PROBABILITY 2021

JUHAN ARU

[1]

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

Section 0

Introduction

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

A bit of history

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

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

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

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

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

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

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

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

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

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

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

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

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

This course

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

The course will be roughly in three chapters:

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

(2) Random variables - random variables are the central objects of probability theory, they are the random numbers, or other random objects that come up in our probabilistic model. We will see how to describe and study random variables, and meet several random variable that come up more frequently.

(3) Limit theorems - a special case of the Law of Large Numbers says that if you keep on tossing a fair coin, then the proportion of tails will get closer and closer to a half. We will be prove this result, but we will also prove a version of the Central Limit Theorem, discussed above.

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

0.1 Some historical probability models and their limitations

In this section we shortly discuss some preliminary probability models.

Laplace model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Now, the number of all possible sequences of n tosses is exactly 2^n as each toss has two options. On the other hand, the number of outcomes with k heads out of n tosses is given by exactly $\binom{n}{k}$. So using Lemma 0.2 several times for disjoint events of exactly k tosses, we can write

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

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

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

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

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

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

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

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

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

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

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

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

An intermediate model

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

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

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

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

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

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

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

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

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

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

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

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

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

Probability vs area: intermediate continuous model

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

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

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

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

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

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

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

Section 1

Basic notions

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

1.1 Basics of measure spaces and probability spaces

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

1.1.1 Definition of a measure space

For a measure space the structure comes in two bits:

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

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

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

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

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

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

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

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

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

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

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

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

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

(1) Then also $\Omega \in \mathcal{F}$ and if $A, B \in \mathcal{F}$, then also $A \setminus B \in \mathcal{F}$.

(2) For any $n \geq 1$, if $A_1, \dots, A_n \in \mathcal{F}$, then also $A_1 \cup \dots \cup A_n \in \mathcal{F}$ and $A_1 \cap \dots \cap A_n \in \mathcal{F}$.

(3) If $A_1, A_2, \dots, \in \mathcal{F}$, then also $\bigcap_{n\geq 1} A_1 \in \mathcal{F}$.

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

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

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

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

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

1.1.2 Definition of a probability space

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

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

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

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

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

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

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

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

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

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

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

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

Finally, let us end this subsection with a little remark saying that we are not generalizing all models we presented before. Indeed, at the end of Section 0 we discussed that a version of continuous probability could be defined using the Riemann integral. However, let us see that such a construction doesn't easily give rise to a probability space with the definition above: consider $\Omega = [0,1]$ and let $\mathcal{F}$ be the subset of all sets A such that $\mathbf{1}_{\{x\in A\}}$ is Riemann-integrable. Then surprisingly $\mathcal{F}$ is not a sigma-algebra, as shown by the following exercise.

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

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

1.1.3 Some general properties of measure spaces

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof. This proof is on the exercise sheet. □

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

1.1.4 Measurable and measure preserving maps

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

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

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

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

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

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

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

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

Proof. The proof is on the exercise sheet. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2 Probability spaces

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

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

1.2.1 Discrete probability spaces

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

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

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

The proof is a rather boring affair:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[⋆ End of the non-examinable part ⋆]

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2.2 A cautionary tail for continuous probability spaces

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.2.3 Borel σ-algebra on topological spaces

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

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

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

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

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

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

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

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

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

Proof. On the exercise sheet. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.3.1 The uniform or Lebesgue measure

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

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

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

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

Notice that the Lebesgue measure is shift invariant:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- For a fair coin toss we can use a probability measure takes values 0 and 1 with probability 1/2. This probability measure on $(\mathbb{R}, \mathcal{F}_E)$ can be defined by setting $\mathbb{P}_C(F) = 0.5|F \cap \{0,1\}|$ for every $F \in \mathcal{F}_E$. You can easily verify that this is a probability measure and then $F(x) = 0.51_{x \geq 0} + 0.5_{x \geq 1}$
- For the random stick-length we can use the uniform measure on $[0,1]$. We defined it on $[0,1]$ above, but we can extend it to $\mathbb{R}$ by setting $\widetilde{\mathbb{P}}_U(F) = \mathbb{P}_U(F \cap [0,1])$ for every $F \in \mathcal{F}_E$. Again, it is an easy check that this indeed gives a probability measure. $\mathbb{P}_U$ corresponds to the function F defined by $F(x) = x1_{x \in [0,1]} + 1_{x>1}$.
- Finally, measurement errors are often described by what is called the Gaussian measure. The standard Gaussian measure on $\mathbb{R}$ by definition corresponds to the function F defined by $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{x^2}{2}) dx$.

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

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

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

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

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

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

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

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

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

But now

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

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

This finishes the existence part of the theorem. □

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

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

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

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

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

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

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

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

So we are left to show that if $H_1, H_2, \dots$ are in $\mathcal{H}$, then so is $\bigcup_{i \geq 1} H_i$. To prove this, let us fix some $\epsilon > 0$. First, as $\mathbb{P}$ is a probability measure and $A_n = \bigcup_{i=1}^n H_i$ are increasing sets with $\bigcup_{n \geq 1} A_n = \bigcup_{i \geq 1} H_i$, we can choose $n \in \mathbb{N}$ such that $\mathbb{P}(\bigcup_{i \geq 1} H_i) \setminus A_n) < \epsilon/2$. Further, for each H_i with $i = 1 \dots n$ we can pick disjoint intervals or half-lines $U_{i,1}, \dots, U_{i,m_i}$ such that $\mathbb{P}(H_i \Delta (U_{i,1} \cup \dots \cup U_{i,m_i})) < \frac{\epsilon}{2n}$.

