STAT 9657 — ADVANCED PROBABILITY

by Hristo Sendov

Contents

1 Background

1.1 General notation

Given a set Ω and subsets $A, B \subseteq \Omega$, the following notation is used

intersection:	$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$
union:	$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$
set-minus:	$A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$
symmetric difference:	$A \Delta B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ but } \omega \notin A \cap B\}$
complement:	$A^c = \{\omega \in \Omega : \omega \notin A\}$
empty set:	$\emptyset =$ the set without any element
real numbers:	$\mathbb{R}$
natural numbers:	$\mathbb{N} = \{1, 2, 3, \dots\}$
rational numbers:	$\mathbb{Q}$
complex numbers:	$\mathbb{C}$
positive infinity:	∞
negative infinity:	$-\infty$
indicator function:	$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$

Operations with infinities. For any $x \in \mathbb{R}$ we have

(i) $-\infty < x < \infty$;

(ii) $\infty + x = x + \infty = \infty$;

(iii) $-\infty + x = x + (-\infty) = -\infty$;

(iv) $\infty + \infty = \infty$; $-\infty + (-\infty) = -\infty$;

(v) $\infty \cdot x = x \cdot \infty = \infty$ if $x > 0$;

(vi) $(-\infty) \cdot x = x \cdot (-\infty) = -\infty$ if $x > 0$;

(vii) $\infty \cdot x = x \cdot \infty = -\infty$ if $x < 0$;

(viii) $(-\infty) \cdot x = x \cdot (-\infty) = \infty$ if $x < 0$;

(ix) $\infty \cdot \infty = \infty$; $(-\infty) \cdot (-\infty) = \infty$.

Note that the expressions $\infty + (-\infty)$, $-\infty + \infty$, $0 \cdot \infty$, $\infty \cdot 0$, $0 \cdot (-\infty)$, $(-\infty) \cdot 0$ are not defined and we will avoid them like the plague.

Given real numbers α and β, we use $\alpha \wedge \beta := \min\{\alpha, \beta\}$.

1.2 Bounded sets, inf, and sup

A set $\Omega \subset \mathbb{R}$ is called bounded from above (resp. below) if there is a constant M such that $\omega \leq M$ (resp. $M \leq \omega$) for every $\omega \in \Omega$. The constant M is called an upper bound for Ω. Clearly, if M is an upper bound for Ω and if $M' \geq M$, then M' is also an upper bound for Ω. Thus, a set bounded from above has many upper bounds. If Ω is bounded from both above and below we say it is bounded, equivalently, there is a constant M such that $|\omega| \leq M$ for every $\omega \in \Omega$.

Theorem 1. If $\Omega \subset \mathbb{R}$ is bounded from above (resp. below) then there is a smallest upper (resp. largest lower) bound. That is, an upper (resp. lower) bound that is smaller (resp. larger) than or equal to any other upper (resp. lower) bound.

The smallest upper bound of a set Ω, bounded from above, is denoted by $\sup\Omega$ and read supremum of Ω. The largest lower bound of a set Ω bounded from below is denoted by $\inf\Omega$ and read infimum of Ω. In other words, if M is an upper bound for Ω then $\sup\Omega \leq M$, and similarly, if M is a lower bound for Ω, then $M \leq \inf\Omega$. For example, if $\Omega = [0,1]$ then $\inf\Omega = 0$ and $\sup\Omega = 1$. If $\Omega = (0,1)$ then $\inf\Omega = 0$ and $\sup\Omega = 1$, again. If $\Omega = \{\text{all rational numbers in } (0,1)\}$ then $\inf\Omega = 0$ and $\sup\Omega = 1$. Thus, the numbers $\inf\Omega$ and $\sup\Omega$ may or may not be elements of Ω. We can extend the definition of inf and sup to unbounded sets. If Ω is not bounded from above then we define $\sup\Omega = \infty$ and if Ω is not bounded from below, we define $\inf\Omega = -\infty$. Finally, if $\Omega = \emptyset$ then any real number M is an upper bound (as well as a lower bound), thus

$$\sup\emptyset = -\infty \text{ and } \inf\emptyset = \infty.$$

If Ω is not empty then $\inf\Omega \leq \sup\Omega$. Note that if A and B are two *non-empty* subsets of $\mathbb{R}$ and

$$\text{if } A \subseteq B, \text{ then } \inf B \leq \inf A \leq \sup A \leq \sup B. \tag{1}$$

1.3 Finite and infinite sets

A set is *finite*, well, if it has finitely many elements, otherwise it is called infinite.

Lemma 2. Suppose $\Omega = \{\omega_1, \dots, \omega_n\}$ is a set with n elements. There are exactly 2^n different subsets (or events) of Ω.

Proof. A subset A of Ω can be specified by stating exactly which elements of Ω are in A and which are not. Consider a $0,1$-vector $(x_1, \dots, x_n)$ with n coordinates, that is $x_i \in \{0,1\}$ for every $i = 1, 2, \dots, n$. Every such vector describes a subset A of Ω. Indeed, we define ω_i to be in A if $x_i = 1$ and ω_i not to be in A if $x_i = 0$. Conversely for any subset A of Ω there is a $0,1$-vector that describes A in the above way. Since there are 2^n different $0,1$ vectors with n coordinates, there are 2^n subsets of Ω. □

The set of all subsets of Ω will be denoted by 2^Ω. For example, if $\Omega = \{a,b,c\}$ has three elements then

$$2^\Omega = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}$$

has $2^3 = 8$ elements. Each element of 2^Ω is a subset of Ω.

In the case when Ω has infinitely many elements, it has infinitely many subsets. Unfortunately, some infinities are larger than other infinities. A set with infinitely many elements is called *countably infinite* if its elements can be ordered in a sequence.

Example 3. The natural numbers form a countably infinite set since we can order them in a sequence $1, 2, 3, \ldots$

Example 4. The integers (positive and negative) $\ldots - 3, -2, -1, 0, 1, 2, 3, \ldots$ are also countably infinite because we can order them as

$$0, 1, -1, 2, -2, 3, -3, \ldots$$

Example 5. The rational numbers (those that can be written as a ratio of two integers, say $1/2$ or $345/45$, or $-3/4$) are also countably infinite. Note that the rational numbers are dense in the sense that every interval (a, b), no matter how small or large contains a rational number. A priori, by looking at the real number line one cannot tell what is the next rational number after, say $1/2$. Their placement on the real number line does not show immediately how to order them in a sequence. Here is how one can order them in a sequence. We will do that only for the positive rational numbers for added simplicity.

	1	2	3	4	5	$\cdots$
1	1/1	1/2	1/3	1/4	1/5	$\cdots$
2	2/1	2/2	2/3	2/4	2/5	$\cdots$
3	3/1	3/2	3/3	3/4	3/5	$\cdots$
4	4/1	4/2	4/3	4/4	4/5	$\cdots$
5	5/1	5/2	5/3	5/4	5/5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We look at the diagonals in this table that go from north-east to south-west and list the numbers in them one diagonal after another:

$$1/1, 1/2, 2/1, 1/3, 2/2, 3/1, 1/4, 2/3, 3/2, 4/1, 1/5, 2/4, 3/3, 4/2, 5/1, \ldots$$

Many numbers in this sequence are repeated, for example $1/1 = 2/2 = 3/3 = 1$ or $1/2 = 2/4$. We delete all repetitions leaving only the first instance of a repeated number to obtain

$$1/1, 1/2, 2/1, 1/3, 3/1, 1/4, 2/3, 3/2, 4/1, 1/5, 5/1, \ldots$$

We obtained a sequence of all positive rational numbers, which is what we wanted.

Lemma 6. Union of countably many sets each one of which has countably many elements is countably infinite.

Proof. We have countably many sets, that is we can order them in a sequence $A_1, A_2, A_3, \ldots$ Each set A_i has countably many elements, say $A_1 = \{a_1, a_2, a_3, \ldots\}$, $A_2 = \{b_1, b_2, b_3, \ldots\}$, $A_3 = \{c_1, c_2, c_3, \ldots\}$, and so on. We need to show that we can order the elements of $\bigcup_{i=1}^{\infty} A_i$ in a sequence as well. We use an idea analogous to the one presented above. Place the elements of the sets A_i in rows one after another

	1	2	3	4	5	$\cdots$
A_1	a_1	a_2	a_3	a_4	a_5	$\cdots$
A_2	b_1	b_2	b_3	b_4	b_5	$\cdots$
A_3	c_1	c_2	c_3	c_4	c_5	$\cdots$
A_5	d_1	d_2	d_3	d_4	d_5	$\cdots$
A_6	e_1	e_2	e_3	e_4	e_5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We look at the diagonals in this table that go from north-east to south-west and list the numbers in them one diagonal after another:

$$a_1, a_2, b_1, a_3, b_2, c_1, a_4, b_3, c_2, d_1, \dots$$

This sequence contains the all elements in the union $\bigcup_{i=1}^{\infty} A_i$. □

Definition 7. A random variable X defined on a sample space Ω is called *discrete* if X takes a finitely many or countably many different values. If the random variable X takes more than countably many different values, for example, if it takes any value in an interval (a, b), then X is called *continuous* random variable.

Every real number r can be represented in a decimal form as $r = a_0.a_1a_2a_3\dots$, where a_0 is an integer and $a_1, a_2, a_3, \dots$ are digits between 0 and 9. What this says in reality is that

$$r = \sum_{i=0}^{\infty} \frac{a_i}{10^i}.$$

It can be shown that a number r is rational if and only if its decimal representation becomes periodic from some point on. For example, $1/2 = 0.5000\dots$, $1/3 = 0.3333\dots$, and $1/13 = 0.07692307692307692\dots$ are periodic decimal representations the first with period (0), the second with period (3), and the third with period (076923). Some rational numbers have two different decimal representations. These are the rational numbers that have finite decimal representation $r = a_0.a_1a_2a_3\dots,$, that is there is an index $N \in \mathbb{N}$, such that $a_n = 0$ for all $n > N$. The two representations are $r = a_0.a_1a_2\dots a_{n-1}a_n000\dots$ and $r = a_0.a_1a_2\dots a_{n-1}(a_n - 1)999\dots$. Such is $1/2$ with two representations $0.5000\dots$ and $0.4999\dots$. In general, a rational number has finite decimal representation if and only if it is of the form $m/10^n$ for some integers n and m, e.g. $1/2 = 5/10$. Two real numbers $a_0.a_1a_2a_3, \dots$ and $b_0.b_1b_2b_3, \dots$ are *equal* if $a_k = b_k$ for all $k = 0, 1, \dots$ or if they are the two different representations of a rational number with finite decimal representation.

Example 8. The real numbers are infinitely many and are more than countable. That is, the real numbers cannot be ordered in a sequence. Let us see that the real numbers in the interval $(0, 1)$ cannot be ordered in a sequence. Consider the decimal representation of the numbers in $(0, 1)$. Suppose the real numbers in $(0, 1)$ can be listed in a sequence

$$\begin{aligned} x_1 &= 0.a_{11}a_{12}a_{13}a_{14}\dots \\ x_2 &= 0.a_{21}a_{22}a_{23}a_{24}\dots \\ x_3 &= 0.a_{31}a_{32}a_{33}a_{34}\dots \end{aligned}$$

$$x_4 = 0.a_{41}a_{42}a_{43}a_{44}\dots$$
$$\vdots$$

Consider a number $r := 0.b_1b_2b_3\dots$ constructed in such a way that b_1 is different from a_{11}, 0, and 9, b_2 is different from a_{22}, 0, and 9, b_3 is different from a_{33}, 0 and 9, and so on. The number r does not contain 0's and 9's after its decimal point. That is, it is not a rational number that has two decimal representations. Hence, r has a unique decimal representation and it is different from all numbers $x_1, x_2, x_3, \dots$ in the sequence, since it differs from the first number in the list by its first digit after the decimal point; it differs from the second number in the list by its second digit, and so on. This contradiction shows that our assumption that the real numbers in $(0, 1)$ can be ordered in a sequence is wrong. (Note that when constructing the number $0.b_1b_2b_3\dots$ we had 7 choices for the digit b_1, 7 choices for b_2, and so on.)

Example 9 (The Cantor set)**.** Start with the interval $C_0 := [0, 1]$ and remove the middle third $(1/3, 2/3)$. The remaining set has now two components $C_1 := [0, 1/3] \cup [2/3, 1]$. From each of those components remove their middle thirds $(1/9, 2/9)$ and $(7/9, 8/9)$. The remaining set is $C_2 := [0, 1/9] \cup [2/9, 1/3] \cup [6/9, 7/9] \cup [8/9, 1]$. Again remove the middle third from each component to obtain

$$C_3 := [0, 1/27] \cup [2/27, 3/27] \cup [6/27, 7/27] \cup [8/27, 9/27] \\ \cup [18/27, 19/27] \cup [20/27, 21/27] \cup [24/27, 25/27] \cup [26/27, 1].$$

Continue in this was indefinitely at every step removing the middle third of each component. We obtain a decreasing sequence of closed sets $C_0 \supset C_1 \supset C_2 \supset C_3 \cdots$. The *Cantor (ternary) set* is defined as the intersection $\cap_{i=0}^{\infty} C_i$.

Being an intersection of closed sets the Cantor set is closed and non-empty, for example 0 and 1 are in it. The "length" of C is

$$1 - \frac{1}{3} - \frac{2}{9} - \frac{4}{27} - \cdots = 1 - \frac{1}{3}\Big(1 + \frac{2}{3} + \Big(\frac{2}{3}\Big)^2 + \cdots\Big) = 1 - \frac{1}{3}\frac{1}{1 - 2/3} = 0.$$

Every number $x \in [0, 1]$ has a ternary expansion of the form

$$x = \sum_{i=1}^{\infty} \frac{x_i}{3^i} \quad \text{with } x_i \in \{0, 1, 2\}. \tag{2}$$

That is, we represent each $x \in [0, 1]$ as a sequence $0.x_1x_2x_3\dots$ where $x_i \in \{0, 1, 2\}$. Conversely, given such a sequence, to obtain the actual number $x \in [0, 1]$ one needs to calculate the sum (2). This expansion is not unique for the numbers in $[0, 1]$ of the form $m/3^n$. For example, $1/3$ admits two expansions, namely $0.1000\dots$ and $.0222\dots$. Non-uniqueness occurs only for those x that admit an expansion ending with an infinite sequence of 0's. We claim (without a proof) that the numbers in the Cantor set have a ternary expansion $0.x_1x_2x_3\dots$ in which $x_n \neq 1$ for all n. If $x \in [0, 1]$ has two ternary expansions, then it is in the Cantor set if *one* of the expansions has no term equal to 1. For example, 0.1 is in the Cantor set since it can be represented also as $0.0222\dots$ and this representation does not involve 1. Numbers such as $0.12, 0.11, 0.0101$ are not in the Cantor set. To get used to these ideas we make the following observations.

Sequences $0.1x_2x_3\ldots = 1/3 + x_2/9 + x_3/27 + \cdots$ correspond to the numbers in $[1/3, 2/3]$.
Sequences $0.01x_3\ldots = 0/3 + 1/9 + x_3/27 + \cdots$ correspond to the numbers in $[1/9, 2/9]$.
Sequences $0.21x_3\ldots = 2/3 + 1/9 + x_3/27 + \cdots$ correspond to the numbers in $[7/9, 8/9]$.

Clearly, the Cantor set has infinitely many numbers in it. But the amazing fact is that modifying the argument given in Example 8, one can see that the Cantor set C is not countable. An uncountable set with length 0!

1.4 Sequences and limits

1.4.1 Sequences

Let $a_1, a_2, a_3, \ldots$ be a sequence of real numbers, denoted for short by $\{a_n\}_{n=1}^{\infty}$, or just $\{a_n\}$. The sequence $\{a_n\}$ is called increasing if $a_n \leq a_{n+1}$ for all $n = 1, 2, 3, \ldots$ It is called decreasing if $a_n \geq a_{n+1}$ for all $n = 1, 2, 3, \ldots$ It is called *monotone* if it is either increasing or decreasing. A sequence $a_1, a_2, a_3, \ldots$ is bounded (resp. above, below) if the set $\Omega := \{a_1, a_2, , a_3, \ldots\}$ is such. The number ℓ is called a limit point of the sequence $\{a_n\}$ if for every $\epsilon > 0$ the interval $(\ell - \epsilon, \ell + \epsilon)$ contains infinitely many elements of the sequence. Formally: ℓ is a limit point of the sequence $\{a_n\}$ if for every $\epsilon > 0$ and every $N \in \mathbb{N}$ there is an index $n \geq N$ such that $a_n \in (\ell - \epsilon, \ell + \epsilon)$.

We need to extend the notion of a limit of a sequence to include the symbols ∞ and $-\infty$. The symbol ∞ is a limit point of the sequence $\{a_n\}$ if it is unbounded from above. (Analogously, the symbol $-\infty$ is a limit point of the sequence $\{a_n\}$ if it is unbounded from below.) Formally, the symbol ∞ is a limit point of the sequence $\{a_n\}$ if for every $M \in \mathbb{R}$ and every $N \in \mathbb{N}$ there is an index $n \geq N$ such that $a_n \geq M$. (Analogously, the symbol $-\infty$ is a limit point of the sequence $\{a_n\}$ if for every $M \in \mathbb{R}$ and every $N \in \mathbb{N}$ there is an index $n \geq N$ such that $a_n \leq M$.

With this extension we have the following theorem.

Theorem 10. Sequence $\{a_n\}$ always has a limit point in $\mathbb{R} \cup \{-\infty, \infty\}$. In particular, if a sequence $\{a_n\}$ is bounded, then it has a limit point $\ell \in \mathbb{R}$.

Note that a sequence $\{a_n\}$ may have many limit points.

Definition 11. The sequence $\{a_n\}$ is *convergent* if it has exactly one limit point. That limit point is denoted by $\lim\limits_{n\to\infty} a_n$.

Note that the notation $\lim\limits_{n\to\infty} a_n$ doesn't make sense if the sequence $\{a_n\}$ is not convergent.

Formally, the sequence $\{a_n\}$ converges to $\ell \in \mathbb{R}$ if for every $\epsilon > 0$ the interval $(\ell - \epsilon, \ell + \epsilon)$ contains all but finitely many elements of the sequence. This implies that if a sequence $\{a_n\}$ converges to a limit $\ell \in \mathbb{R}$ then the set $\{a_n\}$ is bounded.

The sequence $\{a_n\}$ converges to ∞ if for every M, the interval (M, ∞) contains all but finitely many elements of the sequence. The sequence $\{a_n\}$ converges to $-\infty$ if for every M, the interval $(-\infty, M)$ contains all but finitely many elements of the sequence.

Theorem 12. The sequence $\{a_n\}$ has a limit point ℓ if and only if there is a subsequence $\{a_{n_i}\}_{i=1}^{\infty}$ converging to ℓ.

Combining Theorems 10 and 12, we see that every sequence $\{a_n\}$ has a converging (possibly to infinity) subsequence.

Theorem 13. If the sequence $\{a_n\}$ is increasing (resp. decreasing) then it is convergent to a limit equal to $\sup\{a_1, a_2, a_3, \ldots\}$ (resp. $\inf\{a_1, a_2, a_3, \ldots\}$).

Note that if a sequence $\{a_n\}$ is increasing (resp. decreasing) and bounded from above (resp. below) then its limit is in $\mathbb{R}$, that is, it cannot be ∞ (resp. $-\infty$).

Often it is important to be able to tell if a sequence $\{a_n\}$ is convergent or not (without explicitly knowing what its limit point might be). For this reason we define another sequence $b_n := \sup_{k,p\geq 0} |a_{n+k} - a_{n+p}|$ for all $n = 1, 2, \ldots$ Note that the sequence $\{b_n\}$ is decreasing $b_n \geq b_{n+1}$ and bounded from below $b_n \geq 0$. Hence $\{b_n\}$ is convergent to a (non-negative) limit point. This limit point may be strictly positive or zero. We have the following criterion.

Theorem 14 (Cauchy)**.** The sequence $\{a_n\}$ converges to a finite number if and only if $\lim\limits_{n\to\infty} b_n = 0$.

Exercise 15. *Give an example of a sequence $\{a_n\}$ for which $b_n = \infty$ for all n.*

1.4.2 Series

Given a sequence $\{a_n\}$ when does it make sense to sum all its elements? That is, what does it mean to write $\sum_{i=1}^{\infty} a_i$? Let

$$s_n := \sum_{i=1}^{n} a_i$$

denote the sum of the first n elements of the sequence $\{a_n\}$. The sum s_n is also called the *n-th partial sum of* $\{a_n\}$. If the sequence $\{s_n\}$ has a unique limit, that is, if it is convergent, then we define

$$\sum_{i=1}^{\infty} a_i := \lim_{n\to\infty} s_n.$$

According to our definition of a limit point, we allow $\sum_{i=1}^{\infty} a_i \in \mathbb{R} \cup \{-\infty, \infty\}$. From now on, whenever we write $\sum_{i=1}^{\infty} a_i$ we will understand that the sequence $\{s_n\}$ of partial sums is convergent to the number $\sum_{i=1}^{\infty} a_i$. The next lemma says that the tail of a series converges to 0 if the sum of the series is a finite number.

Lemma 16. If $\sum_{i=1}^{\infty} a_i \in \mathbb{R}$, then $\lim_{n\to\infty} \sum_{i=n}^{\infty} a_i = 0$.

Proof. Let $s := \sum_{i=1}^{\infty} a_i$ and let $t_n := \sum_{i=n}^{\infty} a_i$, then $s = s_n + t_{n+1}$. Since s_n converges to s as n approaches infinity, we must have that t_n approaches 0. □

1.4.3 Limit superior and limit inferior

Let $\{a_n\}$ be an arbitrary sequence and define the sequence $A_k := \sup\{a_k, a_{k+1}, a_{k+2}, \ldots\}$ for $k = 1, 2, 3, \ldots$ Since $\{a_{k+1}, a_{k+2}, \ldots\} \subseteq \{a_k, a_{k+1}, a_{k+2}, \ldots\}$ by (1) we obtain that $A_k \geq A_{k+1}$. That is, the sequence $\{A_k\}$ is decreasing, hence by Theorem 13 it has a limit (possibly infinite). This limit is denoted by anyone of the following notations

$$\limsup_{n\to\infty} a_n := \overline{\lim_{n\to\infty}}\, a_n := \lim_{n\to\infty} \sup_{k\geq n} a_k := \lim_{k\to\infty} \sup\{a_k, a_{k+1}, a_{k+2}, \ldots\} := \lim_{k\to\infty} A_k.$$

Analogously, let $\{a_n\}$ be an arbitrary sequence and define the sequence $B_k := \inf\{a_k, a_{k+1}, a_{k+2}, \ldots\}$ for $k = 1, 2, 3, \ldots$ Since $\{a_{k+1}, a_{k+2}, \ldots\} \subseteq \{a_k, a_{k+1}, a_{k+2}, \ldots\}$ by (1) we obtain that $B_k \leq B_{k+1}$. That is, the sequence $\{B_k\}$ is increasing, hence by Theorem 13 it has a limit (possibly infinite). This limit is denoted by anyone of the following notations

$$\liminf_{n\to\infty} a_n := \underline{\lim_{n\to\infty}}\, a_n := \lim_{n\to\infty} \inf_{k\geq n} a_k := \lim_{k\to\infty} \inf\{a_k, a_{k+1}, a_{k+2}, \ldots\} := \lim_{k\to\infty} B_k.$$

Cleraly from the definition, we have $B_k \leq A_k$ for all $k = 1, 2, \ldots$ Since $\{A_k\}$ and $\{B_k\}$ are convergent sequences we can take limits from both sides and the inequality is preserved in the limit:

$$\liminf_{n\to\infty} a_n = \lim_{k\to\infty} B_k \leq \lim_{k\to\infty} A_k = \limsup_{n\to\infty} a_n.$$

Note that the notation $\liminf\limits_{n\to\infty} a_n$ and $\limsup\limits_{n\to\infty} a_n$ always makes sense, that is limsup and liminf of a sequence $\{a_n\}$ always exist (but may be ∞ or $-\infty$).

Theorem 17. Let $\{a_n\}$ be an arbitrary sequence and let L denote the set of all its limit points. (Note that by Theorem 10, the set L is not empty.) Then $\limsup\limits_{n\to\infty} a_n = \sup L$ and $\liminf\limits_{n\to\infty} a_n = \inf L$.

In addition it can be shown that the set L of all limit points of a sequence is always a closed set and as a consequence contains the values $\sup L$ and $\inf L$. Thus, one can view $\limsup_{n\to\infty} a_n$ as the largest limit point of the sequence $\{a_n\}$ and $\liminf_{n\to\infty} a_n$ as its smallest limit point. That is, all limit points of $\{a_n\}$ are in the interval $[\liminf_{n\to\infty} a_n, \limsup_{n\to\infty} a_n]$. Suppose, $\ell := \limsup_{n\to\infty} a_n$ is a finite number, that is $\ell \in \mathbb{R}$. For any $\epsilon > 0$, the sequence $\{a_n\}$ does not have a limit point that is bigger than $\ell + \epsilon$. Hence by Theorem 10, only finitely many elements of $\{a_n\}$ are bigger than $\ell + \epsilon$ (indeed, if infinitely many elements of $\{a_n\}$ are bigger than $\ell + \epsilon$ than that subsequence will have a limit point bigger than $\ell + \epsilon$, contradicting the fact that ℓ is the biggest limit point). Similarly, if $\ell := \liminf_{n\to\infty} a_n$ is finite number, then only finitely many elements of $\{a_n\}$ are smaller than $\ell - \epsilon$.

We have the following corollary.

Corollary 18. A sequence $\{a_n\}$ is convergent if and only if $\limsup\limits_{n\to\infty} a_n = \liminf\limits_{n\to\infty} a_n$ and in that case

$$\lim_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n.$$

Thus, the quantities $\limsup_{n\to\infty} a_n$ and $\liminf_{n\to\infty} a_n$, that, as mentioned above always exist, give us a convenient way to check if a sequence is convergent or not.

Proposition 19. Let $\Omega \subset \mathbb{R}$ be a nonempty set and let $\ell := \sup \Omega$. Then there is an increasing sequence $\{a_n\}$ with elements from Ω converging to ℓ.

Proof. If $\ell = \infty$, then the set Ω is unbounded, that is for every $N \in \mathbb{N}$ there is an element a_n from Ω bigger than N. Choose, a_1 such that $a_1 > 1$. Choose a_2 such that $a_2 > \max\{2, a_1\}$, and continue like that inductively. Once $a_1, \ldots, a_{n-1}$ have been chosen, choose a_n such that $a_n > \max\{n, a_1, a_2, \ldots, a_{n-1}\}$. Clearly, the sequence $\{a_n\}$ is increasing and converges to ∞.

Suppose now $\ell < \infty$. If $\ell \in \Omega$, then we are done since the sequence $\{a_n\}$, where every $a_n := \ell$ is increasing and converging to ℓ. So, suppose in addition that $\ell \notin \Omega$. Recall that ℓ is the smallest upper bound of Ω. That is, for every $N \in \mathbb{N}$ the number $\ell - 1/N$ is not an upper bound of Ω. This means that there is an element a_n from Ω such that $\ell - 1/N \leq a_n < \ell$. When $N = 1$, choose a_1 such that $\ell - 1 \leq a_1 < \ell$. Since $\max\{a_1, \ell - 1/2\} < \ell$, there is an a_2 such that $\max\{a_1, \ell - 1/2\} \leq a_2 < \ell$. Once $a_1, \ldots, a_{n-1}$ have been chosen (all are less than ℓ), choose a_n such that $\max\{a_1, a_2, \ldots, a_{n-1}, \ell - 1/n\} \leq a_n < \ell$. Clearly, the sequence $\{a_n\}$ is increasing and converges to ℓ. □

Example 20. *The sequence* $\{a_n\} = \{+1, -1, +1, -1, +1, -1, \ldots\}$ *is not convergent but has two limit points* $+1$ *and* -1*. In addition,* $\operatorname{liminf}_{n\to\infty} a_n = -1$ *and* $\operatorname{limsup}_{n\to\infty} a_n = +1$. □

Exercise 21. *Consider the sequence* $\{a_n\}$ *on* $[0,1]$ *defined as follows:* $a_1 = 0, a_2 = 1, a_3 = 1/2, a_4 = 1/4, a_5 = 3/4, a_6 = 1/8, a_7 = 3/8, a_8 = 5/8, a_9 = 7/8, \ldots$ *What are the limit points of* $\{a_n\}$*, limsup and liminf?*

Exercise 22. *Consider the sequence* $\{a_n\}$ *on* $[0,1]$ *defined as follows:* $a_1 = 0, a_2 = 1, a_3 = 0, a_4 = 1/2, a_5 = 1, a_6 = 0, a_7 = 1/4, a_8 = 2/4, a_9 = 3/4, a_{10} = 1, a_{11} = 0, a_{12} = 1/8, a_{13} = 2/8, a_{14} = 3/8, a_{15} = 4/8, a_{16} = 5/8, a_{17} = 6/8, a_{18} = 7/8, a_{19} = 1, \ldots$ *What are the limit points of* $\{a_n\}$*, limsup and liminf?*

1.5 Properties of limsup and liminf

Let $\{a_n\}$ be a sequence.

- Let $\ell := \operatorname*{limsup}_{n\to\infty} a_n$. Since ℓ is one of the limit points of the sequence $\{a_n\}$, then there is a subsequence $\{a_{n_i}\}_{i=1}^{\infty}$ converging to ℓ. The situation is analogous for liminf.

- $\operatorname*{limsup}_{n\to\infty} ca_n = c(\operatorname*{limsup}_{n\to\infty} a_n)$ for any constant $c \geq 0$.

- $\operatorname*{limsup}_{n\to\infty}(-a_n) = -\operatorname*{liminf}_{n\to\infty} a_n$.

Let $\{b_n\}$ be another sequence.

- $\operatorname*{limsup}_{n\to\infty}(a_n + b_n) \leq \operatorname*{limsup}_{n\to\infty} a_n + \operatorname*{limsup}_{n\to\infty} b_n$, whenever the right-hand side is not $\infty - \infty$ or $-\infty + \infty$. Because of this property we say that limit superior is subadditive. If one of the sequences, say $\{a_n\}$, converges to a limit a, then the inequality becomes equality and we can replace $\operatorname*{limsup}_{n\to\infty} a_n$ by $\lim_{n\to\infty} a_n = a$.

Exercise 23. *Using the above properties derive the superadditivity of limit inferior:*

$$\liminf_{n\to\infty}(a_n + b_n) \geq \liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n.$$

Write similar relationships for

$$\limsup_{n\to\infty}(a_n - b_n) \text{ and } \liminf_{n\to\infty}(a_n - b_n).$$

1.6 Limits of functions

Let $f : \mathbb{R} \to \mathbb{R}$ be a function.

We say that $\ell \in \mathbb{R} \cup \{-\infty, \infty\}$ is *a limit point* of $f(x)$ as x approaches $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ if *there exists a* sequence $\{x_n\}_{n=1}^{\infty}$ converging to x_0, (with values different from x_0) such that the sequence of function values $\{f(x_n)\}_{n=1}^{\infty}$ converges to ℓ. For short, we say that ℓ is a limit point of $f(x)$ at x_0.

For example, any number in $[-1, 1]$ is a limit point of $\cos(x)$ as x approaches infinity.

If $f(x)$ has exactly one limit point, say ℓ, as x approaches x_0, then we say that ℓ is the limit of $f(x)$ at x_0. Formally, this means that *for every* sequence $\{x_n\}_{n=1}^{\infty}$ converging to x_0, the sequence of function values $\{f(x_n)\}_{n=1}^{\infty}$ converges to ℓ. We denote this by

$$\lim_{x\to x_0} f(x) = \ell.$$

Let L be the set of all limit points of $f(x)$ at x_0. (It is a fact that L is a closed set.) We define

$$\limsup_{x\to x_0} f(x) := \sup L \text{ and } \liminf_{x\to x_0} f(x) := \inf L.$$

So, $\limsup\limits_{x\to x_0} f(x)$ is the largest limit point of $f(x)$ at x_0 and $\liminf\limits_{x\to x_0} f(x)$ is the smallest limit point of $f(x)$ at x_0. By the definition of a limit point, we get the following useful property

- Suppose $\limsup\limits_{x\to x_0} f(x) = \ell$. Then there is a sequence $\{x_n\}$ converging to x_0, such that $\lim\limits_{n\to\infty} f(x_n) = \ell$. Analogously for liminf.

For example

$$\limsup_{x\to\infty} \cos(x) = 1, \qquad \liminf_{x\to\infty} \cos(x) = -1,$$
$$\limsup_{x\to\infty} x\cos(x) = \infty, \qquad \liminf_{x\to\infty} x\cos(x) = -\infty.$$

On the other hand

$$\liminf_{x\to\infty} \log(x) = \infty, \qquad \limsup_{x\to\infty}(-\log(x)) = -\infty.$$

We say that $\ell \in \mathbb{R} \cup \{-\infty, \infty\}$ is the limit of $f(x)$ as x approaches $x_0 \in \mathbb{R} \cup \{\infty\}$ *from the left* if *for every* sequence $\{x_n\}_{n=1}^{\infty}$ converging to x_0 *with smaller values* (that is, $x_n < x_0$ for all n), the sequence of function values $\{f(x_n)\}_{n=1}^{\infty}$ converges to ℓ. We denote this by

$$\lim_{x \to x_0^-} f(x) = \ell.$$

For short, we say that ℓ is the *left limit of* $f(x)$ *at* x_0. If $x_0 = \infty$ then, the left limit of $f(x)$ at ∞ is just its limit there (if it exists).

We say that $\ell \in \mathbb{R} \cup \{-\infty, \infty\}$ is the limit of $f(x)$ as x approaches $x_0 \in \mathbb{R} \cup \{-\infty\}$ *from the right* if *for every* sequence $\{x_n\}_{n=1}^{\infty}$ converging to x_0 *with bigger values* (that is, $x_n > x_0$ for all n), the sequence of function values $\{f(x_n)\}_{n=1}^{\infty}$ converges to ℓ. We denote this by

$$\lim_{x \to x_0^+} f(x) = \ell.$$

For short, we say that ℓ is the *right limit of* $f(x)$ *at* x_0. If $x_0 = -\infty$ then, the right limit of $f(x)$ at $-\infty$ is just its limit there (if it exists).

One can verify that $f(x)$ has a limit at x_0 if and only if

$$\lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^+} f(x).$$

We say that the function $f(x)$ is continuous at x_0 if it has a limit at x_0 and that limit is $f(x_0)$. Note that a function $f(x)$ may have a limit ℓ at x_0 and still not be continuous at x_0. This happens when the limit ℓ at x_0 is not equal to $f(x_0)$.

Functions that always have a left and a right limits are the monotone functions. The function $f(x)$ is *increasing* on (a, b) if $f(x) \leq f(y)$ for all $x \leq y$ in (a, b). The function $f(x)$ is *decreasing* if $f(x) \geq f(y)$ for all $x \leq y$ in (a, b). If it is either increasing or decreasing on (a, b) we say it is *monotone*. If the function $f(x)$ is monotone on (a, b), here $-\infty \leq a < b \leq \infty$, and $x_0 \in (a, b)$ then it has a left and a right limit at x_0. In addition, $f(x)$ has a right limit at a and a left limit at b.

For example, suppose the function $f(x)$ is increasing on (a, b). Then, its right limit at $x_0 = a$ is $\inf\{f(x) : x \in (a, b)\}$ and its left limit at $x_0 = b$ is $\sup\{f(x) : x \in (a, b)\}$. Analogously if the function $f(x)$ is decreasing on (a, b).

The following is a list of the most important properties of limsup and liminf.

- $\liminf\limits_{x \to \infty} f(x) \leq \limsup\limits_{x \to \infty} f(x)$ and equality holds if and only if $f(x)$ has a limit as x approaches infinity. In that case $\lim\limits_{x \to \infty} f(x) = \liminf\limits_{x \to \infty} f(x) = \limsup\limits_{x \to \infty} f(x)$.

- If $f(x) \leq g(x)$ for all $x \in (a, \infty)$ then

$$\limsup_{x \to \infty} f(x) \leq \limsup_{x \to \infty} g(x), \quad \text{and}$$
$$\liminf_{x \to \infty} f(x) \leq \liminf_{x \to \infty} g(x).$$

- $\limsup_{x\to\infty}(cf(x)) = c(\limsup_{x\to\infty} f(x))$, whenever $c > 0$.

- $\limsup_{x\to\infty}(-f(x)) = -\liminf_{x\to\infty} f(x)$.

- Whenever the righthand side is not $\infty - \infty$ or $-\infty + \infty$, we have

  $$\limsup_{x\to\infty}(f(x) + g(x)) \le \limsup_{x\to\infty} f(x) + \limsup_{x\to\infty} g(x).$$

  If $g(x)$ has a limit as x approaches ∞, that is, $\lim_{x\to\infty} g(x) =: \ell$ then the inequality in the previous property becomes equality

  $$\limsup_{x\to\infty}(f(x) + g(x)) = \limsup_{x\to\infty} f(x) + \ell.$$

We conclude with another way to describe liminf and limsup of a function $f(x)$ as x approaches infnity. Let $f(x)$ be arbitrary function defined on (a, ∞). Define a new function $\bar{f}(x)$ on (a, ∞) as follows

$$\bar{f}(x) := \sup\{f(y) : y \in [x, \infty)\}.$$

Check that $\bar{f}(x)$ is a decreasing function on (a, ∞). Hence it has a limit at ∞. It is a fact, that we are not going to prove that

$$\lim_{x\to\infty} \bar{f}(x) = \limsup_{x\to\infty} f(x).$$

Analogously, define a new function $\underline{f}(x)$ on (a, ∞) as follows

$$\underline{f}(x) := \inf\{f(y) : y \in [x, \infty)\}.$$

Check that $\underline{f}(x)$ is a increasing function on (a, ∞). Hence it has a limit at ∞. It is a fact, that we are not going to prove that

$$\lim_{x\to\infty} \underline{f}(x) = \liminf_{x\to\infty} f(x).$$

Because of this liminf and limsup of a function $f(x)$ as x approaches infinity are sometimes denoted by

$$\lim_{x\to\infty} \inf_{t\ge x} f(t) \text{ and } \lim_{x\to\infty} \sup_{t\ge x} f(t).$$

Exercise 24. For any functions $f(x)$ and $g(x)$ defined on (a, ∞) we have

$$\limsup_{x\to\infty}(f(x) + g(x)) \ge \limsup_{x\to\infty} f(x) + \liminf_{x\to\infty} g(x) \ge \liminf_{x\to\infty}(f(x) + g(x)).$$

Exercise 25. Let $f(x)$ and $g(x)$ be any functions defined on (a, ∞) and suppose

$$\liminf_{x\to\infty}(g(x) - f(x)) \ge 0.$$

Show that $\limsup_{x\to\infty} f(x) \le \limsup_{x\to\infty} g(x)$ and $\liminf_{x\to\infty} f(x) \le \liminf_{x\to\infty} g(x)$.

1.7 Functions and sets

Let X and Y be two sets and let $f : X \to Y$ be a function from X to Y. For any subset $A \subseteq X$ we define the image of A under f to be

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

and note that $f(A) \subseteq Y$. For any subset $B \subseteq Y$ we define the preimage of B under f to be

$$f^{-1}(B) := \{x \in X : f(x) \in B\}$$

and note that $f^{-1}(B) \subseteq X$.

Here are several facts that are not difficult to establish and will be used repeatedly. Let A_i, $i \in I$ be a family of subsets of X where I is an index set. Let B_j, $j \in J$ be a family of subsets of Y where J is an index set.

(i) $f\Big(\bigcup_{i\in I} A_i\Big) = \bigcup_{i\in I} f(A_i)$;

(ii) $f\Big(\bigcap_{i\in I} A_i\Big) \subseteq \bigcap_{i\in I} f(A_i)$;

(iii) $f^{-1}\Big(\bigcup_{j\in J} B_j\Big) = \bigcup_{j\in J} f^{-1}(B_j)$;

(iv) $f^{-1}\Big(\bigcap_{j\in J} B_j\Big) = \bigcap_{j\in J} f^{-1}(B_j)$;

(v) $f^{-1}(B^c) = (f^{-1}(B))^c$ for all $B \subseteq Y$.

If $g : Y \to Z$ is another function then $g \circ f : X \to Z$ denotes the composition $(g \circ f)(x) := g(f(x))$. Let $A \subseteq X$, $B \subseteq Y$, and $C \subseteq Z$ be three subsets. One can show that

(i) $(f^{-1} \circ f)(A) \supseteq A$;

(ii) $(f \circ f^{-1})(B) \subseteq B$;

(iii) $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$.

1.8 Stolz, Cesaro, Kronecker

The following theorem can be viewed as a l'Hôpital's rule for sequences.

Theorem 26 (Stolz)**.** Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be sequences of real numbers and let $\{b_n\}_{n=1}^{\infty}$ be strictly increasing and converging to infinity. If

$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n} = \ell,$$

where $\ell \in \mathbb{R} \cup \{-\infty, \infty\}$. Then

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \ell.$$

A corollary of Stolz' theorem is a result for the so-called Cesaro means that will be useful later. The proof is left as an exercise.

Corollary 27 (Cesaro means)**.** *Let* $\{x_n\}$ *be a convergent sequence of real numbers with* $\lim_{n\to\infty} x_n = x$*. Define* $y_n := \frac{1}{n}\sum_{i=1}^{n} x_i$ *for* $n = 1, 2, 3, \ldots$ *Show that* $\lim_{n\to\infty} y_n = x$*.*

Corollary 28 (Kronecker lemma)**.** Let $\lim\limits_{n\to\infty} \sum\limits_{k=1}^{n} x_k = s \in \mathbb{R}$ and $\{b_n\}_{n=1}^{\infty}$ be an increasing sequence of real numbers converging to infinity, then $\lim\limits_{n\to\infty} \dfrac{1}{b_n} \sum\limits_{k=1}^{n} b_k x_k = 0$.

Proof. Define for convenience $b_0 := 0$. Let $s_n := \sum_{k=1}^{n} x_k$ with $s_0 := 0$. Then, $\lim\limits_{n\to\infty} s_n = s$ and

$$\frac{1}{b_n}\sum_{k=1}^{n} b_k x_k = \frac{1}{b_n}\sum_{k=1}^{n} b_k(s_k - s_{k-1}) = s_n - \frac{1}{b_n}\sum_{k=1}^{n}(b_k - b_{k-1})s_{k-1}.$$

Let $a_n := \sum_{k=1}^{n}(b_k - b_{k-1})s_{k-1}$ and apply Stolz theorem:

$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n} = \lim_{n\to\infty} \frac{(b_{n+1} - b_n)s_n}{b_{n+1} - b_n} = \lim_{n\to\infty} s_n = s.$$

This shows that $\lim\limits_{n\to\infty} \dfrac{a_n}{b_n} = \lim\limits_{n\to\infty} \dfrac{1}{b_n}\sum\limits_{k=1}^{n}(b_k - b_{k-1})s_{k-1} = s$ and the result follows. □

2 Measure spaces

A *measure space* is the triple $(\Omega, \mathcal{F}, \mathbb{P})$. We proceed to define and illustrate the constituent components of a probability space.

(1) Ω is a non-empty set of elementary events, or outcomes of an experiment, or states. Those elementary events are denoted ω.

Example 29. (a) If we roll a die, then all possible outcomes are the numbers between 1 and 6. That means that the possible elementary events of the experiment are $\Omega = \{1, 2, 3, 4, 5, 6\}$. (b) If we flip a coin, then the possible outcomes are either 'head' (H) or 'tail' (T), that means that possible elementary events are $\Omega = \{H, T\}$. If we flip two coins, then $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ is the set of all elementary outcomes. (c) If we measure the lifetime of a lightbulb in hours then we can theoretically choose $\Omega = [0, \infty)$. □

The set $\mathcal{F}$ is called a σ-algebra (definition coming soon). This is the collection of observable subsets $A \subseteq \Omega$ also called *events*. The interpretation is that one can usually not decide whether a system is in the particular state $\omega \in \Omega$, but one can decide whether A contains the unknown state ω or not. In the light bulb example above, we can never determine (due to the fact that we cannot measure time with infinite precision), whether a light bulb that just burned out, lasted exactly $\pi = 3.14159\ldots$ hours, but we can determine if it lasted between, say, 3 and 3.2 hours. That is, if

ω is the exact life span of the bulb, we can determine whether or not $\omega \in [3, 3.2]$.

The last component of a measure space is a function $\mathbb{P}$, defined on $\mathcal{F}$ with values in $[0, \infty)$. The function $\mathbb{P}$ is called a *measure*. That is, for every set $A \in \mathcal{F}$ the value $\mathbb{P}(A)$ is a non-negative number.

The pair $(\Omega, \mathcal{F})$ is called *measurable space*. One may have many different measures on the same measurable space.

2.1 σ-algebras

The σ-algebra is a basic tool in probability theory. It is the set the probability measures are defined on. Without this notion it would be impossible to consider the fundamental Lebesgue measure on the interval $[0, 1]$ or to consider Gaussian measures, without which many parts of mathematics can not live.

Definition 30 (algebra)**.** Let Ω be a non-empty set. A collection $\mathcal{F}$ of subsets $A \subseteq \Omega$ is called an *algebra* on Ω if

(i) $\emptyset \in \mathcal{F}$,

(ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$, and

(iii) $A, B \in \mathcal{F}$ implies that $A \cup B \in \mathcal{F}$.

By simple induction, the third condition implies that for any finite number of sets $A_1, \ldots, A_n \in \mathcal{F}$ we have $\bigcup_{i=1}^n A_i \in \mathcal{F}$.

Definition 31 (σ-algebra)**.** Let Ω be a non-empty set. A collection $\mathcal{F}$ of subsets $A \subseteq \Omega$ is called an *σ-algebra* on Ω if

(i) $\emptyset \in \mathcal{F}$,

(ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$, and

(iii) $A_1, A_2, \ldots \in \mathcal{F}$ implies that $\bigcup_{i=1}^\infty A_i \in \mathcal{F}$.

The only difference between algebra and σ-algebra is that the σ-algebra is "closed" under countable union of sets from it. Often algebra and σ-algebra are called *field* and *σ-field* respectively. Clearly every σ-algebra is also an algebra, but the opposite is not true, see example below. Clearly, if an algebra contains only a finite number of sets, then it is a σ-algebra. If $\Omega = \{\omega_1, \ldots, \omega_n\}$ (finitely many elementary events) then any algebra on Ω is automatically a σ-algebra.

Given two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ on Ω we say that $\mathcal{F}_2$ is larger than $\mathcal{F}_1$ if $A \in \mathcal{F}_1$ implies $A \in \mathcal{F}_2$. That is, if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, in other words, $\mathcal{F}_2$ contains more events and that is why we may say that it is *finer* than $\mathcal{F}_1$.

