Probability Theory (MATH-432)

Janne Junnila

December 23, 2021

Contents

Notation

Set theory

$x \in A$	x is an element of the set A
$A \cup B$	union of two sets
$A \cap B$	intersection of two sets
$A \setminus B$	difference of two sets
$A \Delta B$	symmetric difference: $A \Delta B = (A \setminus B) \cup (B \setminus A)$
$\mathcal{P}(A)$	power set of A (the set of all subsets of A)
$(x_n)_{n=1}^N$	finite sequence $x_1, x_2, \dots, x_N$
$(x_n)_{n=1}^\infty$	infinite sequence $x_1, x_2, \dots$
$(x_n)_n$	countable (finite or infinite) family
$(x_i)_{i \in I}$	family indexed by an arbitrary index set I
$\bigcup_{i \in I} A_i$	union of a family of sets
$\bigcap_{i \in I} A_i$	intersection of a family of sets
$\biguplus_{i \in I} A_i$	union of disjoint sets
$\mathrm{Im}(f)$	the image $f(A)$ of a function $f\colon A \to B$
$\lvert A \rvert$	number of elements in the set A

Specific sets

$\mathbb{N}$	natural numbers $1, 2, \dots$
$\mathbb{Z}$	integers
$\mathbb{R}$	real numbers
$\overline{\mathbb{R}}$	extended real numbers $\mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{C}$	complex numbers $x + iy$
$\mathbb{R}^d$	d-dimensional Euclidean space

Measure theory

$\mathcal{F}, \mathcal{G}, \dots$	σ-algebras denoted by fancy letters	Definition 1.1

$\mathcal{B}$	Borel σ-algebra of some topological space	Definition 1.6
$\sigma(\mathcal{A})$	σ-algebra generated by a family of subsets	Definition 1.4
μ, ν	measures	Definition 1.8
$\mathbb{1}_A(x)$	indicator function: 1 if $x \in A$, 0 otherwise	

Probability theory

Ω	sample space	Definition 1.9
ω	outcome, $\omega \in \Omega$	Definition 1.9
$\mathbb{P}$	probability measure	Definition 1.9
$\sigma(X)$	σ-algebra generated by the random variable X	Definition 1.22
$\limsup_n A_n$	the event that infinitely many of A_n happen	Definition 2.1
$\liminf_n A_n$	the event that ultimately all of A_n happen	Definition 2.1

Various notation

$a \wedge b$	minimum of a and b
$a \vee b$	maximum of a and b

Introduction

Probability as an intuitive notion is *probably* very old. Like in that very first sentence, we run into situations where chance has to be estimated all the time – whether it was our gatherer ancestors guessing where to find food, or a modern judge weighing evidence in a court of justice; Games of chance are older than the written history [9].

Finding systematic ways to reason about probabilities on the other hand does not seem to be so old. Gambling was a driving force in early mathematical investigations into probability, starting with Cardano in 16th century who in his study of dice games defined *odds* as the ratio of favorable outcomes to unfavorable ones, and continuing with Pascal and Fermat in 17th century whose work among other things lead to the notion of expected value.

In modern mathematical language we could try to axiomatize some kind of early versions of probability theory suitable for e.g. dice games by postulating the following:

- There is a nonempty finite set Ω of possible outcomes.
- Each outcome $\omega \in \Omega$ has a probability $p_\omega \in [0, 1]$.
- We have $\sum_{\omega\in\Omega} p_\omega = 1$.

One can then proceed by defining other probabilistic concepts such as events, random variables, expectations and independence.

Example. Let $\Omega := \{H, T\} \times \{1, 2, 3, 4, 5, 6\}$ and $p_\omega = 1/12$ for all $\omega \in \Omega$. We can view each pair $(t, d) \in \Omega$ as a simultaneous toss of a fair coin and a 6-sided die.

- A *random variable* is a function $X\colon \Omega \to \mathbb{R}$. For instance $X(t, d) = d$ gives us the value of the die toss, $Y(t, d) = \mathbb{1}_{\{H\}}(t)$ is 1 if the coin toss was heads and 0 otherwise, and $Z(t, d) = 2d - 3\mathbb{1}_{\{T\}}(t)$ would be a random variable which is 2 times the die toss, minus 3 if the coin toss was tails.
- *Events* are subsets of Ω, for instance the event $\{H\} \times \{1, 3, 5\}$ corresponds to tossing a heads and an odd number. The same event could also be defined by using the random variables as $X^{-1}(\{1, 3, 5\}) \cap Y^{-1}(\{1\})$, which we often write as $\{X \in \{1, 3, 5\}\} \cap \{Y = 1\}$.
- The *probability of an event* E is denoted by $\mathbb{P}[E] := \sum_{\omega\in E} p_\omega$.

- The random variables X and Y are *independent*, meaning that

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$$

for all $A, B \subset \mathbb{R}$. This is also reflected in the product structure of the space Ω: Each option for the coin toss has the same options and probabilities for the die throw. ◆

- The *expected value* of a random variable X is the weighted sum

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_\omega,$$

which for our particular X equals $\frac{1}{6}(1 + 2 + \cdots + 6) = \frac{7}{2}$.

The above framework works nicely as long as the sample space Ω is finite (or countable), but modern applications of probability have come a long way from the simple – or often not so simple! – combinatorics of die throws and card games. Indeed, the view that everything is captured by the probabilities of the individual outcomes $\omega \in \Omega$ starts to fall apart when the number of possible outcomes becomes uncountable, such as when trying to choose a random number from the interval $[0, 1]$.

Exercise. Show that one cannot assign probabilities $p_x > 0$ for all $x \in [0, 1]$ in such a way that $\sum_{x \in [0,1]} p_x = 1$. Here we interpret the sum as $\sum_{x \in [0,1]} p_x := \sup_{S \subset [0,1], S \text{ is finite}} \sum_{x \in S} p_x$. ◆

The usual axiomatization of probability theory by Kolmogorov therefore lets go of the idea that the probabilities of individual $\omega \in \Omega$ determine the probabilities of events, and instead directly defines probabilities on the events. This allows us to say for instance that the probability of a uniformly distributed random number on $[0, 1]$ has probability $1/3$ to lie in the interval $[1/3, 2/3]$, even though the probability of hitting any single fixed $x \in [0, 1]$ is 0. This point of view will naturally lead us to measure theory, which will present some technical challenges but also in the end gives a richer framework to work in.[1]

As a result we will see that the underlying probability space Ω largely loses its importance: There are typically infinitely many different ways to choose a set Ω, a bunch of events on it and probabilities for those events, but in the end the only thing we care about it is that Ω is large enough so that it can be used to define the specific events and random variables we want to model in our applications. Thus Ω could intuitively be thought of as an abstract set which

[1] Although the measure theoretic foundation of probability has become the standard, mathematics is flexible and one can wonder about alternative approaches to modeling probability. A curious reader can for instance take a look at [5] for a nice story using nonstandard analysis.

contains the whole future of the universe behind a single ω, and our events and random variables could be just a tiny part of the total randomness included! Yet, it will require from us a considerable amount of measure theoretic work to see that such powerful Ωs do exist within set theory.

After setting up the foundations, we will state and rigorously prove some of the main theorems of basic probability such as the law of large numbers and the central limit theorem. Considerable focus will also be given to different modes of convergence of random variables and associated function spaces of random variables.

Foundations

1.1 σ-algebras

We will start by defining σ-algebras, which will later on be used to model probabilistic events.

Definition 1.1. Let T be any set. A **σ-algebra** on T is a nonempty collection $\mathcal{G}$ of subsets of T that is closed under complementation and taking countable unions. ◆

- Note that it follows from the definition that if $\mathcal{G}$ is a σ-algebra on T, then $\varnothing, T \in \mathcal{G}$, and if $A_1, A_2, \dots \in \mathcal{G}$ then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{G}$.

The pair $(T, \mathcal{G})$ is called a **measurable space** and the elements of $\mathcal{G}$ are called **measurable subsets**. One often says just "measurable subsets of T" if the σ-algebra $\mathcal{G}$ is clear from the context. Any set has at least one σ-algebra on it, as is seen in the following example.

Example 1.2. The set $\{\varnothing, T\}$ is a σ-algebra on T. It is the smallest σ-algebra on T and we call it the **trivial σ-algebra**. There is also a unique largest σ-algebra on T, the power set $\mathcal{P}(T)$. ◆

Lemma 1.3. *Let $(\mathcal{G}_i)_{i\in I}$ be a collection of σ-algebras on T indexed by an arbitrary index set I. Then $\mathcal{G} = \bigcap_{i\in I} \mathcal{G}_i$ is a σ-algebra.*

Proof. As $\varnothing \in \mathcal{G}_i$ for all $i \in I$, we see that $\varnothing \in \mathcal{G}$ so $\mathcal{G}$ is nonempty.

If $A \in \mathcal{G}$ then $A \in \mathcal{G}_i$ for all $i \in I$. Thus $T \setminus A \in \mathcal{G}_i$ for all $i \in I$ and we see that $T \setminus A \in \mathcal{G}$.

Finally if (A_n) is a countable family of sets in $\mathcal{G}$, it is also a countable family of sets in each $\mathcal{G}_i$ and thus its union belongs to each $\mathcal{G}_i$ and hence also to $\mathcal{G}$. □

Definition 1.4. Let $(A_i)_{i\in I}$ be a collection of subsets of T indexed by an arbitrary index set I. Then the **σ-algebra generated by** $(A_i)_{i\in I}$ is given by

$$\sigma((A_i)_{i\in I}) \coloneqq \bigcap \{\mathcal{G} \in \mathcal{P}(\mathcal{P}(T)) : \mathcal{G} \text{ is a } \sigma\text{-algebra on } T \text{ and } \{A_i\}_{i\in I} \subset \mathcal{G}\}. \quad ◆$$

Exercise 1.5 (σ-algebras on finite sets)**.** Assume that T is finite and $\mathcal{G}$ is a σ-algebra on T. Show that there exists a unique way to partition T into disjoint sets $A_1, \dots, A_n \in \mathcal{G}$ such that every set in $\mathcal{G}$ can be expressed as the union of some $A_{i_1}, \dots, A_{i_k}$, $1 \le i_1, \dots, i_k \le n$.

Conversely, show that if $A_1, \dots, A_n$ is a partition of T into disjoint sets, then the collection of all different unions $A_{i_1} \cup \cdots \cup A_{i_k}$ forms a σ-algebra. ◆

A very common situation is when the σ-algebra is generated by a topology.

Definition 1.6. Let T be a topological space. The **Borel σ-algebra** $\mathcal{B}$ on T is the σ-algebra generated by the open sets of T. ◆

It is often easy to find nicer generating collections for the Borel σ-algebra than arbitrary open sets. In particular for the real line we have the following useful lemma.

Lemma 1.7. *The Borel σ-algebra on $\mathbb{R}$ is generated by any of the following collections of sets:*

- *Open intervals (a, b) with $a, b \in \mathbb{Q}$.*
- *Closed intervals $[a, b]$ with $a, b \in \mathbb{Q}$.*
- *Intervals of the form $(-\infty, t)$ with $t \in \mathbb{Q}$.*
- *Intervals of the form $(-\infty, t]$ with $t \in \mathbb{Q}$.*

Proof. Exercise. □

1.2 *Measures*

Having defined σ-algebras we next turn our attention to measures, which are a way to assign a *size* to sets in a σ-algebra in a consistent way.

Definition 1.8. A **measure** μ on a measurable space $(T, \mathcal{G})$ is a countably additive map $\mu\colon \mathcal{G} \to [0, \infty]$. ◆

- Countable additivity means that for any countable (finite or infinite) family (A_n) of *disjoint* sets in $\mathcal{G}$ we have $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.
- By applying this to the empty family we see that $\mu(\varnothing) = 0$.[1]

Given a set T, a σ-algebra $\mathcal{G}$ and a measure μ, we call the triple $(T, \mathcal{G}, \mu)$ a **measure space**. A central object for us will be a special measure space called the probability space.

Definition 1.9. Let Ω be a set, $\mathcal{F}$ a σ-algebra on Ω and $\mathbb{P}\colon \mathcal{F} \to [0, 1]$ a measure such that $\mathbb{P}[\Omega] = 1$. We call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a **probability space**, the set Ω the **sample space** and the measure $\mathbb{P}$ a **probability measure**. The elements of Ω are called **outcomes** and the elements of $\mathcal{F}$ are **events**. For any event $A \in \mathcal{F}$ we call $\mathbb{P}[A]$ the **probability** of A. ◆

[1] Often in the literature countable additivity is defined so that additivity holds for any infinite sequence $A_1, A_2, \dots$ but not a priori for finite families. In this case one needs to also explicitly assume that $\mu(\varnothing) = 0$.

From now on the symbols Ω, $\mathcal{F}$ and $\mathbb{P}$ will always refer to the sample space, the σ-algebra of events and the probability measure of some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Example 1.10. Here are a couple of simple examples of measure spaces.

- Let T be any set and define the **counting measure** μ on the σ-algebra $\mathcal{P}(T)$ by setting
$$\mu(A) \coloneqq \begin{cases} |A|, & \text{if } A \text{ is finite} \\ \infty, & \text{otherwise} \end{cases}$$
for all $A \subset T$.
- If T is finite and nonempty, we may define the **uniform probability measure** ν on T by letting $\nu \coloneqq \mu/|T|$.
- Assume that T is nonempty and fix $x \in T$. The **Dirac delta measure** μ_x at x is defined by setting $\mu_x(A) \coloneqq \mathbb{1}_A(x)$ for all $A \subset T$.
- Assume that Ω is countable and for every $\omega \in \Omega$ we have assigned a probability $p_\omega \in [0,1]$ in such a way that $\sum_{\omega\in\Omega} p_\omega = 1$. Then $\mathbb{P}[A] \coloneqq \sum_{\omega\in A} p_\omega$ is a probability measure on the σ-algebra $\mathcal{F} \coloneqq \mathcal{P}(\Omega)$. ◆

Another central example is the Lebesgue measure, whose existence we will show later on.

Example 1.11. There is a unique measure λ defined on the Borel σ-algebra of $\mathbb{R}^d$ that satisfies

- For any rectangle $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ we have
$$\lambda(R) = \prod_{k=1}^{n} (b_k - a_k).$$

The measure λ is called the **Lebesgue measure** on $\mathbb{R}^d$. ◆

Let us list some basic properties of probability measures.

Lemma 1.12. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

- *If $A, B \in \mathcal{F}$ and $A \subset B$, then $\mathbb{P}[A] \le \mathbb{P}[B]$.*
- *If $(A_n)_{n=1}^\infty$ is a sequence of events, then $\mathbb{P}[\bigcup_{n=1}^\infty A_n] \le \sum_{n=1}^\infty \mathbb{P}[A_n]$.*
- *If $A_1 \subset A_2 \subset A_3 \subset \ldots$, then $\mathbb{P}[\bigcup_{n=1}^\infty A_n] = \lim_{n\to\infty} \mathbb{P}[A_n]$.*
- *If $A_1 \supset A_2 \supset A_3 \supset \ldots$, then $\mathbb{P}[\bigcap_{n=1}^\infty A_n] = \lim_{n\to\infty} \mathbb{P}[A_n]$.*

Proof. Exercise. □

Remark. The above lemma holds also for general measures, provided that in the last item we assume that one of A_n has finite measure – what can go wrong otherwise? ◆

1.3 *Random variables*

Definition 1.13. Let $(T_1, \mathcal{T}_1)$ and $(T_2, \mathcal{T}_2)$ be two measurable spaces.

- A map $f : T_1 \to T_2$ is called **measurable** (w.r.t. the σ-algebras $\mathcal{T}_1$ and $\mathcal{T}_2$) if for all $A \in \mathcal{T}_2$ we have $f^{-1}(A) \in \mathcal{T}_1$.
- If the domain T_1 is a probability space, we call f a **T_2-valued random variable**.
- If furthermore $(T_2, \mathcal{T}_2) = (\mathbb{R}, \mathcal{B})$ we drop "T_2-valued" and simply say that f is a random variable. ◆

Random variables are the meat and butter of probability theory since in the end what really matters are the (joint) laws of the random variables we define. This means that although there are many ways to construct the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, it does not matter which particular way we pick – we are happy to just know that a rich enough probability space exists on which we can define our random variables. In the end we want to be able to simply say something like

> Let X and Y be two independent standard normal random variables and given X and Y let Z be an Poisson random variable with mean $X^2 + Y^2$.

and know that one can construct Ω, $\mathcal{F}$ and $\mathbb{P}$ which are able to host the random variables X, Y and Z. We are not yet there, however.

The following proposition is often useful when checking that a function is measurable.

Proposition 1.14. *Let $(T_1, \mathcal{T}_1)$ and $(T_2, \mathcal{T}_2)$ be two measurable spaces and let $f : T_1 \to T_2$ be a function. Assume that $\mathcal{T}_2$ is generated by some sets $(A_i)_{i \in I}$. Then f is measurable if and only if $f^{-1}(A_i) \in \mathcal{T}_1$ for all $i \in I$.*

Proof. Clearly if f is measurable then the condition holds.

The proof of the opposite direction uses a strategy that is very useful in many theorems regarding σ-algebras, and the reader is advised to memorize it: *To show that some proposition $P(A)$ holds for all sets A in a σ-algebra $\mathcal{T}$ generated by sets $(A_i)_{i \in I}$, it is enough to show that:*

- *The set $\{A \in \mathcal{T} : P(A)\}$ is itself a σ-algebra.*

- *$P(A_i)$ is true for all $i \in I$.*

In the present case $P(A)$ is the proposition "$f^{-1}(A)$ is measurable". Let us thus define

$$\mathcal{G} := \{A \in \mathcal{F}_2 : f^{-1}(A) \in \mathcal{F}_1\}.$$

By assumption we have $f^{-1}(A_i) \in \mathcal{F}_1$ for all $i \in I$. To show that $\mathcal{G}$ is a σ-algebra we check the following:

- $\mathcal{G}$ is closed under complementation: Given $B \in \mathcal{G}$ we have $f^{-1}(B^c) = (f^{-1}(B))^c \in \mathcal{F}_1$, so $B^c \in \mathcal{G}$.
- $\mathcal{G}$ is closed under countable unions: If $(B_n)_n$ is a countable collection of sets in $\mathcal{G}$, then $f^{-1}(\bigcup_n B_n) = \bigcup_n f^{-1}(B_n) \in \mathcal{F}_1$, so $\bigcup_n B_n \in \mathcal{G}$. □

Combining the above proposition with Lemma 1.7 one sees that to check the measurability of a function $f : T \to \mathbb{R}$ it is enough to for example check the measurability of $f^{-1}((-\infty, t))$ for all $t \in \mathbb{R}$ (or $\mathbb{Q}$). This is useful for instance in the proof of the following elementary facts.

Proposition 1.15. *We have the following basic facts on combining random variables.*

- *Assume that X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then $f \circ X$ is a random variable.*
- *Assume that X and Y are random variables, then also $X + Y$, $X - Y$ and XY are random variables.*
- *Assume that X and Y are random variables and $Y(\omega) \neq 0$ for all $\omega \in \Omega$. Then X/Y is a random variable.*

Proof. Exercise. Hint: One can express the set $\{X + Y < t\}$ as the countable union

$$\{X + Y < t\} = \bigcup_{u \in \mathbb{Q}} (\{X < u\} \cap \{Y < t - u\}).$$ □

In addition to sums and products, one often wants to look at the measurability of various limits of random variables. To this end it will be helpful to define the extended reals.

Definition 1.16. We denote by $\overline{\mathbb{R}}$ the **extended real numbers** $\mathbb{R} \cup \{-\infty, \infty\}$ and endow it with the topology generated by the open intervals in $\mathbb{R}$ together with the intervals $[-\infty, x)$ and $(x, \infty]$ for $x \in \mathbb{R}$. The set $\overline{\mathbb{R}}$ with its Borel σ-algebra becomes a measurable space. ◆

Exercise 1.17. Show that X is a $\overline{\mathbb{R}}$-valued random variable if and only if the sets $X^{-1}(\{\pm\infty\})$ and $X^{-1}(A)$ are measurable for all Borel $A \subset \mathbb{R}$. ◆

Proposition 1.18. *Assume that $(X_n)_{n=1}^{\infty}$ is a sequence of random variables. Then*

$$\inf_n X_n, \quad \sup_n X_n, \quad \liminf_{n\to\infty} X_n \quad \textit{and} \quad \limsup_{n\to\infty} X_n$$

are $\overline{\mathbb{R}}$-valued random variables. In particular pointwise limits of random variables are random variables if the limit exists at every point.

Proof. Exercise. □

The most important property of a random variable is its distribution, which is a probability measure on the target space of the variable.

Definition 1.19. Let X be a T-valued random variable. The **law** (or **distribution**) of X is the probability measure $X_*\mathbb{P}$ on T defined by $X_*\mathbb{P}(A) := \mathbb{P}(X^{-1}(A))$.[2] ◆

Example 1.20. Let $\Omega = \{H, T\} \times \{1, \dots, 6\}$ with uniform probability measure on the σ-algebra $\mathcal{P}(\Omega)$, and define the random variables $X(c, d) = \mathbb{1}_{\{H\}}(c)$ and $Y(c, d) = d$. Then the law of $X(c, d)$ equals $\frac{1}{2}(\delta_0 + \delta_1)$ and the law of $Y(c, d)$ equals $\frac{1}{6}\sum_{k=1}^{6} \delta_k$. The random variables $Z = 1 - X$ and $W = \frac{1+(-1)^Y}{2}$ have the same law as X, but their relationships to X are different. For instance $X + Z = 1$ is a constant random variable, while $X + W$ takes value 0 with probability $1/4$, value 1 with probability $1/2$ and value 2 with probability $1/4$. Thus (X, Z) and (X, W) have different joint laws, although their *marginal laws* are the same. ◆

The most common way to define random variables is by giving their law.

Example 1.21. A *standard normal random variable* X is a random variable whose law is given by

$$X_*\mathbb{P}(A) = \int_A \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, dx$$

for any Borel set $A \subset \mathbb{R}$. The function $x \mapsto \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ is called the *probability density function* of X with respect to the Lebesgue measure. ◆

1.4 *Sub-σ-algebras as encoders of information*

Definition 1.22. Let X be a random variable. The **σ-algebra generated by X** is defined by

$$\sigma(X) := \{X^{-1}(A) : A \in \mathcal{B}\},$$

[2] In general if f is a map from a measure space $(T_1, \mathcal{G}_1, \mu)$ to a measurable space $(T_2, \mathcal{G}_2)$, one can define on the latter space a measure ν by setting $\nu(A) = \mu(f^{-1}(A))$ for all $A \in \mathcal{G}_2$. The measure ν is called the **push-forward measure** of μ via the map f and also denoted by $f_*\mu$.

where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$. ◆

Exercise 1.23. Show that $\sigma(X)$ is indeed a σ-algebra. ◆

In the presence of a fixed outcome $\omega \in \Omega$, the σ-algebra $\sigma(X)$ can be thought of as consisting of all the available information about X: If we want to know whether X lies in some particular Borel set A, we can check whether $\omega \in X^{-1}(A)$.

Definition 1.24. Let X and Y be random variables. We say that Y is **measurable w.r.t. X** or **X-measurable** if $\sigma(Y) \subset \sigma(X)$, or equivalently $Y^{-1}(A) \in \sigma(X)$ for all $A \in \mathcal{B}$. ◆

The following theorem makes precise the idea that if Y is X-measurable, then Y can be reconstructed from X.

Theorem 1.25. *Assume that X and Y are random variables such that Y is X-measurable. Then there exists a Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f \circ X$.*

The proof of this theorem will make use of the following approximation result which we will find useful also later on.

Definition 1.26. A random variable X is **simple** if it takes only finitely many different values. ◆

Proposition 1.27. *Let X be a random variable. Then there exists a sequence X_n of simple random variables such that $\lim_{n\to\infty} X_n = X$.*

Proof. Let

$$X_n := \sum_{k=-n^2}^{n^2} \frac{k}{n} \mathbb{1}_{X^{-1}([\frac{k}{n}, \frac{k+1}{n}))}.$$

Then clearly for fixed $\omega \in \Omega$ we have for $n \geq |X(\omega)|$ that

$$X_n(\omega) = \frac{\lfloor nX(\omega) \rfloor}{n},$$

which tends to $X(\omega)$ as $n \to \infty$. □

Proof of Theorem 1.25. Note that X and Y are random variables also in the restricted probability space $(\Omega, \sigma(X), \mathbb{P})$. Thus by Proposition 1.27 there exists a sequence $(Y_n)_{n=1}^{\infty}$ of X-measurable simple functions such that $\lim_{n\to\infty} Y_n = Y$.

Fix $n \geq 1$. Since Y_n is simple, it takes m different values $a_1, \ldots, a_m$ in the sets $A_1, \ldots, A_m \in \sigma(Y_n)$, respectively. As $\sigma(Y_n) \subset \sigma(X)$, we may write $A_k = X^{-1}(B_k)$ for some Borel sets $(B_k)_{k=1}^m$, and hence $Y_n = f_n \circ X$ where f_n is the

Borel measurable function $\mathbb{R} \to \mathbb{R}$ defined by

$$f_n(x) \coloneqq \sum_{k=1}^{m} a_k \mathbb{1}_{B_k}(x).$$

By Proposition 1.18 the function $\tilde{f} = \limsup_{n\to\infty} f_n$ is a measurable function $\mathbb{R} \to \overline{\mathbb{R}}$. We may define a function $f : \mathbb{R} \to \mathbb{R}$ by setting

$$f(x) \coloneqq \begin{cases} 0, & \text{if } \tilde{f}(x) = \pm\infty \\ \tilde{f}(x), & \text{otherwise} \end{cases}.$$

Then f is measurable and $Y = \lim_{n\to\infty} Y_n = \lim_{n\to\infty} f_n \circ X = f \circ X$, since the limit $\lim_{n\to\infty} f_n(z) = f(z)$ holds for all $z \in \mathrm{Im}(X)$. □

Since there is no probability measure involved in the definition, the measurability of one random variable with respect to another does not say much about the distribution of these two variables. Now, the opposite of measurability of X w.r.t. Y would in some sense be to be unable to say anything about X when knowing Y, and interestingly the probability measure becomes important to make a natural definition in this case.

As a first attempt – without defining probabilities – one could try for instance to require that $X(Y^{-1}\{a\})$ does not depend on $a \in \mathbb{R}$, i.e. X always at least has the same possibilities no mattery which value Y takes. This however is not very natural for various reasons, the most important of which is that probabilistically thinking we should not only require that the possible values for X stay the same when conditioning on Y, but also that the probabilities do not depend on Y. This is called independence.

More generally and precisely, two σ-algebras $\mathcal{F}_1, \mathcal{F}_2$ are independent if neither contains probabilistic information about the other, meaning that knowing that $A \in \mathcal{F}_1$ happened does not affect the probability that $B \in \mathcal{F}_2$ happened, i.e. $\mathbb{P}[B|A] = \mathbb{P}[B]$. By the elementary definition of conditional probability we would thus have $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$, and this is what we will actually take as the definition.

Definition 1.28. Let $\mathcal{F}_1, \dots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{F}$. We say that $\mathcal{F}_1, \dots, \mathcal{F}_n$ are **independent** if for all $(A_i)_{i=1}^n \in \prod_{i=1}^n \mathcal{F}_i$ we have

$$\mathbb{P}\Big[\bigcap_{i=1}^{n} A_i\Big] = \prod_{i=1}^{n} \mathbb{P}[A_i].$$

Moreover:

- An arbitrary collection $(\mathcal{F}_i)_{i\in I}$ of σ-algebras is independent if any finite subcollection of it consists of independent σ-algebras.

- Events $(A_i)_{i\in I}$ are independent if $\sigma(\{A_i\})$ are independent.
- Random variables $(X_i)_{i\in I}$ are independent if $\sigma(X_i)$ are independent. ◆

The notion of independence is also tightly tied to products of probability spaces which we will discuss later. The point of the next exercise is to illustrate this idea in the case of at most countable number of outcomes.

Exercise 1.29. Let $(X_k)_{k=1}^n$ be random variables with X_k defined on a probability space $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$ where Ω_k is countable and $\mathcal{F}_k = \mathcal{P}(\Omega_k)$. Define $\Omega \coloneqq \prod_{k=1}^n \Omega_k$, let $\mathcal{F} \coloneqq \mathcal{P}(\Omega)$ and define a function $\mathbb{P}$ on $\mathcal{P}(\Omega)$ by setting

$$\mathbb{P}[\{(\omega_1, \dots, \omega_n)\}] \coloneqq \mathbb{P}_1[\{\omega_1\}] \dots \mathbb{P}_k[\{\omega_n\}]$$

for singletons $(\omega_1, \dots, \omega_n)$ and extending to arbitrary subsets by summation.

Show that $\mathbb{P}$ defines a probability measure and that the random variables $\tilde{X}_k : \Omega \to \mathbb{R}$ given by $\tilde{X}_k((\omega_1, \dots, \omega_n)) \coloneqq X_k(\omega_k)$ are independent and $\tilde{X}_k$ has the same law as X_k. ◆

1.5 *Infinitely many coin tosses*

So far the most complicated random variables we know how to construct are the ones that take at most countably many different values. The purpose of this section is to show how to construct random sequences $(X_n)_{n=1}^\infty$, where X_n are independent Bernoulli random variables. This will have a big impact on our repertoire of random variables, as we will see that by using infinitely many coin tosses we will be able to for instance construct random variables with arbitrary laws on $\mathbb{R}$. With new powers come new responsibilities, however, and we will have to be a bit more careful as somewhat weird phenomena such as nonmeasurable sets will appear as a by-product.

For the construction a natural starting point is to define $\Omega = \{0,1\}^{\mathbb{N}}$ and $X_n((\omega_k)_{k=1}^\infty) \coloneqq \omega_n$. The hard part is in defining the σ-algebra $\mathcal{F}$ and the probability measure $\mathbb{P}$, since the latter cannot be defined on all subsets of Ω simultaneously as is illustrated by the following proposition.

Proposition 1.30. *Let $\Omega = \{0,1\}^{\mathbb{Z}}$ and define the shift-operator $T : \Omega \to \Omega$ which maps $(\omega_k)_{k\in\mathbb{Z}} \mapsto (\omega_{k+1})_{k\in\mathbb{Z}}$.*[3] *By symmetry it would be natural to require that $\mathbb{P}$ is T-invariant and that $\mathbb{P}[\{\omega\}] = 0$ for any single $\omega \in \Omega$. However no such $\mathbb{P}$ can be defined for all sets in $\mathcal{P}(\Omega)$ simultaneously.*

Proof. Assume that such $\mathbb{P}$ exists. Let us say that ω_1 and ω_2 are equivalent if $\omega_1 = T^n\omega_2$ for some $n \in \mathbb{Z}$. By the axiom of choice we can construct a

[3] Note that it does not matter whether we index the sequences using $\mathbb{N}$ or $\mathbb{Z}$ since both are countable. The shift-operator is just easier to define using $\mathbb{Z}$.

set A which contains exactly one representative from each equivalence class. We note that if $\omega = T^n\omega$ for some $n \in \mathbb{Z}$, then ω must be of the periodic form $(\dots, \omega_m, \omega_1, \omega_2, \dots, \omega_m, \omega_1, \dots)$ and the length of the period m divides n. Since there are only countably many such periodic ω's, we may remove them from A without changing the measure of A and obtain a set B. The sets T^nB ($n \in \mathbb{Z}$) are disjoint, there are countably many of them and their union contains every nonperiodic ω. Thus $1 = \mathbb{P}[\bigcup_{n\in\mathbb{Z}} T^nB] = \sum_{n\in\mathbb{Z}} \mathbb{P}[B]$, but this is a contradiction since if $\mathbb{P}[B] \neq 0$, the sum is ∞. □

To overcome this problem our strategy will be to define $\mathbb{P}$ step by step on larger and larger collections of events, eventually ending up with a σ-algebra while all the time carefully ensuring that countable additivity is preserved. To this end, let us begin by calling a set of the form

$$A = \prod_{n=1}^{\infty} A_n, \quad A_i = \{0, 1\} \text{ for all but finitely many } n$$

a **cylinder set** and define

$$\mathbb{P}[A] \coloneqq \prod_{n=1}^{\infty} \mathbb{P}[A_n]$$

for all cylinder sets, where $\mathbb{P}[A_n] \coloneqq |A_n|/2$. Let

$$R \coloneqq \{\bigcup_{k=1}^{n} A^k : A^1, \dots, A^n \text{ are disjoint cylinder sets}\}$$

denote the collection of all finite unions of disjoint cylinder sets. The family R is a nice stepping stone towards a σ-algebra since it forms an *algebra*.

