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

1.1 Definitions

Let Ω be a set. A collection $\mathcal{F}$ of its subsets is called a **σ-algebra** (sometimes also σ-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 coutable unions.

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

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

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

is called a **measure** if

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

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

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. The sets in $\mathcal{F}$ are called **events**. The empty set is called an impossible event because $\mathbb{P}(\varnothing) = 0$. 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) $\mathbb{P}(\bigcup_{i=1}^n A_i) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n)$,

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

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

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

$$B_1 = A_1,\ B_2 = A_2 \setminus A_1,\ B_3 = A_3 \setminus A_2, \ldots$$

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

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 Ω,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

Conversely, since

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

every probability measure is of this form.

1.5 Example. Let Ω 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 Ω, 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 Ω is fairly natural, say $\Omega = [0,1]$ for 1), defining an *appropriate* σ-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 σ-algebra, that is the smallest σ-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 σ-algebra **generated** by a family $\mathcal{A}$ of subsets of Ω (the smallest σ-algebra containing $\mathcal{A}$, which makes sense because intersections σ-algebras are still σ-algebras).

1.6 Example. As a result, for every $x \in [0,1]$, we have

$$\mathbb{P}(\{x\}) = \mathbb{P}\left(\bigcap_{n\geq 1}(x - 1/n, x + 1/n)\right) = \lim_{n\to\infty} \mathbb{P}((x - 1/n, x + 1/n)) = \lim_{n\to\infty} \frac{2}{n} = 0$$

(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. □

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 $0 \notin V$. Let $x \oplus y$ denote the addition modulo 1, that is $x \oplus y = x + y$ if $x + y \le 1$ and $x \oplus y = x + y - 1$ if $x + y > 1$. Consider the translations of V,

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

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

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

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 σ-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 σ-algebra $\mathcal{F}$, provided that certain conditions are met.

A family $\mathcal{A}$ of subsets of a set Ω 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 Ω be a set and let $\mathcal{A}$ be an algebra on Ω. Suppose a function $\mathbb{P}\colon \mathcal{A} \to [0, +\infty)$ 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 σ-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 $\bigcap B_n = \varnothing$, 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 $\bigcap B_n = \varnothing$. 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 $\bigcap B_n \neq \varnothing$. 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

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

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 $K_n = \bigcap_{k \leq n} \mathrm{cl}(C_k)$ 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, $\bigcap_n K_n = \bigcap_{n=1}^{\infty} \mathrm{cl}(C_n)$ is nonempty (recall a simple argument: otherwise $\bigcup_n (\mathrm{cl}(C_n))^c$ covers $[0, 1]$ without any finite subcover). Since $\bigcap B_n$ contains $\bigcap_n \mathrm{cl}(C_k)$, the argument is finished.

Theorem 1.8 provides a unique extension of $\mathbb{P}$ onto the σ-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 $\mathbb{P}(\{0\}) = 0$. $\square$

Given a metric space (E, ρ), the σ-algebra of subsets of E generated by all open sets in E is called the **Borel σ-algebra on** E, denoted $\mathcal{B}(E)$. For example, the σ-algebra constructed in the previous example is exactly $\mathcal{B}([0,1])$.

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 Ω on which defining a finitely additive measure will be intuitive and easy. Let Cyl be the family of all cylinders on Ω, 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,

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

For $A_{\varepsilon_1,\ldots,\varepsilon_n} \in \mathrm{Cyl}$, we set

$$\mathbb{P}\left(A_{\varepsilon_1,\ldots,\varepsilon_n}\right) = \frac{1}{2^n}.$$

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

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

which is a σ-algebra,

$$\mathbb{P}(A) = \mathrm{Leb}(f(A)), \qquad A \in \mathcal{F},$$

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 $\mathbb{P}\left(A_{\varepsilon_1,\ldots,\varepsilon_n}\right) = \frac{1}{2^n}$ and this construction also fulfils our intuitive basic requirement. We need get back to this example when we discuss independence. □

1.3 Conditioning

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

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad A \in \mathcal{F}.$$

It is natural to introduce a σ-algebra of events $\mathcal{F}_B$ *seen* by B, that is

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

1.12 Theorem. *$\mathbb{P}(\cdot|B))$ is a probability measure on $\mathcal{F}$, thus also on $\mathcal{F}_B$.*

The new probability measure $\mathbb{P}(\cdot|B)$ 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* $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$. *Then*

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

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

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A|B_n)\,\mathbb{P}(B_n).$$

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

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

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

1.4 Exercises

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

2. If $A_1, \ldots, A_n$ are events, then we have

   a) $\mathbb{P}(\bigcup_{i=1}^n A_i) \leq \sum_{k=1}^m (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$ for m odd,

   b) $\mathbb{P}(\bigcup_{i=1}^n A_i) \geq \sum_{k=1}^m (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$ 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 σ-algebras on a countable set.

