Probability (graduate class)
Lecture Notes

Tomasz Tkocz*

These lecture notes were written for the graduate course 21-721 Probability that I taught at Carnegie Mellon University in Spring 2020.

*Carnegie Mellon University; ttkocz@math.cmu.edu
## Contents

1 Probability space 4
   1.1 Definitions .............................................. 4
   1.2 Basic examples ........................................... 5
   1.3 Conditioning ............................................. 9
   1.4 Exercises ................................................ 11

2 Random variables 12
   2.1 Definitions and basic properties ....................... 12
   2.2 \( \pi - \lambda \) systems .................................. 14
   2.3 Properties of distribution functions .................... 14
   2.4 Examples: discrete and continuous random variables .. 16
   2.5 Exercises ................................................ 19

3 Independence 20
   3.1 Definitions ............................................... 20
   3.2 Product measures and independent random variables .... 22
   3.3 Examples .................................................. 23
   3.4 Borel-Cantelli lemmas ..................................... 26
   3.5 Tail events and Kolmogorov’s 0–1 law .................... 27
   3.6 Exercises ................................................ 29

4 Expectation 31
   4.1 Definitions and basic properties ....................... 31
   4.2 Variance and covariance .................................. 33
   4.3 Independence again, via product measures .............. 34
   4.4 Exercises ................................................ 37

5 More on random variables 38
   5.1 Important distributions .................................... 38
   5.2 Gaussian vectors .......................................... 42
   5.3 Sums of independent random variables .................. 43
   5.4 Density .................................................... 44
   5.5 Exercises ................................................ 46

6 Important inequalities and notions of convergence 49
   6.1 Basic probabilistic inequalities ......................... 49
   6.2 \( L_p \)-spaces ............................................. 51
   6.3 Notions of convergence .................................... 56
   6.4 Exercises ................................................ 60
1 Probability space

1.1 Definitions

Let $\Omega$ be a set. A collection $\mathcal{F}$ of its subsets is called a \textit{\(\sigma\)-algebra} (sometimes also \(\sigma\)-field) if

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

(ii) for every $A \in \mathcal{F}$, we have $A^c \in \mathcal{F}$, that is $\mathcal{F}$ is closed under taking complements, 

(iii) for every sets $A_1, A_2, \ldots$ in $\mathcal{F}$, we have $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$, that is $\mathcal{F}$ is closed under taking countable unions.

Note that these imply that $\emptyset \in \mathcal{F}$ and that $\mathcal{F}$ is also closed under taking set difference, countable intersections, etc. For instance, $\mathcal{F} = \{\emptyset, \Omega\}$ is the trivial $\sigma$-algebra and $\mathcal{F} = 2^\Omega$ (all the subsets of $\Omega$) is the largest possible $\sigma$-algebra.

Suppose $\mathcal{F}$ is a $\sigma$-algebra on the set $\Omega$. A function

$$
\mu: \mathcal{F} \to [0, +\infty]
$$

is called a \textit{measure} if

(i) $\mu(\emptyset) = 0$, 

(ii) $\mu$ is countably-additive, that is for every pairwise disjoint sets $A_1, A_2, \ldots$ in $\mathcal{F}$, we have

$$
\mu \left( \bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mu(A_n).
$$

The measure $\mu$ is \textit{finite} if $\mu(\Omega) < \infty$, \textit{\(\sigma\)-finite} if $\Omega$ is a countable union of sets in $\mathcal{F}$ of finite measure. The measure $\mu$ is a \textit{probability measure} if $\mu(\Omega) = 1$.

A \textit{probability space} is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega$ is a set, $\mathcal{F}$ is a $\sigma$-algebra on $\Omega$ and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. The sets in $\mathcal{F}$ are called \textit{events}. The empty set is called an impossible event because $\mathbb{P}(\emptyset) = 0$. Set operations have natural interpretations, for instance for “$A \cap B$”, we say “$A$ and $B$ occur”, for “$A \cup B$”, we say “$A$ or $B$ occurs”, for “$A^c$”, we say “$A$ does not occur”, for “$\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$”, we say “infinitely many of the events $A_k$ occur”, etc.

This definition is a starting point of modern probability theory. It was laid as foundations by Kolmogorov who presented his axiom system for probability theory in 1933.

We record some basic and useful properties of probability measures.

1.1 Theorem. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B, A_1, A_2, \ldots$ be events. Then

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
(ii) if \( A \subset B \), then \( \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A) \) and \( \mathbb{P}(A) \leq \mathbb{P}(B) \),

(iii) \( \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \),

(iv) \( \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathbb{P}(A_1 \cap \cdots \cap A_n) \),

(v) \( \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \),

(vi) if \( A_1, \ldots, A_n \) are pairwise disjoint, then \( \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \).

We omit proofs (which are rather standard). Part (iv) is the so-called inclusion-exclusion formula. Part (v) is the so-called union bound.

We also have the following continuity of measure-type results for monotone events.

1.2 Theorem. Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. Let \( A_1, A_2, \ldots \) be events.

(i) if the events \( A_n \) are increasing, that is \( A_1 \subset A_2 \subset \ldots \), then

\[
\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{n \to \infty} \mathbb{P}(A_n),
\]

(ii) if the events \( A_n \) are decreasing, that is \( A_1 \supset A_2 \supset \ldots \), then

\[
\mathbb{P}\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{n \to \infty} \mathbb{P}(A_n),
\]

Proof. (i) It helps consider the events

\[
B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus A_2, \ldots
\]

which are disjoint. We skip the details. Part (ii) can be obtained from (i) by using the complements. \(\square\)

1.3 Remark. Theorem 1.1 and Theorem 1.2 (i) hold for arbitrary measures (the proofs do not need the assumption \( \mathbb{P}(\Omega) = 1 \) of the measure \( \mathbb{P} \) being probabilistic). Theorem 1.2 (ii) holds for arbitrary measures \( \mathbb{P} \) as long as \( \mathbb{P}(A_k) < \infty \) for some \( k \).

1.2 Basic examples

1.4 Example. Let \( \Omega = \{\omega_1, \omega_2, \ldots\} \) be a countable set and \( \mathcal{F} = 2^\Omega \) (all subsets). Defining a probability measure on \((\Omega, \mathcal{F})\) really amounts to specifying a nonnegative sequence \( p_1, p_2, \ldots \) such that \( \sum_i p_i = 1 \) and defining \( \mathbb{P}(\{\omega_i\}) = p_i \). Then for every subset \( A \) of \( \Omega \),

\[
\mathbb{P}(A) = \sum_{i: \omega_i \in A} p_i.
\]
Conversely, since
\[ 1 = \mathbb{P}(\Omega) = \sum_i \mathbb{P}(\{\omega_i\}), \]
every probability measure is of this form.

**1.5 Example.** Let \( \Omega \) be a finite nonempty set and \( \mathcal{F} = 2^\Omega \). The uniform probability measure on \((\Omega, \mathcal{F})\) (sometimes referred to as *classical*) is defined as
\[
\mathbb{P}(A) = \frac{|A|}{|\Omega|},
\]
for every subset \( A \) of \( \Omega \), where here \(|\cdot|\) denotes cardinality.

Our next two examples will require nontrivial constructions. We wish to define two probability spaces which will be reasonable models of
1) selecting a point uniformly at random on the interval \([0, 1]\)
2) tossing a fair coin infinitely many times.

As much as choosing the ground set \( \Omega \) is fairly natural, say \( \Omega = [0, 1] \) for 1), defining an appropriate \( \sigma \)-algebra and a probability measure on it poses certain challenges. Let us first try to illustrate possible subtleties.

Let \( \Omega = [0, 1] \). If \((\Omega, \mathcal{F}, \mathbb{P})\) is meant to be a probability space modelling selecting a point uniformly at random on \([0, 1]\), for \( 0 \leq a < b \leq 1 \), we should have \( \mathbb{P}((a, b)) = b - a \) (the probability that a point is in the interval \((a, b)\) equal its length), and more generally, \( \mathbb{P} \) should be translation-invariant. Thus \( \mathcal{F} \) should at the very least contain all intervals. Thus let \( \mathcal{F} \) be such a \( \sigma \)-algebra, that is the smallest \( \sigma \)-algebra containing all the intervals in \([0, 1]\); we write
\[
\mathcal{F} = \sigma(\mathcal{I}),
\]
where \( \mathcal{I} \) is the family of all the intervals in \([0, 1]\) and in general
\[
\mathcal{F} = \sigma(\mathcal{A})
\]
denotes the \( \sigma \)-algebra generated by a family \( \mathcal{A} \) of subsets of \( \Omega \) (the smallest \( \sigma \)-algebra containing \( \mathcal{A} \), which makes sense because intersections \( \sigma \)-algebras are still \( \sigma \)-algebras).

**1.6 Example.** As a result, for every \( x \in [0, 1] \), we have
\[
\mathbb{P}\{\{x\}\} = \mathbb{P}\left(\bigcap_{n \geq 1} \left(x - 1/n, x + 1/n\right)\right) = \lim_{n \to \infty} \mathbb{P}\left(\left(x - 1/n, x + 1/n\right)\right) = \lim_{n \to \infty} \frac{2}{n} = 0
\]
(recall Theorem 1.2 (ii)), that is, of course, probability of selecting a fixed point is zero. This however indicates why probability measures are defined to be *only* countably additive as opposed to fully additive, because if the latter was the case, we would have
\[
1 = \mathbb{P}\{[0, 1]\} = \mathbb{P}\left(\bigcup_{x \in [0, 1]} \{x\}\right) = \sum \mathbb{P}\{\{x\}\} = 0,
\]
a contradiction. \( \square \)
Moreover, we cannot just crudely take $\mathcal{F} = 2^\Omega$ because of the following construction of the Vitali set.

1.7 Example. For $x, y \in [0, 1]$, let $x \sim y$ if and only if $x - y \in \mathbb{Q}$. This is an equivalence relation and let $V$ be the set of representatives of its abstract classes. Without loss of generality assume that $0 \not\in V$. Let $x \oplus y$ denote the addition modulo 1, that is $x \oplus y = x + y$ if $x + y \leq 1$ and $x \oplus y = x + y - 1$ if $x + y > 1$. Consider the translations of $V$,

$$V \oplus r = \{ v \oplus r, \, v \in V \}, \quad r \in [0, 1] \cap \mathbb{Q}.$$ 

Note that these sets are pairwise disjoint (because if $v_1 \oplus r_1 = v_2 \oplus r_2$ for some $v_1, v_2 \in V$ and $r_1, r_2 \in [0, 1] \cap \mathbb{Q}$, then $v_1 - v_2 \in \mathbb{Q}$, hence $v_1 = v_2$, thus $r_1 = r_2$). Moreover,

$$\bigcup_{r \in [0, 1] \cap \mathbb{Q}} V \oplus r = [0, 1]$$

(because every point in $[0, 1]$ is in a certain abstract class, hence differs from its representative by a rational). Thus, by countable-additivity

$$1 = \mathbb{P} \left( \bigcup_{r \in [0, 1] \cap \mathbb{Q}} V \oplus r \right) = \sum_{r \in [0, 1] \cap \mathbb{Q}} \mathbb{P} (V \oplus r).$$

If $\mathbb{P}$ is translation-invariant, we have $\mathbb{P} (V \oplus r) = \mathbb{P} (V)$ and then the right hand side is either 0 or $+\infty$, a contradiction. □

Summarising, to model a uniform random point on $[0, 1]$, we take $\Omega = [0, 1]$ and $\mathcal{F}$ to be the $\sigma$-algebra generated by all the intervals. We know how to define $\mathbb{P}$ on the generators. Carathéodory’s theorem is an important abstract tool which allows to extend this definition from the generators to the whole $\sigma$-algebra $\mathcal{F}$, provided that certain conditions are met.

A family $\mathcal{A}$ of subsets of a set $\Omega$ is called an algebra if

(i) $\Omega \in \mathcal{A}$,

(ii) if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$,

(iii) if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

1.8 Theorem (Carathéodory). Let $\Omega$ be a set and let $\mathcal{A}$ be an algebra on $\Omega$. Suppose a function $\mathbb{P} : \mathcal{A} \to [0, +\infty)$ satisfies

(i) $\mathbb{P} (\Omega) = 1$,

(ii) $\mathbb{P}$ is finitely additive, that is for every $A_1, \ldots, A_n \in \mathcal{A}$ which are pairwise disjoint, we have

$$\mathbb{P} \left( \bigcup_{i=1}^n A_i \right) = \sum_{i=1}^n \mathbb{P} (A_i).$$
(iii) for every $A_1, A_2, \ldots \in \mathcal{A}$ with $A_1 \subset A_2 \subset \ldots$ such that $A = \bigcup_{n=1}^{\infty} A_n$ is in $\mathcal{A}$, we have

$$\lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}(A).$$

Then $\mathbb{P}$ can be uniquely extended to a probability measure on the $\sigma$-algebra $\mathcal{F} = \sigma(\mathcal{A})$ generated by $\mathcal{A}$.

1.9 Remark. By considering $B_n = A \setminus A_n$, condition (iii) is equivalent to the following:

if $B_1, B_2, \ldots \in \mathcal{F}_0$ such that $B_1 \supset B_2 \supset \ldots$ with $\bigcap B_n = \emptyset$, then $\mathbb{P}(B_n) \to 0$ as $n \to \infty$.

We defer the proof of Carathéodory’s theorem to Appendix A.

1.10 Example. We are ready to construct a probability space modelling a random point uniform on $[0, 1]$. Let $\Omega = [0, 1]$. Let

$$\mathcal{F}_0 = \{\{a_1, b_1\} \cup \cdots \cup \{a_n, b_n\}, \ n \geq 1, 0 \leq a_1 \leq b_1 \leq \cdots \leq a_n \leq b_n \leq 1\}.$$  

It is easy to check that $\mathcal{F}_0$ is an algebra on $\Omega_0 = (0, 1]$. For a set $F$ in $\mathcal{F}_0$, say

$$F = (a_1, b_1) \cup \cdots \cup (a_n, b_n),$$

we define

$$\mathbb{P}(F) = \sum_{i=1}^{n} (b_i - a_i).$$

Clearly $\mathbb{P}$ satisfies conditions (i) and (ii) of Theorem 1.8. We now verify (iii) by means of Remark 1.9. Suppose $B_1 \supset B_2 \supset \ldots$ are in $\mathcal{F}_0$ with $\bigcap B_n = \emptyset$. If it is not the case that $\mathbb{P}(B_n) \to 0$, there is $\varepsilon > 0$ such that $\mathbb{P}(B_k) > \varepsilon$ for infinitely many $k$, say for simplicity for all $k \geq 1$. We show that $\bigcap B_n \neq \emptyset$. For every $k$, there is a set $C_k$ in $\mathcal{F}_0$ whose closure is a subset of $B_k \cap (0, 1)$ and $\mathbb{P}(B_k \setminus C_k) \leq \varepsilon 2^{-k-1}$. Then for every $n$, we have

$$\mathbb{P} \left( B_n \setminus \bigcap_{k \leq n} C_k \right) = \mathbb{P} \left( \bigcup_{k \leq n} B_n \setminus C_k \right) \leq \mathbb{P} \left( \bigcup_{k \leq n} B_k \setminus C_k \right) \leq \sum_{k \leq n} \mathbb{P}(B_k \setminus C_k) \leq \sum_{k \leq n} \varepsilon 2^{-k-1} < \varepsilon / 2.$$

This and $\mathbb{P}(B_n) > \varepsilon$ together give that $\mathbb{P} \left( \bigcap_{k \leq n} C_k \right) > \varepsilon / 2$. In particular, for every $n$, $\bigcap_{k \leq n} C_k$ is nonempty and consequently $K_n = \bigcap_{k \leq n} \text{cl}(C_k)$ is nonempty. Thus $\{K_n\}_{n=1}^{\infty}$ is a decreasing family ($K_1 \supset K_2 \supset \ldots$) of nonempty compact sets. By Cantor’s intersection theorem, $\bigcap_{n} K_n = \bigcap_{n=1}^{\infty} \text{cl}(C_n)$ is nonempty (recall a simple argument: otherwise $\bigcup_{n} (\text{cl}(C_n))^c$ covers $[0, 1]$ without any finite subcover). Since $\bigcap B_n$ contains $\bigcap_{n} \text{cl}(C_k)$, the argument is finished.

Theorem 1.8 provides a unique extension of $\mathbb{P}$ onto the $\sigma$-algebra generated by $\mathcal{F}_0$. This extension is nothing but Lebesgue measure on $(0, 1]$, denoted Leb. We can trivially extend it onto $[0, 1]$ by assigning $\mathbb{P} \left( \{0\} \right) = 0$. \qed
Given a metric space \((E, \rho)\), the \(\sigma\)-algebra of subsets of \(E\) generated by all open sets in \(E\) is called the **Borel \(\sigma\)-algebra on** \(E\), denoted \(B(E)\). For example, the \(\sigma\)-algebra constructed in the previous example is exactly \(B([0, 1])\).

1.11 Example. We construct a probability space modelling an infinite sequence of tosses of a fair coin. Let \(\Omega = \{(\omega_1, \omega_2, \ldots), \omega_1, \omega_2, \ldots \in \{0, 1\}\}\) be the set of all infinite binary sequences. We can proceed as for the random point on \([0, 1]\): we define an algebra of subsets of \(\Omega\) on which defining a finitely additive measure will be intuitive and easy. Let \(Cyl\) be the family of all cylinders on \(\Omega\), that is sets of the form \(A_{\varepsilon_1, \ldots, \varepsilon_n} = \{\omega \in \Omega, \omega_j = \varepsilon_j, j = 1, \ldots, n\}\). We define the algebra of cylinders, that is the family of all finite unions of cylinders,

\[
\mathcal{F}_0 = \{A_1 \cup \cdots \cup A_k, k \geq 1, A_1, \ldots, A_k \in Cyl\}.
\]

For \(A_{\varepsilon_1, \ldots, \varepsilon_n} \in Cyl\), we set

\[
P(A_{\varepsilon_1, \ldots, \varepsilon_n}) = \frac{1}{2^n}.
\]

It remains to apply Theorem 1.8. Checking (iii) proceeds similarly and eventually boils down to a topological argument (by Tikhonov’s theorem \(\Omega = \{0, 1\} \times \{0, 1\} \times \ldots\) is compact with the standard product topology).

Alternatively, a binary expansion of a random point \(x \in (0, 1]\) gives a random sequence which intuitively does the job, too. Formally, let \(f: \Omega \to [0, 1], f(\omega) = \sum_{i=1}^{\infty} \frac{\omega_i}{2^i}\). We define

\[
\mathcal{F} = \{f^{-1}(B), B \in B([0, 1])\},
\]

which is a \(\sigma\)-algebra,

\[
P(A) = \text{Leb}(f(A)), \quad A \in \mathcal{F},
\]

which is a probability measure (\(f\) is surjective, hence \(f(f^{-1}(B)) = B\) for every \(B\)). Note that for cylinders we have that \(f(A_{\varepsilon_1, \ldots, \varepsilon_n})\) is an interval of length \(\frac{1}{2^n}\). Thus \(P(A_{\varepsilon_1, \ldots, \varepsilon_n}) = \frac{1}{2^n}\) and this construction also fulfils our intuitive basic requirement. We need get back to this example when we discuss independence. \(\square\)

1.3 Conditioning

Given a probability space \((\Omega, \mathcal{F}, P)\) and an event \(B\) of positive probability, \(P(B) > 0\), we can define

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad A \in \mathcal{F}.
\]

It is natural to introduce a \(\sigma\)-algebra of events \(\mathcal{F}_B\) seen by \(B\), that is

\[
\mathcal{F}_B = \{A \cap B, A \in \mathcal{F}\}.
\]

1.12 Theorem. \(P(|B)\) is a probability measure on \(\mathcal{F}\), thus also on \(\mathcal{F}_B\).
The new probability measure $P(\cdot | B)$ is referred to as the **conditional probability** given $B$. Introducing it often times makes computations more intuitive. We have several useful facts.

1.13 **Theorem** (Chain rule). Suppose that $A_1, \ldots, A_n$ are events which satisfy the condition $P(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 | A_1) \cdots P(A_n | A_1 \cap \cdots \cap A_{n-1}).$$

1.14 **Theorem** (Law of total probability). Suppose $\{B_n, n = 1, 2, \ldots\}$ is a finite or countable family of events which partition $\Omega$ and $P(B_n) > 0$ for each $n$. Then for every event $A$, we have

$$P(A) = \sum_n P(A | B_n) P(B_n).$$

1.15 **Theorem** (Bayes’ formula). Suppose $\{B_n, n = 1, 2, \ldots\}$ is a finite or countable family of events which partition $\Omega$ and $P(B_n) > 0$ for each $n$. Then for every event $A$ of positive probability and every $k$, we have

$$P(B_k | A) = \frac{P(A | B_k) P(B_k)}{\sum_n P(A | B_n) P(B_n)}.$$

We leave all the proofs as exercise to the dedicated reader.
1.4 Exercises

1. If $A$ and $B$ are events, then $P(A \cap B) \geq P(A) - P(B^c)$.

2. If $A_1, \ldots, A_n$ are events, then we have
   a) $P(\bigcup_{i=1}^n A_i) \leq \sum_{k=1}^m (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k})$ for $m$ odd,
   b) $P(\bigcup_{i=1}^n A_i) \geq \sum_{k=1}^m (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k})$ for $m$ even.
   These are called Bonferroni inequalities.

3. There are $n$ invitation cards with the names of $n$ different people and $n$ envelopes with their names. We put the cards at random into the envelopes, one card per envelope. What is the chance that not a single invitation landed in the correct envelope? What is the limit of this probability as $n$ goes to infinity?

4. Describe all $\sigma$-algebras on a countable set.

5. Is there an infinite $\sigma$-algebra which is countable?

6. Show that the number of $\sigma$-algebras on the $n$-element set equals $\frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$.

7. Prove Theorems 1.12 – 1.15.
2 Random variables

2.1 Definitions and basic properties

Central objects of study in probability theory are random variables. They are simply measurable functions. To put it formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a random variable if for every Borel set $B$ in $\mathbb{R}$, $B \in \mathcal{B}(\mathbb{R})$, we have $X^{-1}(B) \in \mathcal{F}$. In other words, $X$ is a measurable function on $(\Omega, \mathcal{F}, \mathbb{P})$. An $\mathbb{R}^n$-valued random variable, that is a measurable function $X : \Omega \to \mathbb{R}^n$ is called a random vector.

2.1 Example. Let $A$ be an event. We define

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

This is a random variable called the indicator random variable of the event $A$.

2.2 Example. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$ and $\mathbb{P} = \text{Leb}$. Define $X : \Omega \to [0,1]$ as $X(\omega) = \omega$. This is a random variable which intuitively is uniform on $[0,1]$. We will make this precise soon. On the other hand, if $V$ is the Vitali set from Example 1.7, then $X = 1_V$ is not a random variable because $X^{-1}(\{1\}) = V \notin \mathcal{F}$.

We record several very basic facts. One piece of notation: we often write $\{X \leq t\}$, or $\{X \in B\}$, etc. meaning $\{\omega \in \Omega, \ X(\omega) \leq t\} = X^{-1}((-\infty,t])$, or $\{\omega \in \Omega, \ X(\omega) \in B\} = X^{-1}(B)$,etc. Moreover, $\{X \in A, X \in B\}$ means $\{X \in A\} \cap \{X \in B\}$.

2.3 Theorem. If $X : \Omega \to \mathbb{R}$ satisfies: for every $t \in \mathbb{R}$,

$$\{X \leq t\} \in \mathcal{F},$$

then $X$ is a random variable.

Proof. Consider the family $\{A \subset \mathbb{R}, \ X^{-1}(A) \in \mathcal{F}\}$. It is not difficult to check that this is a $\sigma$-algebra. By the assumption, it contains the intervals $(-\infty, t]$, $t \in \mathbb{R}$, which generate $\mathcal{B}(\mathbb{R})$.

2.4 Theorem. If $X,Y$ are random variables (defined on the same probability space), then $X + Y$ and $XY$ are random variables.

Proof. We use Theorem 2.3. Note that

$$\{X + Y > t\} = \bigcup_{q \in \mathbb{Q}} \{X > q, Y > t - q\}$$

and the right hand side is in $\mathcal{F}$ as a countable union of events. Thus $X + Y$ is a random variable. Moreover, for $t \geq 0$,

$$\{X^2 \leq t\} = \{-\sqrt{t} \leq X \leq \sqrt{t}\} = \{X \leq \sqrt{t}\} \setminus \{X < -\sqrt{t}\} \in \mathcal{F}$$
so $X^2$ and $Y^2$ are also random variables. Thus

$$XY = \frac{1}{2}((X + Y)^2 - X^2 - Y^2)$$

is a random variable.