Definition 1.31. Let T be some set and $R \subset \mathcal{P}(T)$. We say that R is an **algebra** if it is nonempty and closed under complementation and taking finite unions. ◆

Exercise 1.32. Show that R is an algebra. Hint: It is probably easiest to do this in steps, showing that:

- Intersection of two cylinder sets is a cylinder set.
- Complement of a cylinder set is in R.
- Deduce the result for arbitrary sets in R. ◆

We next define

$$\mathbb{P}[A] \coloneqq \sum_{k=1}^{n} \mathbb{P}[A^k]$$

for any $A = \biguplus_{k=1}^{n} A^k \in R$. Note that this is well-defined since for m large enough so that $A_j^k = \{0,1\}$ for all $1 \le k \le n$ and $j > m$, we may compute

$$\sum_{k=1}^{n} \mathbb{P}[A^k] = \sum_{k=1}^{n} \prod_{j=1}^{\infty} \frac{|A_j^k|}{2} = \sum_{k=1}^{n} 2^{-m} |\prod_{j=1}^{m} A_j^k| = 2^{-m} |\bigcup_{k=1}^{n} \prod_{j=1}^{m} A_j^k| = 2^{-m} |\pi_m A|,$$

where $\pi_m : \{0,1\}^{\mathbb{N}} \to \{0,1\}^m$ is the projection to the first m coordinates. The right hand side is constant for m large enough, and hence if $A = \biguplus_{k=1}^{\tilde{n}} \tilde{A}^k$ is another representation of A as a union of cylinder sets, we indeed have $\sum_{k=1}^{n} \mathbb{P}[A^k] = \sum_{k=1}^{\tilde{n}} \mathbb{P}[\tilde{A}^k]$. From here it also easily follows that $\mathbb{P}$ is finitely additive on R, namely if $A = \biguplus_{k=1}^{n} A^k$ and $B = \biguplus_{j=1}^{m} B^j$ are two disjoint elements in R, then $A \cup B$ can be represented as $A \cup B = \biguplus_{k=1}^{n} A^k \uplus \biguplus_{j=1}^{m} B^j$ and hence

$$\mathbb{P}[A \cup B] = \sum_{k=1}^{n} \mathbb{P}[A^k] + \sum_{j=1}^{m} \mathbb{P}[A^j] = \mathbb{P}[A] + \mathbb{P}[B].$$

Are we safe now? Remember that we want $\mathbb{P}$ to be countably additive. One thing that could potentially go wrong would be that even though $\mathbb{P}$ is finitely additive on R, there would be some countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint elements of R whose union is also in R but $\mathbb{P}[\biguplus_{n=1}^{\infty} A_n] \neq \sum_{n=1}^{\infty} A_n$. However this is in fact not an issue in our case because of the following lemma.

Lemma 1.33. *There does not exist any infinite sequence $(A_n)_{n=1}^{\infty}$ of disjoint non-empty elements of R such that their union is also in R.*

Proof. Notice that if the union $A = \biguplus_{n=1}^{\infty} A_n$ is in R, then adding $\Omega \setminus A$ to the sequence would give us Ω as the union, so it is enough to prove the claim in the case $A = \Omega$. Moreover, since any element of R is a finite union of cylinder sets, we may without loss of generality assume that all A_n are cylinder sets as well.

Assume that such cylinder sets A^n with $\biguplus_{n=1}^{\infty} A^n = \Omega$ exist. We will construct an element $\omega \in \Omega$ such that $\omega \notin A^n$ for any n, and this will give us the contradiction. The construction proceeds by induction: We let $\omega_1 = 0$ if there are infinitely many A^n such that $0 \in A_1^n$, otherwise there are infinitely A^n for which $1 \in A_1^n$ and we let $\omega_1 = 1$. Similarly, assuming that $\omega_1, \dots, \omega_m$ have been defined, we let $\omega_{m+1} = 0$ if there are infinitely many A^n such that $(\omega_1, \dots, \omega_m, 0) \in A_1^n \times \cdots \times A_{m+1}^n$, otherwise there are infinitely many A^n for which $(\omega_1, \dots, \omega_m, 1) \in A_1^n \times \cdots \times A_{m+1}^n$ and we set $\omega_{m+1} = 1$. But ω so constructed cannot belong to any given A^n, since if m is so large that $A_j^n = \{0,1\}$ for all $j \ge m$, then because the sets were disjoint, no other $A^{n'}$ has any elements with starting coordinates $(\omega_1, \dots, \omega_m)$, which contradicts the construction where at every stage there were infinitely many such $A^{n'}$. □

We have now reached the situation where we have defined $\mathbb{P}$ as a countably additive map[4] on the algebra generated by the cylinder sets. This is actually a scenario where general measure theoretic results start to apply, so let us continue a bit more abstractly and assume that T is some set, R is some algebra of subsets of T and $\mu\colon R \to [0,\infty]$ is a countably additive map with $\mu(T) < \infty$.

The next and final step is to extend μ to the σ-algebra generated by R. To this end we will define the **outer measure** $\mu^*\colon \mathcal{P}(T) \to [0,1]$ by setting

$$\mu^*(A) \coloneqq \inf\Big\{ \sum_{n=1}^{\infty} \mu(A_n) : A_n \in R \text{ for all } n \in \mathbb{N}, A \subset \bigcup_n A_n \Big\}.$$

Lemma 1.34. *The outer measure μ^* satisfies the following properties:*

- $\mu^*(A) < \infty$ *for all* $A \in \mathcal{P}(T)$.
- $\mu^*(A) \leq \mu^*(B)$ *for all* $A, B \in \mathcal{P}(T)$ *such that* $A \subset B$.
- μ^* *is countably subadditive on* $\mathcal{P}(T)$*, meaning that*

  $$\mu^*(\bigcup_n A_n) \leq \sum_n \mu^*(A_n)$$

  for any countable family $(A_n)_n$ *of subsets of* T.
- $\mu^*(A) = \mu(A)$ *for all* $A \in R$.

Proof. Exercise. □

The main idea in extending the domain of μ from R to a σ-algebra consists of extending the domain of μ by setting $\mu(A) \coloneqq \mu^*(A)$ whenever the set A can be approximated by elements of R up to zero μ^*-measure. The key to make this rigorous is to define the pseudometric $d(A,B) = \mu^*(A \Delta B)$ on $\mathcal{P}(T)$ and take the closure of R in this topology. We will next show in a series of claims that $\overline{R}$ is actually a σ-algebra and that μ^* is countably additive when restricted to $\overline{R}$.

Claim: d is indeed a pseudometric.

We leave this as an exercise.

Claim: μ^* is continuous in the pseudometric d.

We see that

$$\mu^*(A) - \mu^*(B) \leq \mu^*(A \cap B) + \mu^*(A \setminus B) - \mu^*(B) \leq \mu^*(A \setminus B) \leq \mu^*(A \Delta B),$$

[4] To be clear – in this context countable additivity means that if $(A_n)_{n=1}^{\infty}$ is a sequence of disjoint sets in R and it also happens that $\bigcup_{n=1}^{\infty} A_n$ is in R (which does not have to be the case in general), then $\mathbb{P}[\bigcup_{n=1}^{\infty} A_n] = \sum_{n=1}^{\infty} \mathbb{P}[A_n]$.

and exchanging the roles of A and B we have

$$|\mu^*(A) - \mu^*(B)| \leq d(A, B).$$

Thus μ^* is Lipschitz and in particular continuous.

Claim: The closure of R under the pseudometric d is a σ-algebra.

- Clearly $\varnothing \in \overline{R}$, so $\overline{R}$ is nonempty.
- If $A \in \overline{R}$, then there exists a sequence $A_n \in R$ such that $d(A_n, A) \to 0$. Since $d(T \setminus A_n, T \setminus A) = d(A_n, A)$, we see that also $T \setminus A_n \to T \setminus A \in \overline{R}$.
- If $A, B \in \overline{R}$, then there exist sequences $(A_n)_{n=1}^{\infty}, (B_n)_{n=1}^{\infty}$ of elements of R such that $A_n \to A$ and $B_n \to B$. We have

  $$\begin{aligned} d(A_n \cup B_n, A \cup B) &= \mu^*((A_n \cup B_n)\Delta(A \cup B)) \leq \mu^*((A_n \Delta A) \cup (B_n \Delta B)) \\ &\leq d(A_n, A) + d(B_n, B), \end{aligned}$$

  which tends to 0 as $n \to \infty$, so $A_n \cup B_n \to A \cup B \in \overline{R}$. By de Morgan's law and the previous bullet we also have $A_n \cap B_n \to A \cap B \in \overline{R}$.
- For A, B, A_n, B_n as above we also have

  $$\begin{aligned} \mu^*(A \cup B) &= \lim_{n\to\infty} \mu(A_n \cup B_n) = \lim_{n\to\infty} (\mu(A_n) + \mu(B_n) - \mu(A_n \cap B_n)) \\ &= \mu^*(A) + \mu^*(B) - \mu^*(A \cap B), \end{aligned}$$

  so in particular for disjoint $A, B \in \overline{R}$ we have $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$. Thus μ^* is finitely additive on $\overline{R}$.
- Finally if $(A_n)_{n=1}^{\infty}$ are disjoint elements of $\overline{R}$, let $B_n := \bigcup_{k=1}^{n} A_k$ (with the convention $B_0 = \varnothing$) and $B = \bigcup_{n=1}^{\infty} A_n$. For any $n \geq 0$ we have

  $$\mu^*(B \Delta B_n) = \mu^*\Big(\bigcup_{k=n+1}^{\infty} A_k\Big) \geq \mu^*\Big(\bigcup_{k=n+1}^{m} A_k\Big) = \sum_{k=n+1}^{m} \mu^*(A_k)$$

  for all $m \geq n + 1$. Thus by letting $m \to \infty$ we get

  $$\mu^*(B \Delta B_n) \geq \sum_{k=n+1}^{\infty} \mu^*(A_k)$$

  and by subadditivity the inequality is actually an equality. Since $\mu^*(B)$ is finite, we see that $\sum_{k=1}^{\infty} \mu^*(A_k) < \infty$, and thus

  $$\lim_{n\to\infty} \mu^*(B \Delta B_n) = \lim_{n\to\infty} \sum_{k=n+1}^{\infty} \mu^*(A_k) = 0.$$

Hence $B_n \to B \in \overline{R}$, which finishes the proof that $\overline{R}$ is a σ-algebra.

Claim: μ^* is a measure when restricted to $\overline{R}$.

Let $(A_k)_{k=1}^{\infty}$ be a sequence of disjoint elements of $\overline{R}$. We saw above that $\lim_{n\to\infty} \bigcup_{k=1}^{n} A_k = \bigcup_{k=1}^{\infty} A_k$, and hence by the continuity and finite additivity of μ^* we have

$$\mu^*(\bigcup_{k=1}^{\infty} A_k) = \lim_{n\to\infty} \mu^*(\bigcup_{k=1}^{n} A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu^*(A_k) = \sum_{k=1}^{\infty} \mu^*(A_k).$$

Since $\sigma(R) \subset \overline{R}$, we have proven the following general extension result, apart from the claim on uniqueness.

Theorem 1.35 (Carathéodory's extension theorem)**.** *Let R be an algebra on T on which a countably additive map $\mu\colon R \to [0,1]$ has been defined with $\mu(T) = 1$. Then μ extends uniquely to a probability measure on $\sigma(R)$.*

The uniqueness will follow from a general result that states that probability measures that agree on a π-system P also agree on the σ-algebra generated by P.

Definition 1.36. Let P be a collection of subsets of a set T. Then P is a **π-system** if $A, B \in P$ implies that $A \cap B \in P$. ◆

Definition 1.37. Let D be a collection of subsets of a set T. Then D is a **λ-system** (or **Dynkin-system**) if

- $\varnothing \in D$,
- if $A \in D$, then $A^c \in D$, and
- if (A_n) is a countable family of disjoint elements of D, then $\bigcup_n A_n \in D$.

◆

These definitions can be thought of as splitting the conditions of a σ-algebra into two separate parts.

Lemma 1.38. *Assume that $\mathcal{G}$ is a collection of subsets of a set T that is both a π-system and a λ-system. Then $\mathcal{G}$ is a σ-algebra.*

Proof. Exercise. □

What makes the separation of conditions useful is that typically checking that something is a π-system is easy, and for λ-systems the fact that you only need to check the countable union condition for disjoint sets typically plays well together with the countable additive condition of measures.

Just as with σ-algebras, any intersection of λ-systems is a λ-system, and hence for any family of subsets $P \in \mathcal{P}(T)$ one can define the smallest λ-system $\lambda(P)$ containing P by taking the intersection of all λ-systems that contain P. The following theorem shows that an analogue of Lemma 1.38 also works when taking λ-extensions of π-systems.

Theorem 1.39 (π-λ theorem)**.** *If P is a π-system, then $\lambda(P) = \sigma(P)$.*

Proof. It is enough to show that $\lambda(P)$ is a π-system. We do this in three steps.

Step I: *For all $B \in \lambda(P)$ the set $\mathcal{G}_B := \{A \in \lambda(P) : A \cap B \in \lambda(P)\}$ is a λ-system.*
We clearly have $\varnothing \in \mathcal{G}_B$. Moreover, if $A \in \mathcal{G}_B$, then

$$A^c \cap B = (A \cup B^c)^c = ((A \cap B) \cup B^c)^c \in \lambda(P)$$

so $A^c \in \mathcal{G}_B$. Finally if $(A_n)_{n=1}^{\infty}$ is a sequence of disjoint elements of $\mathcal{G}$, then

$$\Big(\bigcup_{n=1}^{\infty} A_n\Big) \cap B = \bigcup_{n=1}^{\infty}(A_n \cap B) \in \lambda(P),$$

so $\bigcup_{n=1}^{\infty} A_n \in \mathcal{G}$.

Step II: *If $A \in \lambda(P)$ and $B \in P$, then $A \cap B \in \lambda(P)$.*
Let us fix $B \in P$. In this case clearly $P \subset \mathcal{G}_B$, so we see that $\mathcal{G}_B$ is a λ-system containing P and contained in $\lambda(P)$, so we must have $\mathcal{G}_B = \lambda(P)$.

Step III: *If $A, B \in \lambda(P)$, then $A \cap B \in \lambda(P)$.*
This time we fix $B \in \lambda(P)$. By the second step we again see that $P \subset \mathcal{G}_B$ and hence $\mathcal{G}_B = \lambda(P)$, which proves the claim. □

We are finally ready to show that probability measures are determined by their values on a π-system generating the σ-algebra.

Theorem 1.40. *Assume that $(T, \mathcal{G})$ is a measurable space on which two probability measures μ and ν have been defined. Assume further that $P \subset \mathcal{G}$ is a π-system with $\sigma(P) = \mathcal{G}$ and that $\mu(A) = \nu(A)$ for all $A \in P$. Then $\mu = \nu$.*

Proof. Let $\mathcal{F} := \{A \in \mathcal{G} : \mu(A) = \nu(A)\}$. Then by assumption $P \subset \mathcal{F}$ and by Theorem 1.39 it is enough to show that $\mathcal{F}$ is a λ-system. Clearly $\varnothing \in \mathcal{F}$. Moreover, if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, since

$$\mu(A^c) = 1 - \mu(A) = 1 - \nu(A) = \nu(A^c).$$

Finally, if $(A_n)_{n=1}^{\infty}$ is a sequence of disjoint elements of $\mathcal{F}$, then

$$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n) = \sum_{n=1}^{\infty} \nu(A_n) = \nu\Big(\bigcup_{n=1}^{\infty} A_n\Big),$$

so $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. □

As a corollary the uniqueness part of Theorem 1.35 follows.

Extending measures is a somewhat delicate topic. Let us close the section by giving two exercises for the curious reader willing to delve deeper into the business.

Exercise 1.41 (Carathéodory's condition)**.** In the literature the σ-algebra $\overline{R}$ is often defined via a characterisation called *Carathéodory's condition*. A set $A \subset T$ is said to satisfy Carathéodory's condition if

$$\mu^*(E) = \mu^*(E \setminus A) + \mu^*(E \cap A)$$

for all $E \subset T$. Show that A satisfies Carathéodory's condition if and only if $A \in \overline{R}$.

The advantage of Carathéodory's condition is that it perhaps more easily adapts to the situation where the measure we are extending is not finite. In that case one needs multiple metrics to generate the right topology, see [4].

The disadvantage is that I find it a bit magical/opaque, and for finite measures the closure-approach might actually be a bit faster. ◆

Exercise 1.42. Let T be some set and $P \subset \mathcal{P}(T)$. We say that P is a **prealgebra**[5] if the following conditions hold.

- We have $T \in P$.
- If $A, B \in P$, then $A \cap B$ and $A \setminus B$ can be expressed as finite unions of disjoint sets in P.

Show that Theorem 1.35 still holds if we replace the assumption that P is an algebra by the assumption that P is a prealgebra. ◆

One might also at first think that perhaps we could start with a measure μ defined on a π-system S and then extend it to the σ-algebra generated by S. Indeed, in view of Theorem 1.40 this is a natural thought since the uniqueness of the extension would automatically be guaranteed. Unfortunately the information contained in a π-system does not guarantee the existence of an extension, and indeed there are simple counter examples where an extension does not exist.

Exercise 1.43. Construct a π-system S on some set T and a function $\mu\colon S \to [0, \infty]$ which is countably additive, but which cannot be extended to a measure on $\sigma(S)$. Here countably additive on S means that whenever (A_n) is a countable

[5] This is nonstandard terminology. A similar structure has been used in the lecture notes [8], where it was called a "semi-anneau", but in English semiring usually means a slightly less general structure.

family of disjoint sets in S such that their union A is also in S, then $\mu(A) = \sum_n \mu(A_n)$. ◆

1.6 *Uniform measure on* $[0,1]$

Armed with infinitely many coin tosses it is easy to construct a uniform probability measure λ on $([0,1], \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra generated by the closed intervals $[a,b]$ with $0 \le a \le b \le 1$. The measure λ is called the Lebesgue measure on $[0,1]$.

Theorem 1.44. *There exists a unique measure λ on $([0,1], \mathcal{B})$ which satisfies $\lambda([0,t]) = t$ for all $t \in [0,1]$.*

Proof. The uniqueness is clear since the intervals $[0,t]$ with $t \in [0,1]$ form a π-system that generates $\mathcal{B}$.

To show the existence, let us consider a sequence $(X_n)_{n=1}^{\infty}$ of independent Bernoulli random variables constructed in the previous section and define the random variable

$$X := \sum_{n=1}^{\infty} X_n 2^{-n}.$$

Then $X \in [0,1]$ always and X is measurable since it is the limit of random variables $\sum_{n=1}^{N} X_n 2^{-n}$ as $N \to \infty$.

Let us define $\lambda := X_* \mathbb{P}$ to be the law of X. Assume that $t \in [0,1]$ has the binary representation $t = \sum_{n=1}^{\infty} t_n 2^{-n}$ with $t_n \in \{0,1\}$. Since

$$\lambda(\{t\}) = \mathbb{P}[X = t] = 0,$$

we have $\lambda([0,t]) = \lambda([0,t))$, and

$$\begin{aligned}\lambda([0,t)) &= \mathbb{P}[X < t] = \mathbb{P}[\{X_i < t_i \text{ in the first index } i \text{ where } X_i \neq t_i\}] \\ &= \sum_{i=1}^{\infty} \mathbb{P}[X_i < t_i] \prod_{j=1}^{i-1} \mathbb{P}[X_j = t_j] = \sum_{i=1}^{\infty} 2^{-i} t_i = t.\end{aligned}$$ □

From the proof we also see the following.

Corollary 1.45. *There exists a probability space on which one can define a random variable X with uniform distribution on the interval $[0,1]$.*

Note that one could also ask for a random variable U which is uniform in the open interval $(0,1)$. This can be obtained by setting

$$U(\omega) = \begin{cases} X(\omega), & \text{if } X(\omega) \notin \{0,1\} \\ \frac{1}{2}, & \text{otherwise} \end{cases}.$$

The variables X and U have actually the same law since we are just redefining X in a set of measure 0.

1.7 *Distribution functions and arbitrary laws on* $\mathbb{R}$

Definition 1.46. Let X be a real-valued random variable. The **cumulative distribution function (c.d.f.)** $F_X : \mathbb{R} \to [0,1]$ of X is defined by

$$F_X(x) := \mathbb{P}[X \leq x]. \qquad \blacklozenge$$

The basic properties of F_X are given in the following lemma.

Lemma 1.47. *Let X be a real-valued random variable with c.d.f. F_X. Then*

- *F_X increases monotonically from 0 to 1 as x goes from $-\infty$ to ∞.*
- *F_X is right-continuous.*

Proof. Exercise. □

An important aspect of the c.d.f. is that it determines the law of the random variable.

Theorem 1.48. *Let X and Y be two random variables with $F_X = F_Y$. Then X and Y have the same law.*

Proof. Let P be the π-system formed by the closed intervals $(-\infty, x]$, $x \in \mathbb{R}$. By Lemma 1.7 we have $\sigma(P) = \mathcal{B}$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$. By definition the law of a random variable X is the measure $X_*\mathbb{P}$ defined on $\mathcal{B}$, and we have $X_*\mathbb{P}((-\infty, x]) = F_X(x)$. Thus by Theorem 1.40 if $F_X = F_Y$, the measures $X_*\mathbb{P}$ and $Y_*\mathbb{P}$ agree on P and hence on $\mathcal{B}$. □

We will next look at going from c.d.f.'s to random variables.

Theorem 1.49. *Let $F : \mathbb{R} \to [0,1]$ be a right-continuous monotonically increasing function with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Then there exists a probability space on which one can define a random variable with c.d.f. F.*

Proof. Let us define the quantile function $G(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}$ for $t \in (0,1)$ and let U be a uniform random variable on $(0,1)$. Then by the right-continuity of F the random variable $G(U)$ satisfies

$$\mathbb{P}[G(U) \leq t] = \mathbb{P}[\inf\{x \in \mathbb{R} : F(x) \geq U\} \leq t] = \mathbb{P}[F(t) \geq U] = F(t)$$

so $G(U)$ has the right c.d.f. □

Let us close this section with the following helpful result.

Lemma 1.50. *One can simultaneously construct on a common probability space a countable number of independent random variables $(X_n)_{n=1}^{\infty}$ with c.d.f.s $(F_n)_{n=1}^{\infty}$.*

Proof. Exercise. □

Spaces of random variables

2.1 *Borel–Cantelli lemma*

Given an infinite sequences of events one often wants to know whether infinitely many of them happen, or even more strongly whether there are only finitely many events that do *not* happen.

Definition 2.1. Let $(A_n)_{n=1}^{\infty}$ be a sequence of events.

- The **limsup event** $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ is the event that infinitely many of the events A_n happen simultaneously.
- The **liminf event** $\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$ is the event that eventually (starting from some random index n_0) all the events A_n happen.

◆

The Borel–Cantelli lemma states that if the probabilities of the events A_n decrease quickly enough, then with probability one only finitely many A_n happen.

Theorem 2.2 (Borel–Cantelli lemma)**.** *Let* $(A_n)_{n=1}^{\infty}$ *be events. If* $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$*, then* $\mathbb{P}[\limsup_{n\to\infty} A_n] = 0$.

Proof. Since $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$, for any $\varepsilon > 0$ we may pick an $n_0 \in \mathbb{N}$ such that $\sum_{n=n_0}^{\infty} \mathbb{P}[A_{n_0}] < \varepsilon$. Then

$$\mathbb{P}[\limsup_{n\to\infty} A_n] \leq \mathbb{P}[\bigcup_{k=n_0}^{\infty} A_k] \leq \sum_{k=n_0}^{\infty} \mathbb{P}[A_k] < \varepsilon.$$

As ε was arbitrary, this proves the claim. □

There is also a partial converse of this lemma in the case of independent events.

Theorem 2.3 (Second Borel–Cantelli lemma)**.** *Let* $(A_n)_{n=1}^{\infty}$ *be a sequence of independent events. If* $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$*, then* $\mathbb{P}[\limsup_{n\to\infty} A_n] = 1$.

Proof. Exercise. □

Example 2.4 (Records – taken from [3])**.** Let $X_1, \dots, X_n$ be the winning scores in year n of some annual sports competition. We assume that $X_1, \dots, X_n$ are independent and identically distributed and that their common c.d.f. is continuous.

Let $E_n \coloneqq \{X_n \geq \max(X_1, \dots, X_{n-1})\}$ be the event that a new record is made in year n. We leave it as an exercise to check that $\mathbb{P}[E_n] = \frac{1}{n}$ and that $E_1, E_2, \dots$ are independent.

Then since $\sum_{n=1}^{\infty} \mathbb{P}[E_n] = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$, we see that new records are made infinitely many times.

On the other hand this does not happen too often: Let $F_n = E_n \cap E_{n+1}$ be the event that records are broken in two consecutive years. Then

$$\sum_{n=1}^{\infty} \mathbb{P}[F_n] = \sum_{n=1}^{\infty} \frac{1}{n(n+1)} < \infty,$$

so with probability one this happens only finitely many times. ◆

2.2 *The space L^0 and convergence in probability*

Let U be a uniform random number on the interval $[0, 1]$. Then we saw that for any fixed $x \in [0, 1]$ we have $\mathbb{P}[U = x] = 0$. Thus it follows that

$$\mathbb{P}[U \in [0, 1]] = \mathbb{P}[U \in [0, 1] \setminus \{1/2\}] = 1.$$

In fact even $\mathbb{P}[U \in [0, 1] \setminus \mathbb{Q}] = 1$, since we are only removing countably many points. All three events $\{U \in [0, 1]\}$, $\{U \in [0, 1] \setminus \{1/2\}\}$ and $\{U \in [0, 1] \setminus \mathbb{Q}\}$ are therefore equivalent in a probabilistic sense. This motivates the following definition.

Definition 2.5. Let A be an event.

- We say that A happens **almost surely (a.s.)** if $\mathbb{P}[A] = 1$ and **almost never** if $\mathbb{P}[A] = 0$.
- In the latter case A is called a **null set** of $\mathbb{P}$.
- If $A = \Omega$, then A happens **surely**. ◆

With this terminology established, let us turn to the main topic of this section: How do we tell two random variables apart?

Certainly if two random variables X and Y are equal almost surely, there should not be any difference between them in a probabilistic sense. Thus it is natural to define the quotient space

$$L^0 \coloneqq \{X \colon \Omega \to \mathbb{R} : X \text{ measurable}\}/\sim,$$

where $\sim$ is the equivalence relation that identifies random variables X and Y such that $\mathbb{P}[X \neq Y] = 0$. We will define other L^p-spaces later on for $p > 0$, but for now let us just assume this notation.

Remark. It is important to note that the space L^0 depends not only on the σ-algebra $\mathcal{F}$ but also on the measure $\mathbb{P}$. Thus two random variables which were equivalent under $\mathbb{P}$ might not be equivalent under another probability measure $\widetilde{\mathbb{P}}$ and vice versa. With some abuse of terminology we will however still continue calling the elements of L^0 random variables. ◆

The following exercise shows that L^0 is a vector space.

Exercise 2.6. Let X, X', Y, Y' be random variables such that $X = X'$ and $Y = Y'$ almost surely. Show that $X + Y = X' + Y'$ almost surely and also that if $c \in \mathbb{R}$, then $cX = cX'$ almost surely. ◆

Thus identifying random variables work mostly very nicely. One has to however be a little bit more careful when taking limits.

Proposition 2.7. *Let X_n be random variables that converge almost surely. Then there exists a random variable X such that $X_n \overset{a.s.}{\to} X$, and the same holds if we replace each X_n with a random variable $X_n' \overset{a.s.}{=} X_n$ and X with a variable $X' \overset{a.s.}{=} X$. Thus almost sure convergence is well-defined for elements of L^0.*

Proof. Exercise. □

Remark. Proposition 2.7 has some subtlety to it so for clarity let us note the following.

- It might be good to write out what we mean when we say that X_n converges almost surely. Quite literally, this means that the set $\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}$ has probability 1 (one can show that the set is always measurable).

  Similarly when we say that $X_n \overset{a.s.}{\to} X$, we mean that the set $\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$ is measurable and has probability 1.

- The above proposition *does not* say that if X_n converges almost surely to some function X that then X is measurable. This holds in general if and only if $\mathbb{P}$ is *complete*, meaning that if A has $\mathbb{P}[A] = 0$, then all subsets $B \subset A$ are measurable and $\mathbb{P}[B] = 0$. ◆

By moving from pointwise defined random variables to equivalence classes in L^0 we have thus retained the nice vector space structure and even pointwise limits work nicely when they are replaced by almost sure limits. On the other hand we have ensured that all the elements in L^0 are at least honestly different (without the extra redundancy caused by a.s. equal random variables) and we can now turn to the question of *how different*.

A natural way to compare two random variables $X, Y \in L^0$ would be to ask how big is $\mathbb{P}[|X - Y| > \lambda]$ for any given $\lambda > 0$. Via a nice trick this kind of thinking can even be refined into an honest metric.

Definition 2.8. The **Ky Fan metric** $d_{KF} \colon L^0 \times L^0 \to [0,1]$ is defined by

$$d_{KF}(X,Y) \coloneqq \inf\{\varepsilon \geq 0 : \mathbb{P}[|X - Y| > \varepsilon] \leq \varepsilon\}$$

for $X, Y \in L^0$. ◆

Theorem 2.9. *The pair (L^0, d_{KF}) is a complete metric space.*

Proof. We leave showing that d_{KF} is a metric as an exercise.

Let us show the completeness. Assume that $(X_n)_{n=1}^{\infty}$ is a Cauchy sequence in L^0. Since it is enough to show that X_n has a converging subsequence, we may assume that $d_{KF}(X_n, X_m) < 2^{-m}$ for $n \geq m \geq 1$. Hence we get in particular that $\mathbb{P}[|X_{n+1} - X_n| > 2^{-n}] \leq 2^{-n}$ for all $n \geq 1$. By Borel–Cantelli thus almost surely $|X_{n+1} - X_n| \leq 2^{-n}$ for n large enough, and we see that the series

$$X_1 + \sum_{n=1}^{\infty} (X_{n+1} - X_n) =: X$$

converges almost surely. Finally we note that $X_n \overset{a.s.}{\to} X$ implies that for any $\varepsilon > 0$ we have

$$\mathbb{P}\Big[\bigcap_{n_0} \bigcup_{n \geq n_0} \{|X_n - X| > \varepsilon\}\Big] = 0.$$

Hence we may pick for any $\varepsilon > 0$ an integer n_0 so large that $\mathbb{P}[|X_n - X| > \varepsilon] \leq \varepsilon$ for $n \geq n_0$. This shows that $d_{KF}(X_n, X) \leq \varepsilon$ and that X_n converges to X under the metric d_{KF}. □

From the above proof we also see the following important important facts.

Proposition 2.10. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables.*

- *If X_n converges in L^0, then it has a subsequence that converges almost surely.*
- *If X_n converges almost surely, then it converges in L^0.*

Convergence in the metric d_{KF} is often called convergence in probability, and it is equivalent to the following definition.

Definition 2.11. Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. We say that X_n **converge in probability** to a random variable X if

$$\lim_{n \to \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$$

for all $\varepsilon > 0$. ◆

Proposition 2.12. *A sequence X_n of random variables converges in probability to a random variable X if and only if $d_{KF}(X_n, X) \to 0$.*

Proof. Assume first that X_n converge in probability to X. Fix $\varepsilon > 0$ and choose n_0 so large that $\mathbb{P}[|X_n - X| > \varepsilon] \leq \varepsilon$ for all $n \geq n_0$. Then by definition

$$d_{KF}(X_n, X) = \inf\{\varepsilon : \mathbb{P}[|X_n - X| > \varepsilon] \leq \varepsilon\} \leq \varepsilon$$

for $n \geq n_0$. Since ε was arbitrary we see that $d_{KF}(X_n, X) \to 0$.

Conversely, assume that $d_{KF}(X_n, X) \to 0$. We want to show that

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$$

for any fixed $\varepsilon > 0$. Notice that for any $\delta \in (0, \varepsilon)$ there exists $n_0 \geq 1$ such that for all $n \geq n_0$ there exists $s < \delta$ for which $\mathbb{P}[|X_n - X| > s] \leq s$. But this also implies that $\mathbb{P}[|X_n - X| > \varepsilon] \leq \mathbb{P}[|X_n - X| > s] \leq s \leq \delta$, so since δ was arbitrary, we have that $X_n \xrightarrow{\mathbb{P}} X$. □

We will next consider approximation in L^0 by simple random variables.

Definition 2.13. We say that a random variable $X \in L^0$ is simple if it has a representative that is simple according to Definition 1.26, and we denote the set of all simple variables in L^0 by S. ◆

It is easy to check that S is closed under addition and scalar multiplication, so S is actually a vector subspace of L^0. Since we know (by Proposition 1.27) that for any random variable X there exists a sequence of simple random variables converging to X almost surely, and that almost sure convergence implies convergence in probability, we obtain the following.

