Then the product probability $\mathbb{P}_\Pi$ measure on $(\Pi_{i=1}^n \Omega_i, \mathcal{F}_\Pi)$ exists and is unique.*

Proof. On the example sheet

□

2.4.2 Probability spaces for independent random variables

We will now follow through the philosophy alluded to above:

- if we are given some laws of random variables and we want to construct a common probability space on which all of these random variables are defined and are moreover mutually independent, then we should use product spaces.

We will again state this proposition in a larger generality than we prove it.

Theorem 2.20 (Existence of probability spaces with independent random variables // partly admitted)**.** *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.*

Example 2.21. *Suppose you have a coin that is not fair, but comes up heads with probability $p \in (0, 1)$. How would you model the sequence of independent n such 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 are useful.

Indeed, we can define the probability space as follows:

- *we take the product space of n copies of $(\{0, 1\}, \mathcal{P}(\{0, 1\}), \mathbb{P}_p)$, where $\mathbb{P}_p$ such that it gives 1 with probability p and 0 with probability $1 - p$.*

Notice that in this probability space, the probability of a fixed sequence of n tosses with m heads and tails $n - m$ is exactly $p^m(1 - p)^{n-m}$. If we further want to calculate the probability that we have exactly m heads we have to sum over all sequences with m heads and we get $\binom{n}{m}p^m(1 - p)^{n-m}$. Check that $\sum_{m=0}^n \binom{n}{m}p^m(1 - p)^{n-m} = 1$!

Let us now give the proof of the theorem in the case when all the random variables are defined on discrete probability spaces. For a slightly more natural statement, see the exercise sheet

Proof of Theorem 2.20, case of finite products of random variables on discrete spaces. Suppose we have discrete probability spaces $(\Omega_i, \mathcal{P}(\Omega_i), \mathbb{P}_i)$ and random variables $X_i : \Omega_i \to \mathbb{R}$.

By the Lemma 2.19 above, we can construct the product probability space corresponding to these probability spaces, denoted $(\Omega_\Pi = \Pi_{i=1}^n \Omega_i, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$.

Now, define $\widetilde{X}_i(\omega_1, \ldots, \omega_n) := X_i(\omega_i)$. One can check that $\widetilde{X}_i$ thus defined are all random variables and they are defined to have the same law as X_i. Indeed, by the definition of $\widetilde{X}_i$ and the product measure

$$\mathbb{P}_{\widetilde{X}_i}(E) = \mathbb{P}_\Pi(\Omega_1 \times \Omega_2 \cdots \times X_i^{-1}(E) \times \ldots \cdots \times \Omega_n) = \mathbb{P}_{X_i}(E).$$

Finally, we need to check that the random variables $(\widetilde{X}_i)_{i=1\ldots n}$ are mutually independent on the space $(\Pi_{i=1}^n \Omega_i, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$. From the identity

$$\{w : \Omega_\Pi : \widetilde{X}_j(\omega) \in E_j\} = \{\Omega_1 \times \cdots \times X_i^{-1}(E) \times \cdots \times \Omega_n\}$$

we have that:

$$\mathbb{P}_\Pi(\bigcap_{i=1\ldots n} \{\widetilde{X}_i \in E_i\}) = \mathbb{P}_\Pi(\Pi_{i=1}^n X_i^{-1}(E_i)).$$

By the definition of product measure this equals $\Pi_{1=1}^n \mathbb{P}_{X_i}(E_i)$, which in turn equals $\Pi_{i=1}^n \mathbb{P}_{\widetilde{X}_j}(E_j)$ by equality in law. The last expression is equal to $\Pi_{i=1}^n \mathbb{P}_\Pi(\widetilde{X}_i \in E_i)$ by definition and we conclude. □

Let us finish this section by playing with an important example.

2.4.3 Erdös-Renyi random graph

Our aim in this section is to describe and study random graphs. Graphs are simple mathematical structures that help to describe networks like social networks, or logistic networks or why not the network of neurons in the brain.

Definition 2.22 (Simple graph)**.** *Let $n \in \mathbb{N}$. A simple graph is a pair $G = (V, E)$ where V is a set of points $V = \{v_1, \ldots, v_n\}$, called vertices, and E is a subset of $\{\{v_i, v_j\} : (v_i, v_j) \in V \times V, v_i \neq v_j\}$, i.e. a set of unordered pairs of distinct vertices, called edges.*

You can imagine the graph as drawing all the n points $v_1, \ldots, v_n$ on the plane and then drawing a line between v_i and v_j to say they are connected if and only if $\{v_i, v_j\} \in E$.

If the networks are very big, like the brain or the social network in Facebook, it is both impractical and unfeasible to describe them in all detail. Moreover, it comes out that usually they start resembling certain random networks. Thus in order to understand properties of these real world networks, one often studies the simplified models of random networks.

The easiest model of a random network, or in our mathematical language of a random graph, is the Erdös-Renyi random graph where we include each edge with probability $p > 0$.

Example 2.23 (Erdös-Renyi random graph)**.** *For $n \in \mathbb{N}$ consider a set of vertices V of size n and let E be the set of all undirected edges between these vertices.*

The Erdös-Renyi random graph $G_{n,p}$ of size n and edge parameter $p \in [0, 1]$ is then defined by including each possible edge independently with probability p.

To define the relevant probability space we let

- *The state space should include all possible graphs with the vertex set V. We observe that this can be done by determining the edge set. So we let $\Omega = \{0,1\}^E$ be the set of all possible edge configurations on n vertices - we interpret 1 to mean that an edge is present.*
- *We assume that we can check for any edge if it is present or not, and thus set $\mathcal{F} = \mathcal{P}(\Omega)$*
- *Finally, we set each edge to be present with probability p independently of others. In other words for each $\omega \in \Omega$ we set*

$$\mathbb{P}_p(\{\omega\}) := p^{|\omega|}(1-p)^{|E|-|\omega|},$$

where $\omega \in \Omega$ is an edge configuration and $|\omega|$ is the number of edges in this configuration.

Finally, we can identify each element ω also with the resulting graph $G_{n,p}(\omega) = (V, E(\omega))$.

What are some questions that we would like to look at? Roughly we would like to answer how the graph look likes when n is very large, i.e. tending to infinity. Of course sometimes one could be also interested in n small, but then one could actually explicitly describe the probability of each possible graph and picture it.

Now to describe how the graph looks like we could consider the following questions:

(1) How many edges are present?
(2) Is the graph connected, i.e. can one find for each $v, w \in V$ a set of edges $e_1, \ldots, e_n$ such that each e_i, e_{i-1} share a vertex and e_1 is connected to v and e_n connected to w?
(3) If yes, what is the maximal distance between two vertices?
(4) If no, how many different connected components are there?
(5) What is the biggest connected component?
(6) ...

Each of these questions is about a single graph, i.e. a single configuration ω. Thus in the random graph model they correspond either to an event or random variable, whose probability or law we can study.

For example, $N_E : \Omega \to \mathbb{N}$ given by $N_E(\omega) := |\omega|$ attaches to each ω its number of edges and thus corresponds to the first question. Similarly the event $F := \{\omega \text{ is connected}\}$ corresponds to the second question. Of course there are also more complex questions, which arise when one consideres several questions at the same time.

One is interested in both how the probability of these events behaves for $p \in [0,1]$ fixed and $n \to \infty$, but also how this behaviour changes when we change p. Notice that a priori p does not need to be constant, we can also easily consider a sequence of graphs $G_{n,p(n)}$ where $p(n)$ is a function of n.