2.5 Theorem. If $X_1, X_2, \ldots$ are random variables (defined on the same probability space), then $\inf_n X_n$, $\liminf_n X_n$, $\lim_n X_n$ (if exists, understood pointwise) are random variables.

Proof. For instance \( \{\inf_n X_n \geq t\} = \bigcap_n \{X_n \geq t\} \) justifies that $\inf_n X_n$ is a random variable. We leave the rest as an exercise.

2.6 Theorem. Let $X$ be a random variable. If $f: \mathbb{R} \to \mathbb{R}$ is a (Borel) measurable function, that is $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for every $B \in \mathcal{B}(\mathbb{R})$, then $f(X)$ is a random variable.

Proof. We have $(f(X))^{-1}(B) = X^{-1}(f^{-1}(B))$.

2.7 Example. If $X$ is a random variable, then $|X|^p$, $e^X$, etc. are random variables.

Given a random variable $X$, we define the \textbf{\textit{σ-algebra generated by $X$, denoted $\sigma(X)$ as the smallest $σ$-algebra with respect to which $X$ is measurable}}, that is

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

(the family on the right is a $σ$-algebra). Similarly, given a collection of random variables $\{X_i\}_{i \in I}$ we define its $σ$-algebra as the smallest $σ$-algebra with respect to which every $X_i$ is measurable, that is

$$\sigma(X_i, i \in I) = \sigma(X_i^{-1}(B), B \in \mathcal{B}(\mathbb{R}), i \in I).$$

Let $X$ be a random variable. The \textbf{\textit{law of $X$, denoted $\mu_X$ is the following probability measure on $\langle \mathbb{R}, \mathcal{B}(\mathbb{R}) \rangle$}},

$$\mu_X(B) = \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}).$$

2.8 Example. Let $X$ be a constant random variable a.s., that is $\mathbb{P}(X = a) = 1$ for some $a \in \mathbb{R}$. Its law $\mu_X$ is a very simple measure on $\mathbb{R}$,

$$\mu_X(A) = \begin{cases} 1, & \text{if } a \in A, \\ 0, & \text{if } a \notin A. \end{cases}$$

This measure on $\mathbb{R}$ is called the \textbf{\textit{Dirac delta}} at $a$, denoted $\delta_a$.

The \textbf{\textit{cumulative distribution function of $X$, denoted $F_X$}}, (distribution function or CDF in short) is the following function $F_X: \mathbb{R} \to [0, 1]$,

$$F_X(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$
It is rather clear that for some two random variables $X$ and $Y$, $\mu_X = \mu_Y$ does not imply that $X = Y$ (the random variables may even be defined on different probability spaces). We say that $X$ and $Y$ have the same distribution (law) if $\mu_X = \mu_Y$. Is it clear that $F_X = F_Y$ implies that $X$ and $Y$ have the same distribution? In other words, do CDFs determine distribution? To answer this and many other similar questions, it is convenient to use an abstract tool from measure theory – Dynkin’s theorem on $\pi - \lambda$ systems.

2.2 $\pi - \lambda$ systems

A family $\mathcal{A}$ of subsets of a set $\Omega$ is a $\pi$-system if it is closed under finite intersections, that is for every $A, B \in \mathcal{A}$, we have $A \cap B \in \mathcal{A}$.

A family $\mathcal{L}$ of subsets of a set $\Omega$ is a $\lambda$-system if

(i) $\Omega \in \mathcal{L}$,
(ii) if $A, B \in \mathcal{L}$ and $A \subset B$, then $B \setminus A \in \mathcal{L}$,
(iii) for every $A_1, A_2, \ldots \in \mathcal{L}$ such that $A_1 \subset A_2 \subset \ldots$, we have $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

For example, the family of intervals $\{(−\infty, t], t \in \mathbb{R}\}$ is a $\pi$-system. The importance of this example is that this family generates $\mathcal{B}(\mathbb{R})$.

Note that if a family is a $\pi$-system and a $\lambda$-system, then it is a $\sigma$-algebra.

A fundamental and useful result is the following theorem (see Appendix B for the proof).

2.9 Theorem (Dynkin). If a $\lambda$-system $\mathcal{L}$ contains a $\pi$-system $\mathcal{A}$, then $\mathcal{L}$ contains $\sigma(\mathcal{A})$.

2.3 Properties of distribution functions

Equipped with Dynkin’s theorem, we are able to show that distribution functions indeed determine the distribution, which reverses the trivial implication that if $\mu_X = \mu_Y$, then $F_X = F_Y$.

2.10 Theorem. Let $X$ and $Y$ be random variables (possibly defined on different probability spaces). If $F_X = F_Y$, then $\mu_X = \mu_Y$.

Proof. Let $\mathcal{A} = \{(−\infty, t], t \in \mathbb{R}\}$. This is a $\pi$-system and $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$. Consider

$$\mathcal{L} = \{A \in \mathcal{B}(\mathbb{R}), \mu_X(A) = \mu_Y(A)\}.$$

This is a $\lambda$-system (which easily follows from properties of probability measures). The assumption $F_X = F_Y$ gives $\mathcal{L} \supset \mathcal{A}$. Thus, by Theorem 2.9, we get $\mathcal{L} \supset \sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$. By the definition of $\mathcal{L}$, this gives $\mu_X = \mu_Y$. \qed
2.11 Remark. The same proof gives the following: if for some two probability measures \( \mu, \nu \) defined on the same space, we have \( \mu = \nu \) on a \( \pi \)-system generating a \( \sigma \)-algebra \( \mathcal{F} \), then \( \mu = \nu \) on \( \mathcal{F} \).

We list 3 basic properties of distribution functions.

2.12 Theorem. Let \( X \) be a random variable. Then its distribution function \( F_X \) satisfies

(i) \( F_X \) is nondecreasing, that is for every \( s \leq t \), \( F_X(s) \leq F_X(t) \),

(ii) \( \lim_{t \to -\infty} F_X(t) = 0 \) and \( \lim_{t \to +\infty} F_X(t) = 1 \),

(iii) \( F_X \) is right-continuous, that is for every \( t \in \mathbb{R} \), \( \lim_{s \to t^+} F_X(s) = F_X(t) \).

Proof. Part (i) follows from the inclusion \( \{X \leq s\} \subset \{X \leq t\} \) if \( s \leq t \). Alternatively, \( 0 \leq P(X \in (s, t]) = P(X \leq t) - P(X \leq s) = F_X(t) - F_X(s) \).

Part (ii), (iii) follow from the continuity of probability measures (Theorem 1.2).

These properties in fact characterise distribution functions.

2.13 Theorem. If a function \( F: \mathbb{R} \to [0, 1] \) satisfies (i)-(iii) from Theorem 2.12, then \( F = F_X \) for some random variable \( X \).

Proof. Let \( \Omega = [0, 1] \), \( \mathcal{F} = \mathcal{B}([0, 1]) \) and \( \mathbb{P} = \text{Leb} \). The idea is to define \( X \) as the inverse of \( F \). Formally, we set

\[
X(\omega) = \inf\{y, F(y) \geq \omega\}, \quad \omega \in [0, 1].
\]

By the definition of infimum and (i)-(iii), \( X(\omega) \leq t \) if and only if \( \omega \leq F(t) \) (check!). Thus

\[
F_X(t) = \mathbb{P}(X \leq t) = \text{Leb}\{\omega \in [0, 1], \omega \leq F(t)\} = F(t).
\]

2.14 Remark. There is another construction, sometimes called canonical, based on Carathéodory’s theorem. We set \( \Omega = \mathbb{R}, \mathcal{F} = \mathcal{B}(\mathbb{R}) \), define \( \mathbb{P}((-\infty, t]) = F(t) \) and then extend \( \mathbb{P} \). With such \( \mathbb{P} \), the desired random variable is the canonical one, \( X(x) = x, \quad x \in \mathbb{R} \).

For a random vector \( X = (X_1, \ldots, X_n) \) in \( \mathbb{R}^n \), the cumulative distribution function of \( X \) is the function \( F_X: \mathbb{R}^n \to [0, 1], \)

\[
F_X(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_n).
\]

As before, for random vectors \( X, Y \) in \( \mathbb{R}^n \), \( F_X = F_Y \) implies that \( \mu_X = \mu_Y \). The characterising properties are almost the same – the monotonicity statement is strengthened.
2.15 Theorem. Let $X$ be a random vector in $\mathbb{R}^n$. Then its distribution function $F_X$ satisfies

(i) $F_X$ is nondecreasing, that is for every $s, t \in \mathbb{R}^n$ with $s_i \leq t_i$, $i \leq n$, we have $F_X(s) \leq F_X(t)$. Moreover,

$$\sum_{\varepsilon \in \{0,1\}^n} (-1)^{\sum_{k=1}^n \varepsilon_k} F_X(\varepsilon_1 s_1 + (1 - \varepsilon_1)t_1, \ldots, \varepsilon_n s_n + (1 - \varepsilon_n)t_n) \geq 0,$$

(ii) $F_X(t_1^{(m)}, \ldots, t_n^{(m)}) \xrightarrow{m \to \infty} 0$ provided that $\inf_{k \leq n} t_k^{(m)} \xrightarrow{m \to \infty} -\infty$,

(iii) $F_X(t_1^{(m)}, \ldots, t_n^{(m)}) \xrightarrow{m \to \infty} 1$ provided that $\inf_{k \leq n} t_k^{(m)} \xrightarrow{m \to \infty} +\infty$,

(iv) $F_X$ is right-continuous.

Proof. We only show (i) as the rest is proved in much the same way as in one dimension. The inequality is nothing but the statement that the probability of $X$ being in the box $\prod_{k=1}^n (s_k, t_k]$ is nonnegative (cf. the proof of Theorem 2.13 (i)). To see this, let $A = \bigcap_{k=1}^n \{X_k \leq t_k\}$ and $B = \bigcup_{k=1}^n \{X_k \leq s_k\}$. Then

$$0 \leq P\left( X \in \prod_{k=1}^n (s_k, t_k] \right) = P( A \cap B^c ) = P( A ) - P( A \cap B ).$$

Note that $P(A) = F_X(t)$ and $P(A \cap B) = \bigcup_{k=1}^n \{X_k \leq s_k\} \cap B$. Applying the inclusion-exclusion formula finishes the proof.

Again, these properties characterise distribution functions of random vectors. The proof follows a canonical construction sketched in Remark 2.14. We leave the details as an exercise.

2.16 Theorem. If a function $F: \mathbb{R}^n \to [0, 1]$ satisfies (i)-(iv) from Theorem 2.15, then $F = F_X$ for some random vector $X$ in $\mathbb{R}^n$.

We end with a simple remark which follows from the right-continuity.

2.17 Remark. For a random variable $X$ and $a \in \mathbb{R}$, we have

$$P(X = a) = P(\{X \leq a\} \setminus \{X < a\}) = P(X \leq a) - P(X < a) = F_X(a) - F_X(a-).$$

Now, $P(X = a) > 0$ if and only if $F_X$ is discontinuous at $a$ (has a jump) and the value of the jump is precisely $P(X = a)$. In this case, we say that $X$ has an atom at $a$.

2.4 Examples: discrete and continuous random variables

2.18 Example. We say that a random variable $X$ is discrete if there is a countable subset $A$ of $\mathbb{R}$ such that $P(X \in A) = 1$. Say $A = \{a_1, a_2, \ldots\}$ and denote $p_k = P(X = a_k)$.
We can assume that the $p_k$ are all positive (otherwise, we just do not list $a_k$ in $A$). We have $\sum_k p_k = 1$. The $a_k$ are then the atoms of $X$. The law of $X$ is a mixture of Dirac deltas at the atoms,

$$
\mu_X = \sum p_k \delta_{a_k}.
$$

The CDF of $X$ is a piecewise constant function with jumps at the atoms with the values being the $p_k$.

2.19 Example. We say that a random variable $X$ is continuous if there is an integrable function $f: \mathbb{R} \to \mathbb{R}$ such that

$$
\mu_X(A) = \int_A f, \quad A \in \mathcal{B}(\mathbb{R}).
$$

Then $f$ is called the density of $X$ (note it is not unique – we can modify $f$ on a set of Lebesgue measure zero without changing the above). In particular,

$$
F_X(t) = \mu_X((-\infty, t]) = \int_{-\infty}^t f(x)dx
$$

and necessarily $F_X$ is continuous. We collect basic characterising properties of density functions.

2.20 Theorem. Let $X$ be a continuous random variable and let $f$ be its density function. Then

(i) $\int f = 1$

(ii) $f \geq 0$ a.e.

(iii) $f$ is determined by $X$ uniquely up to sets of measure 0.

Proof. Plainly, $\int f = \mu_X(\mathbb{R}) = 1$, so we have (i). To see (ii), let $A_n = \{f < -1/n\}$ and $A = \{f < 0\} = \bigcup A_n$. We have

$$
0 \leq \mu_X(A_n) = \int_{A_n} f \leq -\frac{1}{n} \text{Leb}(A_n),
$$

so $\text{Leb}(A_n) = 0$ and thus $\text{Leb}(A) = 0$. The proof of (iii) is similar.

2.21 Theorem. Suppose a function $f: \mathbb{R} \to \mathbb{R}$ satisfies properties (i)-(ii) of Theorem 2.20. Then there is a continuous random variable $X$ with density $f$.

Proof. We set $F(x) = \int_{-\infty}^x f$, $x \in \mathbb{R}$ and use Theorem 2.13.

Of course, it is easy to give examples of random variables which are neither discrete nor continuous, say $F(x) = \frac{x}{2}1_{[0,1)}(x) + 1_{[1,\infty)}(x)$ is a distribution function of such a random variable (it is not continuous because $F$ is not continuous and it is not discrete because $F$ is not piecewise constant). We finish off this chapter with an interesting strong example of this sort.
2.22 Example. Let $F : [0, 1] \to [0, 1]$ be the Cantor’s devil’s staircase function (a continuous nondecreasing function which is piecewise constant outside the Cantor set $C \subset [0, 1]$). Extend $F$ on $\mathbb{R}$ by simply putting $0$ on $(-\infty, 0]$ and $1$ on $[1, +\infty)$. Then $F$ is a distribution function of some random variable. It is not discrete because $F$ is continuous and if it was continuous, we would have

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

for some integrable function $f$, but since $f(x) = F'(x) = 0$ for $x \notin C$ ($F$ is constant on $C^c$), we would also have

$$1 = \int_{\mathbb{R}} f = \int_{C} f + \int_{C^c} = 0$$

(the first integral vanishes because $C$ is of measure 0 and the second integral vanishes because as we just saw $f$ vanishes on $C^c$), a contradiction. What is this random variable?
2.5 Exercises

1. Give an example of two different random variables $X$ and $Y$ with $\mu_X = \mu_Y$.

2. Fill out the details in the proof of Theorem 2.13.

3. Prove Theorem 2.16.

4. Show that every random variable has at most countably many atoms.

5. Suppose that a random vector $(X, Y)$ is such that both $X$ and $Y$ are continuous random variables. Does the random vector $(X, Y)$ have to be continuous?

6. Is there a random vector $(X, Y, Z)$ in $\mathbb{R}^3$ such that $aX + bY + cZ$ is a uniform random variable on $[-1, 1]$ for every reals $a, b, c$ with $a^2 + b^2 + c^2 = 1$?

   **Hint:** Archimedes’ Hat-Box Theorem.

7. Let $X$ be a random variable uniform on $[0, 2]$. Find the distribution function of random variables $Y = \max\{1, X\}$, $Z = \min\{X, X^2\}$.

8. Give an example of an uncountable family of random variables $\{X_i\}_{i \in I}$ such that $\sup_{i \in I} X_i$ is not a random variable.

9. Is there a random variable such that the set of the discontinuity points of its distribution function is dense in $\mathbb{R}$?
3 Independence

Recall that two events $A, B$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$, which is equivalent to $\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c) \mathbb{P}(B)$. A good way to generalise this is via $\sigma$-algebras.

3.1 Definitions

For an event $A \in \mathcal{F}$, we define the $\sigma$-algebra generated by it as

$$\sigma(A) = \{\emptyset, \Omega, A, A^c\},$$

that is $\sigma(A) = \sigma(1_A)$, the $\sigma$-algebra generated by the indicator random variable $1_A$. The crucial general definition of independence is as follows.

A family $\{\mathcal{F}_i\}_{i \in I}$ of collections of subsets of $\Omega$ (typically $\sigma$-algebras, $\pi$-systems, etc.) with each $\mathcal{F}_i$ being a subset of $\mathcal{F}$ is called independent if for every $n, i_1, \ldots, i_n \in I$ and every $A_1 \in \mathcal{F}_{i_1}, \ldots, A_n \in \mathcal{F}_{i_n}$, we have

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

A family of events $\{A_i\}_{i \in I}$ is independent (or simply the events $A_i, i \in I$, are independent) if the family of the $\sigma$-algebras generated by them, $\{\sigma(A_i)\}_{i \in I}$ is independent. A family of random variables $\{X_i\}_{i \in I}$ is independent (or simply the random variables $X_i, i \in I$, are independent) if the family of the $\sigma$-algebras generated by them, $\{\sigma(X_i)\}_{i \in I}$ is independent. Note that since $\sigma(A) = \sigma(1_A)$, the events $A_i$ are independent if and only if the random variables $1_{A_i}$ are independent.

As is stated now, to check the independence of say 3 events $\{A_1, A_2, A_3\}$, we have to verify $4^3$ identities of the form $\mathbb{P}(B_1 \cap B_2 \cap B_3) = \mathbb{P}(B_1) \mathbb{P}(B_2) \mathbb{P}(B_3)$, where each $B_i$ is one of the sets $\emptyset, \Omega, A_i, A_i^c$. Of course, many of these identities are either trivial or follow from the other. It turns out that $\pi$-systems can help and we have the following useful general lemma.

3.1 Lemma. Let $\{A_i\}_{i \in I}$ be a family of $\pi$-systems. Then the family $\{\sigma(A_i)\}_{i \in I}$ is independent if and only if the family $\{A_i\}_{i \in I}$ is independent.

Proof. Since the definition of independence involves only finite sub-families, we can assume that $I = \{1, \ldots, n\}$. One implication is clear, so we assume that the $\pi$-systems are independent and want to deduce the independence of the $\sigma$-algebras generated by them. To this end, we shall use Dynkin’s theorem. We define the class

$$\mathcal{L}_1 = \{B_1 \in \mathcal{F} : \forall A_2 \in A_2, \ldots, A_n \in A_n$$

$$\mathbb{P}(B_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(B_1) \mathbb{P}(A_2) \cdots \mathbb{P}(A_n)\}$$
By the assumption \( \mathcal{L}_1 \) contains \( A_1 \). By properties of probability measures, \( \mathcal{L}_1 \) is a \( \lambda \)-system. Hence, by Dynkin’s theorem (Theorem 2.9), \( \mathcal{L}_1 \) contains \( \sigma(A_1) \). It remains to inductively repeat the same argument: suppose we know for some \( k < n \) that
\[
P (B_1 \cap \cdots \cap B_k \cap A_{k+1} \cap \cdots \cap A_n) = \prod_{i=1}^{n} P (A_i) \quad (3.1)
\]
for every \( B_i \in \sigma(A_i), i \leq k \) and \( A_j \in A_j, j > k \). Fix some \( B_i \in \sigma(A_i), i \leq k \). As above, considering \( \mathcal{L}_{k+1} = \{ B_{k+1} \in \mathcal{F} : \forall A_{k+2} \in A_{k+2}, \ldots, A_n \in A_n \}
\)
\[
P (B_1 \cap \cdots \cap B_k \cap B_{k+1} \cap A_{k+2} \cap \cdots \cap A_n) = \prod_{i=1}^{n} P (A_i) \]
shows that (3.1) holds for \( k + 1 \). Thus this holds for \( k = n \).

We note two useful results about *packaging* independence.

3.2 Theorem. Let \( \{F_i\}_{i \in I} \) be a family of independent \( \sigma \)-algebras. Suppose the index set \( I \) is partitioned into nonempty sets \( \{I_j, j \in J\} \). Then the \( \sigma \)-algebras
\[
\mathcal{G}_j = \sigma\left(\{F_i, i \in I_j\}\right), \quad j \in J
\]
are independent.

Proof. For each \( j \in J \), define \( A_j \) to be the \( \pi \)-system of all finite intersections of the form \( B_{i_1} \cap \cdots \cap B_{i_m} \), where \( i_1, \ldots, i_m \in I_j \) and \( B_{i_k} \in F_{i_k}, k = 1, \ldots, m \). We have \( \sigma(A_j) = \mathcal{G}_j \). By the assumption, it follows that the families \( A_j, j \in J \) are independent (check!), so by Lemma 3.1, the \( \mathcal{G}_j \) are independent.

3.3 Theorem. Suppose
\[
X_{1,1}, \ldots, X_{1,n_1}, \\
X_{2,1}, \ldots, X_{2,n_2}, \\
\vdots, \quad \vdots, \quad \vdots, \\
X_{k,1}, \ldots, X_{k,n_k}
\]
are independent random variables and
\[
f_1 : \mathbb{R}^{n_1} \to \mathbb{R}, \\
\vdots \\
f_k : \mathbb{R}^{n_k} \to \mathbb{R}
\]
are measurable functions. Then the random variables
\[
Y_1 = f_1(X_{1,1}, \ldots, X_{1,n_1}), \\
\vdots \\
Y_k = f_k(X_{k,1}, \ldots, X_{k,n_k})
\]
are independent.
Proof. By Theorem 3.2, the \( \sigma \)-algebras \( G_i = \sigma(\sigma(X_{i,1}), \ldots, \sigma(X_{i,n_i})), i \leq k \), are independent. The result follows because \( \{Y_i \leq t\} \in G_i \), so \( \sigma(Y_i) \subset G_i \). \( \square \)

3.2 Product measures and independent random variables

Given two probability measures \( \mu, \nu \) on \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\), recall that their product, denoted \( \mu \otimes \nu \), is the unique measure on \((\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))\) such that

\[
(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B) \quad \text{for all } A, B \in \mathcal{B}(\mathbb{R})
\]

(see Appendix C for the details). Exploiting Lemma 3.1, we derive convenient and important equivalent conditions for independence of random variables. For simplicity we state it just for two random variables, but of course it can be easily generalised to arbitrary many of them.

3.4 Theorem. The following are equivalent

(i) random variables \( X, Y \) are independent,

(ii) \( F_{(X,Y)}(s,t) = F_X(s)F_Y(t), \) for all \( s, t \in \mathbb{R}, \)

(iii) \( \mu(X,Y) = \mu_X \otimes \mu_Y. \)

Proof. (i)⇒(ii) follows from the definition since \( \{X \leq s\} \in \sigma(X) \) and \( \{Y \leq y\} \in \sigma(Y). \)

(ii)⇒(i) follows from Lemma 3.1 (\( \{X \leq s\}_{s \in \mathbb{R}} \) is a \( \pi \)-system generating \( \sigma(X) \)).

(i)⇒(iii) from the definition, \( \mu(X,Y) = \mu \otimes \nu \) on the \( \pi \)-system of the product sets \( A \times B, \)
\( A, B \in \mathcal{B}(\mathbb{R}) \) which generate \( \mathcal{B}(\mathbb{R}^2) \), thus, by Remark 2.11, \( \mu(X,Y) = \mu \otimes \nu \) on \( \mathcal{B}(\mathbb{R}^2) \).

(iii)⇒(ii) follows by applying (iii) to \( A = \{X \leq s\}, B = \{Y \leq t\}. \)

For continuous random variables, we have another convenient criterion in terms of densities.

3.5 Theorem. If \( X_1, \ldots, X_n \) are continuous random variables with densities \( f_1, \ldots, f_n \) respectively, then they are independent if and only if the random vector \((X_1, \ldots, X_n)\) is continuous with density

\[
f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n).
\]
Proof. Suppose we have independence. Then, by Theorem 3.4 (ii) and Fubini’s theorem, for every Borel sets $A_i$ in $\mathbb{R}$, we have

$$P((X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

$$= \prod_{i=1}^{n} \int_{A_i} f_i$$

$$= \int_{A_1 \times \cdots \times A_n} f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n.$$ 

This means that $f_1(x_1) \cdots f_n(x_n)$ is the density of $(X_1, \ldots, X_n)$. To see the opposite implication, simply backtrack the above equalities. \hfill \Box

We leave it as an exercise to prove a discrete analogue.