Proposition 2.14. *The set S is dense in L^0.*

Let us close this section with the following useful result, which also shows that convergence in probability only depends on the null sets of the measure.

Proposition 2.15. *Let $(X_n)_{n=1}$ and X be random variables. Then $X_n \xrightarrow{\mathbb{P}} X$ if and only if for every subsequence $(X_{n_k})_{k=1}^{\infty}$ there exists a further subsequence $(X_{n_{k_m}})_{m=1}^{\infty}$ such that $X_{n_{k_m}} \xrightarrow{a.s.} X$.*

Proof. If $X_n \xrightarrow{\mathbb{P}} X$, then every subsequence of it converges also in probability, hence has a further subsequence that converges almost surely.

Conversely, assume that almost surely converging sub-sub-sequences exist but X_n does not converge in probability to X. Then there exist $\varepsilon > 0$ and a subsequence $(X_{n_k})_{k=1}^{\infty}$ for which $\mathbb{P}[|X_{n_k} - X| > \varepsilon] > \varepsilon$ for all $k \geq 1$. However by

assumption there exists a further subsequence $(X_{n_{k_m}})_{m=1}^{\infty}$ such that $X_{n_{k_m}} \overset{a.s.}{\to} X$, hence $X_{n_{k_m}} \overset{\mathbb{P}}{\to} X$, which is a contradiction. □

As a corollary we easily obtain the following.

Proposition 2.16. *Assume that $X_n \overset{\mathbb{P}}{\to} X$ and $Y_n \overset{\mathbb{P}}{\to} Y$. Then the following hold:*

- $X_n + Y_n \overset{\mathbb{P}}{\to} X + Y$
- $X_n Y_n \overset{\mathbb{P}}{\to} XY$
- *If $Y_n, Y \neq 0$ almost surely, then $X_n/Y_n \overset{\mathbb{P}}{\to} X/Y$.*
- *If $g\colon \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \overset{\mathbb{P}}{\to} g(X)$.*

Proof. Exercise. □

2.3 *The space L^∞*

The next space we will take a look at is the space L^∞, which consists of almost surely bounded random variables.

In L^0 the metric mostly ignored the size of the difference between two random variables – it was enough that they were close in a large set but outside that set they could differ a lot. In L^∞ on the other hand the maximal difference is all that matters.

Definition 2.17. We define

$$L^\infty := \{X \in L^0 : \|X\|_{L^\infty} < \infty\},$$

where

$$\|X\|_{L^\infty} := \inf\{\lambda \geq 0 : \mathbb{P}[|X| > \lambda] = 0\}. \qquad \blacklozenge$$

Theorem 2.18. *The space $(L^\infty, \|\cdot\|_{L^\infty})$ is a Banach space.*[1]

Proof. It is clear that $\|X\|_{L^\infty}$ does not depend on the representative of $X \in L^0$.

To check that $\|\cdot\|_{L^\infty}$ is a norm, let us show the triangle inequality and leave the other properties for the reader to check. We have

$$\begin{aligned}\mathbb{P}[|X+Y| > \|X\|_{L^\infty} + \|Y\|_{L^\infty}] &\leq \mathbb{P}[|X| > \|X\|_{L^\infty} \text{ or } |Y| > \|Y\|_{L^\infty}] \\ &\leq \mathbb{P}[|X| > \|X\|_{L^\infty}] + \mathbb{P}[|Y| > \|Y\|_{L^\infty}] = 0,\end{aligned}$$

[1] Recall that Banach space means a complete normed space.

and hence

$$\|X\|_{L^\infty} + \|Y\|_{L^\infty} \in \{\lambda \geq 0 : \mathbb{P}[|X| > \lambda = 0]\},$$

whence

$$\|X + Y\|_{L^\infty} \leq \|X\|_{L^\infty} + \|Y\|_{L^\infty}.$$

Let us then show completeness. Assume that X_n is a Cauchy sequence in L^∞. Then for all $n, m \geq 1$ the events $A_n := \{|X_n| \leq \|X_n\|_{L^\infty}\}$ and $A_{n,m} := \{|X_n - X_m| \leq \|X_n - X_m\|_{L^\infty}\}$ have probability 1 and thus also the event

$$A := \bigcap_{n=1}^{\infty} A_n \cap \bigcap_{n,m=1}^{\infty} A_{n,m}$$

has probability 1. For all $\omega \in A$ the sequence $(X_n(\omega))_{n=1}^\infty$ is Cauchy in $\mathbb{R}$ since we have the inequality

$$|X_n(\omega) - X_m(\omega)| \leq \|X_n - X_m\|_{L^\infty} \quad (n, m \geq 1) \tag{2.1}$$

and X_n is Cauchy in L^∞. By the completeness of $\mathbb{R}$ we therefore see that X_n converges a.s., and by Proposition 2.7 there exists a measurable X such that $X_n \overset{a.s.}{\to} X$. In fact, X has a pointwise defined representative given by $X(\omega) := \lim_{n\to\infty} X_n(\omega)\mathbb{1}_A(\omega)$ and we will work with this representative. Moreover, since the upper bound (2.1) is uniform in ω, we can pick n_0 so large that $\|X_{n_0} - X_m\|_{L^\infty} \leq 1$ for all $m \geq n_0$ and then

$$|X(\omega)| = \lim_{m\to\infty} |X_m(\omega) - X_{n_0}(\omega) + X_{n_0}(\omega)| \leq \|X_{n_0}\|_\infty + 1$$

for all $\omega \in A$, and thus $X \in L^\infty$. Similarly

$$|X_n(\omega) - X(\omega)| = |X_n(\omega) - \lim_{m\to\infty} X_m(\omega)| \leq \sup_{m\geq n} \|X_n - X_m\|_{L^\infty} \to 0$$

uniformly for $\omega \in A$ as $n \to \infty$, so $\lim_{n\to\infty} \|X_n - X\|_{L^\infty} = 0$. □

From the proof above we see that the convergence in L^∞ is very strong and implies in particular convergence almost surely.

Proposition 2.19. *If $X_n \to X$ in L^∞, then $X_n \overset{a.s.}{\to} X$.*

By looking at the proof of Proposition 1.27 one can easily check that the same proof also shows the following.

Proposition 2.20. *The set of simple random variables S is dense in L^∞.*

2.4 *Expectation and the space L^1*

Our next goal is to define the Lebesgue integral $\int_\Omega X\, d\mathbb{P}$ for so called integrable random variables X. The space of integrable random variables is also called the L^1-space.

In the case of a probability space the integral is called **expectation** and we use the notation $\mathbb{E}[X] \coloneqq \int_\Omega X\, d\mathbb{P}$. We will later see how to define the integral for other measure spaces.

We will construct the space L^1 and the expectation map $\mathbb{E}\colon L^1 \to \mathbb{R}$ by first defining $\mathbb{E}[X]$ on simple random variables $X \in S$ and then approximating other random variables by elements of S.

Let $X \in S$ be a simple random variable and fix a representative X_0 of X that takes only finitely many values. Then we have

$$X_0 = \sum_{k=1}^{n} a_k \mathbb{1}_{E_k}$$

where n is the number of distinct values attained by X_0, $a_k \in \mathbb{R}$ are the values themselves and $E_k = X_0^{-1}(\{a_k\})$. We then define

$$\mathbb{E}[X] \coloneqq \sum_{k=1}^{n} a_k \mathbb{P}[E_k].$$

We note that this definition does not depend on the representative X_0, since the sets E_k with positive measure can only differ by a set of measure 0 between representatives.

Two basic properties of $\mathbb{E}$ for simple functions are given in the following lemma.

Lemma 2.21. *Let $X, Y \in S$. The expectation satisfies the following.*

- *Linearity:* $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ *and* $\mathbb{E}[cX] = c\mathbb{E}[X]$ *for* $c \in \mathbb{R}$.
- *(A special case of) Hölder's inequality:* $|\mathbb{E}[XY]| \leq \mathbb{E}[|X|]\|Y\|_{L^\infty}$.

Remark. Above $\mathbb{E}[|X|]$ is well-defined since $X \in S \Rightarrow |X| \in S$. ◆

Proof. **Linearity:** The scalar multiplication part is clear. For the claim on the sum assume that X and Y have the representations $X = \sum_{k=1}^{n} a_k \mathbb{1}_{E_k}$ and $Y = \sum_{k=1}^{m} b_k \mathbb{1}_{D_k}$. Then the sets $E_j \cap D_k$ partition Ω and on the part $E_j \cap D_k$ the random variable $X + Y$ takes the value $a_j + b_k$. Hence

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_{j,k} (a_j + b_k)\mathbb{P}[E_j \cap D_k] = \sum_{j,k} a_j \mathbb{P}[E_j \cap D_k] + \sum_{j,k} b_k \mathbb{P}[E_j \cap D_k] \\ &= \mathbb{E}[X] + E[Y].\end{aligned}$$

Hölder's inequality: We can w.l.o.g. assume that $\mathbb{P}[D_k] \neq 0$ for all k, and then

$$\begin{aligned}|\mathbb{E}[XY]| &= |\sum_{j,k} a_j b_k \mathbb{P}[E_j \cap D_k]| \leq \sum_{j,k} |a_j| \mathbb{P}[E_j \cap D_k] \max\{|b_k| : k = 1, \dots, m\} \\ &= \mathbb{E}[|X|]\|Y\|_{L^\infty}.\end{aligned}$$

□

The above theorem shows that $\mathbb{E}$ is a linear functional on S which satisfies the triangle inequality $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$ for $X \in S$ (substitute $Y = 1$ in Hölder's inequality). Note also that from the triangle inequality one also gets monotonicity: If $X \leq Y$, then $\mathbb{E}[Y - X] \geq |\mathbb{E}[Y - X]| \geq 0$, so $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Let us define the norm $\|\cdot\|_{L^1} : S \to [0, \infty)$ by setting $\|X\|_{L^1} = \mathbb{E}[|X|]$. It is indeed a norm since by monotonicity we have the triangle inequality

$$\|X + Y\|_{L^1} = \mathbb{E}[|X + Y|] \leq \mathbb{E}[|X| + |Y|] = \|X\|_{L^1} + \|Y\|_{L^1}$$

and the other required properties are easy to see from the definition of $\mathbb{E}$.

We have now defined a normed space $(S, \|\cdot\|_{L^1})$. Our strategy next is to take the Banach space completion of S under the norm $\|\cdot\|_{L^1}$ and show that it can in fact be viewed in a natural way as a subset of L^0, a subset which we will then call L^1. Let us begin by recalling the following basic theorem regarding completions of normed spaces.

Theorem. *Let $(V, \|\cdot\|)$ be a normed vector space over $\mathbb{R}$. Then there exists a Banach space $\hat{V}$ and a mapping $\iota : V \to \hat{V}$ such that ι is a linear isometry and $\iota(V)$ is dense in $\hat{V}$. The space $\hat{V}$ is called a **completion** of V and it is unique up to isometric isomorphisms.*

Moreover, if $f : V \to U$ is a uniformly continuous map to a complete metric space U, then f extends uniquely to a uniformly continuous map $\hat{f} : \hat{V} \to U$.

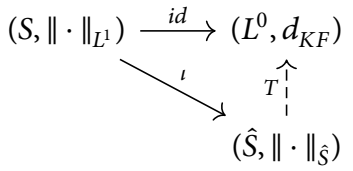
We will take this result for granted. If the reader has not seen it before or just wants to refresh their memory, we direct them to Appendix A for a proof of the second part and Appendix B for a proof of the existence of completions of normed spaces.

Let now $\hat{S}$ be any completion of S under the norm $\|\cdot\|_{L^1}$ and denote the norm in $\hat{S}$ by $\|\cdot\|_{\hat{S}}$. We begin by noting that the Ky Fan metric is weaker than the L^1-norm.

Lemma 2.22. *If $X, Y \in S$, then $d_{KF}(X, Y) \leq \sqrt{\mathbb{E}[|X - Y|]}$.*

Proof. Exercise. Hint: Show that $\mathbb{P}[|Z| > \varepsilon] \leq \varepsilon^{-1}\mathbb{E}[|Z|]$ for all $Z \in S$ and $\varepsilon > 0$ and apply this in the case $Z = X - Y$. □

In particular the identity map $S \to L^0$ is linear and continuous (and hence uniformly continuous), so it admits a continuous linear extension $T : \hat{S} \to L^0$. This is illustrated in the following commutative diagram:



Our next task is to show that T is injective – this will allow us to define $L^1 \coloneqq T(\hat{S})$ and view L^1 as a copy of $\hat{S}$ sitting inside L^0. For this we will need the following lemma.

Lemma 2.23. *Let $(X_n)_{n=1}^\infty$ be a sequence of simple random variables which converges in probability to 0 and also such that $\iota(X_n)$ converges in $\hat{S}$ to some $\hat{X} \in \hat{S}$. Then $\hat{X} = 0$.*

Proof. It is enough to show that $\|X_n\|_{L^1} \to 0$. For any $m \geq 1$ and $\varepsilon > 0$ we have

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}[|X_n|] &\leq \limsup_{n\to\infty} \mathbb{E}[|X_n| \mathbb{1}_{\{|X_n|\leq\varepsilon\}}] + \limsup_{n\to\infty} \mathbb{E}[|X_n| \mathbb{1}_{\{|X_n|>\varepsilon\}}] \\ &\leq \varepsilon + \limsup_{n\to\infty} \mathbb{E}[|X_n - X_m|] + \|X_m\|_{L^\infty} \limsup_{n\to\infty} \mathbb{P}[|X_n| > \varepsilon] \\ &= \varepsilon + \limsup_{n\to\infty} \|\iota(X_n) - \iota(X_m)\|_{\hat{S}} = \varepsilon + \|\hat{X} - X_m\|_{\hat{S}}.\end{aligned}$$

Letting $m \to \infty$ and $\varepsilon \to 0$ on the right hand side shows the claim. □

Using the above lemma it is easy to see that T is injective: Since T is linear, it is enough to show that if $T(\hat{X}) = 0$ for some $\hat{X} \in \hat{S}$ then $\hat{X} = 0$. But this is now clear since for any such $\hat{X}$ we may pick a sequence $(X_n)_{n=1}^\infty$ in S such that $\iota(X_n) \to \hat{X}$ in $\hat{S}$ and then by assumption

$$0 = T(\hat{X}) = \lim_{n\to\infty} T(\iota(X_n)) = \lim_{n\to\infty} X_n,$$

where the limit is in probability, so by the lemma above $\hat{X} = 0$.

We have thus shown that the map T is an injection that continuously embeds $\hat{S}$ into L^0 and we can define

$$L^1 \coloneqq T(\hat{S}).$$

We also extend the definition of $\|\cdot\|_{L^1}$ from S to L^1 by setting

$$\|X\|_{L^1} \coloneqq \|T^{-1}(X)\|_{\hat{S}}$$

for all $X \in L^1 \setminus S$.

This way L^1 has now become another isomorphic completion of S and we may forget about $\hat{S}$.

At this stage it is probably a good idea to pause a little and collect what we have actually shown into a theorem.

Theorem 2.24. *There exists a unique Banach space $(L^1, \|\cdot\|_{L^1})$ satisfying the following properties:*

- *We have $S \subset L^1 \subset L^0$.*
- *L^1 is a completion of S under the norm $\|\cdot\|_{L^1}$.*
- *L^1 is continuously embedded in L^0, or in other words: convergence in L^1 implies convergence in probability.*

Let us also introduce one more term.

Definition 2.25. A random variable $X \in L^0$ is called **integrable** if $X \in L^1$. ◆

We next note that also the expectation can be extended from S to L^1.

Proposition 2.26. *The map $\mathbb{E}\colon S \to \mathbb{R}$ extends uniquely to a continuous linear map $L^1 \to \mathbb{R}$.*

Moreover, Hölder's inequality still holds: if $X \in L^1$ and $Y \in L^\infty$, then $XY \in L^1$ and

$$|\mathbb{E}[XY]| \leq \|X\|_{L^1}\|Y\|_{L^\infty}.$$

Proof. The extension is clear since L^1 is a completion of S and $\mathbb{E}$ is linear and continuous on S under the $\|\cdot\|_{L^1}$-norm.

For the second claim, let $(X_n)_{n=1}^\infty$ and $(Y_n)_{n=1}^\infty$ be two sequences of simple random variables such that $X_n \to X$ in L^1 and $Y_n \to Y$ in L^∞. By Lemma 2.21 we have

$$|\mathbb{E}[X_nY_n]| \leq \|X_n\|_{L^1}\|Y_n\|_{L^\infty}.$$

Clearly the right hand side tends to $\|X\|_{L^1}\|Y\|_{L^\infty}$, so it is enough to check that $X_nY_n \to XY$ in L^1. This is true because $X_nY_n \xrightarrow{\mathbb{P}} XY$ and the sequence is Cauchy in L^1, since

$$\begin{aligned}\mathbb{E}[|X_nY_n - X_mY_m|] &\leq \mathbb{E}[|X_n - X_m||Y_n| + |Y_n - Y_m||X_m|] \\ &\leq \|X_n - X_m\|_{L^1}\|Y_n\|_{L^\infty} + \|Y_n - Y_m\|_{L^\infty}\|X_m\|_{L^1},\end{aligned}$$

and $\sup_{n\geq 1}\|Y_n\|_{L^\infty}$ and $\sup_{m\geq 1}\|X_m\|_{L^1}$ are bounded. □

Here are some further properties of the expectation.

Proposition 2.27. *The following hold:*

- Triangle inequality*: For any $X, Y \in L^1$ we have $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.*
- Monotonicity*: If $X, Y \in L^1$ and $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.*
- L^∞ embeds continuously in L^1*: If $Y \in L^\infty$, then $Y \in L^1$ and $\|Y\|_{L^1} \leq \|Y\|_{L^\infty}$. In particular convergence in L^∞ implies convergence in L^1.*

- Only size matters: *If $X \in L^0$ and $Y \in L^1$ and $|X| \leq |Y|$, then $X \in L^1$.*
- Reality check: *If $X \in L^1$, then $|X| \in L^1$ and $\|X\|_{L^1} = \mathbb{E}[|X|]$.*

Proof. The first two items follow directly from Hölder's inequality like we saw in the case of simple functions above.

Similarly to see that L^∞ embeds continuously in L^1, take $X = 1$ in Hölder's inequality.

For the second last point we can write $X = Yg(Y)$, where $g(Y) = X/Y$ if $Y \neq 0$ and 0 otherwise. Then $g(Y) \in L^\infty$ and the claim again follows from Proposition 2.26.

For the reality check we notice that applying the previous point to $X = |Z|$ and $Y = Z$ we see that the map $Z \mapsto |Z|$ from L^1 to itself is well-defined. It is also continuous since by monotonicity $\mathbb{E}[||X| - |Y||] \leq \mathbb{E}[|X - Y|]$ for all $X, Y \in L^1$. Hence also the composition $X \mapsto \mathbb{E}[|X|]$ is a continuous map $L^1 \to \mathbb{R}$, and as the equality $\|X\|_{L^1} = \mathbb{E}[|X|]$ holds for all $X \in S$ we see that it must by continuity hold for all $X \in L^1$. □

Let us close this section by giving another common metric for L^0.

Proposition 2.28. *The map $(X, Y) \mapsto \mathbb{E}[|X - Y| \wedge 1] =: d_{L^0}$ defines a complete metric on L^0 which is equivalent to the Ky Fan metric.*

Proof. Exercise. □

2.5 *Uniform integrability and convergence theorems*

We have seen that convergence in L^1 implies convergence in probability. It is therefore natural to ask the following question: if we know that a sequence converges in probability, what extra condition is needed for it to converge in L^1? The answer turns out to be *uniform integrability*, which guarantees that the random variables do not concentrate their mass in smaller and smaller subsets of the probability space.

Example 2.29. Consider the probability space $([0, 1], \mathcal{B}, \lambda)$, where λ is the uniform measure on $[0, 1]$. Define a sequence of random variables X_n by setting $X_n(x) = n\mathbb{1}_{[0,1/n]}(x)$. Then $X_n \xrightarrow{\mathbb{P}} 0$ but $\mathbb{E}[X_n] = 1$ for all n so X_n does not converge in L^1. ◆

The above example presents a typical case of a sequence that is *not* uniformly integrable.

Definition 2.30. Let $(X_i)_{i \in I}$ be a family of random variables in L^1. We say that the family is **uniformly integrable** if $\sup_{i \in I} \mathbb{E}[|X_i|] < \infty$ and for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathbb{E}[|X_i|\mathbb{1}_A] < \varepsilon$$

for all $i \in I$ and $A \in \mathcal{F}$ such that $\mathbb{P}[A] \leq \delta$. ◆

Remark. In analysis one often drops the first condition $\sup_{i\in I} \mathbb{E}[|X_i|] < \infty$, but it is standard to include it in probability theory. The most important practical consequence of this condition is that uniformly integrable random variables are *tight*, meaning that

$$\lim_{\lambda\to\infty} \sup_{i\in I} \mathbb{P}[|X_i| > \lambda] = 0.$$

An alternative definition of uniform integrability then is to say that

$$\lim_{\lambda\to\infty} \sup_{i\in I} \mathbb{E}[|X_i| \mathbb{1}_{\{|X_i|>\lambda\}}] = 0,$$

and we leave the proof as an exercise. ◆

Lemma 2.31. *For any $X \in L^1$ the singleton family $\{X\}$ is uniformly integrable.*

Proof. Let $(A_n)_{n=1}^\infty$ be any sequence of events such that $\mathbb{P}[A_n] \to 0$. It is enough to show that $\mathbb{E}[|X|\mathbb{1}_{A_n}] \to 0$ as $n \to \infty$.

Let $(Y_m)_{m=1}^\infty$ be a sequence in L^∞ such that $Y_m \to X$ in L^1. Then

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}[|X|\mathbb{1}_{A_n}] &\leq \limsup_{n\to\infty} (\mathbb{E}[|X - Y_m|\mathbb{1}_{A_n}] + \mathbb{E}[|Y_m|\mathbb{1}_{A_n}]) \\ &\leq \mathbb{E}[|X - Y_m|] + \|Y_m\|_{L^\infty} \limsup_{n\to\infty} \mathbb{P}[A_n] = \|X - Y_m\|_{L^1}\end{aligned}$$

and the claim follows by letting $m \to \infty$. □

Theorem 2.32. *Let $(X_n)_{n=1}^\infty$ be a sequence in L^1. Then X_n converge in L^1 if and only if X_n converge in probability and the sequence is uniformly integrable.*

Proof. Assume first that $X_n \to X$ in L^1. Then we already know that the sequence converges in probability, so it is enough to check that it is uniformly integrable. For any $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for all events A for which $\mathbb{P}[A] < \delta_0$ we have $\mathbb{E}[|X|\mathbb{1}_A] < \varepsilon/2$. Thus there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have

$$\mathbb{E}[|X_n|\mathbb{1}_A] \leq \mathbb{E}[|X_n - X|\mathbb{1}_A] + \mathbb{E}[|X|\mathbb{1}_A] \leq \varepsilon.$$

On the other hand the family $\{X_1, \dots, X_{n_0-1}\}$ consists of just finitely many random variables each of which is individually uniformly integrable, so there exist $\delta_1, \dots, \delta_{n_0-1} > 0$ for which $\mathbb{E}[|X_n|\mathbb{1}_A] < \varepsilon$ when $\mathbb{P}[A] < \delta_n$, $1 \leq n \leq n_0 - 1$. Picking $\delta = \min(\delta_0, \delta_1, \dots, \delta_{n_0-1})$ proves the claim.

Assume then that X_n converge in probability and that they are uniformly

integrable. It is enough to show that X_n is Cauchy in L^1. Let $\varepsilon > 0$. We have

$$\begin{aligned}\mathbb{E}[|X_n - X_m|] &\leq \mathbb{E}[|X_n - X_m| \mathbb{1}_{\{|X_n - X_m| \leq \varepsilon\}}] + \mathbb{E}[|X_n - X_m| \mathbb{1}_{\{|X_n - X_m| > \varepsilon\}}] \\ &\leq \varepsilon + \mathbb{E}[|X_n| \mathbb{1}_{\{|X_n - X_m| > \varepsilon\}}] + \mathbb{E}[|X_m| \mathbb{1}_{\{|X_n - X_m| > \varepsilon\}}].\end{aligned}$$

Since the sequence is Cauchy in probability, we have

$$\mathbb{P}[|X_n - X_m| > \varepsilon] \to 0$$

as $m, n \to \infty$. Thus by uniform integrability we get

$$\mathbb{E}[|X_n - X_m|] \leq 3\varepsilon$$

for large enough n, m. □

A useful criterion for checking uniform integrability is the following.

Lemma 2.33. *Assume that $(X_i)_{i \in I}$ is a family of random variables and that there exists $Y \in L^1$ such that for all $i \in I$ we have $|X_i| \leq Y$ almost surely. Then the family $(X_i)_{i \in I}$ is uniformly integrable.*

Proof. Obvious since for any event A we have $\mathbb{E}[|X_i| \mathbb{1}_A] \leq \mathbb{E}[Y \mathbb{1}_A]$. □

Corollary 2.34 (Dominated convergence theorem)**.** *Assume that $(X_n)_{n=1}^{\infty}$ is a sequence which converges in probability to X and that there exists $Y \in L^1$ such that we have $|X_n| \leq Y$ a.s. for all $n \geq 1$. Then $X \in L^1$ and*

$$\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Proof. By Lemma 2.33 the sequence $(X_n)_{n=1}^{\infty}$ is uniformly integrable and the claim follows from Theorem 2.32. □

In the rest of the section we will look at expectations of non-negative random variables. To this end we make the following definition.

Definition 2.35. Assume that X is an a.s. nonnegative random variable which is not integrable, i.e. $X \notin L^1$. We then define $\mathbb{E}[X] := \infty$. Here we allow X to also take the value ∞ with positive probability.

More generally any random variable X can be split into its positive and negative parts, $X = X^+ - X^-$ with X^+ and X^- non-negative, and if exactly one of X^+ and X^- is not in L^1, we may define $\mathbb{E}[X] := \pm\infty$ accordingly. ◆

To see that the definition is natural, we note the following.

Theorem 2.36 (Monotone convergence theorem)**.** *Assume that $(X_n)_{n=1}^{\infty}$ is a pointwise increasing sequence of random variables taking values in $[0, \infty]$ and*

let X denote the pointwise limit. Then

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Proof. If $X \in L^1$, we are done by the dominated convergence theorem.

If $X \notin L^1$, then it is enough to show that the increasing sequence $\mathbb{E}[X_n]$ is not bounded. To obtain a contradiction, assume that it is. Then for $m \geq n$ we have

$$\mathbb{E}[|X_m - X_n|] = \mathbb{E}[X_m] - \mathbb{E}[X_n],$$

but since the sequence $\mathbb{E}[X_n]$ converges and is therefore Cauchy, we see that X_n is Cauchy in L^1 and hence converges to X in L^1, which is a contradiction. □

We close this section with one more useful result.

Theorem 2.37 (Fatou's lemma)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables taking values in $[0, \infty]$. Then*

$$\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Proof. Let us write $Y_n = \inf_{k\geq n} X_k$. Then $Y_n \leq X_n$ for all $n \geq 1$, and moreover the sequence $(Y_n)_{n=1}^{\infty}$ increases monotonically. Thus by the monotone convergence theorem

$$\mathbb{E}[\liminf_{n\to\infty} X_n] = \mathbb{E}[\lim_{n\to\infty} Y_n] = \lim_{n\to\infty} \mathbb{E}[Y_n] = \liminf_{n\to\infty} \mathbb{E}[Y_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

□

As an easy corollary we have the following version for random variables converging in probability.

Corollary 2.38. *If $(X_n)_{n=1}^{\infty}$ is a sequence of non-negative random variables converging to some random variable X in probability, then*

$$\mathbb{E}[X] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Proof. Let X_{n_k} be a subsequence such that

$$\lim_{k\to\infty} \mathbb{E}[X_{n_k}] = \liminf_{n\to\infty} \mathbb{E}[X_n].$$

This subsequence contains a sub-sub-sequence $X_{n_{k_m}}$ which converges to X a.s., and by choosing suitable representatives we may actually assume that $X_{n_{k_m}} \to X$ surely without altering the value of any of the expectations. The claim then follows from Fatou's lemma. □

2.6 *Integration on general measure spaces*

Although our main interest is in probability spaces, it is still useful to have the Lebesgue integral defined on other measure spaces as well, the most important case being of course $\mathbb{R}^d$ with the Lebesgue measure. In this section we will briefly and without proofs explain how one can accomplish this by using probability measures and the integral we have already defined.

Definition 2.39. Let $(T, \mathcal{G}, \mu)$ be a measure space. We say μ is **σ-finite** if there exists a countable (finite or infinite) partition $(A_n)_n$ of T into disjoint measurable subsets such that $0 < \mu(A_n) < \infty$ for all n. ◆

Let $(T, \mathcal{G}, \mu)$ be a σ-finite measure space with a partition $(A_n)_n$ as above. Define probability measures μ_n on $\mathcal{G}$ by setting $\mu_n(E) := \frac{\mu(E \cap A_n)}{\mu(A_n)}$ and for a non-negative measurable function $f : T \to [0, \infty]$ set

$$\int_T f \, d\mu := \sum_n \mu(A_n) \int_{A_n} f \, d\mu_n.$$

It is an easy exercise to show that the definition of $\int_T f \, d\mu$ does not depend on the partition A_n. If f is a signed measurable function we can write it as a difference of its positive and negative parts f^+ and f^- and define $\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$, provided that at least one of the integrals on the right hand side is finite. If both of them are finite, we say that f is integrable, and this also defines the class $L^1(\mu)$.

Finally (just for completeness since we won't need it later on), let us note that if $(T, \mathcal{G}, \mu)$ is a general measure space, we may for all measurable non-negative f define

$$\int_T f \, d\mu := \int_{\{x \in T : f(x) > 0\}} f \, d\mu,$$

provided that the restriction of μ to the set $\{x \in T : f(x) > 0\}$ is σ-finite. Otherwise we set $\int_T f \, d\mu = \infty$, which makes sense, since if μ restricted to $\{x \in T : f(x) > 0\}$ is not σ-finite, at least one of the sets $\{x \in T : f(x) \in [2^{-n}, 2^{-n+1})\}$ where $n \in \mathbb{Z}$ must have infinite mass.

We leave it for the reader to check that all the basic results for integrals in Proposition 2.26 and Proposition 2.27 continue to hold, with the exception that $L^\infty(\mu)$ is not anymore necessarily a subset of $L^1(\mu)$. Also the monotone and dominated convergence theorems as well as Fatou's lemma still hold.

2.7 *Absolute continuity of measures*

Let us next discuss another way to characterise σ-finiteness as those measures that are in a sense equivalent to a probability measure. Our starting point will

be the following observation.

Lemma 2.40. *Assume that $(T, \mathcal{G}, \mu)$ is a measure space and that $f : T \to [0, \infty]$ is a non-negative measurable function. Then the map $\nu \colon \mathcal{G} \to [0, \infty]$ defined by*

$$\nu(A) = \int_A f \, d\mu \tag{2.2}$$

is a measure. Moreover for any measurable g we have $g \in L^1(\nu)$ if and only if $gf \in L^1(\mu)$, in which case

$$\int g \, d\nu = \int gf \, d\mu.$$

Proof. Let $(A_n)_{n=1}^{\infty}$ be a sequence of disjoint measurable subsets. By the monotone convergence theorem we have

$$\nu\Big(\biguplus_{n=1}^{\infty} A_n \Big) = \int \mathbb{1}_{\bigcup A_n} f \, d\mu = \int \sum_{n=1}^{\infty} \mathbb{1}_{A_n} f \, d\mu = \sum_{n=1}^{\infty} \int_{A_n} f \, d\mu,$$

so countable additivity holds and ν is a measure. Moreover, if $g = \mathbb{1}_E$ is an indicator function then clearly $\int g \, d\nu = \nu(E) = \int \mathbb{1}_E f \, d\mu$ holds. By linearity the identity also holds in the case where g is a simple function and hence by approximation for all $g \in L^1(\nu)$. □

The relationship between the two measures in (2.2) motivates a couple of definitions.

Definition 2.41. Let $(T, \mathcal{G})$ be a measurable space and μ and ν two measures on $\mathcal{G}$.

- We say that ν has the **Radon–Nikodym property** relative to μ if there exists a measurable function $f : T \to [0, \infty]$ such that

  $$\nu(A) = \int_A f \, d\mu$$

  for all $A \in \mathcal{G}$. The function f is called a **density function** or a **Radon–Nikodym derivative** of ν relative to μ and we often write $f = \frac{d\nu}{d\mu}$.

- We say that ν is **absolutely continuous** w.r.t. μ and write $\nu \ll \mu$ if $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathcal{G}$.

- We say that the measures μ and ν are **equivalent** and write $\mu \sim \nu$ if they have the same null sets, i.e. $\nu \ll \mu$ and $\mu \ll \nu$. ◆

Below are a couple of preliminary observations regarding absolute continuity and the Radon–Nikodym property. The first one is a kind of a chain rule.