Studying the properties of Erdös-Renyi random graphs was and still is a very active research topic, with hundreds if not thousands of papers written about them. We will try to just get a very small taste of this research.

Let us concentrate on one notion, that of connectivity and look at some scenarios. Notice that when $p = 1$ then the graph is connected with probability 1 and when $p = 0$ it is disconnected with probability 1. We will try to get a grasp what happens with $p_n \in (0,1)$ possibly changing with n.

Claim 2.24. *Let $p \in (0,1)$ be fixed. Then as $n \to \infty$ the probability of the graph being connected converges to 1 almost surely i.e. with probability 1.*

This is maybe not so surprising as with fixed probability p we will have lots of edges: indeed, if you think of edges as coin tosses, you would expect to have a proportion p of all edges to be present, which makes $pn(n-1)/2$ edges!

Proof. We will prove that $\mathbb{P}_p(\{G_{n,p} \text{ is not connected}\}) \to 0$ as $n \to \infty$. First notice that

$$\{G_{n,p} \text{ is not connected}\} = \cup_{v \neq w \in V}\{ \ v, w \text{ not connected by a path}\}.$$

Thus by the union bound

$$\mathbb{P}_p(\{G_{n,p} \text{ is not connected}\}) \leq 1/2 \sum_{v \neq w \in V} \mathbb{P}_p(\{ \ v, w \text{ not connected by a path}\}),$$

where the $1/2$ comes from the fact that we count each edge twice in the sum. But because of symmetry of the model, each pair of edges is equivalent, so we can write the right hand side as $n(n-1)/4\mathbb{P}_p(\{ \ v, w \text{ not connected by a path}\})$.

Thus we want to bound the probability that v and w are not connected by a path. First, just looking at the edge $\{v, w\}$ is not enough - this edge is absent with probability $1-p$, which doesn't go to zero. However, there are many other ways to connect these two vertices.

One way is to use an intermediate vertex z: then v and are not connected if and only if w there is no vertex z such that both $\{z, w\}$ and $\{z, v\}$ belong to the edge set. Thus we can write

$$\mathbb{P}_p(\{ \ v, w \text{ not connected by a path}\}) \leq \prod_{z \in V \setminus \{v,w\}} \mathbb{P}_p(\{\{v, z\} \notin E\} \cup \{\{w, z\} \notin E\}).$$

But now $\mathbb{P}_p(\{\{v, z\} \notin E\} \cup \{\{w, z\} \notin E\}) = 1 - \mathbb{P}_p(\{\{v, z\} \in E\} \cap \{\{w, z\} \in E\} = 1 - p^2$ and hence

$$\mathbb{P}_p(\{ \ v, w \text{ not connected by a path}\}) \leq (1-p^2)^{n-2}.$$

This clearly goes to zero as $n \to \infty$ and thus any two fixed vertices will be connected with probability going to 1.

We now come back to our initial probability of all pairs being connected and bound:

$$\mathbb{P}_p(\{G_{n,p} \text{ is not connected}\}) \leq n(n-1)(1-p^2)^{n-2}/4.$$

This is also nicely goes to zero!

□

In fact, if we look at the proof more carefully we see that the claim is true as long as $p = p(n)$ goes to zero with n sufficiently slowly. In other words the exact same proof gives us

Claim 2.25. *Let $(p_n)_{n \geq 1}$ be a sequence of numbers in $[0,1]$ satisfying $p_n \geq n^{-1/4}$. Then as $n \to \infty$ the probability of the graph being connected converges to 1 almost surely i.e. with probability 1.*

Proof. We follow the proof above and notice that for $1 \geq p_n \geq n^{-1/4}$ we still have that

$$n(n-1)(1-p^2)^{n-2}/4 \to 0$$

as $n \to \infty$. □

On the other hand, we have that

Claim 2.26. *Let $(p_n)_{n\geq 1}$ be a sequence of numbers in $[0,1]$ such that $p_n \leq n^{-2}$. Then as $n \to \infty$ the probability of the graph being connected converges to 0 almost surely i.e. with probability 1.*

This will be on the exercise sheet. But notice the interesting phenomena: there seems to be a sort of threshold effect. If p_n decays very fast, the probability of connectedness goes to 0; if decays slowly enough it goes to 1. Why doesn't it go to some other number between 0 and 1? Where is the exact threshold? It is a non-trivial theorem that says this threshold is exactly at $p_n = \frac{\log n}{n}$!

SECTION 3

Random variables and random vectors

In this chapter, we will look more closely into random variables and n-tuples of random variables, called random vectors.

3.1 The cumulative distribution function of a random variable

Recall that we call two random variables equal in law, when the probability measures they induce on $(\mathbb{R}, \mathcal{F}_B)$ are equal - this allowed us to compare random variables defined on different probability spaces, coming up in different contexts.

Our first aim is to see how to classify and compare random variables more easily. Indeed, for now one has to actually determine We already saw that the law of each random variable is described by the probability over all possible events, but this is a description that is very difficult to deal with.

It comes out that all the information about the law of a random variable can be uniquely encoded using what is called a cumulative distribution function.

Definition 3.1 (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 $F(x_n) \to F(x)$.

Given a random variable X, we define its cumulative distribution function as follows:

Proposition 3.2 (Cum.dist. function of a random variable)**.** *For each random variable X (defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$), the function $F_X(x) := \mathbb{P}_X((-\infty, x])$ defines a cumulative distribution function (c.d.f).*

Proof. Set $F_X(x) = \mathbb{P}(X \in (-\infty, x])$. Then as $(-\infty, x] \subseteq (-\infty, y]$ for $x \leq y$, we have by (1) of Proposition 1.12 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 $\bigcap_{n\geq 1} A_n = (-\infty, x]$. By continuity of $\mathbb{P}$, i.e. (5) of Proposition 1.12, it follows that $\mathbb{P}_X(A_n) \to \mathbb{P}_X((-\infty, x])$. But now notice that as $x_n \to x$, we have that for any n large enough $\{-\infty, y_n\} \subseteq A_{m_n}$ for some m_n chosen such that $m_n \to \infty$ as $n \to \infty$. It follows that $F_X(y) \leq F_X(y_n) \leq \mathbb{P}_X(A_{m_n})$ and we conclude that $F_X(y_n) \to F_X(y)$ as $n \to \infty$.

The final two claims are on the example sheet.

□

In fact, it comes out the conversely each cumulative distribution function gives rise to a unique law of a random variable.

Theorem 3.3 (Laws of random variable are uniquely determined by c.d.f. // admitted)**.** *Each cumulative distribution function F gives rise to a unique law of a random variable X such that $F_X(x) = \mathbb{P}_X((-\infty, x])$. In other words c.d.f.s are in one to one correspondence with probability measures $\mathbb{P}$ on $(\mathbb{R}, \mathcal{F}_B)$.*

We admit this theorem in the general case, but will again prove the discrete case. Let us look at a simple example:

Example 3.4. *Let us calculate the c.d.f of the so called Bernoulli random variable X that takes value 1 with probability p and 0 with probability $1-p$. Notice that all indicator functions of events correspond to such random variables with $\mathbb{P}(E) = p$.*

We 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 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, i.e. by numbers $x_n < x$.

Lemma 3.5 (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. □

Example 3.6. *Let us also exhibit the c.d.f. of the uniform random variable U taking values uniformly in $[0, 1]$. It is given by $F_U := x1_{x\in[0,1]} + 1_{x>1}$. By the proposition above we can see that for any interval $(a, b) \subseteq [0, 1]$, $\mathbb{P}(U \in (a, b)) = b - a$.*

From above we see that all jumps of F_X correspond to points where $\mathbb{P}_X(X = x) > 0$. In fact there can be only countably many of them.