3.6 Theorem. If $X_1, \ldots, X_n$ are discrete random variables with the atoms in some sets $A_1, \ldots, A_n$ respectively, then they are independent if and only if for every sequence $a_1, \ldots, a_n$ with $a_i \in A_i$ for each $i$, we have

$$P(X_1 = a_1, \ldots, X_n = a_n) = P(X_1 = a_1) \cdots P(X_n = a_n).$$

3.3 Examples

3.7 Example. Let $\Omega = \{0, 1\}^n$, $\mathcal{F} = 2^\Omega$, $P$ is uniform, that is $P(\{\omega\}) = 2^{-n}$ for every $\omega \in \Omega$ (the probability space of $n$ tosses of a fair coin). For $k = 1, \ldots, n$ consider the events

$$A_k = \{\omega \in \Omega, \omega_k = 0\} \quad (k\text{th toss is 0}).$$

We claim that the events $A_1, \ldots, A_n$ are independent. For $1 \leq i_1 < \ldots < i_k \leq n$, we have

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(\{\omega \in \Omega: \omega_{i_1} = \ldots = \omega_{i_k} = 0\}) = \frac{2^{n-k}}{2^n} = 2^{-k} = \prod_{j=1}^{k} P(A_{i_j}).$$

Lemma 3.1 finishes the argument.

3.8 Example. Let $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = 2^\Omega$, $P$ is uniform, that is $P(\{\omega\}) = 1/4$ for every $\omega \in \Omega$ (4 sided fair die). Let $A_i = \{1, i+1\}$, $i = 1, 2, 3$. Then

$$P(A_i) = \frac{1}{2}, \quad i = 1, 2, 3,$$

$$P(A_i \cap A_j) = P(\{1\}) = \frac{1}{4} = P(A_i)P(A_j), \quad i \neq j$$

$$P(A_1 \cap A_2 \cap A_3) = P(\{1\}) = \frac{1}{4} \neq P(A_1)P(A_2)P(A_3),$$

so the events $A_1, A_2, A_3$ are pairwise independent but not independent.
3.9 Example. Let $\Omega = [0,1]^2$, $\mathcal{F} = \mathcal{B}([0,1]^2)$, $\mathbb{P} = \text{Leb}$ (a random point uniformly selected from the unit square $[0,1]^2$). Let $A = B = \{(x,y) \in [0,1]^2, \ x > y\}$ and $C = \{(x,y) \in [0,1]^2, \ x < \frac{1}{2}\}$. Then

$$
\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \mathbb{P}(B) \mathbb{P}(C),
$$

but

$$
\mathbb{P}(A \cap B) \neq \mathbb{P}(A) \mathbb{P}(B), \quad \mathbb{P}(A \cap C) \neq \mathbb{P}(A) \mathbb{P}(C), \quad \mathbb{P}(B \cap C) \neq \mathbb{P}(B) \mathbb{P}(C).
$$

3.10 Example. If for events $A_1, \ldots, A_n$ and every $\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\}$, we have

$$
\mathbb{P}(A_1^{\varepsilon_1} \cap \cdots \cap A_n^{\varepsilon_n}) = \mathbb{P}(A_1^{\varepsilon_1}) \cdots \mathbb{P}(A_n^{\varepsilon_n}),
$$

where $A^\varepsilon = A$ if $\varepsilon = 0$ and $A^\varepsilon = \overline{A}$ if $\varepsilon = 1$, then the family $\{A_i\}_{i \leq n}$ is independent. A simple explanation relies on algebraic manipulations like this one

$$
\mathbb{P}(A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) + \mathbb{P}(A_1^c \cap A_2 \cap \cdots \cap A_n).
$$

We skip the details.

3.11 Example. Let $\Omega = (0,1]$ and $\mathbb{P} = \text{Leb}$ (a random point uniformly selected from the unit interval $(0,1]$). For every point $x \in (0,1]$, we write its binary expansion,

$$
x = \sum_{n=1}^{\infty} d_n(x) \frac{1}{2^n},
$$

where $d_n(x) \in \{0,1\}$ is the $n$th digit of $x$. For uniqueness, say we always write the expansion that has infinitely many $1$’s, e.g. $\frac{1}{2} = 0.0111\ldots$. Consider the events

$$
A_k = \{x \in (0,1], \ d_k(x) = 0\}, \quad k = 1, 2, \ldots
$$

Claim. $\mathbb{P}(A_k) = \frac{1}{2}$ and $\{A_k\}_{k \geq 1}$ are independent.

In other words, the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is also a good model for infinitely many tosses of a fair coin with the events $A_k$ being “$k$th toss is heads”. To prove that $\mathbb{P}(A_k) = \frac{1}{2}$, just note that $A_k$ is the union of $2^k$ intervals $(\sum_{i=1}^{k} \varepsilon_i 2^{-i}, \sum_{i=1}^{k} \varepsilon_i 2^{-i} + 2^{-k}]$, $\varepsilon_1, \ldots, \varepsilon_k \in \{0,1\}$, each of length $2^{-k}$. To prove the independence, note that for fixed $\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\},$

$$
\mathbb{P}(A_1^{\varepsilon_1} \cap \cdots \cap A_n^{\varepsilon_n}) = \text{Leb}\{x \in (0,1], \ d_1(x) = \varepsilon_1, \ldots, d_n(x) = \varepsilon_n\}
$$

$$
= \text{Leb}\left(\sum_{i=1}^{n} \frac{\varepsilon_i}{2^n} \sum_{i=1}^{n} \varepsilon_i \frac{1}{2^n}\right)
$$

$$
= \frac{1}{2^n}
$$

and use Example 3.10.
To put this important example a bit differently, we have constructed the sequence \( d_1, d_2, \ldots \) of independent, identically distributed random variables (i.i.d. for short), each one having equal probability of taking the value 0 and 1 (\( d_k \) tells us the outcome of the \( k \)th toss).

3.12 Example. We construct a sequence \( X_1, X_2, \ldots \) of i.i.d. random variables uniform on [0, 1]. Let as before \( \Omega = (0, 1], \mathcal{F} = \mathcal{B}((0, 1]), \mathbb{P} = \text{Leb} \). For every \( \omega \in \Omega \) we write as before its (unique) binary expansion

\[ \omega = \sum_{i=1}^{\infty} \frac{\omega_i}{2^i} = 0.\omega_1\omega_2 \ldots, \]

where \( \omega_1, \omega_2, \ldots \in \{0, 1\} \) are the consecutive digits of \( \omega \). We define new functions

\[ X_1(\omega) = 0.\omega_1\omega_3\omega_7\omega_{10} \ldots \]
\[ X_2(\omega) = 0.\omega_2\omega_5\omega_9 \ldots \]
\[ X_3(\omega) = 0.\omega_4\omega_8 \ldots \]
\[ X_4(\omega) = 0.\omega_7 \ldots \]

(we put the consecutive indices on the diagonals: 1, then 2, 3, then 4, 5, 6 then 7, 8, 9, 10 and so on). Intuitively

1) \( X_1, X_2, \ldots \) are independent random variables
2) each \( X_i \) is uniform on [0, 1].

The intuition for 1) is that the rows are built on disjoint sequences of the \( \omega_i \). The formal proof follows instantly from Theorem 3.2 about packaging independence.

The intuition for 2) is that each \( \omega_i \) is just a random digit. The formal proof follows from the observation that for every \( k \geq 1 \) and \( j = 0, 1, \ldots, 2^k - 1 \), we have

\[ \mathbb{P} \left( \frac{j}{2^k} < X_i \leq \frac{j+1}{2^k} \right) = \frac{1}{2^k}, \]

so by the continuity of \( \mathbb{P} \), we have \( \mathbb{P}(a < X_i \leq b) = b - a \) for every interval \( (a, b] \subset (0, 1] \).

3.13 Example. Given probability distribution functions \( F_1, F_2, \ldots \), we construct a sequence of independent random variables \( Y_1, Y_2, \ldots \) such that \( F_{Y_i} = F_i \) for each \( i \). We take the sequence \( X_1, X_2, \ldots \) of i.i.d. uniform random variables uniform on [0, 1] from Example 3.12. We set

\[ Y_i = G_i(X_i), \]

where \( G_i : [0, 1] \to \mathbb{R} \) is the inverse function of \( F_i \) defined in the proof of Theorem 2.13, that is

\[ G_i(x) = \inf\{y \in \mathbb{R}, \ F_i(y) \geq x\}. \]

Then (see the proof of Theorem 2.13), we have

\[ F_{Y_i}(t) = \mathbb{P} (Y_i \leq t) = \mathbb{P} (G_i(X_i) \leq t) = \mathbb{P} (X_i \leq F_i(t)) = F_i(t) \]
and the $Y_i$ are independent because the $X_i$ are independent.

### 3.4 Borel-Cantelli lemmas

Recall that for an infinite sequence of events $A_1, A_2, \ldots$, we define

$$\lim sup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

(infinitely many $A_k$ occur) and

$$\lim inf A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

(eventually all the $A_k$ occur, that is only finitely many $A_k$ do not occur). Plainly,

$$\left(\lim sup A_n\right)^c = \lim inf A_n^c.$$

The notation is explained by the following identities involving the indicator functions $1_{A_n}$:

$$\lim sup A_n = \{\omega \in \Omega, \lim sup_{n \to \infty} 1_{A_n}(\omega) = 1\}$$

and

$$\lim inf A_n = \{\omega \in \Omega, \lim inf_{n \to \infty} 1_{A_n}(\omega) = 1\}.$$

### 3.14 Lemma (The first Borel-Cantelli lemma).

If $A_1, A_2, \ldots$ are events such that

$$\sum_n \mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(\lim sup A_n) = 0.$$

**Proof.** By the monotonicity of the events $B_k = \bigcup_{n \geq k} A_n$ and the union bound, we get

$$\mathbb{P}(\lim sup A_n) = \mathbb{P}\left(\bigcap_k B_k\right) = \lim_{k \to \infty} \mathbb{P}(B_k) \leq \lim_{k \to \infty} \sum_{n \geq k} \mathbb{P}(A_n) = 0.$$

\[\square\]

### 3.15 Lemma (The second Borel-Cantelli lemma).

If $A_1, A_2, \ldots$ are independent events such that

$$\sum_n \mathbb{P}(A_n) = \infty,$$

then

$$\mathbb{P}(\lim sup A_n) = 1.$$
Proof. By the monotonicity of the events $B_k = \bigcap_{n \geq k} A_n^c$, we get

$$
P((\limsup A_n)^c) = \mathbb{P}\left( \bigcup_k B_k \right) = \lim_{k \to \infty} \mathbb{P}(B_k).
$$

so it is enough to show that $\mathbb{P}(B_k) = 0$. By independence, for $l \geq k$,

$$
\mathbb{P}(B_k) \leq \mathbb{P}\left( \bigcap_{l \geq n \geq k} A_n^c \right) = \prod_{k \leq n \leq l} \mathbb{P}(A_n^c) = \prod_{k \leq n \leq l} (1 - \mathbb{P}(A_n)).
$$

Thus, by the inequality $1 - x \leq e^{-x}$,

$$
\mathbb{P}(B_k) \leq e^{-\sum_{k \leq n \leq l} \mathbb{P}(A_n)}.
$$

Letting $l \to \infty$ and using that $\sum_{n \geq k} \mathbb{P}(A_n) = \infty$ finishes the proof. \hfill \Box

3.16 Example. Let $X_1, X_2, \ldots$ be i.i.d. random variable with the distribution function specified by the condition $\mathbb{P}(X_k > t) = e^{-t}$, $t > 0$ for each $k$. Fix $\alpha > 0$ and consider the events $A_n = \{X_n > \alpha \log n\}$. Since $\mathbb{P}(A_n) = e^{-\alpha \log n} = n^{-\alpha}$, by the Borel-Cantelli lemmas, we get

$$
\mathbb{P}(X_n > \alpha \log n \text{ for infinitely many } n) = \begin{cases} 
0, & \text{if } \alpha > 1, \\
1, & \text{if } \alpha \leq 1.
\end{cases}
$$

Let

$$
L = \limsup_{n \to \infty} \frac{X_n}{\log n}.
$$

Thus,

$$
\mathbb{P}(L \geq 1) = \mathbb{P}\left( \frac{X_n}{\log n} \text{ for infinitely many } n \right) = 1
$$

and

$$
\mathbb{P}(L > 1) = \mathbb{P}\left( \bigcup_{k \geq 1} \left\{ L > 1 + \frac{1}{k} \right\} \right)
\leq \sum_{k \geq 1} \mathbb{P}\left( \frac{X_n}{\log n} > 1 + \frac{1}{2k} \text{ for infinitely many } n \right)
= 0.
$$

Therefore, $L = 1$ a.s.

3.5 Tail events and Kolmogorov’s 0–1 law

For a sequence of random variables $X_1, X_2, \ldots$, we define its tail $\sigma$-algebra by

$$
\mathcal{T} = \bigcap_{n \geq 1} \sigma(X_{n+1}, X_{n+2}, \ldots).
$$
For example, very natural events such as
\[ \{ \lim_n X_n \text{ exists} \}, \quad \{ \sum_n X_n \text{ converges} \}, \quad \{ \limsup_n X_n > 1 \} \]
belong to \( \mathcal{T} \). If we have independence, the tail \( \sigma \)-algebra carries only trivial events.

**3.17 Theorem** (Kolmogorov’s 0-1 law). If \( X_1, X_2, \ldots \) is a sequence of independent random variables and \( \mathcal{T} \) is its tail \( \sigma \)-algebra, then for every \( A \in \mathcal{T} \), we have that either \( \mathbb{P}(A) = 0 \) or \( \mathbb{P}(A) = 1 \).

**Proof.** Define the \( \sigma \)-algebras
\[ \mathcal{X}_n = \sigma(X_1, \ldots, X_n) \]
and
\[ \mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots). \]
We prove the theorem by establishing the following 4 simple claims.

**Claim 1.** For every \( n \), \( \mathcal{X}_n \) and \( \mathcal{T}_n \) are independent.
Indeed, consider the family \( \mathcal{A} \) of the events of the form \( \{ \forall i \leq n, \; X_i \leq s_i \} \), \( s_i \in \mathbb{R} \) and the family \( \mathcal{B} \) of the events of the form \( \{ \forall n < i < n + m, \; X_i \leq t_i \} \), \( m \geq 1, \; t_i \in \mathbb{R} \). These are \( \pi \)-systems which generate \( \mathcal{X}_n \) and \( \mathcal{T}_n \) respectively. Clearly \( \mathcal{A} \) and \( \mathcal{B} \) are independent, hence \( \mathcal{X}_n \) and \( \mathcal{T}_n \) are independent (Lemma 3.1).

**Claim 2.** For every \( n \), \( \mathcal{X}_n \) and \( \mathcal{T} \) are independent.
This follows instantly because \( \mathcal{T} \subset \mathcal{T}_n \).

**Claim 3.** Let \( \mathcal{X} = \sigma(X_1, X_2, \ldots) \). Then \( \mathcal{X} \) and \( \mathcal{T}_n \) are independent.
Let \( \mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{X}_n \). This is a \( \pi \)-system generating \( \mathcal{X} \). By Claim 2, \( \mathcal{A} \) and \( \mathcal{T} \) are independent, so \( \mathcal{X} \) and \( \mathcal{T} \) are independent (Lemma 3.1).

**Claim 4.** For every \( A \in \mathcal{T} \), \( \mathbb{P}(A) \in \{0, 1\} \).
Since \( \mathcal{T} \subset \mathcal{X} \), by Claim 3, \( \mathcal{T} \) is independent of \( \mathcal{T} \), thus
\[ \mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A) \mathbb{P}(A), \]
hence the result. \( \square \)
3.6 Exercises

1. Define the Rademacher functions \( r_1, r_2, \ldots : [0, 1] \to \{-1, 0, 1\} \) by
\[
r_n(x) = \text{sgn}(\cos(2^n \pi x)), \quad x \in [0, 1],
\]
where \( \text{sgn} \) is the usual signum function. Consider these functions as random variables on the probability space \( ([0, 1], \mathcal{B}([0, 1]), \text{Leb}) \). What is the distribution of \( r_n \)? Show that the family \( \{r_n\}_{n \geq 1} \) is independent.

2. Define the Walsh functions \( w_A : \{-1, 1\}^n \to \{-1, 1\} \) indexed by all subsets \( A \) of \( \{1, \ldots, n\} \),
\[
w_A(x_1, \ldots, x_n) = \prod_{i \in A} x_i, \quad x = (x_1, \ldots, x_n) \in \{-1, 1\}^n
\]
and \( w_{\emptyset} = 1 \) (a constant function). Consider these functions as random variables on \( \{-1, 1\}^n \) equipped with the uniform probability measure. What is the distribution of \( w_A \)? Show that the \( w_A \) are pairwise independent. Are they independent?

3. For independent events \( A_1, \ldots, A_n \),
\[
(1 - e^{-1}) \min \left\{ 1, \sum_{i=1}^{n} \mathbb{P}(A_i) \right\} \leq \mathbb{P} \left( \bigcup_{i=1}^{n} A_i \right) \leq \min \left\{ 1, \sum_{i=1}^{n} \mathbb{P}(A_i) \right\}.
\]

4. Prove the so-called infinite monkey theorem: when we toss a fair coin infinitely many times then the event that “every given finite sequence of heads/tails occurs infinitely many times” is certain.

5. Suppose events \( A_1, A_2, \ldots \) are independent and all have equal probabilities. What is the probability that infinitely many \( A_i \)'s occur?

6. Suppose events \( A_1, A_2, \ldots \) are independent and \( \mathbb{P}(A_n) \in (0, 1) \) for every \( n \). Then infinitely many \( A_n \) occur with probability 1 if and only if at least one \( A_n \) occurs with probability 1.

7. Let \( \Omega \) be the set of positive integers and let \( A_k \) be the set of positive integers divisible by \( k \), \( k \geq 1 \). Is there a probability measure \( \mathbb{P} \) defined on all the subsets of \( \Omega \) such that \( \mathbb{P}(A_k) = \frac{1}{k} \) for every \( k = 1, 2, \ldots \)?

8. Prove Theorem 3.6.

9. Fill out the details in Example 3.10.

10. Let \( X_1, X_2, \ldots \) be a sequence of independent random variables and let \( \mathcal{T} \) be its tail \( \sigma \)-algebra. If a random variable \( Y \) is \( \mathcal{T} \)-measurable, then \( Y \) is a.s. constant.
11. Let $X_1, X_2, \ldots$ be a sequence of independent random variables. Show that the radius of convergence of the power series $\sum_{n=1}^{\infty} X_n z^n$ is a.s. constant.

12. Are there two nonconstant continuous functions $f, g : [0, 1] \to \mathbb{R}$ which, when viewed as random variables on the probability space $([0, 1], \mathcal{B}([0, 1]), \text{Leb})$, are independent?
4 Expectation

4.1 Definitions and basic properties

Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. We say that $X$ is integrable, if

$$\int_{\Omega} |X(\omega)|dP(\omega) < \infty$$

and then define its expectation (also called its mean) as

$$E_X = \int_{\Omega} X(\omega)dP(\omega).$$

Both integrals are Lebesgue integrals (see Appendix E for a construction and basic properties). For a random vector $X$ in $\mathbb{R}^n$, its expectation is defined as the following vector in $\mathbb{R}^n$,

$$EX = [E_{X_1} : \ldots : E_{X_n}].$$

We list the most important basic properties

(i) monotonicity: if $X$ is a nonnegative random variable, then $E_X \geq 0$,

(ii) the triangle inequality: $|E_X| \leq E|X|$,

(iii) linearity: if $X, Y$ are integrable, then for every $a, b \in \mathbb{R}$, $aX + bY$ is integrable and $E(ax + bY) = aE_X + bE_Y$.

We also list the most important limit theorems.

4.1 Theorem (Lebesgue’s monotone convergence theorem). If $X_1, X_2, \ldots$ are nonnegative random variables such that for every $n$, $X_{n+1} \geq X_n$, then

$$E\left(\lim_{n \to \infty} X_n\right) = \lim_{n \to \infty} E_{X_n}$$

(with the provision that the left hand side is $+\infty$ if and only if the right hand side is $+\infty$)

4.2 Theorem (Fatou’s lemma). If $X_1, X_2, \ldots$ are nonnegative random variables, then

$$E\left(\liminf_{n \to \infty} X_n\right) \leq \liminf_{n \to \infty} E_{X_n}.$$ 

4.3 Theorem (Lebesgue’s dominated convergence theorem). If $X, X_1, X_2, \ldots$ are random variables such that $X_n \xrightarrow{n \to \infty} X$ a.s. and for every $n$, $|X_n| \leq Y$ for some integrable random variable $Y$, then

$$E|X_n - X| \xrightarrow{n \to \infty} 0.$$

In particular,

$$E_X = E\left(\lim_{n \to \infty} X_n\right) = \lim_{n \to \infty} E_{X_n}.$$
The proofs are in Appendix E.

We turn to the relation between the expectation of a random variable and the integral against its law.

4.4 Theorem. Let $h: \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Let $X$ be a random variable. Then

$$\mathbb{E}h(X) = \int_{\mathbb{R}} h(x) d\mu_X(x).$$

(The identity should be understood as follows: if the integral on one side exists, then the other one does and they are equal.)

Proof. We leverage the linearity of both sides in $h$ and use a standard method from measure theory of complicating $h$.

I. If $h = 1_A$, for some $A \in \mathcal{B}(\mathbb{R})$, then

$$\mathbb{E}h(X) = \int_{\Omega} 1_A(\omega) d\mathbb{P}(\omega) = \mathbb{P}(A) = \mu_X(A) = \int_{\mathbb{R}} 1_A(x) d\mu_X(x) = \int_{\mathbb{R}} h(x) d\mu_X(x).$$

II. If $h$ is a simple function, that is $h = \sum_{i=1}^{N} x_i 1_{A_i}$ for some $x_1, \ldots, x_N \in \mathbb{R}$ and $A_1, \ldots, A_N \in \mathcal{B}(\mathbb{R})$, then the identity follows from the previous step by linearity.

III. If $h$ is a nonnegative function, then there is a sequence of nonnegative simple functions $h_1, h_2, \ldots$ such that for every $n$, $h_{n+1} \geq h_n$ and $h_n \to h$ (pointwise). Thus, the identity follows in this case from the previous step by Lebesgue’s monotone convergence theorem.

IV. If $h$ is arbitrary, we decompose it into its positive and negative part, $h^+ = \max\{0, h\}$, $h^- = \max\{0, -h\}$,

$$h = h^+ - h^-$$

and the identity follows from the previous step by linearity and the definition of Lebesgue integral. \qed

4.5 Remark. Note that the identity we proved is linear in $h$. The above argument of gradually complicating $h$ is standard in such situations.

4.6 Corollary. If $X$ is a discrete random variable with $p_i = \mathbb{P}(X = x_i) > 0$, $\sum_i p_i = 1$, then since $\mu_X = \sum_i p_i \delta_{x_i}$, we get

$$\mathbb{E}h(X) = \sum_i p_i h(x_i).$$

If $X$ is a continuous random variable with density $f$, then since

$$\int_{\mathbb{R}} h(x) d\mu_X(x) = \int_{\mathbb{R}} h(x) f(x) dx$$

(which can be justified exactly as in the proof of Theorem 4.4), we get

$$\mathbb{E}h(X) = \int_{\mathbb{R}} h(x) f(x) dx.$$
4.2 Variance and covariance

For a random variable \( X \) with \( \mathbb{E}X^2 < \infty \) (as we say, square-integrable), we define its \textbf{variance} as

\[
\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.
\]

Since \((X - \mathbb{E}X)^2 = X^2 - 2(\mathbb{E}X)X + (\mathbb{E}X)^2\), we have the convenient formula,

\[
\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2.
\]

Note that by its definition, the variance is shift invariant,

\[
\text{Var}(X + c) = \text{Var}(X),
\]

for every constant \( c \in \mathbb{R} \), and scales quadratically,

\[
\text{Var}(\lambda X) = \lambda^2 \text{Var}(X),
\]

for every constant \( \lambda \in \mathbb{R} \). Moreover, if \( X \) and \( Y \) are random variables with \( \mathbb{E}X^2, \mathbb{E}Y^2 < \infty \), then because \((X + Y)^2 \leq 2X^2 + 2Y^2\), we have \( \mathbb{E}(X + Y)^2 < \infty \) and denoting \( \bar{X} = X - \mathbb{E}X \), \( \bar{Y} = Y - \mathbb{E}Y \), we obtain

\[
\text{Var}(X + Y) = \mathbb{E}(\bar{X} + \bar{Y})^2 = \mathbb{E}\bar{X}^2 + \mathbb{E}\bar{Y}^2 + 2\mathbb{E}\bar{X}\bar{Y} = \text{Var}(X) + \text{Var}(Y) + 2\mathbb{E}\bar{X}\bar{Y}.
\]

This motives the following definition of the \textbf{covariance} between such two random variables,

\[
\text{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) = \mathbb{E}XY - (\mathbb{E}X)(\mathbb{E}Y).
\]

By the above identity we also obtain the following formula for the variance of the sum.

\[
4.7 \textbf{Theorem. Let } X_1, \ldots, X_n \text{ be square-integrable random variables. Then}
\]

\[
\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j} \text{Cov}(X_i, X_j).
\]

In particular, if the \( X_i \) are \textbf{uncorrelated}, that is \( \text{Cov}(X_i, X_j) = 0 \) for all \( i \neq j \), we have

\[
\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).
\]

For a random vector \( X \) in \( \mathbb{R}^n \) with square-integrable components, we define its \textbf{covariance matrix} as

\[
\text{Cov}(X) = [\text{Cov}(X_i, X_j)]_{i,j \leq n}.
\]

It is convenient to write

\[
\text{Cov}(X) = \mathbb{E}\bar{X}\bar{X}^\top,
\]

with \( \bar{X} = X - \mathbb{E}X \) (here the expectation of the \( n \times n \) matrix \( \bar{X}\bar{X}^\top \) is understood entry-wise). From this and the linearity of expectation, we quickly obtain the following basic properties of covariance matrices.