Lemma 2.42. *If μ, ν and η are measures on the same σ-algebra and η has the Radon–Nikodym property relative to ν with a density $\frac{d\eta}{d\nu}$, and ν in turn has the Radon–Nikodym property relative to μ with a density $\frac{d\nu}{d\mu}$, then η has the Radon–Nikodym property relative to μ with a density $\frac{d\eta}{d\nu}\frac{d\nu}{d\mu}$.*

Proof. Simple unraveling of definitions. □

The second one is that having the Radon–Nikodym property implies absolute continuity.

Lemma 2.43. *If ν has the Radon–Nikodym property relative to μ, then $\nu \ll \mu$.*

Proof. Trivial since anything integrated over a set of measure 0 equals 0. □

A central nontrivial result in measure theory is that if μ is a probability measure then the converse holds.

Theorem 2.44 (Radon–Nikodym theorem)**.** *Let μ be a probability measure, and let ν be absolutely continuous w.r.t. μ. Then ν has the Radon–Nikodym property relative to μ.*

Proof. See Appendix C. □

Finally, let us state a useful characterisation of σ-finite measure spaces.

Proposition 2.45. *A nonzero measure μ on a measurable space $(T, \mathcal{G})$ is σ-finite if and only if there exists a probability measure ν on $\mathcal{G}$ such that $\mu \sim \nu$ and $\frac{d\mu}{d\nu} < \infty$ almost surely.*

Proof. Assume first that such probability measure ν exists. Pick a representative of $\frac{d\mu}{d\nu}$ that is finite everywhere and consider the disjoint sets $A_k = \{\frac{d\mu}{d\nu} \in [k, k+1)\}$. Then we have that $\bigcup_{k=1}^{\infty} A_k = T$ and moreover $\mu(A_k) = \int_{A_k} \frac{d\mu}{d\nu}\, d\nu \leq k+1$, so μ is σ-finite.

Assume then that μ is σ-finite and let $(A_n)_{n=1}^{\infty}$ be a partition of T such that $\mu(A_n) < \infty$ for all $n \geq 1$. We can then define a probability measure ν by setting

$$\nu(E) = C \int_E \sum_{n=1}^{\infty} \frac{\mathbb{1}_{A_n}}{2^n(1+\mu(A_n))}\, d\mu,$$

where $C = (\sum_{n=1}^{\infty} \frac{\mu(A_n)}{2^n(1+\mu_{A_n})})^{-1}$ is a normalising constant (well-defined since μ is not identically 0). Then clearly $\nu(E) = 0$ if and only if $\mu(E) = 0$ and thus ν and μ are equivalent measures. Moreover, we have

$$\frac{d\mu}{d\nu} = C^{-1} \frac{1}{\sum_{n=1}^{\infty} \frac{\mathbb{1}_{A_n}}{2^n(1+\mu(A_n))}} < \infty$$

almost surely. □

2.8 *Lebesgue measure on* $\mathbb{R}$

We have already seen how to construct the uniform measure λ_0 on $[0,1]$ using infinitely many Bernoulli random variables and in a similar manner we can define λ_n for $n \in \mathbb{Z}$ to be the uniform measure on the interval $[n, n+1]$.

Let us now consider the map $\lambda : \mathcal{B} \to [0,\infty]$ given by

$$\lambda(A) := \sum_{n\in\mathbb{Z}} \lambda_n(A \cap [n, n+1]).$$

It is easy to check that λ is countably additive and hence a measure.

Proposition 2.46. *We have* $\lambda([a,b]) = b - a$ *for all* $-\infty < a < b < \infty$.

Proof. Exercise. □

This is also enough to characterise the measure.

Exercise 2.47. Let μ and ν be σ-finite measures on a measurable space $(T, \mathcal{G})$ that agree on a π-system P generating the σ-algebra. Assume further that there exists a sequence $(A_n)_{n=1}^{\infty}$ of sets in P such that $\bigcup_{n=1}^{\infty} A_n = T$ and $\mu(A_n) < \infty$ and show that then $\mu = \nu$. ◆

It is customary to write $\int f(x)\, d\lambda(x) = \int f(x)\, dx$ when the integrating measure is the Lebesgue measure. The following theorem is of huge practical importance.

Theorem 2.48 (Fundamental theorem of calculus)**.** *Assume that* f *is continuously differentiable on an interval* $[a,b] \subset \mathbb{R}$. *Then*

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

Proof. Exercise. □

Remark. Let us mention that in general it is not hard to show that if a function is Riemann integrable, it is also Lebesgue integrable and the integrals agree.

Secondly, let us also mention that the assumption that the derivative is continuous is in fact not needed. It is enough to assume that the derivative is integrable, see e.g. [6, Theorem 7.21]. We leave the further studies of these topics to a real analysis and/or measure theory course. ◆

Having defined the Lebesgue measure we may now formally say what it means for a distribution to have a probability density function.

Definition 2.49. Let $p_X : \mathbb{R} \to [0, \infty)$ be a measurable function with

$$\int_{-\infty}^{\infty} p_X(x)\, dx = 1.$$

We say that a random variable X has a **probability density function (p.d.f.)** p_X if its law satisfies

$$X_* \mathbb{P}(A) = \int_A p_X(x)\, dx$$

for all $A \in \mathcal{B}$. ◆

Let us close this section with the following change-of-variables formula, which is useful when computing expectations in practice.

Proposition 2.50. *Let X be a random variable. Then for any non-negative measurable $F : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}[F(X)] = \int F(x)\, d(X_* \mathbb{P})(x).$$

Moreover, for general measurable F the composition $F \circ X$ is integrable w.r.t. $\mathbb{P}$ if and only if F is integrable w.r.t. $X_* \mathbb{P}$, and in this case the above equality holds. In particular, if X has a probability density function p_X, then

$$\mathbb{E}[F(X)] = \int F(x) p_X(x)\, dx.$$

Proof. It is clear that if we can show the result for non-negative F, then the general signed case follows by splitting into positive and negative parts. Moreover, the second formula in the case where X has a probability density follows from Lemma 2.40 after we have shown the first formula.

Notice first that if $F = \mathbb{1}_A$ is an indicator function of some Borel set $A \subset \mathbb{R}$, then the formula holds since by definition

$$\mathbb{E}[F(X)] = \mathbb{P}[X \in A] = X_* \mathbb{P}(A) = \int_{\mathbb{R}} F\, d(X_* \mathbb{P}).$$

By linearity the formula thus holds whenever F is simple (takes only finitely many values).

If F is a general non-negative measurable function, then we may consider the sequence $F_n = (\lfloor 2^n F \rfloor / 2^n) \wedge n$, which consists of simple functions and converges monotonously to F. By the monotone convergence theorem then

$$\mathbb{E}[F(X)] = \lim_{n\to\infty} \mathbb{E}[F_n(X)] = \lim_{n\to\infty} \int_{\mathbb{R}} F_n\, d(X_* \mathbb{P}) = \int_{\mathbb{R}} F\, d(X_* \mathbb{P}). \qquad \square$$

Remark. The above change-of-variables formula holds also for T-valued ran-

dom variables where $(T, \mathcal{G})$ is some measurable space, and in particular for random vectors having a probability density in $\mathbb{R}^d$ also the analogue of the second formula in Proposition 2.50 holds. The proof is essentially the same. ◆

2.9 *L^p-spaces for general p*

Let us return again to the setting of probability spaces. So far we have defined three L^p spaces, namely when $p \in \{0, 1, \infty\}$. In this section we will complete the picture to any $p \in [0, \infty]$.

Definition 2.51. The space L^p for $p \in (0, \infty)$ is defined by

$$L^p := \{X \in L^0 : \|X\|_{L^p} < \infty\},$$

where

$$\|X\|_{L^p} := \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$ ◆

We will soon see that $\|\cdot\|_{L^p}$ is a norm for $p \geq 1$. This is not however true for $p < 1$ and in this case one has to think of $\|\cdot\|_{L^p}$ as just being no more than notation.

Let us start by looking at a bunch of (very) useful inequalities.

Theorem 2.52 (Jensen's inequality)**.** *Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be convex, meaning that*

$$\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y)$$

for all $x, y \in \mathbb{R}$, $t \in [0, 1]$. Then for any random variable X such that $\mathbb{E}[X]$ is defined (i.e. at least one of $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ is finite) also $\mathbb{E}[\varphi(X)]$ is defined and we have

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

Proof. Exercise. □

As a corollary we see that $L^q \subset L^p$ when $q \geq p$.

Corollary 2.53. *We have $\|X\|_{L^p} \leq \|X\|_{L^q}$ for $0 < p \leq q \leq \infty$.*

Proof. The case $q = \infty$ is easy, so assume that $q < \infty$. Applying Jensen's inequality with the function $x \mapsto x^{q/p}$ we have

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} = \left((\mathbb{E}[|X|^p])^{q/p}\right)^{1/q} \leq (\mathbb{E}[|X|^q])^{1/q} = \|X\|_{L^q}.$$ □

The next inequality is convenient (among other things) when one wants to derive estimates for expectations of products of random variables.

Theorem 2.54 (Hölder's inequality)**.** *Let $p, q \in [1, \infty]$ be such that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

and assume that $X \in L^p$ and $Y \in L^q$. Then $XY \in L^1$ and

$$\|XY\|_{L^1} \leq \|X\|_{L^p} \|Y\|_{L^q}.$$

Proof. We note first that for any $a, b \geq 0$ we have Young's inequality

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

The case when a or b is 0 is clear, and otherwise we may write $a^p = e^s$ and $b^q = e^t$ for some $s, t \in \mathbb{R}$. Then by the convexity of the exponential function we have

$$ab = e^{\frac{1}{p}s + \frac{1}{q}t} \leq \frac{e^s}{p} + \frac{e^t}{q} = \frac{a^p}{p} + \frac{b^q}{q}.$$

Let us now prove the claim itself. Again the case when either $\|X\|_{L^p}$ or $\|Y\|_{L^p}$ equals 0 is trivial since then X or Y is 0 almost surely and also $\mathbb{E}[|XY|] = 0$. We may thus assume by scaling that $\|X\|_{L^p} = \|Y\|_{L^q} = 1$. Then letting $a = |X|$ and $b = |Y|$ and integrating gives us the inequality:

$$\mathbb{E}[|XY|] \leq \mathbb{E}\left[\frac{|X|^p}{p} + \frac{|Y|^q}{q}\right] = \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_{L^p} \|Y\|_{L^q}. \qquad \square$$

Theorem 2.55 (Minkowski inequality)**.** *Let $X, Y \in L^p$ for $p \in (0, \infty]$. Then*

$$\|X + Y\|_{L^p}^{p \wedge 1} \leq \|X\|_{L^p}^{p \wedge 1} + \|Y\|_{L^p}^{p \wedge 1}.$$

Proof. For $p \in (0, 1)$ we have

$$\|X + Y\|_{L^p}^p = \mathbb{E}[|X + Y|^p] \leq \mathbb{E}[|X|^p] + \mathbb{E}[|Y|^p]$$

by the inequality $(a + b)^p \leq a^p + b^p$ for $a, b \geq 0$.

For $p = 1$ and $p = \infty$ we already know the result from earlier sections.

For $p \in (1, \infty)$ let q be such that $\frac{1}{p} + \frac{1}{q} = 1$. We may assume that $\mathbb{E}[|X+Y|^p] \neq 0$, since otherwise the inequality is trivial. Then by the triangle inequality and

Hölder's inequality we have

$$\begin{aligned}\mathbb{E}[|X+Y|^p] &\le \mathbb{E}[|X||X+Y|^{p-1}] + \mathbb{E}[|Y||X+Y|^{p-1}] \\ &\le \|X\|_{L^p}\||X+Y|^{p-1}\|_{L^q} + \|Y\|_{L^p}\||X+Y|^{p-1}\|_{L^q} \\ &= (\|X\|_{L^p} + \|Y\|_{L^p})(\mathbb{E}[|X+Y|^p])^{1/q},\end{aligned}$$

and the claim follows by dividing by $(\mathbb{E}[|X+Y|^p])^{1/q}$. □

Let us next define

$$d_{L^p}(X,Y) \coloneqq \begin{cases} \mathbb{E}[|X-Y| \wedge 1], & \text{if } p = 0 \\ \|X-Y\|_{L^p}^p, & \text{if } p \in (0,1) \\ \|X-Y\|_{L^p}, & \text{if } p \ge 1. \end{cases}$$

Theorem 2.56. *The space (L^p, d_{L^p}) is a complete metric space for all $p > 0$ and in particular a Banach space for $p \ge 1$.*

Proof. Recall that we already know the result for $p \in \{0, 1, \infty\}$. In other cases it follows from the Minkowski inequality that d_{L^p} is a metric. Moreover, for $p \ge 1$ the metric is given by an actual norm, so L^p is a normed space in this case.

To show that the spaces are complete, we note that if X_n is Cauchy in L^p for $p > 0$, then it is in particular Cauchy in L^0 since

$$\mathbb{E}[|X_n - X_m| \wedge 1] \le \mathbb{E}[|X_n - X_m|^p]$$

if $p \in (0,1)$ and

$$\mathbb{E}[|X_n - X_m| \wedge 1] \le \|X_n - X_m\|_{L^1} \le \|X_n - X_m\|_{L^p}$$

if $p \ge 1$. Thus X_n converges in L^0 to some random variable X. By Corollary 2.38 we then have

$$\|X\|_{L^p}^{p\wedge 1} \le \liminf_{k\to\infty} \|X_k\|^{p\wedge 1} \le \liminf_{k\to\infty} (\|X_k - X_{n_0}\|_{L^p}^{p\wedge 1} + \|X_{n_0}\|_{L^p}^{p\wedge 1}) < \infty$$

where $X_{n_0} \in L^p$ is chosen in such a way that $\|X_k - X_{n_0}\| \le 1$ for all $k \ge n_0$. Thus $X \in L^p$ and similarly

$$\mathbb{E}[|X - X_n|^p] \le \liminf_{k\to\infty} \mathbb{E}[|X_k - X_n|^p] \le \sup_{k\ge n} \mathbb{E}[|X_k - X_n|^p] \to 0$$

as $n \to \infty$ so $X_n \to X$ in L^p. □

Let us close this chapter by giving the following summary of the spaces we have studied:

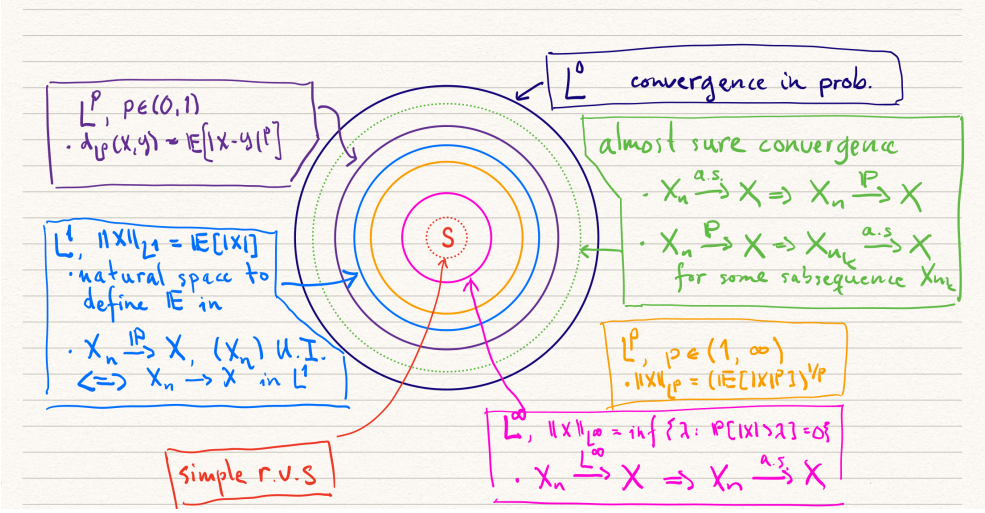


Figure 2.1.: Relationships between L^p-spaces.

- The spaces $(L^p)_{p\in[0,\infty]}$ form a decreasing set of complete metric spaces.
- The embedding $L^p \subset L^q$ for $\infty \geq p \geq q \geq 0$ is continuous. In particular if X_n is a sequence of random variables that converges in L^p, it will also converge in L^q.
- The spaces with $p \geq 1$ are Banach spaces (the metric is given by a norm).
- The smaller spaces are dense inside the larger ones.
- Convergence a.s. does not define a space of its own but is related to the space L^0 in the following way: If a sequence converges a.s., it will converge in L^0. Conversely if a sequence converges in L^0 it will contain a *subsequence* which converges a.s.
- Convergence in L^0 is also known as convergence in probability.

Independence and conditioning

3.1 *Product measures and Fubini theorem*

In elementary geometry one learns that the area of a rectangle equals the product of its width and height. Phrased differently: the two-dimensional measure of the rectangle equals the product of the one dimensional measures of its sides. Taking products of general measures to obtain measures on the product space is a generalization of this simple idea.

Let us first define the product of two σ-algebras.

Definition 3.1. Let $(T_1, \mathcal{G}_1)$ and $(T_2, \mathcal{G}_2)$ be two measurable spaces. The **product σ-algebra** $\mathcal{G}_1 \otimes \mathcal{G}_2$ on $T_1 \times T_2$ is the σ-algebra generated by sets of the form $A_1 \times A_2$ where $A_i \in \mathcal{G}_i$ for $i = 1, 2$. ◆

Product measures are similarly defined by requiring that the measure of a product set is the product of measures.

Definition 3.2. Let $(T_1, \mathcal{G}_1, \mu_1)$ and $(T_2, \mathcal{G}_2, \mu_2)$ be two measure spaces. A measure μ on the product σ-algebra $G_1 \times G_2$ is a **product** of the measures μ_1 and μ_2 if

$$\mu(A_1 \times A_2) = \mu(A_1)\mu(A_2)$$

for all $A_i \in \mathcal{G}_i, i = 1, 2$. In this case $(T_1 \times T_2, \mathcal{G}_1 \otimes \mathcal{G}_2, \mu)$ is called a **product space** of the measure spaces $(T_1, \mathcal{G}_1, \mu_1)$ and $(T_2, \mathcal{G}_2, \mu_2)$. Such a product measure is usually denoted by $\mu_1 \otimes \mu_2$. ◆

The product of two measures is not always unique, but in the case of σ-finite measures this is the case.

Theorem 3.3. *Assume that the measures μ_1 and μ_2 in Definition 3.2 are σ-finite. Then there exists a unique product measure on $\mathcal{G}_1 \times \mathcal{G}_2$.*

Proof when μ_1 and μ_2 are probability measures. Note that the set

$$P \coloneqq \{A_1 \times A_2 : A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2\}$$

is a **semialgebra**, meaning that P is closed under intersections, contains the empty set, and if $A \in P$, then the complement of A can be written as a finite disjoint union of sets in P.

From any semialgebra P one can construct an algebra $\mathcal{A}$ by taking finite unions of sets in P, and one can check that if $\mu\colon P \to [0, \infty]$ is a countably ad-

ditive map, then it admits a countably additive extension to $\mathcal{A}$, and hence one can apply Carathéodory's extension theorem (Theorem 1.35). We leave the details to the reader, since this was also more or less the content of Exercise 1.42.

Anyway, by the above discussion and Theorem 1.35 it is thus enough to check that the map $\mu\colon P \to [0,\infty]$ defined by $\mu(A_1 \times A_2) \coloneqq \mu_1(A_1)\mu_2(A_2)$ is countably additive. Assume thus that $A = B \times C \in P$ is written as $A = \biguplus_{n=1}^{\infty} A_n$ with $A_n = B_n \times C_n \in P$. Then we have

$$\mathbb{1}_B(x)\mathbb{1}_C(y) = \sum_{n=1}^{\infty} \mathbb{1}_{B_n}(x)\mathbb{1}_{C_n}(y)$$

for all $x \in T_1$ and $y \in T_2$. Integrating first over x and then over y and using the monotone convergence theorem gives

$$\mu(A) = \mu_1(B)\mu_2(C) = \sum_{n=1}^{\infty} \mu_1(B_n)\mu_2(B_n) = \sum_{n=1}^{\infty} \mu(A_n),$$

which proves the theorem in the case of probability measures. □

The most important result regarding integration with respect to the product measure is that it can be computed as an iterated integral.

Theorem 3.4 (Fubini's theorem)**.** *Let $(T_1, \mathcal{G}_1, \mu_1)$ and $(T_2, \mathcal{G}_2, \mu_2)$ be σ-finite measure spaces. Then for any integrable $f : T_1 \times T_2 \to \mathbb{R}$ the map $x \mapsto f(x, y)$ is integrable for a.e. $y \in T_2$ and the map $y \mapsto f(x, y)$ is integrable for a.e. $x \in T_1$, the a.e. defined map $x \mapsto \int_{T_2} f(x, y)\, d\mu_2(y)$ is integrable w.r.t. μ_1 and the a.e. defined map $y \mapsto \int_{T_1} f(x, y)\, d\mu_1(x)$ is integrable w.r.t. μ_2, and we have*

$$\begin{aligned}\int_{T_1 \times T_2} f\, d(\mu_1 \otimes \mu_2) &= \int_{T_1} \int_{T_2} f(x, y)\, d\mu_2(y)\, d\mu_1(x) \\ &= \int_{T_2} \int_{T_1} f(x, y)\, d\mu_1(x)\, d\mu_2(y).\end{aligned}$$

Moreover, the above identity also holds for any $f \geq 0$ (even if f is not integrable in which case all integrals are ∞).

Before going to the proof, let us see how knowing the result for probability measures gives us the σ-finite case as well.

Proof of Theorems 3.3 and 3.4 if they hold for probability measures. Let μ_1 and μ_2 be two σ-finite measures. Then by Proposition 2.45 there exist probability measures ν_1 and ν_2 and positive functions f_1 and f_2 so that $d\mu_1 = f_1 d\nu_1$ and

$d\mu_2 = f_2 d\nu_2$. We may then define

$$(\mu_1 \otimes \mu_2)(A) = \int \mathbb{1}_A(x,y) f_1(x) f_2(y) d(\nu_1 \otimes \nu_2)(x,y)$$

which satisfies

$$(\mu_1 \otimes \mu_2)(B \times C) = \int \mathbb{1}_B(x)\mathbb{1}_C(y) f_1(x) f_2(y) d\nu_1(x) d\nu_2(y) = \mu_1(B)\mu_2(C),$$

so it gives a product measure which is unique by Exercise 2.47. Fubini's theorem is also immediate since for any $g \in L^1(\mu_1 \otimes \mu_2)$ we have

$$\begin{aligned}\int g\, d(\mu_1 \otimes \mu_2) &= \int g(x,y) f_1(x) f_2(y) d(\nu_1 \otimes \nu_2) \\ &= \int \int g(x,y) f_1(x)\, d\nu_1(x) f_2(y)\, d\nu_2(y) \\ &= \int \int g(x,y)\, d\mu_1(x)\, d\mu_2(y). \qquad \square\end{aligned}$$

We will now begin preparing for the proof of Theorem 3.4 for probability measures and start with the following useful fact about measurable functions on product spaces.

Proposition 3.5. *Let $(T_1, \mathcal{G}_1)$ and $(T_2, \mathcal{G}_2)$ be measurable spaces and let $(T, \mathcal{G})$ be the product space. Assume that $f : T_1 \times T_2 \to \mathbb{R}$ is $\mathcal{G}$-measurable. Then for a fixed $x \in T_1$ the map $y \mapsto f(x,y)$ is $\mathcal{G}_2$-measurable.*

Proof. Exercise. Hint: Use the fact that every measurable f is a limit of simple functions and deduce that it is enough to prove the theorem in the case where $f = \mathbb{1}_A$ for some set $A \in \mathcal{G}_1 \otimes \mathcal{G}_2$. Use the π-λ theorem. $\square$

The second ingredient we need is a Fubini's theorem for indicator functions.

Lemma 3.6. *Assume that $(T_1, \mathcal{G}_1, \mu_1)$ and $(T_2, \mathcal{G}_2, \mu_2)$ are probability spaces and that $A \in \mathcal{G}_1 \otimes \mathcal{G}_2$. Then the map $x \mapsto \int_{T_2} \mathbb{1}_A(x,y)\, d\mu_2(y)$ is $\mathcal{G}_1$-measurable and*

$$(\mu_1 \otimes \mu_2)(A) = \int_{T_1} \int_{T_2} \mathbb{1}_A(x,y)\, d\mu_2(y)\, d\mu_1(x).$$

Proof. We use the π-λ theorem. Let $\mathcal{A} \subset \mathcal{G}_1 \otimes \mathcal{G}_2$ be the family of all measurable sets for which the claim holds. Clearly all sets of the form $A_1 \times A_2$ with $A_1 \in \mathcal{G}_1$ and $A_2 \in \mathcal{G}_2$ belong to $\mathcal{A}$ so it is enough to check that $\mathcal{A}$ is a λ-system. If $A \in \mathcal{A}$,

then

$$(\mu_1 \otimes \mu_2)(A^c) = 1 - (\mu_1 \otimes \mu_2)(A) = \int_{T_1} \int_{T_2} (1 - \mathbb{1}_A(x,y))\, d\mu_2(y)\, d\mu_1(x)$$
$$= \int_{T_1} \int_{T_2} \mathbb{1}_{A^c}(x,y)\, d\mu_2(y)\, d\mu_1(x),$$

where also the measurability of the map $x \mapsto \int_{T_2} \mathbb{1}_{A^c}(x,y)\, d\mu_2(y)$ is clear so $A^c \in \mathcal{A}$. Finally if $(A_n)_{n=1}^\infty$ are disjoint elements of $\mathcal{A}$, then by the monotone convergence theorem

$$(\mu_1 \otimes \mu_2)(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \int_{T_1} \int_{T_2} \mathbb{1}_{A_n}(x,y)\, d\mu_2(y)\, d\mu_1(x)$$
$$= \int_{T_1} \int_{T_2} \mathbb{1}_{\bigcup_n A_n}(x,y)\, d\mu_2(y)\, d\mu_1(x),$$

where $x \mapsto \int_{T_2} \mathbb{1}_{\bigcup_n A_n}(x,y)\, d\mu_2(y)$ is measurable because it is a sum of measurable functions, so also $\bigcup_{n=1}^\infty A_n \in \mathcal{A}$. □

Proof of Theorem 3.4. It remains to prove the theorem for probability measures. Let us first assume that X is $\mathcal{G}_1 \otimes \mathcal{G}_2$ measurable and non-negative. Then the sequence $X_n = (\lfloor 2^n X \rfloor / 2^n) \wedge n$ of simple functions converges to X pointwise monotonously, and it follows from the monotone convergence theorem that

$$\int\int X\, d\mu_2\, d\mu_1 = \lim_{n\to\infty} \int\int X_n\, d\mu_2\, d\mu_1 = \lim_{n\to\infty} \int X_n\, d(\mu_1 \otimes \mu_2) = \int X\, d(\mu_1 \otimes \mu_2).$$

If X is in $L^1(\mu_1 \otimes \mu_2)$, then the result follows by considering separately the positive and negative parts of X. □

Let us close this section by discussing a little bit products of more than two spaces. In general we have the following definition.

Definition 3.7. Let $(T_i, \mathcal{G}_i)_{i\in I}$ be a family of measurable spaces. Then the **product σ-algebra** $\bigotimes_{i\in I} \mathcal{F}_i$ is the σ-algebra on $\prod_{i\in I} T_i$ generated by the projection maps $\pi_\alpha : (t_i)_{i\in I} \mapsto t_\alpha$ ($\alpha \in I$), i.e.

$$\bigotimes_{i\in I} \mathcal{F}_i := \sigma(\{\pi_\alpha^{-1}(A) : \alpha \in I, A \in \mathcal{G}_\alpha\}). \qquad \blacklozenge$$

In the case where we have finitely many σ-finite measure spaces $(T_k, \mathcal{G}_k, \mu_k)$, $1 \le k \le n$, one can prove that again $\bigotimes_{k=1}^n \mathcal{G}_k$ is generated by measurable rectangles $A_1 \times \cdots \times A_n$, and that there exists a product measure $\mu_1 \otimes \cdots \otimes \mu_n$ on this σ-algebra. Moreover, one can show that if one takes the products iteratively,

then the product is associative and the order does not matter, meaning that $(\mu_1 \otimes \mu_2) \otimes \mu_3 = \mu_1 \otimes (\mu_2 \otimes \mu_3) = \mu_1 \otimes \mu_2 \otimes \mu_3$.

The most important example of a product of multiple measures is of course the d-dimensional Lebesgue measure on $\mathbb{R}^d$.

Definition 3.8. Let λ be the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$. We then define the d-dimensional Lebesgue measure $\lambda_d := \lambda^{\otimes d}$ as the d-fold product of λ with itself.

Moreover, if X is a $\mathbb{R}^n$-valued random variable and f is a measurable function $\mathbb{R}^n \to [0, \infty)$, then we say that X has a p.d.f. f, if $X_* \mathbb{P}$ has the density f w.r.t. λ_n. ◆

Remark. Two technical points are worth mentioning here.

- One can show that the product σ-algebra $\mathcal{B}^{\otimes d}$ is also isomorphic to the Borel σ-algebra on $\mathbb{R}^d$.
- Quite often when people talk about the Lebesgue measure λ_d they mean the measure which is defined on the so called Lebesgue measurable sets, which form a σ-algebra $\mathcal{L}_d \supset \mathcal{B}^{\otimes d}$. The inclusion is strict and in fact one can view $\mathcal{L}_d$ as the completion of $\mathcal{B}^{\otimes d}$ with respect to λ_d, meaning that

  $$\mathcal{L}_d := \{A \cup N : A \in \mathcal{B}^{\otimes d}, N \in \mathcal{N}\},$$

  where

  $$\mathcal{N} := \{N \subset \mathbb{R}^d : \exists N' \in \mathcal{B}^{\otimes d}, N \subset N', \lambda_d(N') = 0\}.$$

  Now, when working with $\mathcal{L}_d$ there is the catch that unlike for the Borel σ-algebras, it is no longer true that $\mathcal{L}_d = \mathcal{L}_n \otimes \mathcal{L}_m$ when $d = n+m$. Thus if one wants to work with the Lebesgue measurable sets instead of Borel sets, then the d-dimensional Lebesgue measure has to be defined as the completion of the product of 1-dimensional measures.

The proofs of the claims in the second bullet point above are not hard but require a bit of work and are more suited to a measure theory course, so we will skip them. The first bullet point however is easy to show and the reader is encouraged to try and prove it. ◆

It is not quite as clear how to extend the definition of a product measure to infinitely many spaces, but in the case of probability spaces this turns out to be possible.

Theorem 3.9 (Product probability spaces)**.** *Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)_{i \in I}$ be a collection of probability spaces indexed by an arbitrary index set I. Then there exists a unique product probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega := \prod_{i \in I} \Omega_i$, $\mathcal{F} := \bigotimes_{i \in I} \mathcal{F}_i$ and $\mathbb{P}$ a*

probability measure that satisfies

$$\mathbb{P}(C) = \prod_{i \in I} \mathbb{P}_i(C_i)$$

for all $C \subset \Omega$ of the form $C = \prod_{i \in I} C_i$ with $C_i \in \mathcal{F}_i$ for all $i \in I$ and $C_i = \Omega_i$ for all but finitely many i's.

Proof. Exercise. Hint: Use Carathéodory's extension theorem. To show countable additivity on the semialgebra generated by the cylinder sets, try to use similar ideas as in Lemma 1.33: Start with the case where the family is countable. If $\biguplus_{n=1}^{\infty} A^n = \Omega$ is a partition of Ω into disjoint cylinder sets but $\sum_{n=1}^{\infty} \mathbb{P}[A_n] \neq 1$, then by Fubini's theorem there exists $\omega_1 \in \Omega_1$ such that we have

$$\sum_{n=1}^{\infty} \mathbb{1}_{A_1^n}(\omega_1) \prod_{k=2}^{\infty} \mathbb{P}_k[A_k^n] \neq 1.$$

By induction one can find $\omega_1, \dots, \omega_m$ such that

$$\sum_{n=1}^{\infty} \mathbb{1}_{A_1^n}(\omega_1) \dots \mathbb{1}_{A_m^n}(\omega_m) \prod_{k=m+1}^{\infty} \mathbb{P}[A_k^n] \neq 1.$$

Derive a contradiction by considering $\omega = (\omega_n)_{n=1}^{\infty} \in \Omega$ and the set A^m to which it belongs. Can you extend to the case of an uncountable product? □

3.2 *Independence and products*

In this section we will show how the distributions of independent random variables are product measures and prove the product formula for expectation of independent random variables.