Example 32. Let Ω be a non-empty set.

(a) Let $\mathcal{F} := 2^\Omega$—the collection of all possible subsets of Ω including the emptyset. This is clearly a σ-algebra and it contains any other σ-algebra on Ω. That is why this is the largest σ-algebra on Ω.

(b) Let $\mathcal{F} := \{\emptyset, \Omega\}$. This is the smallest σ-algebra on Ω. It is contained in every other σ-algebra on Ω.

(c) Fix a subset $A \subseteq \Omega$, then $\mathcal{F} := \{\emptyset, A, A^c, \Omega\}$ is a σ-algebra. □

Exercise 33 (De Morgan's Laws)**.** *For any sets $A_1, A_2, \ldots \subseteq \Omega$ we have*

$$\Big(\bigcup_{i=1}^{\infty} A_i\Big)^c = \bigcap_{i=1}^{\infty} A_i^c \quad \textit{and} \quad \Big(\bigcap_{i=1}^{\infty} A_i\Big)^c = \bigcup_{i=1}^{\infty} A_i^c.$$

Exercise 34 (Restricting a σ-algebra)**.** Suppose $\mathcal{F}$ is a σ-algebra on Ω. Let $\bar{\Omega} \subseteq \Omega$. Show that the collection of sets $\bar{\mathcal{F}} := \{\bar{\Omega} \cap A : A \in \mathcal{F}\}$ is a σ-algebra on $\bar{\Omega}$.

Immediate properties of a σ-algebra.
(a) $\Omega \in \mathcal{F}$. Indeed, since $\emptyset \in \mathcal{F}$ the second rule of Definition 31 says that $\Omega = \emptyset^c \in \mathcal{F}$.
(b) $A_1, A_2, \ldots \in \mathcal{F}$ implies that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$. Indeed, the second rule of Definition 31 says that $A_i^c \in \mathcal{F}$ for all i so by the third rule $\bigcup_{i=1}^{\infty} A_i^c \in \mathcal{F}$. Then, by the above lemma, $\bigcap_{i=1}^{\infty} A_i = \Big(\bigcup_{i=1}^{\infty} A_i^c\Big)^c \in \mathcal{F}$, where we used the second rule again.
(c) If $A, B \in \mathcal{F}$ then $A \setminus B \in \mathcal{F}$. Indeed, $A \setminus B = A \cap B^c \in \mathcal{F}$.

Example 35 (algebra which is not a σ-algebra)**.** Let $\mathcal{F}$ be the collection of all subsets $A \subseteq \mathbb{R}$ such that either A contains only finitely many elements or A^c contains only finitely many elements. □

Example 36 (algebra which is not a σ-algebra)**.** Let $\mathcal{F}$ be the collection of all subsets $A \subseteq \mathbb{R} \cup \{\infty\}$ that can be written as

$$A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n],$$

where $-\infty \leq a_1 \leq b_1 \leq \cdots \leq a_n \leq b_n \leq \infty$ with the convention $(a, a] = \emptyset$. Then, $\mathcal{F}$ is an algebra that is not a σ-algebra (why?). □

Unfortunately, most of the important σ-algebra can not be constructed explicitly. Surprisingly, one can work practically with them nevertheless. In the following we describe a simple procedure which generates σ-algebras. We start with the fundamental fact that intersection of σ-algebras is a σ-algebra. The proof is very easy, but important.

Proposition 37. Let $\mathcal{F}_i$, $i \in I$, $I \neq \emptyset$, be a family of σ-algebras on a nonempty set Ω, where I is an arbitrary index set. Then

$$\mathcal{F} := \bigcap_{i \in I} \mathcal{F}_i$$

is a σ-algebra as well.

Proof. Notice first that $\emptyset \in \mathcal{F}_i$ for all $i \in I$ so $\emptyset \in \bigcap_{i \in I} \mathcal{F}_i$. Next, let $A, A_1, A_2, \ldots \in \bigcap_{i \in I} \mathcal{F}_i$. Hence $A, A_1, A_2, \ldots \in \mathcal{F}_i$ for all $i \in I$, which since $\mathcal{F}_i$ is a σ-algebra implies that $A^c \in F_i$ and $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}_i$ for all $i \in I$. Consequently $A^c \in \bigcap_{i \in I} \mathcal{F}_i$ and $\bigcup_{j=1}^{\infty} A_j \in \bigcap_{i \in I} \mathcal{F}_i$. □

The next proposition explains how an arbitrary collection of sets determines a unique σ-algebra.

Proposition 38. *Let $\mathcal{F}$ be a arbitrary collection of subsets of Ω. Then, there is a unique smallest σ-algebra, denoted by $\sigma(\mathcal{F})$, on Ω such that $\mathcal{F} \subseteq \sigma(\mathcal{F})$.*

Proof. Let J be the set of all σ-algebras on Ω that contain the sets in $\mathcal{F}$, such σ-algebras exist (i.e. the set J is not empty) since the σ-algebra 2^Ω contains the sets in $\mathcal{F}$. Define

$$\sigma(\mathcal{F}) := \bigcap_{C \in J} C$$

to be the intersection of all σ-algebras in J. By Proposition 37, $\sigma(\mathcal{F})$ is a σ-algebra. To show that $\sigma(\mathcal{F})$ is smallest σ-algebra containing the sets in $\mathcal{F}$. Indeed, assume that $\mathcal{G}$ is a σ-algebra containing the sets in $\mathcal{F}$, then by definition $\mathcal{G} \in J$ so that $\sigma(\mathcal{F}) := \bigcap_{C \in J} C \subseteq \mathcal{G}$.

Finally, suppose there is another smallest σ-algebra containing the sets in $\mathcal{F}$. Call it $\mathcal{F}'$. Since $\mathcal{F}' \in J$, we get $\sigma(\mathcal{F}) \subseteq \mathcal{F}'$. Since $\mathcal{F}'$ is smaller than any other σ-algebra containing the sets in $\mathcal{F}$, we get $\mathcal{F}' \subseteq C$ for all $C \in J$. Hence, $\mathcal{F}' \subseteq \sigma(\mathcal{F})$. Therefore $\mathcal{F}' = \sigma(\mathcal{F})$. □

We say that the σ-algebra $\sigma(\mathcal{F})$ is *generated* by the collection of sets $\mathcal{F}$. The construction follows a "top-down" approach. It is elegant but the disadvantage is that there is no explicit formula for the elements of $\sigma(\mathcal{F})$. If $\mathcal{F}$ is a finite collection of sets, then $\sigma(\mathcal{F})$ has finitely many sets and it is possible to describe them explicitly, as the next exercise shows.

Exercise 39. (a) What is the σ-algebra generated by $\emptyset$?
(b) Fix a subset $A \subseteq \Omega$. What is the σ-algebra generated by A?
(c) Fix two subsets $A, B \subseteq \Omega$. What is the σ-algebra generated by $\{A, B\}$?

The next exercise shows that the smallest algebra (defined in an analogous way) generated by a collection of sets can be constructed explicitly from the "bottom-up".

Exercise 40. Let $\mathcal{M}$ be a collection of subsets of Ω. Define

$$\mathcal{M}' := \{\emptyset, \Omega\} \cup \big(\cup_{A \in \mathcal{M}} \{A, A^c\} \big).$$

Next, define $\mathcal{M}''$ to be the set of all finite unions of finite intersections of sets from $\mathcal{M}'$. That is

$$\mathcal{M}'' = \big\{ \cup_{i=1}^n \cap_{j=1}^m A_{ij} : A_{ij} \in \mathcal{M}' \text{ for all } i = 1, \ldots, n \text{ and } j = 1, \ldots, m, \text{ where } n, m \in \mathbb{N} \big\}.$$

(a) Show that $\mathcal{M}''$ is an algebra.
(b) Show that $\mathcal{M}''$ is the smallest algebra containing $\mathcal{M}$.

Exercise 41. Suppose that $\mathcal{M}_1$ and $\mathcal{M}_2$ are collections of subsets of Ω and $\mathcal{F}$ is a σ-algebra. Show that
(a) If $\mathcal{M}_1 \subseteq \mathcal{M}_2$, then $\sigma(\mathcal{M}_1) \subseteq \sigma(\mathcal{M}_2)$.
(b) If $\mathcal{M}_1 \subseteq \mathcal{F}$, then $\sigma(\mathcal{M}_1) \subseteq \mathcal{F}$.

Exercise 42. Suppose that Ω_1 and Ω_2 are two sets and $f : \Omega_1 \to \Omega_2$ is a function between them. Let $\mathcal{M}$ be a collection of subsets of Ω_2 and let $\mathcal{F}$ be a σ-algebra on Ω_2. Show that
(a) $f^{-1}(\mathcal{F})$ is a σ-algebra on Ω_1.
(b) $f^{-1}(\sigma(\mathcal{M})) = \sigma(f^{-1}(\mathcal{M}))$.

2.1.1 The Borel σ-algebra on $\mathbb{R}$

We now turn to one of the most important examples, the Borel σ-algebra on $\mathbb{R}$. To do this we need the notion of open and closed sets.

Definition 43 (open and closed sets)**.** (i) A subset $A \subseteq \mathbb{R}$ is called *open*, if for each $x \in A$ there is an $\epsilon > 0$ such that $(x-\epsilon, x+\epsilon) \subset A$.

(ii) A subset $B \subseteq \mathbb{R}$ is called *closed*, if $A := \mathbb{R} \setminus B$ is open.

Note that $\mathbb{R}$ and $\emptyset$ are the only subsets of $\mathbb{R}$ that are both open and closed. Note that some sets are neither open nor closed, for example $(a, b]$ for $a < b$. The open subsets of $\mathbb{R}$ have a simple description. It is a fact that we are not going to prove, that every open subset of $\mathbb{R}$ is a finite or countable union of disjoint open intervals of the type (a, b). Countable union means that the sets that are being united can be ordered in a sequence. That is, every open subset A of $\mathbb{R}$ can be represented as

$$A = \bigcup_{n=1}^{\infty} (a_i, b_i)$$

for some $-\infty \leq a_i \leq b_i \leq \infty$, $i = 1, 2, \ldots$, such that the intervals $\{(a_i, b_i) : i = 1, 2, \ldots\}$ are disjoint. In this representation, we allow some of the intervals to be unbounded, that is $(-\infty, b)$ or (a, ∞).

The structure of the closed subsets of $\mathbb{R}$ could be quite complicated as the Cantor set shows.

Proposition 44 (Borel σ-algebra on $\mathbb{R}$)**.** Let
$\mathcal{F}_0$ be the collection of all open subsets or $\mathbb{R}$,
$\mathcal{F}_1$ be the collection of all closed subsets or $\mathbb{R}$,
$\mathcal{F}_2$ be the collection of all intervals $(-\infty, b]$ for $b \in \mathbb{R}$,
$\mathcal{F}_3$ be the collection of all intervals $(-\infty, b)$ for $b \in \mathbb{R}$,
$\mathcal{F}_4$ be the collection of all intervals $(a, b]$ for $-\infty < a < b < \infty$,
$\mathcal{F}_5$ be the collection of all intervals (a, b) for $-\infty < a < b < \infty$.
Then $\sigma(\mathcal{F}_0) = \sigma(\mathcal{F}_1) = \sigma(\mathcal{F}_2) = \sigma(\mathcal{F}_3) = \sigma(\mathcal{F}_4) = \sigma(\mathcal{F}_5)$.

Proof. Since $\mathcal{F}_3 \subset \mathcal{F}_0$ we have

$$\sigma(\mathcal{F}_3) \subseteq \sigma(\mathcal{F}_0).$$

Next, we look at $\mathcal{F}_5$. For $-\infty < a < b < \infty$ we have that

$$(a, b) = \bigcup_{n=N}^{\infty} [a + 1/n, b) = \bigcup_{n=N}^{\infty} \Big((-\infty, b) \setminus (-\infty, a + 1/n)\Big) \in \sigma(\mathcal{F}_3),$$

where in the first union we have chosen N to be big enough. The second union shows that $\mathcal{F}_5 \subseteq \sigma(\mathcal{F}_3)$ and thus

$$\sigma(\mathcal{F}_5) \subseteq \sigma(\mathcal{F}_3).$$

Now, let $A \subseteq \mathbb{R}$ be a non-empty open set. By the structure facts about open sets, mentioned above,

$$A = \bigcup_{n=1}^{\infty} (a_i, b_i)$$

for some $-\infty \leq a_i < b_i \leq \infty$, $i = 1, 2, \ldots$ (If there are infinite intervals, $(-\infty, b)$ or (a, ∞), in the representation of A, note that they are already in $\sigma(\mathcal{F}_5)$ as a union of countably many finite intervals.) This shows that $\mathcal{F}_0 \subseteq \sigma(\mathcal{F}_5)$ and hence

$$\sigma(\mathcal{F}_0) \subseteq \sigma(\mathcal{F}_5)$$

Combining the inclusions, we established that

$$\sigma(\mathcal{F}_0) \subseteq \sigma(\mathcal{F}_5) \subseteq \sigma(\mathcal{F}_3) \subseteq \sigma(\mathcal{F}_0)$$

showing that

$$\sigma(\mathcal{F}_0) = \sigma(\mathcal{F}_5) = \sigma(\mathcal{F}_3).$$

Next, since $A \in \mathcal{F}_0$ implies $A^c \in \mathcal{F}_1 \subseteq \sigma(\mathcal{F}_1)$, we get that $A = (A^c)^c \in \sigma(\mathcal{F}_1)$. Hence $\mathcal{F}_0 \subseteq \sigma(\mathcal{F}_1)$ and $\sigma(\mathcal{F}_0) \subseteq \sigma(\mathcal{F}_1)$. The inclusion $\sigma(\mathcal{F}_1) \subseteq \sigma(\mathcal{F}_0)$ can be shown in a similar way. Thus

$$\sigma(\mathcal{F}_0) = \sigma(\mathcal{F}_1).$$

The rest of the proof is left as an exercise. □

Definition 45. The σ-algebra constructed in Proposition 44 is called the *Borel* σ-algebra on $\mathbb{R}$ and is denoted by $\mathcal{B}(\mathbb{R})$.

Example 46. An interesting Borel set is the Cantor set C, described in Example 8. Indeed, there we started from the Borel set $[0, 1]$ and removed from it a countably many open intervals (also Borel sets). Thus, the result must be in the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

It is a fact that not every subset of $\mathbb{R}$ is a Borel set. That is, there are subsets of $\mathbb{R}$ that are not Borel sets. The proof of this fact is beyond our goals and will not be given.

2.1.2 Constructing $\sigma(\mathcal{F})$ from the bottom-up (optional)

A set X is called *totally ordered* if a binary relation is defined on it, denoted $\leq$, that satisfies the following properties for all $a, b, c \in X$.

(i) If $a \leq b$ and $b \leq a$, then $a = b$ (antisymmetry);

(ii) If $a \leq b$ and $b \leq c$, then $a \leq c$ (transitivity);

(iii) Either $a \leq b$ or $b \leq a$.

If in addition, it satisfies the property

(iv) Any non-empty subset $A \subseteq X$ contains an element a, such that $a \leq x$ for all $x \in A$,

then X is called *totally well-ordered* or *well-ordered* for short. This axiom says that every subset of X has a smallest element, which, by the other three axioms, has to be unique. For example, the set $\{0, 1, 2, \ldots\}$ is well-ordered, while the set $[0, 1]$ is totally ordered but not well-ordered, since the subset $(1/2, 3/4)$ does not contain a smallest element. If $a \leq b$ and a is not equal to b, than we say that a is strictly smaller than b, denoted $a < b$. An equivalent way to express axiom (iv) is to say that X does not have a strictly decreasing sequence. The Zermelo's Well-Ordering Theorem states that any set X can be well-ordered, possibly in many different ways. Two well-ordered sets X and Y are isomorphic (equivalent) if there is a one-to-one and onto map f from X to Y that preserves the order: if $a \leq b$ in X, then $f(a) \leq f(b)$ in Y.

Definition 47. A well-ordered set λ is called an *ordinal number* if every element of λ is also a subset of λ: if $a \in \lambda$ then $a \subset \lambda$.

In other words, the elements of an ordinal number λ are sets themselves and λ contains all their elements as well. Moreover, it can be shown, that the order of the set λ is the one generated by set inclusion: for every $a, b \in \lambda$, we have $a \leq b$ if and only if $a \subseteq b$. Another striking fact is that every well-ordered set X is isomorphic to an ordinal number λ. Since we are not going to distinguish between isomorphic well-ordered sets, one can think of every well-ordered set X as an ordinal number and two well-ordered sets X and Y represent the same ordinal number if and only if they are isomorphic.

The ordinal numbers extend the natural numbers. That is why the word 'number' appears in the name. To see that, represent 0 as $\emptyset$ and

- $1 := \{\emptyset\}$
- $2 := \{\emptyset, \{\emptyset\}\}$
- $3 := \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$
- $4 := \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}$,

and continue like that for every natural number. The sets on the right-hand side, whose elements are well-ordered according to set inclusion, are ordinal numbers. For example, the set for 4, contains the element $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, which represents 3, and it also contains all the elements of 3, namely $\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}$. In this way, one can use the funny notation $3 \in 4$.

As another example, the set $\omega := \{0, 1, 2, \ldots\}$ is well-ordered, hence it is an ordinal number. Note that this set is infinite but countable. There are ordinals that are not countable, just any well-order on $[0, 1]$ will turn it into an uncountable ordinal.

We now state the properties of ordinal numbers that are important to us.

1. Every element of an ordinal number is a well-ordered set, hence an ordinal number itself, for example $5 \in \omega$.

2. If λ and μ are ordinal numbers, then exactly one of the following situations holds: (a) λ is isomorphic to an element of μ, or (b) μ is isomorphic to an element of λ, or (c) λ is isomorphic to μ. This allows one to define an order between the ordinal numbers: $\lambda \leq \mu$ if (a) holds, or $\mu \leq \lambda$ if (b) holds, or $\lambda = \mu$ if (c) holds. Thus, any set of ordinal numbers is totally ordered.

   For example $n < \omega$ for all natural numbers n. As another example, take any $a \in \lambda$. Since a is an ordinal number isomorphic to itself, we conclude that $a \leq \lambda$. Moreover, a cannot be equal to λ (no set contains itself), hence $a < \lambda$. To emphasize: the notation $\mu \leq \lambda$ means that μ is isomorphic either to λ or to an element of λ. In the fist case we have $\mu = \lambda$, while in the second (since we do not distinguish between isomorphic sets) $\mu \in \lambda$. This shows, that $\lambda = \{a : a < \lambda\}$, or using funny notation $\lambda = [0, \lambda)$.

3. Any collection of ordinal numbers has a unique smallest element. This implies that, any set of ordinal numbers is well-ordered, hence (isomorphic to) an ordinal number. For example, $\{1, 3\} = \{\{\emptyset\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}\}$ is isomorphic to $\{\emptyset, \{\emptyset\}\} = 2$.

Given an ordinal number λ, the *successor of* λ, denoted by $\lambda + 1$, is defined to be the set $\lambda \cup \{\lambda\}$. That is, the elements of $\lambda + 1$ are those of λ and λ itself, or using the funny notation $\lambda+1 = [0, \lambda]$. Since λ is an element of $\lambda+1$, we have $\lambda < \lambda+1$. For example $\omega+1 = \{0, 1, 2, \ldots, \omega\}$, $\omega + 2 = \{0, 1, 2, \ldots, \omega, \omega + 1\}$, and repeating, we obtain

$$\omega + \omega = \{0, 1, 2, \ldots, \omega, \omega + 1, \omega + 2, \ldots\} =: 2\omega.$$

We can continue $\{0, 1, 2, \ldots, \omega, \omega + 1, \omega + 2, \ldots, 2\omega, 2\omega + 1, 2\omega + 2, \ldots\} =: 3\omega$ and if we add three more dots (think of what that means) one obtains

$$\{0, 1, 2, \ldots, \omega, \omega + 1, \omega + 2, \ldots, 2\omega, 2\omega + 1, 2\omega + 2, \ldots, \ldots\} =: \omega^\omega,$$

which is still a countable set, since it is a countable union of countable sets. One can continue like that until their head explodes (for example, when trying to come to terms with the set $\omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}$, which is still a countable ordinal).

If an ordinal is not the successor of any other ordinal, then it is called a *limit ordinal.* Thus, there are two types of ordinals: successors and limit ordinals. For example, ω and ω^ω are limit ordinals, while 5 is a successor since $4 + 1 = 5$.

The collection of all uncountable ordinals has a smallest element, denoted by ω_1. It is not difficult to show that λ is a countable ordinal if and only if $\lambda \in \omega_1$, that is, ω_1 is the set of all countable ordinals. To see that ω_1 is a limit ordinal (i.e. not a successor), suppose that $\omega_1 = \lambda + 1$ for some λ. Such λ must satisfy $\lambda < \omega_1$, so it is countable. But then $\lambda + 1$ is countable (just one element was added to λ), contradicting the fact that ω_1 is uncountable.

Lemma 48. *Let $\lambda_1, \lambda_2, \ldots$ be a sequence of countable ordinals. Then, there is a countable ordinal μ^*, such that $\lambda_i < \mu^*$ for all $i = 1, 2, \ldots$*

Proof. The set $\omega_1 \setminus \big(\big(\cup_{i=1}^\infty \lambda_i\big) \cup \{\lambda_1, \lambda_2, \ldots\}\big)$ is not empty, since from an uncountable set we subtract a countable one. Let μ^* be an element in the difference. Since $\mu^* \in \omega_1$, it is a countable ordinal. Clearly, $\mu^* \neq \lambda_i$ for all i. Also, one cannot have $\mu^* < \lambda_i$, since then $\mu^* \in \lambda_i$, contradicting the choice of μ^*. Hence, $\lambda_i < \mu^*$ for all $i = 1, 2, \ldots$ □

Exercise 49. *If λ is a limit ordinal, then $\lambda = \bigcup_{\mu<\lambda} \mu$.*

The usefulness of the ordinal numbers is that they themselves are well-ordered. So, just like one can do inductive arguments over the natural numbers, one can do inductive arguments (or definitions) over the ordinal numbers.

Lemma 50 (Transfinite induction)**.** *Consider a property $\mathcal{P}(\lambda)$ that depends on the ordinals. Suppose that $\mathcal{P}(0)$ is true. If one can show that $\mathcal{P}(\lambda)$ is true, whenever $\mathcal{P}(\mu)$ is true for every $\mu < \lambda$, then $\mathcal{P}(\lambda)$ is true for every ordinal λ.*

Proof. Suppose that the property is not true for every ordinal number. Then, the collection of ordinals, for which the property is not true, is not empty. It has a smallest element, call it μ^*. (We have $\mu^* > 0$, since $\mathcal{P}(0)$ is true.) That is, $\mathcal{P}(\mu)$ is true for every $\mu < \mu^*$. But then, by the hypothesis, one can show that $\mathcal{P}(\mu^*)$ is true as well, a contradiction. □

When making the induction step, one usually considers two cases: a successor and a limit case, corresponding to the two different types of ordinal numbers. (Every natural number is a successor, that is why the ordinary induction has only a successor case.) We are now ready to define the sigma algebra $\sigma(\mathcal{F})$. Let $\mathcal{F}$ be a collection of subsets from Ω.

Base case: Let $\sigma_0 := \mathcal{F} \cup \{\emptyset, \Omega\}$.

Successor case: Suppose σ_λ has been defined, then let

$$\sigma_{\lambda+1} := \sigma_\lambda \bigcup \{\cup_{i=1}^\infty A_i : A_i \in \sigma_\lambda\} \bigcup \{(\cup_{i=1}^\infty A_i)^c : A_i \in \sigma_\lambda\}.$$

Limit case: Suppose λ is a limit ordinal and that σ_μ has been defined for all $\mu < \lambda$. Then, let

$$\sigma_\lambda := \bigcup_{\mu<\lambda} \sigma_\mu.$$

In this way, one can construct a collection of sets σ_λ for any ordinal λ. Procedure like that for constructing objects is called *transfinite recursion.* In this case, the procedure implies that if $\mu < \lambda$, then $\sigma_\mu \subseteq \sigma_\lambda$.

Theorem 51. *The collection of sets σ_{ω_1} is the smallest σ-algebra containing $\mathcal{F}$, that is*

$$\sigma(\mathcal{F}) = \sigma_{\omega_1}.$$

Proof. Since ω_1 is a limit ordinal, we have $\sigma_{\omega_1} = \cup_{\mu<\omega_1}\sigma_\mu$. Let us see first that σ_{ω_1} is a σ-algebra containing $\mathcal{F}$. Since $0 < \omega_1$, the limit case of the definition, says that $\sigma_0 \subseteq \sigma_{\omega_1}$. So, $\emptyset, \Omega \in \sigma_{\omega_1}$ and $\mathcal{F} \subseteq \sigma_{\omega_1}$.

If $A \in \sigma_{\omega_1}$, then the limit case of the definition, says that $A \in \sigma_\mu$ for some $\mu < \omega_1$. Then, the successor case implies that $A^c \in \sigma_{\mu+1}$, and since $\mu + 1 < \omega_1$, the limit case of the definition again implies $A^c \in \sigma_{\omega_1}$.

Now let $A_i \in \sigma_{\omega_1}$ for all $i = 1, 2, \ldots$ The limit case of the definition implies that for every $i = 1, 2, \ldots$, there is an ordinal $\mu_i < \omega_1$, such that $A_i \in \sigma_{\mu_i}$. Necessarily, μ_i is countable ordinal for each i. By Lemma 48, there is a countable μ^*, such that $\mu_i \leq \mu^*$. Hence, $A_i \in \sigma_{\mu_i} \subseteq \sigma_{\mu^*}$. Thus, by the successor case $\cup_{i=1}^\infty A_i \in \sigma_{\mu^*+1} \subset \sigma_{\omega_1}$, since $\mu^* + 1 < \omega_1$.

This shows that $\sigma(\mathcal{F}) \subseteq \sigma_{\omega_1}$. To show the opposite inclusion, we use the principle of *transfinite induction.*

As a base case, note that $\sigma_0 \subseteq \sigma(\mathcal{F})$. Consider and ordinal λ and suppose that $\sigma_\mu \subseteq \sigma(\mathcal{F})$ for all $\mu < \lambda$. We need to show that $\sigma_\lambda \subseteq \sigma(\mathcal{F})$. Consider two cases. If λ is a successor, that is $\lambda = \mu + 1$ for some $\mu < \lambda$, then we are done by the successor case of the definition. If λ is a limit ordinal, then we are done by the limit case of the definition. This shows $\sigma_\lambda \subseteq \sigma(\mathcal{F})$ for all ordinals λ. In particular $\sigma_{\omega_1} \subseteq \sigma(\mathcal{F})$. □

Exercise 52. *Use the above ideas to show that $f^{-1}(\sigma(\mathcal{M})) = \sigma(f^{-1}(\mathcal{M}))$, where f and $\mathcal{M}$ are as in Exercise 42.*

2.2 Measures

A collection of subsets $\{A_i\}_{i\in I}$ of Ω is called *disjoint* if $A_i \cap A_j = \emptyset$ for all $i, j \in I$ with $i \neq j$.

Definition 53. Suppose $\mathcal{F}$ is a σ-algebra.
A map $\mathbb{P} : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ is called a *measure on the σ-algebra $\mathcal{F}$* if
1) $\mathbb{P}(\emptyset) = 0$
2) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$; and
3) For any sequence of *disjoint* sets $A_1, A_2, \ldots \in \mathcal{F}$ we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{3}$$

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called *measure space.* The basic properties of a measure are collected below.

Proposition 54. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space. Then the following properties hold*

(i) If $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint sets then $\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i)$;

(ii) If $A, B \in \mathcal{F}$ then $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$;

(iii) If $A, B \in \mathcal{F}$ and $B \subseteq A$ then $\mathbb{P}(B) \leq \mathbb{P}(A)$;

(iv) If $A_1, A_2, \ldots \in \mathcal{F}$ is a sequence of any sets then $\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$;

(v) Continuity from below: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are such that $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big);$$

(vi) Continuity from above: If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ are such that $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ and $\mathbb{P}(A_1) < \infty$, then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i\Big).$$

Proof. (i) Let $A_{n+1} = A_{n+2} = \cdots = \emptyset$ so that

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i),$$

because $\mathbb{P}(\emptyset) = 0$.

(ii) Since $(A \cap B) \cap (A \setminus B) = \emptyset$, we get

$$\mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) = \mathbb{P}((A \cap B) \cup (A \setminus B)) = \mathbb{P}(A).$$

(iii) Immediate from the previous part since $\mathbb{P}(A \setminus B) \geq 0$ and $A \cap B = B$.

(iv) Let $B_1 := A_1$ and $B_i := A_1^c \cap A_2^c \cap \cdots \cap A_{i-1}^c \cap A_i$ for $i = 2, 3, \ldots$ Since $B_i \subseteq A_i$ we get $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ for all i. Note that the sets $B_1, B_2, \ldots$ are disjoint and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Hence

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

(v) Let $B_1 := A_1$ and $B_2 := A_2 \setminus A_1$, $B_3 := A_3 \setminus A_2, \ldots$ Then, we have that the sets $B_1, B_2, \ldots$ are disjoint and

$$\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \bigcup_{i=1}^{n} B_i = A_n.$$

Consequently

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{i=1}^{n} B_i\Big) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(vi) This part is left as an exercise. (Where do you use the requirement that $\mathbb{P}(A_1) < \infty$?)

□

The third property in Proposition 54 implies that for any $A \in \mathcal{F}$ we have $0 \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega)$. Indeed, since $\emptyset \subseteq A \subseteq \Omega$, then $0 = \mathbb{P}(\emptyset) \leq \mathbb{P}(A) \leq P(\Omega)$.

• If $\mathbb{P}(\Omega) = 1$ then the measure $\mathbb{P}$ is called *probability measure* and the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called *probability space*. In that case $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$.

• If $\mathbb{P}(\Omega) < \infty$ then the measure $\mathbb{P}$ is called *finite*.

• A measure $\mathbb{P}$ is called *σ-finite* if there is a sequence of sets $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_3 \subseteq \cdots$ such that $\Omega = \bigcup_{n=1}^{\infty} \Omega_n$ and $\mathbb{P}(\Omega_n) < \infty$ for all $n = 1, 2, \ldots$

Thus, probability measures are a special type of finite measures and both are special type of σ-finite measures.

Note that $\mathbb{P}$ is just a map that assigns non-negative numbers to the sets in $\mathcal{F}$. Nothing in the properties of the map $\mathbb{P}$ does not prevent it from assigning 0 (or any other number) to some of the sets in $\mathcal{F}$. The fact that $\mathbb{P}(A) = 0$ does not imply that A is the empty set. A set $A \in \mathcal{F}$ such that $\mathbb{P}(A) = 0$ is called a *null set.*

Exercise 55. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space. Show that
(1) If $A_1, A_2, \ldots \in \mathcal{F}$ are null sets, then so is $\bigcup_{n=1}^{\infty} A_n$.
(2) If $A, B \in \mathcal{F}$ and A is a null set, then $\mathbb{P}(A \cup B) = \mathbb{P}(B)$.
(2) If $A, B \in \mathcal{F}$ and A is a null set, then $\mathbb{P}(A \cap B) = 0$.

Exercise 56. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space and let $A_1, A_2, \ldots \in \mathcal{F}$ be any sets. Show that
(1) $\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \lim_{m\to\infty} \mathbb{P}\Big(\bigcup_{n=1}^{m} A_n\Big)$.
(2) If $\mathbb{P}(A_1) < \infty$ then $\mathbb{P}\Big(\bigcap_{n=1}^{\infty} A_n\Big) = \lim_{m\to\infty} \mathbb{P}\Big(\bigcap_{n=1}^{m} A_n\Big)$.

Consider a sequence of subsets $A_1, A_2, \dots$ of Ω. It is trivial to see that the set

$$\bigcup_{n=1}^{\infty} A_n$$

consists of those ω's from Ω that are in *at least one* of the sets $\{A_n\}$. Similarly, the set

$$\bigcap_{n=1}^{\infty} A_n$$

consists of those ω's from Ω that are in *all* of the sets $\{A_n\}$.

Now, let $\operatorname{liminf} A_n$ be the set of those ω's from Ω that are in *all but finitely many* of the sets $\{A_n\}$. Let $\operatorname{limsup} A_n$ be the set of those ω's from Ω that are in *infinitely many* of the sets $\{A_n\}$. Clearly, if ω is in all but finitely many of the sets $\{A_n\}$, then it is in infinitely many of them, so

$$\operatorname{liminf} A_n \subseteq \operatorname{limsup} A_n.$$

We have the following representations.

Proposition 57. For any sequence of subsets $A_1, A_2, \dots$ of Ω we have the representations

$$\operatorname{liminf} A_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \quad \text{and} \quad \operatorname{limsup} A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Proof. We show the first equality. The second is left as an exercise.

Let $\omega \in \operatorname{liminf} A_n$. Then ω is in all but finitely many of the sets $A_1, A_2, \dots$. So, for some n large enough, ω is in every set in $A_n, A_{n+1}, \dots$, consequently $\omega \in \bigcap_{k=n}^{\infty} A_k$. Thus, $\omega \in \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$. This shows that $\operatorname{liminf} A_n \subseteq \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$.

Conversely, if $\omega \in \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$, then $\omega \in \bigcap_{k=n}^{\infty} A_k$ for some n. The last inclusion says that ω is in every one of the sets $A_n, A_{n+1}, \dots$. So, ω may not be in the sets $A_1, \dots, A_{n-1}$. These are finitely many, so $\omega \in \operatorname{liminf} A_n$.

Putting the two inclusions together, shows the equality. □

Note that if the sets $A_1, A_2, \dots$ are in a σ-algebra $\mathcal{F}$, then so are the sets $\operatorname{liminf} A_n$ and $\operatorname{limsup} A_n$. We do not use the notation $\operatorname*{liminf}_{n\to\infty} A_n$ in order to emphasize the fact that $\operatorname{liminf} A_n$ is a set and not a limit of some sorts. The reason why this set is given such a weird name is explained by Exercise 93.

Proposition 58 (Fatou lemma: special case)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathbb{P}(\Omega) = 1$ and let $A_1, A_2, \dots \in \mathcal{F}$. Then

$$\mathbb{P}\big(\operatorname{liminf} A_n\big) \leq \operatorname*{liminf}_{n\to\infty} \mathbb{P}(A_n) \leq \operatorname*{limsup}_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}\big(\operatorname{limsup} A_n\big).$$

Proof. Later we will present a theorem from which this proposition is a simple corollary. □

A sequence of sets $\{A_n\}$ is called *convergent* if $\operatorname{liminf} A_n = \operatorname{limsup} A_n =: A$. In that case, by the special case of the Fatou lemma, we have

$$\operatorname*{liminf}_{n\to\infty} \mathbb{P}(A_n) = \operatorname*{limsup}_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

Exercise 59. Let $A_1, A_2, \ldots, A_n$ be n events in Ω. Consider the sequence

$$A_1, A_2, \ldots, A_n, A_1, A_2, \ldots, A_n, A_1, A_2, \ldots, A_n, \ldots$$

What is liminf and limsup of that sequence? When is this sequence convergent?

Exercise 60. Let $A_1, A_2, \ldots, A_n$ be n events in Ω. Consider the sequence

$$A_1, A_2, \ldots, A_n, \emptyset, \emptyset, \emptyset, \ldots$$

What is liminf and limsup of that sequence?

Exercise 61. a) Let $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ be an increasing sequence of events in Ω. What is liminf and limsup of that sequence?
b) Let $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ be a decreasing sequence of events in Ω. What is liminf and limsup of that sequence?

Definition 62. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The events $\{A_i\}_{i \in I} \subseteq \mathcal{F}$, where I is an index set, are called *independent*, if for any choice of distinct indexes $i_1, \ldots, i_n \in I$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_n}). \tag{4}$$

Note that in the above definition, equality (4) has to hold for any $n \in \mathbb{N}$ and any distinct indexes $i_1, \ldots, i_n \in I$. Two events $A, B \in \mathcal{F}$ are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. While, three events $\{A, B, C\}$ are independent if $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, $\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$, and $\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$.

The following lemma is stated without a proof, since soon we will prove a more general result from which this lemma is a particular case.

Lemma 63. If the events $\{A_i\}_{i \in I} \subseteq \mathcal{F}$, where I is an index set, are independent, then so are their complements $\{A_i^c\}_{i \in I} \subseteq \mathcal{F}$.

Definition 64. Let $(\Omega, \mathcal{F}, \mathbb{P})$be a probability space and let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{for } A \in \mathcal{F}$$

is called the *conditional probability of A given B.*

Definition 65. We say that the subsets $A_1, A_2, \ldots, A_n$ of Ω form a *partition of* Ω if they are disjoint and $\Omega = \cup_{i=1}^n A_i$.

The following formula is well-known from undergraduate classes.

Theorem 66 (Bayes' formula)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose $A_1, A_2, \ldots, A_n$ form a partition of Ω. Then for any set $B \in \mathcal{F}$ we have

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(B|A_k)\mathbb{P}(A_k)}{\sum_{i=1}^n \mathbb{P}(B|A_i)\mathbb{P}(A_i)},$$

where we also need to require that $\mathbb{P}(B) > 0$, and $\mathbb{P}(A_i) > 0$ for all $i = 1, 2, \ldots, n$.

In the Bayes' theorem, an event A_k is called *hypothesis* and the probabilities $\mathbb{P}(A_k)$ are called *prior probabilities*. The probabilities $\mathbb{P}(A_k|B)$ are called *posterior probabilities* of A_k.

Next, we present the fundamental Lemma of Borel-Cantelli.

Proposition 67 (Borel-Cantelli lemma)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$be a probability space and $A_1, A_2, \ldots \in \mathcal{F}$.

(1) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup A_n) = 0$.

(2) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $\{A_n\}_{n=1}^{\infty}$ are independent, then $\mathbb{P}(\limsup A_n) = 1$.

Proof. (1) By definition we have $\limsup A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} B_n$, where we define $B_n := \bigcup_{k=n}^{\infty} A_k$. Clearly, $B_{n+1} \subseteq B_n$ for all n and by the continuity of $\mathbb{P}$ from above, Proposition 54, part (vi), we get

$$\mathbb{P}(\limsup A_n) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} B_n \Big) = \lim_{n \to \infty} \mathbb{P}(B_n) = \lim_{n \to \infty} \mathbb{P}\Big(\bigcup_{k=n}^{\infty} A_k \Big) \leq \lim_{n \to \infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0,$$

where the third equality follows from Exercise 56 and the last inequality follows again by Proposition (vi).

(2) Showing that $\mathbb{P}(\limsup A_n) = 1$ is equivalent to showing that $\mathbb{P}((\limsup A_n)^c) = 0$. Now

$$(\limsup A_n)^c = \Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \Big)^c = \bigcup_{n=1}^{\infty} \Big(\bigcup_{k=n}^{\infty} A_k \Big)^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c = \liminf A_n^c,$$

where we used twice Lemma 33. Letting $B_n := \cap_{k=n}^{\infty} A_k^c$ we get an increasing sequence $B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots$, implying, by the continuity below of $\mathbb{P}$, that

$$\mathbb{P}((\limsup A_n)^c) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \Big) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} B_n \Big) = \lim_{n \to \infty} \mathbb{P}(B_n),$$

where the last equality follows from Proposition 54, part (v). The proof will be done if we show that $\mathbb{P}(B_n) = 0$. So fix an $n \in \mathbb{N}$. By Lemma 63, since $A_1, A_2, \ldots$ are independent, then so are their complements $A_1^c, A_2^c, \ldots$. By Exercise 56, and by the independence, we have

$$\begin{aligned}
\mathbb{P}(B_n) = \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c \Big) &= \lim_{m \to \infty} \mathbb{P}\Big(\bigcap_{k=n}^{m} A_k^c \Big) = \lim_{m \to \infty} \big(\mathbb{P}(A_n^c) \mathbb{P}(A_{n+1}^c) \cdots \mathbb{P}(A_m^c) \big) \\
&= \lim_{m \to \infty} \big((1 - \mathbb{P}(A_n))(1 - \mathbb{P}(A_{n+1})) \cdots (1 - \mathbb{P}(A_m)) \big) \\
&\leq \lim_{m \to \infty} \big(e^{-\mathbb{P}(A_n)} e^{-\mathbb{P}(A_{n+1})} \cdots e^{-\mathbb{P}(A_m)} \big) \\
&= \lim_{m \to \infty} e^{-\sum_{k=n}^{m} \mathbb{P}(A_k)} \\
&= e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)} \\
&= e^{-\infty} \\
&= 0,
\end{aligned}$$

where we used the inequality $1 - x \leq e^{-x}$ valid for every $x \in \mathbb{R}$. □

2.2.1 How to construct measures?

In general, it is difficult to construct measures on a σ-algebra explicitly, since we do not have explicit formula for the sets in the σ-algebra. Instead, we construct measures on algebras and use the next theorem, which we state without proof, to extend it to σ-algebras. Informally speaking, algebras are simpler than σ-algebras, and contain "fewer" sets, thus it should be easier to construct measures on algebras. First we need a definition.

Definition 68. Suppose $\mathcal{F}$ is a algebra.
A map $\mathbb{P}_0 : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$ is called a *measure on the algebra* $\mathcal{F}$ if
1) $\mathbb{P}_0(\emptyset) = 0$
2) $\mathbb{P}_0(A) \geq 0$ for all $A \in \mathcal{F}$; and
3) For any sequence of *disjoint* events $A_1, A_2, \ldots \in \mathcal{F}$, such that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$, we have

$$\mathbb{P}_0\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}_0(A_i). \tag{5}$$

Notice the very small difference between a measure on a σ-algebra and a measure on an algebra. We know that, if $\mathcal{F}$ is an algebra then the union of a sequence of sets from $\mathcal{F}$ may not be in $\mathcal{F}$. Thus, in order for the value $\mathbb{P}_0\Big(\bigcup_{i=1}^{\infty} A_i\Big)$ to be defined, we need to require that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- If $\mathbb{P}_0(\Omega) < \infty$ then the measure $\mathbb{P}_0$ is called *finite*.
- The measure $\mathbb{P}_0$ is called *σ-finite* if there is a sequence of sets $\Omega_1 \subseteq \Omega_2 \subseteq \Omega_3 \subseteq \cdots$ such that $\Omega = \bigcup_{n=1}^{\infty} \Omega_n$ and $\mathbb{P}_0(\Omega_n) < \infty$ for all $n = 1, 2, 3 \ldots$

Theorem 69 (Carathéodory's extension theorem)**.** Let $\mathbb{P}_0$ be a finite (resp. σ-finite) measure on an algebra $\mathcal{F}$. Then $\mathbb{P}_0$ has a unique extension to a finite (resp. σ-finite) measure $\mathbb{P}$ on $\sigma(\mathcal{F})$, the σ-algebra generated by $\mathcal{F}$. That is, for all $A \in \mathcal{F}$ we have

$$\mathbb{P}(A) = \mathbb{P}_0(A). \tag{6}$$

Note that if $\mathbb{P}_0(\Omega) = 1$, then its extension $\mathbb{P}$ is a probability measure: just replace A by Ω in (6).

Exercise 70. *Let $\{a_{i,j} : i = 1, 2, \ldots, j = 1, 2, \ldots\}$ be an array of positive numbers, such that $a_{i,j} \leq a_{i+1,j}$ for all i, j. Show*

$$\lim_{i\to\infty} \Big(\sum_{j=1}^{\infty} a_{i,j}\Big) = \sum_{j=1}^{\infty} \big(\lim_{i\to\infty} a_{i,j}\big).$$

Exercise 71. *Let $\{a_{i,j} : i = 1, 2, \ldots, j = 1, 2, \ldots\}$ be an array of positive numbers.*

(a) Suppose that $\lim_{j\to\infty} a_{i,j}$ exists for every $i = 1, 2, \ldots$ and suppose that $a_{i,j} \leq a_{i+1,j}$ for all $i, j = 1, 2, \ldots$ Show that

$$\operatorname*{liminf}_{j\to\infty}(\lim_{i\to\infty} a_{i,j}) \geq \lim_{i\to\infty}(\lim_{j\to\infty} a_{i,j}).$$

(b) Suppose that $a_{i,j} \leq a_{i+1,j}$ and that $a_{i,j} \leq a_{i,j+1}$ for all i, j. That is, for every j the sequence $\{a_{i,j}\}_{i=1}^{\infty}$ is increasing and for every i the sequence $\{a_{i,j}\}_{j=1}^{\infty}$ is increasing. Then

$$\lim_{i\to\infty} \big(\lim_{j\to\infty} a_{i,j}\big) = \lim_{j\to\infty} \big(\lim_{i\to\infty} a_{i,j}\big).$$

Let $F : \mathbb{R} \to \mathbb{R}$ be a function with the following properties: (1) F is increasing, and (2) F is right continuous, that is $\lim_{y\to x^+} F(y) = F(x)$ for all $x \in \mathbb{R}$.

We are going to use the Carathéodory extension theorem to see how such a function F defines a measure on the Borel sets of $\mathbb{R}$. Consider the collection $\mathcal{F}$ of all sets $A \subseteq \mathbb{R} \cup \{\infty\}$ that can be written as

$$A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n],$$

where $-\infty \le a_1 \le b_1 \le \cdots \le a_n \le b_n \le \infty$ with the convention $(a, a] = \emptyset$. Example 36 shows that $\mathcal{F}$ is an algebra. Define the function

$$\mathbb{P}_0 : \mathcal{F} \to \mathbb{R} \cup \{\infty\}$$

that assigns to every set $A \in \mathcal{F}$ of the form

$$A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n],$$

for some $-\infty \le a_1 \le b_1 \le \cdots \le a_n \le b_n \le \infty$, the number

$$\mathbb{P}_0(A) := \sum_{i=1}^{n} \big(F(b_i) - F(a_i)\big). \tag{7}$$

(Think about the fact that the number $\mathbb{P}_0(A)$ does not depend on how you represent a set A as a disjoint union of intervals $(a, b]$.) If the interval $(a, b]$ is infinite, we define

$$\begin{aligned} \mathbb{P}_0((-\infty, b]) &:= F(b) - \lim_{x\to-\infty} F(x), \\ \mathbb{P}_0((a, \infty]) &:= \lim_{x\to\infty} F(x) - F(a), \text{ and} \\ \mathbb{P}_0((-\infty, \infty]) &:= \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x). \end{aligned}$$

In order to use the Carathéodory theorem, we have to show that $\mathbb{P}_0$ is a measure on the algebra $\mathcal{F}$. That is, we need to check that the three conditions in Definition 68 hold.

1) $\mathbb{P}_0(\emptyset) = \mathbb{P}_0((a, a]) = F(a) - F(a) = 0$.