5. Is there an infinite σ-algebra which is countable?

6. Show that the number of σ-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 $X\colon \Omega \to \mathbb{R}$ 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 $X\colon \Omega \to \mathbb{R}^n$ is called a **random vector**.

2.1 Example. Let A be an event. We define

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

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 $X\colon \Omega \to [0,1]$ 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 $X = \mathbf{1}_V$ 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 $\{X \le t\}$, or $\{X \in B\}$, etc. meaning $\{\omega \in \Omega,\ X(\omega) \le t\} = X^{-1}((-\infty, t])$, 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 $X\colon \Omega \to \mathbb{R}$ satisfies: for every $t \in \mathbb{R}$,*

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

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 σ-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 $t \ge 0$,

$$\{X^2 \le t\} = \{-\sqrt{t} \le X \le \sqrt{t}\} = \{X \le \sqrt{t}\} \setminus \{X < -\sqrt{t}\} \in \mathcal{F}$$

so X^2 and Y^2 are also random variables. Thus

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

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 $f\colon \mathbb{R} \to \mathbb{R}$ 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 **σ-algebra generated by** X, denoted $\sigma(X)$ as the smallest σ-algebra with respect to which X is measurable, that is

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

(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

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

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

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

2.8 Example. Let X be a constant random variable a.s., that is $\mathbb{P}\left(X = a\right) = 1$ for some $a \in \mathbb{R}$. Its law μ_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 **Dirac delta** at a, denoted δ_a.

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

$$F_X(t) = \mathbb{P}\left(X \leq t\right), \qquad t \in \mathbb{R}.$$

It is rather clear that for some two random variable 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 Ω is a **π-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 Ω is a **λ-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 π-system. The importance of this example is that this family generates $\mathcal{B}(\mathbb{R})$.

Note that if a family is a π-system and a λ-system, then it is a σ-algebra.

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

2.9 Theorem (Dynkin)**.** *If a λ-system $\mathcal{L}$ contains a π-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 π-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 λ-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 $\mu_X = \mu_Y$. □

2.11 Remark. The same proof gives the following: *if for some two probability measures μ, ν defined on the same space, we have $\mu = \nu$ on a π-system generating a σ-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 $s \le t$, $F_X(s) \le F_X(t)$,*

(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 $\{X \le s\} \subset \{X \le t\}$ if $s \le t$. Alternatively,

$$0 \le \mathbb{P}\left(X \in (s,t]\right) = \mathbb{P}\left(X \le t\right) - \mathbb{P}\left(X \le s\right) = F_X(t) - F_X(s).$$

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 $F\colon \mathbb{R} \to [0,1]$ 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

$$X(\omega) = \inf\{y,\ F(y) \ge \omega\}, \qquad \omega \in [0,1].$$

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

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

□

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 $\mathbb{P}\left((-\infty,t]\right) = F(t)$ and then extend $\mathbb{P}$. With such $\mathbb{P}$, the desired random variable is the canonical one, $X(x) = x$, $x \in \mathbb{R}$.

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

$$F_X(t_1,\dots,t_n) = \mathbb{P}\left(X_1 \le t_1,\dots,X_n \le t_n\right).$$

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 $s_i \le t_i$, $i \le n$, we have $F_X(s) \le F_X(t)$. Moreover,

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

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

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

(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 $A = \bigcap_k \{X_k \le t_k\}$ and $B = \bigcup_k \{X_k \le s_k\}$. Then

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

Note that $\mathbb{P}\left(A\right) = F_X(t)$ and $\mathbb{P}\left(A \cap B\right) = \bigcup_k (\{X_k \le s_k\} \cap B)$. Applying the inclusion-exclusion formula finishes the proof. $\square$

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 $F\colon \mathbb{R}^n \to [0,1]$ 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

$$\mathbb{P}\left(X = a\right) = \mathbb{P}\left(\{X \le a\} \setminus \{X < a\}\right) = \mathbb{P}\left(X \le a\right) - \mathbb{P}\left(X < a\right) = F_X(a) - F_X(a-).$$

Now, $\mathbb{P}\left(X = a\right) > 0$ if and only if F_X is discontinuous at a (has a jump) and the value of the jump is precisely $\mathbb{P}\left(X = a\right)$. 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 $\mathbb{P}\left(X \in A\right) = 1$. Say $A = \{a_1, a_2, \dots\}$ and denote $p_k = \mathbb{P}\left(X = a_k\right)$.

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 $f\colon \mathbb{R} \to \mathbb{R}$ such that

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

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,

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

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) $\int_{\mathbb{R}} f = 1$

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

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

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

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

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

2.21 Theorem. *Suppose a function $f\colon \mathbb{R} \to \mathbb{R}$ 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 $F(x) = \frac{x}{2}\,\mathbf{1}_{[0,1)}(x) + \mathbf{1}_{[1,+\infty)}(x)$ 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 $F\colon [0,1] \to [0,1]$ be the Cantor's devil's staircase function (a continuous nondecreasing function which is piecewise constant outside the Cantor set $\mathcal{C} \subset [0,1]$). 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 $x \notin \mathcal{C}$ (F is constant on $\mathcal{C}^c$), we would also have

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

(the first integral vanishes because $\mathcal{C}$ is of measure 0 and the second integral vanishes because as we just saw f vanishes on $\mathcal{C}^c$), 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 σ-algebras.