Lemma 3.7. *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 3.8 (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*

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

Notice that this set S could be dense, like the set of rational numbers, making it hard to imagine as a staircase function!

In the other extreme F_X could also be everywhere continuous. These observations help us separate out two classes of random variables.

3.1.1 Classification of random variables

Definition 3.9 (Discrete vs continuous random variables)**.** *A random variable is called discrete if its c.d.f. F_X is piece-wise constant changing value at most countable many times. It is called continuous if its c.d.f. F_X is continuous.*

These definitions look a bit abstract / non-telling from the probabilistic perspective and a priori differs from the definition we gave on the example sheet! But no need to worry, it does give the same object:

Exercise 3.1 (Discrete vs random variables ver 2)**.** *Consider a random variable X. Prove that*

- *X is discrete, i.e. its cumulative distribution function F_X is piece-wise constant, if and only if there is a countable set $S \subseteq \mathbb{R}$ with $\mathbb{P}(X \in S) = \mathbb{P}_X(S) = 1$.*
- *X is continuous if and only if for every $y \in \mathbb{R}$, $\mathbb{P}(X = y) = \mathbb{P}_X(\{y\}) = 0$.*

Notice that not every random variable is either discrete or continuous, there could be also mixtures of the two, e.g. one could imagine a c.d.f. given by $F(x) = 0.51_{x\geq 0} + 0.5x1_{x\in[0,1)} + 1_{x\geq 1}$ (What does it correspond to?).

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 3.10. *Any cumulative distribution function F can be written uniquely as convex combination of a continuous c.d.f F_c and a piece-wise constant c.d.f. with countably many jumps F_j i.e. for some $a \in [0, 1]$ we have that $F = aF_j + (1 - a)F_c$.*

Moreover, in a later the exercise sheet you will see how to interpret this as saying that each random variable can be written as a random sum of a continuous and discrete random variable.

Proof. If F is either continuous or piece-wise constant with countably many jumps, the existence of such writing is clear. So suppose that F is neither. Write S for the countable set of jumps of F. Define

$$\widehat{F}_j(x) = \sum_{s\in S} 1_{x\geq s}(F(s) - F(s-)),$$

which is piece-wise continuous with countably many jumps.

We claim that $\widehat{F}_c := F - \widehat{F}_j$ is continuous. Indeed, by definition both F and $\widehat{F}_j$ both right-continuous, and thus is also their difference. Moreover, both are continuous at any continuity point x of F, i.e. when $x \notin S$ as by definition then $F(x) = F(x^-)$ and one can check the same for F_j. Finally, when $s \in S$, then again by definition of $\widehat{F}_j$, we have that

$$F(s) - F(s-) = 1_{s\geq s}(F(s) - F(s-)) = \widehat{F}_j(s) - \widehat{F}_j(s-)$$

and thus $\widehat{F}_c$ is continuous at such s too.

Now, as F is neither continuous nor piece-wise constant increasing with jumps, we have that $0 < \widehat{F}_j(\infty) < 1$ and $0 < \widehat{F}_c(\infty) < 1$. Hence, we can define

$$F_j(x) := \frac{\widehat{F}_j(x)}{\widehat{F}_j(\infty)}$$

and

$$F_c(x) := \frac{\widehat{F}_c(x)}{\widehat{F}_c(\infty)}.$$

By definition both of those are non-decreasing, right-continuous satisfying the correct limits at $\pm\infty$ and hence are c.d.f-s for random variables. As F_j increases only via jumps and F_c is continuous, we have the desired writing with $a = \widehat{F}_j(\infty)$ and $1 - a = \widehat{F}_c(\infty)$.

Uniqueness is left as an exercise. 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.

□

3.2 Examples of 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.

Recall that to characterise the law of a random variable, we can either give the value of $\mathbb{P}_X(F)$ for a sufficiently large set of F (e.g. all intervals) or give the c.d.f. For a discrete random variable X it suffices to just determine the support S, i.e. the smallest set $S \subseteq \mathbb{R}$ such that $\mathbb{P}(X \in S) = 1$ and determine $\mathbb{P}_X(X = s)$ for each $s \in S$ (why?).

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 $\{-1, 1\}$. We then call it the Bernoulli

random variable with values $\{-1, 1\}$.

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.

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 this set. 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.

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. Bernoulli random variable comes up naturally in models of independent coin tosses, 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. Or in other words, the Binomial random variable $Bin(n, p)$ can be seen as a sum of n independent $Ber(p)$ random variables.

Exercise 3.2 (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, 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 3.11 (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 $\mathbb{P}(X > k + l | X > k) = (1 - p)^{k+l-k} = (1 - p)^l = \mathbb{P}(X > l)$ 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

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

As for a memoryless random variable $\mathbb{P}(X > l) = \mathbb{P}(X > 1 + l | X > 1)$, we obtain

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

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 3.12 (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/n 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 $p = \lambda/n$.

Poisson random variables also behave very well under taking independent copies and taking random subsets of them:

Exercise 3.3 (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 $Y_1, Y_2, \ldots$ 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.*

3.2.1 How to choose my distribution - the maximum entropy principle

We have now seen several examples of discrete random variables with special properties, which may or may not have sounded relevant and were a bit different for each of the examples. A generic question is the following. Suppose we want to model some statistical phenomena using a random variable. From the experiments or theory we can deduce some weak constraints on the probability distribution of the variable - for example the support of the distribution, i.e. which values it takes, and maybe some other parameters obtained from repeated experiments like some sort of average. The question is: which probability distribution should we choose as our model under these constraints?

Intuitively, we would like to choose a distribution that takes into account these constraints and nothing more. Already Laplace used such an argument: his principle of insufficient reason says that if we only know that we have n outcomes, we should assign each the probability $1/n$. It comes out that somehow the right generalization of this principle of insufficient reason is the principle of maximum entropy - we should choose the probability distribution with maximal entropy, given the constraints. This feels at least intuitively natural, as we are then maximizing our surprise or uncertainty about what is happening. The principle of maximum entropy was introduced by E. T. Jaynes in the 1950s.

To state this, let us first introduce the concept of entropy in the realm of discrete random variables. This rich concept is also interesting in itself and I believe you have already met it in your course of computer science as introduced by Shannon, although its origins go back much further in thermodynamics. In essence the entropy of a random variable is a way to formalise the notion of information content that we learn by observing an outcome.

Definition 3.13 (Discrete (Shannon) entropy)**.** *For a discrete random variable X with outcome set S we define the entropy*

$$H(X) := -\sum_{s \in S} \mathbb{P}(X = s) \log \mathbb{P}(X = s).$$

We also call it the entropy of the probability distribution $\mathbb{P}$.

The mentioned link to information content can be thought of in two steps:

- First, to each event E, and in particular to each outcome $\{X = s\}$ we assign a measure of information content or surprise: $-\log \mathbb{P}(E)$. We like this precise measure more than just $\mathbb{P}(E)$ basically because log is additive under products and hence the information content of two independent events adds up.

- Then to measure the information content over all outcomes we take the weighted average of $-\log \mathbb{P}(X = s)$ as in the definition above. Such a weighted average is called the mathematical expectation, as we will shortly see.