2) The fact that $\mathbb{P}_0(A) \ge 0$ for all $A \in \mathcal{F}$, follows immediately from definition (7) and the fact that F is increasing function, that is $F(b_i) - F(a_i) \ge 0$ since $a_i \le b_i$ for all $i = 1, 2, \ldots, n$.

The verification of the third condition is more involved and is the content of the next two lemmas. The first lemma considers a special case while the second one tackles the general one.

Lemma 72. *Let $(c_i, d_i]$, $i = 1, 2, \ldots$ be disjoint intervals such that*

$$(a, b] = \bigcup_{i=1}^{\infty} (c_i, d_i],$$

where $-\infty \le a \le b \le \infty$. Then,

$$\mathbb{P}_0((a, b]) = \sum_{i=1}^{\infty} \mathbb{P}_0((c_i, d_i]). \tag{8}$$

Proof. If $a = b$ then $(a, b]$ is the empty set and so all sets $(c_i, d_i]$ have to be empty, i.e. $c_i = d_i$ for all i, and there is nothing to show. So assume that $a < b$. We consider two cases, based on whether or not the interval $(a, b]$ is finite or not.

Case 1. Suppose the interval $(a, b]$ is finite. The right continuity of F implies that for any $\epsilon > 0$ there is a $\delta > 0$ so that

$$F(a + \delta) < F(a) + \epsilon. \tag{9}$$

Similarly, for any $i = 1, 2, 3, \ldots$ (replace now ϵ by $\epsilon/2^i$) there is an $\eta_i > 0$ such that

$$F(d_i + \eta_i) < F(d_i) + \frac{\epsilon}{2^i}. \tag{10}$$

Observe now that

$$[a + \delta, b] \subseteq \bigcup_{i=1}^{\infty} (c_i, d_i + \eta_i).$$

That is the closed and bounded interval $[a + \delta, b]$ is covered by open intervals. By a theorem in real analysis,[1] which we will not discuss, $[a + \delta, b]$ is also covered by a *finite* number of these intervals. Suppose for simplicity of notation, that $[a + \delta, b]$ is covered by the first n open intervals, that is

$$[a + \delta, b] \subseteq \bigcup_{i=1}^{n} (c_i, d_i + \eta_i).$$

Thus, using (10) and the fact that F is increasing, we obtain

$$\begin{aligned} F(b) - F(a + \delta) &\leq \sum_{i=1}^{n} \big(F(d_i + \eta_i) - F(c_i)\big) < \sum_{i=1}^{n} \Big(F(d_i) - F(c_i) + \frac{\epsilon}{2^i}\Big) \\ &< \sum_{i=1}^{n} \big(F(d_i) - F(c_i)\big) + \epsilon \leq \sum_{i=1}^{\infty} \big(F(d_i) - F(c_i)\big) + \epsilon \\ &= \sum_{i=1}^{\infty} \mathbb{P}_0((c_i, d_i]) + \epsilon. \end{aligned}$$

Finally, using (9) and combining with the above, we obtain

$$\mathbb{P}_0((a, b]) = F(b) - F(a) < F(b) - F(a + \delta) + \epsilon < \sum_{i=1}^{\infty} \mathbb{P}_0((c_i, d_i]) + 2\epsilon.$$

This shows, that for every $\epsilon > 0$ we have $\mathbb{P}_0((a, b]) \leq \sum_{i=1}^{\infty} \mathbb{P}_0((c_i, d_i]) + 2\epsilon$, which is equivalent to

$$\mathbb{P}_0((a, b]) \leq \sum_{i=1}^{\infty} \mathbb{P}_0((c_i, d_i]). \tag{11}$$

[1] The theorem says that if a closed and bounded set is covered by open sets, then only a finite number of the open sets also cover it.

To show the opposite inequality, we observe that for any N we have $\cup_{i=1}^{N}(c_i,d_i] \subseteq (a,b]$. Using the fact that F is increasing function and that the intervals $\{(c_i,d_i]\}_{i=1}^{\infty}$ are disjoint, we obtain

$$\mathbb{P}_0((a,b]) = F(b) - F(a) \geq \sum_{i=1}^{N} F(d_i) - F(c_i) = \sum_{i=1}^{N} \mathbb{P}_0((c_i,d_i]).$$

(Note that to obtain the last inequality, we cannot use the properties of a measure, because we have not proved that $\mathbb{P}_0$ is a measure yet.) Letting N approach infinity we get

$$\mathbb{P}_0((a,b]) \geq \sum_{i=1}^{\infty} \mathbb{P}_0((c_i,d_i]).$$

Combine this with (11) to establish (14).

Case 2. Suppose that $(a,b]$ is infinite.

For any $n \in \mathbb{N}$, we have

$$(a,b] \cap (-n,n] = \Big(\bigcup_{i=1}^{\infty}(c_i,d_i]\Big) \cap (-n,n] = \bigcup_{i=1}^{\infty}\big((c_i,d_i] \cap (-n,n]\big).$$

Note that sets in the union on the right are disjoint, that all intersections such as $(a,b] \cap (-n,n]$ are half-open bounded intervals. So by the first case, we obtain

$$\mathbb{P}_0((a,b] \cap (-n,n]) = \sum_{i=1}^{\infty} \mathbb{P}_0((c_i,d_i] \cap (-n,n])$$

for all $n = 1,2,3,\ldots$ Let n approach infinity to obtain[2]

$$\begin{aligned}\mathbb{P}_0((a,b]) = \lim_{n\to\infty} \mathbb{P}_0((a,b] \cap (-n,n]) &= \lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{P}_0((c_i,d_i] \cap (-n,n]) \\ &= \sum_{i=1}^{\infty} \lim_{n\to\infty} \mathbb{P}_0((c_i,d_i] \cap (-n,n]) \\ &= \sum_{i=1}^{\infty} \mathbb{P}_0((c_i,d_i]).\end{aligned}$$

The justification for exchanging the limit with the infinite sum is given in Exercise 70. That establishes (14) in the second case. □

[2] Caution is needed when establishing the first equality. One cannot use the property continuity from below since we have not established that $\mathbb{P}_0$ is a measure yet. Suppose the interval $(a,b]$ is of the form $(-\infty,b]$, where b is a finite number. Then for n larger than b, we have $(a,b] \cap (-n,n] = (-n,b]$, and we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}_0((a,b] \cap (-n,n]) = \lim_{n\to\infty} \mathbb{P}_0((-n,b]) = \lim_{n\to\infty}(F(b) - F(-n)) &= F(b) - \lim_{n\to\infty} F(-n) \\ &= F(b) - \lim_{x\to-\infty} F(x) = \mathbb{P}_0((-\infty,b]) = \mathbb{P}_0((a,b]).\end{aligned}$$

Similar considerations are needed in order to establish $\lim_{n\to\infty} \mathbb{P}_0((c_i,d_i] \cap (-n,n]) = \mathbb{P}_0((c_i,d_i])$.

We now verify that the third condition of Definition 68 holds.

Lemma 73. *For any sequence of disjoint events $A_1, A_2, \ldots \in \mathcal{F}$, such that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$, we have*

$$\mathbb{P}_0\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}_0(A_i). \tag{12}$$

Proof. Let

$$A := \bigcup_{i=1}^{\infty} A_i.$$

On the one hand, since $A \in \mathcal{F}$, the set A is a union of finitely many half-open disjoint intervals

$$A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n].$$

On the other hand, since $A_i \in \mathcal{F}$, the set A_i is a union of finitely many (possibly different) half-open disjoint intervals

$$A_i = (a_1^i, b_1^i] \cup (a_2^i, b_2^i] \cup \cdots \cup (a_{n_i}^i, b_{n_i}^i], \quad i = 1, 2, 3, \ldots \tag{13}$$

(Note that the number n_i of half-open intervals that constitute A_i may vary for different sets in the sequence $A_1, A_2, \ldots$) Consider the collection $\mathcal{C}$ of all half-open intervals that participate in (13) and note that they are all disjoint since the sets $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint. Since

$$A = \bigcup_{i=1}^{\infty} A_i = \bigcup_{(c,d] \in \mathcal{C}} (c, d],$$

we can partition the intervals in $\mathcal{C}$ into n groups, $\mathcal{C}_1, \ldots, \mathcal{C}_n$ such that the union of the intervals in the first group is $(a_1, b_1]$, the union of the intervals in the second group is $(a_2, b_2]$ and so on.

Suppose the intervals in the first group are $\mathcal{C}_1 = \{(c_i, d_i] : i = 1, 2, 3, \ldots\}$. Since their union is

$$(a_1, b_1] = \bigcup_{i=1}^{\infty} (c_i, d_i],$$

by Lemma 72 we have

$$\mathbb{P}_0((a_1, b_1]) = \sum_{i=1}^{n} \mathbb{P}_0((c_i, d_i]). \tag{14}$$

Similarly for the intervals in all n groups. We get

$$\mathbb{P}_0(A) = \sum_{i=1}^{n} \mathbb{P}_0((a_i, b_i]) = \sum_{i=1}^{n} \sum_{(c,d] \in \mathcal{C}_i} \mathbb{P}_0((c, d]) = \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \mathbb{P}_0((a_j^i, b_j^i]) = \sum_{i=1}^{\infty} \mathbb{P}_0(A_i),$$

where the third inequality holds because the two double sums go over all the intervals in the collection $\mathcal{C}$ in two different ways. The fourth inequality holds since

$$\sum_{j=1}^{n_i} \mathbb{P}_0((a_j^i, b_j^i]) = \mathbb{P}_0(A_i)$$

by definition of $\mathbb{P}_0$. □

This concludes the proof that $\mathbb{P}_0$ is a measure on $\mathcal{F}$. Now, notice that this measure is always finite or σ-finite (why?). Thus, the Carathéodory's extension theorem now allows us to extend the measure $\mathbb{P}_0$ on $\mathcal{F}$, in a unique way, to a measure P on $\sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R})$. We formulate this in the next theorem.

Theorem 74. *For any increasing, right-continuous function $F : \mathbb{R} \to \mathbb{R}$, there is a unique measure $\mathbb{P} : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ such that*

$$\mathbb{P}((a, b]) = F(b) - F(a)$$

for all $-\infty < a \leq b < \infty$.

The next exercise summarizes how to calculate the measure of various different Borel subsets of $\mathbb{R}$. Recall that an increasing function has left limit at every point (which may not be equal to the value of the function at that point). For any $b \in \mathbb{R}$, denote that left limit by

$$F(b-) := \lim_{x \to b^-} F(x).$$

Exercise 75. Let $\mathbb{P}$ be the measure in Theorem 74. Show that

(i) $\mathbb{P}((-\infty, b)) = F(b-) - \lim_{x \to -\infty} F(x)$;

(ii) $\mathbb{P}((a, b)) = F(b-) - F(a)$;

(iii) $\mathbb{P}(\{a\}) = F(a) - F(a-)$;

(iv) $\mathbb{P}([a, b)) = F(b-) - F(a-)$;

(v) $\mathbb{P}([a, b]) = F(b) - F(a-)$.

(vi) $\mathbb{P}((a, \infty)) = \lim_{x \to \infty} F(x) - F(a)$;

(vii) $\mathbb{P}([a, \infty)) = \lim_{x \to \infty} F(x) - F(a-)$;

(viii) If, in addition, the function F is continuous, then

  (a) $\mathbb{P}((-\infty, b]) = \mathbb{P}((-\infty, b)) = F(b) - \lim_{x \to -\infty} F(x)$;

  (b) $\mathbb{P}((a, b)) = \mathbb{P}([a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b]) = F(b) - F(a)$;

  (c) $\mathbb{P}(\{a\}) = 0$.

(ix) If $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, show that $\mathbb{P}$ is a probability measure.

The next exercise shows that all finite measures on $\mathcal{B}(\mathbb{R})$ are obtained from an increasing, right-continuous function F as described in Theorem 74.

Exercise 76. Let $\mathbb{Q}$ be a finite measure on $\mathcal{B}(\mathbb{R})$. Define the function $F(x) := \mathbb{Q}((-\infty, x])$. Show that F is increasing, right-continuous, and $\mathbb{Q}((a, b]) = F(b) - F(a)$ for all $-\infty < a \leq b < \infty$.

Example 77 (Lebesgue measure)**.** Let $F : \mathbb{R} \to \mathbb{R}$ be the identity function $F(x) = x$. It is clearly increasing and continuous. The unique measure $\mathbb{P}$ on $\mathcal{B}(\mathbb{R})$, satisfying

$$\mathbb{P}((a, b]) = b - a$$

is called the *Lebesgue* measure on $\mathbb{R}$. It is a σ-finite measure. □

2.2.2 Product measure

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two measure spaces (either both finite or both σ-finite). We construct the product space

$$(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$$

as follows.

(1) $\Omega_1 \times \Omega_2 := \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$. Thus, $\Omega_1 \times \Omega_2$ is the set of all ordered pairs of elements from Ω_1 and Ω_2 much like $\mathbb{R}^2$ is the set of all ordered pairs (x, y) of numbers from $\mathbb{R}$ and $\mathbb{R}$.

(2) Consider the collection of subsets of $\Omega_1 \times \Omega_2$ of the form

$$A \times B := \{(\omega_1, \omega_2) : \omega_1 \in A, \omega_2 \in B\} \text{ with } A \in \mathcal{F}_1, B \in \mathcal{F}_2.$$

The set $A \times B$ is called a rectangle in $\Omega_1 \times \Omega_2$ with sides A and B (much like the rectangles in $\mathbb{R}^2$ with sides parallel to the coordinate axis). Denote by $\mathcal{F}_1 \otimes \mathcal{F}_2$ the σ-algebra on $\Omega_1 \times \Omega_2$ generated by this collection of subsets.

(3) Let $\mathcal{G}$ be the collection of all subsets of $\Omega_1 \times \Omega_2$ that can be represented as a finite, disjoint union of rectangles:

$$(A_1 \times B_1) \cup (A_2 \times B_2) \cup \cdots \cup (A_n \times B_n),$$

where $A_i \in \mathcal{F}_1$, $B_i \in \mathcal{F}_2$, and $(A_i \times B_i) \cap (A_j \times B_j) = \emptyset$ for all $i \neq j$. Finally, define the function $\mathbb{P}_0 : \mathcal{G} \to [0, \infty]$ by

$$\mathbb{P}_0((A_1 \times B_1) \cup (A_2 \times B_2) \cup \cdots \cup (A_n \times B_n)) := \sum_{i=1}^{n} \mathbb{P}_1(A_i)\mathbb{P}_2(B_i).$$

The next proposition introduces the properties of $\mathcal{G}$ and $\mathbb{P}_0$.

Proposition 78. The collection of sets $\mathcal{G}$ is algebra. The map $\mathbb{P}_0 : \mathcal{G} \to [0, \infty]$ is a measure on $\mathcal{G}$.

Note that if both $\mathbb{P}_1$ and $\mathbb{P}_2$ are finite measures then so is $\mathbb{P}_0$ since by its definition $\mathbb{P}_0(\Omega_1 \times \Omega_2) = \mathbb{P}_1(\Omega_1)\mathbb{P}_2(\Omega_2) < \infty$.

If both $\mathbb{P}_1$ and $\mathbb{P}_2$ are σ-finite measures then so is $\mathbb{P}_0$. Indeed, let $\Omega_i^1 \subseteq \Omega_i^2 \subseteq \Omega_i^3 \subseteq \cdots$ be such that $\mathbb{P}_i(\Omega_i^k) < \infty$ and $\Omega_i = \cup_{k=1}^{\infty} \Omega_i^k$ for $i = 1, 2$. Then

$$\Omega_1^1 \times \Omega_2^1 \subseteq \Omega_1^2 \times \Omega_2^2 \subseteq \Omega_1^3 \times \Omega_2^3 \subseteq \cdots,$$

$$\Omega_1 \times \Omega_2 = \bigcup_{k=1}^{\infty} \Omega_1^k \times \Omega_2^k,$$

and for all $k = 1, 2, 3 \ldots$, by definition, we have $\mathbb{P}_0(\Omega_1^k \times \Omega_2^k) = \mathbb{P}_1(\Omega_1^k)\mathbb{P}_2(\Omega_2^k) < \infty$.

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two finite (resp. σ-finite) measure spaces. By the Carathéodory extension theorem, there is a unique finite (resp. σ-finite) measure extending $\mathbb{P}_0$ to the σ-algebra generated by $\mathcal{G}$. That measure is denoted by $\mathbb{P}_1 \times \mathbb{P}_2$ and is called *product measure*. The measure space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ is the *product measure space*. In other words, $\mathbb{P}_1 \times \mathbb{P}_2$ is the unique measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$ satisfying

$$(\mathbb{P}_1 \times \mathbb{P}_2)(A \times B) = \mathbb{P}_1(A)\mathbb{P}_2(B) \text{ for all } A \in \mathcal{F}_1 \text{ and } B \in \mathcal{F}_2.$$

Given a third measure space $(\Omega_3, \mathcal{F}_3, \mathbb{P}_3)$, it can be shown that

$$(\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes \mathcal{F}_3 = \mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes \mathcal{F}_3) \text{ and } (\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3 = \mathbb{P}_1 \times (\mathbb{P}_2 \times \mathbb{P}_3).$$

That is, the product between measure spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2, 3$ is associative. We may first multiply $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ and then multiply the result $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ by $(\Omega_3, \mathcal{F}_3, \mathbb{P}_3)$ or we may first multiply $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ and $(\Omega_3, \mathcal{F}_3, \mathbb{P}_3)$ and then multiply $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ by the result $(\Omega_2 \times \Omega_3, \mathcal{F}_2 \otimes \mathcal{F}_3, \mathbb{P}_2 \times \mathbb{P}_3)$. Thus, we denote the final result by

$$(\Omega_1 \times \Omega_2 \times \Omega_3, \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \mathcal{F}_3, \mathbb{P}_1 \times \mathbb{P}_2 \times \mathbb{P}_3).$$

Iterating this procedure, we can define *finite* products

$$(\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots \otimes \mathcal{F}_n, \mathbb{P}_1 \times \mathbb{P}_2 \times \cdots \times \mathbb{P}_n).$$

This allows us to define the *Borel σ-algebra on $\mathbb{R}^n$* by

$$\mathcal{B}(\mathbb{R}^n) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R}), \quad n \text{ times}.$$

Proposition 79. *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. Let $\mathcal{G}_1$ be a collection of subsets of Ω_1 that generates $\mathcal{F}_1$, that is $\sigma(\mathcal{G}_1) = \mathcal{F}_1$. Let $\mathcal{G}_2$ be a collection of subsets of Ω_2 that generates $\mathcal{F}_2$, that is $\sigma(\mathcal{G}_2) = \mathcal{F}_2$. Then, the collection of rectangles*

$$\{A \times B : A \in \mathcal{G}_1, B \in \mathcal{G}_2\}$$

generates the σ-algebra $\mathcal{F}_1 \otimes \mathcal{F}_2$.

Proposition 79 has a natural analogue for the product of finitely many σ-algebras. Thus, using Proposition 44 we obtain different collections of sets with the property that each collection generates $\mathcal{B}(\mathbb{R}^n)$. For example,

$$\mathcal{B}(\mathbb{R}^n) = \sigma\big(\{(a_1, b_1) \times \cdots \times (a_n, b_n) : -\infty < a_i < b_i < \infty \text{ for all } i = 1, 2, \ldots, n\}\big),$$
$$\mathcal{B}(\mathbb{R}^n) = \sigma\big(\{(-\infty, b_1) \times \cdots \times (-\infty, b_n) : -\infty < b_i < \infty \text{ for all } i = 1, 2, \ldots, n\}\big).$$

It is a fact that every open set in $\mathbb{R}^n$ is a countable union of n-dimensional open rectangles $(a_1, b_1) \times \cdots \times (a_n, b_n)$, hence the Borel σ-algebra on $\mathbb{R}^n$ contains (and is generated by) all open sets in $\mathbb{R}^n$.

2.2.3 Atomic measures (optional)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space. A set $A \in \mathcal{F}$ is called an *atom* if $\mathbb{P}(A) > 0$ and for any subset $B \subseteq A$, $B \in \mathcal{F}$, with $\mathbb{P}(A) > \mathbb{P}(B)$, one has $\mathbb{P}(B) = 0$.

Example 80. Consider the set $\Omega = \{1, 2, \ldots, n\}$ with $\mathcal{F} = 2^\Omega$ and $\mathbb{P}(A) = |A|$ for all $A \in \mathcal{F}$. Then, each set $A = \{k\}$, with one element, is an atom.

Consider the set $\Omega = \mathbb{R}$ with $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P}$ = the Lebesgue measure. This measure space has no atoms.

Consider the set $\Omega = \mathbb{R}$ with $\mathbb{P}$ = the Lebesgue measure. Let $A \in \mathcal{B}(\mathbb{R})$ be such that $\mathbb{P}(A) > 0$. Consider the sigma algebra $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$. Then, the measure space $(\Omega, \mathcal{F}, \mathbb{P})$ has atoms (both A and A^c).

A measure space $(\Omega, \mathcal{F}, \mathbb{P})$ is *non-atomic* if it has no atoms. In other words, a measure space is non-atomic if for any set $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$, there exists a subset $B \subseteq A$, $B \in \mathcal{F}$ such that $\mathbb{P}(A) > \mathbb{P}(B) > 0$.

The examples show that the existence of atoms depends not only on the measure but also on the sigma algebra. The goal of this section is to present the following theorem.

Theorem 81 (Sierpiński)**.** Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is non-atomic measure space. Chose $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ and let $a := \mathbb{P}(A)$. The, for any $t \in [0, a]$ there is a set $A_t \in \mathcal{F}$ such that

(i) $A_0 = \emptyset$ and $A_a = A$;

(ii) $A_{t_1} \subseteq A_{t_2}$ for all $0 \le t_1 \le t_2 \le a$; and

(iii) $\mathbb{P}(A_t) = t$ for all $0 \le t \le a$.

3 Random variables

Given a measurable space $(\Omega, \mathcal{F})$, a subset $A \subseteq \Omega$ is called *measurable* if $A \in \mathcal{F}$. Clearly, every subset of Ω is measurable if and only if $\mathcal{F} = 2^\Omega$. Otherwise, there are subsets of Ω that are not measurable.

Definition 82. Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be two measurable spaces. The function $X : \Omega \to S$ is called *measurable* if

$$X^{-1}(B) \in \mathcal{F} \quad \text{for all } B \in \mathcal{S}. \tag{15}$$

If $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then X is called a *random variable.*

Thus, $X : \Omega \to S$ is a measurable function if the preimage of every measurable subset of S is a measurable subset of Ω. A random variable is simply a measurable function into the real numbers with the Borel σ-algebra. Note that we do not need a measure on $\mathcal{F}$ (or on $\mathcal{S}$) to define a random variable. Sometimes, instead of $X : \Omega \to S$ we write $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ when we want to emphasize the σ-algebras. When $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ we just say that X is a random variable on $(\Omega, \mathcal{F})$.

Example 83. (a) The function $X(\omega) = 2$ for all $\omega \in \Omega$ is a random variable.

(b) If $A \in \mathcal{F}$, then the function

$$X(\omega) := \begin{cases} 5 & \text{if } \omega \in A, \\ 2 & \text{if } \omega \notin A \end{cases}$$

is a random variable.

(c) If $A \in \mathcal{F}$, then the function

$$X(\omega) := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A \end{cases}$$

is a random variable. This special case of example (b), preceding it, is called *indicator random variable* or the *indicator of* A. It is denoted by $\mathbf{1}_A(\omega)$.

(d) If $A, B \in \mathcal{F}$, then the function

$$X(\omega) := \begin{cases} 5 & \text{if } \omega \in A \setminus B, \\ 3 & \text{if } \omega \in A \cap B, \\ 2 & \text{if } \omega \in B \setminus A, \\ 1 & \text{if } \omega \notin A \cup B \end{cases}$$

is a random variable. Note that this random variable can be expressed as a linear combination of indicator random variables:

$$X(\omega) = 5 \cdot \mathbf{1}_{A\setminus B}(\omega) + 3 \cdot \mathbf{1}_{A\cap B}(\omega) + 2 \cdot \mathbf{1}_{B\setminus A}(\omega) + \mathbf{1}_{(A\cup B)^c}(\omega).$$

Of course, this representation may not be unique, for example we also have

$$X(\omega) = \mathbf{1}_{\Omega}(\omega) + \mathbf{1}_A(\omega) + \mathbf{1}_B(\omega) + 3 \cdot \mathbf{1}_{A\setminus B}(\omega).$$ □

The richness of the collection of random variables on $(\Omega, \mathcal{F})$ very much depends on the richness of the σ-algebra $\mathcal{F}$. If $\mathcal{F}$ contains a few sets, then there are only a "few" random variables on $(\Omega, \mathcal{F})$, and vice versa.

Example 84. What are the random variables on $(\Omega, \mathcal{F})$ when $\mathcal{F} = \{\emptyset, \Omega\}$? Let X be a random variable on $(\Omega, \mathcal{F})$. Suppose X takes two different values, say a and b. Then, $X^{-1}(\{a\})$ is a proper subset of Ω and hence not in $\mathcal{F}$. (Proper subset means that it is not $\emptyset$ nor Ω.) This contradiction shows that X is a constant function.

What are the random variables on $(\Omega, \mathcal{F})$ when $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$? Suppose X takes two different values, say a and b. Then, $X^{-1}(\{a\})$ is a proper subset of Ω and has to be in $\mathcal{F}$, so without loss of generality, it is A and then $X^{-1}(\{b\}) = A^c$. One can see that X cannot take a third value, different from a and b. So

$$X = a\mathbf{1}_A + b\mathbf{1}_{A^c}.$$ □

Checking condition (15) looks daunting since there may be many sets B in $\mathcal{S}$. The next proposition simplifies the task.

Proposition 85. If the collection of sets $\mathcal{S}_0$ generates $\mathcal{S}$ (that is, $\sigma(\mathcal{S}_0) = \mathcal{S}$), then $X : \Omega \to S$ is measurable if and only if

$$X^{-1}(B) \in \mathcal{F} \quad \text{for all } B \in \mathcal{S}_0. \tag{16}$$

Proof. Clearly, if X is a measurable function then it satisfies (16). Suppose on the opposite that X satisfies (16). We need to show that is satisfies (15). Let M be the collections of all sets $B \in \mathcal{S}$ such that $X^{-1}(B) \in \mathcal{F}$. By (16) we have $\mathcal{S}_0 \subseteq M$. We show now that M is a σ-algebra. This would imply $\mathcal{S} = \sigma(\mathcal{S}_0) \subseteq M$, showing (15).

Since $X^{-1}(\emptyset) = \emptyset \in \mathcal{F}$ we get $\emptyset \in M$. Recall the properties in Subsection 1.7. If $B \in M$, then $X^{-1}(B^c) = (X^{-1}(B))^c \in \mathcal{F}$, since $\mathcal{F}$ is a σ-algebra, hence $B^c \in M$. If $B_1, B_2, \ldots \in M$, then $X^{-1}\big(\bigcup_{j=1}^{\infty} B_j\big) = \bigcup_{j=1}^{\infty} X^{-1}(B_j) \in \mathcal{F}$, since $\mathcal{F}$ is a σ-algebra, we have $\bigcup_{j=1}^{\infty} B_j \in M$. □

Combining Proposition 44 and Proposition 85, we see that in order to check that $X : \Omega \to \mathbb{R}$ is a random variable we need to check that $X^{-1}((a,b)) \in \mathcal{F}$ for every $-\infty < a \leq b < \infty$. Or, we need to check only that $X^{-1}((-\infty, b]) \in \mathcal{F}$ for every $b \in \mathbb{R}$.

Proposition 86. If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$, and $Y : (S, \mathcal{S}) \to (T, \mathcal{T})$ are measurable functions then the composition $Y \circ X : (\Omega, \mathcal{F}) \to (T, \mathcal{T})$ is a measurable function.

Proof. Exercise. □

Proposition 87. If $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F})$, then the map $(X_1, X_2, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is measurable.

Proof. The Borel σ-algebra on $\mathbb{R}^n$ is generated by the rectangles $A_1 \times A_2 \times \cdots \times A_n$ where A_i is a Borel set on $\mathbb{R}$. Then, in view of Proposition 85, we need to check that the preimage of $A_1 \times A_2 \times \cdots \times A_n$ under $(X_1, X_2, \ldots, X_n)$ is in $\mathcal{F}$. Indeed,

$$\begin{aligned}\{\omega \in \Omega : (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \in A_1 \times A_2 \times \cdots \times A_n\} &= \bigcap_{i=1}^{n} \{\omega \in \Omega : X_i(\omega) \in A_i\} \\ &= \bigcap_{i=1}^{n} X_i^{-1}(A_i) \in \mathcal{F},\end{aligned}$$

since X_i are measurable and $\mathcal{F}$ is a σ-algebra. □

Corollary 88. If $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F})$ and if $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable, then $f(X_1, X_2, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a random variable.

Proof. Note that $f(X_1, X_2, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the composition of the measurable map $(X_1, X_2, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with the measurable function $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Apply Proposition 86. □

Corollary 89. If $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F})$ then so is $X_1 + X_2 + \cdots + X_n$.

Proof. The function $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by $f(x_1, x_2, \ldots, x_n) := x_1 + x_2 + \cdots + x_n$ is measurable. Indeed, the sets $(-\infty, b)$, $b \in \mathbb{R}$ generate the Borel σ-algebra on $\mathbb{R}$. Then, in view of Proposition 85, we need to check that the preimage $f^{-1}((-\infty, b)) = \{x \in \mathbb{R}^n : x_1 + x_2 + \cdots + x_n < b\}$ is in $\mathcal{B}(\mathbb{R}^n)$. But the last set is open and $\mathcal{B}(\mathbb{R}^n)$ contains all open sets in $\mathbb{R}^n$, so we are done. □

It is a fact, that we are not going to prove, that if a function $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is continuous then the preimage of every open set in $\mathbb{R}$ is an open set in $\mathbb{R}^n$. Hence, every continuous function is measurable. This together with Proposition 86 shows that if X is a random variable, then so is cX for all $c \in \mathbb{R}$, X^2, $\sin(X)$, and so on.

Exercise 90. If X and Y are random variables on $(\Omega, \mathcal{F})$, then so are XY, X/Y provided that $Y \neq 0$, and $|X|$.

We say that the sequence of functions $\{X_n\}$ is increasing if $X_n(\omega) \leq X_{n+1}(\omega)$ for all $n = 1, 2, \ldots$ and all $\omega \in \Omega$. Similarly, $\{X_n\}$ is decreasing if $X_n(\omega) \geq X_{n+1}(\omega)$ for all $n = 1, 2, \ldots$ and all $\omega \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be another function. By $X_n \uparrow X$ we denote that $\{X_n\}$ is increasing and $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$. Similarly, $X_n \downarrow X$ denotes that the sequence of functions $\{X_n\}$ is decreasing and $\{X_n(\omega)\}$ converges to $X(\omega)$ for all $\omega \in \Omega$.

Let $X_1, X_2, \ldots$ be a sequence of functions defined on the set Ω with values in $\mathbb{R}$. Define the following functions

$$\begin{aligned}
\inf_n X_n &: \Omega \to \mathbb{R} \quad \text{by} & \inf_n X_n(\omega) &:= \inf\{X_1(\omega), X_2(\omega), \ldots\},\\
\sup_n X_n &: \Omega \to \mathbb{R} \quad \text{by} & \sup_n X_n(\omega) &:= \sup\{X_1(\omega), X_2(\omega), \ldots\},\\
\operatorname*{liminf}_{n\to\infty} X_n &: \Omega \to \mathbb{R} \quad \text{by} & \operatorname*{liminf}_{n\to\infty} X_n(\omega) &:= \lim_{n\to\infty} \inf_{k\geq n} X_k(\omega),\\
\operatorname*{limsup}_{n\to\infty} X_n &: \Omega \to \mathbb{R} \quad \text{by} & \operatorname*{limsup}_{n\to\infty} X_n(\omega) &:= \lim_{n\to\infty} \sup_{k\geq n} X_k(\omega).
\end{aligned}$$

Proposition 91. If $X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{F})$, then so are the functions

$$\inf_n X_n, \quad \sup_n X_n, \quad \operatorname*{liminf}_{n\to\infty} X_n, \quad \text{and} \quad \operatorname*{liminf}_{n\to\infty} X_n.$$

Proof. The infimum of a sequence is strictly less than a if and only if some term of the sequence is strictly less than a. Then, we have

$$\begin{aligned}
\big(\inf_n X_n\big)^{-1}((-\infty, a)) = \{\omega \in \Omega : \inf_n X_n(\omega) < a\} &= \bigcup_{n=1}^{\infty} \{\omega \in \Omega : X_n(\omega) < a\}\\
&= \bigcup_{n=1}^{\infty} X_n^{-1}((-\infty, a)) \in \mathcal{F},
\end{aligned}$$

Since the intervals $(-\infty, a)$, $a \in \mathbb{R}$ generate $\mathcal{B}(\mathbb{R})$, by Proposition 85, $\inf_n X_n$ is measurable. The argument for $\sup_n X_n$ is similar. Next, note that

$$\operatorname*{liminf}_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \inf_{k\geq n} X_k(\omega) = \sup_n \big(\inf_{m\geq n} X_m(\omega)\big),$$

where for the last equality, we used the fact that $Y_n := \inf_{m\geq n} X_m$ forms increasing sequence of functions so the limit is the same as the supremum. By the first part Y_n is a random variable and again by the first part, so is $\sup_n Y_n$. The argument for $\operatorname*{liminf}_{n\to\infty} X_n$ is similar. □

Remark 92. Taking inf, sup, liminf, and limsup of a sequence of random variables, may have a side effect. For example, there may be $\omega \in \Omega$ such that

$$\sup_n X_n(\omega) = \sup\{X_1(\omega), X_2(\omega), \dots\} = \infty.$$

Let us see that the set of those ω's is also measurable. Indeed

$$\begin{aligned}\{\omega \in \Omega : \sup_n X_n(\omega) = \infty\} &= \bigcap_{k=1}^{\infty}\{\omega \in \Omega : \sup_n X_n(\omega) \geq k\} = \Big(\bigcup_{k=1}^{\infty}\{\omega \in \Omega : \sup_n X_n(\omega) < k\}\Big)^c \\ &= \Big(\bigcup_{k=1}^{\infty}\big(\sup_n X_n\big)^{-1}((-\infty, k))\Big)^c \in \mathcal{F},\end{aligned}$$

since all of the sets involved in the last union are in $\mathcal{F}$ (see the proof of Proposition 91). In an analogous way, one can show that the sets

$$\begin{aligned}&\{\omega \in \Omega : \inf_n X_n(\omega) = -\infty\},\\ &\{\omega \in \Omega : \operatorname*{limsup}_{n\to\infty} X_n(\omega) = \infty\}, \quad \{\omega \in \Omega : \operatorname*{limsup}_{n\to\infty} X_n(\omega) = -\infty\},\\ &\{\omega \in \Omega : \operatorname*{liminf}_{n\to\infty} X_n(\omega) = \infty\}, \quad \{\omega \in \Omega : \operatorname*{liminf}_{n\to\infty} X_n(\omega) = -\infty\}\end{aligned}$$

are measurable. □

Exercise 93. Let $A_1, A_2, \ldots \in \mathcal{F}$ be a sequence of sets. Show that

$$\operatorname*{liminf}_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_{\text{liminf} A_n}.$$

Formulate similar equality for limsup.

The most elementary random variables are the step-functions.

Definition 94. A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called a *step-function* if it can be represented as

$$X(\omega) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}(\omega)$$

for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $A_1, \ldots, A_n \in \mathcal{F}$.

Particular examples of step-functions are

$$\begin{aligned}\mathbf{1}_\Omega &\equiv 1,\\ \mathbf{1}_\emptyset &\equiv 0,\\ \mathbf{1}_A + \mathbf{1}_{A^c} &= 1,\\ \mathbf{1}_{A\cap B} &= \mathbf{1}_A \mathbf{1}_B,\\ \mathbf{1}_{A\cup B} &= \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_{A\cap B}.\end{aligned}$$

Exercise 95. Show that step-functions are measurable. Hint: use the fact that by taking all possible intersections of the sets A_i and by adding appropriate complements, we can find a collection of sets $C_1, \dots, C_N \in \mathcal{F}$ such that

(i) $C_j \cap C_k = \emptyset$ for $j \neq k$;

(ii) $\cup_{j=1}^{N} C_j = \cup_{i=1}^{n} A_i$; and

(iii) for every set A_i there is an index set $I_i \subseteq \{1, 2, \dots, N\}$ such that $A_i = \cup_{j \in I_i} C_j$.

Proposition 96. A measurable function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a step-function if and only if it takes finitely many values.

Proof. If X is a step function then it can be represented as $X(\omega) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}(\omega)$ for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $A_1, \dots, A_n \in \mathcal{F}$. So, clearly X takes only finitely many values. (These values maybe difficult to list explicitly since the sets $A_1, \dots, A_n$ may intersect.)

Conversely, if X is measurable and takes only the values $\alpha_1, \dots, \alpha_n$ then the set $A_k := X^{-1}(\{\alpha_k\})$ is in $\mathcal{F}$ (since $\{\alpha_k\}$ is a Borel subset of $\mathbb{R}$) for all $k = 1, 2, \dots, n$. The sets $A_1, \dots, A_n$ are disjoint and we have $X(\omega) = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}(\omega)$. □

Corollary 97. If $X, Y : \Omega \to \mathbb{R}$ are two step functions, then so are $X + Y$, XY, $\min\{X, Y\}$, $\max\{X, Y\}$, and so on.

Given a random variable X on $(\Omega, \mathcal{F})$ and a set $A \subset \mathbb{R}$, often for brevity, the set

$$\{\omega \in \Omega : X(\omega) \in A\} \text{ is going to be denoted as } X^{-1}(A) \text{ or } \{X \in A\}.$$

For example, when $A = (a, b)$, the set $\{\omega \in \Omega : a < X(\omega) < b\}$ will be denoted by $\{a < X < b\}$.

Theorem 98. The function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a random variable if and only if there is a sequence $\{X_n\}$ of step-functions $X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$X(\omega) = \lim_{n \to \infty} X_n(\omega) \text{ for all } \omega \in \Omega.$$

Proof. If there is a sequence $\{X_n\}$ of step-functions $X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$X(\omega) = \lim_{n \to \infty} X_n(\omega) \text{ for all } \omega \in \Omega,$$

then X is a random variable by Exercise 95 and Proposition 91.

Suppose now that X is a random variable. Define the step-functions

$$X_n(\omega) := \sum_{k=-4^n}^{4^n - 1} \frac{k}{2^n} \mathbf{1}_{\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\}}(\omega).$$

We are going to show that $X(\omega) = \lim_{n \to \infty} X_n(\omega)$ for all $\omega \in \Omega$. First of all, notice that $X_n(\omega)$ is indeed a step-function, since the sets $\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\}$ are in $\mathcal{F}$. This is where it is used that X is a random variable. Second, the reason why the sets $\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\}$ are defined with one '$\leq$' and one '$<$' is so that they are disjoint. Fix an $\omega \in \Omega$. For any n the finite sequence of numbers (there are $2 \cdot 4^n + 1$ of them)

$$-\frac{4^n}{2^n}, -\frac{4^n-1}{2^n}, -\frac{4^n-2}{2^n}, \dots, -\frac{1}{2^n}, \frac{0}{2^n}, \frac{1}{2^n}, \dots, \frac{4^n-1}{2^n}, \frac{4^n}{2^n}$$

divides $\mathbb{R}$ into $2 \cdot 4^n + 2$ intervals: the left-most and the right-most of which are infinite and the rest are finite. Only the finite intervals take part in the definition of X_n. When n increases two things happen: (1) The interval $\left[\frac{-4^n}{2^n}, \frac{4^n}{2^n}\right)$, between the smallest and the largest number in the sequence, becomes wider with both endpoints approaching infinity; and (2) The numbers in the sequence get closer together, since the difference between any two consecutive numbers is $1/2^n$. Fix $\omega \in \Omega$ and suppose that n is large enough so that $X(\omega) \in \left[\frac{-4^n}{2^n}, \frac{4^n}{2^n}\right)$, then there is a unique k_n, such that $X(\omega) \in \left[\frac{k_n}{2^n}, \frac{k_n+1}{2^n}\right)$ and as n approach infinity the lower bound of the interval $\frac{k_n}{2^n}$ approaches $X(\omega)$. But $X_n(\omega) = \frac{k_n}{2^n}$, so $X_n(\omega)$ approaches $X(\omega)$. To see the last part better, let us calculate the value of $X_{n+1}(\omega)$. Keep in mind that $X(\omega)$ is fixed. When we go from n to $n+1$, all existing finite intervals are split in half (and new are added, but we are not interested in them at the moment):

$$\left[\frac{k_n}{2^n}, \frac{k_n+1}{2^n}\right) = \left[\frac{2k_n}{2^{n+1}}, \frac{2k_n+1}{2^{n+1}}\right) \cup \left[\frac{2k_n+1}{2^{n+1}}, \frac{2k_n+2}{2^{n+1}}\right).$$

So, if $X_n(\omega) = \frac{k_n}{2^n}$, then for $n+1$ we have

$$X_{n+1}(\omega) = \begin{cases} \frac{2k_n}{2^{n+1}} & \text{if } X(\omega) \in \left[\frac{2k_n}{2^{n+1}}, \frac{2k_n+1}{2^{n+1}}\right), \\ \frac{2k_n+1}{2^{n+1}} & \text{if } X(\omega) \in \left[\frac{2k_n+1}{2^{n+1}}, \frac{2k_n+2}{2^{n+1}}\right). \end{cases}$$

This shows that $|X_n(\omega) - X(\omega)| \leq 1/2^n$, for all n. Therefore, we can conclude that $X_n(\omega)$ converges to $X(\omega)$ for all $\omega \in \Omega$. □

Exercise 99. Let the function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable with $X \geq 0$. Show that there is a sequence $\{X_n\}$ of step-functions $X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $0 \leq X_1(\omega) \leq X_2(\omega) \leq \cdots \leq X(\omega)$ and $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$.

Exercise 100. Let the function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable with $X \geq 0$. Define the step-functions

$$X_n(\omega) := \Big(\sum_{k=1}^{n^2} \frac{k-1}{n} \mathbf{1}_{\{\frac{k-1}{n} \leq X < \frac{k}{n}\}}(\omega) \Big) + n\mathbf{1}_{\{n \leq X\}}$$

for all $n = 1, 2, \dots$. Now try to show that Exercise 99 works with these step-functions. What goes wrong?

Exercise 101. Let the function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable. Show that there is a sequence $\{X_n\}$ of step-functions $X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $0 \leq |X_1(\omega)| \leq |X_2(\omega)| \leq \cdots \leq |X(\omega)|$ and $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$.

Exercise 102. Let $X : \Omega \to S$ be a function between the sets Ω and S. Let $\mathcal{S}$ be a σ-algebra on S. Show that

(i) The collection of subsets $\sigma(X) := \{X^{-1}(A) : A \in \mathcal{S}\}$ is a σ-algebra on Ω. This σ-algebra is called the *σ-algebra generated by X*.

(ii) If $\mathcal{F}$ is a σ-algebra on Ω such that $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is measurable, then $\sigma(X) \subseteq \mathcal{F}$. Hence $\sigma(X)$ is the smallest σ-algebra such that $X : (\Omega, \sigma(X)) \to (S, \mathcal{S})$ is measurable.

(iii) Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable. Let $\mathcal{S}_0 := \{(-\infty, x] : x \in \mathbb{R}\}$. Show that $\sigma(X)$ is generated by $\{X^{-1}(A) : A \in \mathcal{S}_0\}$.

So far the discussion about random variables and measurable functions did not involve any measure. We now include a measure.

Exercise 103. Assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space and let $(S, \mathcal{S})$ be a measurable space. Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable function. For every set $B \in \mathcal{S}$ define

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

Show that $\mathbb{P}_X$ is a measure on $\mathcal{S}$. If $\mathbb{P}$ is a probability measure, show that $\mathbb{P}_X$ is also a probability measure. The measure $\mathbb{P}_X$ is called the *image of* $\mathbb{P}$ *under* X or the *law of* X.

In particular, Exercise 103 says that if $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then any random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ defines a probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$.

The law of a random variable is completely characterized, see Proposition 106 below, by its cumulative distribution function which we introduce now.

Definition 104 (Cumulative distribution function)**.** Any random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ defines a function $F_X : \mathbb{R} \to [0, 1]$ by

$$F_X(x) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x\})$$

called *cumulative distribution function* (c.d.f.) of X.

Proposition 105 (Properties of c.d.f.)**.** Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The cumulative distribution function, $F_X : \mathbb{R} \to [0, 1]$ of X is (a) increasing; (b) right continuous; and (c) $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.

Proof. (a) F_X is increasing since for any $x_1 \le x_2$ we have $\{X \le x_1\} \subseteq \{X \le x_2\}$. Hence $F_X(x_1) = \mathbb{P}(\{X \le x_1\}) \le \mathbb{P}(\{X \le x_2\}) = F_X(x_2)$.

(b) Let $x \in \mathbb{R}$ and let $\{x_n\}$ be a decreasing sequence converging to x from the right. Then, since the sets $A_n := \{\omega \in \Omega : X(\omega) \le x_n\}$ form a decreasing sequence, by Proposition 54, part (vi), we have

$$F(x) = \mathbb{P}(\{X \le x\}) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \le x_n\}\Big) = \lim_{n \to \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \le x_n\}) = \lim_{n \to \infty} F_X(x_n).$$

(c) Exercise. □

The cumulative distribution function F_X of a random variable X satisfies the conditions in Theorem 74 and hence defines a *probability* measure on $\mathcal{B}(\mathbb{R})$. On the other hand, the second part of Exercise 103 shows that the law of X, $\mathbb{P}_X$, is also a measure on $\mathcal{B}(\mathbb{R})$. The next proposition shows that these two measures are the same.

Proposition 106. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the law of X and the measure determined by the cumulative distribution function F_X on the σ-algebra $\mathcal{B}(\mathbb{R})$ are equal.*

Proof. Consider the algebra $\mathcal{F}$ in Example 36. For any $A := (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n] \in \mathcal{F}$ we have

$$\mathbb{P}_X(A) = \mathbb{P}(\{X \in A\}) = \sum_{i=1}^{n} \mathbb{P}(\{a_i < X \leq b_i\}) = \sum_{i=1}^{n} (F_X(b_i) - F_X(a_i)).$$

The last sum is equal to the measure of A under the measure determined by F_X, see Theorem 74. Thus, the two measures coincide on the algebra $\mathcal{F}$ that generates $\mathcal{B}(\mathbb{R})$. Hence, by the uniqueness of the extension in the Carathéodory's extension theorem they must coincide on $\mathcal{B}(\mathbb{R})$. □

Exercise 107. *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures on $\mathcal{B}(\mathbb{R})$. Show that $\mathbb{P}_1 = \mathbb{P}_2$ if and only if $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$.*

Definition 108 (Distribution function)**.** A function $F : \mathbb{R} \to \mathbb{R}$ is called a *distribution function* if it is (a) increasing; (b) right continuous; and (c) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

Exercise 109. Show that there are at most countably many $x \in \mathbb{R}$ where a distribution function F is not continuous.