Theorem 3.10. *Suppose that $X_1, \dots, X_n$ are independent random variables with distributions $\mu_1, \dots, \mu_n$. Then the law of the random vector $(X_1, \dots, X_n)$ is given by $\mu_1 \otimes \cdots \otimes \mu_n$.*

Proof. Let μ be the law of the random vector $(X_1, \dots, X_n)$. Then by definition for Borel sets $A_1, \dots, A_n$ we have

$$\begin{aligned}
\mu(A_1 \times \cdots \times A_n) &= \mathbb{P}[(X_1, \dots, X_n) \in A_1 \times \cdots \times A_n] \\
&= \mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n] \\
&= \mathbb{P}[X_1 \in A_1] \dots \mathbb{P}[X_n \in A_n] \\
&= (\mu_1 \otimes \cdots \otimes \mu_n)(A_1 \times \cdots \times A_n).
\end{aligned}$$

Since the rectangles $A_1 \times \cdots \times A_n$ form a π-system that generates the Borel σ-algebra on $\mathbb{R}^n$, we see that μ and the product measure are equal. □

As a corollary we have the following.

Proposition 3.11. *Assume that $X_1, \dots, X_n$ are random variables with densities $f_1, \dots, f_n$ w.r.t. the Lebesgue measure. Then $X_1, \dots, X_n$ are independent if and only if the random vector $(X_1, \dots, X_n)$ has the density $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$ w.r.t. the Lebesgue measure on $\mathbb{R}^n$.*

Proof. Exercise. □

The main result concerning the expectation of product of two independent random variables is now easy to derive.

Theorem 3.12. *Assume that X and Y are independent random variables and that either $X, Y \geq 0$ or $X, Y \in L^1$. Then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Proof. If X and Y are non-negative, we have by the change-of-variables formula and Fubini's theorem that

$$\mathbb{E}[XY] = \int\int xy d(Y_*\mathbb{P})(y) d(X_*\mathbb{P})(x) = \int x d(X_*\mathbb{P}) \int y d(Y_*\mathbb{P}) = \mathbb{E}[X]\mathbb{E}[Y].$$

For integrable X and Y we may first apply the theorem to $|X|$ and $|Y|$ (note that trivially $\sigma(|X|) \subset \sigma(X)$, so $\sigma(|X|)$ and $\sigma(|Y|)$ are independent), and get that $\mathbb{E}[|XY|] = \mathbb{E}[|X|]\mathbb{E}[|Y|] < \infty$. Hence $XY \in L^1$, and we may again use the change-of-variables formula and Fubini's theorem. □

It is important to note that having $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ does not imply independence of X and Y. Random variables that satisfy the condition are called **uncorrelated**.

Exercise 3.13. Give an example of two random variables X and Y that are uncorrelated but not independent. ◆

The above theorem admits various generalizations. First of all a similar result holds also for n independent random variables $X_1, \dots, X_n$ that are either all non-negative or all integrable, in which case $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$. The proof is essentially the same as the case of two random variables.

A special case of the above is where $X_k = F_k(Y_{k,1}, \dots, Y_{k,m_k})$ for some independent random vectors $Y_k : \Omega \to \mathbb{R}^{m_k}$ and measurable functions $F_k : \mathbb{R}^{m_k} \to \mathbb{R}$. For checking that random vectors are independent the following "grouping lemma" is often useful.

Lemma 3.14. *Assume that $\mathcal{F}_{k,j}$, $1 \leq k \leq n$, $1 \leq j \leq m_k$, are independent σ-algebras. Then the σ-algebras $\mathcal{F}_k := \sigma(\bigcup_{j=1}^{m_k} \mathcal{F}_{k,j})$ are independent.*

As a corollary we have the following.

Corollary 3.15. *Suppose that $X_{k,j}$ ($1 \le k \le n$, $1 \le j \le m_k$, $m_k \ge 1$) are independent random variables. Then the random vectors $(X_{k,1}, \dots, X_{k,m_k})$, $1 \le k \le n$, are independent.*

Proof. Let $\mathcal{T}_{k,j} \coloneqq \sigma(X_{k,j})$. Then the random vector $Y_k \coloneqq (X_{k,1}, \dots, X_{k,m_k})$ is $\sigma(\bigcup_{j=1}^{m_k} \mathcal{T}_{k,j})$-measurable since $Y_k^{-1}(A_1 \times \cdots \times A_{m_k}) = \bigcap_{j=1}^{m_k} X_{k,j}^{-1}(A_j)$ is measurable for all $A_1 \times \cdots \times A_{m_k} \in \mathcal{B}^{\otimes m_k}$ and such rectangles generate the whole product σ-algebra. The claim thus follows from the grouping lemma. □

The proof of the grouping lemma will be based on the following useful result which says that if π-systems are independent, then the σ-algebras generated by them are also independent.

Lemma 3.16. *Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be independent π-systems, meaning that for all $A_k \in \mathcal{A}_k \cup \{\Omega\}$, $1 \le k \le n$, we have*

$$\mathbb{P}[A_1 \cap \cdots \cap A_n] = \prod_{k=1}^{n} \mathbb{P}[A_k].$$

Then $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent.

Proof. Consider the measurable space $(\Omega^n, \mathcal{G})$, where $\mathcal{G} \coloneqq \sigma(\mathcal{A}_1) \otimes \cdots \otimes \sigma(\mathcal{A}_n)$. On $\mathcal{T}^{\otimes n}$ there is of course the product measure $\mathbb{P}^{\otimes n}$ which satisfies

$$\mathbb{P}^{\otimes n}[A_1 \times \cdots \times A_n] = \prod_{k=1}^{n} \mathbb{P}[A_k]$$

for all $A_k \in \sigma(\mathcal{A}_k)$, $1 \le k \le n$. On the other hand we may define a map ν on the semialgebra formed by the product sets $A_1 \times \cdots \times A_n$ by setting

$$\nu(A_1 \times \cdots \times A_n) \coloneqq \mathbb{P}[A_1 \cap \cdots \cap A_n].$$

Note that ν is countably additive since if $A_1 \times \cdots \times A_n = \biguplus_{k=1}^{\infty} B_{k,1} \times \cdots \times B_{k,n}$, then $A_1 \cap \cdots \cap A_n = \biguplus_{k=1}^{\infty} (B_{k,1} \cap \cdots \cap B_{k,n})$. Hence ν extends to a measure on $\mathcal{G}$, but since ν agrees with the product measure on the π-system consisting of sets of the form $A_1 \times \cdots \times A_n$ with $A_k \in \mathcal{A}_k \cup \{\Omega\}$, $1 \le k \le n$, which generates the σ-algebra $\mathcal{G}$, we see that the extension of ν equals the product measure and in particular

$$\mathbb{P}[A_1 \cap \cdots \cap A_n] = \nu(A_1 \times \cdots \times A_n) = \mathbb{P}^{\otimes n}[A_1 \times \cdots \times A_n] = \prod_{k=1}^{n} \mathbb{P}[A_k]$$

for all $A_k \in \sigma(\mathcal{A}_k)$, $1 \le k \le n$, which proves that the σ-algebras $\sigma(A_k)$ are independent. □

Let us next prove the grouping lemma.

Proof of Lemma 3.14. Let let P_k be the π-system formed by all the sets of the form $A_{k,1} \cap \cdots \cap A_{k,m_k}$ with $A_{k,j} \in \mathcal{F}_{k,j}$. Then the π-systems P_k are independent and hence also the σ-algebras $\sigma(P_k) = \sigma(\bigcup_{j=1}^{m_k} \mathcal{F}_{k,j})$ are independent by Lemma 3.16. □

Lemma 3.16 also has the following useful corollary.

Corollary 3.17. *Random variables $X_1, \dots, X_n$ are independent if and only if*

$$\mathbb{P}[X_1 \le t_1, X_2 \le t_2, \dots, X_n \le t_n] = \mathbb{P}[X_1 \le t_1] \cdots \mathbb{P}[X_n \le t_n]$$

for all $t_1, \dots, t_n \in \mathbb{R}$.

Proof. Exercise. □

3.3 *Conditional expectation*

Our last topic in this chapter will be conditional expectation. Conditional expectation $\mathbb{E}[X|\mathcal{G}]$ can be thought of as in some sense the best approximation of a random variable X given the information encoded by a σ-algebra $\mathcal{G}$.

Definition 3.18. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X\colon \Omega \to \mathbb{R}$ an integrable random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. The **conditional expectation** $\mathbb{E}[X|\mathcal{G}]$ of X with respect to $\mathcal{G}$ is the almost surely unique $\mathcal{G}$-measurable random variable which satisfies

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A] \tag{3.1}$$

for all $A \in \mathcal{G}$. ◆

Before showing that the definition makes sense, i.e. that conditional expectation exists and is unique, let us try to gain a bit of intuition. It is quite natural to require that $\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable if we want to take the best approximation of X given the information in $\mathcal{G}$, but it is perhaps less clear how to think about the defining condition (3.1). It might be useful to start with the following example.

Example 3.19. Let $E_1, \dots, E_n$ be a partition of Ω with $\mathbb{P}[E_k] > 0$ for all k, and let $\mathcal{G} \coloneqq \sigma(E_1, \dots, E_n)$. Now if X is a random variable, we have by (3.1) that

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{1}_{E_k}] = \mathbb{E}[X\mathbb{1}_{E_k}].$$

Notice that since $\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, it must be constant on each E_k – let us call these constants a_k. Thus from above we get the equation $a_k\mathbb{P}[E_k] = \mathbb{E}[X\mathbb{1}_{E_k}]$, or

$$a_k = \frac{\mathbb{E}[X\mathbb{1}_{E_k}]}{\mathbb{P}[E_k]}.$$

In other words, $\mathbb{E}[X|\mathcal{G}](\omega) = \frac{\mathbb{E}[X\mathbb{1}_{E_k}]}{\mathbb{P}[E_k]}$ for all $\omega \in E_k$, i.e. $\mathbb{E}[X|\mathcal{G}]$ is obtained by replacing X by its averages over each of the sets E_k. ◆

In the above example we see that taking the conditional expectation is basically just averaging out the extra randomness in order to form a guess based on the information we have. The central property of taking averages is that if you first average over some atomic sets E_k like in the example to obtain $\mathbb{E}[X|\mathcal{G}]$, then the average of $\mathbb{E}[X|\mathcal{G}]$ itself over unions such as $E_1 \cup E_2$ will be the same as the average of the original random variable X over the same set.

Now, in general $\mathcal{G}$ might not be given by such a simple partition as in the example, but we can still ask is there a random variable $\mathbb{E}[X|\mathcal{G}]$ which is $\mathcal{G}$-measurable and preserves the averages over all sets in $\mathcal{G}$. This is exactly the content of the condition (3.1).

Let us next try to show that the definition indeed makes sense.

Theorem 3.20. *For any $X \in L^1$ and any σ-algebra $\mathcal{G} \subset \mathcal{F}$ the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ exists and is unique a.s.*

Moreover, we have the following extension of (3.1)*: If Y is $\mathcal{G}$-measurable, then $XY \in L^1$ if and only if $\mathbb{E}[X|\mathcal{G}]Y \in L^1$, in which case*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]Y] = \mathbb{E}[XY].$$

Proof. We will use the Radon–Nikodym theorem. Assume first that X is non-negative and consider the measure $\nu(A) = \mathbb{E}[X\mathbb{1}_A]$ on $\mathcal{G}$. Clearly $\nu \ll \mathbb{P}$, and thus there exists a unique Radon–Nikodym derivative $\mathbb{E}[X|\mathcal{G}] \coloneqq \frac{d\nu}{d\mathbb{P}}$ such that

$$\mathbb{E}[X\mathbb{1}_A] = \nu(A) = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{1}_A],$$

which is exactly what we wanted. Moreover by Lemma 2.40, we have $Y \in L^1(\nu)$ if and only if $XY \in L^1$, in which case $\mathbb{E}[X|\mathcal{G}]Y \in L^1$ and

$$\mathbb{E}[XY] = \int Y \, d\nu = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]Y].$$

For general $X \in L^1$, we may write $X = X_+ - X_-$ as the difference of its positive and negative parts and define

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X_+|\mathcal{G}] - \mathbb{E}[X_-|\mathcal{G}].$$

By linearity we again have $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]Y] = \mathbb{E}[XY]$ for all $\mathcal{G}$-measurable Y with $XY \in L^1$. □

The first and most fundamental property of conditional expectation is that it is a linear and continuous operator $L^1 \to L^1$.

Proposition 3.21. *Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then the map $L^1 \to L^1$ given by $X \mapsto \mathbb{E}[X|\mathcal{G}]$ is linear and continuous.*

Proof. To see that the map is linear it is enough to show that

$$\mathbb{E}[cX + Y|\mathcal{G}] = c\mathbb{E}[X|\mathcal{G}] + \mathbb{E}[Y|\mathcal{G}]$$

where $c \in \mathbb{R}$ and $X, Y \in L^1$. This is true, since for any $A \in \mathcal{G}$ we have by the linearity of ordinary expectation that

$$\begin{aligned}\mathbb{E}[(c\mathbb{E}[X|\mathcal{G}] + \mathbb{E}[Y|\mathcal{G}])\mathbb{1}_A] &= c\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{1}_A] + \mathbb{E}[\mathbb{E}[Y|\mathcal{G}]\mathbb{1}_A] \\ &= c\mathbb{E}[X\mathbb{1}_A] + \mathbb{E}[Y\mathbb{1}_A] = \mathbb{E}[(cX + Y)\mathbb{1}_A].\end{aligned}$$

For continuity it is enough to show boundedness. This follows easily from the extended formula in Theorem 3.20, since

$$\begin{aligned}\|\mathbb{E}[X|\mathcal{G}]\|_{L^1} &= \mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\operatorname{sgn}(\mathbb{E}[X|\mathcal{G}])] \\ &= \mathbb{E}[X\operatorname{sgn}(\mathbb{E}[X|\mathcal{G}])] \leq \mathbb{E}[|X|] = \|X\|_{L^1}.\end{aligned}$$

□

Proposition 3.22. *Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let X and Y be two random variables such that XY and X are both integrable. Then if Y is $\mathcal{G}$-measurable, we have*

$$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}].$$

Proof. Since for any $A \in \mathcal{G}$ we have $XY\mathbb{1}_A \in L^1$, we have by Theorem 3.20 that

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]Y\mathbb{1}_A] = \mathbb{E}[XY\mathbb{1}_A],$$

which implies that $\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$. □

Computing conditional expectations can often be challenging in practice, but in the case where we have two random variables with a joint p.d.f. and condition one w.r.t. the other we have the following result. Note that by $\mathbb{E}[X|Y]$ we mean $\mathbb{E}[X|\sigma(Y)]$.

Proposition 3.23. *Let $X, Y \in L^1$ be random variables with the joint density $f(x, y)$. Then for any measurable $\varphi\colon \mathbb{R} \to \mathbb{R}$ such that $\varphi(X) \in L^1$ we have $\mathbb{E}[\varphi(X)|Y] = g(Y)$, where*

$$g(y) := \begin{cases} \frac{\int \varphi(x) f(x,y)\,dx}{\int f(x,y)\,dx}, & \text{if } \int f(x,y)\,dx \neq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Proof. The random variable $g(Y)$ is clearly $\sigma(Y)$-measurable. To check (3.1), we note that if $A \in \sigma(Y)$, then $A = Y^{-1}(B)$ for some Borel set $B \subset \mathbb{R}$ and by

the change-of-variables formula and Fubini's theorem we have

$$\begin{aligned}\mathbb{E}[g(Y)\mathbb{1}_A] &= \int_{\mathbb{R}^2} g(y)\mathbb{1}_B(y) f(x,y)\,dx\,dy \\ &= \int_{\mathbb{R}^2} \mathbb{1}_{\{\int f(u,y)\,du>0\}}(y)\mathbb{1}_B(y)\frac{\int_{\mathbb{R}}\varphi(u)f(u,y)\,du}{\int_{\mathbb{R}} f(u,y)\,du} f(x,y)\,dx\,dy \\ &= \int_{\mathbb{R}^2} \mathbb{1}_{\{\int f(u,y)\,du>0\}}(y)\mathbb{1}_B(y)\varphi(u) f(u,y)\,du\,dy \\ &= \int_{\mathbb{R}^2} \varphi(u)\mathbb{1}_B(y) f(u,y)\,du\,dy = \mathbb{E}[\varphi(X)\mathbb{1}_A].\end{aligned}$$

Note that one can justify the use of Fubini's theorem by doing first a similar computation as above but for $\mathbb{E}[|g(Y)|]$ to show that $g(Y) \in L^1$, we leave the details to the reader. □

The following proposition shows that if we condition twice with respect to two σ-algebras $\mathcal{H}$ and $\mathcal{G}$, then the result will always correspond to conditioning with respect to the σ-algebra which contains less information. This is sometimes called the **tower property** of conditional expectation.

Proposition 3.24. *Let $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ and assume that $X \in L^1$. Then*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}].$$

Proof. Exercise. □

Independence and conditioning also works as one would expect: If there is no information, the best guess is just the expectation.

Proposition 3.25. *Let $\mathcal{G} \subset \mathcal{F}$ and assume that $X \in L^1$ is independent of $\mathcal{G}$. Then* $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

Proof. Exercise. □

In particular the above proposition shows that $\mathbb{E}[X|\{\varnothing,\Omega\}] = \mathbb{E}[X]$, so the usual expectation can be viewed as a special case of conditional expectation where we condition w.r.t. the trivial σ-algebra.

Conditional probabilities can also be defined via conditional expectation.

Definition 3.26. The **conditional probability** $\mathbb{P}[A|\mathcal{G}]$ of an event A given a σ-algebra $\mathcal{G}$ is defined by setting

$$\mathbb{P}[A|\mathcal{G}] \coloneqq \mathbb{E}[\mathbb{1}_A|\mathcal{G}]. \qquad \blacklozenge$$

Here are some more properties of conditional expectation. The proofs are mostly trivial and left as an exercise.

Proposition 3.27. *The conditional expectation satisfies the following (X, Y are random variables, $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ are σ-algebras and we assume all the conditional expectations exist. All claims hold almost surely.):*

- *We have* $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.
- *If X is $\mathcal{G}$-measurable then* $\mathbb{E}[X|\mathcal{G}] = X$.
- *If X is independent of $\sigma(\sigma(Y) \cup \mathcal{G})$, then* $\mathbb{E}[XY|\mathcal{G}] = \mathbb{E}[X]\mathbb{E}[Y|\mathcal{G}]$.
- *If $X \geq 0$ then* $\mathbb{E}[X|\mathcal{G}] \geq 0$.
- *If $X \geq Y$ then* $\mathbb{E}[X|\mathcal{G}] \geq \mathbb{E}[Y|\mathcal{G}]$.
- *If $X_n \to X$ in L^1, then* $\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}]$ *in* L^1.

Proof. Exercise. □

As conditional expectations are defined only up to almost sure equivalence, it is not meaningful talk about pointwise convergence of $\mathbb{E}[X_n|\mathcal{G}]$ to $\mathbb{E}[X|\mathcal{G}]$. Almost sure convergence is however still well defined and in particular monotone limits work nicely.

Proposition 3.28. *Let $X_n \in L^1$ be a.s. non-negative and increasing and suppose that the a.s. limit $X = \lim_{n\to\infty} X_n$ is also integrable. Then*

$$\mathbb{E}[X_n|\mathcal{G}] \overset{a.s.}{\to} \mathbb{E}[X|\mathcal{G}].$$

Proof. By monotonicity the sequence $\mathbb{E}[X_n|\mathcal{G}]$ is almost surely increasing and hence converges almost surely to some non-negative $Y \in L^0$. Then by the monotone convergence theorem we have for any $A \in \mathcal{G}$ that

$$\mathbb{E}[Y\mathbb{1}_A] = \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[X_n|\mathcal{G}]\mathbb{1}_A] = \lim_{n\to\infty} \mathbb{E}[X_n\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A],$$

showing that $Y = \mathbb{E}[X|\mathcal{G}]$ almost surely. □

Finally let us consider conditional distributions.

Definition 3.29. Let $\mathcal{B}$ denote the Borel σ-algebra on $\mathbb{R}$. If X is a random variable and $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra, we say that a map $\mu\colon \mathcal{B} \times \Omega \to [0,1]$ is a **regular conditional distribution** for X given $\mathcal{G}$, if the following hold:

- Almost surely for a fixed $\omega \in \Omega$ the map $A \mapsto \mu(A, \omega)$ is a probability measure on $\mathbb{R}$.
- For a fixed $A \in \mathcal{B}$ the map $\omega \mapsto \mu(A, \omega)$ is measurable and we have $\mu(A, \omega) = \mathbb{P}[X \in A|\mathcal{G}](\omega)$ a.s. ◆

Conditional distributions are nice because they let us compute $\mathbb{E}[F(X)|\mathcal{G}]$ for many different F simultaneously.

Proposition 3.30. *Let X be a random variable and $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra, and denote by μ the r.c.d. for X given $\mathcal{G}$. Then for any measurable $\varphi\colon \mathbb{R} \to \mathbb{R}$ such that $\varphi(X) \in L^1$ we almost surely have*

$$\mathbb{E}[\varphi(X)|\mathcal{G}](\omega) = \int \varphi(x)\, d\mu(x,\omega).$$

Proof. We first note that in the case $\varphi = \mathbb{1}_E$ for some $E \in \mathcal{B}$ we have

$$\int \varphi(x)\, d\mu(x,\omega) = \int \mathbb{1}_E\, d\mu(x,\omega) = \mu(E,\omega) = \mathbb{P}[X \in E|\mathcal{G}] = \mathbb{E}[\varphi(X)|\mathcal{G}]$$

by definition. By linearity we see that the claim holds for simple φ and by approximation and monotone convergence one gets the claim for non-negative φ and the final case $\varphi(X) \in L^1$ follows by considering φ_+ and φ_-. □

Let us next show that r.c.d.s exist.

Theorem 3.31. *Let X and $\mathcal{G}$ be as in Definition 3.29. Then there exists a regular conditional distribution μ for X given $\mathcal{G}$.*

Proof. For each $q \in \mathbb{Q}$ let $\omega \mapsto F_0(q,\omega)$ be a fixed pointwise defined representative of $\mathbb{P}[X \le q|\mathcal{G}]$. We next claim that there exists an event $\tilde{\Omega}$ of full probability such that the following hold for all $\omega \in \tilde{\Omega}$:

- $q \mapsto F_0(q,\omega)$ is increasing
- $\lim_{q\to\infty} F_0(q,\omega) = 1$ and $\lim_{q\to-\infty} F_0(q,\omega) = 0$
- $\lim_{q'\downarrow q} F(q',\omega) = F(q,\omega)$

The first bullet point follows since by the monotonicity of conditional expectation we have almost surely $F_0(q,\omega) \le F_0(q',\omega)$ whenever $q < q'$ and to ensure that F_0 is increasing it is enough to consider the intersection of all such events for a countable number of pairs (q,q'). The second point follows from the monotonicity we just showed and the fact that by Proposition 3.28 $\mathbb{P}[X \le n|\mathcal{G}] \to 1$ a.s. as $n \to \infty$ and $\mathbb{P}[X \le n|\mathcal{G}] \to 0$ a.s. as $n \to -\infty$. The third point is similar since a.s. $\lim_{n\to\infty} \mathbb{P}[X \le q + 2^{-n}|\mathcal{G}] = \mathbb{P}[X \le q|\mathcal{G}]$.

Let us next define the map $F\colon \mathbb{R}\times\Omega \to \mathbb{R}$ by setting

$$F(x,\omega) = \inf\{F_0(q,\omega) : \mathbb{Q} \ni q > x\}$$

when $\omega \in \tilde{\Omega}$ and just define it as e.g. $F(x,\omega) = \mathbb{1}_{[0,\infty)}(x)$ for all $\omega \in \Omega\setminus\tilde{\Omega}$. Now it is easy to check that for fixed ω the map $x \mapsto F(x,\omega)$ satisfies the assumptions

of Theorem 1.49, and hence it is the c.d.f. of some random variable, and in particular there exists a map $\mu\colon \mathcal{B}\times\Omega \to [0,1]$ such that for every fixed $\omega \in \Omega$ the map $A \mapsto \mu(A,\omega)$ is a probability measure which satisfies

$$\mu((-\infty, x], \omega) = F(x, \omega).$$

In order to μ be a r.c.d. it remains to check that for fixed $A \in \mathcal{B}$ the map $\omega \mapsto \mu(A,\omega)$ is a random variable which agrees with $\mathbb{P}[X \in A|\mathcal{G}]$ almost surely. Let $\mathcal{A}$ be the set of all A for which the claim holds. Then $\mathcal{A}$ is a λ-system since if $A \in \mathcal{A}$, we have $\mu(A^c,\omega) = 1-\mu(A,\omega) = 1-\mathbb{P}[X \in A|\mathcal{G}](\omega) = \mathbb{P}[X \in A^c|\mathcal{G}]$ a.s. and if $(A_n)_{n=1}^{\infty}$ is a disjoint sequence of elements of $\mathcal{A}$, we have

$$\mu(\biguplus_{n=1}^{\infty} A_n, \omega) = \sum_{n=1}^{\infty} \mu(A_n,\omega) = \sum_{n=1}^{\infty} \mathbb{P}[X \in A_n|\mathcal{G}](\omega) = \mathbb{P}[X \in \biguplus_{n=1}^{\infty} A_n](\omega)$$

almost surely. Moreover $\mathcal{A}$ contains the π-system consisting of intervals of the form $(-\infty, q]$ for some $q \in \mathbb{Q}$, and since these intervals generate $\mathcal{B}$, we have by the π-λ-theorem that $\mathcal{A} = \mathcal{B}$. □

Remark. One can also consider general T-valued random variables, but in that case order to have the existence of r.c.d.s one needs to impose some extra conditions on T. In particular the claim holds when T is a **standard Borel space**, which means that there exists a measurable bijection $\varphi\colon T \to \mathbb{R}$ such that also φ^{-1} is measurable. As one would expect, the r.c.d. in this case is then a map $\mu\colon T\times\Omega \to [0,1]$.

One can show that Borel subsets of complete separable metric spaces are standard Borel spaces when endowed with the σ-algebra generated by the Borel sets. In particular $\mathbb{R}^d$ are standard Borel spaces and thus natural analogues of Theorem 3.31 and Proposition 3.30 hold for vector-valued random variables X, so that for example the formula

$$\mathbb{E}[\varphi(X,Y)|\mathcal{G}](\omega) = \int \varphi(x,y)d\mu(x,y,\omega)$$

holds almost surely whenever $\varphi(X,Y) \in L^1$. We skip the proofs even though they are not difficult – interested readers can try to prove them by themselves or see e.g. [1, Theorem 4.1.17]. ◆

Using r.c.d.s it is easy to generalize many properties of the usual expectation.

Proposition 3.32. *The conditional expectation satisfies:*

- *Jensen's inequality:* $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}])$ *a.s. for convex* φ.

- *Hölder's inequality: For $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\mathbb{E}[|XY||\mathcal{G}] \leq (\mathbb{E}[|X|^p|\mathcal{G}])^{1/p}(\mathbb{E}[|Y|^q|\mathcal{G}])^{1/q}$$

- *Minkowski's inequality: For $p \geq 1$,*

$$(\mathbb{E}[|X+Y|^p|\mathcal{G}])^{1/p} \leq (\mathbb{E}[|X|^p|\mathcal{G}])^{1/p} + (\mathbb{E}[|Y|^p|\mathcal{G}])^{1/p}$$

Proof. Exercise. □

Let us end this section by giving a geometrical interpretation to conditional expectation. This is extra material and we won't use it later, but it is also a good excuse to talk a bit more about L^2 and in general it is good to know these things.

One can show that $X \mapsto \mathbb{E}[X|\mathcal{G}]$ is a continuous operator on L^2 (exercise). The space L^2 on the other hand is special because it is a *Hilbert space*, i.e. its norm is given by the inner product

$$\langle X, Y\rangle_{L^2} \coloneqq \mathbb{E}[XY].$$

Indeed, if $X, Y \in L^2$, then by Hölder's inequality one checks that $XY \in L^1$, so the above definition makes sense, and one can also easily check that $\langle\cdot,\cdot\rangle_{L^2}$ has all the properties of an inner product.

Given an inner product one can say that X and Y are orthogonal if $\langle X, Y\rangle_{L^2} = \mathbb{E}[XY] = 0$. Given a subspace $V \subset L^2$ one can define its orthocomplement $V^\perp \coloneqq \{X \in L^2 : \mathbb{E}[XY] = 0 \text{ for all } Y \in V\}$. Then it is a theorem that any $X \in L^2$ can be written in a unique way in the form $X = X_V + X_{V^\perp}$, where $X_V \in \overline{V}$ and $X_{V^\perp} \in V^\perp$. One can also show that the map $X \mapsto X_V$ is linear and this is called the *orthogonal projection* of X onto $\overline{V}$.

Now consider the closed subspace

$$V = L^2(\mathcal{G}) \coloneqq \{X \in L^2 : X \text{ is } \mathcal{G}\text{-measurable}\}.$$

Conditional expectation is nothing but the orthogonal projection onto the subspace $L^2(\mathcal{G})$.

Another way to express the orthogonal projection of X onto $L^2(\mathcal{G})$ is to say that it is the random variable $Y \in L^2(\mathcal{G})$ which minimizes the distance to X in L^2, or in other words minimizes the variance $\mathbb{E}[|X - Y|^2]$. We leave it as an exercise to try to show this directly without the above theory.

Random series and the law of large numbers

4.1 *Estimating the distribution of random variables*

A task that one often runs into both in theoretical settings as well as in applications is to estimate the probability that a given random variable X lies in a given set $A \subset \mathbb{R}$.

One common task is to get upper bounds for the tail probability $\mathbb{P}[X > \lambda]$ of X being large. A common method to do this is to notice the following: If $\varphi\colon \mathbb{R} \to [0,\infty)$ is an increasing function with $\varphi(\lambda) > 0$, then

$$\mathbb{P}[X \geq \lambda] = \mathbb{P}[\varphi(X) \geq \varphi(\lambda)] = \mathbb{P}\Big[\frac{\varphi(X)}{\varphi(\lambda)} \geq 1\Big] = \mathbb{E}[\mathbb{1}_{\{\frac{\varphi(X)}{\varphi(\lambda)} \geq 1\}}] \leq \frac{\mathbb{E}[\varphi(X)]}{\varphi(\lambda)}.$$

Inequalities resulting from various choices of φ have been given various names:

- Choosing $\varphi(x) = x\mathbb{1}_{[0,\infty)}(x)$ we get **Markov's inequality**

$$\mathbb{P}[|X| \geq \lambda] \leq \frac{\mathbb{E}[|X|]}{\lambda}.$$

- Choosing $\varphi(x) = x^2\mathbb{1}_{[0,\infty)}(x)$ we get **Chebyshev's inequality**

$$\mathbb{P}[|X| \geq \lambda] \leq \frac{\mathbb{E}[X^2]}{\lambda^2}.$$

- Choosing $\varphi(x) = \exp(tx)$ for some $t > 0$ we get the **Chernoff bound**

$$\mathbb{P}[X \geq \lambda] \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t\lambda)}.$$

Apart from tail probabilities another related problem is to show that the distribution is well-concentrated around its mean. Here we will mention the following **Paley–Zygmund inequality** which can be used to show that with a reasonable probability a non-negative random variable does not become too small compared to its mean.

Lemma 4.1. *Let $X \in L^2$ be a non-negative random variable. Then for any $\theta \in$*

$[0,1]$ *we have*

$$\mathbb{P}[X \geq \theta \mathbb{E}[X]] \geq (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Proof. We have by the Cauchy–Schwarz inequality (Hölder's inequality with $p = q = 2$) that

$$\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}_{\{X < \theta \mathbb{E}[X]\}}] + \mathbb{E}[X \mathbb{1}_{\{X \geq \theta \mathbb{E}[X]\}}] \leq \theta \mathbb{E}[X] + \sqrt{\mathbb{E}[X^2]} \sqrt{\mathbb{P}[X \geq \theta \mathbb{E}[X]]}$$

from which the claim follows by subtracting $\theta \mathbb{E}[X]$ on both sides, dividing by $\sqrt{\mathbb{E}[X^2]}$ and squaring. □

4.2 *Strong law of large numbers*

This whole section will be devoted to the proof of the following **strong law of large numbers**.

Theorem 4.2. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed (i.i.d.) random variables in L^1. Then*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mathbb{E}[X_1]$$

almost surely and in L^1.

The proof will be incremental, going from weaker results towards the final one. There will be several clever tricks along the way, but it will also allow us to return to all the theory we have built up so far. Let us fix the notation $S_n := \sum_{k=1}^{n} X_k$. Since we will be interested in the difference between $n^{-1} S_n$ and $\mathbb{E}[X_1]$, it is also useful to define

$$A_n := n^{-1} S_n - \mathbb{E}[X_1] = \frac{1}{n} \sum_{n=1}^{\infty} (X_k - \mathbb{E}[X_k]),$$

where the sum on the right is now over independent random variables with zero expectation. Our goal is equivalent to showing that $A_n \to 0$ a.s. and in L^1.