3.1 Definitions

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

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

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

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

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdot \ldots \cdot \mathbb{P}(A_n).$$

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 σ-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 σ-algebras generated by them, $\{\sigma(X_i)\}_{i \in I}$ is independent. Note that since $\sigma(A) = \sigma(\mathbf{1}_A)$, the events A_i are independent if and only if the random variables $\mathbf{1}_{A_i}$ 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 $\varnothing, \Omega, A_i, A_i^c$. Of course, many of these identities are either trivial of follow from the other. It turns out that π-systems can help and we have the following useful general lemma.

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

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

$$\begin{aligned}\mathcal{L}_1 = \{B_1 \in \mathcal{F} :\ \forall A_2 \in{} &\mathcal{A}_2, \dots, A_n \in \mathcal{A}_n \\ &\mathbb{P}(B_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(B_1)\,\mathbb{P}(A_2) \cdot \ldots \cdot \mathbb{P}(A_n)\}\end{aligned}$$

By the assumption $\mathcal{L}_1$ contains $\mathcal{A}_1$. By properties of probability measures, $\mathcal{L}_1$ is a λ-system. Hence, by Dynkin's theorem (Theorem 2.9), $\mathcal{L}_1$ contains $\sigma(\mathcal{A}_1)$. It remains to inductively repeat the same argument: suppose we know for some $k < n$ that

$$\mathbb{P}\left(B_1 \cap \cdots \cap B_k \cap A_{k+1} \cap \cdots \cap A_n\right) = \mathbb{P}\left(B_1\right) \cdots \mathbb{P}\left(B_k\right) \cdot \mathbb{P}\left(A_{k+1}\right) \cdots \mathbb{P}\left(A_n\right) \tag{3.1}$$

for every $B_i \in \sigma(\mathcal{A}_i)$, $i \le k$ and $A_j \in \mathcal{A}_j$, $j > k$. Fix some $B_i \in \sigma(\mathcal{A}_i)$, $i \le k$. As above, considering

$$\begin{aligned}\mathcal{L}_{k+1} = \{B_{k+1} \in \mathcal{F} :\ \forall A_{k+2} \in \mathcal{A}_{k+2}, \dots, A_n \in \mathcal{A}_n \\ &\mathbb{P}\left(B_1 \cap \cdots \cap B_k \cap B_{k+1} \cap A_{k+2} \cap \cdots \cap A_n\right) \\ &= \mathbb{P}\left(B_1\right) \cdots \mathbb{P}\left(B_k\right) \mathbb{P}\left(B_{k+1}\right) \cdot \mathbb{P}\left(A_{k+2}\right) \cdots \mathbb{P}\left(A_n\right)\}\end{aligned}$$

shows that (3.1) holds for $k+1$. Thus this holds for $k = n$. $\square$

We note two useful results about *packaging* independence.

3.2 Theorem. *Let $\{\mathcal{F}_i\}_{i \in I}$ be a family of independent σ-algebras. Suppose the index set I is partitioned into nonempty sets $\{I_j, j \in J\}$. Then the σ-algebras*

$$\mathcal{G}_j = \sigma\Big(\{\mathcal{F}_i, i \in I_j\}\Big), \qquad j \in J$$

are independent.

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

3.3 Theorem. *Suppose*

$$\begin{matrix} X_{1,1}, & \dots & X_{1,n_1}, \\ X_{2,1}, & \dots & X_{2,n_2}, \\ \vdots & \vdots & \vdots \\ X_{k,1}, & \dots & X_{k,n_k} \end{matrix}$$

are independent random variables and

$$\begin{aligned} &f_1 \colon \mathbb{R}^{n_1} \to \mathbb{R}, \\ &\vdots \\ &f_k \colon \mathbb{R}^{n_k} \to \mathbb{R} \end{aligned}$$

are measurable functions. Then the random variables

$$\begin{aligned} &Y_1 = f_1(X_{1,1}, \dots, X_{1,n_1}), \\ &\vdots \\ &Y_k = f_k(X_{k,1}, \dots, X_{k,n_k}) \end{aligned}$$

are independent.

Proof. By Theorem 3.2, the σ-algebras $\mathcal{G}_i = \sigma(\sigma(X_{i,1}), \ldots, \sigma(X_{i,n_i}))$, $i \le k$, are independent. The result follows because $\{Y_i \le t\} \in \mathcal{G}_i$, so $\sigma(Y_i) \subset \mathcal{G}_i$. $\square$

3.2 Product measures and independent random variables

Given two probability measures μ, ν 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

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

(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)$\Rightarrow$(ii) follows from the definition since $\{X \le s\} \in \sigma(X)$ and $\{Y \le y\} \in \sigma(Y)$.