Remark 3.14. *Often in computer science / information theory one rather uses* $\log_2$ *instead of* log *- this in some sense is just a choice of units for the information content.*

Example 3.15. *One can directly check that for the uniform distribution on* n *points the entropy is* $H = \log n$*. Indeed then* $\mathbb{P}(X = s) = 1/n$ *for any* s *and the claim follows.*

Notice that although we defined the discrete entropy for a random variable, it does not depend on the exact values of the random variable - only on how many values with which probability it takes. In particular for example it is direct to see the following Lemma.

Lemma 3.16. *For any real-valued discrete random variable we have that* $H(X) = H(aX+b)$*.*

Further, the reason of choosing the logarithm in the definition, boils down to the following facts:

Lemma 3.17. *The entropy is non-negative:* $H(X) \geq 0$*. Moreover, for independent discrete random variables* X, Y*, we have that* $H(X, Y) = H(X) + H(Y)$*.*

Proof. The first part is evident from the fact that $-p \log p \geq 0$ for all $p \in [0, 1]$ [7], The proof of the second part is a direct computation on the example sheet. □

There are many ways to give mathematical characterisations of entropy, i.e. to give a set of intuitive conditions for a measure of information content such that they uniquely characterise the entropy functional. We will not do this in our course, but rather explain a property that makes entropy appear - asymptotic equipartition property.

In this respect, consider i.i.d. non-trivial discrete random variables $X_1, X_2, \ldots, X_n$ taking each of the values $s \in S$ with positive probability. Then for each sequence of outcomes $(s_1, s_2, \ldots, s_n)$ we can calculate the probability $\mathbb{P}(X_1 = s_1, \ldots, X_n = s_2)$. By independence, this is given by $\Pi_{i=1}^n \mathbb{P}(X_i = s_i)$. As each of $\mathbb{P}(X_1 = s) \in (0, 1)$, these probabilities decay exponentially with n, i.e. should be roughly of the form $\exp(cn)$. Is this the case, and how does this $c > 0$ behave? For a given sequence, this will clearly depend on the exact sequence $(s_1, s_2, \ldots, s_n)$, so one cannot expect an answer in full generality.

However, one can determine this exponent for a typical outcome sequence and it is given by the entropy of X_1:

Theorem 3.18 (Asymptotic equipartition property)**.** *Let* $X_1, X_2, \ldots$ *be i.i.d. non-trivial discrete random variables, taking values* $s \in S$ *with positive probability. Denote by* $p(s_1, s_2, \ldots, s_n)$ *the probability of the sequence of outcomes* $(s_1, s_2, \ldots, s_n)$*. Then* $\mathbb{P}(|-\frac{1}{n} \log p(X_1, \ldots, X_n) - H(X_1)| > \epsilon)$ *converges to* 0*.*

We will be able to give a simple proof and even a stronger statement of this result towards the end of the course, but I encourage you already to try it out now - this maybe also helps to understand how the notion we are about to introduce simplify our thinking.

As said, entropy has many contexts and many uses, but for us the aim was to help select probability laws on discrete sets. Here are two results in this direction.

[7] If you wish we take the convention that $0 \log 0 = 0$, which of course also makes a lot of sense by taking $x > 0$ and letting it tend to 0.

Lemma 3.19 (Uniform distribution has maximum entropy)**.** *Consider all probability distributions on n points. Among such distributions, the uniform distribution is the unique maximum entropy distribution.*

Proof. Let $Q = (q_i)_{i=1\ldots n}$ denote a probability distribution on n points. We want to prove that

$$H(Q) = -\sum_{i=1}^{n} q_i \log_2 q_i \leq \log_2 n.$$

We use the fact that $\log x$ is concave on $[0, 1]$. Thus we have that

$$H(Q) = \sum_{i=1}^{n}(q_i \log_2 \frac{n}{q_i}) - \log_2 n \leq \log_2(\sum_{i=1}^{n} \frac{q_i n}{q_i}) = \log_2 n^2 - \log_2 n = \log_2 n,$$

where we also used that $\sum_{i=1}^{n} q_i = 1$. Moreover, the equality holds only if $q_i = 1/n$ for all i, so in fact the uniform distribution is the unique maximum entropy distribution □

Similarly we can single out the geometric random variable among all distributions with outcomes in $\mathbb{N}$, however with one extra constraint. Namely, it is maximum entropy distribution on $\mathbb{N}$ among those distributions for which the so called mean or expectation is finite: $\sum_{n\geq 1} n\mathbb{P}(X = n) < \infty$.

3.3 Continuous random variables

Recall that we called a random variable X continuous if F_X was continuous, i.e. without any jumps. From Lemma 3.5 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 3.20 (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 $x \in \mathbb{R}$*

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

[8]

Remark 3.21. *We remark straight away that there are also continuous random variables without a density (see starred section of the exercises).*

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 3.4. *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.*

[8] You might have already heard that there are several notion of an integral. Here the natural integral to use would be Lebesgue integral as then one can integrate over all Borel sets, which as you may have seen, is not possible for the Riemann integral. But in fact for all the examples here thinking of Riemann integral is quite sufficient.

- *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 3.5 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 saw one way of making it precise on Exercise sheet 7.

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 3.5 (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.

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 a more 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...

One of the other important aspects of Gaussians are their intimate relation to linear algebra: Gaussian random variables and random vectors behave extremely well under linear transformations, making them already for this reasons central to many probabilistic models.

Here is a simple lemma in this spirit giving also a meaning to μ and σ^2 as a shift and scaling:

Lemma 3.22. *Let X_{μ,σ^2} be a Gaussian random variable. Further Let X be a standard Gaussian. Then $\sigma X + \mu$ has the same law as X_{μ,σ^2}.*

Proof. This is a direct computation. Let us denote by F_{μ,σ^2} the c.d.f. of X_{μ,σ^2} and by F_X the c.d.f. of X. Pick $\sigma > 0$ and let us calculate the c.d.f. of $\sigma X + \mu$:

$$F_{\sigma X+\mu}(t) = \mathbb{P}(\sigma X + \mu \leq t) = \mathbb{P}(X \leq (t-\mu)/\sigma) = F_X = \int_{-\infty}^{(t-\mu)/\sigma} \frac{1}{2\pi} \exp(-x^2/2)dx.$$

We now make a change of variable $y = \sigma x + \mu$ to see that the right hand side equals

$$\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y-\mu)^2/2\sigma^2) = F_{\mu,\sigma^2}(t).$$

□

3.4 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, it is sometimes useful to think in terms of random vectors:

Definition 3.23 (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 more natural to look at $(X_1, \ldots, X_n)$ not as just a vector of $\mathbb{R}$-valued random variables, but rather as a $\mathbb{R}^n$-valued random variable:

Lemma 3.24 (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 $(\mathbb{R}^n, \mathcal{F}_B)$-valued random variable i.e. the map $\omega \to (X_1(\omega), \ldots, X_n(\omega))$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{F}_B)$. In particular a random vector induces a probability measure $\mathbb{P}_{\overline{X}}$ on $(\mathbb{R}_n, \mathcal{F}_B)$ 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 according to the definition above.

We will not prove this lemma, but just remark that the underlying 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. This should remind you of your topology course and vector-valued continuous functions [9].

This set-up allows us to quickly prove the following basic result:

Lemma 3.25. *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.*

On the exercise sheet you will prove by hand that the sum of two random variables X_1 and X_2 is a random variable - and you will see, it requires patience!

Proof. By above $\overline{X}$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^n, \mathcal{F}_B)$. But now $\Phi : \mathbb{R}^n \to \mathbb{R}$ given by $\Phi(\overline{x}) = \sum_{i=1}^n a_i x_i$ is continuous from $(\mathbb{R}^n, \tau_B)$ to $(\mathbb{R}, \tau_B)$ and in particular it is measurable.