Case of L^2-random variables: Note that if $X_k \in L^2$, then for any $\varepsilon > 0$ and $n \geq 1$ we have

$$\mathbb{E}[|A_n|^2] = n^{-2} \mathbb{E}\Big[\Big| \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \Big|^2 \Big] = n^{-1} \mathbb{E}[|X_1 - \mathbb{E}[X_1]|^2,$$

and as the right hand side tends to 0 we get that $n^{-1} \sum_{k=1}^{n} X_k \to \mathbb{E}[X_1]$ in L^2.

L^1-convergence: Our next step will be to remove the requirement that $X_k \in L^2$ by a truncation argument. Let us fix $\lambda > 0$ and split each X_k as the sum $X_k = X_k \mathbb{1}_{\{|X_k| \leq \lambda\}} + X_k \mathbb{1}_{\{|X_k| > \lambda\}}$. We will also write

$$A_n^{(\leq \lambda)} := \frac{1}{n} \sum_{n=1}^{\infty} (X_k \mathbb{1}_{|X_k| \leq \lambda} - \mathbb{E}[X_k \mathbb{1}_{|X_k| \leq \lambda}])$$

and

$$A_n^{(> \lambda)} := \frac{1}{n} \sum_{n=1}^{\infty} (X_k \mathbb{1}_{|X_k| > \lambda} - \mathbb{E}[X_k \mathbb{1}_{|X_k| > \lambda}])$$

so that $A_n = A_n^{(\leq \lambda)} + A_n^{(> \lambda)}$. We have by the triangle inequality

$$\mathbb{E}[|A_n|] \leq \mathbb{E}[|A_n^{(\leq \lambda)}|] + \mathbb{E}[|A_n^{(> \lambda)}|] \leq \left\| A_n^{(\leq \lambda)} \right\|_{L^2} + 2\mathbb{E}[|X_1| \mathbb{1}_{\{|X_1| > \lambda\}}].$$

Since $X_k \mathbb{1}_{\{|X_k| \leq \lambda\}} \in L^2$, the L^2-result we proved implies that the first term tends to 0 as $n \to \infty$. On the other hand the second term tends to 0 as $\lambda \to \infty$. As λ was arbitrary, we see that $A_n \to 0$ in L^1 as $n \to \infty$. We have now proven the L^1-part of Theorem 4.2, and this also implies convergence in probability – a result which is called the *weak law of large numbers*.

Almost sure convergence along geometric subsequences: Recall that from the convergence in probability which we have now proven it follows that there exists some subsequence $(n_k)_{k=1}^{\infty}$ such that $n_k^{-1} S_{n_k} \to \mathbb{E}[X_1]$ almost surely. We would now like to strengthen this claim to say that this in fact holds for all subsequences of the form $n_k = \lfloor r^k \rfloor$ with $r > 1$. By Borel–Cantelli lemma it is enough to show that for all $\varepsilon > 0$ we have

$$\sum_{k=1}^{\infty} \mathbb{P}[|A_{n_k}| > \varepsilon] < \infty.$$

Indeed, if this holds, then if we let E_j be the event that there exists a random k_j such that $|A_{n_k}| \leq j^{-1}$ for all $k \geq k_j$, we have $\mathbb{P}[E_j] = 1$, and the almost sure convergence along the subsequence n_k will follow by considering the full probability event $\bigcap_{j=1}^{\infty} E_j$.

Since $|A_{n_k}| > \varepsilon$ can only hold if at least one of $|A_{n_k}^{(\leq n_k)}| > \varepsilon/2$ or $|A_{n_k}^{(> n_k)}| > \varepsilon/2$ holds, we have

$$\mathbb{P}[|A_{n_k}| > \varepsilon] \leq \mathbb{P}[|A_{n_k}^{(\leq n_k)}| > \varepsilon/2] + \mathbb{P}[|A_{n_k}^{(> n_k)}| > \varepsilon/2].$$

For the first term we notice that

$$\mathbb{P}[|A_{n_k}^{(\leq n_k)}| > \varepsilon/2] \leq \frac{4}{\varepsilon^2} \mathbb{E}[|A_{n_k}^{(\leq n_k)}|^2] \leq \frac{4\mathbb{E}[|X_1|^2 \mathbb{1}_{\{|X_1| \leq n_k\}}]}{n_k \varepsilon^2}.$$

We then have

$$\sum_{k=1}^{\infty} n_k^{-1}\mathbb{E}[|X_1|^2\mathbb{1}_{\{|X_1|\leq n_k\}}] = \mathbb{E}[|X_1|^2\sum_{k=1}^{\infty}\frac{1}{\lfloor r^k\rfloor}\mathbb{1}_{\{\lfloor r^k\rfloor>|X_1|\}}]$$
$$\leq \mathbb{E}[|X_1|^2\sum_{k=\lfloor\log(|X_1|)/\log(r)\rfloor}^{\infty} r^{-k}] \lesssim \mathbb{E}[|X_1|].$$

For the second term we note that for large enough k we have $|\mathbb{E}[X_1\mathbb{1}_{\{|X_1|>n_k\}}]| < \varepsilon/2$, so in order to have

$$|A_{n_k}^{(>n_k)}| = \Big|\frac{1}{n_k}\sum_{j=1}^{n_k} X_j\mathbb{1}_{\{|X_j|>n_k\}} - \mathbb{E}[X_1\mathbb{1}_{\{|X_1|>n_k\}}]\Big| > \varepsilon/2$$

we actually need to have $|X_j| > n_k$ for at least one $j \in \{1,\dots,n_k\}$. Thus

$$\mathbb{P}[|A_{n_k}^{(>n_k)}| > \varepsilon/2] \leq n_k\mathbb{P}[|X_1| > n_k],$$

but then

$$\sum_{k=1}^{\infty} n_k\mathbb{P}[|X_1| > n_k] = \mathbb{E}[\sum_{k=1}^{\infty} n_k\mathbb{1}_{|X_1|>n_k}] \lesssim \mathbb{E}[\sum_{k=1}^{\log(|X_1|)/\log(r)} r^k] \lesssim \mathbb{E}[|X_1|].$$

Almost sure convergence along the original sequence: The final trick will be to notice that by linearity it is enough to prove the claim in the case where $X \geq 0$. Under this assumption we can then use the fact that

$$\frac{1}{m}\sum_{j=1}^{m} X_j - \frac{1}{n_k}\sum_{j=1}^{n_k} X_j = (\frac{n_k}{m}-1)\frac{1}{n_k}\sum_{j=1}^{n_k} X_j - \frac{1}{m}\sum_{j=m+1}^{n_k} X_j \leq (\frac{n_k}{m}-1)\frac{1}{n_k}\sum_{j=1}^{n_k} X_j$$

for $n_k \geq m$ and

$$\frac{1}{m}\sum_{j=1}^{m} X_j - \frac{1}{n_k}\sum_{j=1}^{n_k} X_j = (\frac{n_k}{m}-1)\frac{1}{n_k}\sum_{j=1}^{n_k} X_j + \frac{1}{m}\sum_{j=n_k+1}^{m} X_j \geq (\frac{n_k}{m}-1)\frac{1}{n_k}\sum_{j=1}^{n_k} X_j$$

for $n_k \leq m$. In the first case choosing $n_k = r^k \geq m$ as small as possible we see that $\frac{n_k}{m} - 1 \leq r - 1$ and thus

$$\frac{1}{m}\sum_{j=1}^{m} X_j \leq \frac{1}{n_k}\sum_{j=1}^{n_k} X_j + O(r-1),$$

as $r > 1$ is arbitrary, we get that $\limsup_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} X_j \leq \mathbb{E}[X_1]$ almost surely. Similarly using the other inequality one obtains that $\liminf_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m} X_j \geq$

$\mathbb{E}[X_1]$, which finishes the proof.

4.3 *Kolmogorov's zero–one law*

Kolmogorov's 0–1 law is a precise statement which roughly says the following: *If whether an event happens can be deduced from vanishing information, then this event either happens almost surely or almost never.*

The idea of "vanishing information" in this case is captured by the concept of tail σ-algebra.

Definition 4.3. Let $(\mathcal{F}_n)_{n=1}^{\infty}$ be a sequence of independent σ-algebras. Then the **tail σ-algebra** $\mathcal{G}$ generated by $(\mathcal{F}_n)_{n=1}^{\infty}$ is the σ-algebra defined by

$$\mathcal{G} \coloneqq \bigcap_{n=1}^{\infty} \sigma\Big(\bigcup_{k=n}^{\infty} \mathcal{F}_k \Big). \qquad \blacklozenge$$

A common situation is the one where we have a sequence $(X_n)_{n=1}^{\infty}$ of independent random variables and $\mathcal{F}_n = \sigma(X_n)$. In this case an event E belongs to the tail σ-algebra $\mathcal{G}$ if it does not depend on the first $X_1, \dots, X_n$ for any $n \geq 1$. A typical example would be the event $E = \{\lim_{n\to\infty} X_n \text{ exists}\}$. The following Kolmogorov's 0–1 law then says that for any $E \in \mathcal{G}$ we must have $\mathbb{P}[E] \in \{0, 1\}$.

Theorem 4.4. *Let $\mathcal{G}$ be a tail σ-algebra according to Definition 4.3 and let $E \in \mathcal{G}$. Then $\mathbb{P}[E] \in \{0, 1\}$.*

Proof. Let $A \in \mathcal{G}$. We will be done if we can show that A is independent of itself, because in that case $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$, which is only possible if $\mathbb{P}[A]$ is either 0 or 1.

Notice that by definition we have $A \in \sigma\Big(\bigcup_{k=n}^{\infty} \mathcal{F}_k \Big)$ for all $n \geq 1$. Consider the σ-algebra $\sigma\Big(\bigcup_{k=1}^{n-1} \mathcal{F}_k \Big)$. It is generated by the π-system consisting of all sets of the form $A_1 \cap \cdots \cap A_{n-1}$, where $A_k \in \mathcal{F}_k$. Similarly $\sigma\Big(\bigcup_{k=n}^{\infty} \mathcal{F}_k \Big)$ is generated by the π-system consisting of all sets of the form $\bigcap_{k=n}^{\infty} A_k$ with $A_k \in \mathcal{F}_k$ and $A_k \neq \Omega$ for only finitely many k. But now it is clear that these two π-systems are independent, so the same holds for the two σ-algebras. Thus in particular A is independent of $\sigma\Big(\bigcup_{k=1}^{n-1} \mathcal{F}_k \Big)$ for all n.

But now one can use a similar argument to show that $\sigma(A)$ is actually independent of $\sigma\Big(\bigcup_{k=1}^{\infty} \mathcal{F}_k \Big)$. Indeed, the latter is generated by a π-system consisting of all sets of the form $\bigcap_{k=1}^{\infty} A_k$ with $A_k \in \mathcal{F}_k$ and $A_k \neq \Omega$ for only finitely many k. Since A is independent of this π-system, it is also independent of the generated σ-algebra.

Finally note that $\mathcal{G} \subset \sigma\Big(\bigcup_{k=1}^{\infty} \mathcal{F}_k\Big)$, so A is both measurable and independent w.r.t. the σ-algebra $\sigma\Big(\bigcup_{k=1}^{\infty} \mathcal{F}_k\Big)$. □

There are many important tail events, and the following exercise presents some of them.

Exercise 4.5. Let $(X_n)_{n=1}^{\infty}$ be independent random variables. Show that the following are tail events and thus have either probability 0 or 1:

(a) $\{\lim_{n\to\infty} X_n \text{ exists}\}$

(b) $\{\lim_{n\to\infty} \sum_{k=1}^{n} X_k \text{ exists}\}$

(c) $\{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k \text{ exists}\}$ ◆

Kolmogorov's 0–1 law is indeed quite strong when applicable, but luckily it does not tell us which of the two possibilities happens for a given tail event, so we still have some interesting math to do. For example, we saw in the proof of the law of large numbers that it still requires quite a bit of work to show that if X_k are identically distributed then the probability of the event in part (c) of the above exercise is indeed 1 and not 0.

4.4 *Kolmogorov's three series theorem*

In this section we will prove Kolmogorov's three series theorem, which provides a sharp answer to the following question: Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables. When does $\sum_{n=1}^{\infty} X_n$ converge almost surely?

Theorem 4.6. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables, $K > 0$ and define $Y_n = X_n \mathbb{1}_{\{|X_n| \le K\}}$ for all $n \ge 1$. Then $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following three deterministic series converge*

$$\sum_{n=1}^{\infty} \mathbb{P}[|X_n| > K], \quad \sum_{n=1}^{\infty} \mathbb{E}[Y_n] \quad \textit{and} \quad \sum_{n=1}^{\infty} \mathbb{E}[|Y_n - \mathbb{E}[Y_n]|^2].$$

The proof of Theorem 4.6 will be based on the following lemma, which shows that for random series consisting of independent terms convergence in probability is equivalent with convergence almost surely.

Lemma 4.7. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables and assume that the random variables $S_n := \sum_{k=1}^{n} X_k$ converge in probability. Then the series $\sum_{n=1}^{\infty} X_n$ converges almost surely.*

Proof. Exercise. □

We will also make use of the following result.

Lemma 4.8. *Let $n \geq 1$ and let $(X_k)_{k=1}^n$ be independent random variables such that for all $1 \leq k \leq n$ we have $\mathbb{E}[X_k] = 0$ and $|X_k| \leq 1$ almost surely. Assume further that*

$$\sum_{k=1}^{n} \mathbb{E}[|X_k|^2] \geq 1.$$

Then there exists a universal constant $\varepsilon > 0$ (not depending on n or the particular random variables $(X_k)_{k=1}^n$) such that

$$\mathbb{P}\Big[\Big|\sum_{k=1}^{n} X_k\Big| \geq \varepsilon\Big] \geq \varepsilon.$$

Proof. Exercise. □

Proof of Theorem 4.6. Let us first show that if the three series converge, then $\sum_{n=1}^{\infty} X_n$ converges almost surely. The first condition $\sum_{n=1}^{\infty} \mathbb{P}[|X_n| > K] < \infty$ implies together with the Borel–Cantelli lemma that almost surely $X_n = Y_n$ for n large enough and hence it is sufficient to show that $\sum_{n=1}^{\infty} Y_n$ converges almost surely. Let us write $S_n = \sum_{k=1}^{n} Y_k$. For $n \geq m$ we have

$$\begin{aligned}
\mathbb{E}[|S_n - S_m|^2] &= \mathbb{E}\Big[\Big|\sum_{k=m+1}^{n} Y_k\Big|^2\Big] = \sum_{j,k=m+1}^{n} \mathbb{E}[Y_j Y_k] \\
&= \sum_{j,k=m+1}^{n} \mathbb{E}[Y_j]\mathbb{E}[Y_k] + \sum_{k=m+1}^{n} (\mathbb{E}[Y_k^2] - \mathbb{E}[Y_k]^2) \\
&= \Big|\sum_{k=m+1}^{n} \mathbb{E}[Y_k]\Big|^2 + \sum_{k=m+1}^{n} \mathbb{E}[|Y_k - \mathbb{E}[Y_k]|^2].
\end{aligned}$$

Since by assumption the series $\sum_{k=1}^{\infty} \mathbb{E}[Y_k]$ and $\sum_{k=1}^{\infty} \mathbb{E}[|Y_k - \mathbb{E}[Y_k]|^2]$ converge, we see that the right hand side tends to 0 as $n, m \to \infty$, which shows that S_n is Cauchy in L^2. Thus S_n converges in probability and by Lemma 4.7 it converges almost surely.

Let us then switch to proving the other direction and assume that $\sum_{n=1}^{\infty} X_n$ converges almost surely and try to show that the three deterministic series converge. The second Borel–Cantelli lemma implies that if we had $\sum_{n=1}^{\infty} \mathbb{P}[|X_n| > K] = \infty$, then a.s. we have $|X_n| > K$ for infinitely many n, but this is not possible since $\sum_{n=1}^{\infty} X_n$ converges almost surely. Thus the first series $\sum_{n=1}^{\infty} \mathbb{P}[|X_n| > K]$ is finite.

Let us next note that if we can show that the third series $\sum_{n=1}^{\infty} \mathbb{E}[|Y_n - \mathbb{E}[Y_n]|^2]$ converges almost surely, then we see that the sequence $U_n = Y_n - \mathbb{E}[Y_n]$ satisfies the assumptions in the first part of the proof (with K replaced by $2K$) and hence $\sum_{n=1}^{\infty} U_n$ converges almost surely. As $\sum_{n=1}^{\infty} Y_n$ converges almost surely, this implies that also $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$ converges. Thus it remains to show the convergence

of the third series.

We can do one more reduction by noting that it is enough to prove the convergence under the extra assumption that $\mathbb{E}[Y_n] = 0$. Indeed, if we let $Z_n = Y_n - Y_n'$, where Y_n' is an independent copy of Y_n for all n, then $\mathbb{E}[Z_n] = 0$ for all $n \geq 1$ and $\sum_{n=1}^{\infty} Z_n$ converges almost surely. Moreover $\sum_{n=1}^{\infty} \mathbb{E}[|Z_n|^2] = 2\sum_{n=1}^{\infty} \mathbb{E}[|Y_n - \mathbb{E}[Y_n]|^2]$, so if we can show the claim for Z_n it will also follow for Y_n.

Assume thus that $\mathbb{E}[Y_n] = 0$ and let again $S_n = \sum_{k=1}^{n} Y_k$, $S_0 = 0$. We have $|Y_n| \leq K$ almost surely and by scaling we may without loss of generality assume that $K = 1$. Suppose, to obtain a contradiction, that $\sum_{n=1}^{\infty} \mathbb{E}[|Y_n|^2] = \infty$. By using induction and Lemma 4.8 we see that there exists $\varepsilon > 0$ and a deterministic sequence $n_1 \leq n_2 \leq \dots$ such that

$$\mathbb{P}[|S_{n_{k+1}} - S_{n_k}| > \varepsilon] \geq \varepsilon$$

for all k. By the second Borel–Cantelli lemma one then has that $|S_{n_{k+1}} - S_{n_k}| > \varepsilon$ happens infinitely often, which contradicts the convergence of S_n. □

Convergence in law and the central limit theorem

5.1 *Convergence in law*

This final chapter is concerned on the convergence of random variables in law. This is a very different and much less probabilistic type of convergence than what we have discussed earlier, since it only looks at what happens with the laws $\mu_n := X_{n*}\mathbb{P}$ of the random variables. Thus one can make sense of this type of convergence even if X_n are defined on different probability spaces.

Definition 5.1. Let $(\mu_n)_{n=1}^{\infty}$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra. We say that μ_n **converge weakly** to a measure μ, if for every bounded continuous function $h\colon \mathbb{R} \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \int h(x)\, d\mu_n(x) = \int h(x)\, d\mu(x).$$

If $(X_n)_{n=1}^{\infty}$ is a sequence of random variables, X is another random variable, and $X_{n*}\mathbb{P}$ converge weakly to $X_*\mathbb{P}$, then we say that X_n **converge in law (or distribution) to** X and write $X_n \xrightarrow{d} X$. By the change-of-variables formula this is equivalent to requiring that

$$\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$$

for every bounded continuous $h\colon \mathbb{R} \to \mathbb{R}$. ◆

Remark. We may define weak convergence and convergence in law in an analogous way for measures on $\mathbb{R}^d$ just by requiring that the convergence holds against continuous and bounded functions $h\colon \mathbb{R}^d \to \mathbb{R}$. ◆

Let us immediately note the following.

Proposition 5.2. *If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{d} X$.*

Proof. Let $h\colon \mathbb{R} \to \mathbb{R}$ be bounded and continuous. By Proposition 2.16 we have $h(X_n) \xrightarrow{\mathbb{P}} h(X)$, and since $h(X_n)$ are bounded we get by the dominated convergence theorem $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$. □

The above definition is the "elegant" one, since it also works analogously for random variables taking values in any metric space. In practice one often

however likes to study distributions via their c.d.f.s, so let us next prove the following characterization of convergence in law.

Theorem 5.3. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables and X another random variable and let F_{X_n} and F_X be their respective c.d.f.s. Then $X_n \xrightarrow{d} X$ if and only if $F_{X_n}(x) \to F(x)$ for all $x \in \mathbb{R}$ such that F is continuous at x.*

Before proving the above theorem, let us note that F being continuous at x is essential. Consider the deterministic random variables $X_n = 1/n$ and $X = 0$. Then the law of X_n is a Dirac delta measure at $1/n$, while the law of X is a Dirac delta measure at 0. Clearly for any continuous and bounded f we have $f(1/n) \to f(0)$, so $X_n \xrightarrow{d} X$. However, $F_{X_n}(0) = 0$ for all n while $F_X(0) = 1$ so $F_{X_n}(0) \not\to F_X(0)$. The failure does not however contradict the theorem because $F_X(x) = \mathbb{1}_{[0,\infty)}(x)$ is not continuous at 0.

Proof of Theorem 5.3. Assume first that $X_n \xrightarrow{d} X$ and that x is a point of continuity of F. Fix $\varepsilon > 0$ and consider the continuous piecewise linear function $h(t)$ which is 1 for $t \leq x$, 0 for $t > x + \varepsilon$ and decreases linearly from 1 to 0 between x and $x + \varepsilon$. Since $X_n \xrightarrow{d} X$, we have $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$. Now $F_{X_n}(x) \leq \mathbb{E}[h(X_n)]$ and $\mathbb{E}[h(X)] \leq F(x + \varepsilon)$, and hence

$$\limsup_{n\to\infty} F_{X_n}(x) \leq F(x + \varepsilon),$$

and by letting $\varepsilon \to 0$ and using the continuity of F at x we see that

$$\limsup_{n\to\infty} F_{X_n}(x) \leq F(x).$$

On the other hand if we let $g(t)$ be the continuous piecewise linear function which is 1 for $t \leq x - \varepsilon$, 0 for $t > x$ and decreases linearly from 1 to 0 between $x - \varepsilon$ and x, then $F_{X_n}(x) \geq \mathbb{E}[g(X_n)]$ and $\mathbb{E}[g(X)] \geq F(x - \varepsilon)$, and hence

$$\liminf_{n\to\infty} F_{X_n}(x) \geq F(x - \varepsilon),$$

and by letting $\varepsilon \to 0$ and using the continuity of F at x we see that

$$\liminf_{n\to\infty} F_{X_n}(x) \geq F(x).$$

Thus $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ as wanted.

The opposite direction follows from the next representation theorem and Proposition 5.2. □

Theorem 5.4 (Skorokhod's representation theorem)**.** *Let $(F_n)_{n=1}^{\infty}$ and F be c.d.f.s*

such that

$$F_n(x) \to F(x)$$

at every continuity point x of F. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $(X_n)_{n=1}^{\infty}$ and X on Ω such that $F_{X_n} = F_n$ for all $n \geq 1$, $F_X = F$ and $X_n \overset{a.s.}{\to} X$.

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which a uniform random variable U taking values on $(0, 1)$ has been defined. Recall the proof of Theorem 1.49, where we defined a random variable X with c.d.f. F by letting $X = G(U)$, where

$$G(t) \coloneqq \inf\{x \in \mathbb{R} : F(x) \geq t\}.$$

Define X_n similarly by letting $X_n = G_n(U)$ (with always the same U), where

$$G_n(t) \coloneqq \inf\{x \in \mathbb{R} : F_n(x) \geq t\}.$$

We claim that $X_n \overset{a.s.}{\to} X$. Let D_F and D_G be the sets of disconcinuity points of F and G respectively. Since F and G are increasing, both D_F and D_G are countable (exercise). Thus U is a continuity point of G almost surely. Since the complement of D_F is dense, for any such $U \notin D_G$ and for any $\varepsilon > 0$ there exists numbers $x_1, x_2 \notin D_F$ such that

$$x_1 < G(U) < x_2 \quad \text{and} \quad |x_1 - x_2| < \varepsilon.$$

This implies that

$$F(x_1) < U < F(x_2),$$

since $U = F(x_2)$ is ruled out by the fact that by $G(U) < x_2$ there exists $x' < x_2$ s.t. $F(x') \geq U$, which means that $\lim_{u \to U^-} G(u) \leq x' < x_2 \leq \lim_{u \to U^+} G(u)$, contradicting the fact that by the continuity of G at U we have

$$\lim_{u \to U-} G(u) = \lim_{u \to U+} G(u).$$

On the other hand since we have $\lim_{n\to\infty} F_n(x_1) = F(x_1)$ and $\lim_{n\to\infty} F_n(x_2) = F(x_2)$, we see that for large enough n

$$F_n(x_1) < U < F_n(x_2),$$

which implies that

$$x_1 \leq G_n(U) \leq x_2,$$

so that $|X_n - X| = |G_n(U) - G(U)| \leq \varepsilon$. □

Remark. Skorokhod's representation theorem also holds for random variables taking values in any separable metric space but the proof is a bit more compli-

cated, see e.g. [2, Theorem 4.30]. In particular for $\mathbb{R}^d$ we have the following: If $(\mu_n)_{n=1}^\infty$ is a sequence of probability measures on $\mathbb{R}^d$ that converges weakly to a probability measure μ, then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{R}^d$-valued random variables $(X_n)_{n=1}^\infty$ and X on Ω such that $X_n \xrightarrow{d} X$ and $(X_n)_* \mathbb{P} = \mu_n$ and $X_* \mathbb{P} = \mu$. ◆

As a corollary we get the following.

Theorem 5.5. *Assume that $(X_n)_{n=1}^\infty$ is a sequence of random variables that converges in law to a random variable X. Then for any continuous $g\colon \mathbb{R} \to \mathbb{R}$ the random variables $g(X_n)$ converge in law to $g(X)$.*

Proof. Skorokhod's representation theorem allows us to assume that in fact $X_n \to X$ almost surely, in which case $g(X_n) \to g(X)$ almost surely, which again implies $g(X_n) \xrightarrow{d} g(X)$. □

Let us next prove the so called Portmanteau theorem which gives us equivalent characterisations of convergence in law.

Definition 5.6. We say that a Borel measurable set $A \subset \mathbb{R}$ is a **continuity set** of a random variable X if $\mathbb{P}[X \in \partial A] = 0$, where ∂A is the (topological) boundary of A. ◆

Theorem 5.7. *Let $(X_n)_{n=1}^\infty$ be a sequence of random variables and X another random variable. Then the following are equivalent:*

(a) $X_n \xrightarrow{d} X$

(b) $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$ *for all bounded and continuous* $h\colon \mathbb{R} \to \mathbb{R}$

(c) $\liminf_{n\to\infty} \mathbb{P}[X_n \in U] \geq \mathbb{P}[X \in U]$ *for all open* $U \subset \mathbb{R}$

(d) $\limsup_{n\to\infty} \mathbb{P}[X_n \in F] \leq \mathbb{P}[X \in F]$ *for all closed* $F \subset \mathbb{R}$

(e) $\lim_{n\to\infty} \mathbb{P}[X_n \in A] = \mathbb{P}[X \in A]$ *for all continuity sets A of X*

Proof. (a) ⇔ (b) is the definition.

(b) ⇒ (c): Let U be open and choose an increasing sequence h_m of continuous and bounded functions such that $h_m \to \mathbb{1}_U$ pointwise. For instance one can set $h_m(x) := (m \operatorname{dist}(x, U^c)) \wedge 1$, where $\operatorname{dist}(x, A) := \inf\{|x - y| : y \in A\}$ is the distance of x from the set $A \subset \mathbb{R}$. Then for any fixed $m \geq 1$ we have $\mathbb{P}[X_n \in U] \geq \mathbb{E}[h_m(X_n)]$ for all n and thus

$$\liminf_{n\to\infty} \mathbb{P}[X_n \in U] \geq \mathbb{E}[h_m(X)].$$

The claim follows by letting $m \to \infty$ and using the monotone convergence theorem.

(c) $\Leftrightarrow$ (d) $\Leftrightarrow$ (e): Exercise.

(c) $\Rightarrow$ (b): Let $h\colon \mathbb{R} \to \mathbb{R}$ be bounded and continuous. By considering the positive and negative parts of h separately it is enough to consider the case $h \geq 0$. Let us denote $\mu_n = (X_n)_* \mathbb{P}$ and $\mu = X_* \mathbb{P}$. We have by Fubini's theorem that

$$\begin{aligned}\mathbb{E}[h(X_n)] &= \int_{\mathbb{R}} h(x) d\mu_n(x) = \int_{\mathbb{R}} \int_0^{\|h\|_\infty} \mathbb{1}_{\{h(x)>t\}}\, dt\, d\mu_n(x) \\ &= \int_0^{\|h\|_\infty} \mathbb{P}[h(X_n) > t]\, dt.\end{aligned}$$

By Fatou's lemma and (c) then

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}[h(X_n)] &\geq \int_0^{\|h\|_\infty} \liminf_{n\to\infty} \mathbb{P}[h(X_n) > t]\, dt \\ &\geq \int_0^{\|h\|_\infty} \mathbb{P}[h(X) > t]\, dt = \mathbb{E}[h(X)],\end{aligned}$$

where we used the fact that $\mathbb{P}[h(X_n) > t] = \mathbb{P}[X_n \in h^{-1}((t,\infty))]$, where by the continuity of h the set $h^{-1}((t,\infty))$ is open. Similarly we can compute that

$$\mathbb{E}[h(X_n)] = \int_0^{\|h\|_\infty} \mathbb{P}[h(X_n) \geq t]\, dt = \|h\|_\infty - \int_0^{\|h\|_\infty} \mathbb{P}[h(X_n) < t]\, dt$$

and thus

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}[h(X_n)] &= \|h\|_\infty - \liminf_{n\to\infty} \int_0^{\|h\|_\infty} \mathbb{P}[h(X_n) < t]\, dt \\ &\leq \|h\|_\infty - \int_0^{\|h\|_\infty} \mathbb{P}[h(X) < t]\, dt \\ &= \int_0^{\|h\|_\infty} \mathbb{P}[h(X) \geq t]\, dt = \mathbb{E}[h(X)].\end{aligned}$$

Thus $\limsup_{n\to\infty} \mathbb{E}[h(X_n)] = \liminf_{n\to\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ as wanted. □

Remark. The above proof works also for $\mathbb{R}^d$-valued random variables. ◆

5.2 *Tightness*

Often when one wants to show convergence in distribution for a given sequence $(X_n)_{n=1}^\infty$ of random variables, it is useful to split the proof into two parts: First one shows that the sequence is *tight*, which means that no probability mass escapes to infinity. Secondly, we will soon see that tightness then

implies that there is a subsequence of X_{n_k} that converges in distribution and to show convergence of the original sequence it is then enough to show that each converging subsequence converges to the same limit.

Definition 5.8. A sequence $(\mu_n)_{n=1}^{\infty}$ of probability measures on $\mathbb{R}$ is **tight** if for every $\varepsilon > 0$ there exists a compact subset $K \subset \mathbb{R}$ such that $\mu_n(K) \geq 1 - \varepsilon$ for all $n \geq 1$.

A sequence $(x_n)_{n=1}^{\infty}$ of random variables is tight if their laws form a tight sequence of probability measures. ◆

Remark. The above definition works also for random variables taking values in $\mathbb{R}^d$ or more generally any metric space. ◆

Theorem 5.9 (Prokhorov's theorem)**.** *A sequence $(X_n)_{n=1}^{\infty}$ of random variables is tight if and only if for every subsequence X_{n_k} of X_n there exists a further subsubsequence $X_{n_{k_j}}$ which converges in law.*

Proof. We prove that if $(X_n)_{n=1}^{\infty}$ is tight then there exists a subsequence which converges in law. The other direction is left as an exercise.

Assume that the sequence $(X_n)_{n=1}^{\infty}$ is tight and let F_n be the c.d.f. of X_n. We will next apply the following lemma.

Lemma 5.10 (Helly's selection theorem)**.** *Every sequence $(F_n)_{n=1}^{\infty}$ of c.d.f.s contains a subsequence F_{n_k} that converges to some right-continuous increasing function F at every continuity point x of F.*

Assuming this lemma for now we will be done if F is a c.d.f., i.e. if

$$\lim_{x\to-\infty} F(x) = 0 \quad \textit{and} \quad \lim_{x\to\infty} F(x) = 1.$$

But this is clear by tightness since for any $\varepsilon > 0$ if $M > 0$ is so large that $F_{n_k}(x) \geq 1 - \varepsilon$ for all $x \geq M$ and x is a continuity point of F then

$$F(x) = \lim_{k\to\infty} F_{n_k}(x) \geq 1 - \varepsilon,$$

showing that $\lim_{x\to\infty} F(x) = 1$ and similar argument works to show the other limit.