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

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

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

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

and hence

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

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

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

and the proposition follows. □

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

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

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

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

Hence we can write

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

1.4.1 Product σ-algebras

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.4.2 Product probability measures on product σ-algebras

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

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

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

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

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

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

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

We will soon see that the product measure describes the situation where all probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ describe independent experiments. Namely, in the product space of $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ for $i \neq j$, consider the events $E_1 = F_1 \times \Omega_2$ and $E_2 = \Omega_1 \times F_1$ with $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$. Then by definition of product measure

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The non-examinable proof of this theorem uses slickly the fact that $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$ are in bijection and the Dyadic lemma we proved in the opposite direction. Namely, from any countable product of infinite fair coin tosses we can construct the uniform probability measure on $([0,1], \mathcal{F}_E)$. But now by looking at this countable product as a countable product of countable products, we can get a countable product of such uniform probability measures! In other words, we can construct a nice measure on $[0,1]^{\mathbb{N}}$. The general case then proceeds as in Theorem 1.26.

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

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

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

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

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

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

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

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

Step 2: the general case.

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

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

The rest follows similarly to Step 1.

□

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

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

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

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

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

1.5 Conditional probability and independence

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

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

1.5.1 Conditional probability

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

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

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

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

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

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

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

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

and countable additivity follows.

□

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

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

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

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

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

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

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

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

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

Proposition 1.45 (Law of total probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Further, let I be countable and $(E_i)_{i\in I}$ be disjoint events with positive probability and such that $\Omega = \bigcup_{i\in I} E_i$. 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 that

$$F = F \cap (\bigcup_{i\in I} E_i) = \bigcup_{i\in I}(F \cap E_i).$$

But $(E_i)_{i\in I}$ are disjoint, so are $(F \cap E_i)_{i\in I}$. Hence by countable additivity

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

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

□

1.5.2 Independence

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

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

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

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

Here are some basic properties of independence:

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

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

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

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

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

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

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

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

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

as desired.

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

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

- mutual independence
- and pairwise independence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.5.3 Bayes' rule

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

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

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

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

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

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

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

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

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

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

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

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

Most often, one hears about Bayes' rule though in the realm of medicine. Let us give an example of this from late spring this year:

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

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

Section 2

Random variables and random vectors

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

Random variables are measurable functions

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

Events as random variables

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

Many more random variables

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

Random variables vs probability measures

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

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

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

Why random variables at all?

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

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

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

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

2.1 (Real) random variables

For concreteness, let us define again:

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

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

the law or the distribution of the random variable X.

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

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

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

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

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

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

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

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

  and

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

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

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

  and

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

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

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

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

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

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

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

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

2.1.1 The cumulative distribution function of a random variable

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

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

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

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

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

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

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

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

Proof. This is on exercise sheet. □

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

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

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

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

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

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

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

Definition 2.6 (Discrete and continuous random variable)**.** *Let X be a random variable. If F_X is piece-wise constant changing value at most countable many times, we then call the X a discrete random variable. F_X is continuous, we call X a continuous random variable.*

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

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

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

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

Proposition 2.7. *Any cumulative distribution function F_X of a random variable X can be written uniquely as convex combination of cumulative distribution functions of a continuous random variable Y_1 and of a discrete random variable Y_2 - i.e. for some $a \in [0,1]$ we have that $F_X = aF_{Y_1} + (1-a)F_{Y_2}$.*

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

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

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

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

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

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

and

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

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

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

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

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

□

We will see how to interpret this result by saying any random variable can be seen as combination of a discrete and continuous random variable. However, to get there we first have to develop some theory, e.g. the notion of independence for random variables. Let us start by just meeting some more random variables.

2.1.2 Discrete random variables

There are several families of laws of discrete random variables that come up again and again. As we will see, sometimes these laws also have very nice mathematical characterizations:

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

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

We use this family of random variables every time we have no a priori reason to prefer one outcome over the other. A fancy mathematical way of saying this would be to say that the uniform law is the only probability law on a finite set that is invariant under permutations of the points.

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

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

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

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

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

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

For a concrete lively example, let's go back to the Erdos-Renyi random graph on n vertices from the example sheet, where each edge is independently included with probability

p. We can then fix some vertex v and consider the random variable M_v giving the number of vertices adjacent to v, i.e. linked to v by an edge. The exercise above shows that this random variable has law $Bin(n-1,p)$.

Geometric random variable

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

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

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

Lemma 2.8 (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 $\mathbb{P}_X(X = 1) = 1$, then X is a degenerate geometric random variable with $p = 1$. So we can suppose that we work in the case $\mathbb{P}_X(X > 1) > 0$.

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 $\mathbb{P}_X(X > l) = (1-p)^l$ for some $p \in (0,1]$. As by the definition of conditional probability

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

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

Now, let us show that each random variable satisfying the memoryless property has the law of a geometric random variable. We have that

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

Thus for a memoryless random variable

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

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

Poisson random variable

Finally, we consider the Poisson random variable: a discrete random variable with values in $\{0\} \cup \mathbb{N}$ and taking the value k with probability

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

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

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

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

Lemma 2.9 (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 the Poisson distribution 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 events described before one could think as follows: suppose that you cut a time-window $[0, 1]$ into n equal pieces of length $1/n$ and in each time window of length $1/n$ the probability of an arrival (say, a visitor coming) is independently λ/n. Then the total number of occurring events is $Bin(n, \lambda/n)$.

But now why did we choose exactly to cut time into n pieces? Maybe it is reasonable to expect that you could cut into arbitrarily small time intervals and the number of arrivals still behaves independently on each interval, and the probability of an arrival scales linearly with time-length. This would correspond to taking the limit $n \to \infty$ in the description, and hence by the previous lemma we see that the Poisson distribution $Poi(\lambda)$ describes the number of events that occurs in the whole time-interval $[0, 1]$.