But it is a direct check a concatenation $f_2 \circ f_1$ of measurable maps $f_1 : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}_1)$, $f_2 : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ is $(\Omega, \mathcal{F}) \to (\Omega_2, \mathcal{F}_2)$-measurable. Thus $\sum_{i=1}^n a_i X_i = \Phi(\overline{X})$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \tau_E)$ and hence a random variable. □

3.4.1 Joint cumulative distribution function

Similarly to the case of a single random variable, random vectors can be characterised by a certain family of functions.

Definition 3.26 (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 $(x_1^m, \ldots, x_n^m)_{m \geq 1}$ 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:

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

As mentioned, each random vector uniquely identifies a joint c.d.f. and vice-versa. One part of the proposition is again easy:

Proposition 3.27 (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*

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

gives rise to a joint cumulative distribution function.

Proof. This is left as an exercise. □

[9] Indeed, the statement of interest here is the following. If $(\Omega, \mathcal{F})$ and $((\Omega_i, \mathcal{F}_i))_{1 \leq i \leq n}$ are 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. Compare this 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 $f = (f_1, \ldots, f_n)$.

However, the existence and uniqueness part given the joint c.d.f. is technical and thus admitted.

Theorem 3.28 (Existence and uniqueness of random vectors via joint c.d.f. (admitted))**.** *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 independence:

Lemma 3.29 (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 $\overline{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$.*

Many relevant examples come actually from joint laws, where each marginal law is different. However, the case of Gaussian vectors is well-spread in machine learning / statistics and elsewhere. To state this, we first define the notion of density for random vectors.

Definition 3.30 (Random vectors with density)**.** *Let $\overline{X} = (X_1, \dots, X_n)$ be a random vector and let $f_{\overline{X}}$ be a non-negative integrable function* [10] *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 $(a_1, b_1] \times \dots (a_n, b_n]$*

$$\mathbb{P}_{\bar{X}}(X_1 \in (a_1, b_1], \dots, X_n \in (a_n, b_n]) = \int_{(a_1,b_1]\times\cdots\times(-a_n,b_n]} f_{\overline{X}}(\bar{x})d\bar{x}. \tag{3.1}$$

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 $\bar{t} = (t_1, \dots, t_n)$. Indeed, if $f_{\overline{X}}$ is continuous, then you can check that we have

$$\mathbb{P}_{\overline{X}}((X_1, \dots, X_n) \in (t_1, \dots, t_n) + [-\epsilon/2, \epsilon/2]^n) = f_{\overline{X}}(t_1, \dots, t_n)\epsilon^n + o(\epsilon^n). \tag{3.2}$$

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

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

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.

We can now state the key example:

Gaussian random vector. The Gaussian (or also normal) random vector is denoted by $\mathcal{N}(\overline{\mu}, C)$, 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:

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

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 $\mathbb{R}^n$. In fact all Gaussian vectors in $\mathbb{R}^n$ are given by just linear transformations of the standard Gaussian - this is what also makes the Gaussians ubiquitous, they behave

[10] Again, you can assume we are using the Riemann integral. In fact one could give a more natural definition via Lebesgue integral, but this one works fine too.

very well under linear transformations.

Section 4

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. In many senses mathematical expectation of a probability distribution is the number that one should give if asked for one single number to describe the distribution.

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.

4.1 Expected value of a discrete random variable

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

Definition 4.1 (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 4.2. *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 $\{1, 2, \ldots, n\}$. 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 4.1 (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 $1/p$.*

As mentioned, the expected value is in some sense the best single number to describe a probability distribution. There are several reasons to say that and first is the following: it minimizes the expected error we make in estimating the value of X just using one deterministic number, when we measure the error in terms of average square differences.

Lemma 4.3. *Let X be an integrable discrete random variable with support S. Suppose that also X^2 is integrable. Then $c = \mathbb{E}(X)$ minimizes the expression $g(c) := \sum_{x\in S}(x-c)^2\mathbb{P}(X = x)$.*

Moreover, show from the definition that the value of $g(\mathbb{E}(X))$ can be written as $\mathbb{E}((X - \mathbb{E}(X)^2)$. This is called the variance of X.

Proof. This is on the example sheet.

□

Another good reason for liking expectation is the fact that it is a linear operator on random variables. Together with this, let us also verify some other simple properties.

Proposition 4.4. *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

$$\sum_{s \in S_{X+Y}} |s| \mathbb{P}(X+Y=s) \leq \sum_{s \in S_{X+Y}} \sum_{x \in S_X} \sum_{y \in S_Y} (|x| + |y|) \mathbb{P}(X=x, Y=y) 1_{x+y=s}. \tag{4.1}$$

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 (4.1) 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 $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

The rest of the exercise is on the example sheet.

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

□

A very similar proof gives that if X, Y are independent and integrable discrete random variables, then XY is integrable and $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Exercise 4.2. *Let X, Y be independent and integrable discrete random variables. Then XY is integrable and $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

This allows us to come to the other fundamental property of the expectation - the empirical average converges to the mathematical expectation, allowing us to justify why we would should maybe be happy to pay any less than 3.5 francs to repeatedly be able to play the dice came from above...

Theorem 4.5 (A version of law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. integrable discrete random variables such that X_1^2 is also integrable. Then for every $\epsilon > 0$*

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)| > \epsilon) \to 0$$

as $n \to \infty$

Roughly, this law of large numbers says that if you repeat the same random experiment independently n times to obtain i.i.d random variables $X_1, X_2, \ldots, X_n$ 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.

We need one final ingredient before proving this:

Proposition 4.6 (Markov)**.** *Let X be a non-negative integrable discrete random variable. Then $\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}(X)$.*

Remark 4.7. *This and the independence claim of course hold also for the general random variables, we just need to first define their expectation!*

Proof of Theorem. By assumption there is some C such that $\mathbb{E}X_1^2 < C$. Let $S_n = n^{-1}\sum_{i=1}^{n} X_i$.

Our aim is to use the Markov's inequality. However, as absolute value is hard to work with we will instead use it for the square, which amends itself to linearity of exepctation and the property of independence from above:

$$\mathbb{P}(|S_n - \mathbb{E}(X_1)| > \epsilon) = \mathbb{P}((S_n - \mathbb{E}(X_1))^2 > \epsilon^2) \leq \mathbb{E}((S_n - \mathbb{E}(X_1))^2)/\epsilon^2.$$

So let us calculate $\mathbb{E}((S_n - \mathbb{E}X_1)^2)$. First by writing out S_n, opening the brackets inside expectation and then using linearity of expectation we have

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = \sum_{i,j \leq n} n^{-2}\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right].$$

We have that $\mathbb{E}X_j = \mathbb{E}X_1$. Thus we see that by linearity

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right] = \mathbb{E}(X_iX_j) + (\mathbb{E}(X_1))^2 - 2(\mathbb{E}(X_1))^2 = \mathbb{E}(X_iX_j) - (\mathbb{E}(X_1))^2.$$

But for $i \neq j$, by independence also $\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\mathbb{E}(X_j) = (\mathbb{E}(X_1))^2$, giving us

$$\mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1)\right] = 0$$

for $i \neq j$. Hence

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = n^{-2}\sum_{i=1}^{n}\left(\mathbb{E}(X_i^2) - (\mathbb{E}(X_1))^2\right) = n^{-2}n^{-1}C \to 0$$

as $n \to \infty$. Hence we see that

$$\mathbb{P}(|S_n - \mathbb{E}X_1| > \epsilon) \leq \epsilon^{-2}n^{-1}C \to 0$$

and the theorem follows.