(ii)$\Rightarrow$(i) follows from Lemma 3.1 ($\{X \le s\}_{s \in \mathbb{R}}$ is a π-system generating $\sigma(X)$).

(i)$\Rightarrow$(iii) from the definition, $\mu_{(X,Y)} = \mu \otimes \nu$ on the π-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)$\Rightarrow$(ii) follows by applying (iii) to $A = \{X \le s\}$, $B = \{Y \le t\}$. $\square$

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

$$\begin{aligned}\mathbb{P}\left((X_1,\dots,X_n)\in A_1\times\dots\times A_n\right) &= \prod_{i=1}^{n}\mathbb{P}\left(X_i\in A_i\right)\\ &= \prod_{i=1}^{n}\int_{A_i} f_i\\ &= \int_{A_1\times\dots\times A_n} f_1(x_1)\dots f_n(x_n)\mathrm{d}x_1\dots\mathrm{d}x_n.\end{aligned}$$

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

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

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

$$\mathbb{P}\left(X_1=a_1,\dots,X_n=a_n\right)=\mathbb{P}\left(X_1=a_1\right)\cdots\mathbb{P}\left(X_n=a_n\right).$$

3.3 Examples

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

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

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

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

Lemma 3.1 finishes the argument.

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

$$\begin{aligned}\mathbb{P}\left(A_i\right)&=\frac{1}{2}, && i=1,2,3,\\ \mathbb{P}\left(A_i\cap A_j\right)=\mathbb{P}\left(\{1\}\right)&=\frac{1}{4}=\mathbb{P}\left(A_i\right)\mathbb{P}\left(A_j\right), && i\neq j\\ \mathbb{P}\left(A_1\cap A_2\cap A_3\right)=\mathbb{P}\left(\{1\}\right)&=\frac{1}{4}\neq\mathbb{P}\left(A_1\right)\mathbb{P}\left(A_2\right)\mathbb{P}\left(A_3\right),\end{aligned}$$

so the events A_1,A_2,A_3 are pairwise independent but not independent.

3.9 Example. Let $\Omega = [0,1]^2$, $\mathcal{F} = \mathcal{B}([0,1]^2)$, $\mathbb{P} = \mathrm{Leb}$ (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

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

but

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

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

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

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

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

We skip the details.

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

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

where $d_n(x) \in \{0,1\}$ is the nth 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

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

Claim. $\mathbb{P}\left(A_k\right) = \frac{1}{2}$ 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 "kth toss is heads". To prove that $\mathbb{P}\left(A_k\right) = \frac{1}{2}$, 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 $\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\}$,

$$\begin{aligned}
\mathbb{P}\left(A_1^{\varepsilon_1} \cap \ldots \cap A_n^{\varepsilon_n}\right) &= \mathrm{Leb}\{x \in (0,1],\ d_1(x) = \varepsilon_1, \ldots, d_n(x) = \varepsilon_n\} \\
&= \mathrm{Leb}\left(\left(\sum_{i=1}^n \frac{\varepsilon_i}{2^i}, \sum_{i=1}^n \frac{\varepsilon_i}{2^i} + \frac{1}{2^n}\right]\right) \\
&= \frac{1}{2^n}
\end{aligned}$$

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 kth 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 ω. We define new functions

$$\begin{aligned}
&X_1(\omega) = 0.\omega_1\omega_3\omega_6\omega_{10}\ldots \\
&X_2(\omega) = 0.\omega_2\omega_5\omega_9\ldots \\
&X_3(\omega) = 0.\omega_4\omega_8\ldots \\
&X_4(\omega) = 0.\omega_7\ldots \\
&\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

(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 ω_i. The formal proof follows instantly from Theorem 3.2 about packaging independence.

The intuition for 2) is that each ω_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

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

(infinitely many A_k occur) and

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

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

$$(\limsup A_n)^c = \liminf A_n^c.$$

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

$$\limsup A_n = \{\omega \in \Omega,\ \limsup_{n\to\infty} \mathbf{1}_{A_n}(\omega) = 1\}$$

and

$$\liminf A_n = \{\omega \in \Omega,\ \liminf_{n\to\infty} \mathbf{1}_{A_n}(\omega) = 1\}.$$

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

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

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

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

□

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

$$\mathbb{P}(\limsup A_n) = 1.$$

Proof. By the monotonicity of the events $B_k = \bigcap_{n \geq k} A_n^c$, we get

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

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. $\square$

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

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

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 σ-algebra** by

$$\mathcal{T} = \bigcap_{n \geq 1} \sigma(X_{n+1}, X_{n+2}, \ldots).$$

For example, very natural events such as

$$\{\lim_n X_n \text{ exists}\}, \quad \{\sum X_n \text{ converges}\}, \quad \{\limsup_n X_n > 1\}$$

belong to $\mathcal{T}$. If we have independence, the tail σ-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 σ-algebra, then for every $A \in \mathcal{T}$, we have that either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.*

Proof. Define the σ-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 π-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 π-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 o $\mathcal{T}$, 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

$$r_n(x) = \operatorname{sgn}\Big(\cos(2^n \pi x)\Big), \qquad x \in [0,1],$$

where sgn is the usual signum function. Consider these functions as random variables on the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{Leb})$. What is the distribution of r_n? Show that the family $\{r_n\}_{n\geq 1}$ is independent.