33
4.8 Theorem. Let $X$ be a random vector in $\mathbb{R}^n$ with square-integrable components. Then

(i) $\text{Cov}(X)$ is a symmetric positive semi-definite matrix

(ii) $\text{Cov}(X + b) = \text{Cov}(X)$, for every (deterministic) vector $b \in \mathbb{R}^n$,

(iii) $\text{Cov}(AX) = A \text{Cov}(X) A^\top$, for every $n \times n$ matrix $A$,

(iv) if $r = \text{rank}(\text{Cov}(X))$, then $\mathbb{P}(X \in H) = 1$ for some $r$-dimensional affine subspace of $\mathbb{R}^n$.

Proof. We show (iv) and leave the rest as an exercise. Let $M = \text{Cov}(X)$. If $M$ has rank $r$, then there are $n - r$ linearly independent vectors in its kernel, say $v_1, \ldots, v_{n-r}$. Since $Mv_i = 0$, we have

$$0 = v_i^\top Mv_i = \mathbb{E}(v_i^\top \bar{X}^\top v_i) = \mathbb{E}((\bar{X}^\top v_i)^2),$$

so the nonnegative random variable $(\bar{X}^\top v_i)^2$ whose expectation is 0 therefore has to be 0 a.s. This holds for every $i$, thus $\mathbb{P}(\forall i \leq n - r \, \bar{X}^\top v_i = 0) = 1$ and we can take $H = \{x \in \mathbb{R}^n, \forall i \leq n - r \, (x - \mathbb{E}X)^\top v_i = 0\}$. \qed

4.3 Independence again, via product measures

Given two probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i), i = 1, 2$, we define their product by taking, of course,

$$\Omega = \Omega_1 \times \Omega_2$$

and

$$\mathcal{F} = \sigma(\mathcal{A}_1 \times \mathcal{A}_2, \quad A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2)$$

which is called the product $\sigma$-algebra, denoted

$$\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2.$$

Then the product measure $\mathbb{P}$, denoted

$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2,$$

is the unique probability measure on $\mathcal{F}$ such that for all $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$,

$$\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2).$$

Its existence is related to Fubini’s theorem (see Appendix C). It plainly generalises to finite products.
4.9 Example. Thanks to separability, we have
\[ B(\mathbb{R}^n) = B(\mathbb{R}) \otimes \cdots \otimes B(\mathbb{R}) \]
(with the right hand side usually denoted \( B(\mathbb{R})^\otimes n \)). One inclusion relies only on the definition of the product topology, that is
\[ B(\mathbb{R})^\otimes n \subset B(\mathbb{R}^n) \]
holds because if \( A_1, \ldots, A_n \) are open, then \( A_1 \times \cdots \times A_n \) is open, so the generators of the left hand side belong to \( B(\mathbb{R}^n) \). The opposite inclusion,
\[ B(\mathbb{R}^n) \subset B(\mathbb{R})^\otimes n \]
holds because an open set in \( \mathbb{R}^n \) is a countable union of the sets of the form \( \prod_{i=1}^n (a_i, b_i) \), by separability, thus the generators of the left hand side belong to \( B(\mathbb{R})^\otimes n \).

For infinite products, we have the following result, which also gives a canonical construction of an infinite sequence of i.i.d. random variables with specified arbitrary laws.

4.10 Theorem. Let \( \mu_1, \mu_2, \ldots \) be probability measures on \((\mathbb{R}, B(\mathbb{R}))\). We set
\[ \Omega = \prod_{i=1}^\infty \mathbb{R} = \mathbb{R} \times \mathbb{R} \times \ldots, \]
\[ X_n(\omega_1, \omega_2, \ldots) = \omega_n, \quad (\omega_1, \omega_2, \ldots) \in \Omega, \]
and
\[ \mathcal{F} = \sigma(X_1, X_2, \ldots). \]
There is a unique probability measure \( P \) on \((\Omega, \mathcal{F})\) such that for every \( k \geq 1 \) and \( A_1, \ldots, A_k \in B(\mathbb{R}) \), \( w \) have
\[ P(A_1 \times \cdots \times A_k \times \mathbb{R} \times \ldots) = \mu_1(A_1) \cdots \mu_k(A_k). \]
Moreover, \( X_1, X_2, \ldots \) are independent random variables on \((\Omega, \mathcal{F}, P)\) with \( \mu_{X_i} = \mu_i \).

We defer its proof to Appendix D. It is based on Carathéodory’s theorem.

Recall that random variables \( X_1, \ldots, X_n \) are independent if and only if its joint law \( \mu_{(X_1, \ldots, X_n)} \) is the product measure \( \mu_{X_1} \otimes \cdots \otimes \mu_{X_n} \) (Theorem 3.4). Using this, we prove one of the most significant consequences of independence: the expectation of the product is the product of the expectations.

4.11 Theorem. Let \( X_1, \ldots, X_n \) be integrable random variables. If they are independent, then \( X_1 \cdots X_n \) is integrable and
\[ \mathbb{E}(X_1 \cdots X_n) = \mathbb{E}X_1 \cdots \mathbb{E}X_n. \]
Proof. We have,

\[
E|X_1 \cdot \ldots \cdot X_n| = \int_{\mathbb{R}^n} |x_1 \cdot \ldots \cdot x_n|d\mu_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n) \\
= \int_{\mathbb{R}^n} |x_1 \cdot \ldots \cdot x_n|d\mu_{X_1}(x_1) \ldots d\mu_{X_n}(x_n) \\
= \prod_{i=1}^n \int_{\mathbb{R}} |x_i|d\mu_{X_i}(x_i),
\]

where in the second equality we use independence and in the last one – Fubini’s theorem. This shows that \(X_1 \cdot \ldots \cdot X_n\) is integrable. The proof of the identity then follows exactly the same lines.

Of course, the converse statement is not true. Take for instance a uniform random variable \(X\) on \((-1, 0, 1)\) and \(Y = |X|\). Then \(E(XY) = 0 = E(X) \cdot E(Y)\), but \(X\) and \(Y\) are not independent.

As a useful corollary, independent random variables are uncorrelated, so we also have that the variance of the sum of independent random variables is the sum of their variances (recall Theorem 4.7).

4.12 Corollary. If \(X_1, X_2\) are independent, then \(\text{Cov}(X_1, X_2) = 0\). In particular, if \(X_1, \ldots, X_n\) are independent square-integrable, then

\[
\text{Var} \left( \sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var}(X_i).
\]

Since \(F_X\) determines the law of \(X\), it should be possible to express \(E(X)\) using it. We finish this chapter with such a formula which is obtained from a simple trick that \(x = \int_0^x dt, x \geq 0\), combined with Fubini’s theorem.

4.13 Theorem. If \(X\) is a nonnegative random variable, then

\[
E(X) = \int_0^\infty P(X > t) dt.
\]

Proof. We have,

\[
E(X) = E \left( \int_0^X dt \right) = E \int_0^\infty 1_{X>t} dt = \int_0^\infty E1_{X>t} dt = \int_0^\infty P(X > t) dt,
\]

where the usage of Fubini’s theorem is justified because \(1_{X>t}\) is a nonnegative function.
4.4 Exercises

1. An urn contains \( N \) balls among which exactly \( b \) are yellow. We pick uniformly at random \( n \) (\( n \leq N \)) balls without replacement. Let \( X \) be the number of yellow balls picked. Find the expectation and variance of \( X \).

2. Show that a nonnegative random variable \( X \) is integrable if and only if
\[
\sum_{n=1}^{\infty} P(X > n) < \infty.
\]

3. Let \( p > 0 \). If \( X \) is a nonnegative random variable, then
\[
\mathbb{E}X^p = p \int_0^\infty t^{p-1} P(X > t) \, dt.
\]
Give an analogous formula for \( \mathbb{E}f(X) \) for an arbitrary increasing and differentiable function \( f : [0, \infty) \to [0, \infty) \) with \( f(0) = 0 \).

4. Let \( p > 0 \) and \( X \) be a random variable with \( \mathbb{E}|X|^p < \infty \). Then
\[
\lim_{t \to \infty} t^p P(|X| > t) = 0.
\]

5. Suppose \( X \) and \( Y \) are independent random variables and the distribution function of \( X \) is continuous. Then \( P(X = Y) = 0 \).

6. Let \( X \) and \( Y \) be independent random variables taking values in \( S = \{ z \in \mathbb{C}, |z| = 1 \} \). If \( X \) is uniform, then \( XY \) is also uniform.

7. Suppose \( X \) and \( Y \) are positive random variables with the same distribution. Does it follow that \( \mathbb{E}X^{X+Y} = \mathbb{E}Y^{X+Y} \)?

8. Let \( X \) and \( Y \) be bounded random variables. Show that \( X \) and \( Y \) are independent if and only if for every positive integers \( m, n \), we have \( \mathbb{E}(X^m Y^n) = \mathbb{E}X^m \mathbb{E}Y^n \).

9. Let \( X \) be a square-integrable random variable. Find \( \min_{x \in \mathbb{R}} \mathbb{E}(X - x)^2 \).

10. Let \( X \) be an integrable random variable. Show that \( \min_{x \in \mathbb{R}} \mathbb{E}|X - x| \) is attained at \( x = \text{Med}(X) \), the median of \( X \), that is any number \( m \) for which \( P(X \geq m) \geq \frac{1}{2} \) and \( P(X \leq m) \geq \frac{1}{2} \).

11. Let \( X \) be a square-integrable random variable. We have, \( |\mathbb{E}X - \text{Med}(X)| \leq \sqrt{\text{Var}(X)} \).

12. Prove properties (i)-(iii) of covariance matrices from Theorem 4.8.

13. Suppose there is a countable family of disjoint open disks with radii \( r_1, r_2, \ldots \), all contained in the unit square \([0,1]^2\) on the plane. If the family covers \([0,1]^2\) up to a set of (Lebesgue) measure 0, then \( \sum_i r_i = \infty \).
5 More on random variables

5.1 Important distributions

We list several discrete and continuous laws of random variables that appear very often in probability theory.

1) The Dirac delta distribution. For \( a \in \mathbb{R} \), let \( X \) be an a.s. constant random variable, \( \mathbb{P}(X = a) = 1 \).

Then

\[
\mu_X = \delta_a
\]

is the Dirac delta distribution at \( a \). We have,

\[
\mathbb{E}X = a, \quad \text{Var}(X) = 0.
\]

2) The Bernoulli distribution. For \( p \in [0, 1] \), let \( X \) be a random variable taking two values 0 and 1 with probabilities \( \mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p \).

Then

\[
\mu_X = (1 - p)\delta_0 + p\delta_1
\]

is the Bernoulli distribution with parameter \( p \). We have,

\[
\mathbb{E}X = p, \quad \text{Var}(X) = p(1 - p).
\]

Notation: \( X \sim \text{Ber}(p) \).

3) The binomial distribution. For an integer \( n \geq 1 \) and \( p \in [0, 1] \), let \( X \) be a random variable taking values \( \{0, 1, \ldots, n\} \) with probabilities

\[
\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad 0 \leq k \leq n.
\]

Then

\[
\mu_X = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} \delta_k
\]

is the Binomial distribution with parameters \( n \) and \( p \). It can be directly checked that

\( X \) has the same law as \( X_1 + \cdots + X_n \), where \( X_1, \ldots, X_n \) are i.i.d. Bernoulli random variables with parameter \( p \), which gives a very convenient probabilistic representation of \( X \). In other words,

\( X \) is the number of successes in \( n \) independent Bernoulli trials.
We have,
\[ EX = E(X_1 + \ldots + X_n) = nEX_1 = np \]
and
\[ \text{Var}(X) = n \text{Var}(X_1) = np(1 - p). \]
Notation: \( X \sim \text{Bin}(n, p). \)

4) The Poisson distribution. For \( \lambda > 0 \), let \( X \) be a random variable taking values \( \{0, 1, 2, \ldots\} \) with probabilities
\[ \mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0. \]
Then
\[ \mu_X = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta_k \]
is the Poisson distribution with parameter \( \lambda \). We will see later that this distribution arises as an appropriate limit of the Binomial distribution with parameters \( n \) and \( \lambda/n \) as \( n \to \infty \). In other words, \( X \) is “the number of successes in \( n \) independent Bernoulli trials each with probability of success \( \frac{\lambda}{n} \) as \( n \to \infty \), so that the rate of success is \( \lambda \). This distribution models well the number of events occurring in a fixed interval of time if these events occur with a constant mean rate \( \lambda \), independently of the time since the last event, say the number of calls in a busy call centre.

We have,
\[ \mathbb{E}X = \lambda, \quad \text{Var}(X) = \lambda. \]
Notation: \( X \sim \text{Pois}(\lambda). \)

5) The geometric distribution. For \( p \in [0, 1] \), let \( X \) be a random variable taking values \( \{1, 2, \ldots\} \) with probabilities
\[ \mathbb{P}(X = k) = (1 - p)^{k-1}p, \quad k \geq 1. \]
Then
\[ \mu_X = \sum_{k=1}^{\infty} (1 - p)^{k-1}p \delta_k \]
is the Geometric distribution with parameter \( p \). It can be directly checked that
\( X \) has the same law as \( \inf\{n \geq 1, \ X_n = 1\} \), where \( X_1, X_2, \ldots \) are i.i.d. Bernoulli random variables with parameter \( p \). In other words,
\( X \) is the number of trials in independent Bernoulli trials until first success.

We have,
\[ \mathbb{E}X = \frac{1}{p}, \quad \text{Var}(X) = \frac{1 - p}{p^2}. \]
Notation: \( X \sim \text{Geom}(p). \)
6) **The uniform distribution.** For a Borel set $K$ in $\mathbb{R}^n$ of positive finite Lebesgue measure (volume) $|K|$, let $X$ be a random variable with density function

$$f(x) = \frac{1}{|K|}1_K(x), \quad x \in \mathbb{R}^n.$$ 

Then

$$\mu_X(A) = \int_A f(x)dx = \frac{|A \cap K|}{|K|}, \quad A \in \mathcal{B}(\mathbb{R}^n).$$

We say that $\mu_X$ is the uniform measure on $K$. We have,

$$\mathbb{E}X = \frac{1}{|K|} \int_K xdx \quad \text{(the barycentre of } K).$$

Notation: $X \sim \text{Unif}(K)$.

In particular, if $K = [0, 1]$ in $\mathbb{R}$, $X$ is uniform on the unit interval $[0, 1]$ and we have

$$\mathbb{E}X = \frac{1}{2}, \quad \text{Var}(X) = \frac{1}{12}.$$  

7) **The exponential distribution.** For $\lambda > 0$, let $X$ be a random variable with density function

$$f(x) = \lambda e^{-\lambda x}1_{(0, \infty)}(x), \quad x \in \mathbb{R}.$$ 

We say that $\mu_X$ (or $X$) has the exponential distribution with parameter $\lambda$. This is a continuous analogue of the geometric distribution. It has the so-called memory-less property: for every $s, t > 0$,

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)$$

which characterises it uniquely among continuous distributions (see exercises). We have,

$$\mathbb{E}X = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$ 

Notation: $X \sim \text{Exp}(\lambda)$.

8) **The gamma distribution.** For $\beta, \lambda > 0$, let $X$ be a random variable with density function

$$f(x) = \frac{\lambda^\beta}{\Gamma(\beta)} x^{\beta-1}e^{-\lambda x}1_{(0, \infty)}(x), \quad x \in \mathbb{R},$$

where

$$\Gamma(\beta) = \int_0^\infty t^{\beta-1}e^{-t}dt,$$

is the Gamma function. We say that $\mu_X$ (or $X$) has the Gamma distribution with parameters $\beta$ and $\lambda$. When $\beta = n$ is a positive integer, we have a nice probabilistic representation,

$$X \text{ has the same law as } X_1 + \cdots + X_n,$$
where \(X_1, \ldots, X_n\) are i.i.d. exponential random variables with parameter \(\lambda\). We have,

\[
\mathbb{E}X = \frac{\beta}{\lambda}, \quad \text{Var}(X) = \frac{\beta}{\lambda^2}.
\]

Notation: \(X \sim \text{Gamma}(\beta, \lambda)\).

9) The beta distribution. For \(\alpha, \beta > 0\), let \(X\) be a random variable with density function

\[
f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}1_{(0,1)}(x), \quad x \in \mathbb{R},
\]

where

\[
B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}
\]

is the Beta function. We say that \(\mu_X\) (or \(X\)) has the Beta distribution with parameters \(\alpha, \beta\). This distribution appears naturally as a marginal of a random vector uniform on the centred unit Euclidean ball. We have,

\[
\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\]

Notation: \(X \sim \text{Beta}(\alpha, \beta)\).

10) The Cauchy distribution. Let \(X\) be a random variable with density function

\[
f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.
\]

We say that \(\mu_X\) (or \(X\)) has the standard Cauchy distribution. It has the following stability property: for every \(a_1, \ldots, a_n \in \mathbb{R}\),

\[
a_1X_1 + \cdots + a_nX_n \text{ has the same law as } (\sum |a_i|)X,
\]

where \(X_1, \ldots, X_n\) are i.i.d. copies of \(X\). Cauchy random variables are not integrable.

Notation: \(X \sim \text{Cauchy}(1)\).

11) The Gaussian distribution. Let \(X\) be a random variable with density function

\[
f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad x \in \mathbb{R}.
\]

We say that \(\mu_X\) (or \(X\)) has the standard Gaussian (or normal) distribution. We have,

\[
\mathbb{E}X = 0, \quad \text{Var}(X) = 1.
\]

Notation: \(X \sim \text{N}(0,1)\).

For \(\mu \in \mathbb{R}\) and \(\sigma > 0\) consider

\[
Y = \mu + \sigma X.
\]

This a Gaussian random variable with parameters \(\mu\) and \(\sigma\). It has density

\[
g(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.
\]
We have,

\[ EY = \mu, \quad \text{Var}(Y) = \sigma^2. \]

Notation: \( Y \sim N(\mu, \sigma^2). \)

The key property of the Gaussian distribution is that \textit{sums of independent Gaussians are Gaussian}. Formally, let \( Y_1 \sim N(\mu_1, \sigma_1^2), Y_2 \sim N(\mu_2, \sigma_2^2) \) be two independent Gaussian random variables. Then,

\[ Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \quad (5.1) \]

We prove this later. Because of the central limit theorem, Gaussian random variables are ubiquitous.

5.2 Gaussian vectors

Let \( X_1, \ldots, X_n \) be i.i.d. standard Gaussian random variables. The vector

\[ X = (X_1, \ldots, X_n) \]

is called a standard Gaussian random vector in \( \mathbb{R}^n \). It has density

\[ f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = (2\pi)^{-n/2} e^{-|x|^2/2}, \quad x \in \mathbb{R}^n \]

(here \(|x| = \sqrt{\sum x_i^2}\) is the Euclidean norm of \( x \)). Note that \( X \) enjoys at the same time two important features: 1) \( X \) has independent components (its law is a product measure), 2) because the density of \( X \) is rotationally invariant, so is \( X \), that is for every orthonormal matrix \( U \in O(n) \),

\[ UX \text{ has the same law as } X. \]

We have,

\[ EY = b, \quad Q = \text{Cov}(Y) = AA^\top. \]

Notation: \( Y \sim N(b, Q). \)
In particular, if \( m = n \) and \( A \) is nonsingular, then \( Y \) has density
\[
g(x) = \frac{1}{\sqrt{2\pi} \sqrt{\det Q}} e^{-\frac{1}{2}(Q^{-1}(x-b),(x-b))}, \quad x \in \mathbb{R}^n
\]
where
\[
(x, y) = \sum_{i=1}^{n} x_i y_i
\]
is the standard scalar product on \( \mathbb{R}^n \).

All of the claims made here are standard but very important computations and we leave the details as exercise.

5.3 Sums of independent random variables

Recall that the convolution of two integrable functions \( f, g : \mathbb{R} \to \mathbb{R} \) is defined as a function
\[
x \mapsto (f \ast g)(x) = \int_{\mathbb{R}} f(x-y)g(y) \, dy
\]
which by Fubini’s theorem is well-defined because \( w(x, y) = f(x)g(y) \) is integrable on \( \mathbb{R}^2 \), so \( w(x - y, y) \) is also integrable on \( \mathbb{R}^2 \).

Convolutions appear naturally when we take sums of independent random variables.

5.1 Theorem. Let \( X \) and \( Y \) be independent random variables. Then the law of \( X + Y \) is given by
\[
\mu_{X+Y}(A) = \int_{A} \mu_Y(A-x) \, d\mu_X(x) = \int_{A} \mu_X(A-y) \, d\mu_Y(y), \quad A \in \mathcal{B}(\mathbb{R}).
\]
In particular, if \( X \) has density \( f \), then \( X + Y \) has density
\[
h(x) = \int_{\mathbb{R}} f(x-y) \, d\mu_Y(y).
\]
If both \( X, Y \) have densities, say \( f, g \) respectively, then
\[
X + Y \text{ has density } f \ast g.
\]

Proof. By independence, \( \mu_{(X,Y)} = \mu_X \otimes \mu_Y \), thus
\[
\mu_{X+Y}(A) = \mu_{(X,Y)} \{ (x, y) \in \mathbb{R}^2, \ x + y \in A \} = \int_{\{x, y \in \mathbb{R}^2, \ x + y \in A\}} d\mu_X(x) \, d\mu_Y(y)
\]
\[
= \int_{x \in \mathbb{R}} \left[ \int_{y \in A-x} d\mu_Y(y) \right] d\mu_X(x),
\]
where the last equality follows by Fubini’s theorem. Since \( \int_{y \in A-x} d\mu_Y(y) = \mu_Y(A-x) \), the first identity follows. Note that swapping the roles of \( X \) and \( Y \) above gives also the identity
\[
\mu_{X+Y}(A) = \int_{\mathbb{R}} \mu_X(A-y) \, d\mu(y).
\]
If $X$ has density $f$, we have $\mu_X(A - y) = \int_{A-y} f(x)dx$, so by a change of variables $x = z - y$ and Fubini’s theorem, we get

$$
\mu_{X+Y}(A) = \int_R \int_A f(x)dx d\mu_Y(y) = \int \int_A f(z - y) dz d\mu_Y(y) = \int_A \left[ \int_R f(z - y) d\mu_Y(y) \right] dz,
$$
so $h(z) = \int_R f(z - y) d\mu_Y(y) = Ef(z - Y)$ is the density of $X + Y$. Finally, if $Y$ has also density, say $g$, then this becomes $h(z) = \int_R f(z - y) g(y) dy$, that is $h = f * g$.

Sometimes we use the notation $\mu_X \star \mu_Y$ to denote $\mu_{X+Y}$. To illustrate this theorem, we consider the example of sums of independent Gaussians.

5.2 Example. Let $X \sim N(0,1), Y \sim N(0,\sigma^2)$ be independent. The densities of $X$ and $Y$ are respectively $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $g(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-y^2/2}$. Thus the density of $X + Y$ is given by

$$
h(z) = \int f(z - x) g(x) dx = \int_R \int e^{-\frac{1}{2}(z-x)^2 - \frac{1}{2\sigma^2}x^2} dx = \int_R e^{-\frac{1}{2}(z-x)^2 - \frac{1}{2\sigma^2}x^2} dx \cdot \int_R e^{-u^2} du = \frac{1}{\sqrt{2\pi\sqrt{1+\sigma^2}}} e^{-\frac{1}{2}(z-x)^2 - \frac{1}{2\sigma^2}x^2}.
$$

that is

$$
X + Y \sim N(0,1 + \sigma^2).
$$

Using this, linearity and the fact that for $Y \sim N(\mu,\sigma^2)$ we can write $Y = \mu + \sigma X$ for a standard Gaussian $X$, we can easily deduce (5.1).