Lemma 110. If $F : \mathbb{R} \to \mathbb{R}$ is a distribution function, then there is a random variable X with cumulative distribution function $F_X = F$.

Proof. Let $\Omega := (0, 1)$, $\mathcal{F} := \mathcal{B}(0, 1)$ the Borel sets on $(0, 1)$, and $\mathbb{P} :=$ the Lebesgue measure on $(0, 1)$ (that is, $\mathbb{P}((a, b]) = b - a$). If $\omega \in (0, 1)$ let

$$X(\omega) := \sup\{y : F(y) < \omega\}.$$

We need to show two things: 1) that X is a random variable, and 2) that $F_X = F$. For 1) it suffices to show the inclusion in

$$X^{-1}((-\infty, x]) = \{\omega \in (0, 1) : X(\omega) \leq x\} \in \mathcal{F}.$$

For 2), we need to show the second equality in

$$F_X(x) = \mathbb{P}(\{\omega \in (0, 1) : X(\omega) \leq x\}) = F(x).$$

We can kill the two birds with one stone if we showed that

$$\{\omega \in (0, 1) : X(\omega) \leq x\} = \{\omega \in (0, 1) : \omega \leq F(x)\} \text{ for all } x \in \mathbb{R}. \tag{17}$$

This is indeed the case since

$$\mathbb{P}(\{\omega \in (0,1] : \omega \leq F(x)\}) = \mathbb{P}(\{\omega \in (0,1) : \omega \leq F(x)\}) = F(x).$$

We now focus on showing (17). Fix $x \in \mathbb{R}$.

Let $\omega \in (0,1)$ be a point in the left-hand side of (17), that is, $X(\omega) \leq x$. But $X(\omega) = \sup\{y : F(y) < \omega\}$, so for every $\epsilon > 0$, $x + \epsilon$ is not in the set $\{y : F(y) < \omega\}$, hence $F(x+\epsilon) \geq \omega$. Since F is right-continuous letting ϵ approach 0, we get $F(x) \geq \omega$, showing that ω is in the set on the right-hand side.

If $\omega \in (0,1)$ is such that $\omega \leq F(x)$ then x is not in the set $\{y : F(y) < \omega\}$. Since F is increasing, that set cannot contain a number bigger than x. That is, x is an upper bound for the set $\{y : F(y) < \omega\}$. Hence $X(\omega) = \sup\{y : F(y) < \omega\} \leq x$, showing that x is in the set on the left-hand side of (17). □

The random variable X defined in the above lemma is sometimes denoted by

$$F^{-1}(\omega) := \sup\{y : F(y) < \omega\} \quad \text{for all } \omega \in (0,1) \tag{18}$$

and called the *quantile function.* (The notation F^{-1} is suggestive, it does not mean that F has an inverse. Still, we will see later that F^{-1} shares some properties with an inverse function.) Thus, the quantile function $F^{-1}(\omega)$ can be viewed as a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega := (0,1)$, $\mathcal{F} := \mathcal{B}(0,1)$ the Borel sets on $(0,1)$, and $\mathbb{P} :=$ the Lebesgue measure on $(0,1)$.

Theorem 111 (Helly's selection theorem)**.** Let $\{F_n\}$ be a sequence of distribution functions. Then there is a subsequence $\{F_{n_i}\}_{i=1}^{\infty}$ and an increasing, right continuous function F so that $\lim_{i\to\infty} F_{n_i}(y) = F(y)$ for every y that is a point of continuity of F.

3.1 Simple classification of random variables

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and X is a random variable on it.

Definition 112. A random variable X that takes on finite or at most a countable number of different values is said to be *discrete.* The *probability mass function* $p(a)$ of X is defined by

$$p(a) := \mathbb{P}(X = a) := \mathbb{P}(\{\omega \in \Omega : X(\omega) = a\}).$$

So, $p(a)$ is just the probability that X takes the value a. If X takes only the values $x_1, x_2, \ldots$, then $p(x_i) \geq 0$ for all $i = 1, 2, \ldots$ and $p(x) = 0$ for all other x's. Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

The cumulative distribution function of a discrete random variable X with values $x_1, x_2, \ldots$ is

$$F_X(x) = \sum_{i : x_i \leq x} p(x_i).$$

Being a cumulative distribution function, $F_X(x)$ is increasing and right-continuous. It is a fact that this $F_X(x)$ has a jump at every x_i and the size of the jump is

$$F_X(x_i) - F_X(x_i-) = p(x_i).$$

Example 113. Consider a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with the following random variable X on it. Let $x_1, x_2, \ldots$ be the set of all rational numbers in $\mathbb{R}$ enumerated in any way. Suppose

$$\mathbb{P}(X = x_i) = 1/2^i \text{ for } i = 1, 2, \ldots$$

Then, the cumulative distribution function of X is

$$F_X(x) = \sum_{i: x_i \leq x} 1/2^i.$$

It is right-continuous everywhere, and has a jump at every rational number. That is one jumpy function!

Note that if Ω is finite or countably infinite set, then any random variable X on Ω is necessarily discrete.

Proposition 114. *If X and Y are two discrete random variables, then so is $X + Y$.*

Proof. Let $\{x_1, x_2, \ldots\}$ be all the values that X takes. We can order them in a sequences (possibly finite one) since X is discrete random variable. Let $\{y_1, y_2, \ldots\}$ be all the values that Y takes. Then, the values that $X + Y$ takes are given in the interior of the following table

	x_1	x_2	x_3	x_4	x_5	$\cdots$
y_1	$x_1 + y_1$	$x_2 + y_1$	$x_3 + y_1$	$x_4 + y_1$	$x_5 + y_1$	$\cdots$
y_2	$x_1 + y_2$	$x_2 + y_2$	$x_3 + y_2$	$x_4 + y_2$	$x_5 + y_2$	$\cdots$
y_3	$x_1 + y_3$	$x_2 + y_3$	$x_3 + y_3$	$x_4 + y_3$	$x_5 + y_3$	$\cdots$
y_4	$x_1 + y_4$	$x_2 + y_4$	$x_3 + y_4$	$x_4 + y_4$	$x_5 + y_4$	$\cdots$
y_5	$x_1 + y_5$	$x_2 + y_5$	$x_3 + y_5$	$x_4 + y_5$	$x_5 + y_5$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

See Example 5 to find out how to order all values $x_i + y_j$ in a sequence. Then, you may want to remove duplicate values, if any. This shows that $X + Y$ is a discrete random variable. □

Definition 115. A random variable X that takes uncountably many different values, for example, if it takes any value in an interval (a, b), is called *continuous* random variable.

Continuous random variables fall into two classes: those that have density and those that do not.

Definition 116 (Density)**.** A random variable X is called *absolutely continuous* if its cumulative distribution function $F_X(x)$ can be represented as

$$F_X(x) = \int_{-\infty}^{x} f(y)\, dy \tag{19}$$

for some function $f(y) \geq 0$. The function $f(x)$ is called a *(probability) density function of X*. In this case, we also say that $F_X(x)$ is *absolutely continuous* function.

If X is absolutely continuous with density $f(x)$, then

$$\begin{aligned}\mathbb{P}_X((a,b]) &= \mathbb{P}(a < X \le b) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \le b\} \setminus \{\omega \in \Omega : X(\omega) \le a\}) \\ &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \le b\}) - \mathbb{P}(\{\omega \in \Omega : X(\omega) \le a\}) = F_X(b) - F_X(a) \\ &= \int_a^b f(y)\,dy.\end{aligned}$$

Definition 117. A function $f : \mathbb{R} \to \mathbb{R}$ is called a *density function* if

1.) $f(x) \ge 0$ for all x, and

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

Even though the cumulative density $F_X(x)$ is uniquely determined by X, the density function is not. You can change a density function at finitely many points (in fact, on a set with Lebesgue measure zero) without changing the integral (19). For example the following three functions give the same integrals

$$f(x) = \begin{cases} e^{-x} & x > 0, \\ 0 & x \le 0; \end{cases} \qquad f(x) = \begin{cases} e^{-x} & x \ge 0, \\ 0 & x < 0; \end{cases} \qquad f(x) = \begin{cases} e^{-x} & x \ge 0, x \ne 4, \\ -10 & x = 4, \\ 0 & x < 0; \end{cases}$$

The first two examples show two different density functions, the last function is, by definition, not a density function since it takes a negative value.

Clearly, every function $f(x)$ that satisfies the conditions in Definition 117 is a density of a random variable. Indeed, the function defined by (19) is increasing and continuous, hence, by Lemma 110, it is the cumulative distribution function of a random variable, having density $f(x)$. Note that the function $f(x)$, itself, does not have to be continuous, all that is required from it is that it is non-negative and integrable to 1.

Example 118. For any $p \in (0,1)$ let
(1) $\Omega := \{0, 1, 2, \ldots, n\}$.
(2) $\mathcal{F} := 2^{\Omega}$ (the collection of all subsets of Ω).
(3) For any $A \in \mathcal{F}$ define $\mathbb{P}(A) := \sum_{k \in A} \binom{n}{k} p^k (1-p)^{n-k}$.
Then, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. □

Interpretation: A sequence of n identical and independent experiments. Each experiment results in success with probability p and failure with probability $1-p$. Then, $\mathbb{P}(\{k\})$ is the probability of exactly k successes in the n experiments. Notice that there is no random variable defined in Example 118. To say that a random variable has a binomial distribution, we need the following definition.

Definition 119. A discrete random variable X defined on a (abstract) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ has *binomial distribution* with parameters n and p, if X has probability mass function

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for all } k = 0, 1, 2, \ldots, n,$$

where n is a positive integer and $p \in (0,1)$.

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ described in Example 118. What would be a binomially distributed random variable on that space? Simple, let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be defined by $X(k) = k$ for all $k = 0, 1, \ldots, n$. Then, X has a binomial distribution.

Example 120. For any $\lambda > 0$ let
(1) $\Omega := \{0, 1, 2, 3, \ldots\}$.
(2) $\mathcal{F} := 2^{\Omega}$ (the collection of all subsets of Ω).
(3) For any $A \in \mathcal{F}$ define $\mathbb{P}(A) := \sum_{k \in A} \frac{e^{-\lambda}\lambda^k}{k!}$.
Then, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. □

The Poisson distribution is used, for example, to model stochastic processes with a continuous time parameter and jumps: the probability that the process jumps k times between the time-points s and t with $0 \le s < t < \infty$ is equal to $\mathbb{P}(\{k\})$.

Definition 121. A discrete random variable X defined on a (abstract) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ has *Poisson distribution* with parameter $\lambda > 0$, if X has probability mass function

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for all } k = 0, 1, 2, \ldots$$

Definition 122. We say that a continuous random variable X is *uniformly distributed* in the interval $[a, b]$ if it has probability density function

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

Definition 123. We say that a continuous random variable X is *normally distributed* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ if it has probability density function

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

It is clearly increasing and continuous.

Definition 124. We say that a continuous random variable X is *exponentially distributed* with parameter $\lambda > 0$ if it has probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0, \\ 0 & \text{if } x < 0. \end{cases}$$

If a random variable X has density, then by (19), its cumulative distribution function $F_X(x)$ is continuous. So, looking at the cumulative distribution function, if it has jumps, then the random variable cannot be absolutely continuous.

If a random variable has density, then it must be continuous random variable, but the opposite is not true. The next example shows a concrete continuous random variable that is not absolutely continuous.

Example 125. Consider a random variable X whose values are the outcomes of the following experiment. Throw a fair coin. If the outcome is H then pick a random number uniformly distributed in $[0, 1]$. If the outcome is T then pick a random number from the set $\{2, 3\}$ with probabilities $2/3, 1/3$, respectively. This random variable is continuous since it can take uncountably many values, namely any value in $[0, 1] \cup \{2, 3\}$ but it is not absolutely continuous as we will now see.

Let Y be a random variable representing the coin flip, that is Y takes values H and T with probabilities $1/2$, $1/2$ respectively. Let Z be a random variable uniformly distributed in $[0, 1]$ and let T be a random variable taking values $\{2, 3\}$ with probabilities $2/3, 1/3$, respectively. We consider several cases.

1) Let $x < 0$, then $F(x) = P(X \leq x) = 0$ since the r.v. X never takes values smaller than 0.

2) Let $0 \leq x \leq 1$, then $\{X \leq x\} = \{Y = H\} \cap \{Z \leq x\}$ where the last two events are independent. Hence

$$F(x) = P(X \leq x) = P(\{Y = H\} \cap \{Z \leq x\}) = P(Y = H)P(Z \leq x) = \frac{1}{2}\frac{x-0}{1-0} = \frac{x}{2}.$$

3) Let $1 \leq x < 2$, then $\{X \leq x\} = \{X \leq 1\}$ and $F(x) = P(X \leq x) = P(X \leq 1) = F(1) = 1/2$, where $F(1)$ was computed in case 1.

4) Let $2 \leq x < 3$, then

$$\{X \leq x\} = \{X \leq 2\} = \{X \leq 1\} \cup \{1 < X \leq 2\} = \{X \leq 1\} \cup \{X = 2\},$$

where the union is disjoint. Hence, $F(x) = P(\{X \leq 1\} \cup \{X = 2\}) = P(\{X \leq 1\}) + P(\{X = 2\}) = 1/2 + P(\{X = 2\})$. Now, $\{X = 2\} = \{Y = T\} \cap \{T = 2\}$, where the last two events are independent, implying that $P(X = 2) = P(Y = T)P(T = 2) = (1/2)(2/3) = 1/3$. Thus, in this case $F(x) = 1/2 + 1/3 = 5/6$.

5) Let $3 \leq x$, then

$$\{X \leq x\} = \{X \leq 3\} = \{X \leq 1\} \cup \{1 < X \leq 2\} \cup \{2 < X \leq 3\} = \{X \leq 1\} \cup \{X = 2\} \cup \{X = 3\}.$$

But we do not need the last union at all. Since X can takes only values that are always less-than-or-equal to 3, we have $F(x) = P(X \leq 3) = 1$.

We summarize all cases in

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/2 & \text{if } 0 \leq x < 1, \\ 1/2 & \text{if } 1 \leq x < 2, \\ 5/6 & \text{if } 2 \leq x < 3, \\ 1 & \text{if } 3 \leq x. \end{cases}$$

Since $F(x)$ is not continuous (note that it always has to be right-continuous), we conclude that X does not have a probability density function. So this is an example of a continuous random variable that is not absolutely continuous. □

Suppose X is absolutely continuous random variable, that is, X has a probability density function $f(x)$. So how can we obtain $f(x)$ if we know $F_X(x)$? The answer lies in the integral representation (19) of $F_X(x)$ and the fundamental theorem of calculus.

Theorem 126 (The fundamental theorem of calculus)**.** If $f(x)$ is an integrable function, continuous at x_0, then the function

$$F(x) = \int_{-\infty}^{x} f(y)\,dy$$

is differentiable at x_0 and $F'(x_0) = f(x_0)$.

In particular, if $f(x)$ is continuous for all x, then $F'(x) = f(x)$ for all x.

3.2 Full classification of random variables (optional)

For a detailed discussion of the results in this section refer to [4]. For any sequence of points $x_1, x_2, \ldots$ and corresponding values $h_1, h_2, \ldots$, define the function

$$j(x) := \sum_{i: x_i \leq x} h_i.$$

A function that can be represented in this way is called a *jump function.* Note that the two sequences may be finite, in which case, $j(x)$ will have finitely many jumps.

A function $s(x) : \mathbb{R} \to \mathbb{R}$ is called *singular* if it is continuous, has bounded variation, and has a derivative equal to zero, almost everywhere.

Finally, a function $\phi(x) : \mathbb{R} \to \mathbb{R}$ is called *absolutely continuous* if it can be represented as an integral

$$\phi(x) = \int_{-\infty}^{x} \psi(t)\,dt.$$

Theorem 127 (Lebesgue decomposition)**.** Any distribution function $F(x)$ can be decomposed in a unique way (up to adding or subtracting constants) as a sum

$$F(x) = \phi(x) + j(x) + s(x),$$

for some absolutely continuous function $\phi(x)$, a jump function $j(x)$, and a singular function $s(x)$.

Now, let X be a random variable with cumulative distribution function $F_X(x)$. Let

$$F_X(x) = \phi_X(x) + j_X(x) + s_X(x),$$

be the Lebesgue decomposition of $F_X(x)$.

- If $\phi_X(x) = 0$ and $s_X(x) = 0$, then X is a discrete random variable.

- If $j_X(x) = 0$ and $s_X(x) = 0$, then X is an absolutely continuous random variable.

- If $\phi_X(x) = 0$ and $j_X(x) = 0$, then X is called *singular* random variable.

- If $\phi_X(x) \neq 0$ or $s_X(x) \neq 0$, then X is continuous random variable (but may not be absolutely continuous).

Singular random variables, are a bizarre bunch. They are characterized by the following property. Singular random variable X is such that $\mathbb{P}(X = a) = 0$ for all $a \in \mathbb{R}$, but there is a Borel set $A \in \mathcal{B}(\mathbb{R})$ with Lebesgue measure 0, such that $\mathbb{P}(X \in A) = 1$.

Example 128. Recall the description of the Cantor set C given in Example 9. Consider now an infinite sequence of independent tosses of a fair coin. If the i-th toss results in tails, record $x_i = 0$; if it results in heads, record $x_i = 2$. Use the x_i's to form a number x,

$$x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}.$$

This defines a random variable X, whose range is the set C. The probability law of this random variable is therefore concentrated on the "zero-length" set C. At the same time, $P(X = x) = 0$ for every x, because any particular sequence of heads and tails has zero probability. For the explanation why C is a Borel set, see Example 46.

3.3 Expected value of a random variable

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on it, in this section, we define the expected value of X denoted by

$$EX = \int_\Omega X \, d\mathbb{P} = \int_\Omega X(\omega) \, d\mathbb{P}(\omega)$$

and study its basic properties. The definition of the expected value is done in three steps.

Step 1. If X is a step-function with representation $X(\omega) = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}(\omega)$ for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $A_1, \dots, A_n \in \mathcal{F}$, then we define

$$\mathbb{E}X = \int_\Omega X \, d\mathbb{P} := \sum_{i=1}^n \alpha_i \mathbb{P}(A_i).$$

We have to check that the definition is correct, since it might be that different representations of X give different expected values EX. However, this is not the case as shown by the next lemma.

Lemma 129. If $X = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i} = \sum_{j=1}^m \beta_j \mathbf{1}_{B_j}$ then $\sum_{i=1}^n \alpha_i \mathbb{P}(A_i) = \sum_{j=1}^m \beta_j \mathbb{P}(B_j)$.

Proof. By subtracting the right-hand side from the left-hand one, in both equations, we only need to show that

$$\text{if } \sum_{i=1}^N \alpha_i \mathbf{1}_{A_i} = 0 \text{ then } \sum_{i=1}^N \alpha_i \mathbb{P}(A_i) = 0.$$

Taking all possible intersections of the sets A_i and by adding appropriate complements, we can find a collection of sets $C_1, \dots, C_M \in \mathcal{F}$ such that

(i) $C_j \cap C_k = \emptyset$ for $j \neq k$;

(ii) for every set A_i there is an index set $I_i \subseteq \{1, 2, \dots, M\}$ such that $A_i = \cup_{j \in I_i} C_j$.

After these preparations, we have

$$0 = \sum_{i=1}^N \alpha_i \mathbf{1}_{A_i} = \sum_{i=1}^N \alpha_i \Big(\sum_{j \in I_i} \mathbf{1}_{C_j} \Big) = \sum_{i=1}^N \sum_{j \in I_i} \alpha_i \mathbf{1}_{C_j} = \sum_{j=1}^M \Big(\sum_{i : j \in I_i} \alpha_i \Big) \mathbf{1}_{C_j}.$$

Since the sets $C_1, \dots, C_M$ are disjoint we conclude that either $\sum_{i: j \in I_i} \alpha_i = 0$ or $C_j = \emptyset$ for all $j = 1, \dots, M$. From this, we get

$$\sum_{i=1}^{N} \alpha_i \mathbb{P}(A_i) = \sum_{i=1}^{N} \alpha_i \Big(\sum_{j \in I_i} \mathbb{P}(C_j) \Big) = \sum_{i=1}^{N} \sum_{j \in I_i} \alpha_i \mathbb{P}(C_j) = \sum_{j=1}^{M} \Big(\sum_{i: j \in I_i} \alpha_i \Big) \mathbb{P}(C_j) = 0.$$

This is what we had to show. □

Exercise 130. Imitating the proof of Lemma 129, show that if X and Y are step functions with $X \leq Y$, then $\mathbb{E}X \leq \mathbb{E}Y$.

Corollary 131. If X and Y are step functions then $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$ for all $a, b \in \mathbb{R}$

Proof. Suppose $X = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{A_i}$ and $Y = \sum_{j=1}^{m} \beta_j \mathbf{1}_{B_j}$. Then

$$\mathbb{E}(aX + bY) = \mathbb{E}\Big(\sum_{i=1}^{n} a\alpha_i \mathbf{1}_{A_i} + \sum_{j=1}^{m} b\beta_j \mathbf{1}_{B_j} \Big) = \sum_{i=1}^{n} a\alpha_i \mathbb{P}(A_i) + \sum_{j=1}^{m} b\beta_j \mathbb{P}(B_j) = a\mathbb{E}X + b\mathbb{E}Y,$$

where in the second equality we used Lemma 129 saying that the expectation of the step function $X + Y$ is the same, no matter what is the representation of $X + Y$. □

Example 132. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable on Ω taking the finitely many distinct values $\{x_1, x_2, \dots, x_n\}$ with probabilities $p_1, p_2, \dots, p_n$, (where $p_i \in [0,1]$ and $\sum_{i=1}^{n} p_i = 1$). Define the subsets $A_i := X^{-1}(\{x_i\})$ for $i = 1, \dots, n$. Observe that the sets $A_1, \dots, A_n$ are disjoint and that $\cup_{i=1}^{n} A_i = \Omega$ (indeed, for every $\omega \in \Omega$, we have that $X(\omega)$ is equal to exactly one x_i, hence $\omega \in A_i$.) Thus, X can be represented as

$$X(\omega) = \sum_{k=1}^{n} x_k \mathbf{1}_{A_k}(\omega)$$

and

$$\mathbb{E}X = \int_{\Omega} X \, d\mathbb{P} = \int_{\Omega} \sum_{k=1}^{n} x_k \mathbf{1}_{A_k}(\omega) \, d\mathbb{P} = \sum_{i=1}^{n} x_i \mathbb{P}(A_i) = \sum_{i=1}^{n} x_i p_i.$$

□

Step 2. If the random variable $X : \Omega \to \mathbb{R}$ is non-negative, $X(\omega) \geq 0$ for all $\omega \in \Omega$, then define

$$\mathbb{E}X = \int_{\Omega} X \, d\mathbb{P} := \sup\{\mathbb{E}Z : 0 \leq Z(\omega) \leq X(\omega) \text{ for all } \omega \in \Omega \text{ and } Z \text{ is a step function}\}. \tag{20}$$

Note that it is possible to end up with $\mathbb{E}X = \infty$ in this definition.

Exercise 133. Using Exercise 130, show that Step 2 is "backwards" compatible with Step 1. That is, if X is a non-negative step function, then $\mathbb{E}X = \sup\{\mathbb{E}Z : 0 \leq Z(\omega) \leq X(\omega)$ for all $\omega \in \Omega;\ Z$ is a step function$\}$.

In order to describe the third and last step on the definition of the expected value we need to define

$$X^+(\omega) := \max\{X(\omega), 0\} \geq 0 \text{ and } X^-(\omega) := -\min\{X(\omega), 0\} \geq 0$$

and observe that

$$X(\omega) = X^+(\omega) - X^-(\omega).$$

By Proposition 91, $X^+(\omega)$ and $X^-(\omega)$ are non-negative *random variables* (that is, they are measurable functions from $(\Omega, \mathcal{F})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$) hence the expectations EX^+ and $\mathbb{E}X^-$ are defined in Step 2.

Step 3. If $\mathbb{E}X^+ < \infty$ or $\mathbb{E}X^- < \infty$ then we say that the *expected value* $\mathbb{E}X$ *of* X *exists* and we define

$$\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^- \in \mathbb{R} \cup \{-\infty, \infty\}.$$

The random variable X is called *integrable* if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$, that is, if $\mathbb{E}X \in \mathbb{R}$. If $\mathbb{E}X^+ = \infty$ and $\mathbb{E}X^- < \infty$, then $\mathbb{E}X = \infty$. If $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- = \infty$, then $\mathbb{E}X = -\infty$.

If $A \in \mathcal{F}$ then we define the integral over A by

$$\int_A X \, d\mathbb{P} := \int_\Omega X \mathbf{1}_A \, d\mathbb{P} := \int_\Omega X(\omega) \mathbf{1}_A(\omega) \, d\mathbb{P}(\omega).$$

Before we can give more elaborate examples of integration we need to investigate the properties of the integral.

3.4 Properties of the expected value

We say that a property $\mathcal{P}(\omega)$, depending on ω, holds $\mathbb{P}$-*almost surely* or *almost surely* (a.s.) if the set $\{\omega \in \Omega : \mathcal{P}(\omega) \text{ holds}\}$ is in $\mathcal{F}$ and has measure one. For example, we say that $X = 0$ a.s. if the set $\{\omega \in \Omega : X(\omega) = 0\}$ is in $\mathcal{F}$ and has measure one. We say that the sequence of functions $\{X_n\}$, defined on Ω, is *increasing a.s.* if the set of all $\omega \in \Omega$ for which $X_1(\omega) \leq X_2(\omega) \leq \ldots$ is in $\mathcal{F}$ and has measure one. We say that the *sequence* $\{X_n\}$ *converges to* X *a.s.* if the set of all $\omega \in \Omega$ for which $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ is in $\mathcal{F}$ and has measure one.

Below we summarize basic properties of the expected value.

Proposition 134. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be two random variables.

(i) If $0 \leq X(\omega) \leq Y(\omega)$, then $0 \leq \mathbb{E}X \leq \mathbb{E}Y$.

(ii) If $X = 0$ a.s. then $\mathbb{E}X = 0$.

(iii) If $X \geq 0$ a.s. and $\mathbb{E}X = 0$, then $X = 0$ a.s.

Proof. (i) Using definition (20) in Step 2, since $0 \leq X(\omega) \leq Y(\omega)$ we get the set inclusion

$$\{\mathbb{E}Z : 0 \leq Z(\omega) \leq X(\omega) \text{ for all } \omega \in \Omega \text{ and } Z \text{ is a step function}\}$$
$$\subseteq \{\mathbb{E}Z : 0 \leq Z(\omega) \leq Y(\omega) \text{ for all } \omega \in \Omega \text{ and } Z \text{ is a step function}\}.$$

Hence the supremum of the first set, giving $\mathbb{E}X$, is smaller that the supremum of the second set, giving $\mathbb{E}Y$.

(ii) Case 1. Suppose that X is a step function. Represent it as $X = \sum_{k=1}^{n} \alpha_k \mathbf{1}_{A_k}$ using disjoint sets $A_1, \ldots, A_n$. Since $X = 0$ a.s., we have that $\alpha_k \neq 0$ implies that $\mathbb{P}(A_k) = 0$, hence $\mathbb{E}X = 0$.

Case 2. Suppose that X is non-negative. If Z is a step function with $0 \leq Z(\omega) \leq X(\omega)$ for all $\omega \in \Omega$, then $Z = 0$ a.s. Hence, by Case 1, $\mathbb{E}Z = 0$. Thus by (20) we have $\mathbb{E}X = 0$.

Case 3. Suppose that X is arbitrary. Represent it as $X = X^+ - X^-$. Since $X = 0$ a.s. then $X^+ = 0$ a.s. and $X^- = 0$ a.s. By Case 2, we have $\mathbb{E}X^+ = 0$ and $\mathbb{E}X^- = 0$, implying that $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^- = 0$.

(iii) Exercise.

□

Lemma 135. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be step functions with $0 \leq X_n \uparrow X(\omega)$. Then

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

Proof. We prove the lemma in the case when X is an indicator function, that is $X(\omega) = \mathbf{1}_A(\omega)$ for some $A \in \mathcal{F}$. Let $\epsilon \in (0,1)$, define the set

$$B_\epsilon^n := \{\omega \in A : 1 - \epsilon \leq X_n(\omega)\}$$

and observe that $(1-\epsilon)\mathbf{1}_{B_\epsilon^n}(\omega) \leq X_n(\omega) \leq \mathbf{1}_A(\omega)$. Hence

$$(1-\epsilon)\mathbb{P}(B_\epsilon^n) = \mathbb{E}\big((1-\epsilon)\mathbf{1}_{B_\epsilon^n}(\omega)\big) \leq \mathbb{E}X_n \leq \mathbb{E}(\mathbf{1}_A) = \mathbb{P}(A).$$

Since $B_\epsilon^n \subseteq B_\epsilon^{n+1}$ and $\cup_{n=1}^{\infty} B_\epsilon^n = A$ we get, by the continuity of the measure from below, that $\lim_{n\to\infty} \mathbb{P}(B_\epsilon^n) = \mathbb{P}(A)$. Taking the limit as n approaches infinity, we obtain

$$(1-\epsilon)\mathbb{P}(A) \leq \lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{P}(A).$$

Since ϵ was arbitrary, we have $\mathbb{P}(A) \leq \lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{P}(A)$ showing that $\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{P}(A) = \mathbb{E}X$.

The general case when X is an arbitrary step function is left as an exercise. □

Lemma 136. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable with $X \geq 0$. There is a sequence of step functions $\{Y_n\}$ such that $0 \leq Y_n \uparrow X$ and

$$\lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}X.$$

Proof. Recall that (20) defines $\mathbb{E}X$ as a supremum of a subset of $\mathbb{R}$. By Proposition 19, there is a sequence of step functions $0 \leq Z_n \leq X$ such that the sequence $\mathbb{E}Z_n$ is increasing and converging to $\mathbb{E}X$. The problem is that the sequence of functions Z_n may not be increasing nor we know if $Z_n(\omega)$ converges to $X(\omega)$ for every $\omega \in \Omega$.

In order to remedy these deficiencies, let $\{X_n\}$ be a sequence of non-negative step-functions such that $X_n \uparrow X(\omega)$. Such a sequence exists by Exercise 99.

Finally, we define

$$Y_n := \max\{X_n, Z_1, Z_2, \ldots, Z_n\},$$

and note that as a maximum of step functions, Y_n is a step function for every n and $Y_n \geq 0$. Note also that $Y_1 \leq Y_2 \leq \ldots$. Since $X_n(\omega) \leq Y_n(\omega) \leq X(\omega)$ for all $\omega \in \Omega$ and since $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ we obtain that $\lim_{n\to\infty} Y_n(\omega) = X(\omega)$. Since $Z_n(\omega) \leq Y_n(\omega) \leq X(\omega)$ by Proposition 134, we get $\mathbb{E}Z_n \leq \mathbb{E}Y_n \leq \mathbb{E}X$. Finally, from $\lim_{n\to\infty} \mathbb{E}Z_n = \mathbb{E}X$ and the last inequality, we obtain $\lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}X$. Thus, sequence $\{Y_n\}$ has all the required properties. □

The next lemma shows that instead of evaluating the expectation $\mathbb{E}X$ of a positive random variable by formula (20), we can instead take an increasing sequence of step functions converging to X (see Exercise 99), and evaluate the limit of their expectations.

Lemma 137. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable with $X \geq 0$. For *any* sequence $\{X_n\}$ of step functions such that $0 \leq X_n \uparrow X$, we have

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

Proof. Let $\{X_n\}$ be a sequence of step functions such that $0 \leq X_n \uparrow X$, we have to show that $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n$. Let $\{Y_n\}$ be the sequence of step functions from Lemma 136 and define the step functions

$$W_{k,n} := \min\{X_k, Y_n\}.$$

First, note that for any fixed k, the sequence $\{W_{k,n}\}_{n=1}^{\infty}$ is increasing and that for any fixed n, the sequence $\{W_{k,n}\}_{k=1}^{\infty}$ is increasing. Second, observe that for any fixed k we have $\lim\limits_{n\to\infty} W_{k,n}(\omega) = X_k(\omega)$ and for any fixed n we have $\lim\limits_{k\to\infty} W_{k,n}(\omega) = Y_n(\omega)$. Hence, by Lemma 135 we have $\lim\limits_{n\to\infty} \mathbb{E}W_{k,n} = \mathbb{E}X_k$ and $\lim\limits_{k\to\infty} \mathbb{E}W_{k,n} = \mathbb{E}Y_n$. Finally, we get

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}Y_n = \lim_{n\to\infty} \big(\lim_{k\to\infty} \mathbb{E}W_{k,n}\big) = \lim_{k\to\infty} \big(\lim_{n\to\infty} \mathbb{E}W_{k,n}\big) = \lim_{k\to\infty} \mathbb{E}X_k,$$

where in the third equality we used Exercise 71. □

We are now ready to establish more properties of the expected value.

Proposition 138. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be two random variables.

(i) If $X \geq 0$ and $Y \geq 0$, then $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

(ii) X is integrable if and only if $|X|$ is and in that case we have $|\mathbb{E}X| \leq \mathbb{E}|X|$.

(iii) If $\mathbb{E}X$ exists and $c \in \mathbb{R}$ then $\mathbb{E}(cX)$ exists and $\mathbb{E}(cX) = c\mathbb{E}X$.

(iv) If $X \leq Y$ and both $\mathbb{E}X$, $\mathbb{E}Y$ exist, then $\mathbb{E}X \leq \mathbb{E}Y$.

(v) If $X = Y$ a.s. and the expectation of one of them exists, then so does the expectation of the other and $\mathbb{E}X = \mathbb{E}Y$.

(vi) If X and Y are integrable then $aX + bY$ is integrable and $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$ for all $a, b \in \mathbb{R}$.

Proof. We are only going to establish the first two properties. The rest are intuitively clear and/or follow similarly.

(i) Let $X \geq 0$ and $Y \geq 0$, then there are increasing sequences of non-negative step functions $\{X_n\}$ and $\{Y_n\}$ converging to X and Y respectively. Hence, $\{X_n + Y_n\}$ is an increasing sequence of non-negative step functions converging to $X + Y$ and

$$\mathbb{E}(X+Y) = \lim_{n\to\infty} \mathbb{E}(X_n + Y_n) = \lim_{n\to\infty} (\mathbb{E}X_n + \mathbb{E}Y_n) = \lim_{n\to\infty} \mathbb{E}X_n + \lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}X + \mathbb{E}Y,$$

where in the second equality we used Corollary 131.

(ii) By definition X is integrable if and only if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$. Since $|X| = X^+ + X^-$, by the previous property, $\mathbb{E}|X| = \mathbb{E}X^+ + \mathbb{E}X^-$. Hence, $|X|$ is integrable if and only if $\mathbb{E}X^+ + \mathbb{E}X^- < \infty$ which happens if and only if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$. To show the inequality, observe that

$$|\mathbb{E}X| = |\mathbb{E}X^+ - \mathbb{E}X^-| \leq |\mathbb{E}X^+| + |\mathbb{E}X^-| = \mathbb{E}X^+ + \mathbb{E}X^- = \mathbb{E}|X|.$$ □

The reader may have noticed the increasing generality in Lemma 135 and Lemma 137. In Lemma 135 the increasing sequence and its limit are step functions. In Lemma 137 the increasing sequence is of step functions but its limit is a positive random variable. Now we are ready to state the most general version, in which the increasing sequence is of positive random variables and its limit is a positive random variable. We increase the generality a bit more by requiring that the sequence is increasing almost surely and converges to its limit almost surely. We begin with a lemma.

Lemma 139. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables such that $0 \leq X_n \uparrow X$. Then, there is a sequence $\{Y_n\}$ of *step functions*, such that $0 \leq Y_n \uparrow X$ and

$$Y_n \leq X_n \text{ for every } n = 1, 2, \ldots$$

Proof. For each X_n take a sequence of step functions $\{X_{n,k}\}_{k=1}^{\infty}$ such that $0 \leq X_{n,k} \uparrow X_n$ as $k \to \infty$ (see Exercise 99). Define the step functions

$$Y_N := \max_{1\leq n,k\leq N} X_{n,k}$$

and observe that $Y_{N-1} \leq Y_N \leq \max_{1\leq n\leq N} X_n = X_N$. Let $Y := \lim_{N\to\infty} Y_N$. For $1 \leq n \leq N$ we have that

$$X_{n,N} \leq Y_N \leq X_N$$

so, taking $N \to \infty$ we obtain $X_n \leq Y \leq X$ and hence

$$X = \lim_{n\to\infty} X_n \leq Y \leq X,$$

showing that $Y = X$. Thus, for the step functions $\{Y_N\}$ we have $0 \leq Y_N \uparrow X$. □

Theorem 140 (Monotone convergence)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. If $0 \leq X_n \uparrow X$ a.s., then

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

Proof. Suppose first that $0 \leq X_n(\omega) \uparrow X(\omega)$ for all $\omega \in \Omega$. Let $\{Y_n\}$ be a sequence of step functions with properties as in Lemma 139. Then, Lemma 137 implies that $\lim\limits_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}X$. On the one hand, $Y_n \leq X_n$ implies that $\mathbb{E}Y_n \leq \mathbb{E}X_n$ and taking limit infimum from both sides gives

$$\mathbb{E}X = \liminf_{n\to\infty} \mathbb{E}Y_n \leq \liminf_{n\to\infty} \mathbb{E}X_n,$$

where the equality comes from the fact that $\{\mathbb{E}Y_n\}$ is a convergent sequence with limit $\mathbb{E}X$. On the other hand, $X_n \leq X$ implies that $\mathbb{E}X_n \leq \mathbb{E}X$ and taking limit superior from both sides (the right-hand side is a constant) gives

$$\limsup_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X.$$

Combining the last two displayed lines shows that

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

Hence, we must have equality throughout, showing that the sequence $\{\mathbb{E}X_n\}$ is convergent with limit $\mathbb{E}X$.

Suppose now that $0 \leq X_n \uparrow X$ a.s.. By definition, this means that there is a set A of measure 1 such that $0 \leq X_n(\omega) \uparrow X(\omega)$ for all $\omega \in A$. Hence, $0 \leq X_n(\omega)\mathbf{1}_A(\omega) \uparrow X(\omega)\mathbf{1}_A(\omega)$ for all $\omega \in \Omega$. By the first part of the proof we have

$$\lim_{n\to\infty} \mathbb{E}(X_n\mathbf{1}_A) = \mathbb{E}(X\mathbf{1}_A).$$

Since $X_n\mathbf{1}_A = X_n$ a.s. and $X\mathbf{1}_A = X$ a.s., by Proposition 138 we get $\mathbb{E}(X_n\mathbf{1}_A) = \mathbb{E}X_n$ and $\mathbb{E}(X\mathbf{1}_A) = \mathbb{E}X$. We are done. □

Exercise 141. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. If $0 \geq X_n \downarrow X$ a.s., then

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

Theorem 142 (Fatou lemma)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables with $X_n \geq 0$. Then

$$\liminf_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E}\big(\liminf_{n\to\infty} X_n\big).$$

Proof. Define $Y_n := \inf_{m\geq n} X_m$ and $Y := \lim_{n\to\infty} Y_n$. By the definition of limit infimum we know that $Y = \liminf_{n\to\infty} X_n$. On the one hand, since $0 \leq Y_n \uparrow Y$, then by the Monotone Convergence Theorem we obtain

$$\lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}Y = \mathbb{E}\big(\liminf_{n\to\infty} X_n\big).$$

On the other hand, since $X_n \geq Y_n$, then $\mathbb{E}X_n \geq \mathbb{E}Y_n$ and taking limit infimum from both sides gives

$$\liminf_{n\to\infty} \mathbb{E}X_n \geq \liminf_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}Y,$$

since the sequence $\{\mathbb{E}Y_n\}$ converges to $\mathbb{E}Y$. Combining the two displayed lines concludes the proof. □

Theorem 143 (Lebesgue's dominated convergence theorem)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $Y, X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. If $\lim_{n\to\infty} X_n = X$ a.s., $|X_n| \leq Y$, and Y is integrable, then X is integrable and

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

Proof. To show that X is integrable, we apply Fatou's lemma to the sequence $\{|X_n|\}_{n=1}^{\infty}$ of positive random variables:

$$\infty > \mathbb{E}Y \geq \liminf_{n\to\infty} \mathbb{E}|X_n| \geq \mathbb{E}\big(\liminf_{n\to\infty} |X_n|\big) = \mathbb{E}|X|.$$

This means that $|X|$ is integrable, which is equivalent to X being integrable.

Next, note that $|X_n| \leq Y$ and the integrability of Y implies that $|X_n|$ is integrable and so is X_n. Also, $|X_n| \leq Y$ implies that $X_n + Y \geq 0$, so by Fatou's lemma we get

$$\begin{aligned}\mathbb{E}Y + \liminf_{n\to\infty} \mathbb{E}X_n &= \liminf_{n\to\infty} \big(\mathbb{E}Y + \mathbb{E}X_n\big) = \liminf_{n\to\infty} \mathbb{E}(Y + X_n) \\ &\geq \mathbb{E}\big(\liminf_{n\to\infty}(Y + X_n)\big) = \mathbb{E}(Y + X) = \mathbb{E}Y + \mathbb{E}X.\end{aligned}$$

Subtracting $\mathbb{E}Y$ from both sides we obtain

$$\liminf_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E}X. \tag{21}$$

Now, $|X_n| \leq Y$ also implies that $-X_n + Y \geq 0$, so repeating the above argument with $-X_n$ we obtain

$$\liminf_{n\to\infty} \mathbb{E}(-X_n) \geq \mathbb{E}(-X),$$

equivalently

$$-\limsup_{n\to\infty} \mathbb{E}X_n \geq -\mathbb{E}X,$$

or equivalently

$$\limsup_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X. \tag{22}$$

Combining (21) with (22) we get

$$\mathbb{E}X \leq \liminf_{n\to\infty} \mathbb{E}(X_n) \leq \limsup_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X.$$

There are equalities throughout, showing that $\{\mathbb{E}X_n\}$ is a convergent sequence with limit $\mathbb{E}X$. □

Example 144. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \ldots$ be a sequence of disjoint subsets of Ω. If X is an integrable random variable, then

$$\int_A X \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_{A_i} X \, d\mathbb{P},$$

where $A := \cup_{i=1}^{\infty} A_i$. Indeed, by definition

$$\int_A X \, d\mathbb{P} = \int_\Omega X \mathbf{1}_A \, d\mathbb{P} = \int_\Omega \sum_{i=1}^{\infty} X \mathbf{1}_{A_i} \, d\mathbb{P}.$$

Define the sequence of random variables $X_n := \sum_{i=1}^{n} X \mathbf{1}_{A_i}$. Clearly, $\lim_{n\to\infty} X_n = X\mathbf{1}_A$ and $|X_n| \le |X|$. But X is integrable and so $|X|$ is integrable. Then, by the Dominated Convergence Theorem we have

$$\begin{aligned} \int_A X \, d\mathbb{P} = \mathbb{E}(X\mathbf{1}_A) = \lim_{n\to\infty} \mathbb{E}X_n = \lim_{n\to\infty} \mathbb{E}\Big(\sum_{i=1}^{n} X\mathbf{1}_{A_i}\Big) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}(X\mathbf{1}_{A_i}) \\ = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{A_i} X \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_{A_i} X \, d\mathbb{P}. \end{aligned}$$

□

Exercise 145. Let $X_1, X_2, \ldots$ be positive random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Show that

$$\sum_{k=1}^{\infty} \mathbb{E}X_k = \mathbb{E}\Big(\sum_{k=1}^{\infty} X_k\Big).$$

Example 146. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable on Ω taking the distinct values $\{x_1, x_2, \ldots\}$ with probabilities $p_1, p_2, \ldots$, (where naturally $p_i \in [0,1]$ and $\sum_{i=1}^{\infty} p_i = 1$). Define the subsets $A_i := X^{-1}(x_i) := \{\omega \in \Omega : X(\omega) = x_i\}$. Observe that the sets $A_1, A_2, \ldots$ are disjoint and that $\cup_{i=1}^{\infty} A_i = \Omega$ (indeed, for every $\omega \in \Omega$, we have that $X(\omega)$ is in $\{x_1, x_2, \ldots\}$ and hence ω is in one of the sets $\{A_n\}$. Thus, using the previous example, we have

$$\begin{aligned} \mathbb{E}X = \int_\Omega X \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_{A_i} X \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_\Omega X\mathbf{1}_{A_i} \, d\mathbb{P} = \sum_{i=1}^{\infty} \int_\Omega X(\omega)\mathbf{1}_{A_i}(\omega) \, d\mathbb{P}(\omega) \\ = \sum_{i=1}^{\infty} \int_\Omega x_i \mathbf{1}_{A_i}(\omega) \, d\mathbb{P}(\omega) = \sum_{i=1}^{\infty} x_i \mathbb{P}(A_i) = \sum_{i=1}^{\infty} x_i p_i. \end{aligned}$$

□

A random variable X that takes finitely many values, as in Example 132, or countably infinitely many, as in Example 146 is called a *discrete random variable.* The expectation of a discrete random variable calculated in these examples is given by a familiar formula.

Example 147. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$:=the Borel sets on $[0,1]$ and $\mathbb{P}$:= the Lebesgue measure on $[0,1]$. Define the random variable

$$X(\omega) := \begin{cases} 0 & \text{if } \omega \in \big[1/2, 1\big), \\ 2^{n+1}/n & \text{if } \omega \in \big[1/2^{n+1}, 3/2^{n+2}\big), \\ -2^{n+1}/n & \text{if } \omega \in \big[3/2^{n+2}, 1/2^{n}\big), \end{cases}$$

where $n = 1, 2, \ldots$ Thus, $\mathbb{P}(X = 0) = 1/2$, $\mathbb{P}(X = 2^{n+1}/n) = 3/2^{n+2} - 1/2^{n+1} = 1/2^{n+2}$, and $\mathbb{P}(X = -2^{n+1}/n) = 1/2^n - 3/2^{n+2} = 1/2^{n+2}$. We have

$$X^+(\omega) = \begin{cases} 0 & \text{if } \omega \in \big[1/2, 1\big), \\ 2^{n+1}/n & \text{if } \omega \in \big[1/2^{n+1}, 3/2^{n+2}\big), \\ 0 & \text{if } \omega \in \big[3/2^{n+2}, 1/2^{n}\big), \end{cases} \quad \text{and } X^-(\omega) = \begin{cases} 0 & \text{if } \omega \in \big[1/2, 1\big), \\ 0 & \text{if } \omega \in \big[1/2^{n+1}, 3/2^{n+2}\big), \\ 2^{n+1}/n & \text{if } \omega \in \big[3/2^{n+2}, 1/2^{n}\big), \end{cases}$$

Hence, $\mathbb{E}X^+ = \sum_{n=1}^{\infty} \frac{2^{n+1}}{n}\left(\frac{1}{2^{n+2}}\right) = \frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n} = \infty$. Similarly, one sees that $\mathbb{E}X^- = \infty$. Thus, $\mathbb{E}X$ does not exist.