2. Define the Walsh functions $w_A \colon \{-1,1\}^n \to \{-1,1\}$ indexed by all subsets A of $\{1, \ldots, n\}$,

$$w_A(x_1, \ldots, x_n) = \prod_{i \in A} x_i, \qquad x = (x_1, \ldots, x_n) \in \{-1,1\}^n$$

and $w_{\varnothing} = 1$ (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 Ω 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 Ω 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 σ-algebra. If a random variable Y is $\mathcal{T}$-measurable, then Y is a.s. constant.

11. Let $X_1, X_2, \dots$ 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 $f, g\colon [0,1] \to \mathbb{R}$ which, when viewed as random variables on the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{Leb})$, are independent?

4 Expectation

4.1 Definitions and basic properties

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

$$\int_\Omega |X(\omega)| \mathrm{d}\mathbb{P}(\omega) < \infty$$

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

$$\mathbb{E}X = \int_\Omega X(\omega) \mathrm{d}\mathbb{P}(\omega).$$

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 $\mathbb{R}^n$,

$$\mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{bmatrix}.$$

We list the most important basic properties

(i) monotonicity: if X is a nonnegative random variable, then $\mathbb{E}X \geq 0$,

(ii) the triangle inequality: $|\mathbb{E}X| \leq \mathbb{E}|X|$,

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

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*

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

(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*

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

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*

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

In particular,

$$\mathbb{E}X = \mathbb{E}\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} \mathbb{E}X_n.$$

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 $h\colon \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Let X be a random variable. Then*

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

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

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

II. If h is a simple function, that is $h = \sum_{i=1}^N x_i \mathbf{1}_{A_i}$ 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. $\square$

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

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

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

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

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

4.2 Variance and covariance

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

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

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

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

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

$$\mathrm{Var}(X + c) = \mathrm{Var}(X),$$

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

$$\mathrm{Var}(\lambda X) = \lambda^2 \,\mathrm{Var}(X),$$

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

$$\mathrm{Var}(X+Y) = \mathbb{E}(\bar{X} + \bar{Y})^2 = \mathbb{E}\bar{X}^2 + \mathbb{E}\bar{Y}^2 + 2\mathbb{E}\bar{X}\bar{Y} = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathbb{E}\bar{X}\bar{Y}.$$

This motives the following definition of the **covariance** between such two random variables,

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

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

4.7 Theorem. *Let $X_1, \dots, X_n$ be square-integrable random variables. Then*

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

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

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

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

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

It is convenient to write

$$\mathrm{Cov}(X) = \mathbb{E}\bar{X}\bar{X}^\top,$$

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

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

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

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

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

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

Proof. We show (iv) and leave the rest as an exercise. Let $M = \mathrm{Cov}(X)$. 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

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

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

□

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

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

which is called the **product σ-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 $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$,

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

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

$$\mathcal{B}(\mathbb{R}^n) = \mathcal{B}(\mathbb{R}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R})$$

(with the right hand side usually denoted $\mathcal{B}(\mathbb{R})^{\otimes n}$). One inclusion relies only on the definition of the product topology, that is

$$\mathcal{B}(\mathbb{R})^{\otimes n} \subset \mathcal{B}(\mathbb{R}^n)$$

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 $\mathcal{B}(\mathbb{R}^n)$. The opposite inclusion,

$$\mathcal{B}(\mathbb{R}^n) \subset \mathcal{B}(\mathbb{R})^{\otimes n}$$

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 $\mathcal{B}(\mathbb{R})^{\otimes n}$.

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 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We set*

$$\Omega = \prod_{i=1}^{\infty} \mathbb{R} = \mathbb{R} \times \mathbb{R} \times \ldots,$$

$$X_n(\omega_1, \omega_2, \ldots) = \omega_n, \qquad (\omega_1, \omega_2, \ldots) \in \Omega,$$

and

$$\mathcal{F} = \sigma(X_1, X_2, \ldots).$$

There is a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that for every $k \geq 1$ and $A_1, \ldots, A_k \in \mathcal{B}(\mathbb{R})$, w have

$$\mathbb{P}\left(A_1 \times \cdots \times A_k \times \mathbb{R} \times \ldots\right) = \mu_1(A_1) \cdot \ldots \cdot \mu_k(A_k).$$