5.4 Density

Recall that a random variable $X$ has density $f$ if for every $t \in \mathbb{R}$,

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

How to find out whether $X$ has density and if that is the case, determine it using its distribution function $F_X$?

5.3 Lemma. Let $F: \mathbb{R} \to \mathbb{R}$ be a nondecreasing, right-continuous function such that $F'$ exists a.e. Then for every $a < b$, we have

$$
\int_a^b F' \leq F(b) - F(a).
$$
Proof. By Fatou’s lemma,
\[
\int_a^b F'(t)dt = \int_a^b \liminf_{\delta \to 0^+} \frac{F(t + \delta) - F(t)}{\delta} dt \leq \liminf_{\delta \to 0^+} \int_a^b \frac{F(t + \delta) - F(t)}{\delta} dt
\]
\[
= \liminf_{\delta \to 0^+} \frac{1}{\delta} \left( \int_b^{b+\delta} F(t) dt - \int_a^{a+\delta} F(t) dt \right)
\]
and the right hand side equals \(F(b) - F(a)\) by the right-continuity of \(F\).

5.4 Corollary. Under the assumptions of Lemma 5.3, if additionally \(\lim_{t \to -\infty} F(t) = 0\) and \(\lim_{t \to +\infty} F(t) = 1\), then for every \(x \in \mathbb{R}\), we have
\[
\int_{-\infty}^x F(t) \leq F(x) \quad \text{and} \quad \int_x^\infty F(t) \leq 1 - F(x).
\]

5.5 Theorem. If \(X\) is a random variable such that \(F'_X\) exists a.e. and \(\int_{-\infty}^\infty F'_X = 1\), then \(X\) is continuous with density
\[
f(x) = \begin{cases} 
F'_X(x), & \text{if } F'_X(x) \text{ exists,} \\
0, & \text{otherwise.}
\end{cases}
\]

Proof. By Corollary 5.4, it remains to show that for every \(x \in \mathbb{R}\), we have \(\int_x^\infty F'_X = F(x)\).
This follows from
\[
\int_{-\infty}^x F'_X + \int_x^\infty F'_X = \int_{-\infty}^\infty F'_X = 1
\]
and \(\int_x^\infty F'_X \leq 1 - F_X(x)\). □
5.5 Exercises

1. There are \( n \) different coupons and each time you obtain a coupon it is equally likely to be any of the \( n \) types. Let \( Y_i \) be the additional number of coupons collected, after obtaining \( i \) distinct types, before a new type is collected (including the new one). Show that \( Y_i \) has the geometric distribution with parameter \( \frac{n-i}{n} \) and find the expected number of coupons collected before you have a complete set.

2. The double exponential distribution with parameter \( \lambda > 0 \) has density \( f(x) = \frac{\lambda}{2}e^{-\lambda|x|} \). Find its distribution function, sketch its plot, find the mean and variance. Let \( X \) and \( Y \) be i.i.d. exponential random variables with parameter 1. Find the distribution of \( X - Y \).

3. Let \( X \) and \( Y \) be independent Poisson random variables with parameters \( \mu \) and \( \lambda \). Show that \( X + Y \) is a Poisson random variable with parameter \( \mu + \lambda \).

4. Let \( X \) be a uniform random variable on \((0, 1)\). Find the distribution function and density of \( Y = -\ln X \). What is the distribution of \( Y \) called?

5. Let \( X \) be a Poisson random variable with parameter \( \lambda \). Show that \( P(X \geq k) = P(Y \leq \lambda) \), for \( k = 1, 2, \ldots \), where \( Y \) is a random variable with the Gamma distribution with parameter \( k \).

6. Let \( X \) and \( Y \) be independent exponential random variables with parameters \( \lambda \) and \( \mu \). Show that \( \min\{X, Y\} \) has the exponential distribution with parameter \( \lambda + \mu \).

7. Let \( X_1, X_2, \ldots \) be independent exponential random variables with parameter 1. Show that for every \( n \), the distribution of \( X_1 + \ldots + X_n \) is Gamma(\( n \)). Generalise this to sums of independent random variables with Gamma distributions: if \( X_1, \ldots, X_n \) are independent with \( X_i \sim \Gamma(\beta_i) \), then \( \sum_{i=1}^{n} X_i \sim \Gamma(\sum_{i=1}^{n} \beta_i) \).

8. Let \((X, Y)\) be a random vector in \( \mathbb{R}^2 \) with density \( f(x, y) = cxy1_{0<x<y<1} \). Find \( c \) and \( P(X + Y < 1) \). Are \( X \) and \( Y \) independent? Find the density of \((X/Y, Y)\). Are \( X/Y \) and \( Y \) independent?

9. Let \( X \) and \( Y \) be independent standard Gaussian random variables. Show that \( X/Y \) has the Cauchy distribution.

10. Let \( X = (X_1, \ldots, X_n) \) be a random vector in \( \mathbb{R}^n \) uniformly distributed on the simplex \( \{x \in \mathbb{R}^n, x_1 + \ldots + x_n \leq 1, x_1, \ldots, x_n \geq 0\} \). Find \( \text{E}X_1, \text{E}X_i^2, \text{E}X_1X_2 \), the covariance matrix of \( X \) and its determinant.

11. Let \( U_1, \ldots, U_n \) be a sequence of i.i.d. random variables, each uniform on \([0, 1]\). Let \( U^*_1, \ldots, U^*_n \) be its nondecreasing rearrangement, that is \( U^*_1 \leq \ldots \leq U^*_n \). In particular, \( U^*_1 = \min\{U_1, \ldots, U_n\} \) and \( U^*_n = \max\{U_1, \ldots, U_n\} \). Show that the vector
(\(U_1^*, \ldots, U_n^*\)) is uniform on the simplex \(\{x \in \mathbb{R}^n, 0 \leq x_1 \leq \ldots \leq x_n \leq 1\}\). Find \(\mathbb{E} U_k^*\) for \(1 \leq k \leq n\).

12. Show the lack of memory property characterises the exponential distribution. Specifically, let \(X\) be a random variable such that for every positive \(s\) and \(t\), \(\mathbb{P}(X > s) > 0\) and \(\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)\). Show that \(X\) has the exponential distribution.

13. Let \(X\) be a random variable such that there is a function \(g: \mathbb{R} \to \mathbb{R}\) such that \(F_X(t) = \int_{-\infty}^{t} g(x) dx\) for every \(t \in \mathbb{R}\). Then \(X\) is continuous and \(g\) is the density of \(X\).

14. Let \(U_1, U_2, U_3\) be independent uniform random variables on \([-1, 1]\). Find the density of \(U_1 + U_2\) and \(U_1 + U_2 + U_3\).

15. Let \(X\) and \(Y\) be independent random variables with densities \(f\) and \(g\) respectively. Show that \(Z = X/Y\) has density \(h(z) = \int_{-\infty}^{\infty} |y| f(yz) g(y) dy\), \(z \in \mathbb{R}\).

16. Let \(X\) be a standard Gaussian random variable and \(Y\) be an exponential random variable with parameter 1, independent of \(X\). Show that \(\sqrt{2Y}X\) has the symmetric (two-sided) exponential distribution with parameter 1.

17. Let \(X_1, X_2, X_3\) be i.i.d. standard Gaussian random variables. Find the mean and variance of \(Y = 3X_1 - X_2 + 2X_3\). Find its density.

18. Show that a continuous Gaussian random vector in \(\mathbb{R}^n\) has independent components if and only if they are uncorrelated.

19. Give an example of a random vector \((X, Y)\) such that \(X\) and \(Y\) are uncorrelated Gaussian random variables but \(X\) and \(Y\) are not independent.

20. Let \((X, Y)\) be a standard Gaussian random vector in \(\mathbb{R}^2\). Let \(\rho \in (-1, 1)\) and define

\[
\begin{bmatrix}
U \\ V
\end{bmatrix} = \begin{bmatrix}
\frac{1}{\sqrt{1 - \rho^2}} X + \frac{\rho}{\sqrt{1 - \rho^2}} Y \\
\frac{1}{\sqrt{1 - \rho^2}} Y + \frac{\rho}{\sqrt{1 - \rho^2}} X
\end{bmatrix}.
\]

Find the density of \((U, V)\). Is this a Gaussian random vector? What is its covariance matrix? What is the distribution of \(U\) and \(V\)? Determine the values of \(\rho\) for which \(U\) and \(V\) are independent.

21. Let \(\rho \in (-1, 1)\) and let \((U, V)\) be a random vector in \(\mathbb{R}^2\) with density

\[
f(u, v) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{ -\frac{1}{2(1 - \rho^2)} (u^2 - 2\rho uv + v^2) \right\}, \quad (u, v) \in \mathbb{R}^2.
\]

Is it a Gaussian random vector? Find the covariance matrix of \((U, V)\). Find the distributions of the marginals \(U\) and \(V\). Determine the values of \(\rho\) for which \(U\) and \(V\) are independent.
22. Suppose \((X, Y)\) is a centred (i.e., \(EX = EY = 0\)) Gaussian random vector in \(\mathbb{R}^2\) with \(\text{Cov}([X, Y]) = [\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}]\). Find, a) the density of \((X, Y)\), b) the density of \(X + 3Y\), c) all \(\alpha \in \mathbb{R}\) for which \(X + Y\) and \(X + \alpha Y\) are independent.

23. Let \(G\) be a standard Gaussian vector in \(\mathbb{R}^n\) and let \(U\) be an \(n \times n\) orthogonal matrix. Find the density of \(UG\). Are the components of this vector independent?

24. Let \(g\) be a standard Gaussian random variable. Show that \(Eg^{2m} = 1 \cdot 3 \cdot \ldots \cdot (2m - 1), m = 1, 2, \ldots\).

25. Using Fubini’s theorem and the fact that the standard Gaussian density integrates to 1, find the volume of a Euclidean ball in \(\mathbb{R}^n\) of radius 1. What is the radius of a Euclidean ball of volume 1? What is its asymptotics for large \(n\)?
6 Important inequalities and notions of convergence

6.1 Basic probabilistic inequalities

One of the simplest and very useful probabilistic inequalities is a tail bound by expectation: the so-called Chebyshev’s inequality.

6.1 Theorem (Chebyshev’s inequality). If $X$ is a nonnegative random variable, then for every $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{1}{t} \mathbb{E}X.$$  

Proof. Since $X \geq X1_{\{X \geq t\}} \geq t1_{\{X \geq t\}}$, taking the expectation yields

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

There are several variants, easily deduced from Chebyshev’s inequality by monotonicity of certain functions. For a nonnegative random variable $X$ and $t > 0$, using the power function $x^p$, $p > 0$, we get

$$\mathbb{P}(X \geq t) = \mathbb{P}(X^p \geq t^p) \leq \frac{1}{t^p} \mathbb{E}X^p. \quad (6.1)$$

For a real-valued random variable $X$, every $t \in \mathbb{R}$ and $\lambda > 0$, using the exponential function $e^{\lambda x}$, we have

$$\mathbb{P}(X \geq t) = \mathbb{P}(\lambda X \geq \lambda t) \leq \frac{1}{e^{\lambda t}} \mathbb{E}e^{\lambda X}. \quad (6.2)$$

For a real-valued random variable $X$, every $t \in \mathbb{R}$, using the square function $x^2$ and variance, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{1}{t^2} \mathbb{E}|X - \mathbb{E}X|^2 = \frac{1}{t^2} \text{Var}(X). \quad (6.3)$$

Another general and helpful inequality is about convex functions. Recall that a function $f: \mathbb{R} \to \mathbb{R}$ is convex if $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for every $\lambda \in [0, 1]$ and $x, y \in \mathbb{R}$. By induction, this can be extended to

$$f \left( \sum_{i=1}^{n} \lambda_i x_i \right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

for every $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\sum_{i=1}^{n} \lambda_i = 1$ and every $x_1, \ldots, x_n \in \mathbb{R}$. The weights $\lambda_i$ can of course be interpreted in probabilistic terms: if $X$ is a random variable taking the value $x_i$ with probability $\lambda_i$, then $\sum \lambda_i x_i = \mathbb{E}X$, whereas $\sum \lambda_i f(x_i) = \mathbb{E}f(X)$, so we have

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

This generalises to arbitrary random variables and is called Jensen’s inequality.
6.2 Theorem (Jensen’s inequality). If \( f : \mathbb{R} \to \mathbb{R} \) is a convex function and \( X \) is a random variable such that both \( \mathbb{E}X \) and \( \mathbb{E}f(X) \) exist, then \[
f(\mathbb{E}X) \leq \mathbb{E}f(X).
\]

We shall present two proofs.

Proof 1. Suppose \( f \) is differentiable. Then by convexity, a tangent line at \( x_0 \) is below the graph, so \[
f(x) \geq f(x_0) + f'(x_0)(x - x_0)
\]
(which holds for every \( x_0 \) and \( x \)). We set \( x = X, \ x_0 = \mathbb{E}X \) and take the expectation of both sides to get

\[
\mathbb{E}f(X) \geq \mathbb{E}[f(\mathbb{E}X) + f'(\mathbb{E}X)(X - \mathbb{E}X)] = f(\mathbb{E}X) + f'(\mathbb{E}X)\mathbb{E}(X - \mathbb{E}X) = f(\mathbb{E}X).
\]

If \( f \) is not differentiable, this argument can be rescued by using the fact that convex functions have left and right derivatives defined everywhere (because the divided differences of convex functions are monotone).

Proof 2. Recall that a function is convex if and only if its epigraph is a convex set. By a separation type argument, this gives that the convex function is a pointwise supremum over a countable collection of linear functions. Specifically, let \( f : \mathbb{R} \to \mathbb{R} \) be a convex function and consider the family of linear functions with rational coefficients which are below \( f \),

\[
\mathcal{A} = \{ \ell : \mathbb{R} \to \mathbb{R}, \ \ell(x) = ax + b, \ a, b \in \mathbb{Q}, \ \ell \leq f \}.
\]

Then

\[
f(x) = \sup_{\ell \in \mathcal{A}} \ell(x), \quad x \in \mathbb{R}.
\]

Jensen’s inequality follows: for every \( \ell \in \mathcal{A} \), by linearity, \( \mathbb{E}\ell(X) = \ell(\mathbb{E}X) \), thus

\[
\mathbb{E}f(X) = \mathbb{E}\sup_{\ell \in \mathcal{A}} \ell(X) \geq \sup_{\ell \in \mathcal{A}} \mathbb{E}\ell(X) = \sup_{\ell \in \mathcal{A}} \ell(\mathbb{E}X) = f(\mathbb{E}X).
\]

The so-called Hölder’s inequality is a very effective tool used to factor out the expectation of a product.

6.3 Theorem (Hölder’s inequality). Let \( p, q > 1 \) be such that \( \frac{1}{p} + \frac{1}{q} = 1 \). For random variables \( X \) and \( Y \), we have

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

In particular, when \( p = q = 2 \), this gives the Cauchy-Schwarz inequality

\[
\mathbb{E}|XY| \leq \sqrt{\mathbb{E}|X|^2 \mathbb{E}|Y|^2}.
\]
Proof 1. We can assume without loss of generality that $E|X|^p$ and $E|Y|^q$ are finite (otherwise the right hand side is $+\infty$ and there is nothing to prove). The key ingredient is an elementary inequality for numbers.

Claim. For $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ and $x, y \geq 0$, we have

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$  

Proof. By the concavity of the log function, we have

$$\log \left( \frac{x^p}{p} + \frac{y^q}{q} \right) \geq \frac{1}{p} \log x^p + \frac{1}{q} \log y^q = \log xy.$$  

Setting $x = \frac{|X|^p}{(E|X|^p)^{1/p}}$, $y = \frac{|Y|^q}{(E|Y|^q)^{1/q}}$, taking the expectation and simplifying yields the desired inequality.  

Proof 2. By homogeneity we can assume that $E|X|^p = 1$ and $E|Y|^q = 1$. We can also assume that $|Y| > 0$ a.e. (otherwise we consider $\max\{|Y|, \frac{1}{q}\}$ and pass to the limit by Lebesgue’s monotone convergence theorem). Define a new probability measure $\tilde{\mathbb{P}}(A) = E|Y|^q 1_A$, $A \in \mathcal{F}$. In other words, $\hat{E}Z = E|Z|^q$ for every ($\tilde{\mathbb{P}}$-integrable) random variable $Z$. Then, by the convexity of $x \mapsto x^p$ and Jensen’s inequality,

$$\left( E|X||Y| \right)^p = \left( \hat{E}|X||Y|^{1-q} \right)^p \leq \hat{E}|X|^p|Y|^{(1-q)p} = E|X|^p|Y|^{(1-q)p+q} = E|X|^p = 1.$$  

6.2 $L_p$-spaces

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $p \in (0, \infty)$, we define

$$L_p = L_p(\Omega, \mathcal{F}, \mathbb{P}) = \{X: \Omega \to \mathbb{R}, \text{ $X$ is a random variable with $E|X|^p < \infty$}\}$$

which is called the $L_p$ space (on $\Omega$). Technically, $L_p$ is defined as the set of the abstract classes of random variables which are equal a.e., but we tacitly assume that and skip such details. We set

$$\|X\|_p = (E|X|^p)^{1/p}, \quad X \in L_p.$$  

We also extend this to $p = \infty$ by setting

$$L_\infty = \{X: \Omega \to \mathbb{R}, \text{ $X$ is a random variable with $|X| \leq M$ a.s., for some $M > 0$}\}$$

and

$$\|X\|_\infty = \text{ess sup}X = \inf\{M \geq 0, \text{ $|X| \leq M$ a.s.}\}.$$
the essential supremum of $X$) with the usual convention that $\inf \emptyset = +\infty$. Equivalently,

$$\|X\|_{\infty} = \inf \{ t \in \mathbb{R}, \ F_X(t) = 1 \}$$

(exercise). We also have

$$\|X\|_p \xrightarrow{p \to \infty} \|X\|_{\infty}$$

(another exercise). The quantity $\|X\|_p$ is called the $p$-th moment of $X$. It is monotone in $p$, which is an easy consequence of Jensen’s inequality.

6.4 Example. Let $0 < p < q$. Take $r = \frac{q}{p}$ and $f(x) = |x|^r$ which is convex. Thus for a random variable $X$ which is in $L_q$, using Jensen’s inequality, we have

$$E|X|^q = E f(|X|^p) \geq f(E|X|^p) = (E|X|^p)^{q/p},$$

equivalently,

$$\|X\|_q \geq \|X\|_p.$$

In other words, the function $p \mapsto \|X\|_p$ of moments of the random variable $X$ is nondecreasing.

6.5 Example. Hölder’s inequality can be restated as: for random variables $X$ and $Y$ and $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$E|XY| \leq \|X\|_p \|Y\|_q.$$  \hspace{1cm} (6.4)

The case $p = 1, q = \infty$ follows by taking the limit in Hölder’s inequality.

Hölder’s inequality gives the following helpful variational formula for $p$thm moments, $p \in [1, \infty]$.

6.6 Theorem. Let $p \in [1, \infty]$. For $X \in L_p$, we have

$$\|X\|_p = \sup \{ EXY, \ Y \text{ is a random variable with } E|Y|^q \leq 1 \},$$

de where $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. To see that the supremum does not exceed the $p$th moment, simply apply Theorem 6.3. To see the opposite inequality, consider $Y = \text{sgn}(X)|X|^{p-1}||X||_{p/q}^{-p/q}$. Then $EXY = \|X\|_p$, so in fact we can write “max ”instead of “sup ”in (6.5). Using this linearisation, we can effortlessly establish the triangle inequality for the $p$th moment, the so-called Minkowski’s inequality. \hfill \Box

6.7 Theorem (Minkowski’s inequality). Let $p \in [1, \infty]$. Let $X$ and $Y$ be random variables. Then

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$
Proof. Invoking (6.5),
\[ \|X + Y\|_p = \sup \{ \mathbb{E}(X + Y)Z \mid \mathbb{E}|Z|^p \leq 1 \}. \]
By linearity, \( \mathbb{E}(X + Y)Z = \mathbb{E}XZ + \mathbb{E}YZ \). Using that \( \sup \{ f + g \} \leq \sup f + \sup g \) and applying again (6.5) finishes the proof. \( \square \)

6.8 Remark. For every \( 0 < p < 1 \) Minkowski’s inequality fails (for instance, take \( X \) and \( Y \) to be i.i.d. \( \text{Ber}(\alpha) \)). Let us derive its analogue. Observe that for \( 0 < p < 1 \) and every real numbers \( x, y \), we have
\[ |x + y|^p \leq |x|^p + |y|^p. \] (6.6)
If \( x + y = 0 \), the inequality is trivial. Otherwise, note that \( |t|^p \geq |t| \) for \( |t| \leq 1 \), so using this and the triangle inequality yields
\[ \left( \frac{|x|}{|x + y|} \right)^p + \left( \frac{|y|}{|x + y|} \right)^p \geq \frac{|x|}{|x + y|} + \frac{|y|}{|x + y|} = \frac{|x| + |y|}{|x + y|} \geq \frac{|x + y|}{|x + y|} = 1. \]
Given two random variables, applying (6.6) for \( x = X(\omega), y = Y(\omega) \) and taking the expectation gives
\[ \mathbb{E}|X + Y|^p \leq \mathbb{E}|X|^p + \mathbb{E}|Y|^p, \quad p \in (0, 1]. \] (6.7)
In other words,
\[ \|X + Y\|^p \leq \|X\|^p + \|Y\|^p, \quad p \in (0, 1]. \] (6.8)

The next two theorems justify that \( L_p \) are in fact Banach spaces (normed spaces which are complete, that is every Cauchy sequence converges).

6.9 Theorem. For every \( p \in [1, \infty] \), \( (L_p, \| \cdot \|_p) \) is a normed space.

Proof. To check that \( X \mapsto \|X\|_p \) is a norm on \( L_p \), it is to be verified that
1) \( \|X\|_p \geq 0 \) with equality if and only if \( X = 0 \) a.s.
2) \( \|\lambda X\|_p = |\lambda| \|X\|_p \), for every \( \lambda \in \mathbb{R} \)
3) \( \|X + Y\|_p \leq \|X\|_p + \|Y\|_p \).

1) and 2) follow easily from the properties of integral and essential supremum. 3) follows from Minkowski’s inequality. \( \square \)

6.10 Theorem. Let \( p \in [1, \infty] \). If \( (X_n)_{n \geq 1} \) is a Cauchy sequence in \( L_p \), that is for every \( \varepsilon > 0 \), there is a positive integer \( N \) such that for every \( n, m \geq N \), we have \( \|X_n - X_m\|_p \leq \varepsilon \), then there is a random variable \( X \) in \( L_p \) such that \( \|X_n - X\|_p \to 0 \). In other words, \( (L_p, \| \cdot \|_p) \) is complete, hence it is Banach space.
Proof. Assume first that $1 \leq p < \infty$. By the Cauchy condition, there is a subsequence $n_k$ such that

$$
\|X_{n_{k+1}} - X_{n_k}\|_p \leq 2^{-k}, \quad k = 1, 2, \ldots
$$

Let

$$
Y_k = \sum_{j=1}^k |X_{n_{j+1}} - X_{n_j}|.
$$

The sequence $(Y_k)$ is nondecreasing, hence it pointwise converges, say to $Y$, $\lim_{k \to \infty} Y_k = Y$. Since $\|Y_k\|_p \leq 1$, by Fatou’s lemma,

$$
\mathbb{E} Y^p = \mathbb{E} \liminf_{k} Y_k^p \leq \liminf_{k} \mathbb{E} Y_k^p \leq 1,
$$

that is $Y \in L_p$. In particular, $Y < \infty$ a.s. Consequently, the sequence

$$
X_{n_k} = X_{n_1} + \sum_{j<k} (X_{n_{j+1}} - X_{n_j})
$$

converges a.s., say to $X$. It remains to show that $\|X_n - X\|_p \to 0$. For a fixed $m$, by Fatou’s lemma, we get

$$
\mathbb{E}|X_m - X|^p = \mathbb{E} \liminf_k |X_m - X_{n_k}|^p \leq \liminf_k \mathbb{E}|X_m - X_{n_k}|^p,
$$

thus by the Cauchy condition, for every $\varepsilon > 0$, there is $N$ such that for every $m > N$,

$$
\mathbb{E}|X_m - X|^p \leq \varepsilon.
$$

This finishes the argument.