3.4.1 Poisson's Theorem (optional)

Before we show the main theorem of this section, we need a fact from Calculus.

Lemma 148. For all $x \in \mathbb{R}$ with $|x| < 1$ we have

$$\log(1 + x) = x + o(x),$$

where $o(x)$ is a function such that $\lim_{x\to 0} o(x)/x = 0$.

The following lemma is a standard fact from Calculus as well, which we include with its proof.

Lemma 149. If $\lim\limits_{n\to\infty} a_n = a \in \mathbb{R}$ then $\lim\limits_{n\to\infty}\left(1 + \dfrac{a_n}{n}\right)^n = e^a$.

Proof. Since the logarithm function is invertible and continuous on $(0, \infty)$, we can take logarithms and try to prove the equivalent limit $\lim\limits_{n\to\infty} n \log\left(1 + \dfrac{a_n}{n}\right) = a$. Let $b_n := \frac{a_n}{n}$ and note that $\lim\limits_{n\to\infty} nb_n = a$. Since $\lim\limits_{n\to\infty} b_n = 0$, replacing x by b_n in Lemma 148 we obtain

$$\lim_{n\to\infty} n\log(1 + b_n) = \lim_{n\to\infty} n(b_n + o(b_n)) = \lim_{n\to\infty}\left(nb_n + nb_n\frac{o(b_n)}{b_n}\right) = a,$$

where we used that $\lim\limits_{n\to\infty}\dfrac{o(b_n)}{b_n} = 0$. □

Theorem 150 (Poisson's theorem). *Let $\lambda > 0$ and let $p_n \in (0,1)$ for all $n = 1, 2, \ldots$ If $\lim\limits_{n\to\infty} np_n = \lambda$, then*

$$\lim_{n\to\infty}\binom{n}{k}p_n^k(1 - p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!} \quad \textit{for all } k = 0, 1, 2, \ldots$$

Proof. Fix an integer $k \geq 0$. Then

$$\begin{aligned}\binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\ &= \frac{1}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} (np_n)^k (1-p_n)^{n-k}.\end{aligned}$$

Note that $\lim_{n\to\infty} (np_n)^k = \lambda^k$ and $\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} = 1$ so all that is left to show is $\lim_{n\to\infty} (1-p_n)^{n-k} = e^{-\lambda}$. Using $\lim_{n\to\infty} np_n = \lambda$ we conclude that $\lim_{n\to\infty} p_n = 0$, hence $\lim_{n\to\infty} (1-p_n)^{-k} = 1$ and we are left with proving that $\lim_{n\to\infty} (1-p_n)^n = e^{-\lambda}$. The last limit follows from Lemma 149 applied with $a_n := -np_n$ and after noting that $(1-p_n)^n = \left(1 + \frac{-np_n}{n}\right)^n$. □

Poisson theorem seems to have nothing to do with probability. But a deeper look reveals the following interpretation.

Corollary 151. Let X_n be a binomial random variable with parameters (n, p_n) and let Y be a Poisson random variable with parameter $\lambda > 0$. If $\lim_{n\to\infty} np_n = \lambda$, then

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(Y = k) \quad \text{for all } k = 0, 1, 2, \ldots$$

In the corollary, the random variables X_n and Y may not be defined on the same probability space. Since $\mathbb{E}X_n = np_n$ and $\mathbb{E}Y = \lambda$, the condition $\lim_{n\to\infty} np_n = \lambda$ can be re-written as

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}Y.$$

4 Advanced topics

4.1 Inequalities for random variables

Proposition 152 (Chebyshev's inequality)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X \geq 0$ be a random variable on Ω. Then, for all $\lambda > 0$ we have

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}X}{\lambda}.$$

Proof. Recall that $\mathbb{P}(X \geq \lambda)$ is a short-hand notation for $\mathbb{P}(\{\omega \in \Omega : X(\omega) \geq \lambda\})$. We simply have

$$\begin{aligned}\lambda\mathbb{P}(\{\omega \in \Omega : X(\omega) \geq \lambda\}) &= \lambda\mathbb{E}(\mathbf{1}_{\{\omega\in\Omega:X(\omega)\geq\lambda\}}) = \mathbb{E}(\lambda\mathbf{1}_{\{\omega\in\Omega:X(\omega)\geq\lambda\}}) \\ &\leq \mathbb{E}(X\mathbf{1}_{\{\omega\in\Omega:X(\omega)\geq\lambda\}}) \leq \mathbb{E}X.\end{aligned}$$

□

Definition 153. A function $f : \mathbb{R} \to \mathbb{R}$, defined on an open set $D \subset \mathbb{R}$ is *convex on D* if and only if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

for all $\alpha \in [0, 1]$ and all $x, y \in D$.

For the inequality in the definition to make sense, it is necessary that the point $\alpha x + (1-\alpha)y$ is in D all $\alpha \in [0,1]$ and all $x, y \in D$. That is, the domain D of a convex function has to be a convex set. It is a calculus fact that if $f(x)$ is twice continuously differentiable, then $f(x)$ is convex on D if and only if $f''(x) \geq 0$ for all $x \in D$.

Convex functions are a very good tool for proving inequalities.

Example 154. For any $x, y \geq 0$ and any positive numbers a, b with $a+b=1$ we have the inequality

$$x^a y^b \leq ax + by. \tag{23}$$

Indeed, if $x = 0$ or $y = 0$ the inequality holds trivially. So assume now that $x, y > 0$. The function $f(x) := -\log(x)$ is convex on $(0,\infty)$ (since $f''(x) = 1/x^2 \geq 0$ for all $x \in (0,\infty)$), hence

$$-\log(ax+by) \leq a(-\log(x))+b(-\log(y)) = -(a\log(x)+b\log(y)) = -(\log(x^a)+\log(y^b)) = -\log(x^a y^b).$$

The inequality follows. □

Example 155. For any $x, y \geq 0$ and any $p \geq 1$ we have the inequality

$$(x+y)^p \leq 2^{p-1}(x^p + y^p). \tag{24}$$

The inequality is trivial if $x = 0$ or $y = 0$. So, suppose $x, y > 0$. Since the function $f(x) = x^p$ is convex on $(0,\infty)$, we have

$$\left(\frac{x+y}{2}\right)^p = f((1/2)x + (1/2)y) \leq (1/2)(f(x) + f(y)) = \frac{x^p + y^p}{2}.$$

Multiplying the beginning and the end by 2^p gives (24).

It is a fact that every convex function $f : \mathbb{R} \to \mathbb{R}$ is continuous and thus measurable as a function from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It is a fact that $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if for every $x_0 \in \mathbb{R}$ there is a *supporting line* to the graph of $f(x)$ at the point $(x_0, f(x_0))$. That means, for every $x_0 \in \mathbb{R}$ there are constants $a, b \in \mathbb{R}$ such that $ax + b \leq f(x)$ for all $x \in \mathbb{R}$ and $ax_0 + b = f(x_0)$.

If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking only two values x and y with probabilities α and $(1-\alpha)$ respectively, then by the definition of convexity we have

$$f(\mathbb{E}X) = f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) = \mathbb{E}f(X).$$

Turns out this holds in general.

Proposition 156 (Jensen's inequality)**.** If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}|X| < \infty$, then

$$f(\mathbb{E}X) \leq \mathbb{E}f(X),$$

where the expected value on the right-hand side maybe infinity.

Proof. Let $x_0 := \mathbb{E}X$. Let $ax + b$ be the supporting line at the point $(x_0, f(x_0))$ on the graph of $f(x)$. That is, $ax + b \leq f(x)$ for all $x \in \mathbb{R}$ and $ax_0 + b = f(x_0)$. It follows that $aX(\omega) + b \leq f(X(\omega))$ for all $\omega \in \Omega$ and hence $f(\mathbb{E}X) = a\mathbb{E}X + b = \mathbb{E}(aX + b) \leq \mathbb{E}(f(X))$. □

Exercise 157 (Jensen generalized)**.** If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $X_1, \ldots, X_n$ are random variables. Show that $f(\mathbb{E}X_1, \ldots, \mathbb{E}X_n) \leq \mathbb{E}f(X_1, \ldots, X_n)$ provided that $\mathbb{E}|X_i| < \infty$ for all $i = 1, \ldots, n$. Hint: use the fact that at every $x_0 = (x_{0,1}, \ldots, x_{0,n}) \in \mathbb{R}^n$ there is a supporting hyperplane to the graph of $f(x)$ at the point $(x_0, f(x_0))$. That means, that there is a vector $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$ such that $a_1x_1 + \cdots + a_nx_n + b \leq f(x_1, \ldots, x_n)$ for every vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $a_1x_{0,1} + \cdots + a_nx_{0,n} + b = f(x_{0,1}, \ldots, x_{0,n})$.

Example 158. (i) The function $f(x) = |x|$ is convex, so for every integrable random variable X we have $|\mathbb{E}X| \leq \mathbb{E}|X|$.

(ii) For any $p \in [1, \infty)$ the function $f(x) = |x|^p$ is convex, so for any integrable random variable X, applying Jensen's inequality to $|X|$ we obtain

$$(\mathbb{E}|X|)^p \leq \mathbb{E}(|X|^p).$$

□

The second example above is a particular case of the following more general inequality.

Proposition 159 (Hölder's inequality)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X, Y be two random variables. If $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|Y|^q)^{\frac{1}{q}},$$

provided that the product on the right-hand side is not $0 \cdot \infty$ or $\infty \cdot 0$.

Proof. If $\mathbb{E}|X|^p = 0$ then $|X|^p = 0$ a.s. and so $XY = 0$ a.s. implying that $\mathbb{E}|XY| = 0$. Thus, in this case the inequality holds trivially. Similarly if $\mathbb{E}|Y|^q = 0$. So assume now that $\mathbb{E}|X|^p > 0$ and $\mathbb{E}|Y|^q > 0$. If $\mathbb{E}|X|^p = \infty$ or $\mathbb{E}|Y|^q = \infty$ then the inequality holds trivially since the right hand side is infinity. So suppose now that $\mathbb{E}|X|^p, \mathbb{E}|Y|^q \in (0, \infty)$. Define new random variables

$$\tilde{X} := \frac{X}{(\mathbb{E}|X|^p)^{\frac{1}{p}}} \quad \text{and} \quad \tilde{Y} := \frac{Y}{(\mathbb{E}|Y|^q)^{\frac{1}{q}}}.$$

Setting $x := |\tilde{X}|^p$, $y := |\tilde{Y}|^q$, $a := \frac{1}{p}$, and $b := \frac{1}{q}$ and substituting those values in inequality (23) we get

$$|\tilde{X}\tilde{Y}| = x^a y^b \leq ax + by = \frac{1}{p}|\tilde{X}|^p + \frac{1}{q}|\tilde{Y}|^q$$

and taking expectations we obtain

$$\mathbb{E}|\tilde{X}\tilde{Y}| \leq \frac{1}{p}\mathbb{E}|\tilde{X}|^p + \frac{1}{q}\mathbb{E}|\tilde{Y}|^q = \frac{1}{p} + \frac{1}{q} = 1.$$

On the other side, we have

$$\mathbb{E}|\tilde{X}\tilde{Y}| = \frac{\mathbb{E}|XY|}{(\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|Y|^q)^{\frac{1}{q}}}.$$

Combining the two displayed lines we establish the required inequality. □

Exercise 160. For $0 < p < q < \infty$ we have that $(\mathbb{E}|X|^p)^{1/p} \leq (\mathbb{E}|X|^q)^{1/q}$

Exercise 160 shows that the function $t \mapsto (\mathbb{E}|X|^t)^{1/t}$ is increasing for $t \in (0, \infty)$, thus it has a limit at infinity, denoted by

$$\|X\|_\infty := \lim_{t\to\infty} (\mathbb{E}|X|^t)^{1/t}. \tag{25}$$

Exercise 161. Show that $\|X\|_\infty = \inf\{x \in \mathbb{R} : \mathbb{P}(|X| > x) = 0\}$. The number on the right-hand side is called the *essential supremum* of X.

With definition (25), one can take the limit as q approaches infinity in Hölder's inequality to see that it holds when $p = 1$ and $q = \infty$ (resp. $p = \infty$ and $q = 1$) provided that in the limit the product on the right-hand side is not $0 \cdot \infty$ (resp. $\infty \cdot 0$).

Corollary 162. Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers. If $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\sum_{n=1}^{\infty} |a_n b_n| \le \Big(\sum_{n=1}^{\infty} |a_n|^p\Big)^{\frac{1}{p}} \Big(\sum_{n=1}^{\infty} |b_n|^q\Big)^{\frac{1}{q}},$$

provided that the product on the right-hand side is not $0 \cdot \infty$ or $\infty \cdot 0$.

Proof. Consider the first N elements of the sequences $\{a_n\}$ and $\{b_n\}$. Let $\Omega = \{1, 2, \ldots, N\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(\{k\}) = 1/N$. Then the functions $X, Y : \Omega \to \mathbb{R}$ defined by $X(k) := a_k$ and $Y(k) = b_k$ are (discrete) random variables. By Hölder's inequality we get

$$\frac{1}{N}\sum_{n=1}^{N} |a_n b_n| \le \Big(\frac{1}{N}\sum_{n=1}^{N} |a_n|^p\Big)^{\frac{1}{p}} \Big(\frac{1}{N}\sum_{n=1}^{N} |b_n|^q\Big)^{\frac{1}{q}}.$$

Multiplying both sides by N and letting N approach infinity proves the result. □

Corollary 163. Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers. Let $\{p_n\}$ be a sequence of non-negative numbers with $\sum_{n=1}^{\infty} p_n = 1$. If $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\sum_{n=1}^{\infty} p_n |a_n b_n| \le \Big(\sum_{n=1}^{\infty} p_n |a_n|^p\Big)^{\frac{1}{p}} \Big(\sum_{n=1}^{\infty} p_n |b_n|^q\Big)^{\frac{1}{q}},$$

provided that the product on the right-hand side is not $0 \cdot \infty$ or $\infty \cdot 0$.

Proof. Let $\Omega = \{1, 2, \ldots\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(\{k\}) = p_k$. Then the functions $X, Y : \Omega \to \mathbb{R}$ defined by $X(k) := a_k$ and $Y(k) = b_k$ are (discrete) random variables. By Example 146 and Hölder's inequality we get the reslt. □

Theorem 164 (Minkowski's inequality)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X, Y be two random variables. If $p \in [1, \infty)$, then

$$\big(\mathbb{E}|X + Y|^p\big)^{\frac{1}{p}} \le (\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}.$$

Proof. If $\mathbb{E}|X|^p = \infty$ or $\mathbb{E}|Y|^p = \infty$, then the inequality holds trivially with infinity right-hand side, so assume that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^p < \infty$.

When $p = 1$ the inequality follows from $|X + Y| \le |X| + |Y|$. So, assume that $p \in (1, \infty)$ and take a $q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$. That is $q = \frac{p}{p-1}$, and continue:

$$\begin{aligned}\mathbb{E}|X + Y|^p &= \mathbb{E}(|X + Y||X + Y|^{p-1}) \le \mathbb{E}\big((|X| + |Y|)|X + Y|^{p-1}\big)\\ &= \mathbb{E}(|X||X + Y|^{p-1}) + \mathbb{E}(|Y||X + Y|^{p-1})\\ &\le (\mathbb{E}|X|^p)^{\frac{1}{p}}\big(\mathbb{E}|X + Y|^{(p-1)q}\big)^{\frac{1}{q}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}\big(\mathbb{E}|X + Y|^{(p-1)q}\big)^{\frac{1}{q}},\end{aligned}$$

where for the last inequality, we used the Hölder's inequality. Finally, since $(p - 1)q = p$, dividing both sides by $(\mathbb{E}|X + Y|^p)^{\frac{1}{q}}$ concludes the proof.

But why can we divide by $(\mathbb{E}|X + Y|^p)^{\frac{1}{q}}$? What if it is 0 or ∞? If $\mathbb{E}|X + Y|^p = 0$ then Minkowski's inequality holds trivially. If $\mathbb{E}|X + Y|^p = \infty$ then we need to prove Minkowski's inequality by different means. Indeed, by (24), we get

$$|X + Y|^p \le (|X| + |Y|)^p \le 2^{p-1}(|X|^p + |Y|^p)$$

and taking expectations of both sides

$$\mathbb{E}|X + Y|^p \le 2^{p-1}(\mathbb{E}|X|^p + \mathbb{E}|Y|^p)$$

shows that $\mathbb{E}|X|^p + \mathbb{E}|Y|^p = \infty$ which in its turn implies that $(\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}} = \infty$. So, Minkowski's inequality holds again with both sides equal to infinity. □

4.2 Change of variables and Fubini's theorem

In this section we prove a *change of variable formula* of the integral $\mathbb{E}X = \int_\Omega X\, d\mathbb{P}$. In many cases, this is the only formula that allows us to explicitly compute the expected value $\mathbb{E}X$. This formula allows us to convert any integral of the form $\int_\Omega X\, d\mathbb{P}$ into an integral over the real line $\mathbb{R}$ of the form $\int_\mathbb{R} x\, dF(x)$. It is important that the reader recalls Exercise 103.

Theorem 165. Assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and let $(S, \mathcal{S})$ be a measurable space. Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable function and let $g : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function such that either $g \ge 0$ or $\mathbb{E}|g(X)| < \infty$. Then, we have

$$\int_\Omega g(X(\omega))\, d\mathbb{P}(\omega) = \int_S g(s)\, d\mathbb{P}_X(s). \tag{26}$$

If one of the integrals exists then the other exists and their values are equal.

Before we prove the proposition, let us look at a particular case.

Example 166. *If we take $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, that is, if X is a random variable, then*

$$\int_\Omega g(X(\omega))\, d\mathbb{P}(\omega) = \int_\mathbb{R} g(s)\, d\mathbb{P}_X(s),$$

where by Proposition 106, the law of X, denoted $\mathbb{P}_X(s)$, is equal to the measure determined by the c.d.f. of X, denoted $F_X(x)$. The last integral $\int_\mathbb{R} g(s)\, d\mathbb{P}_X(s)$ is often written as $\int_\mathbb{R} g(s)\, dF_X(s)$. Hence, computing the expectation $\mathbb{E}g(X)$ reduces to calculating an integral over the real line $\mathbb{R}$ with a measure on the Borel sets determined by the c.d.f., $F_X(s)$ of X.

Proof of Theorem 165. (1) Suppose first that $g(s) \geq 0$ for all $s \in S$. Let g_n be a sequence of step functions on S such that $0 \leq g_n \uparrow g$. Such a sequence exists by Exercise 99. It should be clear that then $0 \leq g_n(X(\omega)) \uparrow g(X(\omega))$ for all $\omega \in \Omega$ and by Proposition 96 $g_n(X(\omega))$ is a step function on Ω for all $n = 1, 2, \ldots$ Thus, if we show that

$$\int_\Omega g_n(X(\omega))\, d\mathbb{P}(\omega) = \int_S g_n(s)\, d\mathbb{P}_X(s) \qquad \text{for all } n = 1, 2, \ldots,$$

then by Lemma 137, taking limits of both sides as n approaches infinity, we obtain (26). It is enough to check the last equality for $g_n(s) = \mathbf{1}_A(s)$ for some $A \in \mathcal{S}$. (Indeed, then one can multiply $\mathbf{1}_A(s)$ by a real number α and take sums to obtain the equality for general step function $\sum_{i=1}^k \alpha_i \mathbf{1}_{A_i}(s)$.) Then, we obtain

$$\begin{aligned}\int_\Omega g_n(X(\omega))\, d\mathbb{P}(\omega) &= \int_\Omega \mathbf{1}_A(X(\omega))\, d\mathbb{P}(\omega) = \int_\Omega \mathbf{1}_{X^{-1}(A)}(\omega)\, d\mathbb{P}(\omega) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}_X(A) \\ &= \int_S \mathbf{1}_A(s)\, d\mathbb{P}_X(s) = \int_S g_n(s)\, d\mathbb{P}_X(s).\end{aligned}$$

(2) Suppose now that $g(s)$ is arbitrary, then note that $g(X(\omega)) = g^+(X(\omega)) - g^-(X(\omega))$. Repeating the above arguments for the non-negative functions $g^+(s)$ and $g^-(s)$ and using the additivity of the integral, we conclude the proof. The fact that $\mathbb{E}|g(X)| < \infty$ is used to ensure that $\mathbb{E}g^+(X) < \infty$ and $\mathbb{E}g^-(X) < \infty$. □

The following proposition is stated without proof. It must look familiar from elementary probability courses.

Proposition 167. Assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable and let $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function such that either $g \geq 0$ or $\mathbb{E}|g(X)| < \infty$. If the c.d.f. of X has density, that is

$$F_X(x) = \int_{-\infty}^x f(s)\, ds,$$

then, we have

$$\int_\Omega g(X(\omega))\, d\mathbb{P}(\omega) = \int_\mathbb{R} g(x) f(x)\, dx. \tag{27}$$

If one of the integrals exists then the other exists and their values are equal.

The proof of Proposition 167 follows along the same lines as the proof of Theorem 165, only that this time a few extra steps are required since the integral on the right-hand side of (27) is not over a probability space and its formal definition requires a few extra steps. Yet, we can work with the result, since in the majority of cases, the integral on the right-hand side of (27) is a Riemann integral.

What Proposition 167 says is that when the c.d.f. F_X has density $f(s)$ then

$$\int_\mathbb{R} g(s)\, dF_X(s) = \int_\mathbb{R} g(s) f(s)\, ds. \tag{28}$$

Informally, we write $dF_X(s) = f(s)ds$.

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two measurable spaces (finite or σ-finite) and recall the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$.

Theorem 168 (Fubini's theorem)**.** Let $X : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a measurable function such that either $X \geq 0$ or

$$\int_{\Omega_1 \times \Omega_2} |X(\omega_1, \omega_2)| \, d(\mathbb{P}_1 \times \mathbb{P}_2)(\omega_1, \omega_2) < \infty$$

then

(i) There are $\mathbb{P}_i$-measurable, non-negative (resp. integrable) functions $h_i(\omega_i)$, $i = 1, 2$, such that

$$\begin{aligned} h_1(\omega_1) &= \int_{\Omega_2} X(\omega_1, \omega_2) \, d\mathbb{P}_2(\omega_2) \quad \mathbb{P}_1\text{-almost surely and} \\ h_2(\omega_2) &= \int_{\Omega_1} X(\omega_1, \omega_2) \, d\mathbb{P}_1(\omega_1) \quad \mathbb{P}_2\text{-almost surely} \end{aligned}$$

(ii) Integrals with respect to the product measure $\mathbb{P}_1 \times \mathbb{P}_2$ can be evaluated iteratively:

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) \, d(\mathbb{P}_1 \times \mathbb{P}_2)(\omega_1, \omega_2) &= \int_{\Omega_2} \Big(\int_{\Omega_1} X(\omega_1, \omega_2) \, d\mathbb{P}_1(\omega_1) \Big) \, d\mathbb{P}_2(\omega_2) \\ &= \int_{\Omega_1} \Big(\int_{\Omega_2} X(\omega_1, \omega_2) \, d\mathbb{P}_2(\omega_2) \Big) \, d\mathbb{P}_1(\omega_1). \end{aligned}$$

Corollary 169. *Let $X \geq 0$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then*

$$\begin{aligned} \mathbb{E}X &= \int_\Omega X(\omega) \, dP(\omega) = \int_\Omega \int_0^\infty \mathbf{1}_{[0, X(\omega))}(x) \, dx \, dP(\omega) \\ &= \int_0^\infty \int_\Omega \mathbf{1}_{[0, X(\omega))}(x) \, dP(\omega) \, dx = \int_0^\infty \int_\Omega \mathbf{1}_{\{\omega \in \Omega : X(\omega) > x\}}(\omega) \, dP(\omega) \, dx \\ &= \int_0^\infty \mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) \, dx = \int_0^\infty \mathbb{P}(X > x) \, dx = \int_0^\infty (1 - F_X(x)) \, dx. \end{aligned}$$

Exercise 170. *Let $X \geq 0$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $p > 0$, then*

$$\mathbb{E}X^p = \int_0^\infty p x^{p-1} \mathbb{P}(X > x) \, dx.$$

4.3 Independence

Recal Definition 62. The collections of sets $\{\mathcal{F}_i\}_{i \in I}$ from $\mathcal{F}$, where I is an index set, are *independent* if for any $n \in \mathbb{N}$, any distinct indexes $i_1, \ldots, i_n \in I$, and any $A_{i_k} \in \mathcal{F}_{i_k}$, $k = 1, 2, \ldots, n$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_n}). \tag{29}$$

This definition simplifies when the index set $I = \{1, \dots, N\}$ is finite and when $\mathcal{F}_i$ contains Ω for all $i = 1, \dots, N$. Then, the collections $\{\mathcal{F}_i\}_{i=1}^N$ are independent if and only if

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_N) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdots \mathbb{P}(A_N) \tag{30}$$

for any $A_i \in \mathcal{F}_i$, $i = 1, \dots, N$. That is because in (30) we can let some of the sets $A_i := \Omega$ and since $\mathbb{P}(\Omega) = 1$, we can obtain (29) for any distinct indexes $i_1, \dots, i_n \in \{1, \dots, N\}$.

The random variables $\{X_i\}_{i \in I}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are *independent* if for any $n \in \mathbb{N}$, any distinct indexes $i_1, \dots, i_n \in I$, and any $A_k \in \mathcal{B}(\mathbb{R})$, $k = 1, 2, \dots, n$, we have

$$\mathbb{P}(X_{i_1} \in A_1, X_{i_2} \in A_2, \dots, X_{i_n} \in A_n) = \mathbb{P}(X_{i_1} \in A_1)\mathbb{P}(X_{i_2} \in A_2) \cdots \mathbb{P}(X_{i_n} \in A_n).$$

At first glance these two definitions are quite different. In order to reconcile them, recall Exercise 102, explaining how every random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ generates a σ-algebra $\sigma(X)$, contained in $\mathcal{F}$.

Exercise 171. Show that random variables $\{X_i\}_{i \in I}$ are independent if and only if the collection of σ-algebras $\{\sigma(X)\}_{i \in I}$ are independent.

We are going to use the next theorem without a proof.

Theorem 172. Suppose the collection of sets $\mathcal{F}_1, \dots, \mathcal{F}_n$ are independent and suppose every $\mathcal{F}_i$ contains Ω and is closed under intersection (that means: for every $A, B \in \mathcal{F}_i$ we have $A \cap B \in \mathcal{F}_i$). Then $\sigma(\mathcal{F}_1)$,...,$\sigma(\mathcal{F}_n)$ are independent.

Definition 173. If $X_1, \dots, X_n$ are random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ then the function $F : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F(x_1, \dots, x_n) := \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n)$$

is called the *joint cumulative distribution function* of $X_1, \dots, X_n$.

Theorem 174. *Random variables $X_1, \dots, X_n$ are independent if and only if*

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

for all $x_1, \dots, x_n \in (-\infty, \infty]$.

Proof. If $X_1, \dots X_n$ are independent then the equality clearly holds. For the opposite direction, suppose that the equality holds

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n).$$

This means that the collections of sets $\mathcal{F}_i := \big\{\{X_i \le x\} : x \in \mathbb{R} \cup \{\infty\}\big\}$, $i = 1, \dots, n$ are independent. Note that each $\mathcal{F}_i$ contains Ω and since

$$\{X_i \le x\} \cap \{X_i \le y\} = \{X_i \le \min\{x, y\}\} \in \mathcal{F}_i$$

we see that each $\mathcal{F}_i$ is closed under intersection. By Theorem 172, we see that $\sigma(\mathcal{F}_1), \dots, \sigma(\mathcal{F}_n)$ are independent.

By Exercise 102, the sets $\mathcal{F}_i = \big\{\{X_i \le x\} : x \in \mathbb{R} \cup \{\infty\}\big\}$ generate the σ-algebra $\sigma(X_i)$, that is $\sigma(\mathcal{F}_i) = \sigma(X_i)$. Hence, the σ-algebras $\sigma(X_1)$,...,$\sigma(X_n)$ are independent, and that is equivalent to $X_1, \dots, X_n$ being independent, according to Exercise 171. □

Exercise 175. Let $X_1, \ldots, X_n$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose the joint cumulative distribution function $F(x_1, \ldots, x_n)$ can be expressed as

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(s_1, \ldots, s_n)\, ds_1 \cdots ds_n,$$

for some measurable function $f : \mathbb{R}^s \to [0, \infty)$. Such an f is called *joint density* of $X_1, \ldots, X_n$.

Show that $X_1, \ldots, X_n$ are independent if and only if $f(x_1, \ldots, x_n) = g_1(x_1) \cdots g_n(x_n)$, for some measurable functions $g_i \geq 0$, $i = 1, \ldots, n$. In such case, show that up to a multiplicative constant, g_i is the density of X_i.

Given n random variables $X_1, \ldots, X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we obtain n probability spaces $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_{X_i})$, $i = 1, \ldots, n$. Their product, see Subsection 2.2.2, is the probability space

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}). \tag{31}$$

On the other hand, again applying Exercise 103, but this time to the measurable function $(X_1, \ldots, X_n) : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, defines the probability space

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_{(X_1, \ldots, X_n)}), \tag{32}$$

where

$$\mathbb{P}_{(X_1, \ldots, X_n)}(B) = \mathbb{P}\big((X_1, \ldots, X_n) \in B\big) \text{ for all } B \in \mathcal{B}(R^n)$$

is the image of $\mathbb{P}$ under the map $(X_1, \ldots, X_n)$, or the law of $(X_1, \ldots, X_n)$.

The probability spaces (31) and (32) differ only by their measures, which, in general are not equal, unless the random variables $X_1, \ldots, X_n$ are independent.

Remark 176. Theorem 74 describes the measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by the c.d.f F_X of X and Proposition 106 shows that it is equal to $\mathbb{P}_X$, the law of X. Analogously, the joint cumulative distribution function of $X_1, \ldots, X_n$ given in Definition 173 defines a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. The details are beyond the scope of these notes, but it suffices to say that, in the general case, this measure is equal to $\mathbb{P}_{(X_1, \ldots, X_n)}$, the law of $(X_1, \ldots, X_n)$.

For example, when $n = 2$, let $F(x_1, x_2)$ be the joint c.d.f. of X_1, X_2. One can define a measure $\mathbb{P}$ on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ such that on rectangles $(a_1, b_1] \times (a_2, b_2]$ its value is

$$\mathbb{P}((a_1, b_1] \times (a_2, b_2]) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$$

and it can be shown that this measure is exactly $\mathbb{P}_{(X_1, X_2)}$, the law of (X_1, X_2). □

Theorem 177. If the random variables $X_1, \ldots, X_n$ are independent then the measures $\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$ and $\mathbb{P}_{(X_1, \ldots, X_n)}$ are equal on $\mathcal{B}(\mathbb{R}^n)$.

Proof. Let us show first that the measures are equal on the rectangles $A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{B}(\mathbb{R})$ for all $i = 1, \ldots, n$. Indeed

$$\begin{aligned}\mathbb{P}_{(X_1, \ldots, X_n)}(A_1 \times \cdots \times A_n) &:= \mathbb{P}\big((X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n\big) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) \\ &= \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n) = \mathbb{P}_{X_1}(A_1) \cdots \mathbb{P}_{X_n}(A_n)\end{aligned}$$

$$= (\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n})(A_1 \times \cdots \times A_n),$$

where we used the fact that $X_1, \dots, X_n$ are independent. This shows that the two measures coincide on all sets in $\mathcal{B}(\mathbb{R}^n)$ that are finite union of disjoint rectangles. But those sets form an algebra which generates $\mathcal{B}(\mathbb{R}^n)$, hence by the uniqueness of extension in the Carathéodory extension theorem, the two measures must coinside everywhere on $\mathcal{B}(\mathbb{R}^n)$. □

The last theorem which appears unnecessarily theoretical, allows us to compute expectations of functions of several independent random variables by computing consecutively several integrals.

Theorem 178. Suppose X and Y are independent random variables with laws $\mathbb{P}_X$ and $\mathbb{P}_Y$. If $h : \mathbb{R}^2 \to \mathbb{R}$ is a measurable function such that either $h \geq 0$ or $\mathbb{E}|h(X,Y)| < \infty$, then

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y)\, d\mathbb{P}_X(x)\, d\mathbb{P}_Y(y). \tag{33}$$

Proof. Consider the measurable functions $(X,Y) : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ and $h : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The conditions of Proposition 165 are satisfied and we obtain from it

$$\mathbb{E}h(X,Y) = \int_{\Omega} h(X,Y)\, d\mathbb{P} = \int_{\mathbb{R}^2} h(x,y)\, d\mathbb{P}_{(X,Y)}.$$

By Theorem 177, the independence of X and Y imply that $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \times \mathbb{P}_Y$ so we get

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\, d(\mathbb{P}_X \times \mathbb{P}_Y)(x,y).$$

Now we are in the realm of the Fubini's theorem which gives

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\, d(\mathbb{P}_X \times \mathbb{P}_Y)(x,y) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y)\, d\mathbb{P}_X(x)\, d\mathbb{P}_Y(y).$$

□

Remark 179. Another notation for the integral (33) is

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} h(x,y)\, dF_X(x) \Big)\, dF_Y(y),$$

where F_X and F_Y are the c.d.f.'s of X and Y respectively. In particular, if F_X and F_Y have densities:

$$F_X(s) = \int_{-\infty}^{s} f(x)\, dx \text{ and } F_Y(s) = \int_{-\infty}^{s} g(y)\, dy$$

then consulting with (28) we get

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} h(x,y) f(x) g(y)\, dx \Big)\, dy,$$

which must be a well-known formula from elementary probability courses. □

Remark 180. Note the first three lines in the proof of Theorem 178. They do not use the fact that X and Y are independent. Thus, in the case when X and Y are not independent, the only way to calculate the expectation $\mathbb{E}h(X,Y)$ is

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\,d\mathbb{P}_{(X,Y)} = \int_{\mathbb{R}^2} h(x,y)\,dF(x,y),$$

where in the last integral $F(x,y)$ is the joint c.d.f. of X, Y and $dF(x,y)$ denotes the measure that it defines on $\mathbb{R}^2$, which we know is the same as the measure $\mathbb{P}_{(X,Y)}$. In particular, if the joint c.d.f. $F(x,y)$ has density

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy,$$

then

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)f(x,y)\,dx\,dy.$$

The last formula requires a proof but we will not do that here. □

Exercise 181 (Jensen reloaded)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function convex in each variable separately, that is

$$f(x_1,\ldots,\alpha y_i + (1-\alpha)z_i,\ldots,x_n) \le \alpha f(x_1,\ldots,y_i,\ldots,x_n) + (1-\alpha)f(x_1,\ldots,z_i,\ldots,x_n)$$

for every $\alpha \in [0,1]$, $y_i, z_i \in \mathbb{R}$, and every $i = 1,\ldots,n$. Let $X_1,\ldots,X_n$ be independent random variables. Show that $f(\mathbb{E}X_1,\ldots,\mathbb{E}X_n) \le \mathbb{E}f(X_1,\ldots,X_n)$ provided that $\mathbb{E}|f(X_1,\ldots,X_n)| < \infty$ and $\mathbb{E}|X_i| < \infty$ for all $i = 1,\ldots,n$.

Exercise 182. Show that if $X \ge 0$, then $\int_{\mathbb{R}} x1_{[0,\infty)}(x)\,dP_X(x) = \int_{\mathbb{R}} x\,dP_X(x)$. Hint: apply the change of variable formula to both sides.

Theorem 183. *If $X_1,\ldots,X_n$ are independent random variables on $(\Omega,\mathcal{F},\mathbb{P})$ and either a) $X_i \ge 0$ for all $i = 1,\ldots,n$ or b) $\mathbb{E}|X_i| < \infty$ for all $i = 1,\ldots,n$, then*

$$\mathbb{E}(X_1X_2\cdots X_n) = \mathbb{E}X_1\mathbb{E}X_2\cdots\mathbb{E}X_n.$$

Proof. The proof is by induction on n. We first verify the case $n = 2$. Suppose $X_1 \ge$ and $X_2 \ge 0$. Apply Theorem 178 to the positive function $h(x,y) := xy1_{[0,\infty)}(x)1_{[0,\infty)}(y)$ to obtain

$$\begin{aligned}
\mathbb{E}(X_1X_2) = \mathbb{E}h(X_1,X_2) &= \int_{\mathbb{R}}\Big(\int_{\mathbb{R}} xy1_{[0,\infty)}(x)1_{[0,\infty)}(y)\,dP_{X_1}(x)\Big)\,dP_{X_2}(y)\\
&= \int_{\mathbb{R}} y1_{[0,\infty)}(y)\Big(\int_{\mathbb{R}} x1_{[0,\infty)}(x)\,dP_{X_1}(x)\Big)\,dP_{X_2}(y)\\
&= \int_{\mathbb{R}} y1_{[0,\infty)}(y)\Big(\int_{\mathbb{R}} x\,dP_{X_1}(x)\Big)\,dP_{X_2}(y)\\
&= \int_{\mathbb{R}} y1_{[0,\infty)}(y)\mathbb{E}X_1\,dP_{X_2}(y) = \mathbb{E}X_1\int_{\mathbb{R}} y1_{[0,\infty)}(y)\,dP_{X_2}(y)\\
&= \mathbb{E}X_1\int_{\mathbb{R}} y\,dP_{X_2}(y) = \mathbb{E}X_1\mathbb{E}X_2,
\end{aligned}$$

where we used Exercise 182 twice.

Suppose now, $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}|X_2| < \infty$. First, repeat the above steps with the positive independent random variables $|X_1|$ and $|X_2|$ to obtain that $\mathbb{E}|X_1X_2| = \mathbb{E}|X_1||X_2| = \mathbb{E}|X_1|\mathbb{E}|X_2| < \infty$. Second, let $h(x,y) = xy$ and note that $\mathbb{E}|h(X_1,X_2)| = \mathbb{E}|X_1X_2| < \infty$. This verifies that the integrability condition in Theorem 178 holds. Repeat the above calculations one more time with $h(x,y) = xy$ to conclude.

Suppose the result holds for any $n-1$ independent random variables that satisfy condition a) or b). Then, if $X_1, \dots, X_n$ are independent and satisfy condition a) or b), we have

$$\mathbb{E}(X_1X_2\cdots X_n) = \mathbb{E}X_1\mathbb{E}(X_2\cdots X_n) = \mathbb{E}X_1\mathbb{E}X_2\cdots\mathbb{E}X_n,$$

where in the first equality we used the case $n=2$, together with the fact that X_1 and $X_2\cdots X_n$ are independent (see Lemma 271 in Appendix A), while in the second equality we used the induction hypothesis. □

The theorem implies, that if X_1 and X_2 are independent and integrable, then X_1X_2 is integrable. This is not true if independence is removed. For example take $X_1 = X_2 = 1/\sqrt{\omega}$ on $(0,1)$ with the Borel sets and the Lebesgue measure. Then, X_1 and X_2 are not independent, $\mathbb{E}X_1 = \mathbb{E}X_2 = 2$, but $\mathbb{E}X_1X_2 = \infty$ (why?).

Corollary 184. Suppose X and Y are independent random variables and $f, g : \mathbb{R} \to \mathbb{R}$ are measurable functions such that either a) $f \geq 0$ and $g \geq 0$; or b) $\mathbb{E}|f(X)| < \infty$ and $\mathbb{E}|g(Y)| < \infty$. Then

$$\mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(Y).$$

Proof. By Lemma 271 in Appendix A, we have that $f(X)$ and $g(Y)$ are independent random variables. □

Example 185. This example shows that it is possible to have $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$ even though X and Y are not independent. Let the joint distribution of X and Y be as follows

		Y		
		1	0	-1
X	1	0	a	0
	0	b	c	b
	-1	0	a	0

where the numbers a, b, c are strictly positive with $2a + 2b + c = 1$. The random variables X and Y are dependent, since

$$0 = \mathbb{P}(X = 1, Y = 1) \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = ab > 0.$$

Show that $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$. □

Definition 186. Two random variables X and Y with $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$ are called *uncorrelated* if $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Remark 187. The reason why conditions $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$ are required is so that all expectations are finite. Indeed, on the one hand, by Exercise 160 with $p = 1$ and $q = 2$ we get

$$\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{\frac{1}{2}} < \infty \quad \text{and} \quad \mathbb{E}|Y| \leq (\mathbb{E}|Y|^2)^{\frac{1}{2}} < \infty,$$

hence X and Y are integrable. On the other hand, by the Hölder's inequality with $p = q = 2$ we have

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^2)^{\frac{1}{2}} (\mathbb{E}|Y|^2)^{\frac{1}{2}} < \infty,$$

hence the random variable XY is integrable. □

The next theorem gives a formula for the cumulative distribution function of the sum of two independent random variables.

Theorem 188. If X and Y are independent random variables, then

$$\mathbb{P}(X + Y \leq z) = \int_{\mathbb{R}} F_X(z - y)\, dF_Y(y).$$

Proof. Fix a value of z and let $h(x, y) := \mathbf{1}_{\{(x,y)\in\mathbb{R}^2 : x+y\leq z\}}(x, y)$. We have

$$\begin{aligned}\mathbb{P}(X + Y \leq z) &= \int_{\Omega} \mathbf{1}_{\{\omega\in\Omega : X(\omega)+Y(\omega)\leq z\}}(\omega)\, d\mathbb{P}(\omega) = \int_{\Omega} h(X(\omega), Y(\omega))\, d\mathbb{P}(\omega) = \mathbb{E}(h(X, Y)) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x, y)\, d\mathbb{P}_X(x)\, d\mathbb{P}_Y(y),\end{aligned}$$

where in the last equality we used Theorem 178. Next, we evaluate the inside integral. For any fixed y we have

$$\begin{aligned}\int_{\mathbb{R}} h(x, y)\, d\mathbb{P}_X(x) &= \int_{\mathbb{R}} \mathbf{1}_{\{x\in\mathbb{R} : x\leq z-y\}}(x)\, d\mathbb{P}_X(x) = \int_{\mathbb{R}} \mathbf{1}_{(-\infty,z-y]}(x)\, d\mathbb{P}_X(x) \\ &= \mathbb{P}_X((-\infty, z - y]) = \mathbb{P}(X \leq z - y) = F_X(z - y).\end{aligned}$$

The result follows since $dF_Y(y)$ is just another notation for $d\mathbb{P}_Y(y)$. □

The next corollary allows us to calculate specific examples.

Corollary 189. Suppose X, Y are independent random variables. Suppose the c.d.f. of X has density f. Then $X + Y$ has density

$$h(x) = \int_{\mathbb{R}} f(x - y)\, dF_Y(y).$$

In addition, if the c.d.f. of Y has density g, then

$$h(x) = \int_{\mathbb{R}} f(x - y)g(y)\, dy.$$

Proof. First, note that

$$F_X(z-y) = \int_{-\infty}^{z-y} f(x)\,dx = \int_{-\infty}^{z} f(t-y)\,dt,$$

where we performed a change of variables $t := x + y$. Applying Theorem 188 we have

$$\begin{aligned}\mathbb{P}(X+Y \leq z) &= \int_{\mathbb{R}} F_X(z-y)\,dF_Y(y) = \int_{\mathbb{R}} \Big(\int_{-\infty}^{z} f(t-y)\,dt \Big)\,dF_Y(y) \\ &= \int_{-\infty}^{z} \Big(\int_{\mathbb{R}} f(t-y)\,dF_Y(y) \Big)\,dt = \int_{-\infty}^{z} h(t)\,dt,\end{aligned}$$

where we used Fubini's theorem since the function f is positive.

The second formula follows from (28). □

We finish this section with an important theorem that will be proved in Subsection 5.3.2.

Theorem 190 (Kolmogorov)**.** Let $\{F_n\}_{n=1}^{\infty}$ be a given sequence of distribution functions. Then there exists a sequence of independent random variables $\{X_n\}_{n=1}^{\infty}$ such that the cumulative density function of X_n is F_n.

4.4 Modes of convergence

Definition 191 (Types of convergence)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables.

(i) The sequence $\{X_n\}$ converges *almost surely* (a.s.) to X (denoted by $X_n \xrightarrow{a.s.} X$) if

$$\mathbb{P}\big(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\big) = 1.$$

(ii) The sequence $\{X_n\}$ converges in *probability* to X (denoted $X_n \xrightarrow{\mathbb{P}} X$) if for every $\epsilon > 0$

$$\mathbb{P}\big(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}\big) \to 0 \text{ as } n \to \infty.$$

(iii) The sequence $\{X_n\}$ converges in *L_p-mean* to X (denoted $X_n \xrightarrow{L_p} X$) if $\mathbb{E}|X_n|^p < \infty$ for all n and

$$\mathbb{E}|X_n - X|^p \to 0 \text{ as } n \to \infty,$$

where $p \in (0, \infty)$.

Exercise 192. Show that if $X_n \xrightarrow{L_p} X$, then $\mathbb{E}|X|^p < \infty$.

Exercise 193. Show that $\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$ and $\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$ are measurable sets.

Lemma 194. If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{\mathbb{P}} X$.

Proof. Fix an $\epsilon > 0$ and define the sets

$$A_n := \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$$

and let $A := \limsup A_n$. If $\omega \in A$ then ω is in infinitely many of the sets A_n meaning that the sequence $\{X_n(\omega)\}$ doesn't converge to $X(\omega)$. Since $\{X_n(\omega)\}$ doesn't converge to $X(\omega)$ on a set of measure 0, we conclude that $\mathbb{P}(A) = 0$. Applying Proposition 58 we get

$$0 \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \leq \limsup_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = \mathbb{P}(A) = 0.$$

This shows that $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$ which is what we needed. □

Example 195. This example shows that there is a sequence $\{X_n\}$ converging in probability to X but not converging to X a.s. That is, the converse of the last lemma is not true. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$:=the Borel sets on $[0,1]$ and $\mathbb{P}$:= the Lebesgue measure on $[0,1]$. Define the sequence of random variables

$$\begin{array}{llll} X_1 := \mathbf{1}_{[0,1/2]}, & X_2 := \mathbf{1}_{[1/2,1]}, & & \\ X_3 := \mathbf{1}_{[0,1/4]}, & X_4 := \mathbf{1}_{[1/4,1/2]}, & X_5 := \mathbf{1}_{[1/2,3/4]}, & X_6 := \mathbf{1}_{[3/4,1]} \\ X_7 := \mathbf{1}_{[0,1/8]}, \ldots & & & \end{array}$$

Then $X_n \xrightarrow{\mathbb{P}} 0$ since for every small $\epsilon > 0$ we get

$$\mathbb{P}(\{\omega \in \Omega : |X_n(\omega)| > \epsilon\}) = \mathbb{P}(\{\omega \in [0,1] : X_n(\omega) \neq 0\}) = \begin{cases} 1/2 & \text{if } n = 1,2 \\ 1/4 & \text{if } n = 3,4,5,6 \\ 1/8 & \text{if } n = 7, \ldots \\ \vdots & \end{cases}$$

But the sequence $\{X_n\}$ doesn't converge to 0 a.s.. In fact, $\lim\limits_{n\to\infty} X_n(\omega)$ doesn't exist for every $\omega \in [0,1]$. □

Lemma 196. For any $p \in (0,\infty)$, if $X_n \xrightarrow{L_p} X$ then $X_n \xrightarrow{\mathbb{P}} X$.