Moreover, $X_1, X_2, \ldots$ are independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $X_1 \cdot \ldots \cdot X_n$ is integrable and*

$$\mathbb{E}(X_1 \cdot \ldots \cdot X_n) = \mathbb{E}X_1 \cdot \ldots \cdot \mathbb{E}X_n.$$

Proof. We have,

$$\begin{aligned}\mathbb{E}|X_1 \cdot \ldots \cdot X_n| &= \int_{\mathbb{R}^n} |x_1 \cdot \ldots \cdot x_n| \mathrm{d}\mu_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) \\ &= \int_{\mathbb{R}^n} |x_1 \cdot \ldots \cdot x_n| \mathrm{d}\mu_{X_1}(x_1) \ldots \mathrm{d}\mu_{X_n}(x_n) \\ &= \prod_{i=1}^n \int_{\mathbb{R}} |x_i| \mathrm{d}\mu_{X_i}(x_i),\end{aligned}$$

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 $\mathbb{E}(XY) = 0 = \mathbb{E}X \cdot \mathbb{E}Y$, 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* $\mathrm{Cov}(X_1, X_2) = 0$. *In particular, if $X_1, \ldots, X_n$ are independent square-integrable, then*

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

Since F_X determines the law of X, it should be possible to express $\mathbb{E}X$ using it. We finish this chapter with such a formula which is obtained from a simple trick that $x = \int_0^x \mathrm{d}t$, $x \geq 0$, combined with Fubini's theorem.

4.13 Theorem. *If X is a nonnegative random variable, then*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t) \,\mathrm{d}t.$$

Proof. We have,

$$\mathbb{E}X = \mathbb{E}\left(\int_0^X \mathrm{d}t\right) = \mathbb{E}\int_0^\infty \mathbf{1}_{X>t} \,\mathrm{d}t = \int_0^\infty \mathbb{E}\,\mathbf{1}_{X>t} \,\mathrm{d}t = \int_0^\infty \mathbb{P}(X > t) \,\mathrm{d}t,$$

where the usage of Fubini's theorem is justified because $\mathbf{1}_{X>t}$ 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
$$\sum_{n=1}^{\infty} \mathbb{P}(X > n) < \infty.$$

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

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

5. Let X be a random variable satisfying $\lim_{t\to\infty} t^p \mathbb{P}(|X| > t) = 0$. Show that for every $0 < \delta < 1$, we have $\mathbb{E}|X|^{1-\delta} < \infty$. Give an example of such a random variable for which $\mathbb{E}|X| = +\infty$.

6. Suppose X and Y are independent random variables and the distribution function of X is continuous. Then $\mathbb{P}(X = Y) = 0$.

7. 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.

8. Suppose X and Y are positive random variables with the same distribution. Does it follow that $\mathbb{E}\frac{X}{X+Y} = \mathbb{E}\frac{Y}{X+Y}$?

9. 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 $\mathbb{E}(X^m Y^n) = \mathbb{E}X^m \mathbb{E}Y^n$.

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

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

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

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

14. 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,

$$\mathbb{E}X = a, \qquad \mathrm{Var}(X) = 0.$$

2) *The Bernoulli distribution.* For $p \in [0, 1]$, let X be a random variable taking two values 0 and 1 with probabilities

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p.$$

Then

$$\mu_X = (1 - p)\delta_0 + p\delta_1$$

is the Bernoulli distribution with parameter p. We have,

$$\mathbb{E}X = p, \qquad \mathrm{Var}(X) = p(1 - p).$$

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

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

$$X \text{ has the same law as } X_1 + \cdots + X_n,$$

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,

$$X \text{ is the number of successes in } n \text{ independent Bernoulli trials.}$$

We have,

$$\mathbb{E}X = \mathbb{E}(X_1 + \ldots + X_n) = n\mathbb{E}X_1 = np$$

and

$$\operatorname{Var}(X) = n\operatorname{Var}(X_1) = np(1-p).$$

Notation: $X \sim \operatorname{Bin}(n,p)$.

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 λ. We will see later that this distribution arises as an appropriate limit of the Binomial distribution with parameters n and λ/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 λ. This distribution models well the number of events occuring in a fixed interval of time if these events occur with a constant mean rate λ, independently of the time since the last event, say the number of calls in a busy call centre.

We have,

$$\mathbb{E}X = \lambda, \qquad \operatorname{Var}(X) = \lambda.$$

Notation: $X \sim \operatorname{Poiss}(\lambda)$.

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

$$X \text{ has the same law as } \inf\{n \geq 1,\ X_n = 1\},$$

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,

$$\mathbb{E}X = \frac{1}{p}, \qquad \operatorname{Var}(X) = \frac{1-p}{p^2}.$$

Notation: $X \sim \operatorname{Geom}(p)$.