For $p = \infty$, we consider the sets

$$
A_k = \{|X_k| > \|X_k\|_{\infty}\},
$$

$$
B_{n,m} = \{|X_m - X_n| > \|X_m - X_n\|_{\infty}\}.
$$

Their union $E$ is of measure zero, whereas on $E^c$, the variables $X_n$ converge uniformly to a bounded random variable $X$ (because $\mathbb{R}$ is complete).

The case $p = 2$ is the most important because $L_2$ is Hilbert space. The scalar product $\langle \cdot, \cdot \rangle: L_2 \times L_2 \to \mathbb{R}$ is defined by

$$
\langle X, Y \rangle = \mathbb{E}XY, \quad X, Y \in L_2.
$$

Then

$$
\|X\|_2 = \sqrt{\langle X, X \rangle}.
$$

Crucially, we have the parallelogram identity: for $X, Y \in L_2$, we have

$$
\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2(\|X\|_2^2 + \|Y\|_2^2), \quad (6.9)
$$

A consequence of this is that balls in $L_2$ are round and orthogonal projection is well defined.
6.11 Theorem. Let $H$ be a complete linear subspace of $L_2$. Then for every random variable $X$ in $L_2$, there is a unique random variable $Y \in H$ such that the following two conditions hold

(i) $Y$ is closest to $X$ in $H$, that is

$$\|X - Y\|_2 = \inf\{\|X - Z\|_2, Z \in H\},$$

(ii) for every $Z \in H$, $\langle X - Y, Z \rangle = 0$, that is $X - Y$ is orthogonal to $H$.

The uniqueness is understood as follows: if $\tilde{Y} \in H$ satisfies either (i) or (ii), then $\|Y - \tilde{Y}\|_2 = 0$, that is $Y = \tilde{Y}$ a.s.

Proof. Let $d$ denote the infimum in (i). Then, there are $Y_n \in H$ such that $\|X - Y_n\|_2 \to d$. By the parallelogram law,

$$\|X - Y_n\|_2^2 + \|X - Y_m\|_2^2 = 2 \left( \|X - \frac{Y_n + Y_m}{2}\|_2^2 + \|\frac{Y_n - Y_m}{2}\|_2^2 \right) \geq 2d + \|Y_n - Y_m\|_2.$$

Since the left hand side converges to $2d$ as $m, n \to \infty$, we conclude that $(Y_n)$ is a Cauchy sequence in $H$. Since $H$ is assumed to be complete, $\|Y_n - Y\|_2 \to 0$ for some $Y \in H$. Thus, $\|X - Y\|_2 = d$, which establishes (i).

To get (ii), fix $Z \in H$ and note that for every $t \in \mathbb{R}$, by (i), we have

$$\|X - (Y + tZ)\|_2 \geq \|X - Y\|_2,$$

which after squaring and rearranging gives

$$t^2 \|Z\|_2^2 - 2t \langle X - Y, Z \rangle \geq 0.$$

Since this holds for all small $t$ (both positive and negative), necessarily the linear term has to vanish, that is $\langle X - Y, Z \rangle = 0$.

For the uniqueness, suppose $\tilde{Y}$ satisfies (i). Then, by the parallelogram law,

$$2d = \|X - Y\|_2^2 + \|X - \tilde{Y}\|_2^2 = 2 \left( \|X - \frac{Y + \tilde{Y}}{2}\|_2^2 + \|\frac{Y - \tilde{Y}}{2}\|_2^2 \right) \geq 2d + \|Y - \tilde{Y}\|_2^2,$$

so $\|Y - \tilde{Y}\|_2^2 \leq 0$, hence $\|Y - \tilde{Y}\|_2 = 0$ and consequently, $\tilde{Y} = Y$ a.s. If $\tilde{Y}$ satisfies (ii), then since $\tilde{Y} - Y \in H$, we get $\langle X - \tilde{Y}, \tilde{Y} - Y \rangle = 0$. Since also $\langle X - Y, \tilde{Y} - Y \rangle = 0$, we get $\langle \tilde{Y} - Y, \tilde{Y} - Y \rangle = 0$, so $\tilde{Y} = Y$ a.s. \qed
6.3 Notions of convergence

A sequence of random variables \((X_n)\) converges to a random variable \(X\)

a) **almost surely** if \(\mathbb{P} \left( \{\omega \in \Omega, \lim_{n \to \infty} X_n(\omega) = X(\omega) \} \right) = 1\), denoted \(X_n \xrightarrow{a.s.} n \to \infty X\)

b) **in probability** if for every \(\varepsilon > 0\), \(\mathbb{P} \left( |X_n - X| > \varepsilon \right) \xrightarrow{n \to \infty} 0\), denoted \(X_n \xrightarrow{P} n \to \infty 0\)

c) **in \(L_p\)**, \(p > 0\), if \(\mathbb{E} |X_n - X|^p \xrightarrow{n \to \infty} 0\), denoted \(X_n \xrightarrow{L_p} n \to \infty X\).

For instance, let \(\Omega = \{1, 2\}\) and \(\mathbb{P}(1) = \mathbb{P}(2) = \frac{1}{2}\), \(X_n(1) = -\frac{1}{n}\), \(X_n(2) = \frac{1}{n}\).

We have

a) \(X_n \xrightarrow{a.s.} n \to \infty 0\) because \(X_n(\omega) \to 0\) for every \(\omega \in \Omega\),

b) \(X_n \xrightarrow{P} n \to \infty 0\) because \(\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P} \left( \frac{1}{n} > \varepsilon \right) \to 0\),

c) \(X_n \xrightarrow{L_p} n \to \infty 0\) because \(\mathbb{E}|X_n|^p = 2 \frac{1}{n^p} \to 0\).

We have two results, saying that the convergence in probability is the weakest among the three.

6.12 Theorem. **If a sequence of random variables \((X_n)\) converges to \(X\) a.s. then it also converges in probability, but in general not conversely.**

Proof. By the definition of the limit of a sequence,

\[
\{ \lim_n X_n = X \} = \bigcap_{l \geq 1} \bigcup_{N \geq 1} \bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\}.
\]

For any events \(A_l\), \(\mathbb{P}\left( \bigcap_{l \geq 1} A_l \right) = 1\) if and only if \(\mathbb{P}(A_l) = 1\) for all \(l \geq 1\). Therefore, \(X_n \xrightarrow{a.s.} n \to \infty 0\) is equivalent to: for every \(l \geq 1\),

\[
\mathbb{P}\left( \bigcup_{N \geq 1} \bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\} \right) = 1.
\]

By monotonicity with respect to \(N\),

\[
\mathbb{P}\left( \bigcup_{N \geq 1} \bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\} \right) = \lim_{N \to \infty} \mathbb{P}\left( \bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\} \right).
\]

Finally, observe that by the inclusion \(\bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\} \subseteq \left\{ |X_N - X| < \frac{1}{l} \right\}\), we have

\[
1 = \lim_{N \to \infty} \mathbb{P}\left( \bigcap_{n \geq N} \left\{ |X_n - X| < \frac{1}{l} \right\} \right) \leq \lim_{N \to \infty} \mathbb{P}\left( \left\{ |X_N - X| < \frac{1}{l} \right\} \right),
\]

56
so passing to the complements, for every \( l \geq 1 \),

\[
0 \leq \lim_{N \to \infty} \mathbb{P} \left( \left\{ |X_N - X| \geq \frac{1}{l} \right\} \right) \leq 0.
\]

Therefore, for every \( \varepsilon > 0 \), \( \lim_{N \to \infty} \mathbb{P} (|X_N - X| \geq \varepsilon) = 0 \), that is \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} 0 \). The following example of a sequence convergent in probability but not a.s. finishes the proof.

### 6.13 Example

Let \( \Omega = [0, 1] \) and \( \mathbb{P} (\cdot) \) be the uniform probability measure. Let

\[
X_1 = 1, \quad X_2 = 1_{[0, 1/2]}, \quad X_3 = 1_{[1/2, 1]}, \quad X_4 = 1_{[1/4, 1/2]}, \quad X_5 = 1_{[1/4, 1/2]}, \quad X_6 = 1_{[1/2, 3/4]}, \quad X_7 = 1_{[3/4, 1]}, \quad \text{etc.,} \quad X_{2^n}, X_{2^n + 1}, \ldots, X_{2^{n+1}-1}
\]

are indicators of a wandering interval of length \( 2^{-n} \) shifting to right by \( 2^{-n} \) every increment of the index. We have

a) \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} 0 \) because for every \( \varepsilon > 0 \), \( \mathbb{P} (|X_n| > \varepsilon) \leq 2^{-k} \) when \( 2^k \leq n < 2^{k+1} \), which goes to 0 as \( n \) goes to \( \infty \).

b) \( X_n \xrightarrow{n \to \infty \text{ a.s.}} X \) because for every \( \omega \in (0, 1) \), the sequence \((X_n(\omega))\) contains infinitely many 0 and 1, so it is not convergent; moreover, if \( X_n \xrightarrow{n \to \infty \text{ a.s.}} X \) for some random variable \( X \) other than 0, then by Theorem 6.12, \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} X \) and from the uniqueness of limits in probability (homework!), \( X = 0 \) a.s., contradiction.

c) \( X_n \xrightarrow[n \to \infty]{L_p} X \) because \( \mathbb{E}|X_n|^p = 2^{-kp} \) when \( 2^k \leq n < 2^{k+1} \), which goes to 0 as \( n \) goes to \( \infty \).

\[ \square \]

### 6.14 Theorem

If a sequence of random variables \((X_n)\) converges to \( X \) in \( L_p \) for some \( p > 0 \), then it also converges in probability, but in general not conversely.

**Proof.** By Chebyshev’s inequality (6.1),

\[
\mathbb{P} (|X_n - X| > \varepsilon) \leq \frac{1}{\varepsilon^p} \mathbb{E}|X_n - X|^p \xrightarrow[n \to \infty]{\mathbb{P}} 0,
\]

so \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} X \). The following example of a sequence convergent in probability but not in \( L_p \) finishes the proof. \[ \square \]

### 6.15 Example

Let \( \Omega = [0, 1] \) and \( \mathbb{P} (\cdot) \) be the uniform probability measure. Let

\[
X_n = n^{1/p} 1_{[0, 1/n]}.
\]

We have

a) \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} 0 \) because for every \( \varepsilon > 0 \), \( \mathbb{P} (|X_n| > \varepsilon) \leq \frac{1}{\varepsilon} \) which goes to 0 as \( n \) goes to \( \infty \).

b) \( X_n \xrightarrow[n \to \infty]{L_p} X \) because \( \mathbb{E}|X_n|^p = n^{1/p} = 1 \); moreover, if \( X_n \xrightarrow[n \to \infty]{L_p} X \) for some random variable \( X \) other than 0, then by Theorem 6.14, \( X_n \xrightarrow[n \to \infty]{\mathbb{P}} X \) and from the uniqueness of limits in probability (homework!), \( X = 0 \) a.s., contradiction.
c) \( X_n \overset{a.s.}{\longrightarrow} 0 \) because for every \( \omega > 0 \), the sequence \( X_n(\omega) \) becomes eventually constant 0.

Theorems (6.12), (6.14) and Examples 6.13, 6.15 can be summarised in the following diagram.

![Diagram]

We record a few basic algebraic properties of the three notions of convergence.

1) If \( X_n \) converges to \( X \) a.s./in probability/in \( L_p \) and \( Y_n \) converges to \( Y \) a.s./in probability/in \( L_p \), then \( X_n + Y_n \) converges to \( X + Y \) a.s./in probability/in \( L_p \).

2) If \( X_n \) converges to \( X \) a.s./in probability and \( Y_n \) converges to \( Y \) a.s./in probability, then \( X_n \cdot Y_n \) converges to \( X \cdot Y \) a.s./in probability.

3) If \( 0 < p < q \) and \( X_n \) converges to \( X \) in \( L_q \), then \( X_n \) converges to \( X \) in \( L_p \).

Immediately, 1) and 2) for the almost sure convergence follow from those statements for sequences of numbers since the intersection of two events of probability 1 is of probability 1.

Property 1) for \( L_p \) convergence follows from Minkowski’s inequality (Theorem 6.7) and Property 3) follows from the monotonicity of moments (Example 6.4).

Establishing 1) and 2) directly from definition is cumbersome. Instead, we first prove a convenient equivalent condition for convergence in probability in terms of almost sure convergence.

**6.16 Theorem (Riesz).** If a sequence \((X_n)\) of random variables converges to a random variable \(X\) in probability, then there is a subsequence \((X_{n_k})_k\) which converges to \(X\) almost surely.

**Proof.** Since for every \( \varepsilon \), \( \mathbb{P}(|X_n - X| > \varepsilon) \to 0 \), then we can find an index \( n_1 \) such that \( \mathbb{P}(|X_n - X| > 2^{-1}) < 2^{-1} \). By the same logic, we can find an index \( n_2 > n_1 \) such that \( \mathbb{P}(|X_{n_2} - X| > 2^{-2}) < 2^{-2} \), etc. We get a subsequence \((X_{n_k})_k\) such that \( \mathbb{P}(|X_{n_k} - X| > 2^{-k}) < 2^{-k} \) for every \( k \). Since the series \( \sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| > 2^{-k}) \) converges, by the first Borel-Cantelli lemma (Lemma 3.14), with probability 1 only finitely many events \( A_k = \{|X_{n_k} - X| > 2^{-k}\} \) occur. When this happens, \( X_{n_k} \to X \), so \( X_{n_k} \overset{k \to \infty}{\longrightarrow} X \).
### 6.17 Theorem

A sequence \((X_n)\) of random variables converges to a random variable \(X\) in probability if and only if every subsequence \((X_{n_k})_k\) contains a further subsequence \((X_{n_{k_l}})_{l}\) which converges to \(X\) almost surely.

**Proof.** \((\Rightarrow)\) It follows directly from Theorem 6.16.

\((\Leftarrow)\) If \((X_n)\) does not converge to \(X\) in probability, then there is \(\varepsilon > 0\) such that \(\mathbb{P}(|X_n - X| > \varepsilon) \not\to 0\). Consequently, there is \(\varepsilon' > 0\) and a subsequence \((X_{n_k})\) for which \(\mathbb{P}(|X_{n_k} - X| > \varepsilon) > \varepsilon'\). By the assumption, there is a subsequence \((X_{n_{k_l}})_{l}\) convergent to \(X\) almost surely, in particular, in probability, so \(\mathbb{P}(|X_{n_{k_l}} - X| > \varepsilon) \to 0\). This contradiction finishes the proof.

Going back to the algebraic properties 1) and 2) for convergence in probability, we can easily justify them using that they hold for convergence almost surely. For 1), say \(S_n = X_n + Y_n\) does not converge in probability to \(S = X + Y\). Then as in the proof of Theorem 6.17, \(\mathbb{P}(|S_{n_k} - S| > \varepsilon) > \varepsilon'\) for some \(\varepsilon, \varepsilon' > 0\) and a subsequence \((n_k)\). Using Theorem 6.17, there is a further subsequence \((n_{k_l})\) such that \((X_{n_{k_l}})_{l}\) converges to \(X\) a.s. and a further subsequence (for simplicity, denote it the same) such that \((Y_{n_{k_l}})_{l}\) converges to \(Y\) a.s. Then \(S_{n_{k_l}} \xrightarrow{a.s.} S\), which contradicts \(\mathbb{P}(|S_{n_k} - S| > \varepsilon) > \varepsilon'\).
6.4 Exercises

1. Show that the probability that in \( n \) throws of a fair die the number of sixes lies between \( \frac{1}{3} n - \sqrt{n} \) and \( \frac{1}{3} n + \sqrt{n} \) is at least \( \frac{31}{36} \).

2. Let \( X \) be a random variable with density \( \frac{1}{2} e^{-|x|} \) on \( \mathbb{R} \). Show that for every \( p \geq 1 \),
   \[ c_1 p \leq \|X\|_p \leq c_2 p \]
   for some absolute constants \( c_1, c_2 > 0 \).

3. Let \( g \) be a standard Gaussian random variable. Show that for every \( p \geq 1 \),
   \[ c_1 p \leq \|g\|_p \leq c_2 p \]
   for some universal constants \( c_1, c_2 > 0 \).

4. Show that for every random variable \( X \), we have
   \[ \|X\|_{\infty} = \inf \{ t \in \mathbb{R}, F_X(t) = 1 \} \]

5. Show that for every random variable \( X \), we have
   \[ \|X\|_p \longrightarrow_{p \to \infty} \|X\|_{\infty} \]

6. If \( E|X|^{p_0} < \infty \) for some \( p_0 > 0 \), then \( E \log |X| \) exists and
   \[ (E|X|^p)^{1/p} \longrightarrow_{p \to 0^+} e^{E \log |X|} \]
   (thus it makes sense to define the 0th moment as \( \|X\|_0 = e^{E \log |X|} \)).

7. Let \( X \) be a random variable with values in an interval \([0, a]\). Show that for every \( t \)
   in this interval, we have
   \[ P(X \geq t) \geq \frac{E X - t}{a - t} \]

8. Prove the Payley-Zygmund inequality: for a nonnegative random variable \( X \) and every \( \theta \in [0, 1] \), we have
   \[ P(X > \theta E X) \geq (1 - \theta)^2 \frac{(E X)^2}{E X^2} \]

9. Prove that for nonnegative random variables \( X \) and \( Y \), we have
   \[ E \frac{X}{Y} \geq \frac{(E \sqrt{X})^2}{E Y} \]

10. Let \( p \in (0, 1) \) and \( q < 0 \) be such that \( \frac{1}{p} + \frac{1}{q} = 1 \). Then for every random variables \( X \) and \( Y \), we have
    \[ E|X Y| \geq (E|X|^p)^{1/p}(E|Y|^q)^{1/q} \]

11. Let \( X_1, X_2, \ldots \) be i.i.d. positive random variables with \( E X_1^3 < \infty \). Let
    \[ a_n = E \left( \frac{X_1 + \ldots + X_n}{n} \right)^3 \]
    Prove that \( a_n^2 \leq a_{n-1} a_{n+1}, n \geq 2 \).
12. Let $X, X_1, X_2, \ldots$ be identically distributed random variables such that $P(X > t) > 0$ for every $t > 0$. Suppose that for every $\eta > 1$, we have $\lim_{t \to \infty} \frac{P(X > \eta t)}{P(X > t)} = 0$. For $n \geq 1$, let $a_n$ be the smallest number $a$ such that $nP(X > a) \leq 1$. Show that for every $\varepsilon > 0$, we have $\max_{1 \leq i \leq n} X_i \leq (1 + \varepsilon)a_n$ with high probability as $n \to \infty$, i.e. $P(\max_{1 \leq i \leq n} X_i \leq (1 + \varepsilon)a_n) \to 1$.

13. Let $X$ be a random variable such that $Ee^{\delta |X|} < \infty$ for some $\delta > 0$. Show that $E|X|^p < \infty$ for every $p > 0$.

14. Let $X$ be a random variable such that $Ee^{tX} < \infty$ for every $t \in \mathbb{R}$. Show that the function $t \mapsto \log Ee^{tX}$ is convex on $\mathbb{R}$.

15. Let $X$ be a random variable such that $E|X|^p < \infty$ for every $p > 0$. Show that the function $p \mapsto \log \|X\|_p$ is convex on $(0, \infty)$.

16. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent random signs, that is $P(\varepsilon_i = -1) = \frac{1}{2} = P(\varepsilon_i = 1)$, $i \leq n$. Prove that there is a positive constant $c$ such that for every $n \geq 1$ and real numbers $a_1, \ldots, a_n$, we have

$$P \left( \left| \sum_{i=1}^{n} a_i \varepsilon_i \right| > \frac{1}{2} \sqrt{\sum_{i=1}^{n} a_i^2} \right) \geq c.$$

17. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs. Show that there is a constant $c > 0$ such that for every $n \geq 1$ and reals $a_1, \ldots, a_n$, we have

$$P \left( \left| \sum_{i=1}^{n} a_i \varepsilon_i \right| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \right) \geq c.$$

18. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs. Show that there is a constant $c > 0$ such that for every $n \geq 1$ and reals $a_1, \ldots, a_n$, we have

$$P \left( \left| \sum_{i=1}^{n} a_i \varepsilon_i \right| \geq \sqrt{\sum_{i=1}^{n} a_i^2} \right) \geq c.$$

19. The goal is to prove Bernstein's inequality: for every $n$, every real numbers $a_1, \ldots, a_n$ and $t > 0$, we have

$$P \left( \left| \sum_{i=1}^{n} a_i \varepsilon_i \right| > t \right) \leq 2 \exp \left\{ - \frac{t^2}{2 \sum_{i=1}^{n} a_i^2} \right\},$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. symmetric random signs.

a) Show that $\cosh(t) \leq e^{t^2/2}$, $t \in \mathbb{R}$.

b) Find $Ee^{a_i \varepsilon_i}$.

c) Let $S = \sum a_i \varepsilon_i$. Show that for every $t, \lambda > 0$, we have $P(S > t) \leq e^{-\lambda S} e^{\lambda S}$. 61
d) Optimising over $\lambda$ conclude that $\mathbb{P}(S > t) \leq e^{-t^2/(2 \sum a_i^2)}$.
e) Using symmetry, conclude that $\mathbb{P}(|S| > t) \leq 2e^{-t^2/(2 \sum a_i^2)}$.

20. Hoeffding’s lemma: for a random variable $X$ such that $a \leq X \leq b$ a.s. for some $a < b$, we have $\mathbb{E} e^{u(X - \mathbb{E}X)} \leq \exp\left\{\frac{u^2(b-a)^2}{8}\right\}$, $u \in \mathbb{R}$.

21. Hoeffding’s inequality: for independent random variables $X_1, \ldots, X_n$ such that $a_i \leq X_i \leq b_i$ a.s. for some $a_i < b_i$, $i \leq n$, we have
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i\right| > t\right) \leq 2 \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}, \quad t \geq 0.$$

22. Khinchin’s inequality: for every $p > 0$, there are positive constants $A_p, B_p$ which depend only on $p$ such that for every $n$ and every real numbers $a_1, \ldots, a_n$, we have
$$A_p \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_{i=1}^{n} a_i \varepsilon_i\right|^p\right)^{1/p} \leq B_p \left(\sum_{i=1}^{n} a_i^2\right)^{1/2},$$
where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. symmetric random signs.

23. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs. Show that
$$\mathbb{P}\left(\limsup_{n \to \infty} \frac{\varepsilon_1 + \ldots + \varepsilon_n}{\sqrt{2n \log n}} \leq 1\right) = 1.$$

24. Let $X$ be an integrable random variable and define
$$X_n = \begin{cases} -n, & X < -n \\ X, & |X| \leq n \\ n, & X > n. \end{cases}$$

Does the sequence $X_n$ converge a.s., in $L_1$, in probability?

25. Show that if $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, then $\mathbb{P}(X = Y) = 1$ (in other words, the limit in probability is unique).

26. Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables. Prove that $\frac{1}{n} \max_{k \leq n} |X_k|$ converges to 0 in probability.

27. Show that if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_nY_n \xrightarrow{p} XY$.

28. Prove that a sequence of random variables $X_n$ converges a.s. if and only if for every $\varepsilon > 0$, $\lim_{N \to \infty} \mathbb{P}\left(\cap_{n, m \geq N} |X_n - X_m| < \varepsilon\right) = 1$ (the Cauchy condition).

29. Does a sequence of independent random signs $\varepsilon_1, \varepsilon_2, \ldots$ converge a.s.?

30. Let $X_1, X_2, \ldots$ be independent random variables, $X_n \sim \text{Pois}(1/n)$. Does the sequence $X_n$ converge a.s., in $L_1$, in $L_2$, in probability?
31. Show that if for every \( \delta > 0 \) we have \( \sum_{n=1}^{\infty} P(|X_n - X| > \delta) < \infty \), then \( X_n \xrightarrow{n \to \infty} a.s. \ X \).

32. Show that if there is a sequence of positive numbers \( \delta_n \) convergent to 0 such that \( \sum_{n=1}^{\infty} P(|X_n - X| > \delta_n) < \infty \), then \( X_n \xrightarrow{n \to \infty} a.s. \ X \).

33. Let \( X_1, X_2, \ldots \) be i.i.d. random variables such that \( P(|X_i| < 1) = 1 \). Show that \( X_1 X_2 \cdots X_n \) converges to 0 a.s. and in \( L_1 \).