Proof. Fix an $\epsilon > 0$. By Chebyshev's inequality we get

$$\begin{aligned} \mathbb{P}\big(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}\big) &= \mathbb{P}\big(\{\omega \in \Omega : |X_n(\omega) - X(\omega)|^p > \epsilon^p\}\big) \\ &\leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p} \to 0 \text{ as } n \to \infty. \end{aligned}$$

□

Example 197. This example shows that there is a sequence $\{X_n\}$ converging in probability to X but not converging to X in L_p-mean. That is, the converse of the last lemma is not true. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$:=the Borel sets on $[0,1]$ and $\mathbb{P}$:= the Lebesgue measure on $[0,1]$. Define the sequence of random variables

$$X_1 := 2^{1/p} \cdot \mathbf{1}_{[0,1/2]}, \qquad X_2 := 2^{1/p} \cdot \mathbf{1}_{[1/2,1]},$$

$$
\begin{aligned}
&X_3 := 4^{1/p}\cdot \mathbf{1}_{[0,1/4]}, \qquad X_4 := 4^{1/p}\cdot \mathbf{1}_{[1/4,1/2]}, \qquad X_5 := 4^{1/p}\cdot \mathbf{1}_{[1/2,3/4]}, \qquad X_6 := 4^{1/p}\cdot \mathbf{1}_{[3/4,1]}\\
&X_7 := 8^{1/p}\cdot \mathbf{1}_{[0,1/8]}, \ldots
\end{aligned}
$$

Then $X_n \xrightarrow{\mathbb{P}} 0$ since for every small $\epsilon > 0$ we get

$$
\mathbb{P}(\{\omega \in \Omega : |X_n(\omega)| > \epsilon\}) = \mathbb{P}(\{\omega \in [0,1] : X_n(\omega) \neq 0\}) = \begin{cases} 1/2 & \text{if } n = 1,2\\ 1/4 & \text{if } n = 3,4,5,6\\ 1/8 & \text{if } n = 7,\ldots\\ \vdots \end{cases}
$$

But the sequence $\{X_n\}$ doesn't converge to 0 in L_p-mean since $\mathbb{E}|X_n|^p = 1$ for all n. □

Example 195 shows that the converse of Lemma 194 is not true, but the next result is a partial converse.

Lemma 198. If $X_n \xrightarrow{\mathbb{P}} X$ then there is a subsequence $\{X_{n_k}\}_{k=1}^{\infty}$ converging to X a.s.

Proof. Let $\{\epsilon_k\}$ be a sequence of positive numbers converging to 0. For each k there is an index n_k so that $\mathbb{P}\big(\{\omega \in \Omega : |X_{n_k}(\omega) - X(\omega)| > \epsilon_k\}\big) \leq 1/2^k$. (Note that n_k is possibly much bigger than k.) Moreover, we can choose those indexes so that $n_1 < n_2 < n_3 < \ldots$ Let $A_k := \{\omega \in \Omega : |X_{n_k}(\omega) - X(\omega)| > \epsilon_k\}$. Since

$$
\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty
$$

by the Borel-Cantelli lemma, we obtain $\mathbb{P}(\limsup A_k) = 0$, that is $\mathbb{P}(\liminf A_k^c) = 1$. This means that for almost every $\omega \in \Omega$, there are finitely many indexes k for which $|X_{n_k}(\omega) - X(\omega)| > \epsilon_k$ and for the rest of the indexes we have $|X_{n_k}(\omega) - X(\omega)| \leq \epsilon_k$. Since $\epsilon_k \to 0$ we get $X_{n_k}(\omega) \to X(\omega)$ as $k \to \infty$. That is exactly, $\{X_{n_k}\}$ converges to X a.s. as k approaches infinity. □

Finally, $X_n \xrightarrow{L_p} X$ doesn't imply that $X_n \xrightarrow{a.s.} X$ and vice versa $X_n \xrightarrow{a.s.} X$ doesn't imply that $X_n \xrightarrow{L_p} X$ as the next two examples show.

Example 199. This example shows a sequence $\{X_n\}$ such that $X_n \xrightarrow{a.s.} X$ but that doesn't converge to X in L_p-mean. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$:=the Borel sets on $[0,1]$ and $\mathbb{P}$:= the Lebesgue measure on $[0,1]$. Define the sequence of random variables

$$
\begin{aligned}
&X_1 := 2^{1/p}\cdot \mathbf{1}_{[0,1/2)}, \qquad && X_2 := 4^{1/p}\cdot \mathbf{1}_{[0,1/4)},\\
&X_3 := 8^{1/p}\cdot \mathbf{1}_{[0,1/8)}, && X_4 := 16^{1/p}\cdot \mathbf{1}_{[0,1/16)},\\
&\ldots
\end{aligned}
$$

Then $X_n \xrightarrow{a.s.} 0$ (so $X := 0$) but $\mathbb{E}|X_n|^p = 1$ for all $n = 1,2,\ldots$ and all $p \in (0,\infty)$, that is $\mathbb{E}|X_n - X|^p = \mathbb{E}|X_n|^p$ doesn't converge to 0 as n approaches infinity. □

Example 200. This example shows a sequence $\{X_n\}$ such that $X_n \xrightarrow{L_p} X$ but that doesn't converge to X almost surely. In fact, the sequence in Example 195 is just that (why?). □

In view of Example 195 and Lemma 198, the next result, which we state without a proof, appears almost shocking.

Theorem 201 (P. Lévy)**.** Suppose $X_1, X_2, \dots$ is a sequence of independent random variables and let $S_n := X_1 + \cdots + X_n$. If $S_n \xrightarrow{\mathbb{P}} S$, then $S_n \xrightarrow{a.s.} S$.

Proposition 202. Suppose that $X_n \xrightarrow{(*)} X$ and $Y_n \xrightarrow{(*)} Y$ where $(*)$ stand for either 'a.s.' or '$\mathbb{P}$', or 'L_p'. Then for any numbers $a, b \in \mathbb{R}$ we have $aX_n + bY_n \xrightarrow{(*)} aX + bY$.

Proof. (a.s.) Let $(*)$ stand for 'a.s.'. Then the fact that $X_n \xrightarrow{a.s.} X$ means that there is a set A with measure 1 such that $X_n(\omega) \to X(\omega)$ for all $\omega \in A$. Similarly the fact that $Y_n \xrightarrow{a.s.} Y$ means that there is a set B with measure 1 such that $Y_n(\omega) \to Y(\omega)$ for all $\omega \in B$. By the rules for manipulating sequences or numbers we have $aX_n(\omega) + bY_n(\omega) \to aX(\omega) + bY(\omega)$ for all $\omega \in A \cap B$. But $\mathbb{P}(A \cap B) = 1$ so we are done.

($\mathbb{P}$) Let $(*)$ stand for '$\mathbb{P}$'. Define the sets

$$\begin{aligned} C_n &:= \{\omega \in \Omega : |aX_n(\omega) + bY_n(\omega) - (aX(\omega) + bY(\omega))| > \epsilon\}, \\ A_n &:= \{\omega \in \Omega : |a||X_n(\omega) - X(\omega)| > \epsilon/2\}, \\ B_n &:= \{\omega \in \Omega : |b||Y_n(\omega) - Y(\omega)| > \epsilon/2\}. \end{aligned}$$

Since

$$|aX_n(\omega) + bY_n(\omega) - (aX(\omega) + bY(\omega))| \leq |a||X_n(\omega) - X(\omega)| + |b||Y_n(\omega) - Y(\omega)|,$$

we see that $C_n \subset A_n \cup B_n$ (if $\omega \notin A_n \cup B_n$ then $\omega \notin C_n$). Hence, $0 \leq \mathbb{P}(C_n) \leq \mathbb{P}(A_n) + \mathbb{P}(B_n) \to 0$ and we are done.

(L_p) Let $(*)$ stand for 'L_p'. Then by the Minkowski's inequality we have

$$\mathbb{E}|aX_n + bY_n - (aX + bY)|^p \leq \Big(|a|\big(\mathbb{E}|X_n - X|^p\big)^{\frac{1}{p}} + |b|\big(\mathbb{E}|Y_n - Y|^p\big)^{\frac{1}{p}}\Big)^p \to 0$$

which is what we needed to show. □

For the above types of convergence the random variables have to be defined on the same probability space. There is another type of convergence that does not require even that.

Definition 203 (Weak convergence)**.** A sequence of distribution functions $\{F_n\}$ *converges weakly* to a distribution F (denoted $F_n \xrightarrow{w} F$) if $\lim_{n\to\infty} F_n(x) = F(x)$ for every x that is a point of continuity of F.

Definition 204 (Convergence in distribution)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be probability spaces and let $X : \Omega \to \mathbb{R}$ and $X_n : \Omega_n \to \mathbb{R}$ be random variables with distribution functions F and F_n respectively $n = 1, 2, \dots$ Then the sequence $\{X_n\}$ *converges in distribution* to X (denoted $X_n \xrightarrow{d} X$) if $F_n \xrightarrow{w} F$.

Exercise 205. For any random variables X and Y, any $x \in \mathbb{R}$ and $\epsilon > 0$ we have

$$\begin{aligned} \{Y \leq x\} &\subseteq \{X \leq x + \epsilon\} \cup \{|Y - X| > \epsilon\}, \\ \{X \leq x - \epsilon\} &\subseteq \{Y \leq x\} \cup \{|Y - X| > \epsilon\}. \end{aligned}$$

Proposition 206. If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{d} X$.

Proof. Apply Exercise 205 with $Y := X_n$ and $X := X$. Taking probabilities from both sides of both inclusions gives

$$\begin{aligned} \mathbb{P}(\{X_n \leq x\}) &\leq \mathbb{P}(\{X \leq x + \epsilon\}) + \mathbb{P}(\{|X_n - X| > \epsilon\}), \text{ and} \\ \mathbb{P}(\{X \leq x - \epsilon\}) &\leq \mathbb{P}(\{X_n \leq x\}) + \mathbb{P}(\{|X_n - X| > \epsilon\}). \end{aligned}$$

Taking limsup from both sides of the first inequality and liminf from both sides of the second, and using the fact that $\lim_{n\to\infty} \mathbb{P}(\{|X_n - X| > \epsilon\}) = 0$, since $X_n \xrightarrow{\mathbb{P}} X$, results in

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) &\leq \mathbb{P}(\{X \leq x + \epsilon\}), \text{ and} \\ \mathbb{P}(\{X \leq x - \epsilon\}) &\leq \liminf_{n\to\infty} \mathbb{P}(\{X_n \leq x\}), \end{aligned}$$

or combining

$$\mathbb{P}(\{X \leq x - \epsilon\}) \leq \liminf_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) \leq \limsup_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) \leq \mathbb{P}(\{X \leq x + \epsilon\}).$$

This holds for every $\epsilon > 0$. Since the c.d.f. is right continuous, we have $\lim_{\epsilon\to 0^+} \mathbb{P}(\{X \leq x + \epsilon\}) = \mathbb{P}(\{X \leq x\})$. If x is a point where the c.d.f. of X is continuous then $\lim_{\epsilon\to 0^+} \mathbb{P}(\{X \leq x - \epsilon\}) = \mathbb{P}(\{X \leq x\})$. Thus, letting $\epsilon \to 0^+$ we obtain

$$\mathbb{P}(\{X \leq x\}) \leq \liminf_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) \leq \limsup_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) \leq \mathbb{P}(\{X \leq x\}),$$

showing that $\lim_{n\to\infty} \mathbb{P}(\{X_n \leq x\}) = \mathbb{P}(\{X \leq x\})$ for every x where the c.d.f. of X is continuous. □

Example 207. This example shows that $X_n \xrightarrow{d} X$ does not imply $X_n \xrightarrow{\mathbb{P}} X$. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$:=the Borel sets on $[0,1]$ and $\mathbb{P}$:= the Lebesgue measure on $[0,1]$. Define the sequence of random variables

$$\begin{aligned} X_1(\omega) &:= \mathbf{1}_{(0,1/2]}, \\ X_2(\omega) &:= \mathbf{1}_{(0,1/4]\cup(2/4,3/4]}, \\ &\vdots \\ X_n(\omega) &:= \mathbf{1}_{(0,1/2^n]\cup(2/2^n,3/2^n]\cup\cdots\cup((2^n-2)/2^n,(2^n-1)/2^n]}, \\ &\vdots \end{aligned}$$

Let $X := X_1$. If F is the c.d.f. of X and F_n is the c.d.f. of X_n then it is easy to see that

$$F(x) = F_n(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/2 & \text{if } 0 \le x < 1, \\ 1 & \text{if } 1 \le x, \end{cases}$$

so $F_n(x) \to F(x)$ for every x showing that $X_n \xrightarrow{d} X$.

On the other hand, for every $\epsilon \in (0,1)$ we have $\mathbb{P}(\{|X_n - X| > \epsilon\}) = 1/2$ for all $n = 1, 2, \ldots$ showing that $\{X_n\}$ does not converge in probability to X. □

The next exercise shows that if X is a constant random variable then the converse of Proposition 206 holds.

Exercise 208. If $X_n \xrightarrow{d} X$, where $X(\omega) = c$ for every $\omega \in \Omega$, then $X_n \xrightarrow{\mathbb{P}} X$.

Analogue of Proposition 202 for convergence in distribution doesn't hold. Instead we have the following results.

Exercise 209. Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, where $Y(\omega) = c$ for every $\omega \in \Omega$. Show that

(i) $X_n + Y_n \xrightarrow{d} X + Y$;

(ii) If $Z_n - X_n \xrightarrow{d} 0$ then $Z_n \xrightarrow{d} X$;

(iii) $X_n Y_n \xrightarrow{d} XY$. (To avoid complications, assume that $Y_n \ge 0$ and $c > 0$.)

Exercise 210. Let F be a distribution function. Recall that the quantile function is defined by $F^{-1}(\omega) := \sup\{x : F(x) < \omega\}$ for all $\omega \in (0,1)$. Show that

(i) F^{-1} is increasing, left-continuous function on $(0,1)$

(ii) If $x > F^{-1}(\omega)$, then $F(x) \ge \omega$

(iii) If F^{-1} is continuous at ω, then $x > F^{-1}(\omega)$ implies that $F(x) > \omega$.

(iv) If $F(x) > \omega$, then $x \ge F^{-1}(\omega)$

(v) If F is continuous at x, then $F(x) > \omega$ implies that $x > F^{-1}(\omega)$

(vi) If $x < F^{-1}(\omega)$, then $F(x) < \omega$

(vii) If $F(x) < \omega$, then $x < F^{-1}(\omega)$

(viii) $F^{-1} \circ F(x) \le x$ for all $x \in \mathbb{R}$, such that $0 < F(x) < 1$. (What is wrong when $F(x)$ is 0 or 1?)

(ix) $F \circ F^{-1}(\omega) \ge \omega$ for all $\omega \in (0,1)$

(x) $F \circ F^{-1} \circ F(x) = F(x)$ for all $x \in \mathbb{R}$, such that $0 < F(x) < 1$

(xi) $F^{-1} \circ F \circ F^{-1}(\omega) = F^{-1}(\omega)$ for all $\omega \in (0,1)$

(xii) If F has a jump at x, then F^{-1} is constant on $(F(x-), F(x)]$

(xiii) If F^{-1} has a jump at ω, then F is constant on $[F^{-1}(\omega), F^{-1}(\omega+))$

(xiv) The right limit of F^{-1} at ω is $F^{-1}(\omega+) = \inf\{t : F(t) > \omega\}$

(xv) $Y(\omega) := F^{-1}(\omega)$ is a random variable on $\big((0,1), \mathcal{B}(0,1)\big)$ with the Lebesgue measure. Show that $F_Y = F$.

Lemma 211. If F^{-1} is continuous at ω, then $x > F^{-1}(\omega)$ implies that $F(x) > \omega$.

Proof. Fix, $x > F^{-1}(\omega)$. Recall that F^{-1} is always left continuous. So, F^{-1} is continuous at ω if and only if it is right continuous. That is, for all small $\epsilon > 0$, $F^{-1}(\omega + \epsilon)$ is close to $F^{-1}(\omega)$. That is, for all small $\epsilon > 0$, we have

$$x > F^{-1}(\omega + \epsilon) = \sup\{y : F(y) < \omega + \epsilon\}$$

This means, that x is not in the set on the right-hand side (on which supremum is taken). Hence, $F(x) \geq \omega + \epsilon > \omega$. □

Exercise 212. Why is Lemma 211 not true, if F^{-1} is not continuous at ω?

Exercise 213. If F is continuous at x, then $F(x) > \omega$ implies that $x > F^{-1}(\omega)$. Why is this statement not true, if F is not continuous at x?

Proposition 214. If $F_n \xrightarrow{w} F$ then $\lim_{n\to\infty} F_n^{-1}(\omega) = F^{-1}(\omega)$ for almost all $\omega \in (0,1)$.

Proof. Let A be the set of all ω where F^{-1} is continuous. Since F^{-1} can have only countably many discontinuities and the Lebesgue measure of a countable subset of $(0,1)$ is zero, we get that the Lebesgue measure of A is 1. We show that for every $\omega \in A$ we have $F_n^{-1}(\omega) \to F^{-1}(\omega)$ as $n \to \infty$. So fix an $\omega \in A$.

Take an $x \in \mathbb{R}$, where F is continuous, such that $x < F^{-1}(\omega)$ and note that by Exercise 210 we have $F(x) < \omega$. It is given that $F_n(x) \to F(x)$, so for all n large enough we also have $F_n(x) < \omega$ implying that $x < F_n^{-1}(\omega)$. Hence

$$x \leq \liminf_{n\to\infty} F_n^{-1}(\omega).$$

Letting the point x approach $F^{-1}(\omega)$ from below (keeping the fact that x is a point of continuity for F) we get

$$F^{-1}(\omega) \leq \liminf_{n\to\infty} F_n^{-1}(\omega).$$

On the other hand, take a $x \in \mathbb{R}$ where F is continuous and such that $x > F^{-1}(\omega)$. Since F^{-1} is continuous at $\omega \in A$, by Lemma 211, we get $F(x) > \omega$. It is given that $F_n(x) \to F(x)$, so for all n large enough we also have $F_n(x) > \omega$ implying that $x \geq F_n^{-1}(\omega)$, by Exercise 210. Hence

$$x \geq \limsup_{n\to\infty} F_n^{-1}(\omega).$$

Finally, let the point x approach $F^{-1}(\omega)$ from above (keeping the fact that x is a point of continuity for F) to get

$$F^{-1}(\omega) \geq \limsup_{n\to\infty} F_n^{-1}(\omega).$$

Combining the two findings we get

$$F^{-1}(\omega) \geq \limsup_{n\to\infty} F_n^{-1}(\omega) \geq \liminf_{n\to\infty} F_n^{-1}(\omega) \geq F^{-1}(\omega).$$

We must have equality throughout showing that $\lim_{n\to\infty} F_n^{-1}(\omega) = F^{-1}(\omega)$. □

Theorem 215 (Convergence in distribution)**.** We have $X_n \xrightarrow{d} X$ if and only if $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for every bounded, continuous function $g : \mathbb{R} \to \mathbb{R}$.

Proof. Suppose $F_n(x) \xrightarrow{w} F(x)$ and let $Y := F^{-1}$ and $Y_n := F_n^{-1}$, $n = 1, 2, \ldots$ By Exercise 210, the random variable Y has c.d.f F and Y_n has c.d.f. F_n. By Proposition 214, we have $Y_n \xrightarrow{a.s.} Y$. Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded continuous function, then $g(Y_n) \xrightarrow{a.s.} g(Y)$, because g is continuous, and since g is bounded, we can apply Lebesgue's Dominated Convergence Theorem:

$$\mathbb{E}g(X_n) = \mathbb{E}g(Y_n) \to \mathbb{E}g(Y) = \mathbb{E}g(X).$$

To show the opposite direction, suppose that $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$ for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$. Fix an $x \in \mathbb{R}$ where F is continuous and an $\epsilon > 0$. Define the continuous function

$$g_{x,\epsilon}(y) = \begin{cases} 1 & \text{if } y \leq x, \\ 0 & \text{if } y \geq x + \epsilon, \\ \text{linear} & \text{if } x \leq y \leq x + \epsilon. \end{cases}$$

Note that $g_{x,\epsilon} \geq 0$ is bounded and continuous and in particular we have

$$g_{x-\epsilon,\epsilon}(y) = \begin{cases} 1 & \text{if } y \leq x - \epsilon, \\ 0 & \text{if } y \geq x, \\ \text{linear} & \text{if } x - \epsilon \leq y \leq x. \end{cases}$$

Note that for all $y \in \mathbb{R}$, we have

$$\mathbf{1}_{\{y: y \leq x-\epsilon\}}(y) \leq g_{x-\epsilon,\epsilon}(y) \leq \mathbf{1}_{\{y: y \leq x\}}(y) \leq g_{x,\epsilon}(y) \leq \mathbf{1}_{\{y: y \leq x+\epsilon\}}(y).$$

Replacing y in the first and last of these inequalities by $X(\omega)$ and replacing y in the second and third of these inequalities by $X_n(\omega)$, we get that

$$g_{x-\epsilon,\epsilon}(X_n(\omega)) \leq \mathbf{1}_{\{\omega\in\Omega: X_n(\omega) \leq x\}}(\omega) \leq g_{x,\epsilon}(X_n(\omega)),$$
$$g_{x,\epsilon}(X(\omega)) \leq \mathbf{1}_{\{\omega\in\Omega: X(\omega) \leq x+\epsilon\}}(\omega),$$
$$\mathbf{1}_{\{\omega\in\Omega: X(\omega) \leq x-\epsilon\}}(\omega) \leq g_{x-\epsilon,\epsilon}(X(\omega)),$$

hold for all $\omega \in \Omega$. Hence, taking expectations from all sides

$$\mathbb{E}\big(g_{x-\epsilon,\epsilon}(X_n)\big) \leq \mathbb{P}(X_n \leq x) \leq \mathbb{E}\big(g_{x,\epsilon}(X_n)\big),$$
$$\mathbb{E}\big(g_{x,\epsilon}(X)\big) \leq \mathbb{P}(X \leq x + \epsilon),$$
$$\mathbb{P}(X \leq x - \epsilon) \leq \mathbb{E}\big(g_{x-\epsilon,\epsilon}(X)\big).$$

So

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}(X_n\le x)\le\limsup_{n\to\infty}\mathbb{E}\big(g_{x,\epsilon}(X_n)\big)=\mathbb{E}\big(g_{x,\epsilon}(X)\big)\le\mathbb{P}(X\le x+\epsilon)\\ \liminf_{n\to\infty}\mathbb{P}(X_n\le x)\ge\liminf_{n\to\infty}\mathbb{E}\big(g_{x-\epsilon,\epsilon}(X_n)\big)=\mathbb{E}\big(g_{x-\epsilon,\epsilon}(X)\big)\ge\mathbb{P}(X\le x-\epsilon).\end{aligned}$$

Thus, for every $x\in\mathbb{R}$ and $\epsilon>0$ we obtained that

$$\mathbb{P}(X\le x-\epsilon)\le\liminf_{n\to\infty}\mathbb{P}(X_n\le x)\le\limsup_{n\to\infty}\mathbb{P}(X_n\le x)\le\mathbb{P}(X\le x+\epsilon).$$

In other words, if F is the c.d.f. of X and F_n is the c.d.f. of X_n, we have

$$F(x-\epsilon)\le\liminf_{n\to\infty}F_n(x)\le\limsup_{n\to\infty}F_n(x)\le F(x+\epsilon).$$

Since F is continuous at x, letting $\epsilon\to0$ we get

$$F(x)\le\liminf_{n\to\infty}F_n(x)\le\limsup_{n\to\infty}F_n(x)\le F(x).$$

This shows that $\lim_{n\to\infty}F_n(x)=F(x)$ at every x at which F is continuous. □

5 Limit theorems

5.1 Weak and strong laws of large numbers

For a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ we denote

$$S_n:=X_1+\cdots+X_n\quad n=1,2,\dots$$

Lemma 216. Suppose the random variables $\{X_n\}$ are independent and integrable, then

$$\operatorname{Var}S_n=\sum_{k=1}^{n}\operatorname{Var}X_k.$$

Proof. We calculate

$$\begin{aligned}\operatorname{Var}S_n=\mathbb{E}(S_n-\mathbb{E}S_n)^2&=\mathbb{E}\Big(\sum_{k=1}^{n}(X_k-\mathbb{E}X_k)\Big)^2\\&=\sum_{k=1}^{n}\mathbb{E}(X_k-\mathbb{E}X_k)^2+2\sum_{1\le j,k\le n}\mathbb{E}((X_j-\mathbb{E}X_j)(X_k-\mathbb{E}X_k)).\end{aligned}$$

But the fact that E_j and E_k are independent implies

$$\begin{aligned}\mathbb{E}((X_j-\mathbb{E}X_j)(X_k-\mathbb{E}X_k))&=\mathbb{E}(X_jX_k-X_j\mathbb{E}X_k-X_k\mathbb{E}X_j+\mathbb{E}X_j\mathbb{E}X_k)\\&=\mathbb{E}(X_jX_k)-\mathbb{E}X_j\mathbb{E}X_k-\mathbb{E}X_k\mathbb{E}X_j+\mathbb{E}X_j\mathbb{E}X_k=0,\end{aligned}$$

since $\mathbb{E}(X_jX_k)=\mathbb{E}X_j\mathbb{E}X_k$ □

The lemma, combined with the Chebyshev's inequality immediately gives the following bound.

Corollary 217. Suppose the random variables $\{X_n\}$ are independent and integrable, then

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq \epsilon) \leq \frac{1}{\epsilon^2}\sum_{k=1}^{n} \operatorname{Var} X_k. \tag{34}$$

Corollary 218 (Weak law of large numbers)**.** Suppose the random variables $\{X_n\}$ are independent and integrable.

$$\text{If } \lim_{n\to\infty} \sum_{k=1}^{n} \frac{\operatorname{Var} X_k}{n^2} = 0 \text{ then } \frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0.$$

Proof. Just apply inequality (34) with ϵ replaced by $n\epsilon$

$$\mathbb{P}\Big(\Big|\frac{S_n - \mathbb{E}S_n}{n}\Big| \geq \epsilon\Big) \leq \frac{1}{\epsilon^2}\sum_{k=1}^{n} \frac{\operatorname{Var} X_k}{n^2} \to 0$$

as n approaches infinity. Since the limit is zero for every $\epsilon > 0$ the result follows. □

The weak law of large numbers is called 'weak' because the convergence in the conclusion is in probability. The proof of the weak law relies crucially on inequality (34). In order to prove a 'strong' law of large numbers, one in which the convergence in the conclusion is almost sure, we need a stronger version of inequality (34).

Lemma 219 (Kolmogorov inequality)**.** Suppose the random variables $\{X_n\}$ are independent and integrable. Then for every $\epsilon > 0$ we have

$$\mathbb{P}(\max_{k\leq n} |S_k - \mathbb{E}S_k| \geq \epsilon) \leq \frac{1}{\epsilon^2}\sum_{k=1}^{n} \operatorname{Var} X_k. \tag{35}$$

Proof. Without loss of generality we can assume that $\mathbb{E}X_n = 0$, for all n.[3] In that case (35) becomes

$$\mathbb{P}(\max_{k\leq n} |S_k| \geq \epsilon) \leq \frac{1}{\epsilon^2}\sum_{k=1}^{n} \operatorname{Var} X_k. \tag{36}$$

Fix $\epsilon > 0$. Let $A_0 := \Omega$ and $A_n := \{\max_{k\leq n} |S_k| < \epsilon\}$, clearly we have $A_{n-1} \supseteq A_n$ for $n \geq 1$. Let $B_n := A_{n-1} \setminus A_n$ and note that the sets B_n are disjoint and $\cup_{k=1}^{n} B_k = A_n^c$ and

$$B_k = \{|S_1| < \epsilon, \ldots, |S_{k-1}| < \epsilon, |S_k| \geq \epsilon\}. \tag{37}$$

Equality (37) implies that $\epsilon^2 \mathbf{1}_{B_k} \leq (S_k \mathbf{1}_{B_k})^2$. Taking expectation of both sides gives

$$\epsilon^2 \mathbb{P}(B_k) = \mathbb{E}(\epsilon^2 \mathbf{1}_{B_k}) \leq \mathbb{E}(S_k \mathbf{1}_{B_k})^2 \leq \mathbb{E}(S_k \mathbf{1}_{B_k})^2 + \mathbb{E}((S_n - S_k)\mathbf{1}_{B_k})^2 = \mathbb{E}(S_n \mathbf{1}_{B_k})^2. \tag{38}$$

[3] Indeed, otherwise we let $X_n' := X_n - \mathbb{E}X_n$ then the random variables $X_1', X_2', \ldots$ are independent and integrable. Moreover $\operatorname{Var} X_n' = \operatorname{Var} X_n$ and $S_k' - \mathbb{E}S_k' = S_k - \mathbb{E}S_k$, where $S_k' = X_1' + \cdots + X_k'$. So proving (35) is the same as proving (35) with X_n replaced by X_n'.

The last equality is not quite obvious, but let us return to it at the end. Using it, we estimate

$$\epsilon^2 \mathbb{P}(A_n^c) = \epsilon^2 \sum_{k=1}^{n} \mathbb{P}(B_k) \leq \sum_{k=1}^{n} \mathbb{E}(S_n \mathbf{1}_{B_k})^2 \leq \mathbb{E}S_n^2 = \sum_{k=1}^{n} \operatorname{Var} X_k,$$

where in the last equality we used Lemma 216.

Let us justify now the last equality in (38). Square both sides of $S_k \mathbf{1}_{B_k} + (S_n - S_k)\mathbf{1}_{B_k} = S_n \mathbf{1}_{B_k}$, then take expectation throughout:

$$\mathbb{E}(S_k \mathbf{1}_{B_k})^2 + \mathbb{E}(2 S_k \mathbf{1}_{B_k} (S_n - S_k)\mathbf{1}_{B_k}) + \mathbb{E}((S_n - S_k)\mathbf{1}_{B_k})^2 = \mathbb{E}(S_n \mathbf{1}_{B_k})^2.$$

The middle term on the left-hand side is zero since

$$\mathbb{E}(2 S_k \mathbf{1}_{B_k} (S_n - S_k)\mathbf{1}_{B_k}) = 2\mathbb{E}(S_k \mathbf{1}_{B_k} (S_n - S_k)) = 2\mathbb{E}(S_k \mathbf{1}_{B_k})\mathbb{E}(S_n - S_k) = 0,$$

where we used that $S_k \mathbf{1}_{B_k}$ and $(S_n - S_k)$ are independent random variables. Indeed, Equality (37) shows that the random variable $S_k \mathbf{1}_{B_k}$ is a function of $X_1, \ldots, X_k$,[4] while the random variable $(S_n - S_k)$ is a function of $X_{k+1}, \ldots, X_n$. Since the random variables $\{X_n\}_{n=1}^{\infty}$ are independent, by Lemma 271 we see that $S_k \mathbf{1}_{B_k}$ and $(S_n - S_k)$ are independent. □

Corollary 220. Suppose the random variables $\{X_n\}$ are independent and integrable.

$$\text{If } \sum_{k=1}^{\infty} \operatorname{Var} X_k < \infty \text{ then } S_n - \mathbb{E}S_n \text{ converges a.s. to a random variable as } n \to \infty.$$

Proof. Without loss of generality assume, again, that $\mathbb{E}X_n = 0$ for all n. We need to show that S_n converges a.s. to a random variable as $n \to \infty$. Define the random variable

$$W_m := \sup_{k,p \geq 0} |S_{m+k} - S_{m+p}|.$$

Note that for every $\omega \in \Omega$ the sequence $\{W_m(\omega)\}_{m=1}^{\infty}$ is decreasing and bounded from below by 0. Hence, it has a limit. We are going to show that $W_m(\omega) \to 0$ for almost all $\omega \in \Omega$. By Theorem 14, this implies that $\{S_n(\omega)\}_{n=1}^{\infty}$ converges for almost all $\omega \in \Omega$.

The rest of the proof shows that $W_m \xrightarrow{a.s.} 0$. If we remove the first m random variables from the sequence $\{X_n\}$, inequality (36) still holds for the truncated sequence $X_{m+1}, X_{m+2}, \ldots$ But $X_{m+1} + X_{m+2} + \cdots + X_{m+k} = S_{m+k} - S_m$, so we get

$$\mathbb{P}(\max_{k=0,\ldots,n} |S_{m+k} - S_m| \geq \epsilon) \leq \frac{1}{\epsilon^2} \sum_{k=m+1}^{m+n} \operatorname{Var} X_k.$$

[4] Indeed, $S_k \mathbf{1}_{B_k} = S_k g(|S_1|, \ldots, |S_{k-1}|, |S_k|)$, where

$$g(x_1, \ldots, x_{k-1}, x_k) := \mathbf{1}_{(-\infty,\epsilon)\times\cdots\times(-\infty,\epsilon)\times[\epsilon,\infty)}(x_1, \ldots, x_{k-1}, x_k).$$

Note that $g : \mathbb{R}^k \to \mathbb{R}$ is measurable since the set $(-\infty, \epsilon) \times \cdots \times (-\infty, \epsilon) \times [\epsilon, \infty)$ is a measurable subset of $\mathbb{R}^k$.

Since n can be arbitrary large, letting $n \to \infty$, we obtain

$$\mathbb{P}(\sup_{k\geq 0} |S_{m+k} - S_m| \geq \epsilon) \leq \frac{1}{\epsilon^2} \sum_{k=m+1}^{\infty} \operatorname{Var} X_k.$$

Since $\sum_{k=m+1}^{\infty} \operatorname{Var} X_k \to 0$ as $m \to \infty$ we conclude that $\mathbb{P}(\sup\limits_{k\geq 0} |S_{m+k} - S_m| \geq \epsilon) \to 0$ as $m \to \infty$. Next, observe that

$$\begin{aligned} W_m &= \sup_{k,p\geq 0} |S_{m+k} - S_m + S_m - S_{m+p}| \leq \sup_{k,p\geq 0} \big(|S_{m+k} - S_m| + |S_m - S_{m+p}|\big) \\ &= \sup_{k\geq 0} |S_{m+k} - S_m| + \sup_{p\geq 0} |S_m - S_{m+p}|. \end{aligned}$$

Thus, $\{W_m > 2\epsilon\} \subseteq \{\sup\limits_{k\geq 0} |S_{m+k} - S_m| > \epsilon\} \cup \{\sup\limits_{p\geq 0} |S_{m+p} - S_m| > \epsilon\}$ and taking probability from both sides shows that

$$\mathbb{P}(W_m > 2\epsilon) \leq \mathbb{P}(\sup_{k\geq 0} |S_{m+k} - S_m| > \epsilon) + \mathbb{P}(\sup_{p\geq 0} |S_{m+p} - S_m| > \epsilon) \to 0 \text{ as } m \to \infty.$$

This shows that the sequence $\{W_m\}_{m=1}^{\infty}$ converges to 0 in probability. By Lemma 198 there is a subsequence $\{W_{m_k}\}_{k=1}^{\infty}$ converging to 0 a.s.. But since the whole sequence $\{W_m(\omega)\}_{m=1}^{\infty}$ is convergent and a subsequence converges to 0, the whole sequence must converge to zero. □

Corollary 221 (Strong law of large numbers)**.** Suppose the random variables $\{X_n\}$ are independent and integrable.

$$\text{If } \sum_{k=1}^{\infty} \frac{\operatorname{Var} X_k}{k^2} < \infty \text{ then } \frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{a.s.} 0.$$

Proof. Apply Corollary 220 to the random variables $X'_k := X_k/k$. Since $\sum\limits_{k=1}^{\infty} \operatorname{Var} X'_k = \sum\limits_{k=1}^{\infty} \dfrac{\operatorname{Var} X_k}{k^2} < \infty$, then $S'_n - \mathbb{E}S'_n$ converges a.s. to a random variable. That is for almost all $\omega \in \Omega$ we have $S'_n(\omega) - \mathbb{E}S'_n \to s_\omega$ as $n \to \infty$, where s_ω is a some real number. But

$$S'_n(\omega) - \mathbb{E}S'_n = \sum_{k=1}^{n} (X'_k(\omega) - \mathbb{E}X'_k) = \sum_{k=1}^{n} \frac{X_k(\omega) - \mathbb{E}X_k}{k}.$$

Applying Corollary 28 with $x_k := (X_k(\omega) - \mathbb{E}X_k)/k$ and $b_k := k$ we conclude that $(S_n(\omega) - \mathbb{E}S_n)/n \to 0$. Since this holds for almost all $\omega \in \Omega$, we are done. □

The strong law of large numbers given in Corollary 221 may not be the familiar one but it applies to any sequence $\{X_n\}$ of random variables having varied distributions. Note that if any of the random variables has an infinite variance then the result cannot be applied.

Corollary 222 (Strong law of large numbers, classical form, finite variance)**.** Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent, identically distributed random variables with $\mathbb{E}(X_i^2) < \infty$ and $\mathbb{E}X_i = \mu \in \mathbb{R}$. Then

$$\frac{S_n - n\mu}{n} \xrightarrow{a.s.} 0.$$

Proof. Since $\mathbb{E}(X_i^2) < \infty$, we have that $m := \operatorname{Var} X_i \leq \infty$. Then $\sum_{k=1}^{\infty} \frac{\operatorname{Var} X_k}{k^2} = m \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty$ and we conclude using Corollary 221, keeping in mind that $\mathbb{E}S_n/n = \mu$. □

Exercise 223. Show that

$$\text{if } \sum_{k=1}^{\infty} \frac{\operatorname{Var} X_k}{k^2} < \infty \text{ then } \lim_{n\to\infty} \sum_{k=1}^{n} \frac{\operatorname{Var} X_k}{n^2} = 0.$$

This means that the condition in Corollary 221 is stronger that the condition in Corollary 218. Thus, in Corollary 221 we require more but get a stronger conclusion in return too.

Exercise 224. For any random variable X we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) \leq \mathbb{E}|X| \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n).$$

Exercise 225. If the random variables $\{X_n\}_{n=1}^{\infty}$ are independent and $X_n \xrightarrow{a.s.} 0$, then $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq c) < \infty$ for any $c > 0$.

It turns out that if the random variables $\{X_n\}_{n=1}^{\infty}$ are independent and identically distributed, then we do not need to insist on a finite variance, in order to have the same conclusion.

Theorem 226 (Strong law of large numbers, classical form, any variance)**.** Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent, identically distributed random variables with $\mathbb{E}X_i = \mu \in \mathbb{R}$. Then

$$\frac{S_n - n\mu}{n} \xrightarrow{a.s.} 0.$$

Proof. We need to show that $S_n/n \xrightarrow{a.s.} \mu$. Let X be a random variable having the same distribution as X_i. Since $\mathbb{E}X_i = \mu \in \mathbb{R}$, then X_i is integrable and so X is, i.e. $\mathbb{E}|X| < \infty$. Observe by Exercise 224, that $\sum_{k=1}^{\infty} \mathbb{P}(|X_k| \geq k) = \sum_{k=1}^{\infty} \mathbb{P}(|X| \geq k) \leq E|X| < \infty$. Thus, by the Borel-Cantelli Lemma $\mathbb{P}(\text{limsup}\{|X_k| \geq k\}) = 0$. This means that for almost all $\omega \in \Omega$, ω is in finitely many of the sets $\{|X_k| \geq k\}$, $k = 1, 2, \ldots$ Then

$$\frac{S_n}{n} = \frac{1}{n}\Big(\sum_{k=1}^{n} X_k\Big) = \frac{1}{n}\sum_{k=1}^{n} X_k \mathbf{1}_{\{|X_k|<k\}} + \frac{1}{n}\sum_{k=1}^{n} X_k \mathbf{1}_{\{|X_k|\geq k\}} =: \frac{S_n'}{n} + \frac{S_n''}{n}.$$

By the above observations, for almost all $\omega \in \Omega$ only finitely many terms in the sum $S_n''(\omega)$ are non-zero, hence as n goes to infinity, $S_n''/n \xrightarrow{a.s.} 0$. Thus, in order to show that $S_n/n \xrightarrow{a.s.} \mathbb{E}X$ it is enough to show that $S_n'/n \xrightarrow{a.s.} \mathbb{E}X$.

Next, since $X\mathbf{1}_{\{|X|<k\}} \xrightarrow{a.s.} X$ as $k \to \infty$, $|X\mathbf{1}_{\{|X|<k\}}| \leq X$, and $\mathbb{E}|X| < \infty$, by the Dominated Convergence Theorem we get

$$\mathbb{E}\big(X_k \mathbf{1}_{\{|X_k|<k\}}\big) = \mathbb{E}\big(X\mathbf{1}_{\{|X|<k\}}\big) \to \mathbb{E}X \quad \text{as } k \to \infty.$$

Hence, by Corollary 27 we obtain $\lim\limits_{n\to\infty} \dfrac{\mathbb{E}S_n'}{n} = \mathbb{E}X$. Thus, it is sufficient to show that $(S_n' - \mathbb{E}S_n')/n \xrightarrow{a.s.} 0$. In order to do that, we use Corollary 221.

In the rest of the proof, we show that $\sum_{k=1}^{\infty}\frac{\operatorname{Var}X_k'}{k^2}<\infty$, where $X_k':=X_k\mathbf{1}_{\{|X_k|<k\}}$. Using Exercise 145 to exchange the summation with the expectation, we estimate

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{\operatorname{Var}X_k'}{k^2}&\le\sum_{k=1}^{\infty}\frac{\mathbb{E}\big((X_k')^2\big)}{k^2}=\sum_{k=1}^{\infty}\mathbb{E}\Big(\frac{(X_k')^2}{k^2}\Big)=\sum_{k=1}^{\infty}\mathbb{E}\Big(\frac{X_k^2}{k^2}\mathbf{1}_{\{|X_k|<k\}}\Big)\\&=\sum_{k=1}^{\infty}\mathbb{E}\Big(\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\Big)=\mathbb{E}\Big(\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\Big),\end{aligned}\tag{39}$$

where we used that X_k has the same distribution as X. In order to estimate the infinite sum, define the sets $B_m:=\{m-1\le|X|<m\}$ and note that

$$\{|X|<k\}\cap B_m=\begin{cases}\emptyset & \text{if } k<m,\\ B_m & \text{if } k\ge m.\end{cases}$$

Thus

$$\begin{aligned}\Big(\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\Big)\mathbf{1}_{B_m}&=\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\mathbf{1}_{B_m}=\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}\cap B_m}=\sum_{k=m}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{B_m}\\&=\Big(\sum_{k=m}^{\infty}\frac{1}{k^2}\Big)X^2\mathbf{1}_{B_m}\le\Big(\sum_{k=m}^{\infty}\frac{1}{k^2}\Big)m^2\mathbf{1}_{B_m}.\end{aligned}$$

We now bound the sum separately

$$\begin{aligned}\Big(\sum_{k=m}^{\infty}\frac{1}{k^2}\Big)m^2&=1+m^2\Big(\frac{1}{(m+1)^2}+\frac{1}{(m+2)^2}+\cdots\Big)\\&\le1+m^2\Big(\int_m^{m+1}\frac{1}{x^2}\,dx+\int_{m+1}^{m+2}\frac{1}{x^2}\,dx+\cdots\Big)\\&=1+m^2\Big(\int_m^{\infty}\frac{1}{x^2}\,dx\Big)=1+m^2\Big(\frac{1}{m}\Big)=1+m.\end{aligned}$$

Putting the last two estimates together we obtain

$$\Big(\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\Big)\mathbf{1}_{B_m}\le(1+m)\mathbf{1}_{B_m}\le(2+|X|)\mathbf{1}_{B_m}.$$

Notice that the sets B_m are disjoint and $\cup_{m=1}^{\infty}B_m=\Omega$, thus $\sum_{m=1}^{\infty}\mathbf{1}_{B_m}=1_{\Omega}\equiv1$. Summing the last displayed bound over m we get

$$\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\le2+|X|.$$

Substituting into (39) we finally obtain

$$\sum_{k=1}^{\infty}\frac{\operatorname{Var}X_k'}{k^2}\le\mathbb{E}\Big(\sum_{k=1}^{\infty}\frac{X^2}{k^2}\mathbf{1}_{\{|X|<k\}}\Big)\le\mathbb{E}(2+|X|)=2+\mathbb{E}|X|<\infty.$$

With this the proof of the strong law of large numbers is complete. □

Remark 227. Theorem 226 still holds (with a different proof) if we assume that the random variables $\{X_n\}_{n=1}^{\infty}$ are pairwise independent and identically distributed. As we will see, this is not true for the Central Limit Theorem. For the Central Limit Theorem, we have to require that all $\{X_n\}_{n=1}^{\infty}$ be independent. □

5.2 The central limit theorem

5.2.1 Characteristic functions

This section introduces the characteristic function of a random variable and its main properties with minimum emphasis on proofs. We begin by giving the necessary background on complex numbers.

A complex number $z \in \mathbb{C}$ has representation $z = a + ib$ where $a, b \in \mathbb{R}$ and i is the imaginary unit $i = \sqrt{-1}$. Clearly, we have $i^2 = -1, i^3 = -i, i^4 = +1, i^5 = i$, and so on. The complex numbers can be viewed as points in $\mathbb{R}^2$ to each $z = a + ib \in \mathbb{C}$ corresponds the point $(a, b) \in \mathbb{R}^2$.