□

We are still to prove the claim in Exercise 4.2 and the Markov's inequality. The first one will be on the next example sheet, Markov's inequality comes now:

Proof of Markov's inequality: Let X be a non-negative discrete integrable random variable. Then $Y_t = X1_{X\geq t}$ is also a non-negative discrete integrable random variable as $Y_t \leq X$. But now observe that $Y_t \geq t1_{X\geq t}$ and thus

$$\mathbb{E}(X) \geq \mathbb{E}(Y_t) \geq \mathbb{E}(t1_{X\geq t}).$$

But $\mathbb{E}(t1_{X\geq t}) = t\mathbb{P}(X \geq t)$ by linearity and the fact that 1_E is Bernoulli random variable We obtain $\mathbb{E}(X) \geq t\mathbb{P}(X \geq t)$ as desired. □

Hopefully you got convinced that the notion of mathematical expectation is pretty useful. We will now see how to generalize it to arbitrary, not necessarily discrete random variables.

4.2 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:

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

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 4.3 (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 4.8 (Expected value of a random variable)**.** *Let X be a random variable defined on some probability space. If $\mathbb{E}(|X_m|) < \infty$ for some m, then $\mathbb{E}(|X_n|) < \infty$ for all n 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 4.9. *Observe again that the expectation only depends on the law of X and not on the underlying probability space: this is clear in the case of discrete random variables, but now notice that if X and Y have the same law, then so do the discretizations X_n and Y_n.*

Remark 4.10. *A peek into future: if you consider $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{F}_L, \mathbb{P}_U)$ where $\mathcal{F}_L$ is the Lebesgue $\sigma-$algebra and $\mathbb{P}_U$ the Lebesgue measure (we also called it uniform measure). Then for any integrable random variable X, which is just a measurable function from $([0,1], \mathcal{F}_L)$ to $([0,1], \mathcal{F}_E)$, $\mathbb{E}X$ is its Lebesgue integral. You will see a more general construction in your Analysis IV course using a larger family of approximations.*

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 4.3 above we see that $X_1 - 1 \leq X_n \leq X_1 + 1$ and hence $|X_n| \leq |X_1| + 1$. Thus from Proposition 4.4 it follows that $\mathbb{E}(|X_n|) < \infty$ if and only if $\mathbb{E}(|X_1|) < \infty$ giving the first claim.

We now claim that $\mathbb{E}(X_n)$ is a Cauchy sequence. So consider $m \geq n$. Then from Proposition 4.4 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 4.3. 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 4.1 of $\mathbb{E}(X)$ and the definition of $\mathbb{E}(X)$ by Proposition 4.8 agree for any discrete random variable X. This is on the example sheet.

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 4.11. *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 $\mathbb{E}(|X|) \geq |\mathbb{E}(X)|$.*

Further also the Markov inequality holds.

Proof. All these points follow from Proposition 4.4 via discretizations and Exercise 4.3. 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 4.11 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 $\mathbb{E}(X) \geq 0$.

Markov's inequality can be proved either by discretization or in fact by exactly the same proof we gave above. □

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 4.12 (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

$$\mathbb{E}(|X_1|) = \sum_{k \geq 0} k2^{-1} \int_{k2^{-1}}^{(k+1)2^{-1}} f_X(x) dx + \sum_{k \geq 1} k2^{-1} \int_{-k2^{-1}}^{(-k+1)2^{-1}} f_X(x) dx.$$

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 $n \to \infty$. □

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:

$$\mathbb{E}(U) = (b-a)^{-1} \int_{\mathbb{R}} x 1_{x \in [a,b]} dx = (b-a)^{-1} \int_a^b x dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

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 4.11 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 4.11 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.

4.3 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, we have already seen that 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 that generalizes the exercise showing that for discrete random variables $\mathbb{E}\left((X-s)^2\right) = \sum_{x \in S_X} (x-s)^2 \mathbb{P}(X = x)$, i.e. that gives us a nice way to calculate expectations of functions of a r.v.:

Proposition 4.13. *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 $X_1, \ldots, X_n$ 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 $\overline{s} = (s_1, \ldots, s_n) \in \mathbb{R}^n$ 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*

$$\mathbb{E}(\phi(\overline{X})) = \int_{\mathbb{R}^n} \phi(\overline{x}) f_{\overline{X}}(\overline{x}) d\overline{x}.$$

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. Let us start from the discrete case, which works exactly like Exercise 3 on Sheet 10 after checking that if ϕ is measurable then $\phi(X)$ is still a discrete random variable. Let us still spell it out in the notes.

So let S_ϕ denote the support of $\phi(\overline{X})$. By definition, $\phi(\overline{X})$ is integrable iff

$$\sum_{s\in S_\phi} |s|\mathbb{P}(\phi(\overline{X})=s)<\infty$$

and then

$$\mathbb{E}(\phi(\overline{X}))=\sum_{s\in S_\phi} s\mathbb{P}(\phi(\overline{X})=s).$$

By the law of total probability we can write $\mathbb{P}(\phi(\overline{X})=s)=\sum_{\overline{x}\in S_{\overline{X}}} 1_{\phi(\overline{x})=s}$ and thus the the whole expression can be written as

$$\sum_{s\in S_\phi} x\sum_{\overline{x}\in S_{\overline{X}}} 1_{\phi(\overline{x})=s}\mathbb{P}(\overline{X}=\overline{x})=\sum_{\overline{x}\in S_{\overline{X}}}\mathbb{P}(\overline{X}=\overline{x})\sum_{s\in S_\phi} s1_{\phi(\overline{x})=s},$$

where we can change the order of summation as the series is absolutely summable. To conclude, notice that for any fixed $\overline{x}\in\mathbb{R}^n$, we have that $\sum_{s\in S_\phi} s1_{\phi(\overline{x})=s}=\phi(\overline{x})$.

The case of the random variables with density is admitted i.e. non-examinable, but I still give the proof for those interested.

To prove the case for random variables with density, we use discretizations - we set $\phi_n(\overline{x})=2^{-n}\lfloor\phi(\overline{x})2^n\rfloor$. Then - given integrability - we have that

$$\mathbb{E}(\phi_n(\overline{X}))=\sum_{k\in\mathbb{Z}} k2^{-n}\mathbb{P}(\phi_n(\overline{X})=k2^{-n}).$$

Now, given that $\phi^{-1}([a,b))$ are Riemann-measurable, we can write

$$k2^{-n}\mathbb{P}(\phi_n(\overline{X})=k2^{-n})=\int_{\mathbb{R}^n} 1_{\overline{x}\in\phi^{-1}([k2^{-n},(k+1)2^{-n}))}k2^{-n}f_{\overline{X}}(\overline{x})d\bar{x}.$$

Again by absolute summability [11] we can switch the order of sum and integration to get

$$\mathbb{E}(\phi_n(\overline{X}))=\int_{\mathbb{R}^n} f_{\overline{X}}(\overline{x})\sum_{k\in\mathbb{Z}} 1_{\overline{x}\in\phi^{-1}([k2^{-n},(k+1)2^{-n}))}k2^{-n}d\bar{x}.$$

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

$$\sum_{k\in\mathbb{Z}} 1_{\overline{x}\in\phi^{-1}([k2^{-n},(k+1)2^{-n}))}k2^{-n}=\phi_n(\overline{x}).$$

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 4.12. □

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 $\sum_{n\geq1}\int_{\mathbb{R}}|f_n(x)|dx<\infty$ or $\int_{\mathbb{R}}\sum_{n\geq1}|f_n(x)|dx<\infty$, then $\int_{\mathbb{R}}\sum_{n\geq1}f_n(x)dx=\sum_{n\geq1}\int_{\mathbb{R}}f_n(x)dx$. You have met the analogous result for swapping two sums $\sum_{k\geq1}\sum_{n\geq1}$, and the proof is basically the same.