Thus it remains to show Lemma 5.10. The proof will be based on a diagonalization argument. Let $(q_n)_{n=1}^{\infty}$ be an enumeration of the rationals. We will inductively construct a sequence $((n_k^{(i)})_{k=1}^{\infty})_{i=1}^{\infty}$ of sequences and then look at the diagonal $n_k^{(k)}$. We begin by simply setting $n_k^1 = k$. Then assume that $n_k^{(i)}$ has been constructed and let $n_k^{(i+1)}$ be a subsequence of $n_k^{(i)}$ such that $F_{n_k^{(i+1)}}(q_i)$ converges to a limit $F(q_i)$ as $k \to \infty$. This is possible since $F_{n_k^i}(q_i)$ is a bounded sequence of real numbers. We thus see that the diagonal subsequence satisfies

$F_{n_k^k}(q_i) \to F(q_i)$ for all $i \geq 1$. Extend then the definition of F to $\mathbb{R}$ by setting

$$F(x) \coloneqq \inf_{y \in \mathbb{Q}, y \geq x} F(y)$$

for $x \in \mathbb{R}$. The function F is increasing since $F(q_i) \leq F(q_j)$ for $q_i < q_j$ and F is clearly right-continuous.

It remains to show that if x is a continuity point of F then $\lim_{k\to\infty} F_{n_k^{(k)}}(x) \to F(x)$. Let $\varepsilon > 0$ and pick a rational $q_+ \geq x$ such that $F(q) - F(x) \leq \varepsilon$. Then $\limsup_{k\to\infty} F_{n_k^{(k)}}(x) \leq F(q) \leq F(x) + \varepsilon$, so by letting $\varepsilon \to 0$ we get that

$$\limsup_{k\to\infty} F_{n_k^{(k)}}(x) \leq F(x).$$

Similarly considering $q \leq x$ such that $F(x) - F(q) \leq \varepsilon$ we get

$$\liminf_{k\to\infty} F_{n_k^{(k)}}(x) \geq F(x),$$

which proves the claim. □

Remark. Prokhorov's theorem also holds for random variables taking values in a separable metric space, in particular $\mathbb{R}^d$. The proof is a bit more involved, see e.g. [2, Theorem 5.19] for the $\mathbb{R}^d$ case. ◆

5.3 *Characteristic functions*

A useful tool in the study of the distribution of a random variable is its Fourier transform or as people in probability like to say *characteristic function*.

Definition 5.11. Let X be a random variable. The **characteristic function** of X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\varphi_X(t) \coloneqq \mathbb{E}[e^{itX}]. \quad ◆$$

A couple of remarks concerning complex valued random variables are in order.

- The expectation of complex valued random variable $Z = X + iY$ with real and imaginary parts $X, Y \in L^1$ can be defined as $\mathbb{E}[Z] = \mathbb{E}[X] + i\mathbb{E}[Y]$.
- One can extend the definition of the L^p spaces ($p \in [0, \infty]$) to contain complex valued random variables by saying that $Z \in L^p$ if both $X \in L^p$ and $Y \in L^p$.
- The definitions of the norms/metrics in these spaces stay basically the same formally since in the end they are all based on looking at absolute distances $|X(\omega) - Y(\omega)|$ of two real numbers, but the absolute value

makes sense for complex numbers as well. Thus for example $\|Z\|_{L^p} = (\mathbb{E}[|Z|^p])^{1/p}$ for $p \geq 1$ or $d_{L^0}(Z, W) = \mathbb{E}[|Z - W| \wedge 1]$.

- Similarly a family $(Z_i)_{i \in I}$ of complex valued random variables is uniformly integrable if for all $\varepsilon > 0$ there exists $\delta > 0$ such that $\mathbb{E}[|Z_i| \mathbb{1}_A] < \varepsilon$ for all events A with $\mathbb{P}[A] < \delta$.

- Since the definitions are formally the same except that the codomain of random variables has been changed from $\mathbb{R}$ to $\mathbb{C}$, most of the proofs of basic theorems concerning expectations and uniform integrability also carry through word-by-word just by changing the codomains of the random variables from $\mathbb{R}$ to $\mathbb{C}$. Arguments which split a function to its negative and positive parts to reduce to the case of non-negative functions also usually work since in the complex case one can simply split into 4 parts corresponding positive/negative real/imaginary parts.

Now, with the above clarifications in mind, we see that the characteristic function is well defined since $e^{itX} \in L^\infty$ for all $t \in \mathbb{R}$.

A simple but very useful feature of characteristic functions is that they work very nicely with sums of independent random variables.

Proposition 5.12. *Let X and Y be independent random variables. Then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_y(t)$$

for all $t \in \mathbb{R}$.

Proof. Clear. □

One of the main properties of characteristic functions is that they characterise the distribution. In fact we have the following inversion theorem.

Theorem 5.13. *Let X be a random variable and $\mu = X_* \mathbb{P}$ be its law. Then*

$$\lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} \varphi_X(t)\, dt = \mu((a, b)) + \frac{1}{2}\mu(\{a, b\}).$$

Exercise 5.14. Show that if one knows $\mu((a, b)) + \frac{1}{2}\mu(\{a, b\})$ for all real numbers $a < b$, then one can recover the probability measure μ. ◆

Proof of Theorem 5.13. Let use write

$$I_T = \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} \varphi_X(t)\, dt.$$

Note that $\frac{e^{-iat}-e^{-ibt}}{it} = \int_a^b e^{-ixt}\,dx$ is bounded, so by Fubini's theorem we have

$$I_T = \int_{-T}^{T}\int_{-\infty}^{\infty} \frac{e^{-iat}-e^{-ibt}}{it} e^{itx}\,d\mu(x)\,dt = \int_{-\infty}^{\infty}\int_{-T}^{T} \frac{e^{-iat}-e^{-ibt}}{it} e^{itx}\,dt\,d\mu(x).$$

By doing the change of variables $t \mapsto -t$ we have that

$$\int_{-T}^{T} \frac{e^{-i(a-x)t}-e^{-i(b-x)t}}{it}\,dt = \int_{-T}^{T} \frac{-e^{i(a-x)t}+e^{i(b-x)t}}{it}\,dt$$

so taking the average of the two sides

$$\int_{-T}^{T} \frac{e^{-i(a-x)t}-e^{-i(b-x)t}}{it}\,dt = \int_{-T}^{T} \frac{e^{-i(a-x)t}-e^{i(a-x)t}+e^{i(b-x)t}-e^{-i(b-x)t}}{2it}\,dt$$

and hence recalling that $(e^{ix}-e^{-ix})/(2i) = \sin(x)$ we have

$$\begin{aligned} I_T &= \int_{-\infty}^{\infty}\Big(\int_{-T}^{T} \frac{\sin((x-a)t)}{t}\,dt - \int_{T}^{T} \frac{\sin((x-b)t)}{t}\,dt\Big)\,d\mu(x) \\ &= 2\int_{-\infty}^{\infty}\Big(\int_{0}^{T} \frac{\sin((x-a)t)}{t}\,dt - \int_{0}^{T} \frac{\sin((x-b)t)}{t}\,dt\Big)\,d\mu(x) \\ &= 2\int_{-\infty}^{\infty}\Big(\int_{0}^{T(x-a)} \frac{\sin(t)}{t}\,dt - \int_{0}^{T(x-b)} \frac{\sin(t)}{t}\,dt\Big)\,d\mu(x). \end{aligned}$$

Let us denote

$$S(u) \coloneqq \int_0^u \frac{\sin(t)}{t}\,dt = \operatorname{sgn}(u)\int_0^{|u|} \frac{\sin(t)}{t}\,dt.$$

Then

$$I_T = 2\int_{-\infty}^{\infty} (S(T(x-a)) - S(T(x-b)))\,d\mu(x).$$

Since $\lim_{u\to\pm\infty} S(u) = \pm\frac{\pi}{2}$ (exercise), we have that

$$\lim_{T\to\infty} (S(T(x-a)) - S(T(x-b))) = \begin{cases} 0, & \text{if } x < a \text{ or } x > b \\ \frac{\pi}{2}, & \text{if } x \in \{a,b\} \\ \pi, & \text{if } a < x < b \end{cases}$$

and the claim follows by the dominated convergence theorem. □

In the case where φ_X is integrable the inversion theorem gets a simpler form.

Theorem 5.15. *Let X be a random variable with characteristic function φ_X. If*

$\int |\varphi_X(t)|\,dt < \infty$, *then* X *has a p.d.f.* f *which is bounded and continuous and*

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)\,dt.$$

Proof. Let μ be the law of X. We may write for any $a < b$ the inversion theorem as

$$\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\int_a^b e^{-ixt}\varphi_X(t)\,dx\,dt.$$

Note that since φ_X is integrable and $\left|\int_a^b e^{-ixt}\right| \le |a-b|$, by the dominated convergence theorem we can take the limit as $T\to\infty$ and apply Fubini's theorem to get

$$\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \frac{1}{2\pi}\int_a^b\int_{-\infty}^{\infty} e^{-ixt}\varphi_X(t)\,dt\,dx.$$

Letting $b\to a$ here shows that $\mu(\{a\}) = 0$ so there are no atoms, and

$$\mu((a,b)) = \int_a^b f(x)\,dx$$

with $f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)\,dt$ as wanted. Clearly $|f(x)| \le \frac{1}{2\pi}\int_{-\infty}^{\infty}|\varphi_X(t)|\,dt$ so f is bounded, and if $x_n \to y$, then by dominated convergence theorem $f(x_n)\to f(y)$, so f is continuous as well. □

The characteristic function is basically the *Fourier transform* of the distribution of the random variable. Let us list a few useful properties of Fourier transforms.

Definition 5.16. Let $f:\mathbb{R}\to\mathbb{R}$ be integrable. The Fourier transform of f is the function $\hat{f}:\mathbb{R}\to\mathbb{C}$ given by

$$\hat{f}(t) = \int_{\mathbb{R}} f(x)e^{-itx}\,dx.$$ ◆

Note the minus sign in the exponential function. In particular if f is the probability density of some random variable X, then

$$\hat{f}(t) = \overline{\varphi_X(t)}.$$

Theorem 5.17. *The Fourier transform satisfies the following:*

(a) If $\hat{f}$ *is integrable, then* $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}}\hat{f}(t)e^{itx}\,dt$.

(b) If f *is compactly supported and smooth*[1]*, then for any* $N \ge 1$ *there exists* $C > 0$ *such that* $|\hat{f}(t)| \le \frac{C}{1+|t|^N}$ *for all* $t\in\mathbb{R}$.

[1] Smooth means that f has derivatives of all orders.

(c) *If μ is the law of some random variable X with characteristic function φ_X and f is an integrable function such that also $\hat{f}$ is integrable, then*

$$\mathbb{E}[f(X)] = \int f(x)d\mu(x) = \frac{1}{2\pi}\int \hat{f}(t)\varphi_X(t)\,dt.$$

Proof. (a) This is basically just a rephrasing of the already proven inversion theorem in the case where the characteristic function was integrable and can be proven in a similar manner.

(b) Note that by integration by parts

$$|\hat{f}(t)| \le \Big|\int f(x)e^{-itx}\,dx\Big| = \Big|\int \frac{d}{d^n}f(x)\frac{e^{-itx}}{(-it)^n}\,dx\Big| \le |t|^{-n}\int |\frac{d}{d^n}f(x)|\,dx.$$

for any $n \ge 1$.

(c) By using Fubini's theorem we have

$$\frac{1}{2\pi}\int \hat{f}(t)\varphi_X(t)\,dt = \frac{1}{2\pi}\int\int \hat{f}(t)e^{itx}d\mu(x)\,dt = \int f(x)\,d\mu(x). \qquad \square$$

As a corollary we obtain the following.

Theorem 5.18. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables.*

(a) If $X_n \xrightarrow{d} X$ for some random variable X, then $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$.

(b) If φ_{X_n} converge to some function $\varphi\colon \mathbb{R} \to \mathbb{C}$ pointwise and φ is continuous at 0, then φ is the characteristic function of some random variable X and $X_n \xrightarrow{d} X$.

Proof. (a) Since $x \mapsto \exp(itx)$ is bounded, the claim follows from the definition of convergence in distribution.

(b) If we can show that X_n is tight, then we are done by part (a) since every subsequential limit of X_n must converge to a random variable with the same characteristic function.

Consider a smooth function $f\colon \mathbb{R} \to \mathbb{R}$ such that $0 \le f \le \mathbb{1}_{[-1,1]}$ and $f(0) = 1$. Note that $\hat{f}$ is integrable, indeed, it decays faster than any polynomial. Fix $\varepsilon > 0$. We have for any $K > 0$ that

$$\mathbb{P}[|X_n| \le K] \ge \mathbb{E}\Big[f\big(\frac{X_n}{K}\big)\Big].$$

Since the characteristic function of X_n/K equals $\varphi_{X_n}(t/K)$, we have

$$\mathbb{E}\Big[f\big(\frac{X_n}{K}\big)\Big] = \frac{1}{2\pi}\int \hat{f}(t)\varphi_{X_n}(t/K)\,dt.$$

Choose next K so large that

$$\frac{1}{2\pi}\int \hat{f}(t)(\varphi_X(t/K)-1)\,dt \le \varepsilon/2.$$

This is possible because of the continuity of φ_X at 0. By the dominated convergence theorem we also have for large enough n that

$$\Big|\frac{1}{2\pi}\int \hat{f}(t)\varphi_{X_n}(t/K)\,dt - \frac{1}{2\pi}\int \hat{f}(t)\varphi_X(t/K)\,dt\Big| \le \varepsilon/2$$

and since $(2\pi)^{-1}\int \hat{f}(t)\,dt = f(0) = 1$ we get

$$\frac{1}{2\pi}\int \hat{f}(t)\varphi_{X_n}(t/K) \ge \frac{1}{2\pi}\int \hat{f}(t)\varphi_X(t/K)\,dt - \varepsilon/2 \ge 1-\varepsilon.$$

Thus $\mathbb{P}[|X_n| \le K] \ge 1-\varepsilon$ for large enough n and by choosing an even larger K if needed we can ensure this for all n, thus showing that the sequence is tight. □

5.4 *Characteristic function on $\mathbb{R}^d$ and the Cramér–Wold theorem*

In this section we will shortly discuss the characteristic functions of $\mathbb{R}^d$-valued random variables.

Definition 5.19. Let X be an $\mathbb{R}^d$-valued random variable. Then the characteristic function of X is the map $\varphi_X\colon \mathbb{R}^d \to \mathbb{C}$ given by

$$\varphi_X(t) \coloneqq \mathbb{E}[e^{it\cdot X}],$$

where $t\cdot X = t_1X_1 + \cdots + t_nX_n$ is the dot product. ◆

An analogue of the inversion formula in this case is as follows.

Theorem 5.20. *Let X be an $\mathbb{R}^d$-valued random variable and $\mu = X_*\mathbb{P}$ be its law. Then*

$$\lim_{T_1,\dots,T_d\to\infty} \frac{1}{(2\pi)^d}\int_{-T_1}^{T_1}\cdots\int_{-T_d}^{T_d}\prod_{k=1}^{d}\frac{e^{-ia_kt_k}-e^{-ib_kt_k}}{it_k}\varphi_X(t)\,dt_1\dots dt_d$$
$$= \mu([a_1,b_1]\times\cdots\times[a_d,b_d])$$

assuming that $[a_1,b_1]\times\cdots\times[a_d,b_d]$ is a continuity set of X.

Proof. We skip the proof, but one can essentially mimic the one we did in the 1-dimensional case. □

Again one sees that the characteristic function of $\mathbb{R}^d$-valued random variable determines its law and that Theorem 5.18 holds.

As our first application of characteristic functions we will prove the following quite useful theorem which reduces showing convergence in law from d-dimensions to 1 dimension.

Theorem 5.21 (Cramér–Wold theorem)**.** *Let $(X_n)_{n=1}^{\infty}$, X be a $\mathbb{R}^d$-valued random variables. Then $X_n \xrightarrow{d} X$ if and only if for every $t \in \mathbb{R}^d$ we have $t \cdot X_n \xrightarrow{d} t \cdot X$.*

Proof. If $X_n \xrightarrow{d} X$, then since the map $t \mapsto t \cdot X$ is continuous, from (a $\mathbb{R}^d$-valued version) of Theorem 5.5 we get that $t \cdot X_n \xrightarrow{d} t \cdot X$.

For the other direction, note that if $t \cdot X_n \xrightarrow{d} t \cdot X$, then

$$\varphi_{X_n}(t) = \varphi_{t \cdot X_n}(1) \to \varphi_{t \cdot X}(1) = \varphi_X(t),$$

so the characteristic functions converge pointwise and hence $X_n \xrightarrow{d} X$. □

5.5 *The moment problem*

As another application of characteristic functions we will look at the moment problem.

Definition 5.22. Let X be a random variable and $n \geq 1$. If $\mathbb{E}[|X|^n] < \infty$, we say call the number $M_n := \mathbb{E}[X^n]$ the ***n*th moment** of X. ◆

The moment problem asks whether it is possible to construct from a list of moments $(M_n)_{n=1}^{\infty}$ a random variable X with the given moments, and if the answer is "yes", whether the law of X is unique. In general the answer to the first question is no, and even if such a random variable exists it may fail to be unique.

For the existence it is not so easy to give good conditions[2], and we will focus on showing that under some mild assumption on the growth of the moments the random variable is indeed unique.

Theorem 5.23. *Let X and Y be two random variables such that all the moments $M_n = \mathbb{E}[X^n] = \mathbb{E}[Y^n]$ are finite. If*

$$\limsup_{n \to \infty} \frac{|M_n|^{1/n}}{n} < \infty$$

then X and Y have the same law.

[2] One can show that M_n is a sequence of moments of some measure on real line if and only if the infinite matrix $(M_{n+m})_{n,m}$ is positive semi-definite, but this is usually not a very practical condition to work with.

Proof. It is enough to show that M_n determine the characteristic function of X. Note that the condition implies that there exists $K > 0$ such that $|M_n| \leq K^n n^n$. Now if $a \in \mathbb{R}$ and $z \in \mathbb{C}$ is such that $|z-a| < 1/(Ke)$ then by Stirling's formula we have $n! \geq n^n/e^n$ for large enough n (in fact this holds for all $n \geq 1$) and thus

$$\sum_{n=1}^{\infty} \frac{|z-a|^n K^n n^n}{n!} \lesssim \sum_{n=1}^{\infty} (|z-a|Ke)^n < \infty,$$

so by Fubini's theorem we have

$$\begin{aligned}\varphi_X(z) = \mathbb{E}[e^{i(z-a)X} e^{iaX}] &= \mathbb{E}[\sum_{n=1}^{\infty} \frac{(i(z-a)X)^n}{n!} e^{iaX}] \\ &= \sum_{n=1}^{\infty} \frac{(i(z-a))^n \mathbb{E}[X^n e^{iaX}]}{n!}.\end{aligned}$$

Thus φ_X is an analytic function in $U = \{z \in \mathbb{C} : |\operatorname{Im}(z)| < 1/(Ke)\}$, and in particular by setting $a = 0$ we see that its values in $B(0, 1/K)$ are determined by the sequence M_n. By analytic continuation we thus see that M_n determine the values $\varphi_X(z)$ for all $z \in U$ and in particular for all $z \in \mathbb{R}$. □

Exercise 5.24. The condition in Theorem 5.23 is not optimal and can be sharpened e.g. to **Carleman's condition**

$$\sum_{n=1}^{\infty} M_{2n}^{-1/(2n)} = \infty.$$

This condition is however not optimal either and the proof is more complicated and out of the scope of this course. ◆

5.6 *Central limit theorem*

Like the law of large numbers, the central limit theorem is a statement about sums $S_n = \sum_{k=1}^n X_k$ of i.i.d. random variables X_k. Where the law of large numbers looks at $n^{-1}S_n$ where we have normalized the sum to have a constant mean, the central limit theorem instead looks at the fluctuations of S_n around its mean on the level where it has a constant variance, namely $n^{-1/2}(S_n - \mathbb{E}[S_n])$. The surprising result is then that this converges in distribution to a normal random variable with variance $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]$.

This can also be seen as a heuristic justification for the choice of normal random variables as a model of statistical experiments where randomness from many independent sources is added up.

Theorem 5.25. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of i.i.d. random variables in L^2. Then*

$$\frac{S_n - n\mathbb{E}[X_1]}{\sqrt{n}}$$

converges in law to a normal random variable with mean 0 *and variance* $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2]$.

Proof. We may without loss of generality assume that $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$. Let $Z_n = n^{-1/2}S_n$. Our aim is to show that Z_n converges in distribution to a standard normal random variable Z. It is thus enough to show that the characteristic functions $\varphi_{Z_n}(t)$ converge pointwise to $e^{-\frac{t^2}{2}} = \mathbb{E}[e^{itZ}]$. Since Z_n is a sum of i.i.d. random variables, we have that

$$\varphi_{Z_n}(t) = (\varphi_{n^{-1/2}X_1}(t))^n = (\varphi_{X_1}(t/\sqrt{n}))^n.$$

Let us next consider $\varphi_{X_1}(t) = \mathbb{E}[e^{itX}]$. Note that it is differentiable since by linearity

$$\frac{\mathbb{E}[e^{i(t+h)X}] - \mathbb{E}[e^{itX}]}{h} = \mathbb{E}[\frac{e^{i(t+h)X} - e^{itX}}{h}]$$

and $|e^{i(t+h)X} - e^{itX}| = |e^{ihX} - 1| \leq |hX|$ so one can apply the dominated convergence theorem. (Note that $|hX|$ is the length of the arc on the unit circle connecting the points 1 and e^{ihX}.) Thus $\varphi'_{X_1}(t) = \mathbb{E}[iXe^{itX}]$ and in particular $\varphi'_{X_1}(0) = \mathbb{E}[iX] = 0$. Similarly we see that $\varphi^{(2)}_{X_1}$ exists and $\varphi^{(2)}_{X_1}(0) = -\mathbb{E}[X^2] = -1$. Thus by Taylor's theorem

$$\varphi_{X_1}(t) = 1 - \frac{t^2}{2} + o(t^2)$$

as $t \to 0$. This means in particular that

$$\varphi_{Z_n}(t) = \varphi_{X_1}(t/\sqrt{n})^n = (1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big))^n.$$

We next note that for any $a, b \in \mathbb{C}$ with $|a|, |b| \leq 1$ we have the general bound

$$|a^n - b^n| = |(a-b)(a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1})| \leq n|a-b|.$$

Applying this with $a = \varphi_{X_1}(t/\sqrt{n})$ and $b = 1 - \frac{t^2}{2n}$ (for large enough n so that $|b| \leq 1$) we see that

$$|\varphi_{Z_n}(t) - (1 - \frac{t^2}{2n})^n| \leq no\Big(\frac{t^2}{n}\Big) \to 0$$

as $n \to \infty$. Since famously $(1 - \frac{t^2}{2n})^n \to e^{-\frac{t^2}{2}}$, the proof is completed. □

The conditions of the central limit theorem can be relaxed in various ways. Let us mention the following where the variables are no longer assumed to be identically distributed and take values in $\mathbb{R}^d$.

Let us recall the definition of a multivariate normal random vector.

Definition 5.26. Let $C \in \mathbb{R}^{d\times d}$ be a positive definite matrix, i.e. $v^T C v \geq 0$ for all $v \in \mathbb{R}^d$. An $\mathbb{R}^d$-valued random variable X is said to have normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix C if X has the p.d.f.

$$p_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(C)}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}.$$

This distribution is often denoted by $N(\mu, C)$. ◆

Let us also denote by $\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] \in \mathbb{R}^{d\times d}$ the covariance between two $\mathbb{R}^d$-valued random variables. Here the expectation is taken coordinate wise.

Theorem 5.27 (Lindeberg–Feller theorem)**.** *Assume that for every $n \geq 1$ the $\mathbb{R}^d$-valued random variables $X_{n,1}, \dots, X_{n,k_n}$ ($k_n \geq 1$) are independent. Furthermore, suppose that*

(a) $\sum_{j=1}^{k_n} \mathbb{E}[X_j] \to \mu \in \mathbb{R}^d$

(b) $\sum_{j=1}^{k_n} \mathrm{Cov}(X_{n,j}, X_{n,j}) \to C \in \mathbb{R}^{d\times d}$.

(c) For all $\varepsilon > 0$ we have $\sum_{j=1}^{k_n} \mathbb{E}[|X_{n,j}|^2 \mathbb{1}_{|X_{n,j}|>\varepsilon}] \to 0$, where $|\cdot|$ is the Euclidean norm (or more generally any norm) on $\mathbb{R}^d$.

Then

$$S_n \coloneqq \sum_{j=1}^{k_n} X_{n,j} \xrightarrow{d} N(\mu, C).$$

Proof sketch. First of all we note that by the Cramér–Wold theorem it is enough to consider the case $d = 1$. Indeed if we look at the scalars $Y_{n,j} \coloneqq t \cdot X_{n,j}$ for some $t \in \mathbb{R}^d$, then $\sum_{j=1}^{k_n} \mathbb{E}[Y_{n,j}] \to t \cdot \mu$ and $\sum_{j=1}^{k_n} \mathbb{E}[Y_{n,j}^2] \to t^T C t$. The third condition is also satisfied since it clearly holds if $t = 0$ and if $t \neq 0$ we have by Cauchy–Schwarz that

$$\sum_{j=1}^{k_n} \mathbb{E}[|Y_{n,j}|^2 \mathbb{1}_{\{|Y_{n,j}|>\varepsilon\}}] \leq |t|^2 \sum_{j=1}^{k_n} \mathbb{E}[|X_{n,j}|^2 \mathbb{1}_{\{|X_{n,j}|>\varepsilon|t|^{-1}\}}] \to 0.$$

Hence from the one dimensional result we would get

$$t \cdot \sum_{j=1}^{k_n} X_{n,j} = \sum_{j=1}^{k_n} Y_{n,j} \xrightarrow{d} N(t \cdot \mu, t^T C t)$$

for all $t \in \mathbb{R}^d$ and hence since $N(t \cdot \mu, t^T C t) \overset{d}{=} t \cdot N(\mu, C)$ the result would follow from Cramér–Wold.

To show the one dimensional case one can pretty much just follow along the lines of the proof of Theorem 5.25. Note that we can assume that $\mu = 0$ by subtracting from each $X_{n,j}$ its mean $\mathbb{E}[X_{n,j}]$. We again look at the characteristic function of S_n,

$$\varphi_{S_n}(t) = \prod_{j=1}^{k_n} \varphi_{X_{n,j}}(t).$$

This time one needs to look at the Taylor expansion of $\varphi_{X_{n,j}}(t)$ a bit more carefully. Let us fix $t \in \mathbb{R}$. Note that

$$\varphi_{X_{n,j}}(t) = \mathbb{E}[e^{itX_{n,j}} \mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] + \mathbb{E}[e^{itX_{n,j}} \mathbb{1}_{\{|X_{n,j}| > \varepsilon\}}].$$

In the first term we may use a second order Taylor expansion $e^{ixt} = 1 + ixt - x^2t^2/2 + O(x^3)$ to get

$$\begin{aligned}\mathbb{E}[e^{itX_{n,j}} \mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] = \mathbb{E}[\mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] + \mathbb{E}[itX_{n,j} \mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] \\ - \frac{t^2}{2} \mathbb{E}[X_{n,j}^2 \mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] + O(\varepsilon \mathbb{E}[X_{n,j}^2]),\end{aligned}$$

while for the second term we use first order Taylor expansion to get

$$\mathbb{E}[e^{itX_{n,j}} \mathbb{1}_{\{|X_{n,j}| > \varepsilon\}}] = \mathbb{E}[\mathbb{1}_{\{|X_{n,j}| > \varepsilon\}}] + \mathbb{E}[itX_{n,j} \mathbb{1}_{\{|X_{n,j}| \leq \varepsilon\}}] + O(\mathbb{E}[|X_{n,j}|^2 \mathbb{1}_{\{|X_{n,j}| > \varepsilon\}}]).$$

Adding these up we get

$$\varphi_{X_{n,j}}(t) = 1 - \frac{t^2}{2} \mathbb{E}[X_{n,j}^2] + O(\varepsilon \mathbb{E}[|X_{n,j}|^2]) + O(\mathbb{E}[|X_{n,j}|^2 \mathbb{1}_{\{|X_{n,j}| > \varepsilon\}}]).$$

Next we would like to show that

$$\prod_{j=1}^{k_n} \varphi_{X_{n,j}}(t) - \prod_{j=1}^{k_n} (1 - \frac{t^2}{2} \mathbb{E}[X_{n,j}^2])$$

goes to 0 as $n \to \infty$. Here instead of the inequality $|a^n - b^n| \leq n|a - b|$ that we

had before we can use a more general inequality

$$\Big|\prod_{j=1}^{n} a_j - \prod_{j=1}^{n} b_j\Big| \leq \sum_{j=1}^{n} |a_j - b_j|$$

for all complex numbers $(a_j)_{j=1}^n$, $(b_j)_{j=1}^n$ with $|a_j|, |b_j| \leq 1$ (exercise). Thus we get

$$\Big|\prod_{j=1}^{k_n} \varphi_{X_{n,j}}(t) - \prod_{j=1}^{k_n} \big(1 - \frac{t^2}{2}\mathbb{E}[|X_{n,j}|^2]\big)\Big| \lesssim \sum_{j=1}^{k_n} (\varepsilon \mathbb{E}[|X_{n,j}|^2] + \mathbb{E}[|X_{n,j}|^2 \mathbb{1}_{\{|X_{n,j}|>\varepsilon\}}]),$$

where the right hand side tends to εC as $n \to \infty$. This proves the claim since we can choose ε as small as we wish. The proof is finished either by taking logarithms and doing Taylor expansion, or by looking at the above computation in the case where $X_{n,j}$ are normal r.v.s, to check that

$$\prod_{j=1}^{k_n} \big(1 - \frac{t^2}{2}\mathbb{E}[|X_{n,j}|^2]\big) \to e^{-\frac{t^2}{2}C}. \qquad \square$$

5.7 *Stable laws and further limit theorems*

The CLT shows that for centered i.i.d. random variables $(X_n)_{n=1}^\infty$ with finite variance there is a universal limit for the normalized sum $n^{-1/2}(X_1 + \cdots + X_n)$. Next it would be natural to ask what happens if the variables do not have finite variance. For simplicity we will assume that they are however symmetric, i.e. $X_1 \overset{d}{=} -X_1$.

To answer the question, it is helpful to start by thinking backwards – what can we say if we have a CLT-type result, saying that

$$c_n(X_1 + \cdots + X_n) \overset{d}{\to} Y$$

for some nonzero random variable Y and normalizing constants $c_n > 0$? We may rewrite this for $2n$ random variables as

$$\frac{c_{2n}}{c_n}\big(c_n(X_1 + \cdots + X_n) + c_n(X_{n+1} + \cdots + X_{2n})\big) \overset{d}{\to} Y. \tag{5.1}$$

Now one can check that $c_n(X_1 + \cdots + X_n) + c_n(X_{n+1} + \cdots + X_{2n}) \overset{d}{\to} Y_1 + Y_2$ where Y_1 and Y_2 are independent and have the same distribution as Y. Furthermore, it is also not hard to prove that in order to have convergence in (5.1) also $\frac{c_{2n}}{c_n}$ has to converge to some positive constant which we will call d_2. Hence we get

in particular the functional equation

$$\varphi_Y(t) = (\varphi_Y(d_2 t))^2$$

for all $t \in \mathbb{R}$.

Next one can ask what are all the functions φ_Y that satisfy this functional equation. For instance we can check that if $Y = c \neq 0$ is a constant, then this implies that $e^{itc} = e^{i2d_2tc}$ for all t, so that $d_2 = 1/2$ and in particular $c_{2^n} = 2^{-n}c_1 + o(2^{-n})$ corresponds to the normalization appearing in the law of large numbers and we would have $Y = 0$ which is a contradiction. Thus we may from now on assume that Y is not constant.

In general it however unfortunately seems like there might still be quite a few solutions, since basically given e.g. any positive real function f on $[d_2, 1]$ with $f(d_2)^2 = f(1)$ one can uniquely extend f into a function $f\colon (0, \infty]$ that satisfies the functional equation for positive t. One also has $\lim_{t\to 0+} f(t) = 1$ and setting $f(t) = f(-t)$ for $t < 0$ gives a function which satisfies the functional equation and even has $f(0) = 1$. This is still not necessarily a characteristic function of some probability measure, but it indicates that perhaps a bit more information is needed to nail things down.

Luckily we can do a similar splitting as before, but this time into 3 parts to get an additional functional equation

$$\varphi_Y(t) = (\varphi_Y(d_3 t))^3.$$

This should really help us fix things since now every $\varphi_Y(t)$ can be related to $\varphi_Y(1)$ by multiplying by suitable powers of d_2 and d_3.

Lemma 5.28. *Let x and y be two positive real numbers.*

(a) Then there exist two sequences $(a_k)_{k=1}^\infty$ and $(b_k)_{k=1}^\infty$ of nonzero integers such that $x^{a_k} y^{b_k} \to 1$.