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

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

Then

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

We say that μ_X is the uniform measure on K. We have,

$$\mathbb{E}X = \frac{1}{|K|}\int_K x\mathrm{d}x \qquad \text{(the barycentre of } K\text{)}.$$

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

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

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

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

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

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

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

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

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

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

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

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

where

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

is the Gamma function. We say that μ_X (or X) has the Gamma distribution with parameters β and λ. 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 λ. We have,

$$\mathbb{E}X = \frac{\beta}{\lambda}, \qquad \mathrm{Var}(X) = \frac{\beta}{\lambda^2}.$$

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

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

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha - 1}(1-x)^{\beta - 1}\, \mathbf{1}_{(0,1)}(x), \qquad x \in \mathbb{R},$$

where

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

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

$$\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

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

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

$$f(x) = \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R}.$$

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

$$a_1 X_1 + \cdots + a_n X_n \text{ has the same law as } \left(\sum |a_i|\right)X,$$

where $X_1, \ldots, X_n$ are i.i.d. copies of X. Cauchy random variables are *not* integrable. Notation: $X \sim \mathrm{Cauchy}(1)$.

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

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

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

$$\mathbb{E}X = 0, \qquad \mathrm{Var}(X) = 1.$$

Notation: $X \sim N(0,1)$.

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

$$Y = \mu + \sigma X.$$

This a Gaussian random variable with parameters μ and σ. It has density

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}.$$

We have,

$$\mathbb{E}Y = \mu, \qquad \mathrm{Var}(Y) = \sigma^2.$$

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

The key property of the Gaussian distribution is that *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,

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{5.1}$$

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

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

(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,

$$\mathbb{E}X = 0 \;\; (\in \mathbb{R}^n), \qquad \mathrm{Cov}(X) = \mathrm{Id}_{n\times n}.$$

Notation: $X \sim N(0, I_{n\times n})$.

We say that a random vector Y in $\mathbb{R}^m$ is **Gaussian**, if

$$Y \text{ has the same law as } AX + b,$$

for some $m \times n$ matrix A and vector $b \in \mathbb{R}^m$, where $X \sim N(0, I_{n\times n})$. In other words, Gaussian vectors are defined as affine images of standard Gaussian vectors. We have,

$$\mathbb{E}Y = b, \qquad Q = \mathrm{Cov}(Y) = AA^\top.$$

Notation: $Y \sim N(b, Q)$.

In particular, if $m = n$ and A is nonsingular, then Y has density

$$g(x) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det Q}} e^{-\frac{1}{2}\left\langle Q^{-1}(x-b),(x-b)\right\rangle}, \qquad x \in \mathbb{R}^n$$

where

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

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 $f, g\colon \mathbb{R} \to \mathbb{R}$ is defined as a function

$$x \mapsto (f \star g)(x) = \int_{\mathbb{R}} f(x-y)g(y)\mathrm{d}y$$

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*

$$\mu_{X+Y}(A) = \int_A \mu_Y(A-x)\mathrm{d}\mu_X(x) = \int_A \mu_X(A-y)\mathrm{d}\mu_Y(y), \qquad A \in \mathcal{B}(\mathbb{R}).$$

In particular, if X has density f, then $X+Y$ has density

$$h(x) = \int_{\mathbb{R}} f(x-y)\mathrm{d}\mu_Y(y).$$

If both X, Y have densities, say f, g respectively, then

$$X + Y \textit{ has density } f \star g.$$

Proof. By independence, $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$, thus

$$\begin{aligned}\mu_{X+Y}(A) = \mu_{(X,Y)}\left\{(x,y) \in \mathbb{R}^2,\ x+y \in A\right\} &= \iint_{(x,y)\in\mathbb{R}^2,\ x+y\in A} \mathrm{d}\mu_X(x)\mathrm{d}\mu_Y(y) \\ &= \int_{x\in\mathbb{R}} \left[\int_{y\in A-x} \mathrm{d}\mu_Y(y)\right] \mathrm{d}\mu_X(x),\end{aligned}$$

where the last equality follows by Fubini's theorem. Since $\int_{y\in A-x} \mathrm{d}\mu_Y(y) = \mu_Y(A-x)$, the first identity follows. Note that swapping the roles of X and Y above gives also the identity

$$\mu_{X+Y}(A) = \int_{\mathbb{R}} \mu_X(A-y)\mathrm{d}\mu(y).$$

If X has density f, we have $\mu_X(A-y) = \int_{A-y} f(x)\mathrm{d}x$, so by a change of variables $x = z - y$ and Fubini's theorem, we get

$$\begin{aligned}\mu_{X+Y}(A) &= \int_{\mathbb{R}} \int_{A-y} f(x)\mathrm{d}x\mathrm{d}\mu_Y(y) \\ &= \int_{\mathbb{R}} \int_{A} f(z-y)\mathrm{d}z\mathrm{d}\mu_Y(y) \\ &= \int_{A} \left[\int_{\mathbb{R}} f(z-y)\mathrm{d}\mu_Y(y)\right] \mathrm{d}z,\end{aligned}$$

so $h(z) = \int_{\mathbb{R}} f(z-y)\mathrm{d}\mu_Y(y) = \mathbb{E}f(z-Y)$ is the density of $X+Y$. Finally, if Y has also density, say g, then this becomes $h(z) = \int_{\mathbb{R}} f(z-y)g(y)\mathrm{d}y$, that is $h = f \star g$. $\square$