Proposition 4.14. *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. □

We already saw that if X, Y are independent, then their product factorises. But in fact there is a sort of converse too - X, Y are independent if the expectation factorizes for all continuous functions!

Proposition 4.15. *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*

$$\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}g(X)\mathbb{E}h(Y), \tag{4.2}$$

  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 (4.2) holds.*

Proof. The first part follows similarly to the last proposition:

From Lemma 3.29 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 4.14. 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

$$F_{(X,Y)}(s, t) \leq \mathbb{E}g_{s,n}(X)g_{t,n}(Y) \leq F_{(X,Y)}(s + 2^{-n}, t + 2^{-n}).$$

By assumption

$$\mathbb{E}g_{s,n}(X)g_{t,n}(Y) = \mathbb{E}g_{s,n}(X)\mathbb{E}g_{t,n}(Y)$$

. Now by right-continuity of $F_{(X,Y)}$, we know that $F_{(X,Y)}(s + 2^{-n}, t + 2^{-n})$ converges to $F_{(X,Y)}(s, t)$ and hence also $\mathbb{E}g_{s,n}(X)g_{t,n}(Y)$ does. Further we have seen that $\mathbb{E}g_{s,n}(X)$ converges to $F_X(s)$ and similarly $\mathbb{E}g_{t,n}(Y)$ converges to $F_X(t)$. Thus we conclude that $F_{(X,Y)}(s, t) = F_X(s)F_X(t)$ as desired.

For the other direction, we first observe the following (this will be on the exercise sheet):

Exercise 4.4. *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 independent random variables X, Y we have that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. The discrete case was on the exercise sheet 11.

The general case proceeds again via approximation and is left as an exercise.

□

4.4 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).

This measures the deviation from the mean, and in fact we already saw it when characterising the expectation as a minimzer of deviation:

Definition 4.16 (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 4.13, 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, whose proof was on the example sheet.

Lemma 4.17 (Chebyshev's inequality)**.** *Let X be an integrable random variable with finite variance. Then $\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{Var(X)}{t^2}$.*

4.4.1 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. The precise relation of two random variables is encoded in their joint law, but that can be quite complicated.

Here we introduce a simpler and weaker measure of how two random variables are related, and a way to in some sense measure the level of dependence.

Definition 4.18 (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 4.19 (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.

Lemma 4.20 (Covariance and dependence)**.** *Let X, Y be two random variables of finite positive variance defined on the same probability space.*

- *Then the correlation $\rho(X, Y) \in [-1, 1]$. Further, it is equal to 1 if and only if there exist some $\lambda > 0, c \in \mathbb{R}$ such that $X = \lambda Y + c$ almost surely; it is equal to -1 if and only if there exist some $\lambda > 0, c \in \mathbb{R}$ such that $X = -\lambda Y + c$ almost surely;*
- *Further, if X, Y are independent, integrable with finite variance, then their covariance is zero.*

Proof. The first part follows from the Cauchy-Schwarz inequality.

For the second part we calculate:

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}X)(Y-\mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

But by independence of X, Y we know that $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$ and we conclude. □

Given a random vector, it is often useful to define the covariance between each pair of components.

Definition 4.21 (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 4.5 (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 very nice indeed!

4.5 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 4.22 (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?)

Jensen's inequality in the probabilistic set-up is stated as follows:

Theorem 4.23 (Jensen's inequality)**.** *Let X be an integrable random variable and ϕ a convex function such that $\phi(X)$ is also integrable* [12]*. Then*

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

[12] Recall that a convex function is continuous and thus if X is a random variable, then so is $\phi(X)$

Notice the similarity with 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 hope 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.

As a corollary we have the following simple lemma, saying that the existence of higher moments implies the existence of lower moments too:

Lemma 4.24. *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 $\mathbb{E}|X|^n \geq (\mathbb{E}(|X|^m))^{n/m}$.*

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, we conclude 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 4.25. *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 beautiful result, saying that one can approximate each continuous function on a finite interval arbitrary well using polynomials:

Theorem 4.26 (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. There is a short probabilistic, but non-examinable proof at the end of the subsection. Let us here see how it implies the proposition.

Proof of Proposition 4.25. 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 4.14, 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, $|\mathbb{E}\left(g(X) - P_\epsilon(X)\right)| \leq \mathbb{E}|g(X) - P_\epsilon(X)|$. 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 4.11 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 4.14. □

For variables that do not have finite support, this characterisation can fail for several reasons. First, of course all moments might not exist and then only the few existing moments might not characterize the distribution. Second, even if all moments exist, they might grow too quickly to characterize the distribution:

Exercise 4.6 (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 $n \in \mathbb{N}$, $\mathbb{E}W^n = \mathbb{E}exp(Xn) = \mathbb{E}Y_a^n$.

4.5.1 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 4.27 (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 4.28 (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 4.29 (MGF for random vectors (admitted))**.** *Let $\overline{X}$ be a random vector taking values in $\mathbb{R}^n$ such that $\mathbb{E}e^{\langle \bar{t}, \bar{x}\rangle} < \infty$ for $\bar{t}$ in some open neighbourhood of 0.*[13] *We then call $M_{\overline{X}}(\bar{t}) = \mathbb{E}e^{\langle \bar{t}, \bar{x}\rangle}$ 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 4.30 (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 $t, s \in I$, $M_X(t)M_Y(s) = M_{(X,Y)}((t, s))$.*

I didn't have time to do this proof in the course, so it is admitted. But I will still give the proof here.

[13] Here $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^n$

Proof. Firstly, if X, Y are independent then the condition follows directly from Proposition 4.15. 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 4.29: 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 4.29 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:

Exercise 4.7. *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 4.8. *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:

Definition 4.31 (Characteristic function)**.** *Let X be a random variable. Then*

$$c_X(t) = \mathbb{E}e^{itX} = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$$

is called the characteristic function of X.

The nice thing is that the characteristic function exists for all $t \in \mathbb{R}$ as both $\cos(tX)$ and $\sin(tX)$ are trivially bounded. Moreover, it uniquely characterizes the law of the random variable and in case of random variables with density, it corresponds to the Fourier transform of the density. But this and much more will already topic of a future course...

4.6 ⋆ 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

$$\mathbb{E}(|\widehat{X}\widehat{Y}|) \leq 1. \tag{4.3}$$

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 4.3 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 $\phi(x + y) \geq \phi(x) + c_x y$.

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 4.26. 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 4.13 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 5

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 $A_1, A_2, \dots$ of a sequence happen? This could for example be used to model whether an infectious disease will only have a limited spread
- When can we say something about the limit of the sequence of random variables $X_1, X_2, \dots$? In what senses can we talk about convergence? We have already seen some vague statements in the lines that $Bin(n, \lambda/n)$ converge to Poisson or that the empirical average of i.i.d. random variables converges to its expectation. What are the right mathematical notions and statements?