The norm of $z = a + ib$ is $|z| := \sqrt{a^2 + b^2}$. Note that this is just the norm of vector $(a, b) \in \mathbb{R}^2$, or in other words, the distance of (a, b) to the origin of the coordinate system $(0, 0)$. Thus, the norm of a complex number z is the distance from z to $(0, 0)$. In particular, the complex numbers with norm 1 are the points on the circle with radius 1 and centered at $(0, 0)$ in $\mathbb{R}^2$. If we have two complex numbers $z = a + ib$ and $w = c + id$ then the difference is $z - w = (a - c) + i(b - d)$, so the norm of $z - w$ is $|z - w| = \sqrt{(a - c)^2 + (b - d)^2}$. This is the distance between the vectors (a, b) and (c, d), or in other words, the distance between z and w. It can be shown that for any complex numbers z, w one has $|z + w| \leq |z| + |w|$, $|zw| = |z||w|$, and $|z/w| = |z|/|w|$. Thus, for every $z \in \mathbb{C}$, the number $z/|z|$ has norm 1. Complex numbers cannot be compared, for $z, w \in \mathbb{C}$, it doesn't make sense to say $z \leq w$ or $w \leq z$. But their norms, being real numbers, can be compared.

The complex conjugate of $z = a + ib$, denoted $\bar{z}$, is $\bar{z} := a - ib$. Note that $\bar{z}$ is the reflexion of z across the x-axis in $\mathbb{R}^2$. The properties of conjugation are $z\bar{z} = |z|^2$; $\overline{z + w} = \bar{z} + \bar{w}$; $\overline{zw} = \bar{z}\bar{w}$; and $|\bar{z}| = |z|$.

Complex numbers of the form $z = \cos(t) + i \sin(t)$, for $t \in \mathbb{R}$, always have norm 1 since $\cos^2(t) + \sin^2(t) = 1$. In fact, every complex number with norm 1 can be represented in this way for a unique $t \in (-\pi, \pi]$.

The exponent of a complex number $z = a + ib$ is defined to be

$$\begin{aligned} e^z &:= 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + \cdots \\ &= 1 + \frac{a + ib}{1!} + \frac{(a + ib)^2}{2!} + \frac{(a + ib)^3}{3!} + \frac{(a + ib)^4}{4!} + \frac{(a + ib)^5}{5!} + \cdots \end{aligned}$$

(It can be shown that this power series converges for every $z \in \mathbb{C}$.) If $a = 0$, we obtain that for any $b \in \mathbb{R}$ we have

$$\begin{aligned} e^{ib} &= 1 + \frac{ib}{1!} + \frac{(ib)^2}{2!} + \frac{(ib)^3}{3!} + \frac{(ib)^4}{4!} + \frac{(ib)^5}{5!} + \cdots \\ &= 1 + i\frac{b}{1!} - \frac{b^2}{2!} - i\frac{b^3}{3!} + \frac{b^4}{4!} + i\frac{b^5}{5!} + \cdots \\ &= \Big(1 - \frac{b^2}{2!} + \frac{b^4}{4!} - \cdots\Big) + i\Big(\frac{b}{1!} - \frac{b^3}{3!} + \frac{b^5}{5!} - \cdots\Big) \end{aligned}$$

$$= \cos(b) + i\sin(b),$$

where we used that the power series expansions of $\cos(b)$ and $\sin(b)$, for any $b \in \mathbb{R}$, are

$$\cos(b) = 1 - \frac{b^2}{2!} + \frac{b^4}{4!} - \cdots \qquad \text{and} \qquad \sin(b) = \frac{b}{1!} - \frac{b^3}{3!} + \frac{b^5}{5!} - \cdots$$

Using the power series expansion of e^z it can be shown (but we will not) that for any $z, w \in \mathbb{C}$ we have

$$e^{z+w} = e^z e^w. \tag{40}$$

Therefore, if $z = a + ib$ we have

$$e^z = e^{a+ib} = e^a e^{ib} := e^a(\cos(b) + i\sin(b)).$$

That is, e^a is a real number and $e^{ib} = \cos(b) + i\sin(b)$ having norm $|e^{ib}| = 1$. That is, if $b \in \mathbb{R}$ then e^{ib} is a complex number on the circle with radius 1, centered at 0 (called, the *unit circle*). The number 1 is also on the unit circle. Thus, the distance between e^{ib} and 1 is

$$|e^{ib} - 1| \leq 2, \tag{41}$$

since the diameter of the unit circle is 2. In addition, we have

$$e^{\bar{z}} = e^{a-ib} = e^a e^{i(-b)} = e^a(\cos(-b) + i\sin(-b)) = e^a(\cos(b) - i\sin(b)) = \overline{e^z}.$$

The following fact will be useful, but we are not going to prove it. Compare(!) with Lemma 149.

Lemma 228. If $\lim_{n\to\infty} z_n = z \in \mathbb{C}$ then $\lim_{n\to\infty} \left(1 + \frac{z_n}{n}\right)^n = e^z$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{C}$ is measurable if it is measurable as a function from Ω into $\mathbb{R}^2$ (with the Borel σ-algebra). In that case we say that X is a complex-valued random variable. For every $\omega \in \Omega$, we have $X(\omega) = A(\omega) + iB(\omega)$ where $A : \Omega \to \mathbb{R}$ and $B : \Omega \to \mathbb{R}$ are measurable functions (why?). The expectation of X is defined to be

$$\mathbb{E}X := \mathbb{E}A + i\mathbb{E}B, \tag{42}$$

provided that the expectations $\mathbb{E}A$ and $\mathbb{E}B$ exist and are finite.

Now, let $X : \Omega \to \mathbb{R}$ be a (real-valued) random variable. The characteristic function (ch.f.) of X is defined to be

$$\phi_X(t) := \mathbb{E}\big(e^{itX}\big) \quad \text{for } t \in \mathbb{R}.$$

By definition, we have $\mathbb{E}\big(e^{itX}\big) = \mathbb{E}(\cos(tX) + i\sin(tX)) = \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX))$. Now, for every $\omega \in \Omega$, we have $|\cos(tX(\omega))| \leq 1$ and $|\sin(tX(\omega))| \leq 1$ so for every $t \in \mathbb{R}$ the random variables $\cos(tX)$ and $\sin(tX)$ are integrable. That is, the expectations $\mathbb{E}(\cos(tX))$ and $\mathbb{E}(\sin(tX))$ exist and are finite. Hence the characteristic function is well-defined for every $t \in \mathbb{R}$. Other notations for the characteristic function are

$$\phi_X(t) = \mathbb{E}\big(e^{itX}\big) = \int_\Omega e^{itX}\, d\mathbb{P} = \int_\mathbb{R} e^{itx}\, dF_X(x),$$

where F_X is the c.d.f. of X, see the first comment after Proposition 165.

The characteristic function is a useful tool for proving results about convergence in distribution. Since it is just a tool, we are not going to go into the deep details. We are going to show the elementary properties of the characteristic function and then just state the deeper results.

Proposition 229. The characteristic function has the following properties:

(i) $\phi_X(0) = 1$;

(ii) $\phi_X(-t) = \overline{\phi_X(t)}$;

(iii) $|\phi_X(t)| = \big|\mathbb{E}\big(e^{itX}\big)\big| \le \mathbb{E}\big|e^{itX}\big| = 1$;

(iv) $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$;

(v) If X and Y are independent with characteristic functions $\phi_X(t)$ and $\phi_Y(t)$, then $X+Y$ has characteristic function $\phi_X(t)\phi_Y(t)$.

Proof. (i) Easy. (ii) $\phi(-t) = \mathbb{E}(\cos(-tX) + i\sin(-tX)) = \mathbb{E}(\cos(tX) - i\sin(tX)) = \overline{\mathbb{E}(\cos(tX)) - i\mathbb{E}(\sin(tX))} = \overline{\phi(t)}$. (iii) Since $|a+ib| := \sqrt{a^2+b^2}$ we have

$$\begin{aligned}|\phi_X(t)| &= \big|\mathbb{E}\big(e^{itX}\big)\big| = \big|\mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX))\big| = \Big(\big(\mathbb{E}(\cos(tX))\big)^2 + \big(\mathbb{E}(\sin(tX))\big)^2\Big)^{1/2}\\ &= h(\mathbb{E}(\cos(tX)), \mathbb{E}(\sin(tX))),\end{aligned}$$

where $h(x,y) = (x^2+y^2)^{1/2}$. It is a fact that the function $h(x,y) : \mathbb{R}^2 \to \mathbb{R}$ is convex, hence by Exercise 157, for the random variables $\cos(tX)$ and $\sin(tX)$, we have

$$\begin{aligned}h(\mathbb{E}(\cos(tX)), \mathbb{E}(\sin(tX))) &\le \mathbb{E}h(\cos(tX), \sin(tX)) \le \mathbb{E}\big(\cos^2(tX) + \sin^2(tX)\big)^{1/2}\\ &= \mathbb{E}|\cos(tX) + i\sin(tX)| = \mathbb{E}\big|e^{itX}\big|.\end{aligned}$$

In addition, note that $\mathbb{E}\big(\cos^2(tX) + \sin^2(tX)\big)^{1/2} = \mathbb{E}(1) = 1$. (iv) Exercise. (v) $\mathbb{E}e^{it(X+Y)} = \mathbb{E}(e^{itX}e^{itY}) = \mathbb{E}e^{itX}\mathbb{E}e^{itY}$, where we used, (42), Corollary 184 with condition b). □

Theorem 230 (Continuity theorem)**.** Let $X, X_1, X_2, \ldots$ be a sequence of random variables with characteristic functions $\phi(t), \phi_1(t), \phi_2(t), \ldots$ respectively.
(a) If $X_n \xrightarrow{d} X$, then $\phi_n(t) \to \phi(t)$ for all t.
(b) If $\phi_n(t) \to \phi(t)$ for all t and $\phi(t)$ is continuous at $t = 0$, then $X_n \xrightarrow{d} X$.

Theorem 231 (Inversion theorem)**.** Let X and Y are two random variables with c.d.f.'s F_X and F_Y, and characteristic functions ϕ_X and ϕ_Y. Then $F_X = F_Y$ if and only if $\phi_X = \phi_Y$.

Theorem 232 (Moments and derivatives theorem)**.** If $\mathbb{E}|X|^n < \infty$ then the characteristic function $\phi_X(t)$ is n times continuously differentiable, with derivatives given by

$$\phi_X^{(k)}(t) = \int_{\mathbb{R}} (ix)^k e^{itx}\, dF_X(x), \quad k = 0, 1, \ldots, n. \tag{43}$$

In particular, we obtain that $\phi_X^{(k)}(0) = \int_{\mathbb{R}} (ix)^k\, dF_X(x) = \mathbb{E}(iX)^k$.

Theorem 232 says that if $\mathbb{E}|X|^n < \infty$ then the characteristic function $\phi(t)$ of X has n-th order Taylor expansion around x:

$$\phi(x+t) = \phi(x) + \frac{t}{1!}\phi^{(1)}(x) + \frac{t^2}{2!}\phi^{(2)}(x) + \cdots + \frac{t^n}{n!}\phi^{(n)}(x) + o(t^n),$$

where "error" term $o(t^n)$ is an (unknown) function with the property that $o(t^n)/t^n \to 0$ as $t \to 0$. In particular, the Taylor expansion at $x = 0$ is

$$\begin{aligned} \phi(t) &= \phi(0) + \frac{t}{1!}\phi^{(1)}(0) + \frac{t^2}{2!}\phi^{(2)}(0) + \cdots + \frac{t^n}{n!}\phi^{(n)}(0) + o(t^n) \\ &= 1 + \frac{t}{1!}\mathbb{E}(iX) + \frac{t^2}{2!}\mathbb{E}(iX)^2 + \cdots + \frac{t^n}{n!}\mathbb{E}(iX)^n + o(t^n). \end{aligned} \tag{44}$$

5.2.2 The central limit theorem

Let $N(0,1)$ denote a standard normal random random variable (that is, a normal random variable with mean 0 and standard deviation 1.) Refer to Example 123 for the c.d.f. of a normal random variable.

Theorem 233 (Central limit theorem)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with $\mathbb{E}X_i = \mu \in \mathbb{R}$ and $\operatorname{Var} X_i = \sigma^2 \in (0,\infty)$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Proof. Without loss of generality assume $\mathbb{E}X_i = 0$ or otherwise replace X_n by $X_n' := X_n - \mu$. Let X be a random variable with the same distribution as X_i. Since $\mathbb{E}X_i^2 = \operatorname{Var} X_i + (\mathbb{E}X_i)^2 = \sigma^2 + 0 < \infty$ then by (44) with $n = 2$ we get that the characteristic function of X satisfies

$$\phi_X(t) = \phi_X(0) + \frac{t}{1!}\phi_X^{(1)}(0) + \frac{t^2}{2!}\phi_X^{(2)}(0) + o(t^2) = 1 - \frac{t^2}{2}\sigma^2 + o(t^2).$$

Hence by Proposition 229, part (iv) we obtain

$$\phi_{X/\sigma\sqrt{n}}(t) = \phi_X(t/\sigma\sqrt{n}) = 1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{\sigma^2 n}\Big).$$

So by Proposition 229, part (v), using the fact that $X_1, \ldots, X_n$ are independent, we obtain

$$\phi_{S_n/\sigma\sqrt{n}}(t) = \phi_{X_1/\sigma\sqrt{n}+\cdots+X_n/\sigma\sqrt{n}}(t) = \phi_{X_1/\sigma\sqrt{n}}(t)\cdots\phi_{X_n/\sigma\sqrt{n}}(t) = \Big(1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{\sigma^2 n}\Big)\Big)^n = \Big(1 + \frac{c_n}{n}\Big)^n,$$

where we defined the complex number

$$c_n := n\Big(-\frac{t^2}{2n} + o\Big(\frac{t^2}{\sigma^2 n}\Big)\Big) = -\frac{t^2}{2} + \frac{t^2}{\sigma^2}\frac{o(\frac{t^2}{\sigma^2 n})}{\frac{t^2}{\sigma^2 n}}.$$

As n approaches infinity, $\frac{t^2}{\sigma^2 n}$ approaches 0 and hence $\frac{o(\frac{t^2}{\sigma^2 n})}{\frac{t^2}{\sigma^2 n}}$ approaches 0. Hence $c_n \to -t^2/2$ as $n \to \infty$. By Lemma 228 we see that for every fixed t we have

$$\phi_{S_n/\sigma\sqrt{n}}(t) \to e^{-t^2/2}.$$

Since $e^{-t^2/2}$ is the characteristic function of $N(0,1)$, and it is continuous at $t = 0$, by Theorem 230, part b), we conclude. □

Example 234. Let $\{X_n\}$ be independent and identically distributed random variables with $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^2 = \sigma^2 \in (0, \infty)$. We will show that

$$\frac{\sum_{k=1}^n X_k}{\left(\sum_{k=1}^n X_k^2\right)^{1/2}} \xrightarrow{d} N(0,1).$$

Indeed

$$\frac{\sum_{k=1}^n X_k}{\left(\sum_{k=1}^n X_k^2\right)^{1/2}} = \frac{\sum_{k=1}^n X_k}{\sigma\sqrt{n}} \frac{\sigma}{\left(\left(\sum_{k=1}^n X_k^2\right)/n\right)^{1/2}}.$$

By the central limit theorem

$$\frac{\sum_{k=1}^n X_k}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \tag{45}$$

By the strong law of large numbers, applied to the random variables $\{X_n^2\}$ we have $\left(\sum_{k=1}^n X_k^2\right)/n \xrightarrow{a.s.} \sigma^2$. Hence

$$\frac{\sigma}{\left(\left(\sum_{k=1}^n X_k^2\right)/n\right)^{1/2}} \xrightarrow{a.s.} 1.$$

But almost sure convergence implies convergence in probability, which implies convergence in distribution, hence

$$\frac{\sigma}{\left(\left(\sum_{k=1}^n X_k^2\right)/n\right)^{1/2}} \xrightarrow{d} 1. \tag{46}$$

Evoke Exercise 209, part (iii) to conclude that the product of the left-hand sides of (45) and (46) converges in distribution to $N(0,1)$. □

5.2.3 Other limit theorems (optional)

The requirement in the central limit theorem that the random variables have to be identically distributed can be removed with a small price to pay for that. The next theorem does that. Define

$$s_n := \sqrt{\operatorname{Var} S_n}$$

to be the standard deviation of S_n. If the random variables $\{X_n\}$ are independent, then by Lemma 216, we have $s_n = \sqrt{\operatorname{Var} X_1 + \operatorname{Var} X_2 + \cdots + \operatorname{Var} X_n}$. If the random variables $\{X_n\}$ are independent and identically distributed with variance σ^2, then $s_n = \sigma\sqrt{n}$.

The proof of the next theorem is given in Appendix B.

Theorem 235 (Central limit theorem, not identically distributed random variables)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and integrable random variables.

$$\text{If for some } \delta \in (0,1] \text{ we have } \lim_{n\to\infty} \sum_{k=1}^{n} \frac{\mathbb{E}|X_k - \mathbb{E}X_k|^{2+\delta}}{s_n^{2+\delta}} = 0 \text{ then } \frac{S_n - \mathbb{E}S_n}{s_n} \xrightarrow{d} N(0,1).$$

Remark 236. Note also that if the random variables are identically distributed with mean μ and variance σ^2, then $s_n = \sigma\sqrt{n}$ and the conclusion of the theorem is exactly the same as that of Theorem 233. The condition $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{\mathbb{E}|X_k - \mathbb{E}X_k|^{2+\delta}}{s_n^{2+\delta}} = 0$ is reminiscent of the condition in Corollary 218. (Recall that $\operatorname{Var} X_k = \mathbb{E}|X_k - \mathbb{E}X_k|^2$). □

Let us compare Theorem 226 and Theorem 233. If $\sigma \in (0,\infty)$ is any constant, then the strong law of large numbers can be rewritten in an equivalent form as follows.

Theorem 237 (Strong law of large numbers, classical form)**.** Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent, identically distributed random variables with $\mathbb{E}X_i = \mu \in \mathbb{R}$. Then

$$\frac{S_n - n\mu}{\sigma n} \xrightarrow{a.s.} 0.$$

In the strong law of large numbers the denominator n grows much faster (as n approaches infinity) that the denominator $\sigma\sqrt{n}$ in the central limit theorem. That faster growth "kills" all randomness in the random walk $S_n - n\mu$ and hence the "normalized" $((S_n - n\mu)/n)$ random walk almost always ends at 0. In the central limit theorem the slower growth of the denominator is not capable of annihilating the random walk and that is why the limiting distribution of the "normalized" $((S_n - n\mu)/\sigma\sqrt{n})$ is non-trivial (that is, it is not a constant). This raises the question, if there are other functions of n that we can put in the denominator and still have a meaningful result. One answer is given by the next deep theorem.

Theorem 238 (Law of iterated logarithm)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with $\mathbb{E}X_i = \mu \in \mathbb{R}$ and $\operatorname{Var} X_i = \sigma^2 \in (0,\infty)$. Then

$$\mathbb{P}\Big(\limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{n\log\log n}} = \sqrt{2}\Big) = \mathbb{P}\Big(\liminf_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{n\log\log n}} = -\sqrt{2}\Big) = 1.$$

Notice first that for every $n \geq 16$ the denominator in the law of iterated logarithm is between $\sigma\sqrt{n}$ and σn, that is

$$\sigma\sqrt{n} \leq \sigma\sqrt{n\log\log n} \leq \sigma n.$$

Hence the conclusion in Theorem 238 is in a sense "between" those of Theorems 226 and 233. It says that for almost all $\omega \in \Omega$, we have

$$\limsup_{n\to\infty} \frac{S_n(\omega) - n\mu}{\sigma\sqrt{n\log\log n}} = \sqrt{2} \quad \text{and} \quad \liminf_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{n\log\log n}} = -\sqrt{2}.$$

This means that for every $\epsilon > 0$ the sequence $\frac{S_n(\omega)-n\mu}{\sigma\sqrt{n\log\log n}}$ does not have a limit point bigger than $\sqrt{2}+\epsilon$ and does not have a limit point smaller than $-\sqrt{2}-\epsilon$. But that can happen if and only if the sequence $\frac{S_n(\omega)-n\mu}{\sigma\sqrt{n\log\log n}}$ has finitely many terms above $\sqrt{2}+\epsilon$ and below $-\sqrt{2}-\epsilon$. (If it has infinitely many terms above $\sqrt{2}+\epsilon$ then it must have a limit point bigger than $\sqrt{2}+\epsilon$ contradicting the fact that limsup is $\sqrt{2}$.) In other words, for almost all $\omega \in \Omega$ the random walk $S_n(\omega) - n\mu$ satisfies

$$(-\sqrt{2}-\epsilon)\sigma\sqrt{n\log\log n} \le S_n(\omega) - n\mu \le (\sqrt{2}+\epsilon)\sigma\sqrt{n\log\log n}$$

for all n large enough. For example, suppose $\sigma = 1$ and $\epsilon = 0.01$, then almost surely the random walk $S_n - n\mu$ is between the graphs of the functions $f(x) := (-\sqrt{2}-0.01)\sqrt{x\log\log x}$ and $g(x) := (\sqrt{2}+0.01)\sqrt{x\log\log x}$ for all n large enough.

It is important to know how good is the approximation in the central limit theorem. The next theorem does that and we will omit the proof. Recall that

$$\Phi(x) := \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy$$

is the cumulative distribution function of a standard normal random variable.

Theorem 239 (Berry-Essen)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with $\mathbb{E}X_i = \mu \in \mathbb{R}$ and $\operatorname{Var} X_i = \sigma^2 \in (0,\infty)$. Suppose also that $\mathbb{E}|X_i - \mu|^3 = \rho \in (0,\infty)$. Then

$$\Big|\mathbb{P}\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\Big) - \Phi(x)\Big| \le \frac{3\rho}{\sigma^3\sqrt{n}}.$$

5.3 Applications

5.3.1 Glivenko-Cantelli theorem

Let $\{X_n\}$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. The function

$$F_n(x,\omega) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X_k(\omega),\infty)}(x). \tag{47}$$

is called *empirical distribution function*, it is a function of both ω and x. For every x and ω, the empirical distribution function counts how many of the numbers $X_1(\omega),\dots,X_n(\omega)$ are less than or equal to x and divides by n.

We give some motivation for considering the function $F_n(x,\omega)$. Suppose we do not know the real distribution $F(x)$ of the random variables $\{X_n\}$ and suppose that we can *sample* as many as we like of the variables $\{X_n\}$ at one and the same ω. That is, suppose for any ω we can measure the numbers $X_1(\omega), X_2(\omega),\dots,X_n(\omega)$. How can we estimate $F(x)$ using the sample? The goal of this section is to prove the Glivenko-Cantelli theorem, which says that for every x the empirical distribution function converges to $F(x)$ for almost all $\omega \in \Omega$. In fact, Glivenko-Cantelli theorem proves more than that as we will see.

Since $\mathbf{1}_{[X_k(\omega),\infty)}(x) = \mathbf{1}_{\{X_k \le x\}}(\omega)$ we also have the representation

$$F_n(x,\omega) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k \le x\}}(\omega).$$

Exercise 240. Let $\{X_n\}_{n=1}^{\infty}$ be random variables. Fix $n \in \{1, 2, \ldots\}$ and $\omega \in \Omega$. Show that

(i) As a function of x, $F_n(x, \omega)$ is a distribution function;

(ii) For every x, $F_n(x-, \omega) := \lim_{y \uparrow x} F_n(y, \omega) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} \mathbf{1}_{\{X_k < x\}}(\omega)$.

To simplify the notation we are going to stop writing the ω but keep in mind that it is there. That is, we denote

$$F_n(x) := F_n(x, \omega) \quad \text{and} \quad F_n(x-) := F_n(x-, \omega).$$

Lemma 241. Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. Then for every $x \in \mathbb{R}$, $F_n(x) \xrightarrow{a.s.} F(x)$ as $n \to \infty$.

Proof. Let X be a random variable having distribution $F(x)$ as well. Fix $x \in \mathbb{R}$ and define the random variables $Y_n := \mathbf{1}_{\{X_n \le x\}}$, $n = 1, 2, \ldots$ The random variables $\{Y_n\}_{n=1}^{\infty}$ are independent (why?) and identically distributed. Indeed,

$$\begin{aligned}
\mathbb{P}(Y_n \le y) &= \mathbb{P}(\mathbf{1}_{\{X_n \le x\}} \le y) \\
&= \mathbb{P}(\mathbf{1}_{\{X \le x\}} \le y) = \begin{cases} \mathbb{P}(\Omega) & \text{if } y \ge 1, \\ \mathbb{P}(X > x) & \text{if } 0 \le y < 1, \\ \mathbb{P}(\emptyset) & \text{if } y < 0. \end{cases} = \begin{cases} 1 & \text{if } y \ge 1, \\ 1 - F(x) & \text{if } 0 \le y < 1, \\ 0 & \text{if } y < 0. \end{cases}
\end{aligned}$$

Next, $\mathbb{E}Y_n = \mathbb{P}(X_n \le x) = F(x) < \infty$. So by the strong law of large numbers, Theorem 226, we have $\frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{a.s.} F(x)$. □

Remark 242. The lemma says that for every $x \in \mathbb{R}$ there is a measurable subset $A_x \subseteq \Omega$ with $\mathbb{P}(A) = 1$ such that $\lim_{n\to\infty} F_n(x, \omega) = F(x)$ for all $\omega \in A_x$. Note that this set A_x depends on x. What do we get if we intersect all these A_x? That is, define $A := \cap_{x \in \mathbb{R}} A_x$. If $\omega \in A$, then $\lim_{n\to\infty} F_n(x, \omega) = F(x)$ for all $x \in \mathbb{R}$. This is almost what we want to show, but the problem is that, being an intersection of more than countably many sets, A may not be measurable and may not have probability 1. □

Lemma 243. Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. Then for every $x \in \mathbb{R}$, $F_n(x-) \xrightarrow{a.s.} F(x-)$ as $n \to \infty$.

Proof. Let X be a random variable having distribution $F(x)$ as well. Fix $x \in \mathbb{R}$ and define the random variables $Z_n := \mathbf{1}_{\{X_n < x\}}$, $n = 1, 2, \ldots$. The random variables $\{Z_n\}_{n=1}^{\infty}$ are independent (why?) and identically distributed (why?). By Exercise 240, we have $F_n(x-, \omega) = \frac{1}{n}\sum_{k=1}^{n} Z_k$. Next, $\mathbb{E}Z_n = \mathbb{P}(X_n < x) = F(x-) < \infty$. So by the strong law of large numbers, Theorem 226, we have $\frac{1}{n}\sum_{k=1}^{n} Z_k \xrightarrow{a.s.} F(x-)$. □

Combining Lemmas 241 and 243 in a clever way, gives the result that we wanted.

Theorem 244 (Glivenko-Cantelli)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. Then

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

Proof. Fix a positive integer $k \geq 1$. For every integer $j = 1, \ldots, k-1$, define the point

$$x_{j,k} := F^{-1}(j/k).$$

Let $A_{j,k} \subseteq \Omega$ be the set of measure 1 such that $F_n(x_{j,k}) \to F(x_{j,k})$ for all $\omega \in A_{j,k}$ as per Lemma 241. Let $B_{j,k} \subseteq \Omega$ be the set of measure 1 such that $F_n(x_{j,k}-) \to F(x_{j,k}-)$ for all $\omega \in B_{j,k}$ as per Lemma 243. Let

$$C_k := \Big(\bigcap_{1 \leq j \leq k-1} A_{j,k} \Big) \bigcap \Big(\bigcap_{1 \leq j \leq k-1} B_{j,k} \Big),$$

and note that $\mathbb{P}(C_k) = 1$ (the complement of C_k is a finite union of sets of measure 0.) The point of defining C_k is that for any $\omega \in C_k$ we have

$$F_n(x_{j,k}) \to F(x_{j,k}) \text{ and } F_n(x_{j,k}-) \to F(x_{j,k}-)$$

for all $x_{j,k}$ with $j = 1, \ldots, k-1$..

Fix an integer $k \geq 1$ and $\omega \in C_k$. Since there are finitely many $x_{j,k}$ for that fixed k (remember that $j = 1, 2, \ldots, k-1$), there is an integer $N = N(\omega, k)$ such that for all $n \geq N$ we have

$$|F_n(x_{j,k}) - F(x_{j,k})| < 1/k \text{ and } |F_n(x_{j,k}-) - F(x_{j,k}-)| < 1/k \tag{48}$$

for all $j = 1, 2, \ldots, k-1$. If we let $x_{0,k} := -\infty$ and $x_{k,k} := \infty$ then the first inequality in (48) still holds for $j = 0$ and $j = k$ (indeed $F_n(-\infty) - F(-\infty) = 0 - 0 = 0$ and $F_n(\infty) - F(\infty) = 1 - 1 = 0$), while the second holds for $j = k$. The numbers $x_{0,k}, x_{1,k}, \ldots, x_{k-1,k}, x_{k,k}$ partition the real line $\mathbb{R}$ into the intervals $[x_{0,k}, x_{1,k})$, $[x_{1,k}, x_{2,k})$,…,$[x_{k-1,k}, x_{k,k})$. We now show a bound on the growth of the function F on each of these intervals.

Claim: For any $j = 1, \ldots, k$ we have $F(x_{j,k}-) - F(x_{j-1,k}) \leq 1/k$.

Proof: First, if $j = 1, 2, .., k-1$ then $F(x_{j,k}-) = \lim_{x \uparrow x_{j,k}} F(x)$, and since for $x < x_{j,k} = F^{-1}(j/k)$, by Lemma 210, part (vi) we have $F(x) < j/k$, hence $F(x_{j,k}-) \leq j/k$.

Second, if $j = 2, \ldots, k$, then $F(x_{j-1,k}) = F(F^{-1}((j-1)/k)) \geq (j-1)/k$, where we used Lemma 210, part (ix). Putting these observations together we get:

If $j = 1$ we have $F(x_{j,k}-) - F(x_{j-1,k}) = F(x_{1,k}-) - 0 \leq 1/k$;

If $j = 2, 3, \ldots, k-1$ we have $F(x_{j,k}-) - F(x_{j-1,k}) \leq j/k - (j-1)/k = 1/k$;

If $j = k$ we have $F(x_{j,k}-) - F(x_{j-1,k}) = 1 - F(x_{k-1,k}) \leq 1 - (k-1)/k = 1/k$. □

Every $x \in \mathbb{R}$ must belong to one of the intervals $[x_{j-1,k}, x_{j,k})$ for some $j \in \{1, \ldots, k\}$. Using inequalities (48), the Claim, and the monotonicity of $F(x)$, we estimate

$$F_n(x) \leq F_n(x_{j,k}-) \leq F(x_{j,k}-) + 1/k \leq F(x_{j-1,k}) + 2/k \leq F(x) + 2/k,$$
$$F_n(x) \geq F_n(x_{j-1,k}) \geq F(x_{j-1,k}) - 1/k \geq F(x_{j,k}-) - 2/k \geq F(x) - 2/k.$$

This shows that $|F_n(x) - F(x)| \leq 2/k$. Since this is true for every $x \in \mathbb{R}$ we see that

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq 2/k. \tag{49}$$

We showed that for any $k \geq 1$ and any $\omega \in C_k$, there is an $N = N(k, \omega)$ such that for all $n \geq N$ inequality (49) holds. Finally, let

$$C := \bigcap_{k=1}^{\infty} C_k.$$

Since C is a countable intersection of sets of measure 1, we have $\mathbb{P}(C) = 1$.

Fix any $\omega \in C$ and any $\epsilon > 0$. Since $\omega \in C_k$ for all $k \geq 1$, we can choose k such that $2/k < \epsilon$. By the above, there is an $N = N(k, \omega)$ such that for all $n \geq N$,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq 2/k < \epsilon.$$

This means that for all $\omega \in C$, we have $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0$ as $n \to \infty$. We are done. □

Thus, the Glivenko-Cantelli theorem is a significant strengthening of Lemma 241. We conclude with a statement of a strengthening of Glivenko-Cantelli theorem, quantifying the rate of convergence as n tends to infinity.

Theorem 245 (Dvoretzky-Kiefer-Wolfowitz inequality)**.** Let $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. Then

$$\mathbb{P}\big(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \epsilon\big) \leq 2e^{-2n\epsilon^2},$$

for any $\epsilon > 0$.

The next exercise shows that the Dvoretzky-Kiefer-Wolfowitz' inequality indeed strengthens the Glivenko-Cantelli theorem.

Exercise 246. Show that Theorem 244 follows from the Dvoretzky-Kiefer-Wolfowitz' inequality.

5.3.2 Kolmogorov's extension theorem

Suppose you want to find out what is the distribution of an unknown random variable X. Suppose you can *sample* this random variable as much as you like. Say you sampled it n times and obtained the values x_1, x_2,…,x_n. How would you estimate the distribution function from the sample you have got? From undergraduate courses, one "knows" that the probability $\mathbb{P}(X \leq x)$ is *approximately equal* to the observed frequency of the sampled values that are less than or equal to x:

$$\frac{\text{The number of sample values that are less than or equal to } x}{n}. \tag{50}$$

So we should have that

$$F(x) := \mathbb{P}(X \leq x) \tag{51}$$

is *approximately equal* to (50). The good news is that we can write (50) as a function of x:

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{[x_k, \infty)}(x). \tag{52}$$

Indeed, $\mathbf{1}_{[x_k,\infty)}(x)$ is equal to 1 if $x_n \le x$ and is equal to 0 otherwise. So the sum $\sum_{k=1}^n \mathbf{1}_{[x_k,\infty)}(x)$ is equal to the number of sample values that are less than or equal to x. The bad news is that we do not know what it means to sample a random variable and we do not know if $\mathbb{P}(X \le x)$ is approximately equal to (50) and in what sense.

To sample a random variable means to perform an experiment with an unpredictable outcome and then measure that outcome to get x_1. (If you throw a die, usually the outcome of the experiment is the side of the die that points upwards and the measurement is the count of the dots on that side. If you are interested in the lifetime of a light bulb, the experiment is to pick a random light bulb from an inventory of identical light bulbs and turn it on until it expires. The measurement is the estimation of the (approximate) time when it burnt out.) Denote the set of all possible outcomes of the experiment by Ω. The outcome of the experiment is $\omega_1 \in \Omega$ and the result of the measurement is $X(\omega_1)$.

To sample again, perform the experiment again, independently of what happened before, measure the outcome to get the number x_2. But the outcome of the experiment, when performed for the second time is $\omega_2 \in \Omega$ and the result of the measurement is $x_2 := X(\omega_2)$. And so on, after performing the experiment n times we collect the sample

$$x_1 = X(\omega_1), x_2 = X(\omega_2), \dots, x_n = X(\omega_n).$$

So, how can we estimate the distribution function $F(x)$ using the sample? We will answer this question if we show that (52), or equivalently (50), converges to $F(x)$ as n approaches infinity. Formula (52) can be written as

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[x_k,\infty)}(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(\omega_k),\infty)}(x)$$

Thus, we want to show that for every $x \in \mathbb{R}$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(\omega_k),\infty)}(x) \xrightarrow{a.s.} F(x) \text{ as } n \to \infty \tag{53}$$

The function on the left-hand side of (53) is quite different from the function defined by (47). The function in (53) is a function of x and $\omega_1, \omega_2, \dots, \omega_n$, and as $n \to \infty$ the number of arguments of that function increases to infinity as well. Another difficulty that we have to clarify is what we mean by almost sure convergence in (53). In the Glivenko-Cantelly theorem, the almost sure convergence refers to the fact that for almost all $\omega \in \Omega$ we have $F_n(x,\omega) \to F(x)$ as $n \to \infty$. But now, the arguments in

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(\omega_k),\infty)}(x)$$

are x and in fact a whole sequence of omegas $\omega_1, \omega_2, \dots, \omega_n, \dots$. Thus we need to define a measure on the set of all sequences of omegas.

Suppose a sequence of probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ for $i = 1, 2\dots$ is given. We will explain what it means to form their product (a product of countably many probability spaces). Define the set

$$\Omega^{\mathbb{N}} := \{(\omega_1, \omega_2, \dots) : \omega_i \in \Omega_i \text{ for all } i = 1, 2, \dots\}.$$

The set $\Omega^{\mathbb{N}}$ is the collection of all sequences, in which the i-th element is from Ω_i. To define a σ-algebra on $\Omega^{\mathbb{N}}$, first consider the subsets of $\Omega^{\mathbb{N}}$ of the form

$$\Pi_{A_1,\ldots,A_n} := \{(\omega_1, \omega_2, \ldots) \in \Omega^{\mathbb{N}} : \omega_1 \in A_1, \ldots, \omega_n \in A_n\}$$

for some n and some $A_i \in \mathcal{F}_i$ for all $i = 1, \ldots, n$. That is, $\Pi_{A_1,\ldots,A_n}$ is the set of all sequences such that their first term is in A_1, second term is in A_2,..., and n-th term is in A_n. The subsequent terms of the sequences have no restrictions. Let

$$\mathcal{F}^{\mathbb{N}} := \sigma(\{\Pi_{A_1,\ldots,A_n} : A_i \in \mathcal{F}_i \text{ for all } i = 1, \ldots, n,\ n = 1, 2, \ldots\}).$$

Theorem 247 (Kolmogorov extension theorem)**.** There is a unique probability measure $\mathbb{P}^{\mathbb{N}}$ on $(\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}})$, such that

$$\mathbb{P}^{\mathbb{N}}(\Pi_{A_1,\ldots,A_n}) = \mathbb{P}_1(A_1) \cdots \mathbb{P}_n(A_n),$$

for any n and any $A_i \in \mathcal{F}_i$, where $i = 1, \ldots, n$.

You may not realize that you have been exposed to these ideas before. Here is one example from undergraduate probability classes.

Example 248. An experiment has outcomes $\Omega = \{\omega_1, \omega_2, \ldots\}$ occurring with probabilities $P(\omega_k) = p_k$. Perform the experiment repeatedly. What is the probability that ω_i occurs before ω_j, where $i \neq j$.

Solution. We have a 'little' experiment with outcomes in $\Omega = \{\omega_1, \omega_2, \ldots\}$ occurring with probabilities $P(\omega_k) = p_k$, but we repeat it endlessly, one little experiment after another. So, we end up with a *big* experiment having an outcome that is a sequence of ω's. The sample space of this big experiment is the set of all possible sequences of ω's.

We are interested in the event, call it E, of all sequences in which ω_i occurs before ω_j. But the first time ω_i occurs in a sequence could be on position 1, or 2, or any. Let E_n be the event consisting of all sequences in which ω_i occurs *for the first time* on the n-th position and ω_j does not occur on any of the previous $(n-1)$ positions. Note that the events $E_1, E_2, \ldots$ are disjoint (that is there is no sequence of ω's that is in say E_1 and E_2) and

$$E = \bigcup_{i=1}^{\infty} E_i.$$

The probability that neither ω_i nor ω_j occurs in one little experiment is $1 - p_i - p_j$. Thus, E_n is the event consisting of all sequences in which neither ω_i nor ω_j occur on positions $1, 2, \ldots, n-1$, and ω_i occurs on position n. Thus,[5]

$$P(E_n) = (1 - p_i - p_j)^{n-1} p_i.$$

[5] Let us carefully match this example with the developments before it. We have $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i) = (\Omega, 2^{\Omega}, P)$ for all $i = 1, 2, \ldots$ The sample space $\Omega^{\mathbb{N}}$ is the set of all sequences with elements from Ω. The event E_n is in the σ-algebra $\mathcal{F}^{\mathbb{N}}$. More precisely, let $A_k := \Omega \setminus \{\omega_i, \omega_j\}$ for all $k = 1, \ldots, n-1$ and let $A_n := \{\omega_i\}$. Then $E_n = \Pi_{A_1,\ldots,A_n}$. Hence, $\mathbb{P}^{\mathbb{N}}(E_n) = P(A_1) \cdots P(A_{n-1}) P(A_n) = (1 - p_i - p_j)^{n-1} p_i$. So, the measure, P, that is used is in fact $\mathbb{P}^{\mathbb{N}}$ but nobody told you that.

Thus, we finally have

$$\begin{aligned}P(E) &= P\Big(\bigcup_{n=1}^{\infty} E_n\Big) = \sum_{n=1}^{\infty} P(E_n) = \sum_{n=1}^{\infty}(1-p_i-p_j)^{n-1}p_i \\ &= p_i\sum_{n=1}^{\infty}(1-p_i-p_j)^{n-1} = p_i\frac{1}{1-(1-p_i-p_j)} = \frac{p_i}{p_i+p_j}.\end{aligned}$$

Denote by $\mathbf{w}$ a generic element of $\Omega^{\mathbb{N}}$, that is, $\mathbf{w} = (\omega_1, \omega_2, \ldots)$. Consider the function

$$P_n : \Omega^{\mathbb{N}} \to \Omega_n, \text{ defined by } P_n(\mathbf{w}) := \omega_n.$$

Lemma 249. The functions $\{P_n\}$ have the following properties.

(i) Each P_n is measurable;

(ii) The law of P_n is the measure $\mathbb{P}_n$;

(iii) The measurable functions $\{P_n\}$ are independent.

Proof. (i) Let $A_n \in \mathcal{F}_n$ then $P_n^{-1}(A_n) = \{\mathbf{w} \in \Omega^{\mathbb{N}} : \omega_n \in A_n\} = \Pi_{A_1,\ldots,A_n} \in \mathcal{F}^{\mathbb{N}}$ with $A_1 := \Omega_1, \ldots, A_{n-1} := \Omega_{n-1}$.

(ii) Let $A_n \in \mathcal{F}_n$, then $\mathbb{P}^{\mathbb{N}}(P_n^{-1}(A_n)) = \mathbb{P}^{\mathbb{N}}(\Pi_{A_1,\ldots,A_n}) = \mathbb{P}_1(A_1)\cdots\mathbb{P}_n(A_n) = \mathbb{P}_n(A_n)$ since $A_1 := \Omega_1, \ldots, A_{n-1} := \Omega_{n-1}$. So the measurable function P_n is distributed with law $\mathbb{P}_n$.

(iii) For any fixed n, let $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$. Then,

$$\begin{aligned}\mathbb{P}^{\mathbb{N}}(P_1 \in A_1, \ldots, P_n \in A_n) &= \mathbb{P}^{\mathbb{N}}(\{\mathbf{w} \in \Omega^{\mathbb{N}} : P_1(\mathbf{w}) \in A_1, \ldots, P_n(\mathbf{w}) \in A_n\}) = \mathbb{P}^{\mathbb{N}}(\Pi_{A_1,\ldots,A_n}) \\ &= \mathbb{P}_1(A_1)\cdots\mathbb{P}_n(A_n) = \mathbb{P}^{\mathbb{N}}(P_1 \in A_1)\cdots\mathbb{P}^{\mathbb{N}}(P_n \in A_n),\end{aligned}$$

where in the last equality we used part (ii) of the proof. □

To summarize: $(\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathbb{P}^{\mathbb{N}})$ is a probability space and the measurable functions $\{P_n\}$ are independent and distributed with law $\mathbb{P}_n$. Suppose now that we are also given random variables $X_n : (\Omega_n, \mathcal{F}_n, \mathbb{P}_n) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, for all $n = 1, 2, \ldots$ Consider the random variables

$$\mathbf{X}_n : (\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathbb{P}^{\mathbb{N}}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ defined by } \mathbf{X}_n(\mathbf{w}) := X_n(P_n(\mathbf{w})) \text{ for all } n = 1, 2, \ldots$$

They are independent, for reasons similar to those given in Lemma 271. We show now, that $\mathbf{X}_n$ has the same distribution as X_n.

Lemma 250. The cumulative distribution function of $\mathbf{X}_n$ is equal to the cumulative distribution function of X_n. Hence $\mathbb{E}\mathbf{X}_n = \mathbb{E}X_n$.

Proof. Indeed

$$\mathbb{P}^{\mathbb{N}}(\mathbf{X}_n \le x) = \int_{\Omega^{\mathbb{N}}} \mathbf{1}_{\{\mathbf{X}_n \le x\}}(\mathbf{w})\, d\mathbb{P}^{\mathbb{N}}(\mathbf{w}) = \int_{\Omega^{\mathbb{N}}} \mathbf{1}_{(-\infty,x]}(\mathbf{X}_n(\mathbf{w}))\, d\mathbb{P}^{\mathbb{N}}(\mathbf{w})$$

$$= \int_{\Omega^{\mathbb{N}}} \mathbf{1}_{(-\infty,x]}(X_n(P_n(\mathbf{w})))\, d\mathbb{P}^{\mathbb{N}}(\mathbf{w}) = \int_{\Omega_n} \mathbf{1}_{(-\infty,x]}(X_n(\omega))\, d\mathbb{P}_n(\omega),$$

where in the last equality, we performed a change of variable, according to Proposition 165, with $g(\omega) := \mathbf{1}_{(-\infty,x]}(X_n(\omega)) \geq 0$, and using the fact that the law of P_n is exactly the measure $\mathbb{P}_n$. Thus, we continue

$$\mathbb{P}^{\mathbb{N}}(\mathbf{X}_n \leq x) = \int_{\Omega_n} \mathbf{1}_{(-\infty,x]}(X_n(\omega))\, d\mathbb{P}_n(\omega) = \int_{\Omega_n} \mathbf{1}_{\{X_n \leq x\}}(\omega)\, d\mathbb{P}_n(\omega) = \mathbb{P}_n(X_n \leq x).$$

The fact that $\mathbb{E}\mathbf{X}_n = \mathbb{E}X_n$ should be now clear. □

Exercise 251. *Prove Theorem 190.*

Now, return to (53). We apply the above discussion with $X_i = X$ and $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i) = (\Omega, \mathcal{F}, \mathbb{P})$ for all $i = 1, 2 \ldots$ Let $\mathbf{w} = (\omega_1, \omega_2, \ldots)$, then one can rewrite the function on the left-hand side of (53)

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(\omega_k),\infty)}(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(P_k(\mathbf{w})),\infty)}(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[\mathbf{X}_k(\mathbf{w}),\infty)}(x).$$

Since the random variables $\{\mathbf{X}_n\}$ are independent and identically distributed, by the Glivenko-Cantelli theorem, we have that for almost all $\mathbf{w} \in \Omega^{\mathbb{N}}$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[\mathbf{X}_k(\mathbf{w}),\infty)}(x) \text{ converges to the c.d.f. of } \mathbf{X}_n \text{ as } n \to \infty$$

But the c.d.f. of $\mathbf{X}_n$ is exactly equal to the c.d.f. of X, see Lemma 250. This establishes (53) and clarifies that the almost sure convergence in it, has to be taken with respect to the measure $\mathbb{P}^{\mathbb{N}}$ on the set of all sequences of elementary events $(\omega_1, \omega_2, \ldots)$.

In practice events with measure zero never occur. Thus, in practice, if we keep performing the experiment over and over again obtaining a sequence of outcomes $(\omega_1, \omega_2, \ldots)$ and measurements $X(\omega_1), X(\omega_2), \ldots$ then

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{[X(\omega_k),\infty)}(x)$$

converges, as $n \to \infty$, to the distribution of the unknown random variable X.

6 Conditional expectation

Let us start with special cases.