(b) Moreover if $\log(x)/\log(y)$ is irrational, then for any $t > 0$ there exist two sequences $(a_k)_{k=1}^\infty$ and $(b_k)_{k=1}^\infty$ such that $x^{a_k} y^{b_k} \to t$.

Proof. Exercise. □

To characterise φ it is perhaps easiest to start with the function $h(t) := |\varphi_Y(t)|$, since then we can take roots and obtain the more general functional equation

$$h(t) = (h(d_2^a d_3^b t))^{2^a 3^b}$$

for all $a, b \in \mathbb{Z}$.

Since Y is not a constant, there exists $t \in \mathbb{R}$ such that $0 < |h(t)| < 1$. [3] Then by choosing a_k and b_k as in the lemma, we see that

$$h(t) = (h(d_2^{a_k} d_3^{b_k} t))^{2^{a_k} 3^{b_k}}.$$

Since $d_2^{a_k} d_3^{b_k} t \to t$, we must have $2^{a_k} 3^{b_k} \to 1$. Taking logarithms we see that $a_k \log(2) + b_k \log(3) \to 0$, and $a_k \log(d_2) + b_k \log(d_3) \to 0$. This implies that $\frac{\log(d_2)}{\log(d_3)} = \frac{\log(2)}{\log(3)}$. In particular if we write $d_2 = 2^{-\beta}$ for some β, then $d_3 = 3^{-\beta}$.

Having identified the relationship between d_2 and d_3 we now actually see that since $\log(2)/\log(3)$ is irrational, for any $t > 0$ we can in fact choose sequences $(a_k)_{k=1}^{\infty}$ and $(b_k)_{k=1}^{\infty}$ so that $2^{-a_k\beta} 3^{-b_k\beta} \to t$. The equation

$$h(1) = h(2^{-a_k\beta} 3^{-b_k\beta})^{2^{a_k} 3^{b_k}}$$

then implies by continuity that

$$h(t) = h(1)^{t^{1/\beta}}.$$

Writing $h(1) = e^{-c}$ for some $c > 0$, we have that

$$h(t) = e^{-ct^{1/\beta}}.$$

Finally since φ_Y is continuous and real (since Y is symmetric), we must have that $\varphi_Y(t) = e^{-c|t|^{1/\beta}}$ for some $c, \beta > 0$. Probability distribution with such characteristic functions are called *stable laws*.

Definition 5.29. For $\alpha \in (0, 2]$ and $c \geq 0$ random variable Y is said to have a **symmetric α-stable distribution with parameter c** if its characteristic function is of the form

$$\varphi_Y(t) = e^{-c|t|^{\alpha}}. \qquad \blacklozenge$$

The word stable refers to the fact that sums of independent α-stable random variables stay α-stable.

There is no point in extending the definition to $\alpha > 2$ since in this case φ_Y will be twice differentiable which will imply that $\mathbb{E}[Y^2] = 0$, so $Y \overset{a.s.}{=} 0$ (details left to the reader).

To see that for $\alpha \leq 2$ the function $e^{-c|t|^{\alpha}}$ is a characteristic function of a random variable is a bit tricky since for $\alpha \notin \{1/2, 1, 2\}$ the corresponding p.d.f. has no formula in terms of elementary functions. To show the existence of such stable laws one can e.g. take (X_n) a sequence of i.i.d. random variables with

[3] Indeed, if one picks two disjoint intervals $[a, b]$ and $[c, d]$ such that Y positive probability of hitting either of them, then for small enough $t > 0$ we have $[ta, tb], [tc, td] \subset [-\pi, \pi]$ and disjoint, so $|\mathbb{E}[e^{itY}]| < 1$ because $|\mathbb{E}[e^{itY}]| = 1$ holds if and only if e^{itY} is constant almost surely.

$\mathbb{P}[X_n > \lambda] = \mathbb{P}[X_n < -\lambda] = x^{-\alpha}/2$ for all $x > 1$. Then the characteristic functions of $(X_1 + \cdots + X_n)/n^{1/\alpha}$ tend to $e^{-C|t|^\alpha}$ for some $C > 0$. See [1, Section 3.8] for details and more about stable laws.

The end.

Metrics and pseudometrics

In this appendix we give a very minimal review of metric and pseudometric spaces.

Definition A.1. Let X be a set. A **pseudometric** d on X is a map $d\colon X \times X \to [0, \infty)$ such that for all $x, y, z \in X$ we have

- $d(x, x) = 0$ (*identity*)
- $d(x, y) = d(y, x)$ (*symmetry*)
- $d(x, y) \leq d(x, z) + d(z, y)$ (*triangle inequality*)

The pseudometric d is a **metric** if it satisfies the stronger property

- $d(x, y) = 0$ if and only if $x = y$. (*identity and indiscernibles*)

The pair (X, d) is called a **(pseudo)metric space**. ◆

Any pseudometric space can be made into a metric space by identifying the points with distance 0 from each other, i.e. considering the equivalence relation $x \sim y \Leftrightarrow d(x, y) = 0$ and defining on the set of equivalence classes $X/\sim$ the metric $\tilde{d}([x], [y]) = d(x, y)$. We leave it to the reader to check that $\sim$ is an equivalence relation and that $\tilde{d}$ is a well-defined metric. For instance the Ky Fan metric d_{KF} from Section 2.2 is only a pseudometric on the space of all random variables, but becomes a metric once we identify almost surely equal random variables.

A pseudometric d induces a topology τ_d on X where a set $U \subset X$ is open if and only if for every $x \in U$ there exists $r > 0$ such that the open ball $B_d(x, r) \coloneqq \{y \in X : d(x, y) < r\}$ is contained in U. This topology is Hausdorff if and only if d is a metric.

In particular one easily checks the following.

Proposition A.2. *Let (X, d_X) and (Y, d_Y) be two pseudometric spaces and let $f\colon X \to Y$ be a function. Then f is continuous at $x \in X$ if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B_{d_X}(x, \delta)) \subset B_{d_Y}(f(x), \varepsilon)$.*

One of the most useful properties of pseudometric spaces is that their topology can alternatively be characterized using sequences.

Definition A.3. Let (X, d) be a pseudometric space and suppose that $(x_n)_{n=1}^{\infty}$ is a sequence of points in X. Then we say that x_n **converges** to a point $x \in X$ and write $x_n \to x$ if and only if $d(x_n, x) \to 0$. ◆

In a metric space the limits of sequences are unique but note that if we have $d(x, y) = 0$ for some $x \neq y$ in a pseudometric space, then every sequence that converges to x also converges to y.

The closure of a set is nicely described using sequences.

Proposition A.4. *Let (X, d) be a pseudometric space and let $A \subset X$. Then $x \in \bar{A}$ (the topological closure of A) if and only if there exists a sequence $(a_n)_{n=1}^{\infty}$ of points in A such that $a_n \to x$.*

Proof. Suppose first that $x \in \bar{A}$. Then every open ball $B_n = B_d(x, 1/n)$ intersects A and we may pick $a_n \in B_n \cap A$. Clearly $a_n \to x$.

Conversely suppose that $a_n \to x$ for some sequence $(a_n)_{n=1}^{\infty}$ of points in A. If U is any open neighbourhood of x then it contains an open ball $B_d(x, r)$ for some $r > 0$ and by definition $a_n \in B_d(x, r)$ for all n large enough. Hence U intersects A and as U was arbitrary we have $x \in \bar{A}$. □

In particular a set $A \subset X$ is closed if and only if the limits of all converging sequences in A stay in A. Thus closed sets are determined by converging sequences and as the open sets are exactly the complements of closed sets, also the topology is determined by converging sequences.

Basic topological notions such as continuity can also be stated in terms of sequences.

Proposition A.5. *Let (X, d_X) and (Y, d_Y) be pseudometric spaces and $f : X \to Y$ a function. Then f is continuous at $x \in X$ if and only if for every sequence $x_n \to x$ we have $f(x_n) \to f(x)$.*

Proof. If f is continuous at x and $x_n \to x$, then for any $\varepsilon > 0$ we may pick $\delta > 0$ such that $f(B_{d_X}(x, \delta)) \subset f(B_{d_Y}(f(x), \varepsilon))$. As $x_n \in B_X(x, \delta)$ eventually, we see that $f(x_n) \in B_{d_Y}(f(x), \varepsilon)$ eventually which (since $\varepsilon > 0$ was arbitrary) implies that $f(x_n) \to f(x)$.

Conversely suppose that f is not continuous at x. Then we may pick $\varepsilon > 0$ such that for every $n \geq 1$ we have $f(B_{d_X}(x, 1/n)) \not\subset f(B_{d_Y}(f(x), \varepsilon))$. Letting $x_n \in B_{d_X}(x, 1/n)$ be such that $d_Y(f(x_n), f(x)) > \varepsilon$ we thus get a sequence $x_n \to x$ such that $f(x_n) \not\to f(x)$. □

A stronger property than continuity is *uniform continuity*, which is not anymore a purely topological notion.

Definition A.6. Let (X, d_X) and (Y, d_Y) be pseudometric spaces and $f : X \to Y$ a function. We say that f is **uniformly continuous** if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_X(x, x') < \delta$ implies $d_Y(f(x), f(x')) < \varepsilon$. ◆

A function $f : X \to Y$ between pseudometric spaces is **Lipschitz** if and only if there exists $K > 0$ such that $d_Y(f(x), f(x')) \leq K d_X(x, x')$ for all $x, x' \in X$. Lipschitz functions are hence uniformly continuous (choose $\delta = \varepsilon/K$ in the

definition). A very important special case is that of an **isometry**, which means that $d_Y(f(x), f(x')) = d_X(x, x')$ for all $x, x' \in X$.

More generally uniformly continuous functions can be characterized as follows.

Proposition A.7. *Let (X, d_X) and (Y, d_Y) be pseudometric spaces. A function $f\colon X \to Y$ is uniformly continuous if and only if there exists an increasing function $\varphi\colon [0,\infty) \to [0,\infty]$ (possibly allowing the value ∞) such that $\varphi(t) \to 0$ as $t \to 0$, $\varphi(0) = 0$ and*

$$d_Y(f(x), f(x')) \leq \varphi(d_X(x, x'))$$

for all $x, x' \in X$.

Proof. If such φ exists, then for every $\varepsilon > 0$ we can find $\delta > 0$ such that $\varphi(t) < \varepsilon$ when $t < \delta$, and thus $d_Y(f(x), f(x')) < \varepsilon$ when $d_X(x, x') < \delta$, showing that f is uniformly continuous. If on the other hand f is uniformly continuous then defining

$$\varphi(t) = \sup\{d_Y(f(x), f(x')) : d_X(x, x') \leq t\}.$$

it is easy to check that φ satisfies the needed properties. □

A function φ as in the above proposition is called a **modulus of continuity** of f.

Uniform continuous functions can be thought of as functions that at least on small enough scales stretch the distances in a uniform manner. This makes them useful for example when studying Cauchy sequences.

Definition A.8. Let (X, d) be a pseudometric space. A sequence $(x_n)_{n=1}^{\infty}$ of points in X is **Cauchy** if and only if for every $\varepsilon > 0$ there exists $N \geq 1$ such that $d(x_n, x_m) < \varepsilon$ for all $n, m \geq N$. ◆

Note that if $f\colon X \to Y$ is uniformly continuous and $(x_n)_{n=1}^{\infty}$ is a Cauchy sequence in X then $f(x_n)$ is a Cauchy sequence in Y.

Cauchy sequences give an intrinsic way of saying which sequences *should* converge because their points get closer and closer together. It can however happen that the space is missing the anticipated limit point. For instance in $(\mathbb{Q}, |\cdot|)$ the recursively defined sequence $x_1 = 1$, $x_n = 1 + \frac{1}{x_{n-1}}$ for $n \geq 2$ is Cauchy but it converges to the golden ratio $\frac{1+\sqrt{5}}{2}$ which is an irrational number.

Definition A.9. A pseudometric space (X, d) is **complete** if every Cauchy sequence in X converges. ◆

A central result is that any metric space can be completed in an essentially unique way by adding some points. For pseudometric spaces it is a bit less clear what would be the right completion since one can always add more points with distance 0 to some existing point. There is a unique Hausdorff completion of a

pseudometric space but this is just the completion of the induced metric space $X/\sim$ mentioned above and therefore loses the original information on equidistant points. Therefore we will only discuss completions of metric spaces.

Definition A.10. A **completion** of a metric space X is a pair $(\hat{X}, \iota)$ where $\hat{X}$ is a complete metric space and $\iota\colon X \to \hat{X}$ is an isometry such that $\iota(X)$ is dense in $\hat{X}$. ◆

The main result for completions is the following.

Theorem A.11. *Every metric space has a completion.*

We will skip the proof since you have probably already seen it and also because we will only need completions of normed spaces which we will discuss in Appendix B. The standard and perhaps most principled way of constructing the completion would be to define $\hat{X}$ as the set of Cauchy sequences on X, modulo the equivalence relation that two Cauchy sequences $(x_n)_{n=1}^\infty$ and $(y_n)_{n=1}^\infty$ are equivalent if $d_X(x_n, y_n) \to 0$. One can then proceed to define

$$\iota(x) \coloneqq [(x, x, \dots)] \quad \text{and} \quad d_{\hat{X}}([x_n], [y_n]) \coloneqq \lim_{n\to\infty} d_X(x_n, y_n)$$

and show that this metric is well-defined and complete and that ι is an isometry. Another somewhat shorter proof of existence goes through the so called Kuratowski embedding, which embeds X isometrically into a complete metric space. One can then define the completion of X by taking the closure of X inside this bigger space.

The specific construction of the completion is however usually not important since once we know that they exist there are cleaner ways to characterise them:

Proposition A.12. *A completion $(\hat{X}, \iota)$ of a metric space X satisfies and is characterised up to an isomorphism[1] by the following universal property: If $f\colon X \to Y$ is a uniformly continuous map from X to a complete metric space Y then there is a unique uniformly continuous extension $\hat{f}\colon \hat{X} \to Y$ satisfying $f = \hat{f} \circ \iota$.*

Proof. Suppose first that $f\colon X \to Y$ is a uniformly continuous map. As $\iota(X)$ is dense in $\hat{X}$, for any $\hat{x} \in \hat{X}$ there is a sequence $x_n \in X$ such that $\iota(x_n) \to \hat{x}$. Then as f is uniformly continuous, the sequence $f(x_n)$ is Cauchy and thus converges to some limit in Y which we call $\hat{f}(x)$. Moreover if we had picked another sequence $\iota(x_n') \to x$, then $d_Y(f(x_n'), f(x_n)) \le \varphi(d_X(x_n, x_n')) \to 0$, where φ is the modulus of continuity of f. Thus $\hat{f}$ is a well-defined function and clearly $f = \hat{f} \circ \iota$ as well. It is moreover uniformly continuous with the same modulus of continuity since for any $\hat{x}, \hat{y} \in \hat{X}$ we may pick sequences $(x_n)_{n=1}^\infty$

[1] An isomorphism between two such pairs $(\hat{X}_1, \iota_1)$ and $(\hat{X}_2, \iota_2)$ is a map $S\colon \hat{X}_1 \to \hat{X}_2$ which is an isometric isomorphism and satisfies $\iota_2 = S \circ \iota_1$.

and $(y_n)_{n=1}^{\infty}$ with $\iota(x_n) \to \hat{x}$ and $\iota(y_n) \to \hat{y}$ and then

$$d_Y(\hat{f}(\hat{x}), \hat{f}(\hat{y})) = \lim_{n\to\infty} d_Y(f(x_n), f(y_n)) \leq \lim_{n\to\infty} \varphi(d_X(x_n, y_n)) = \varphi(d_{\hat{X}}(x, y)).$$

The map $\hat{f}$ is unique since it is continuous and has to equal $f \circ \iota^{-1}$ on $\iota(X)$ which is dense.

To show that the universal property characterises the completions of X, suppose that $(\hat{X}_1, \iota_1)$ and $(\hat{X}_2, \iota_2)$ are two completions of X. Then applying the universal property to the maps ι_2 and ι_1 we get two maps $S_1 : \hat{X}_1 \to \hat{X}_2$ and $S_2 : \hat{X}_2 \to \hat{X}_1$ respectively with $\iota_2 = S_1 \circ \iota_1$ and $\iota_1 = S_2 \circ \iota_2$. But this shows that $\iota_1 = S_2 \circ S_1 \circ \iota_1$ so that $S_2 \circ S_1$ is identity on $\iota_1(X)$ and by uniqueness has to be identity on whole $\hat{X}_1$. Similarly $S_1 \circ S_2$ is the identity map on $\hat{X}_2$. Thus S_1 is a bijection with $S_1^{-1} = S_2$. Moreover $d_{\hat{X}_2}(S_1(\iota_1(x)), S_1(\iota_1(y))) = d_{\hat{X}_2}(\iota_2(x), \iota_2(y)) = d_X(x, y) = d_{\hat{X}_1}(\iota_1(x), \iota_1(y))$ for all $x, y \in X$ so S_1 is an isometry when restricted to $\iota_1(X)$, but then by continuity and density of $\iota_1(X)$ in $\hat{X}_1$ it has to be an isometry on whole $\hat{X}_1$. Hence $\hat{X}_1$ is isometrically isomorphic to $\hat{X}_2$ which finishes the proof. □

Normed spaces and completions

In this appendix we have gathered a small amount of basic facts about normed vector spaces.

Definition B.1. Let X be a (real) vector space and $\|\cdot\|\colon X \to [0,\infty)$ a norm. Then the pair $(X, \|\cdot\|)$ is called a **normed space**. ◆

If X is a normed space, then $d(x,y) \coloneqq \|x-y\|$ defines a metric on X. We endow X with the topology induced by d, this topology is also called the **norm topology** or **strong topology** on X.

Definition B.2. A complete normed space is called a **Banach space**. ◆

The natural maps to study in this setting are the continuous linear maps. A basic result is that continuity is equivalent to *boundedness*.

Proposition B.3. *Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces and $T\colon X \to Y$ a linear map. Then T is continuous if and only if it is **bounded**, meaning that there exists a constant $C > 0$ such that*

$$\|Tx\|_Y \leq C\|x\|_X$$

for all $x \in X$.

Proof. If $\|Tx\|_Y \leq C\|x\|_X$ for all $x \in X$, then we have

$$\|Tx - Ty\|_Y = \|T(x-y)\| \leq C\|x-y\|_X$$

and T is Lipschitz and hence continuous.

Conversely assume that T is continuous and that there exists a sequence $(x_n)_{n=1}^{\infty}$ of nonzero elements of X such that $\|Tx_n\|_Y/\|x_n\|_X \to \infty$. By picking a subsequence we may actually assume that $\|Tx_n\|_Y/\|x_n\|_X \geq n$. Let us define $y_n = \frac{x_n}{n\|x_n\|_X}$. Then we have $\|y_n\|_X = 1/n$ and $\|Ty_n\|_Y \geq 1$, but this contradicts the continuity of T since now $y_n \to 0$ and hence also $Ty_n \to 0$. □

The smallest constant C for which the inequality $\|Tx\|_Y \leq C\|x\|_X$ holds is called the **norm** of T and denoted by $\|T\|$.

It is also helpful to note that linearity actually boosts the continuity to uniform continuity.

Lemma B.4. *A continuous linear map $T\colon X \to Y$ between two normed spaces is uniformly continuous.*

Proof. For any $x, y \in X$ we have $\|Tx - Ty\|_Y = \|T(x-y)\|_Y \leq \|T\|\|x-y\|_X$ so T has a global modulus of continuity $\varphi(s) = \|T\|s$ $(s \geq 0)$. □

Particularly important maps are the ones that preserve distances.

Definition B.5. A linear map $T\colon X \to Y$ between two normed spaces is an **isometry** if $\|Tx\|_Y = \|x\|_X$ for all $x \in X$. ◆

Clearly an isometry is automatically an injection and hence a linear embedding of X into Y. The isometric property ensures that also the norm structures in X and $T(X)$ agree, and hence $T(X)$ can be viewed as an isomorphic copy of the normed space X sitting inside Y.

Let us finally discuss completions of normed spaces.

Definition B.6. Let $(X, \|\cdot\|_X)$ be a normed space. Any pair $(\bar{X}, T)$ where $\bar{X}$ is a Banach space and $T\colon X \to \bar{X}$ is a linear isometry such that $T(X)$ is dense in $\bar{X}$ is called a **completion** of X. ◆

Given a completion $\bar{X}$ of X we usually view X as a subset of $\bar{X}$, in a similar manner as we view the rational numbers as a subset of the real numbers, even if the particular construction we used for the real numbers might actually not possess such an inclusion relation in a purely set theoretic sense.

The main theorem of this appendix is that completions exist. Before that, let us note that as in the case of metric spaces there is a nice way to characterise completions:

Proposition B.7. *A completion $(\hat{X}, \iota)$ of a normed space X satisfies and is characterised up to an isomorphism by the following universal property: If Y is any Banach space and $A\colon X \to Y$ is a continuous linear map, then A extends uniquely to a continuous linear map $\hat{A}\colon \hat{X} \to Y$ such that $A = \hat{A} \circ T$. Moreover, the norms of A and $\hat{A}$ are equal.*

Proof. Mimic the proof of Proposition A.12. □

We will end this appendix with a proof of the existence of completions.

Theorem B.8. *Any normed space X has a completion.*

Before the proof let us recall a bit of functional analysis.

Definition B.9. Let X be a normed space. The **(continuous) dual space** of X is the space

$$X^* \coloneqq \{\varphi\colon X \to \mathbb{R} : \varphi \text{ is linear and continuous}\}$$

endowed with the norm

$$\|\varphi\|_{X^*} = \sup_{x \in X, \|x\| \leq 1} |\varphi(x)|.$$ ◆

The norm in X^* is the same as defined above for continuous linear maps $T\colon X \to Y$ in the special case where $Y = \mathbb{R}$. It is simple to check that it indeed defines a norm on X^*.

We note that duals are always complete.

Proposition B.10. *Let X be a normed space. Then X^* is a Banach space.*

Proof. Suppose that $(\varphi_n)_{n=1}^{\infty}$ is a Cauchy sequence in X^*. Then for any fixed $x \in X$ the sequence $\varphi_n(x)$ is also Cauchy in $\mathbb{R}$ and converges to some limit which we denote $\varphi(x)$. It is easy to check that the map $\varphi\colon X \to \mathbb{R}$ is linear and satisfies $|\varphi(x)| \le \sup_{n\ge 1} \|\varphi_n\|_{X^*}\|x\|_X$ for any $x \in X$, so that $\varphi \in X^*$. Finally let $\varepsilon > 0$ and $x \in X$ with $\|x\| \le 1$ be arbitrary and choose $n \ge 1$ so large that $\|\varphi_n - \varphi_m\| < \varepsilon$ for all $m \ge n$. (Note that n does not depend on x.) Then

$$|\varphi_n(x) - \varphi(x)| \le \limsup_{m\to\infty}(\|\varphi_n - \varphi_m\| + |\varphi_m(x) - \varphi(x)|) < \varepsilon$$

and as x was arbitrary we see that $\|\varphi_n - \varphi\| < \varepsilon$. Since also ε was arbitrary we have $\varphi_n \to \varphi$ in X^*. □

The final ingredient we will need for the proof of Theorem B.8 is the Hahn–Banach theorem.

Theorem B.11 (Hahn–Banach)**.** *Let X be a normed space and $E \subset X$ be a vector subspace. Suppose that $f\colon E \to \mathbb{R}$ is a continuous linear map (w.r.t. the norm on X). Then there exists a continuous linear map $F\colon X \to \mathbb{R}$ with $F(x) = f(x)$ for all $x \in E$.*

We will skip the proof. In fact Hahn–Banach is only needed if one wants to prove the general version of Theorem B.8. When we actually use it in Section 2.4, it is only in the case where X is the set of simple random variables under the L^1-norm. We will indicate after the proof below how to handle this special case without using Hahn–Banach.

Proof of Theorem B.8. Let X^* be the dual of X and X^{**} be the dual of X^* (also known as the *bidual* of X). Let us define the map $\iota\colon X \to X^{**}$ by mapping x to the map $u_x\colon X^* \to \mathbb{R}$ given by $u_x(\varphi) := \varphi(x)$ for all $\varphi \in X^*$. The map ι is linear since for all $x, y \in X$ and $a, b \in \mathbb{R}$ we have $u_{ax+by}(\varphi) = \varphi(ax + by) = a\varphi(x) + b\varphi(y) = au_x(\varphi) + bu_y(\varphi)$ for all $\varphi \in X^*$ so that $\iota(ax + by) = u_{ax+by} = au_x + bu_y = a\iota(x) + b\iota(y)$.

Next we claim that ι is an isometry. Note first that we have the upper bound

$$\|\iota(x)\|_{X^{**}} = \|u_x\|_{X^{**}} = \sup_{\varphi\in X^*, \|\varphi\|_{X^*}\le 1} |u_x(\varphi)| \le \|x\|_X.$$

It remains to show that there exists some φ with $\|\varphi\| \le 1$ such that $|u_x(\varphi)| = |\varphi(x)| = \|x\|_X$. This is where Hahn–Banach will enter: We will simply extend

the linear functional $\varphi_0(tx) = t\|x\|_X$ defined on the subspace $\{tx : t \in \mathbb{R}\} \subset X$ to a continuous linear functional φ on whole X so that it becomes an element of X^*.

Finally we define $\hat{X}$ as the closure of $\iota(X)$ in X^{**}. As X^{**} is complete by Proposition B.10, also $\hat{X}$ as a closed subspace is complete and by definition $\iota(X)$ is dense in $\hat{X}$. □

In the special case where $X = S$ is the space of simple random variables with the L^1-norm, we can avoid the use of Hahn–Banach as follows: Note that any $y \in S$ gives an element φ_y in the dual of S by setting $\varphi_y(x) = \mathbb{E}[yx]$ for all $x \in S$. Indeed, this map is clearly linear and it is continuous since $|\mathbb{E}[yx]| \leq \|y\|_{L^\infty}\|x\|_{L^1}$. Thus in particular if we choose $y = \operatorname{sgn}(x)$ we get $\varphi_y(x) = \mathbb{E}[\operatorname{sgn}(x)x] = \mathbb{E}[|x|] = \|x\|_{L^1}$, showing that $u_x(\varphi_y) = \|x\|_X$ as was needed in the proof.

Radon–Nikodym theorem

In this appendix we will provide a proof of the Radon–Nikodym theorem. It is mainly based on [7].

Theorem C.1. *Let $(T, \mathcal{G}, \mu)$ be a probability space and assume that ν is another measure on T such that $\nu \ll \mu$. Then there exists a measurable function $f : T \to [0, \infty]$ such that*

$$\nu(A) = \int_A f \, d\mu$$

for all $A \in \mathcal{G}$.

Remark. For simplicity we will prove this theorem in the case where ν is also a probability measure. It is easy to show that the theorem is true when μ and ν are both σ-finite measures, and with some extra work one can get rid of this assumption for ν. ◆

Let us start with the following lemma.

Lemma C.2. *Let $\mathcal{A} \subset \mathcal{G}$ be a collection such that*

- *If $\mu(A) = 0$, then $A \in \mathcal{A}$.*
- *If $\mu(A) > 0$, then there exists $B \subset A$ with $\mu(B) > 0$ and $B \in \mathcal{A}$.*
- *$\mathcal{A}$ is closed under countable disjoint unions.*

Then $\mathcal{A} = \mathcal{G}$.

Proof. Let us fix $E \in \mathcal{G}$ and try to show that $E \in \mathcal{A}$. Consider all collections $(A_i)_{i \in I}$ of disjoint sets $A_i \subset E$ with $\mu(A_i) > 0$ and $A_i \in \mathcal{A}$. We can order such collections by inclusion, and by Zorn's lemma there exists a collection $(E_i)_{i \in I}$ which is maximal. Now the index set has to be countable since $\mu(E_i) > 0$ for all $i \in I$ and E has a finite measure. Thus $\tilde{E} \coloneqq \biguplus_{i \in I} E_i$ belongs to $\mathcal{A}$. We cannot have $\mu(E \setminus \tilde{E}) > 0$ since otherwise by assumption there would be some $E' \subset E \setminus \tilde{E}$ which we could add to the collection $(E_i)_{i \in I}$, contradicting the maximality. Since $E \setminus \tilde{E}$ has 0 measure, also $E = \tilde{E} \cup (E \setminus \tilde{E})$ belongs to $\mathcal{A}$. □

Proof of Theorem C.1. As indicated in the remark above, we will assume that ν is also a probability measure. Consider the following set of functions:

$$H \coloneqq \{f : T \to [0, \infty] : f \text{ measurable}, \int_E f \, d\mu \leq \nu(E) \text{ for all } E \in \mathcal{G}\}.$$

The idea is roughly to take the largest of all functions in H. One cannot simply take the pointwise supremum, though, since this could easily be the constant function ∞ if all the singletons of T have zero measure. We will thus go in a bit roundabout way and define instead the maximal total mass

$$M \coloneqq \sup\{\int_T f\,d\mu : f \in H\} \leq \nu(T) = 1.$$

Note that if $f, g \in H$, then also $f \vee g \in H$, since

$$\begin{aligned}\int_E (f \vee g)\,d\mu &= \int_{E\cap\{f\geq g\}} f\,d\mu + \int_{E\cap\{f<g\}} g\,d\mu \\ &\leq \nu(E\cap\{f\geq g\}) + \nu(E\cap\{f<g\}) = \nu(E).\end{aligned}$$

Thus there exists a pointwise increasing sequence $f_n \in H$ such that $\int_T f_n\,d\mu \to M$. Let us now define $f \coloneqq \sup_n f_n$ and claim that f is a Radon–Nikodym derivative of ν with respect to μ.

By the monotone convergence theorem it is clear that f satisfies $\int_E f\,d\mu \leq \nu(E)$, so it is enough to show the opposite inequality. Assume in contrary that there exists a set E and $\varepsilon > 0$ such that $\int_E f\,d\mu < \nu(E) - 2\varepsilon$. We claim that inside E there exists a subset $F \subset E$ with $\mu(F) > 0$ and such that $f + \varepsilon\mathbb{1}_F \in H$. If not, then for all $F \subset E$ of positive μ-measure there exists a set G such that $\int_G (f + \varepsilon\mathbb{1}_F)\,d\mu \geq \nu(G)$. We see then from

$$\nu(G\cap F) + \nu(G\setminus F) = \nu(G) \leq \int_G (f+\varepsilon\mathbb{1}_F)\,d\mu \leq \int_{G\cap F} (f+\varepsilon\mathbb{1}_{G\cap F})\,d\mu + \nu(G\setminus F)$$

that also the set $\tilde{G} = G\cap F \subset F$ satisfies that $\int_{\tilde{G}}(f+\varepsilon\mathbb{1}_{\tilde{G}}) \geq \nu(\tilde{G})$. Hence the collection

$$\mathcal{A} = \{G \subset E : \int_G (f+\varepsilon\mathbb{1}_G)\,d\mu \geq \nu(G)\}$$

satisfies the second bullet in Lemma C.2. It is also clear that $\mathcal{A}$ contains the sets of measure 0 inside E and is closed under countable unions. Thus in fact $\mathcal{A}$ contains all the measurable subsets of E, including E itself, but this is a contradiction since then

$$\int_E f\,d\mu + \varepsilon \geq \int_E (f+\varepsilon\mathbb{1}_E)\,d\mu \geq \nu(E) \geq \int_E f\,d\mu + 2\varepsilon.$$

Thus there must exist a set $F \subset E$ with $\mu(F) > 0$ and $f + \varepsilon\mathbb{1}_F \in H$, but this now contradicts the fact that $\int f\,d\mu = M$ is the supremum of total masses over H, since $\int (f+\varepsilon\mathbb{1}_F)\,d\mu \geq M + \varepsilon\mu(F)$. □

Bibliography

[1] R. Durrett. *Probability: Theory and Examples*. 5th ed. 2019.

[2] Olav Kallenberg. *Foundations of modern probability*. Springer, 2002.

[3] Kalle Kytölä. *Probability theory*. 2019. URL: https://math.aalto.fi/~kkytola/files_KK/ProbaTh2019/ProbaTh-2019.pdf.

[4] Dorothy Maharam. *From finite to countable additivity*. Portugaliae mathematica 44.3 (1987), 265–282.

[5] Edward Nelson. *Radically elementary probability theory*. Vol. 20. 1989, 240–243.

[6] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill, 2006.

[7] Anton R. Schep. *A newer addendum to "And still one more proof of the Radon–Nikodym theorem"* (2012). URL: https://people.math.sc.edu/schep/Radon-update-2.pdf.

[8] Jean Schmets. *Theorie de la mesure*. 2004.

[9] David G. Schwartz. *Roll the bones: The history of gambling*. Winchester Books, 2013.

Index