34. Let \( V \) be the linear space of all random variables on a probability space \( (\Omega, F, P) \) (two random variables are considered equal if they are equal a.s.). Define \( \rho: V \times V \to \mathbb{R} \),

\[
\rho(X, Y) = E\frac{|X - Y|}{1 + |X - Y|}.
\]

Show that this a metric on \( V \), \( (V, \rho) \) is complete and \( X_n \xrightarrow{p \text{-} n \to \infty} \ X \) if and only if \( \rho(X_n, X) \to 0 \).

35. Let \( (\Omega, F, P) \) be a discrete probability space. Show that for every sequence of random variables \( (X_n) \) on this space, \( X_n \xrightarrow{p \text{-} n \to \infty} \ X \) if and only if \( X_n \xrightarrow{a.s. \, n \to \infty} \ X \).

36. Show that in general almost sure convergence is not metrisable.
7 Laws of large numbers

Suppose we roll a die $n$ times and the outcomes are $X_1, X_2, \ldots, X_n$. We expect that the average $\frac{X_1 + \ldots + X_n}{n}$ should be approximately 3.5 (the expectation of $X_1$) as $n$ becomes large. Laws of large numbers establish that rigorously, in a fairly general situation.

Formally, we say that a sequence of random variables $X_1, X_2, \ldots$ satisfies the weak law of large numbers if $\frac{X_1 + \ldots + X_n}{n} - \mathbb{E} \frac{X_1 + \ldots + X_n}{n}$ converges to 0 in probability and the sequence satisfies the strong law of large numbers if the convergence is almost sure. In particular, for a sequence of identically distributed random variables, we ask whether $\frac{X_1 + \ldots + X_n}{n} \rightarrow \mathbb{E} X_1$. Consider two examples when no reasonable law of large numbers holds and the opposite.

7.1 Example. Let $X_1, X_2, \ldots$ be i.i.d. standard Cauchy random variables. Then it can be checked that $\bar{S}_n = \frac{X_1 + \ldots + X_n}{n}$ has the same distribution as $X_1$, so $\bar{S}_n$ is a “well spread out” random variable which in no reasonable sense should be close to its expectation (which in fact does not exists!), or any other constant.

7.2 Example. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs, that is $P(\varepsilon_i = \pm 1) = \frac{1}{2}$. Let $\bar{S}_n = \frac{\varepsilon_1 + \ldots + \varepsilon_n}{n}$. By Bernstein’s inequality (Exercise 6.19), $P(|\bar{S}_n| > t) \leq 2e^{-nt^2/2}$, so the series $\sum_{n=1}^{\infty} P(|\bar{S}_n| > t)$ converges, so $\bar{S}_n \xrightarrow{a.s.} 0 = \mathbb{E} \varepsilon_1$ (check!). In other words, the sequence $(\varepsilon_n)$ satisfies the strong law of large numbers.

7.1 Weak law of large numbers

Using the second moment, we can easily get a very simple version of the weak law of large numbers for uncorrelated random variables with uniformly bounded variance.

7.3 Theorem (The $L_2$ law of large numbers). Let $X_1, X_2, \ldots$ be random variables such that $\mathbb{E}|X_i|^2 < \infty$ for every $i$. If

$$\frac{1}{n^2} \text{Var}(X_1 + \ldots + X_n) \xrightarrow{n \to \infty} 0,$$

then denoting $S_n = X_1 + \ldots + X_n$,

$$\frac{S_n}{n} - \mathbb{E} \frac{S_n}{n} \xrightarrow{L_2} 0.$$

In particular, this holds when the $X_i$ are uncorrelated with bounded variance, that is $\text{Var}(X_i) \leq M$ for every $i$ for some $M$.

Proof. We have

$$\mathbb{E} \left| \frac{S_n}{n} - \mathbb{E} \frac{S_n}{n} \right|^2 = \frac{1}{n^2} \mathbb{E}|S_n - \mathbb{E}S_n|^2 = \frac{1}{n^2} \text{Var}(X_1 + \ldots + X_n) \xrightarrow{n \to \infty} 0.$$

Since

$$\text{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j),$$

64
when the $X_i$ are uncorrelated with bounded variance, we have
\[
\frac{1}{n^2} \text{Var}(X_1 + \ldots + X_n) \leq \frac{Mn}{n^2} = \frac{M}{n}
\]
which goes to 0 as $n \to \infty$. \hfill \square

Since convergence in $L_2$ implies convergence in probability, the above is in fact stronger then a weak law of large numbers.

7.4 Example. Let $X$ be a random vector in $\mathbb{R}^n$ uniformly distributed on the cube $[-1,1]^n$, that is $X = (X_1, \ldots, X_n)$ with the $X_i$ being i.i.d. uniform on $[-1,1]$. The assumptions of the above $L_2$ law of large numbers are satisfied for $X_1^2, X_2^2, \ldots$, so in particular
\[
\frac{X_1^2 + \ldots + X_n^2}{n} - \mathbb{E}X_1^2 \xrightarrow{p} 0
\]
Note that $\mathbb{E}X_1^2 = \frac{1}{3}$. By definition, this convergence in probability means that for every $\varepsilon > 0$,
\[
\mathbb{P} \left( \left| \frac{X_1^2 + \ldots + X_n^2}{n} - \frac{1}{3} \right| > \varepsilon \right) \xrightarrow{n \to \infty} 0,
\]
or equivalently,
\[
\mathbb{P} \left( \sqrt{n(1/3 - \varepsilon)} < \sqrt{X_1^2 + \ldots + X_n^2} < \sqrt{n(1/3 + \varepsilon)} \right) \xrightarrow{n \to \infty} 1.
\]
In words, a random point in a high dimensional cube is typically near the boundary of the Euclidean ball centered at 0 of radius $\sqrt{n/3}$.

7.5 Example. Let $X_1, X_2, \ldots$ be i.i.d. random variables uniform on $\{1, \ldots, n\}$. For $k \geq 1$, let
\[
\tau_k = \inf \{ m \geq 1, |\{X_1, \ldots, X_m\}| = k \}.
\]
This random variable can be though of as the first index (time) when we have collected $k$ coupons if the $X_i$ are though of as coupons given to us one by one and selected uniformly at random (with replacement) among $n$ different coupons. We are interested in the behaviour of $\tau_n$ as $n \to \infty$ (the time needed to collect the entire set of $n$ coupons). For convenience we set $\tau_0 = 0$ and of course $\tau_1 = 1$. Let
\[
T_k = \tau_k - \tau_{k-1}, \quad k \geq 1,
\]
which is time we wait to get a coupon of a next type after we have collected $k-1$ different coupons. We have,
\[
\mathbb{P} (T_k = l) = \left( \frac{k-1}{n} \right)^{l-1} \left( 1 - \frac{k-1}{n} \right), \quad l = 1, 2, \ldots,
\]
that is
\[
T_k \sim \text{Geom} \left( 1 - \frac{k-1}{n} \right)
\]
and $T_1, \ldots, T_n$ are independent. Plainly,
\[ \tau_n = T_1 + \ldots + T_n. \]

Thus
\[ \mathbb{E} \tau_n = \sum_{k=1}^{n} \mathbb{E} T_k = \sum_{k=1}^{n} \left( 1 - \frac{k-1}{n} \right) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k}, \]
so for large $n$, we have $\mathbb{E} \tau_n \sim n \log n$. Moreover, thanks to independence,
\[ \text{Var}(\tau_n) = \sum_{k=1}^{n} \text{Var}(T_k) \leq \sum_{k=1}^{n} \left( 1 - \frac{k-1}{n} \right)^{-2} = n^2 \sum_{k=1}^{n} \frac{1}{k^2} < 2n^2. \]

If we let
\[ t_n = \frac{\tau_n}{n \log n}, \]
we obtain
\[ \mathbb{E}(t_n - 1)^2 = \frac{1}{n^2 \log^2 n} \mathbb{E}(\tau_n - n \log n)^2 \leq \frac{2}{n^2 \log^2 n} \left( \mathbb{E}(\tau_n - \mathbb{E} \tau_n)^2 + (\mathbb{E} \tau_n - n \log n)^2 \right) \leq \frac{4}{\log^2 n} + \frac{2}{\log^2 n}. \]

This gives that $t_n \to 1$ in $L_2$ and in probability.

Our goal is to prove the weak law of large numbers for i.i.d. sequences under optimal assumptions on integrability.

7.6 Theorem (The weak law of large numbers). If $X_1, X_2, \ldots$ are i.i.d. random variables such that
\[ t \mathbb{P}(|X_1| > t) \xrightarrow{t \to \infty} 0, \quad (7.1) \]
then
\[ \frac{X_1 + \ldots + X_n}{n} - \mu_n \xrightarrow{p \to \infty} 0, \quad (7.2) \]
where $\mu_n = \mathbb{E} X_1 1_{(|X_1| \leq n)}$.

7.7 Remark. The assumption is optimal in the following sense: condition (7.1) is necessary for existence of a sequence $a_n$ such that
\[ \frac{X_1 + \ldots + X_n}{n} - a_n \xrightarrow{p \to \infty} 0 \]
(see exercises).

To prove the theorem, we first establish a fairly general lemma.
7.2 Exercises

1.
A Appendix: Carathéodory’s theorem

Our goal here is to give a proof of Carathéodory’s theorem about extensions of measures.

A.1 Theorem (Carathéodory). Let \( \Omega \) be a set and let \( \mathcal{A} \) be an algebra on \( \Omega \). Suppose a function \( \mathbb{P} : \mathcal{A} \to [0, +\infty) \) satisfies

(i) \( \mathbb{P}(\Omega) = 1 \),

(ii) \( \mathbb{P} \) is finitely additive, that is for every \( A_1, \ldots, A_n \in \mathcal{A} \) which are pairwise disjoint, we have

\[
\mathbb{P}\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} \mathbb{P}(A_i),
\]

(iii) for every \( A_1, A_2, \ldots \in \mathcal{A} \) with \( A_1 \subset A_2 \subset \cdots \) such that \( A = \bigcup_{n=1}^{\infty} A_n \) is in \( \mathcal{A} \), we have

\[
\lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}(A).
\]

Then \( \mathbb{P} \) can be uniquely extended to a probability measure on the \( \sigma \)-algebra \( \mathcal{F} = \sigma(\mathcal{A}) \) generated by \( \mathcal{A} \).

Proof. We break the proof into 3 steps.

I. We define a nonnegative function \( \mathbb{P}^* \) on all subsets of \( \Omega \) satisfying: \( \mathbb{P}^*(\Omega) = 1 \), \( \mathbb{P}^* \) is monotone and subadditive (the so-called exterior or outer measure).

II. We define a family of subsets \( \mathcal{M} \) of \( \Omega \) which is a \( \sigma \)-algebra and \( \mathbb{P}^* \) is countably-additive on \( \mathcal{M} \).

III. We show that \( \mathbb{P}^* \) agrees with \( \mathbb{P} \) on \( \mathcal{M} \) and that \( \mathcal{M} \) contains \( \mathcal{A} \).

We proceed with proving the steps I, II, III. Then we argue about the uniqueness.

I. For a subset \( A \) of \( \Omega \), we define

\[
\mathbb{P}^*(A) = \inf \sum_{n} \mathbb{P}(A_n),
\]

where the infimum is taken over all sets \( A_1, A_2, \ldots \in \mathcal{A} \) such that \( \bigcup_{n} A_n \supset A \).

Clearly, \( \mathbb{P}^* \) is nonnegative. Since \( \emptyset \in \mathcal{A} \), we have \( \mathbb{P}^*(\emptyset) = 0 \). It is also clear that \( \mathbb{P}^* \) is monotone, that is if \( A \subset B \), then \( \mathbb{P}^*(A) \leq \mathbb{P}^*(B) \). Finally, we show that \( \mathbb{P}^* \) is subadditive, that is for every sets \( A_1, A_2, \ldots \), we have

\[
\mathbb{P}^* \left( \bigcup_{n} A_n \right) \leq \sum_{n} \mathbb{P}^*(A_n).
\]

Indeed, by the definition of \( \mathbb{P}^* \), for \( \varepsilon > 0 \), there are sets \( B_{n,k} \in \mathcal{A} \) such that \( A_n \subset \bigcup_{k} B_{n,k} \) and \( \sum_{k} \mathbb{P}(B_{n,k}) < \mathbb{P}^*(A_n) + \varepsilon 2^{-n} \). Then \( \bigcup_{n} A_n \subset \bigcup_{n,k} B_{n,k} \) and consequently,

\[
\mathbb{P}^* \left( \bigcup_{n} A_n \right) \leq \sum_{n,k} \mathbb{P}(B_{n,k}) < \sum_{n} \mathbb{P}^*(A_n) + \varepsilon.
\]
Since $\varepsilon > 0$ is arbitrary, we get the desired inequality.

II. We define the following class of subsets of $\Omega$,

$$\mathcal{M} = \{ A \subset \Omega, \ \forall E \subset \Omega \ P^*(A \cap E) + P^*(A^c \cap E) = P^*(E) \}.$$ 

Since $P^*$ is subadditive, $A \in \mathcal{M}$ is equivalent to the so-called Carathéodory’s condition: for all $E \subset \Omega$,

$$P^*(A \cap E) + P^*(A^c \cap E) \leq P^*(E).$$ (A.1)

First we show that $\mathcal{M}$ is an algebra and $P^*$ is finitely additive on $\mathcal{M}$. Clearly, $\Omega \in \mathcal{M}$ and if $A \in \mathcal{M}$, then $A^c \in \mathcal{M}$. Let $A, B \in \mathcal{M}$. Then for an arbitrary subset $E$ of $\Omega$, we have

$$P^*(E) = P^*(B \cap E) + P^*(B^c \cap E) = P^*(A \cap E) + P^*(B \cap (A^c \cap E)) + P^*(B^c \cap E) \geq P^*(A \cap B \cap E) + P^*(A^c \cap E) \cup (A \cap B^c \cap E) \cup (A^c \cap B^c \cap E) \geq P^*((A \cap B) \cap E) + P^*((A \cap B)^c \cap E).$$

Thus $A \cap B \in \mathcal{M}$ and consequently, $\mathcal{M}$ is an algebra.

To prove the finite additivity of $P^*$ on $\mathcal{M}$, take $A, B \in \mathcal{M}$ with $A \cap B = \emptyset$ and note that since $A \in \mathcal{M}$, we have

$$P^*(A \cup B) = P^*(A \cap (A \cup B)) + P^*(A^c \cap (A \cup B)) = P^*(A) + P^*(B).$$

By induction, we easily get the desired finite additivity.

Now we argue that $P^*$ is in fact countably additive on $\mathcal{M}$. If $A_1, A_2, \ldots \in \mathcal{M}$ are pairwise disjoint and we let $A = \bigcup_{k=1}^{\infty} A_k$, then

$$\sum_{k=1}^{n} P^*(A_k) = P^*(\bigcup_{k=1}^{n} A_k) = P^*(A \cap \bigcup_{k=1}^{n} A_k) \leq P^*(A)$$

because $P^*$ is monotone (see I.). Taking the limit $n \to \infty$, we get $\sum_{k=1}^{\infty} P^*(A_k) \leq P^*(A)$. By the subadditivity of $P^*$, we also have the reverse inequality, hence we have equality and the countable additivity of $P^*$ follows.

It remains to show that $\mathcal{M}$ is a $\sigma$-algebra. It is enough to consider pairwise disjoint sets $A_1, A_2, \ldots \in \mathcal{M}$ and argue that $A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{M}$ (if they are not disjoint, we consider $B_n = A_n \cap A_{n-1}^c \cap \ldots \cap A_1^c$ which are pairwise disjoint, which are in $\mathcal{M}$.
and \( \bigcup_n B_n = A \). To this end, we want to verify (A.1) for \( A \). Fix \( E \subset \Omega \) and let \( K_n = \bigcup_{k=1}^n A_k \). By induction, we show that

\[
P^\ast(K_n \cap E) = \sum_{k=1}^n P^\ast(A_k \cap E).
\]

The base case \( n = 1 \) is clear. Further,

\[
P^\ast(K_{n+1} \cap E) = P^\ast(K_n \cap K_{n+1} \cap E) + P^\ast(K_n^c \cap K_{n+1} \cap E) \\
= P^\ast(K_n \cap E) + P^\ast(A_{n+1} \cap E) \\
= \sum_{k=1}^n P^\ast(A_k \cap E) + P^\ast(A_{n+1} \cap E),
\]

where in the last equality we used the inductive hypothesis. This finishes the inductive argument. Since \( K_n \in \mathcal{M} \), we obtain

\[
P^\ast(E) = \sum_{k=1}^n P^\ast(A_k \cap E) + P^\ast(E \cap A^c) \geq \sum_{k=1}^n P^\ast(A_k \cap E) + P^\ast(E \cap K_n^c) \\
\geq \sum_{k=1}^n P^\ast(A_k \cap E) + P^\ast(E \cap A^c),
\]

where the last inequality holds because \( P^\ast \) is monotone (see I.) and \( K_n \subset A \). Letting \( n \to \infty \) and using subadditivity, we get

\[
P^\ast(E) \geq \sum_{k=1}^\infty P^\ast(A_k \cap E) + P^\ast(E \cap A^c) \geq P^\ast(E \cap A) + P^\ast(E \cap A^c),
\]

so \( A \) satisfies (A.1).

**III.** We show 1) \( A \subset \mathcal{M} \) which also gives \( \sigma(A) \subset \mathcal{M} \) because \( \mathcal{M} \) is a \( \sigma \)-algebra. Moreover, we show 2) \( P^\ast = P \) on \( \mathcal{A} \), so \( P^\ast \) is the desired extension of \( P \) on \( \sigma(A) \). The uniqueness follows immediately from Dynkin’s theorem on \( \pi-\lambda \) systems (see Appendix B and Remark 2.11).

To prove 1), take \( A \in \mathcal{A} \) and an arbitrary subset \( E \) of \( \Omega \). Fix \( \varepsilon > 0 \). By the definition of \( P^\ast \), there are sets \( B_1, B_2, \ldots \in \mathcal{A} \) such that \( E \subset \bigcup_n B_n \) and \( \sum_{n=1}^\infty P(B_n) \leq P^\ast(E) + \varepsilon \). Since \( E \cap A \subset \bigcup_n (B_n \cap A) \) and \( E \cap A^c \subset \bigcup_n (B_n \cap A^c) \) and \( B_n \cap A, B_n \cap A^c \in \mathcal{A} \), by the definition of \( P^\ast \),

\[
P^\ast(E \cap A) \leq \sum_n P(B_n \cap A)
\]

and similarly

\[
P^\ast(E \cap A^c) \leq \sum_n P(B_n \cap A^c).
\]

Adding these up and using the additivity of \( P \) on \( \mathcal{A} \), we get

\[
P^\ast(E \cap A) + P^\ast(E \cap A^c) \leq \sum_n (P(B_n \cap A) + P(B_n \cap A^c)) = \sum_n P(B_n) \leq P^\ast(E) + \varepsilon,
\]

70
so (A.1) holds, so $A \in \mathcal{M}$.

To prove 2), take $A \in \mathcal{A}$. By the definition of $\mathbb{P}^*$, clearly, $\mathbb{P}^*(A) \leq \mathbb{P}(A)$. To argue for the opposite inequality, suppose $A \subseteq_n A_n$ for some $A_1, A_2, \ldots \in \mathcal{A}$. Let $C_n = \bigcup_{k=1}^n (A \cap A_k)$. We have that $C_1, C_2, \ldots$ are all in $\mathcal{A}$, $C_1 \subseteq C_2 \subseteq \ldots$ and $\bigcup_n C_n = \bigcup_{k=1}^\infty (A \cap A_k) = A \cap \bigcup_{k=1}^\infty A_k = A$ is also in $\mathcal{A}$. Using finite subadditivity of $\mathbb{P}$ on $\mathcal{A}$ and its monotonicity, we have

$$\mathbb{P}(C_n) \leq \sum_{k=1}^n \mathbb{P}(A \cap A_k) \leq \sum_{k=1}^n \mathbb{P}(A_k).$$

Letting $n \to \infty$, by assumption (ii) (finally used for the first and last time!), we obtain

$$\mathbb{P}(A) = \lim_{n \to \infty} \mathbb{P}(C_n) \leq \sum_{k=1}^\infty \mathbb{P}(A_k).$$

After taking the infimum over the $A_k$, this gives $\mathbb{P}(A) \leq \mathbb{P}(A^*)$, hence $\mathbb{P}(A) = \mathbb{P}(A^*)$. This finishes the whole proof. $\square$
B Appendix: Dynkin’s theorem

Recall that a family $A$ of subsets of a set $\Omega$ is a $\pi$-system if it is closed under finite intersections, that is for every $A, B \in A$, we have $A \cap B \in A$. A family $L$ of subsets of a set $\Omega$ is a $\lambda$-system if $\Omega \in L$, for every $A, B \in L$ with $A \subset B$, we have $B \setminus A \in L$ and for every $A_1, A_2, \ldots \in L$ such that $A_1 \subset A_2 \subset \ldots$, we have $\bigcup_{n=1}^{\infty} A_n \in L$.

B.1 Remark. If a family is a $\pi$-system and a $\lambda$-system, then it is a $\sigma$-algebra.

B.2 Theorem (Dynkin). Let $\Omega$ be a set. If a $\lambda$-system $L$ on $\Omega$ contains a $\pi$-system $A$ on $\Omega$, then $L$ contains $\sigma(A)$.

Proof. Let $L_0$ be the smallest $\lambda$-system containing $A$. By Remark B.1, it suffices to show that $L_0$ is a $\pi$-system. To this end, we first consider the family

$$C = \{ A \subset \Omega, \ A \cap B \in L_0 \text{ for every } B \in A \}.$$ 

Clearly, $C$ contains $A$. Moreover, $C$ is a $\lambda$-system. Indeed,

(i) $\Omega \in C$ because $A \subset L_0$,

(ii) let $U, V \in C$ with $U \subset V$, then for $B \in A$,

$$(V \setminus U) \cap B = (U \cap B) \setminus (V \cap B)$$

which is in $L_0$ because $U \cap B \subset V \cap B$ and $L_0$ is a $\lambda$-system

(iii) let $A_1, A_2, \ldots \in C$ with $A_1 \subset A_2 \subset \ldots$, then for $B \in A$, we have $A_1 \cap B \subset A_2 \cap B \subset \ldots$ and

$$\left( \bigcup_{i=1}^{n} A_i \right) \cap B = \bigcup_{i=1}^{n} (A_i \cap B)$$

which is in $L_0$ because $A_i \cap B$ are in $L_0$ and it is a $\lambda$-system.

We thus get that $C$, as a $\lambda$-system containing $A$, contains the smallest $\lambda$-system containing $A$, that is $L_0$. This means that $A \cap B \in L_0$ whenever $A \in L_0$ and $B \in A$.

The rest of the proof is a repetition of the same argument. We consider the family

$$\tilde{C} = \{ A \subset \Omega, \ A \cap B \in L_0 \text{ for every } B \in L_0 \}.$$ 

By the previous step, we know that $\tilde{C} \supset A$. We show that $\tilde{C}$ is a $\lambda$-system, hence, as above, it contains $L_0$. Therefore, for every $A, B \in L_0$, $A \cap B \in L_0$, that is $L_0$ is a $\pi$-system, as required. □
Appendix: Fubini’s theorem

Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2$ be two probability measures. Let 
$$
\Omega = \Omega_1 \times \Omega_2.
$$

Define the product $\sigma$-algebra 
$$
\mathcal{F} = \sigma(A_1 \times A_2, A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2),
$$

denoted $\mathcal{F}_1 \otimes \mathcal{F}_2$.

For $A \subset \Omega$, define the sections of $A$,
$$
A_{\omega_1} = \{ \omega_2 \in \Omega_2, (\omega_1, \omega_2) \in A \}, \quad \omega_1 \in \Omega_1,
$$
$$
A^{\omega_2} = \{ \omega_1 \in \Omega_1, (\omega_1, \omega_2) \in A \}, \quad \omega_2 \in \Omega_2.
$$

Similarly, for a function $X: \Omega \to \mathbb{R}$, define its section functions
$$
X_{\omega_1}: \Omega_2 \to \mathbb{R}, \quad X_{\omega_1}(\omega_2) = X(\omega_1, \omega_2), \quad \omega_1 \in \Omega_1,
$$
$$
X^{\omega_2}: \Omega_1 \to \mathbb{R}, \quad X^{\omega_2}(\omega_1) = X(\omega_1, \omega_2), \quad \omega_2 \in \Omega_2.
$$

We have the following lemma about $\mathcal{F}$-measurability.