Let $B \in \mathcal{F}$ have positive probability $\mathbb{P}(B) > 0$. Then, the conditional probability of $A \in \mathcal{F}$, given B, is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{\mathbb{P}(B)} \int_{\Omega} \mathbf{1}_A \mathbf{1}_B\, d\mathbb{P} = \frac{1}{\mathbb{P}(B)} \int_{A} \mathbf{1}_B\, d\mathbb{P}.$$

We have known for some time that $\mathbb{P}(\cdot|B)$ is a measure on $\mathcal{F}$. Denote this measure by $\mathbb{P}_B$ for short. That is

$$\mathbb{P}_B(A) := \mathbb{P}(A|B).$$

What we see now is that $\mathbb{P}_B(A)$ can be represented in the form of Problem (viii) from Homework 5 with X raplaced by $\mathbf{1}_B$:

$$\mathbb{P}_B(A) = \frac{1}{\mathbb{E}(\mathbf{1}_B)} \int_A \mathbf{1}_B \, d\mathbb{P}.$$

So far, $\mathbb{P}_B(A)$ is nothing more than the conditional probability of A given B. Now, the second part of this homework problem is really interesting. It says that for any random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, non-negative or integrable with respect to $\mathbb{P}_B$, we have

$$E_{\mathbb{P}_B}(X) = \int_\Omega X \, d\mathbb{P}_B = \frac{1}{\mathbb{P}(B)} \int_\Omega X \mathbf{1}_B \, d\mathbb{P} = \frac{1}{\mathbb{P}(B)} \int_B X \, d\mathbb{P},$$

which, by analogy, is called the *conditional expectation of* X *given* B and denoted by $E(X|B)$. That is

$$E(X|B) := \frac{1}{\mathbb{P}(B)} \int_B X \, d\mathbb{P}, \tag{54}$$

provided that $\mathbb{P}(B) > 0$. This is a number. Similarly, we can define the number $E(X|B^c)$, the conditional expectation of X given B^c by

$$\mathbb{E}(X|B^c) := \frac{1}{\mathbb{P}(B^c)} \int_{B^c} X \, d\mathbb{P}.$$

Now, consider the random variable $X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, defined by

$$X_0 := E(X|B)\mathbf{1}_B + E(X|B^c)\mathbf{1}_{B^c}. \tag{55}$$

It is not only measurable with with respect to $\mathcal{F}$ (meaning that the preimage under X_0 of a Borel set is in $\mathcal{F}$) but also measurable with respect to $\sigma(\{B\}) = \{\emptyset, B, B^c, \Omega\}$. That is, $X_0 : (\Omega, \sigma(\{B\}), \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is also a random variable. It is trivial to see that

$$\int_C X_0 \, d\mathbb{P} = \int_C X \, d\mathbb{P} \quad \text{holds for all } C \in \sigma(\{B\}) \tag{56}$$

(There are four choices for C in this case.) Conversely, a random variable $X_0 : (\Omega, \sigma(\{B\}), \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying (56) is of the form (55) almost surely. In other words, there is a unique (to within almost sure equality) $\sigma(\{B\})$-measurable random variable that satisfies (56). For this reason, X_0 is called the *conditional expectation of* X *given* $\sigma(\{B\})$, denoted by $\mathbb{E}(X|\sigma(\{B\}))$. We should emphasize: X_0 is a $\sigma(\{B\})$-measurable random variable.

What is the expectation of X_0? Using the definitions of the numbers $E(X|B)$ and $E(X|B^c)$, we get

$$\mathbb{E}(X_0) = E(X|B)\mathbb{P}(B) + E(X|B^c)\mathbb{P}(B^c) = \int_B X \, d\mathbb{P} + \int_{B^c} X \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P} = \mathbb{E}X.$$

Note that $\{B, B^c\}$ is a partition of Ω. The above arguments can be made for any partition $\{B_1, B_2, \ldots, B_n\}$ of Ω, where $\mathbb{P}(B_k) > 0$ for all $k = 1, \ldots, n$. (That is, the sets $\{B_1, B_2, \ldots, B_n\}$ are disjoint and their union is Ω.) In this case, the step-function

$$X_0 := \sum_{i=1}^{n} E(X|B_i)\mathbf{1}_{B_i}, \tag{57}$$

where

$$E(X|B_i) := \frac{1}{\mathbb{P}(B_i)} \int_{B_i} X \, d\mathbb{P}$$

is called the *conditional expectation of* X *given* $\sigma(\{B_i : i = 1, \ldots, n\})$. It is a $\sigma(\{B_i : i = 1, \ldots, n\})$-measurable and satisfies

$$\int_C X_0 \, d\mathbb{P} = \int_C X \, d\mathbb{P} \quad \text{for all } C \in \sigma(\{B_i : i = 1, \ldots, n\})$$

Exercise 252. *If the random variables* $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *are integrable and satisfy*

$$\int_A X \, d\mathbb{P} = \int_A Y \, d\mathbb{P} \quad \textit{for all } A \in \mathcal{F}.$$

then $X = Y$ *a.s.*

In general, we have the following theorem that we will not prove.

Theorem 253. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a non-negative (resp. integrable) random variable. Then, for every σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, there is a non-negative (resp. integrable) random variable $X_0 : (\Omega, \mathcal{G}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, unique, to within almost sure equality with respect to $\mathbb{P}$, such that

$$\int_C X_0 \, d\mathbb{P} = \int_C X \, d\mathbb{P} \quad \text{for all } C \in \mathcal{G}. \tag{58}$$

If X is both non-negative and integrable, then so is X_0.

The uniqueness in Theorem 253 is easy to show. Suppose there are two $\mathcal{G}$-measurable random variables X_0 and X_0' that satisfy equations (58). Substituting X_0 and X_0' into (58) and comparing the left-hand sides we find that

$$\int_C X_0 \, d\mathbb{P} = \int_C X_0' \, d\mathbb{P} \quad \text{for all } C \in \mathcal{G}.$$

Using Exercise 252, we conclude that $X_0 = X_0'$ a.s.

Definition 254. Suppose that $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is non-negative or integrable random variable. Suppose that $\mathcal{G}$ is a σ-algebra contained in $\mathcal{F}$. A random variable $X_0 : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called the *conditional expectation of* X *given* $\mathcal{G}$ if

(i) X_0 is $\mathcal{G}$-measurable; and

(ii) X_0 satisfies (58).

It is denoted by $\mathbb{E}(X|\mathcal{G})$.

• Theorem 253 guarantees that a random variable, satisfying the two conditions in Definition 254, exists. But it also states that it may not be unique. If two random variables X_0 and Y_0 satisfy the two conditions in Definition 254, then $X_0 = Y_0$ a.s. (But, if you change the σ-algebra $\mathcal{G}$, then the conditional expectation may change completely.)

• One should remember that $\mathbb{E}(X|\mathcal{G})$ stands for a $\mathcal{G}$-measurable function, that is, it is a random variable on $(\Omega, \mathcal{G}, \mathbb{P})$. Since $\mathcal{G} \subseteq \mathcal{F}$, it is obvious that $\mathbb{E}(X|\mathcal{G})$ is also $\mathcal{F}$-measurable. Hence, $\mathbb{E}(X|\mathcal{G})$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ as well.

• The random variable $\mathbb{E}(X|\mathcal{G})$ is only defined to within almost sure equality with respect to $\mathbb{P}$. This means that if $X_0, Y_0 : (\Omega, \mathcal{G}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are two random variables that satisfy (58), then $X_0 = Y_0$ $\mathbb{P}$-almost surely. That is why statements about conditional expectation, often hold only $\mathbb{P}$-almost surely.

• What happens if we take $\mathcal{G} = \{\emptyset, \Omega\}$? If $\mathcal{G} = \{\emptyset, \Omega\}$, then a random variable is $\mathcal{G}$-measurable if and only if it is equal to a constant almost surely, so $X_0 = c$ a.s. To find that constant, let $C := \Omega$ in (58):

$$c = \int_\Omega X_0 \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P} = \mathbb{E}X.$$

• What happens if we take $\mathcal{G} = \{\emptyset, A, A^c, \Omega\}$? If $\mathcal{G} = \{\emptyset, A, A^c, \Omega\}$, then, Example 84, a random variable X_0 is $\mathcal{G}$-measurable if and only if it is $X_0 = a\mathbf{1}_A + b\mathbf{1}_{A^c}$. Now, if $X_0 = \mathbb{E}(X|\mathcal{G})$, then we can find the values of the constants a and b using (58) with C replaced by A and then by A^c:

$$a\mathbb{P}(A) = \int_A X_0 \, d\mathbb{P} = \int_A X \, d\mathbb{P} = \mathbb{P}(A)\mathbb{E}(X|A),$$
$$b\mathbb{P}(A^c) = \int_{A^c} X_0 \, d\mathbb{P} = \int_{A^c} X \, d\mathbb{P} = \mathbb{P}(A^c)\mathbb{E}(X|A^c).$$

Thus,

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X|A)\mathbf{1}_A + \mathbb{E}(X|A^c)\mathbf{1}_{A^c}.$$

• The situation in the last two bullets generalizes easily. Let I be a finite or countably infinite index set, that is $I = \{1, 2, \dots, n\}$ or $I = \{1, 2, \dots\}$. Let $\{B_i : i \in I\}$ be sets in $\mathcal{F}$ with $\mathbb{P}(B_i) > 0$ for all $i \in I$. Suppose the sets $\{B_i : i \in I\}$ are disjoint and their union is Ω. If $\mathcal{G} = \sigma(B_i : i \in I)$, then

$$\mathbb{E}(X|\mathcal{G}) = \sum_{i=1}^{\infty} E(X|B_i)\mathbf{1}_{B_i}. \tag{59}$$

This example generalizes (57) since now the σ-algebra $\mathcal{G}$ may be generated by countably many sets.

• What happens if we take $\mathcal{G} = \sigma(X)$? First, since $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a random variable, that is $\mathcal{F}$-measurable, we know that $\sigma(X) \subseteq \mathcal{F}$, so $\sigma(X)$ is a legitimate choice for $\mathcal{G}$ in Theorem 253. Since, X is $\sigma(X)$-measurable, and it obviously satisfies (58), when put in place of X_0, then the uniqueness part of Theorem 253 guarantees that $X = \mathbb{E}(X|\mathcal{G})$ a.s.

• The arguments in the previous bullet apply to any σ-algebra $\mathcal{G}$ satisfying $\sigma(X) \subseteq \mathcal{G} \subseteq \mathcal{F}$. For such $\mathcal{G}$, we have $\mathbb{E}(X|\mathcal{G}) = X$ a.s.

• If the σ-algebra $\mathcal{G}$ contains a minimal set A, that is, no proper subsets of A are in $\mathcal{G}$, then a $\mathcal{G}$-measurable random variable X_0 must be a constant on A. If, in addition, X_0 satisfies (58), then the value of this constant is $\mathbb{E}(X|A)$.

• The proceeding bullets should clarify what is the intuitive meaning of the random variable $\mathbb{E}(X|\mathcal{G})$: its values are the average values of X over the sets in $\mathcal{G}$.

• Often one encounters expression like $\mathbb{E}(X|Y)$, where X and Y are random variables. This is a short-hand notation for $\mathbb{E}(X|\sigma(Y))$.

• Similarly to the previous bullet, if $\{Y_i : i \in I\}$ is any family of random variables (finitely many or not), by $\mathbb{E}(X|Y_i, i \in I)$ one understands $\mathbb{E}(X|\mathcal{G})$, where $\mathcal{G}$ is the σ-algebra generated by $\{Y_i : i \in I\}$ on Ω.

The following properties of conditional expectation follow easily from the definition and the corresponding properties of general random variables.

Proposition 255. Assume that $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are random variables that are either both non-negative or both integrable. Let $\mathcal{G} \subseteq \mathcal{F}$. Then

(i) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$

(ii) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.

(iii) If $X = Y$ a.s. then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(Y|\mathcal{G})$ a.s.

(iv) If $X = c$ a.s. then $\mathbb{E}(X|\mathcal{G}) = c$ a.s.

(v) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$ a.s., where $a, b \geq 0$ if X and Y are both non-negative, or $a, b \in \mathbb{R}$ if X and Y are both integrable.

(vi) If $X \leq Y$ a.s. then $\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})$ a.s.

(vii) $|\mathbb{E}(X \,|\, \mathcal{G})| \leq \mathbb{E}(|X| \,|\, \mathcal{G})$ a.s.

(viii) If $\{X_n\}_{n=1}^{\infty}$ is an increasing sequence of non-negative random variables converging to X, then

$$\lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

(ix) If $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables converging to X a.s. and if there is an integrable random variable Y such that $|X_n| \leq Y$ for all n, then

$$\lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

(x) For any $\epsilon > 0$

$$\mathbb{E}(\mathbf{1}_{\{X\geq\epsilon\}}|\mathcal{G}) \leq \frac{\mathbb{E}(X^2|\mathcal{G})}{\epsilon^2}$$

(xi) If $f : \mathbb{R} \to \mathbb{R}$ is a convex function, then

$$f(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(f(X)|\mathcal{G}).$$

(xii) If $p \geq 1$ and $\mathbb{E}|X|^p < \infty$, then

$$|\mathbb{E}(X|\mathcal{G})|^p \leq \mathbb{E}(|X|^p|\mathcal{G}) \text{ a.s.}$$

Taking expectation from both sides of the last inequality and using property (i), gives

$$\mathbb{E}(|\mathbb{E}(X|\mathcal{G})|^p) \leq \mathbb{E}(|X|^p) < \infty.$$

(xiii) If $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q < \infty$, where $p, q \in (1, \infty)$ satisfy $1/p + 1/q = 1$, then

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

Proof. (i) Take $C := \Omega$ in (58) to get

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X_0) = \int_\Omega X_0 \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P} = \mathbb{E}X.$$

The rest is left for homework. □

Lemma 256. Suppose $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a non-negative (resp. integrable) random variable and $\mathcal{G} \subseteq \mathcal{F}$. Let X_0 be a $\mathcal{G}$-measurable random variable. Then $X_0 = \mathbb{E}(X|\mathcal{G})$ a.s. if and only if

$$\int_\Omega Z X_0 \, d\mathbb{P} = \int_\Omega Z X \, d\mathbb{P} \tag{60}$$

for all $\mathcal{G}$-measurable non-negative (resp. bounded) random variables Z.

Proof. (⇐) Just take $Z = \mathbf{1}_C$ for $C \in \mathcal{G}$ and we obtain (58). (⇒) Suppose now (58) holds. Then, (60) holds right away for indicator functions $Z = \mathbf{1}_C$ for $C \in \mathcal{G}$ and by linearity, it holds for all $\mathcal{G}$-step-functions Z. If $X \geq 0$, then $X_0 = \mathbb{E}(X|\mathcal{G}) \geq 0$. Take any $Z \geq 0$. Let $\{Z_n\}$ be an increasing sequence of positive step-functions converging to Z. By the above, we have

$$\int_\Omega Z_n X_0 \, d\mathbb{P} = \int_\Omega Z_n X \, d\mathbb{P}.$$

Taking the limit as n goes to infinity and using the Monotone Convergence Theorem shows that (60) holds for all $Z \geq 0$.

If X is integrable, we want to show that (60) holds for all bounded $\mathcal{G}$-measurable Z. First, take any bounded $Z \geq 0$, that is $|Z| \leq C$. Let $\{Z_n\}$ be an increasing sequence of positive step-functions converging to Z. Then, $|Z_n X_0| \leq C|X_0|$ and $|Z_n X| \leq C|X|$ hold for every n. Since X and X_0 are integrable, by the Dominated Convergence Theorem we see that (60) holds for the bounded $Z \geq 0$. Finally, take any bounded Z. Using the decomposition $Z = Z^+ - Z^-$ and the fact that both Z^+ and Z^- are non-negative and bounded, implies that (60) holds for Z^+ and Z^-, and hence for Z. □

All properties listed in Proposition 255 are extensions of corresponding properties of ordinary random variables. Lemma 256 allows us to derive new ones.

Proposition 257. Let X and Y be either both non-negative or satisfy $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$, and $\mathbb{E}|XY| < \infty$.

(i) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(XY|\mathcal{G}) = X\mathbb{E}(Y|\mathcal{G})$ a.s..

(ii) $\mathbb{E}(X\mathbb{E}(Y|\mathcal{G})|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})\mathbb{E}(Y|\mathcal{G})$ a.s..

Proof. (i) Suppose first that both X and Y are non-negative. Let Z be any non-negative, $\mathcal{G}$-measurable random variable. Apply Lemma 256 with Z replaced by the non-negative random variable ZX and X_0 replaced by the $\mathcal{G}$-measurable random variable $\mathbb{E}(Y|\mathcal{G})$, to get

$$\int_\Omega ZX\mathbb{E}(Y|\mathcal{G})\, d\mathbb{P} = \int_\Omega ZXY\, d\mathbb{P}. \tag{61}$$

Since $X\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable random variable, then by Lemma 256 again, but this time applied to Z itself and X_0 replaced by $X\mathbb{E}(Y|\mathcal{G})$, we conclude that $X\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(XY|\mathcal{G})$.

Suppose now, $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$, and $\mathbb{E}|XY| < \infty$. Let Z be any bounded, $\mathcal{G}$-measurable random variable. Define the bounded random variables $X_n := X\mathbf{1}_{\{-n\leq X\leq n\}}$. Apply Lemma 256 with Z replaced by the bounded random variable ZX_n and X_0 replaced by the $\mathcal{G}$-measurable random variable $\mathbb{E}(Y|\mathcal{G})$, to get

$$\int_\Omega ZX_n\mathbb{E}(Y|\mathcal{G})\, d\mathbb{P} = \int_\Omega ZX_nY\, d\mathbb{P}.$$

Since $X_n\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable random variable, then by Lemma 256 again, but this time applied to Z itself and X_0 replaced by $X_n\mathbb{E}(Y|\mathcal{G})$, we conclude that

$$X_n\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(X_nY|\mathcal{G}) \quad \text{for all } n = 1, 2, \ldots \tag{62}$$

Let n approach infinity. Since, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$ the left-hand side of (62) approaches $X\mathbb{E}(Y|\mathcal{G})$. Since, $|X_nY| \leq |XY|$ and $|XY|$ is integrable, by Proposition 255, part (ix), the right-hand side of (62) approaches $\mathbb{E}(XY|\mathcal{G})$.

(ii) Apply the first part, together with the fact that $\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable. □

Proposition 258. If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$, then

(i) $\mathbb{E}(\mathbb{E}(X|\mathcal{G}_1)|\mathcal{G}_2) = \mathbb{E}(X|\mathcal{G}_1)$;

(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1) = \mathbb{E}(X|\mathcal{G}_1)$.

Proof. (i) By definition, $\mathbb{E}(X|\mathcal{G}_1)$ is $\mathcal{G}_1$-measurable and hence it is $\mathcal{G}_2$-measurable. Apply part (i) of Proposition 257.

(ii) We check the two conditions in Definition 254. First, $\mathbb{E}(X|\mathcal{G}_1)$ is $\mathcal{G}_1$-measurable. Second, for all $C \in \mathcal{G}_1$, we have

$$\int_C \mathbb{E}(X|\mathcal{G}_1)\, d\mathbb{P} = \int_C X\, d\mathbb{P} = \int_C \mathbb{E}(X|\mathcal{G}_2)\, d\mathbb{P}.$$

The first equality holds since $\mathbb{E}(X|\mathcal{G}_1)$ is the conditional expectation of X. The second, holds since $\mathbb{E}(X|\mathcal{G}_2)$ is the conditional expectation of X and $C \in \mathcal{G}_2$ (recall that $\mathcal{G}_1 \subseteq \mathcal{G}_2$). This implies that $\mathbb{E}(X|\mathcal{G}_1)$ is the conditional expectation of $\mathbb{E}(X|\mathcal{G}_2)$, given $\mathcal{G}_1$. □

Proposition 259. Suppose X is non-negative or integrable. Suppose $\sigma(X)$ is independent of $\mathcal{G}$ (in that case, we say that X and $\mathcal{G}$ are independent), then

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X).$$

Proof. The random variable $\mathbb{E}(X)$ (it is a constant) is $\mathcal{G}$-measurable. For every $C \in \mathcal{G}$, we have

$$\int_C X\,d\mathbb{P} = \int_\Omega X\mathbf{1}_C\,d\mathbb{P} = \mathbb{E}(X\mathbf{1}_C) = \mathbb{E}(X)\mathbb{E}(\mathbf{1}_C) = \mathbb{E}(X)\int_\Omega \mathbf{1}_C\,d\mathbb{P} = \int_C \mathbb{E}(X)\,d\mathbb{P},$$

where we used that the random variables X and $\mathbf{1}_C$ are independent. This shows that $\mathbb{E}(X) = \mathbb{E}(X|\mathcal{G})$ a.s. □

We give another important characterization of conditional expectation.

Proposition 260. Suppose $\mathbb{E}|X|^2 < \infty$. The conditional expectation $\mathbb{E}(X|\mathcal{G})$ is the unique $\mathcal{G}$-measurable random variable (up to almost sure equality) that minimizes the expectation

$$\mathbb{E}\big((X - X_0)^2\big)$$

over all X_0 with $\mathbb{E}|X_0|^2 < \infty$.

Proof. Let Y be a $\mathcal{G}$-measurable random variable with $\mathbb{E}|Y|^2 < \infty$. Let $X_0 := \mathbb{E}(X|\mathcal{G})$. We want to show that

$$\mathbb{E}\big((X - X_0)^2\big) \leq \mathbb{E}\big((X - Y)^2\big). \tag{63}$$

In fact, we show the stronger identity from which the inequality follows

$$\mathbb{E}\big((X - X_0)^2\big) + \mathbb{E}\big((X_0 - Y)^2\big) = \mathbb{E}\big((X - Y)^2\big).$$

Expanding all squares, gives

$$\mathbb{E}(X^2) - 2\mathbb{E}(XX_0) + \mathbb{E}(X_0^2) + \mathbb{E}(X_0^2) - 2\mathbb{E}(X_0Y) + \mathbb{E}(Y^2) = \mathbb{E}(X^2) - 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$$

which simplifies to

$$\mathbb{E}(XY) + \mathbb{E}(X_0^2) = \mathbb{E}(XX_0) + \mathbb{E}(X_0Y) \tag{64}$$

On the one hand, since X_0 is $\mathcal{G}$-measurable, we have

$$\mathbb{E}(XX_0) = \mathbb{E}\big(\mathbb{E}(XX_0|\mathcal{G})\big) = \mathbb{E}\big(X_0\mathbb{E}(X|\mathcal{G})\big) = \mathbb{E}(X_0^2).$$

On the other hand, since Y is $\mathcal{G}$-measurable, we have

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY|\mathcal{G})) = \mathbb{E}(Y\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(YX_0).$$

Substituting into (64), establishes its validity. Note that (63) holds with equality if and only if $\mathbb{E}\big((X_0 - Y)^2\big) = 0$, that is, if and only if $Y = X_0$ a.s. □

Proposition 260 says that for a random variable X with $\mathbb{E}|X|^2 < \infty$, the conditional expectation $\mathbb{E}(X|\mathcal{G})$ is precisely the best (least-squares) approximation of X by a $\mathcal{G}$-measurable, square-integrable, random variable.

We conclude this section with a case study, designed to better understand the role of condition (i) in Definition 254. Let $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two integrable random variables and let $\mathcal{G}$ be a σ-algebra contained in $\mathcal{F}$. Suppose that

$$\int_C X \, d\mathbb{P} = \int_C Y \, d\mathbb{P} \quad \text{for all } C \in \mathcal{G}.$$

Can we conclude that one of X and Y is the conditional expectation of the other?

The answer varies. If both X and Y are $\mathcal{G}$-measurable, then by Exercise 252, we see that $X = Y$ a.s.. If X is $\mathcal{G}$-measurable, but Y is not, then by Definition 254, we have that $X = \mathbb{E}(Y|\mathcal{G})$. Similarly, if Y is $\mathcal{G}$-measurable, but X is not, then $Y = \mathbb{E}(X|\mathcal{G})$. Finally, if both X and Y are not $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(Y|\mathcal{G})$.

6.1 Special case: conditional probability

In the definition of conditional expectation, we may specialize X to be an indicator random variable. Thus, we arrive at the concept of conditional probability.

Definition 261. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra and let $A \in \mathcal{F}$. The *conditional probability of A, given* $\mathcal{G}$ is defined by

$$\mathbb{P}(A|\mathcal{G}) := \mathbb{E}(\mathbf{1}_A|\mathcal{G}).$$

Thus, $\mathbb{P}(A|\mathcal{G})$ is a $\mathcal{G}$-measurable, (a.s.) non-negative random variable, satisfying

$$\int_C \mathbb{P}(A|\mathcal{G}) \, d\mathbb{P} = \int_C \mathbf{1}_A \, d\mathbb{P} = \mathbb{P}(A \cap C).$$

Using properties of conditional expectation, we easily get the following properties of conditional probability.

Proposition 262. Conditional probability satisfies the following properties.

(i) $0 \le \mathbb{P}(A|\mathcal{G}) \le 1$ a.s.;

(ii) $\mathbb{P}(\emptyset|\mathcal{G}) = 0$ a.s. and $\mathbb{P}(\Omega|\mathcal{G}) = 1$ a.s.;

(iii) If $A_1 \subset A_2$, then $\mathbb{P}(A_1|\mathcal{G}) \le \mathbb{P}(A_2|\mathcal{G})$ a.s.;

(iv) For every sequence $A_1, A_2, \ldots$ of pairwise disjoint sets from $\mathcal{F}$, we have

$$\mathbb{P}\Big(\bigcup_{k=1}^{\infty} A_k|\mathcal{G}\Big) = \sum_{k=1}^{\infty} \mathbb{P}(A_k|\mathcal{G}) \quad \text{a.s.}$$

The properties listed in Proposition 262 do not imply that $A \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ is a probability measure on $\mathcal{F}$ for almost every $\omega \in \Omega$, because in each property, the exceptional $\mathbb{P}$-null set depends on the events A, A_1, A_2 etc., involved. The union of these (possibly uncountably many) null sets may not be a null set.

For example, let $\{B_1, B_2, \ldots\}$ be a sequence of disjoint sets with $\mathbb{P}(B_k) > 0$, for all $k = 1, \ldots$, and consider the σ-algebra $\mathcal{G} := \sigma(B_1, B_2, \ldots)$. We know that since $\mathbb{P}(A|\mathcal{G})$ is $\mathcal{G}$-measurable random variable, it has the form

$$\mathbb{P}(A|\mathcal{G}) = \sum_{k=1}^{\infty} b_k \mathbf{1}_{B_k}$$

for some constants $b_1, b_2, \ldots$ Integrate both sides over the set B_k to find $b_k := \mathbb{P}(A|B_k)$. That is,

$$\mathbb{P}(A|\mathcal{G}) = \sum_{k=1}^{\infty} \mathbb{P}(A|B_k) \mathbf{1}_{B_k}.$$

This formula, as should be expected, is just a special case of (59) for $X = \mathbf{1}_A$.

6.2 Special case: conditional expectation of X given Y

Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be an integrable random variable. In this subsection, we consider the conditional expectation of X with respect to a σ-algebra generated by a measurable function $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$.

We need the following theorem, a special case of which appeared on Problem Set 4 (see Problem (viii) there). Note that the theorem is purely functional one and has nothing to do with probability.

Theorem 263 (Factorization lemma)**.** Let $X : \Omega \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $Y : \Omega \to (S, \mathcal{S})$ be two functions on the set Ω, where $(S, \mathcal{S})$ is a measurable space. Then, $\sigma(X) \subseteq \sigma(Y)$ if and only if there is a measurable function $g : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that

$$X = g \circ Y.$$

If X is non-negative, then g is non-negative.

Note that the function g is uniquely determined on the set $Y(\Omega) \subseteq S$. Indeed, suppose there is another function $h : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying $X = h \circ Y$. For every $y \in Y(\Omega)$ there is an $\omega \in \Omega$ such that $Y(\omega) = y$. Then $g(y) = g(Y(\omega)) = X(\omega) = h(Y(\omega)) = h(y)$.

Theorem 264. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be an integrable random variable and let $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ be a measurable map. There is an integrable function $g : (S, \mathcal{S}, \mathbb{P}_Y) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that

$$\mathbb{E}(X|Y) = g \circ Y. \tag{65}$$

In addition, g satisfies

$$\int_B g \, d\mathbb{P}_Y = \int_{\{Y \in B\}} X \, d\mathbb{P} \quad \text{for all } B \in \mathcal{S}. \tag{66}$$

Conversely, if $g : (S, \mathcal{S}, \mathbb{P}_Y) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is an integrable function that satisfies (66), then

$$\mathbb{E}(X|Y) = g \circ Y \text{ a.s.}$$

Proof. Since, by definition, $\mathbb{E}(X|Y)$ is $\sigma(Y)$-measurable, the factorization lemma implies that there is a measurable function $g : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that (65) holds.

To show that g is integrable, apply the change of variables formula to the non-negative function $|g|$, as follows:

$$\int_S |g|\, d\mathbb{P}_Y = \int_\Omega |g(Y)|\, d\mathbb{P} = \int_\Omega |\mathbb{E}(X|Y)|\, d\mathbb{P} < \infty,$$

since $\mathbb{E}(X|Y)$ is integrable whenever X is.

Next, for every $B \in \mathcal{S}$, by the change of variables formula again, we have

$$\begin{aligned} \int_B g\, d\mathbb{P}_Y &= \int_S \mathbf{1}_B g\, d\mathbb{P}_Y = \int_\Omega (\mathbf{1}_B \circ Y)(g \circ Y)\, d\mathbb{P} = \int_{\{Y \in B\}} g \circ Y\, d\mathbb{P} \\ &= \int_{\{Y \in B\}} \mathbb{E}(X|Y)\, d\mathbb{P} = \int_{\{Y \in B\}} X\, d\mathbb{P}, \end{aligned}$$

establishing (66).

Conversely, suppose $g : (S, \mathcal{S}, \mathbb{P}_Y) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is an integrable function satisfying (66). Clearly, $g \circ Y$ is $\sigma(Y)$-measurable. Also, for all $B \in \mathcal{S}$, by the change of variable formula, we have

$$\int_{\{Y \in B\}} g \circ Y\, d\mathbb{P} = \int_B g\, d\mathbb{P}_Y = \int_{\{Y \in B\}} X\, d\mathbb{P},$$

which by definition of $\sigma(Y)$ is just

$$\int_C g \circ Y\, d\mathbb{P} = \int_C X\, d\mathbb{P} \quad \text{for all } C \in \sigma(Y).$$

This means $g \circ Y = \mathbb{E}(X|Y)$ a.s.. □

We know that the function g in (65) is uniquely determined on the set $Y(\Omega)$. Notice that this is the same as saying that g is uniquely determined $\mathbb{P}_Y$-almost everywhere, since $\mathbb{P}_Y(Y(\Omega)) = 1$.

Definition 265. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be an integrable random variable and let $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ be a measurable map. Let $g : (S, \mathcal{S}, \mathbb{P}_Y) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the integrable function satisfying (65). For every $y \in S$, we define

$$\mathbb{E}(X|Y = y) := g(y)$$

and call it the *conditional expectation of X given that $Y = y$.*

Note that, while $\mathbb{E}(X|Y)$ is a random variable, $\mathbb{E}(X|Y = y)$ is a number. By the comment before the definition, the function $y \in S \mapsto \mathbb{E}(X|Y = y) \in \mathbb{R}$ is uniquely defined only on the set $Y(\Omega)$, which is enough because it has $\mathbb{P}_Y$-measure one. By (65), for every $\omega \in \Omega$, we have

$$\mathbb{E}(X|Y)(\omega) = g(Y(\omega)) = \mathbb{E}(X|Y = Y(\omega)). \tag{67}$$

In particular, if $Y(\omega_1) = Y(\omega_2)$, then $\mathbb{E}(X|Y)(\omega_1) = \mathbb{E}(X|Y)(\omega_2)$. That is, $\mathbb{E}(X|Y)$ is constant on every set $\{\omega \in \Omega : Y(\omega) = y\}$.

Suppose that the singleton set $\{y\}$ is an element of $\mathcal{S}$ for some $y \in S$. Then, according to (66), applied with $B = \{y\}$, we have

$$g(y)\mathbb{P}(Y = y) = \int_{\{Y=y\}} X \, d\mathbb{P}.$$

If, in addition, $\mathbb{P}(Y = y) = \mathbb{P}_Y(\{y\}) > 0$, we obtain

$$\mathbb{E}(X|Y = y) = g(y) = \frac{1}{\mathbb{P}(Y = y)} \int_{\{Y=y\}} X \, d\mathbb{P} = \mathbb{E}(X|\{Y = y\}).$$

This is precisely formula (54) with $B = \{Y = y\}$. This shows that, if $\{y\} \in \mathcal{S}$ and $\mathbb{P}(Y = y) > 0$, the two notations $\mathbb{E}(X|Y = y)$ and $\mathbb{E}(X|\{Y = y\})$ stand for the same thing and can be used interchangeably.

Example 266. Let I be a finite of countably infinite index set, that is $I = \{1, 2, \ldots, n\}$ or $I = \{1, 2, \ldots\}$. Let $\{B_i : i \in I\}$ be sets in $\mathcal{F}$ with $\mathbb{P}(B_i) > 0$ for all $i \in I$. Suppose the sets $\{B_i : i \in I\}$ are disjoint and their union is Ω.

Consider the measurable space $(S, \mathcal{S}) := (I, 2^I)$ and the measurable function $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (I, 2^I)$ defined by $Y(\omega) = i$ for all $\omega \in B_i$. We know that $\sigma(Y) = \sigma(B_i : i \in I)$ and that

$$\mathbb{E}(X|Y) = \sum_{i=1}^{\infty} \mathbb{E}(X|B_i)\mathbf{1}_{B_i}.$$

The relevant g is given by $g(i) = \mathbb{E}(X|B_i)$ for all $i \in I$.

We conclude with another very useful and intuitive result facilitating the computation of conditional expectation. Try to verify this result independently, when X and Y have joint density. It is much easier in this case.

Proposition 267. Suppose X and Y are independent random variables. Let $h : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \to (\mathbb{R}, \mathcal{B})$ be measurable and either non-negative or $\mathbb{E}|h(X, Y)| < \infty$. Then

$$\mathbb{E}(h(X, Y)|Y)(\omega) = \mathbb{E}(h(X, Y(\omega))) \quad \text{for almost all } \omega \in \Omega. \tag{68}$$

Proof. Recall that

$$\mathbb{E}h(X, Y) = \int_{\mathbb{R}^2} h(x, y) \, d\mathbb{P}_{(X,Y)}$$

and since X and Y are independent, we have $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \times \mathbb{P}_Y$. So, by the Fubini's theorem, we have

$$\begin{aligned}
\mathbb{E}h(X, Y) &= \int_{\mathbb{R}^2} h(x, y) \, d\mathbb{P}_{(X,Y)} = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} h(x, y) \, d\mathbb{P}_X(x) \Big) \, d\mathbb{P}_Y(y) \\
&= \int_{\mathbb{R}} \Big(\int_{\Omega} h(X, y) \, d\mathbb{P} \Big) \, d\mathbb{P}_Y(y) = \int_{\mathbb{R}} \mathbb{E}h(X, y) \, d\mathbb{P}_Y(y).
\end{aligned}$$

The first part of Fubini's theorem, says that the function $g(y) := \mathbb{E}h(X,y)$ exists for $\mathbb{P}_Y$-almost all $y \in \mathbb{R}$ and that (if we define $g(y) := 0$ for those y's for which $\mathbb{E}h(X,y)$ does not exists) it is measurable (and non-negative or integrable). Hence, $g(Y)$ is $\sigma(Y)$-measurable.

Let $C \in \sigma(Y)$, that is $C = \{Y \in B\}$ for some $B \in \mathcal{B}(\mathbb{R})$. We have

$$\begin{aligned}\int_C h(X,Y)\,d\mathbb{P} &= \int_\Omega h(X,Y)\mathbf{1}_C\,d\mathbb{P} = \int_\Omega h(X,Y)\mathbf{1}_B(Y)\,d\mathbb{P} = \int_\Omega h(X,Y)\mathbf{1}_{\mathbb{R}\times B}(X,Y)\,d\mathbb{P}\\ &= \int_{\mathbb{R}^2} h(x,y)\mathbf{1}_{\mathbb{R}\times B}(x,y)\,d\mathbb{P}_{(X,Y)},\end{aligned}$$

where we used the change of variable formula. Since X and Y are independent random variables, we know that $\mathbb{P}_{(X,Y)} = \mathbb{P}_X \times \mathbb{P}_Y$. So, we can continue using the Fubini's theorem and then twice the change of variable formula

$$\begin{aligned}\int_C h(X,Y)\,d\mathbb{P} &= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} h(x,y)\mathbf{1}_{\mathbb{R}\times B}(x,y)\,d\mathbb{P}_X\Big)\,d\mathbb{P}_Y = \int_{\mathbb{R}} \mathbb{E}(h(X,y)\mathbf{1}_{\mathbb{R}\times B}(X,y))\,d\mathbb{P}_Y\\ &= \int_{\mathbb{R}} \mathbb{E}(h(X,y)\mathbf{1}_B(y))\,d\mathbb{P}_Y\\ &= \int_{\mathbb{R}} g(y)\mathbf{1}_B(y)\,d\mathbb{P}_Y\\ &= \int_\Omega g(Y)\mathbf{1}_B(Y)\,d\mathbb{P}\\ &= \int_\Omega g(Y)\mathbf{1}_C\,d\mathbb{P}\\ &= \int_C g(Y)\,d\mathbb{P}.\end{aligned}$$

By the second part of Definition 254, this shows

$$\mathbb{E}(h(X,Y)|Y) = g(Y) \text{ almost surely.}$$

Evaluate both sides of this equality at ω to see that this is the same as (68). □

6.2.1 The case when X and Y have joint density

In this section, we assume that $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are random variables and X is integrable. (That is, we consider the special case of the previous section, when $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$)

In addition, we suppose that X and Y are jointly continuously distributed. That is, there is a non-negative function $f : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\,dsdt.$$

In that case, Y also has density

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx.$$

The goal of this section is to prove the following theorem.

Theorem 268. If $f_Y(y) > 0$ for all $y \in \mathbb{R}$, then the following formula holds.

$$\mathbb{E}(X|Y) = \frac{1}{f_Y(Y)} \int_{-\infty}^{\infty} x f(x,Y)\, dx \quad \mathbb{P}\text{-almost surely.}$$

In particular, we have

$$\mathbb{E}(X|Y=y) = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f(x,y)\, dx \quad \text{for } \mathbb{P}_Y\text{-almost all } y \in \mathbb{R}. \tag{69}$$

Proof. The second formula follows from the first together with (67).

To prove the first formula, start by recalling that $\mathbb{P}_{(X,Y)}$ denotes the measure induced by $(X,Y) : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ on $\mathcal{B}(\mathbb{R}^2)$. By the change of variable theorem, for every $B \in \mathcal{B}(\mathbb{R})$. we have

$$\begin{aligned}\int_{\{Y\in B\}} X\, d\mathbb{P} &= \int_\Omega X \mathbf{1}_B(Y)\, d\mathbb{P} = \int_\Omega X \mathbf{1}_{\mathbb{R}\times B}(X,Y)\, d\mathbb{P} = \int_{\mathbb{R}^2} x \mathbf{1}_{\mathbb{R}\times B}(x,y)\, d\mathbb{P}_{(X,Y)}(x,y) \\ &= \int_{\mathbb{R}\times B} x\, d\mathbb{P}_{(X,Y)}(x,y) = \int_{\mathbb{R}\times B} x f(x,y)\, d(x,y).\end{aligned}$$

The integration in the last integral is with respect to the Lebesgue measure on $\mathbb{R}^2$. Since X is integrable, we may repeat the same calculations with $|X|$ in stead of X and $B = \mathbb{R}$, to conclude that

$$\int_{\mathbb{R}\times\mathbb{R}} |x| f(x,y)\, d(x,y) < \infty.$$

The Lebesgue measure on $\mathbb{R}^2$ is a product measure. (It is the product of the Lebesgue measure on $\mathbb{R}$ and the Lebesgue measure on $\mathbb{R}$.) Thus, by Fubini's theorem, we have

$$\int_{\{Y\in B\}} X\, d\mathbb{P} = \int_{\mathbb{R}\times B} x f(x,y)\, d(x,y) = \int_B \Big(\int_{\mathbb{R}} x f(x,y)\, dx \Big) dy.$$

Define the function

$$g(y) := \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f(x,y)\, dx.$$

We continue, using the change of variable formula in the third equality:

$$\begin{aligned}\int_{\{Y\in B\}} X\, d\mathbb{P} &= \int_B g(y) f_Y(y)\, dy = \int_{\mathbb{R}} \mathbf{1}_B(y) g(y) f_Y(y)\, dy = \int_{\mathbb{R}} \mathbf{1}_B(y) g(y)\, d\mathbb{P}_Y(y) = \int_B g(y)\, d\mathbb{P}_Y(y) \\ &= \int_{\{Y\in B\}} g(Y)\, d\mathbb{P},\end{aligned}$$

that is, we have $\mathbb{E}(X|Y) = g(Y)$. □

Recall that if $f_X(x)$ is the density of X, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

In order to extend this analogy, if $f_Y(y) > 0$ for all $y \in \mathbb{R}$, one defines the function

$$f(x|y) := \frac{f(x,y)}{f_Y(y)},$$

called the *conditional density of X given that $Y = y$.* With it, formula (69) becomes

$$\mathbb{E}(X|Y = y) = \int_{-\infty}^{\infty} x f(x|y)\, dx.$$

You may want to check (and it is simple) that for every y, the function $x \mapsto f(x|y)$ is indeed a density function. If you want to abuse notation even further, you may say that '$X|Y = y$' is the random variable having density $x \mapsto f(x|y)$.

References

[1] H. Bauer, *Probability Theory. de Gruyter Studies in Mathematics 23*, Walter de Gruyter, 1996.

[2] C. Geiss, S. Geiss, *An Introduction to Probability Theory*, 2009, http://www.math.jyu.fi/%7Egeiss/scripts/introduction-probability.pdf.

[3] R. Durrett, *Probability, Theory and Examples, 4th ed.*, Cambridge University Press, 2010.

[4] A.N. Kolmogorov, S.V. Fomin, *Introductory Real Analysis*, Dover Publications, Inc., 1975.

[5] M. Loève, *Probability theory 1, 4th ed.*, Springer Verlag, 1977.

7 Appendix A: Functions of independent random variables

The lemmas in this Appendix are needed in for the proof of Theorem 183. The first result is a corollary from Theorem 172.

Corollary 269. Suppose $\mathcal{F}_{i,j}$, $1 \le i \le n$, $1 \le j \le m_i$ are independent σ-algebras. Let $\mathcal{G}_i := \sigma(\cup_{j=1}^{m_i} \mathcal{F}_{i,j})$ for $i = 1, \ldots, n$. Then $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are independent σ-algebras.

Proof. We want to use Theorem 172. The claim would be trivial if $\cup_{j=1}^{m_i} \mathcal{F}_{i,j}$ were closed under intersection, but it need not be so. Let $\mathcal{A}_i$ be the collection of sets of the form $\cap_{j=1}^{m_i} A_{i,j}$ where $A_{i,j} \in \mathcal{F}_{i,j}$. Clearly, $\mathcal{A}_i \subseteq \sigma(\cup_{j=1}^{m_i} \mathcal{F}_{i,j})$. Now $\mathcal{A}_i$ is closed under intersection, contains Ω, and contains $\cup_{j=1}^{m_i} \mathcal{F}_{i,j}$. Hence $\sigma(\mathcal{A}_i) = \mathcal{G}_i$, for $i = 1, \ldots, n$. Since the collections of sets $\mathcal{A}_i$, $i = 1, 2, \ldots, n$ are independent, so are $\mathcal{G}_i$, for $i = 1, \ldots, n$, by Theorem 172. □

Lemma 270. Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{F}_i := \sigma(X_i)$, $i = 1, \ldots, n$ be the σ-algebras that they generate. Consider the function $(X_1, \ldots, X_n) : \Omega \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Show that the σ-algebra that it generates on Ω is

$$\sigma((X_1, \ldots, X_n)) = \sigma(\cup_{i=1}^{n} \mathcal{F}_i).$$

Proof. Since the rectangles $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{B}(\mathbb{R})$, generate $\mathcal{B}(\mathbb{R}^n)$ their preimages under $(X_1, \ldots, X_n)$, i.e. $\cap_{i=1}^{n} X_i^{-1}(A_i)$, generate $\sigma((X_1, \ldots, X_n))$. But $X_i^{-1}(A_i) \in \mathcal{F}_i$ so $\cap_{i=1}^{n} X_i^{-1}(A_i) \in \sigma(\cup_{i=1}^{n} \mathcal{F}_i)$ and thus $\sigma((X_1, \ldots, X_n)) \subseteq \sigma(\cup_{i=1}^{n} \mathcal{F}_i)$.

For the opposite inclusion, we show that $\mathcal{F}_i \subseteq \sigma((X_1, \ldots, X_n))$ for all $i = 1, \ldots, n$, thus $\sigma(\cup_{i=1}^{n} \mathcal{F}_i) \subseteq \sigma((X_1, \ldots, X_n))$. Indeed, the preimage of the rectangle $A_1 \times \Omega \times \cdots \times \Omega$ under $(X_1, \ldots, X_n)$ is $X_1^{-1}(A_1)$ which is in $\sigma((X_1, \ldots, X_n))$. But the sets $\{X_1^{-1}(A_1) : A_1 \in \mathcal{B}(\mathbb{R})\}$ generate $\mathcal{F}_1$ so $\mathcal{F}_1 \subseteq \sigma((X_1, \ldots, X_n))$. To show that the other $\mathcal{F}_i$'s are in $\sigma((X_1, \ldots, X_n))$ is analogous. □

Lemma 271. Suppose $X_{i,j}$, $1 \le i \le n$, $1 \le j \le m_i$ are independent random variables. Suppose the functions $f_i : \mathbb{R}^{m_i} \to \mathbb{R}$ are measurable, and let $Y_i := f_i(X_{i,1}, \ldots, X_{i,m_i})$ for $i = 1, \ldots, , n$. Then the random variables $Y_1, \ldots, Y_n$ are also independent.

Proof. Let $\mathcal{F}_{i,j} := \sigma(X_{i,j})$ and let $\mathcal{G}_i := \sigma(\cup_{j=1}^{m_i} \mathcal{F}_{i,j})$ for $i = 1, \ldots, n$. Then by Corollary 269, $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are independent σ-algebras. It can be shown, using Lemma 270, that $\sigma(Y_i) \subseteq \mathcal{G}_i$ hence $\sigma(Y_1), \ldots, \sigma(Y_n)$ are independent σ-algebras. The latter fact is equivalent to $Y_1, \ldots, Y_n$ being independent random variables. □