5.1 Infinite collections of events and random variables

Before stating a few interesting limit theorems, 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. [14]

Recall that if we say E_i is an event we mean 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 $(E_n)_{n\geq 1}$ 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: when we toss 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 infinitely many events E_i happen. Let us first formalise it: one can check that

$$\{\omega \in \Omega : \omega \in E_i \text{ for infinitely many } i\} = \bigcap_{m\geq 1} \bigcup_{n\geq m} E_n.$$

  The event described this way 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

[14] As discussed, it is not trivial to construct a probability space on which we would have an infinite sequence of independent coin tosses, but here we take this for granted.

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 5.1. *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 5.2 (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. □

The short proof might make you suspicious if it is of any use, but we will see shortly how it is for example to obtain convergence of random variables.

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 5.3 (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. On the exercise sheet □

5.2 Convergence of random variables

When we switch from events to sequences of random variables $X_1, X_2, \dots$, the first question is again - which questions can we even ask?

For example some questions that we might be interested in are:

- Is some value $\geq k$ 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 $\bigcup_{n\geq 1}\{X_i \geq k\}$, 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. The fourth one we state as a lemma, but it is easy to check:

Lemma 5.4. *Let $X, X_1, X_2, \dots$ be random variables on a common probability space. Show that the sets $E := \{\omega : (X_i(\omega))_{i\geq 1}$ converges$\}$ and $E_\infty := \{\omega : (X_i(\omega))_{i\geq 1}$ converges to $X(\omega)\}$ are events, i.e. are measurable.*

Proof. Is left to you to do. □

5.2.1 Almost sure convergence and the Law of large numbers

The exercise above introduces also what is maybe the most natural notion of convergence for random variables on the same probability space. 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 $\{\omega \in \Omega : X_n(\omega)$ converges$\}$ as above. 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?

Definition 5.5 (Almost sure convergence)**.** *Let $X, X_1, X_2, \dots$ be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If 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.*

We saw that these events are indeed measurable. An useful and nice criteria for almost sure convergence comes from the Borel-Cantelli I:

Lemma 5.6. *Let $X, X_1, X_2, \dots$ be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If we can find a decreasing sequence $(a_n)_{n\geq 1}$ of non-negative numbers converging to zero such that for the events $E_n := \{\omega : |X_n(\omega) - X(\omega)| > a_n\}$ we have $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, then X_i converges almost surely to X.*

Proof. By Borel-Cantelli I, only finitely many of E_n happen almost surely, i.e. $\mathbb{P}(\cap_{m\geq 1} \cup_{k\geq m} E_k) = 0$. But now observe that

$$\{\omega : X_n(\omega) \text{ does not converge to } X(\omega)\} \subseteq \cap_{m\geq 1} \cup_{k\geq m} E_k.$$

Indeed, for every ω such that $X_n(\omega)$ does not converge to $X(\omega)$, there is some $\epsilon > 0$ and some subsequence $(n_l)_{l\geq 1}$ with $|X_{n_l}(\omega) - X(\omega)| > \epsilon$. In particular if we let k be such with $a_k < \epsilon$, we have that $\omega \in E_{n_l}$ for all $n_l > k$ and thus we conclude. □

One of the most important examples of almost sure convergence is the law of large numbers, of which we already saw one version in Theorem 4.5.

5.2.2 Law of large numbers

Let us first restate a more general version of the Weak law of large numbers, i.e. Theorem 4.5.

Theorem 5.7 (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 $\mathbb{E}X_1$.

As mentioned below, its proof in the case of variables with finite variance is exactly as in the proof of Theorem 4.5. It can be strenghtened to give the strong law of large numbers.

Theorem 5.8 (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 $\mathbb{E}X_1$.

Roughly, both theorems 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. But 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 does not 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.

For the sake of completeness, we rewrite the proof of the special case of WLLN again too, although we stress it is the same proof as that of Theorem 4.5.

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

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^2) = \sum_{i,j \leq n} n^{-2} \mathbb{E}\left[(X_i - \mathbb{E}X_1)(X_j - \mathbb{E}X_1).\right]$$

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. The proof starts very similarly, but then we apply the corollary of Borel-Cantelli lemma from above.

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

$$\mathbb{E}(|S_n - \mathbb{E}X_1|^4) \leq Cn^{-2}.$$

We now apply Lemma 5.6 with sequence $a_n = n^{-1/8}$. Indeed, by Markov's inequality

$$\mathbb{P}(E_n) = (|S_n - \mathbb{E}X_1| > n^{-1/8}) = \mathbb{P}(|S_n - \mathbb{E}X_1|^4 > n^{-1/2}) \leq \frac{\mathbb{E}|S_n - \mathbb{E}X_1|^4}{n^{-1/2}} \leq Cn^{-3/2}$$

and thus $\sum_{n \geq 1} \mathbb{P}(E_n) < \infty$. Hence this Lemma applies the almost sure convergence of S_n to $\mathbb{E}X_1$. □

Remark 5.9. *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 5.10. *Let $E_1, E_2, \ldots$ 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 $1_{E_1}, 1_{E_2}, \ldots$ 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.

5.2.3 Convergence in law

The other important notion of convergence is that of 'convergence in law'. This is a convergence statement about just the laws of random variables and thus applies also to sequences of random variables defined on different probability spaces. Geometrically you think of it as the convergence of either cumulative distribution functions or maybe even more graphically 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 your toss outcomes 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.

In fact, we have already seen this notion when we talked about the convergence of certain Binomial random variables to Poisson random variables, or discrete uniform random variables to discrete continuous random variables.

Definition 5.11 (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! 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 5.12. *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. This is not hard, but is admitted this year.

□

Remark 5.13. *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.*

The most important example of convergence in law is the Central limit theorem.

5.3 Central limit theorem

The final result of the course is the Central Limit Theorem (CLT).

Theorem 5.14 (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 non-trivial result:

Corollary 5.15. *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 $\mathrm{Var}(Y_i) = p(1-p)$.

□

Further if we consider ± 1 valued random variables, then $\sum X_i$ is exactly equal to the number of ones obtained minus the number of minus ones obtained. The law of large numbers says that this number will be roughly $n(p - 1/2)$, where p is the probability of obtaining 1; the CLT describes the fluctuations.

There are many interesting aspects in the statement:

- The scaling factor $1/\sqrt{n}$. This can be explained by a variance calculation. We have
$$\mathrm{Var}(c_n \sum_{i=1}^n (X_i - \mu)) = c_n^2 n \mathrm{Var}(X_1),$$
which forces $c_n = 1/\sqrt{n}$ if we hope to have something of $O(1)$
- Why Gaussian? To see this we observe that if $X_1, X_2, \dots$ are independent centred Gaussians of variance σ^2, then so is $n^{-1/2}\sum_{i=1}^n X_i$! In fact Gaussians are the only random variables satisfying this property!

The proof of the CLT is not examinable this year, but in fact departs from this same idea.

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: for Gaussians the result holds by above. 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 in the non-examinable proof.

5.3.1 Proof of CLT (non-examinable)

The key step described above is encapsulated in the following proposition - we gave a more general statement that implies it in the class.

Proposition 5.16 (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 4.14 - we approximate indicator functions $1_{X<x}$ via smooth functions and thus obtain the convergence the c.d.f at all continuity points.

Lemma 5.17. *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 runs by approximating $F_X(t), F_{X_n}(t)$ by $\mathbb{E}g_t(X), \mathbb{E}g_{t,n}(X)$ for well-chosen g_t and $g_{t,n}$, quite similarly to what we've seen. □

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 5.16 and Lemma 5.17 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

$$f(S_n) - f(Y) = \sum_{k=1}^n f(S_{n,k+1}) - f(S_{n,k}). \tag{5.1}$$

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!