C.1 Lemma. For every $A \in \mathcal{F}$, every $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, we have
$$
A_{\omega_1} \in \mathcal{F}_2, \quad A^{\omega_2} \in \mathcal{F}_1.
$$

For every $\mathcal{F}$-measurable function $X: \Omega \to \mathbb{R}$, every $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, we have
$$
X_{\omega_1} \text{ is } \mathcal{F}_2\text{-measurable}, \quad X^{\omega_2} \text{ is } \mathcal{F}_1\text{-measurable}.
$$

If moreover $X$ is nonnegative, we have that
$$
\omega_1 \mapsto \int_{\Omega_2} X_{\omega_1}(\omega_2) d\mathbb{P}_2(\omega_2) \text{ is } \mathcal{F}_1\text{-measurable}
$$
and
$$
\omega_2 \mapsto \int_{\Omega_1} X^{\omega_2}(\omega_1) d\mathbb{P}_1(\omega_1) \text{ is } \mathcal{F}_2\text{-measurable}
$$

Proof. Let $\mathcal{M}$ be the class of all subsets $A$ of $\Omega$ such that for every $\omega_1$, $A_{\omega_1}$ is $\mathcal{F}_2$-measurable. Clearly $\mathcal{M}$ contains product sets $B_1 \times B_2$, $B_i \in \mathcal{F}_i$, $i = 1, 2$ which form a $\pi$-system generating $\mathcal{F}$. Moreover, it is easy to check that $\mathcal{M}$ is a $\sigma$-algebra. Thus $\mathcal{M} \supset \mathcal{F}$. We argue similarly about $A^{\omega_2}$.

To prove the $\mathcal{F}_2$-measurability of $X_{\omega_1}$, note that for $B \in \mathcal{B}(\mathbb{R})$,
$$
X_{\omega_1}^{-1}(B) = \{ \omega_2 \in \Omega_2, X(\omega_1, \omega_2) \in B \} = X^{-1}(B)_{\omega_1}.
$$
which is in $F_2$ by the previous part because $X^{-1}(B) \in F$. The $F_1$-measurability of $X^{\omega_2}$, we proceed in the same way.

Finally, if $X = 1_{B_1 \times B_2}$ for some $B_i \in F_i$, $i = 1, 2$, we have

$$\int_{\Omega_2} X_{\omega_1}(\omega_2) dP_2(\omega_2) = 1_{B_1}(\omega_1) \int_{\Omega_2} 1_{B_2}(\omega_2) dP_2(\omega_2),$$

which is clearly $F_1$-measurable. Thus, by the standard arguments (see the proof of Theorem E.6), the same holds when $X$ is a simple function and consequently, thanks to Lebesgue’s monotone convergence theorem, when $X$ is nonnegative.

We define $P: F \to [0, 1]$ as follows: for $A \in F$, let $X = 1_A$ and

$$P(A) = \int_{\Omega_1} \left( \int_{\Omega_2} X_{\omega_1}(\omega_2) dP_2(\omega_2) \right) dP_1(\omega_1).$$

We have the following important result saying that $P$ is the so-called product measure on $\Omega$.

**C.2 Theorem** (The uniqueness of product measures). The set function $P$ is a unique probability measure on $(\Omega, F)$ such that for every $A_1 \in F_1$, $A_2 \in F_2$, we have

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2).$$

Moreover,

$$P(A) = \int_{\Omega_2} \left( \int_{\Omega_1} X^{\omega_2}(\omega_1) dP_1(\omega_1) \right) dP_2(\omega_2).$$

**Proof.** By Lemma C.1, the inner integral in the definition of $P$ is an $F_2$-measurable function, thus $P$ is well defined on $F$. Clearly, $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$, so in particular $P(\Omega) = 1$. If $B_1, B_2, \ldots \in F$ are disjoint, then so are their sections, that is we have $1_{\bigcup_n (B_n)} = \sum_n 1_{(B_n)}$, consequently, by the linearity of integrals, we get that $P$ is countably-additive. The uniqueness follows from the fact that the product sets $A_1 \times A_2$, $A_i \in F_i$, form a $\pi$-system generating $F$, combined with Remark 2.11. The formula with the integrals over $\Omega_1$ and $\Omega_2$ swapped follows by considering

$$\tilde{P}(A) = \int_{\Omega_2} \left( \int_{\Omega_1} X^{\omega_2}(\omega_1) dP_1(\omega_1) \right) dP_2(\omega_2),$$

checking that $\tilde{P}$ satisfies the same defining property, $\tilde{P}(A_1 \times A_2) = P_1(A_1)P_2(A_2)$ and using the uniqueness.

We say that $P$ is the product of $P_1$ and $P_2$, denoted

$$P = P_1 \otimes P_2.$$

**C.3 Theorem** (Fubini). Let $X: \Omega \to \mathbb{R}$ be $F$-measurable.
(i) If $X \geq 0$, then
\[
\int_{\Omega_1 \times \Omega_2} X \, dP = \int_{\Omega_1} \left( \int_{\Omega_2} X_{\omega_1}(\omega_2) \, dP_2(\omega_2) \right) \, dP_1(\omega_1) = \int_{\Omega_2} \left( \int_{\Omega_1} X_{\omega_2}(\omega_1) \, dP_1(\omega_1) \right) \, dP_2(\omega_2).
\]

(ii) If
\[
\int_{\Omega_1} \left( \int_{\Omega_2} |X_{\omega_1}(\omega_2)| \, dP_2(\omega_2) \right) \, dP_1(\omega_1) < \infty,
\]

or
\[
\int_{\Omega_2} \left( \int_{\Omega_1} |X_{\omega_2}(\omega_1)| \, dP_1(\omega_1) \right) \, dP_2(\omega_2) < \infty,
\]

then
\[
\int_{\Omega_1 \times \Omega_2} |X| \, dP < \infty,
\]
that is $X$ is $(\Omega, F, P)$-integrable.

(iii) If $X$ is $(\Omega, F, P)$-integrable, then
\[
P_1 \left\{ \omega_1 \in \Omega_1, \int_{\Omega_2} |X_{\omega_1}(\omega_2)| \, dP_2(\omega_2) < \infty \right\} = 1,
\]
\[
P_2 \left\{ \omega_2 \in \Omega_2, \int_{\Omega_1} |X_{\omega_2}(\omega_1)| \, dP_1(\omega_1) < \infty \right\} = 1
\]
and (i) holds.

Proof. (i) By Theorem C.2, the formula holds for $X = 1_A, A \in F$. Thus it holds for simple functions and by Lebesgue’s monotone convergence theorem, it holds for nonnegative functions.

(ii) Follows from (i) applied to $|X|$.

(iii) By the construction of Lebesgue integrals, $|X|$ being integrable gives
\[
\int_{\Omega} X^+ \, dP < \infty \quad \text{and} \quad \int_{\Omega} X^- \, dP < \infty.
\]
Thus from (a) applied to $X^+$,
\[
\int_{\Omega} X^+ \, dP = \int_{\Omega_1} \left( \int_{\Omega_2} X^+_{\omega_1}(\omega_2) \, dP_2(\omega_2) \right) \, dP_1(\omega_1),
\]
which by basic properties of Lebesgue integrals means that
\[
\int_{\Omega_2} X^+_{\omega_1}(\omega_2) \, dP_2(\omega_2) < \infty
\]
for $P_1$-a.e. $\omega_1$. Similarly for $X^-$. Therefore,
\[
P_1 \left\{ \omega_1 \in \Omega_1, \int_{\Omega_2} |X_{\omega_1}(\omega_2)| \, dP_2(\omega_2) < \infty \right\} = 1.
\]
In particular, for every \( \omega_1 \) in this event,
\[
\int_{\Omega_2} X_{\omega_1}(\omega_2) dP_2(\omega_2) = \int_{\Omega_2} X_{\omega_1}^+(\omega_2) dP_2(\omega_2) - \int_{\Omega_2} X_{\omega_1}^-(\omega_2) dP_2(\omega_2).
\]

For the remaining \( \omega_1 \), we can set all these integrals to be 0 and then we get
\[
\int_{\Omega_1} \left( \int_{\Omega_2} X_{\omega_1}(\omega_2) dP_2(\omega_2) \right) dP_1(\omega_1) = \int_{\Omega_1} \left( \int_{\Omega_2} X_{\omega_1}^+(\omega_2) dP_2(\omega_2) \right) dP_1(\omega_1) - \int_{\Omega_1} \left( \int_{\Omega_2} X_{\omega_1}^-(\omega_2) dP_2(\omega_2) \right) dP_1(\omega_1)
= \int_{\Omega_1 \times \Omega_2} X^+ dP - \int_{\Omega_1 \times \Omega_2} X^- dP
= \int_{\Omega_1 \times \Omega_2} X dP.
\]

We proceed in the same way for the swapped order of taking the integrals over \( \Omega_1 \) and \( \Omega_2 \).

Fubini’s theorem generalises to \( \sigma \)-finite measures as well as products of more than two but finitely many measures. Extensions to products of infinitely many measures are more delicate and are handled in the next appendix.
D Appendix: Infinite products of measures
E Appendix: Construction of expectation

The goal of this section is to define expectation of random variables and establish its basic properties. We shall only consider real-valued random variables. Recall that a function $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a random variable if for every $x \in \mathbb{R}$, the preimage $\{X \leq x\} = \{\omega \in \Omega, X(\omega) \leq x\} = X^{-1}((\infty, x])$ is an event (belongs to the sigma-field $\mathcal{F}$).

A random variable $X$ is called simple if its image $X(\Omega)$ is a finite set, that is

$$X = \sum_{k=1}^{n} x_k 1_{A_k},$$

for some distinct $x_1, \ldots, x_n \in \mathbb{R}$ (values) and events $A_1, \ldots, A_n$ which form a partition of $\Omega$ (we have, $A_k = \{X = x_k\}$).

The expectation of the simple random variable $X$, denoted $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \sum_{k=1}^{n} x_k \mathbb{P}(A_k).$$

The expectation of a nonnegative random variable $X$ is defined as

$$\mathbb{E}X = \sup\{\mathbb{E}Z, \ Z \text{ is simple and } Z \leq X\}.$$  

Note that $\mathbb{E}X \geq 0$ because we can always take $Z = 0$. We can have $\mathbb{E}X = +\infty$ (for instance, for a discrete random variable $X$ with $\mathbb{P}(X = k) = \frac{1}{k(k-1)}, k = 2, 3, \ldots$). For an arbitrary random variable $X$, we write

$$X = X^+ - X^-,$$

where

$$X^+ = \max\{X, 0\} = X 1_{\{X \geq 0\}}$$

is the positive part of $X$ and

$$X^- = -\min\{X, 0\} = -X 1_{\{X \leq 0\}}$$

is the negative part of $X$. These are nonnegative random variables and the expectation of $X$ is defined as

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$$

provided that at least one of the quantities $\mathbb{E}X^+, \mathbb{E}X^-$ is finite (to avoid $\infty - \infty$). We say that $X$ is integrable if $\mathbb{E}|X| < \infty$. Since $|X| = X^+ + X^-$, we have that $X$ is integrable if and only if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$.

One of the desired properties of expectation is linearity. It of course holds for simple random variables.
Theorem. Let $X$ and $Y$ be simple random variables. Then $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

Proof. Let $X = \sum_{k=1}^{m} x_k 1_{A_k}$ and $Y = \sum_{l=1}^{n} y_l 1_{B_l}$ for some reals $x_k, y_l$ and events $A_k$ and $B_l$ are such that the $A_k$ partition $\Omega$ and the $B_l$ partition $\Omega$. Then the events $A_k \cap B_l$, $k \leq m$, $l \leq n$ partition $\Omega$ and

$$X + Y = \sum_{k \leq m, l \leq n} (x_k + y_l) 1_{A_k \cap B_l}.$$ 

This is a simple random variable with

$$\mathbb{E}(X + Y) = \sum_{k \leq m, l \leq n} (x_k + y_l) \mathbb{P}(A_k \cap B_l)$$

$$= \sum_{k \leq m, l \leq n} x_k \mathbb{P}(A_k \cap B_l) + \sum_{k \leq m, l \leq n} y_l \mathbb{P}(A_k \cap B_l)$$

$$= \sum_{k \leq m} x_k \sum_{l \leq n} \mathbb{P}(A_k \cap B_l) + \sum_{l \leq n} y_l \sum_{k \leq m} \mathbb{P}(A_k \cap B_l)$$

$$= \sum_{k \leq m} x_k \mathbb{P} \left( \bigcup_{l \leq n} B_l \right) + \sum_{l \leq n} y_l \mathbb{P} \left( \bigcup_{k \leq m} A_k \cap B_l \right)$$

$$= \sum_{k \leq m} x_k \mathbb{P}(A_k) + \sum_{l \leq n} y_l \mathbb{P}(B_l),$$

which is $\mathbb{E}X + \mathbb{E}Y$ and this finishes the proof.

Nonnegative random variables

Our main goal is to prove linearity of expectation. We first establish a few basic properties of expectation for nonnegative random variables.

Theorem. Let $X$ and $Y$ be nonnegative random variables. We have

(a) if $X \leq Y$, then $\mathbb{E}X \leq \mathbb{E}Y$,

(b) for $a \geq 0$, $\mathbb{E}(a + X) = a + \mathbb{E}X$ and $\mathbb{E}(aX) = a\mathbb{E}X$,

(c) if $\mathbb{E}X = 0$, then $X = 0$ a.s. (i.e. $\mathbb{P}(X = 0) = 1$)

(d) if $A$ and $B$ are events such that $A \subset B$, then $\mathbb{E}X 1_A \leq \mathbb{E}X 1_B$.

Proof. (a) Let $\varepsilon > 0$. By definition, there is a simple random variable $Z$ such that $Z \leq X$ and $\mathbb{E}Z > \mathbb{E}X - \varepsilon$. Then also $Z \leq Y$, so by the definition of $\mathbb{E}Y$, we have $\mathbb{E}Z \leq \mathbb{E}Y$. Thus $\mathbb{E}X - \varepsilon < \mathbb{E}Y$. Sending $\varepsilon$ to 0 finishes the argument.

(b) For a simple random variable $Z$, clearly $\mathbb{E}(a + Z) = a + \mathbb{E}Z$ and $\mathbb{E}(aZ) = a\mathbb{E}Z$. It remains to follow the proof of (a).
(c) For \( n \geq 1 \), we have \( X \geq X_{1}{\{X \geq 1/n\}} \geq \frac{1}{n} \mathbf{1}{\{X \geq 1/n\}} \), so by (a) we get
\[
0 = \mathbb{E}X \geq \mathbb{E}\frac{1}{n} \mathbf{1}{\{X \geq 1/n\}} = \frac{1}{n} \mathbb{P}(X \geq 1/n),
\]
thus \( \mathbb{P}(X \geq 1/n) = 0 \), so by (a) we get
\[
0 = \mathbb{E}X \geq \mathbb{E}\frac{1}{n} \mathbf{1}{\{X \geq 1/n\}} = \frac{1}{n} \mathbb{P}(X \geq 1/n),
\]
thus \( \mathbb{P}(X \geq 1) = 0 \), so
\[
\mathbb{P}(X > 0) = \mathbb{P}\left(\bigcap_{n \geq 1} \{X \geq 1/n\}\right) = \lim_{n \to \infty} \mathbb{P}(X \geq 1/n) = 0.
\]

(d) follows immediately from (a).

The following lemma gives a way to approximate nonnegative random variables with monotone sequences of simple ones.

**E.3 Lemma.** If \( X \) is a nonnegative random variable, then there is a sequence \((Z_n)\) of nonnegative simple random variables such that for every \( \omega \in \Omega \), \( Z_n(\omega) \leq Z_{n+1}(\omega) \) and \( Z_n(\omega) \xrightarrow{n \to \infty} X(\omega) \).

**Proof.** Define
\[
Z_n = \sum_{k=1}^{n-2^n} \frac{k-1}{2^n} \mathbf{1}\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\} + n \mathbf{1}\{X \geq n\}.
\]
Fix \( \omega \in \Omega \). Then \( Z_n(\omega) \) is a nondecreasing sequence (check!). Since \( n > X(\omega) \) for large enough \( n \), we have for such \( n \) that \( 0 \leq X(\omega) - Z_n(\omega) \leq 2^{-n} \).

The following is a very important and useful tool allowing to exchange the order of taking the limit and expectation for monotone sequences.

**E.4 Theorem** (Lebesgue’s monotone convergence theorem). If \( X_n \) is a sequence of nonnegative random variables such that \( X_n \leq X_{n+1} \) and \( X_n \xrightarrow{n \to \infty} X \), then
\[
\mathbb{E}X_n \xrightarrow{n \to \infty} \mathbb{E}X.
\]

**Proof.** By E.2 (a), \( \mathbb{E}X_n \leq \mathbb{E}X_{n+1} \) and \( \mathbb{E}X_n \leq \mathbb{E}X \), so \( \lim_n\mathbb{E}X_n \) exists and is less than or equal to \( \mathbb{E}X \). It remains to show that \( \mathbb{E}X \leq \lim_n\mathbb{E}X_n \). Take a simple random variable \( Z \) such that \( 0 \leq Z \leq X \), with the largest value say \( K \). Observe that for every \( n \geq 1 \) and \( \varepsilon > 0 \),
\[
Z \leq (X_n + \varepsilon) \mathbf{1}\{X < X_n+\varepsilon\} + K \mathbf{1}\{Z \geq X_n+\varepsilon\}. \tag{E.1}
\]

**Claim.** For nonnegative random variables \( X, Y \) and an event \( A \), we have
\[
\mathbb{E}(X \mathbf{1}_A + Y \mathbf{1}_{A^c}) \leq \mathbb{E}X \mathbf{1}_A + \mathbb{E}Y \mathbf{1}_{A^c}.
\]

**Proof of the claim.** Fix \( \varepsilon > 0 \). Take a simple random variable \( Z \) such that \( Z \leq X \mathbf{1}_A + Y \mathbf{1}_{A^c} \) and \( \mathbb{E}Z > \mathbb{E}(X \mathbf{1}_A + Y \mathbf{1}_{A^c}) - \varepsilon \). Note that
\[
Z \mathbf{1}_A \leq X \mathbf{1}_A \quad \text{and} \quad Z \mathbf{1}_{A^c} \leq Y \mathbf{1}_{A^c}.
\]
Thus by E.2 (a),
\[ EZ1_A \leq EX1_A \quad \text{and} \quad EZ1_{A^c} \leq EY1_{A^c}. \]

Adding these two inequalities together and using that \( EZ1_A + EZ1_{A^c} = EZ \), which follows from linearity of expectation for simple random variables (Theorem E.1), we get
\[ E(X1_A + Y1_{A^c}) - \varepsilon < EZ \leq EX1_A + EY1_{A^c}. \]

Sending \( \varepsilon \to 0 \) finishes the argument. \( \square \)

Applying the claim to (E.1), we obtain
\[ EZ \leq EX_n + \varepsilon + KP(Z \geq X_n + \varepsilon). \]

The events \( \{ Z \geq X_n + \varepsilon \} \) form a decreasing family (because \( X_n \leq X_{n+1} \) and their intersection is \( \{ Z \geq X + \varepsilon \} = \emptyset \) (because \( X_n \to X \) and \( Z \leq X \)). Therefore taking \( n \to \infty \) in the last inequality gives
\[ EZ \leq \lim_{n} EX_n + \varepsilon. \]

Taking the supremum over simple random variables \( Z \leq X \) gives
\[ EX \leq \lim_{n} EX_n + \varepsilon. \]

Letting \( \varepsilon \to 0 \), we finish the proof. \( \square \)

As a corollary we obtain a result about the limit inferior of nonnegative random variables and its expectation.

**E.5 Theorem** (Fatou’s lemma). If \( X_1, X_2, \ldots \) are nonnegative random variables, then
\[ \mathbb{E}\liminf_{n \to \infty} X_n \leq \liminf_{n \to \infty} \mathbb{E}X_n. \]

*Proof.* Let \( Y_n = \inf_{k \geq n} X_k \). Then this is a nondecreasing sequence which converges to \( \liminf_{n \to \infty} X_n \) and \( Y_n \leq X_n \). Note that
\[ \liminf_{n \to \infty} \mathbb{E}X_n \geq \liminf_{n \to \infty} \mathbb{E}Y_n = \lim_{n \to \infty} \mathbb{E}Y_n, \]
where the last equality holds because the sequence \( \mathbb{E}Y_n \), as nondecreasing, is convergent. By Lebesgue’s monotone converge theorem,
\[ \lim_{n \to \infty} \mathbb{E}Y_n = \mathbb{E}\left( \lim_{n \to \infty} Y_n \right) = \mathbb{E}\liminf_{n \to \infty} X_n, \]
which in view of the previous inequality finishes the proof. \( \square \)

We are ready to prove linearity of expectation for nonnegative random variables.
**E.6 Theorem.** Let $X$ and $Y$ be nonnegative random variables. Then

$$E(X + Y) = EX + EY.$$  

**Proof.** By Lemma E.3, there are nondecreasing sequences $(X_n)$ and $(Y_n)$ of nonnegative simple random variables such that $X_n \to X$ and $Y_n \to Y$. Then the sequence $(X_n + Y_n)$ is also monotone and $X_n + Y_n \to X + Y$. By Theorem E.1,

$$E(X_n + Y_n) = EX_n + EY_n.$$  

Letting $n \to \infty$, by the virtue of Lebesgue’s monotone convergence theorem, we get in the limit $E(X + Y) = EX + EY$.  

**E.2 General random variables**

Key properties of expectation for general random variables are contained in our next theorem.

**E.7 Theorem.** If $X$ and $Y$ are integrable random variables, then

(a) $X + Y$ is integrable and $E(X + Y) = EX + EY$,

(b) $E(aX) = aEX$ for every $a \in \mathbb{R}$,

(c) if $X \leq Y$, then $EX \leq EY$,

(d) $|E|X| \leq |EX|$.  

**Proof.** (a) By the triangle inequality Theorem E.2 (a) and Theorem E.6,

$$E|X + Y| \leq E(|X| + |Y|) = E|X| + E|Y|$$

and the right hand side is finite by the assumption, thus $X + Y$ is integrable.

To show the linearity, write $X + Y$ in two different ways

$$(X + Y)^+ - (X + Y)^- = X + Y = X^+ - X^- + Y^+ - Y^-,$$

rearrange

$$(X + Y)^+ + X^- + Y^- = (X + Y)^- + X^+ + Y^+,$$

to be able to use the linearity of expectation for nonnegative random variables (Theorem E.6) and get

$$E(X + Y)^+ + E X^- + E Y^- = E(X + Y)^- + E X^+ + E Y^+,$$

which rearranged again gives $E(X + Y) = EX + EY$.

(b) We leave this as an exercise.
(c) Note that $X \leq Y$ is equivalent to saying that $X^+ \leq Y^+$ and $X^- \geq Y^-$ (if $X = X^+$, then $X \leq Y$ implies that $Y = Y^+$, hence $X^+ \leq Y^+$; similarly, if $Y = -Y^-$, then $X \leq Y$ implies that $X = -X^-$, hence $X^- \geq Y^-$). It remains to use Theorem E.2 (a).

(d) Since $-|X| \leq X \leq |X|$, by (c) we get $-E|X| \leq E X \leq E|X|$, that is $|E X| \leq E|X|$. ☐

E.3 Lebesgue’s dominated convergence theorem

We finish with one more limit theorem, quite useful in various applications; we also show one of them.

E.8 Theorem (Lebesgue’s dominated convergence theorem). If $(X_n)$ is a sequence of random variables and $X$ is a random variable such that for every $\omega \in \Omega$, we have $X_n(\omega) \longrightarrow X(\omega)$ and there is an integrable random variable $Y$ such that $|X_n| \leq Y$, then

$$
E|X_n - X| \longrightarrow 0.
$$

In particular,

$$
EX_n \longrightarrow EX.
$$

Proof. Since $|X_n| \leq Y$, taking $n \to \infty$ yields $|X| \leq Y$. In particular, $X$ is integrable as well. By the triangle inequality,

$$
|X_n - X| \leq 2Y
$$

and Fatou’s lemma (Theorem E.5) gives

$$
E(2Y) = E \lim inf(2Y - |X_n - X|) \leq \lim inf E(2Y - |X_n - X|)
$$

$$
= 2EY - \lim sup E|X_n - X|.
$$

As a result, $\lim sup E|X_n - X| \leq 0$, so

$$
E|X_n - X| \longrightarrow 0.
$$

In particular, since by Theorem E.7 (d),

$$
|E(X_n - X)| \leq E|X_n - X|,
$$

we get that the left hand side goes to 0, that is $EX_n \to EX$. ☐
References