Sometimes we use the notation $\mu_X \star \mu_Y$ to denote μ_{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

$$\begin{aligned}h(z) = \int_{\mathbb{R}} f(z-x)g(x)\mathrm{d}x &= \frac{1}{2\pi\sigma}\int_{\mathbb{R}} e^{-\frac{1}{2}(z-x)^2-\frac{1}{2\sigma^2}x^2}\mathrm{d}x \\ &= \frac{1}{2\pi\sigma}\int_{\mathbb{R}} e^{-\frac{1}{2}\frac{1+\sigma^2}{\sigma^2}\left(x-\sqrt{\frac{\sigma^2}{1+\sigma^2}}z\right)^2} e^{-\frac{1}{2}\frac{1}{1+\sigma^2}z^2}\mathrm{d}x \\ &= \frac{1}{2\pi\sigma}e^{-\frac{1}{2}\frac{1}{1+\sigma^2}z^2}\int_{\mathbb{R}} e^{-\frac{1}{2}u^2}\mathrm{d}u \cdot \sqrt{\frac{\sigma^2}{1+\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}}e^{-\frac{1}{2}\frac{1}{1+\sigma^2}z^2},\end{aligned}$$

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 $t \in \mathbb{R}$,

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

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

$$\int_a^b F' \le F(b) - F(a).$$

Proof. By Fatou's lemma,

$$\int_a^b F'(t)\mathrm{d}t = \int_a^b \liminf_{\delta\to 0+} \frac{F(t+\delta)-F(t)}{\delta}\mathrm{d}t \le \liminf_{\delta\to 0+} \int_a^b \frac{F(t+\delta)-F(t)}{\delta}\mathrm{d}t$$
$$= \liminf_{\delta\to 0+} \frac{1}{\delta}\left(\int_b^{b+\delta} F(t)\mathrm{d}t - \int_a^{a+\delta} F(t)\mathrm{d}t\right)$$

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*

$$\int_{-\infty}^x F' \le F(x) \qquad \textit{and} \qquad \int_x^\infty F' \le 1 - F(x).$$

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*

$$f(x) = \begin{cases} F_X'(x), & \textit{if } F_X'(x) \textit{ exists,} \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. By Corollary 5.4, it remains to show that for every $x \in \mathbb{R}$, we have $\int_\infty^x F_X' \ge F(x)$. This follows from

$$\int_{-\infty}^x F_X' + \int_x^\infty F_X' = \int_{-\infty}^\infty F_X' = 1$$

and $\int_x^\infty F_X' \le 1 - F_X(x)$. □

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 μ and λ. 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 λ. Show that $\mathbb{P}(X \geq k) = \mathbb{P}(Y \leq \lambda)$, 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 λ and μ. 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 $f(x, y) = cxy\,\mathbf{1}_{0<x<y<1}$. Find c and $\mathbb{P}(X + Y < 1)$. 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 $\mathbb{E}X_1$, $\mathbb{E}X_1^2$, $\mathbb{E}X_1X_2$, 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 $\{x \in \mathbb{R}^n,\ 0 \le x_1 \le \ldots \le x_n \le 1\}$. Find $\mathbb{E}U_k^*$ for $1 \le k \le n$.

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 $g \colon \mathbb{R} \to \mathbb{R}$ such that $F_X(t) = \int_{-\infty}^t g(x)\mathrm{d}x$ 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 $h(z) = \int_{-\infty}^{\infty} |y| f(yz) g(y) \mathrm{d}y$, $z \in \mathbb{R}$.

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

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

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 ρ 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

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

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 ρ for which U and V are independent.

22. Suppose (X, Y) is a centred (i.e., $\mathbb{E}X = \mathbb{E}Y = 0$) Gaussian random vector in $\mathbb{R}^2$ with $\mathrm{Cov}([\begin{smallmatrix} X \\ Y \end{smallmatrix}]) = [\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}]$. 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 $\mathbb{E}g^{2m} = 1 \cdot 3 \cdot \ldots \cdot (2m-1)$, $m = 1, 2, \ldots$.

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 $t > 0$,*

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

Proof. Since $X \geq X\,\mathbf{1}_{\{X \geq t\}} \geq t\,\mathbf{1}_{\{X \geq t\}}$, taking the expectation yields

$$\mathbb{E}X \geq \mathbb{E}t\,\mathbf{1}_{\{X \geq t\}} = t\mathbb{P}\left(X \geq t\right).$$

□

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

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

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

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

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

$$\mathbb{P}\left(|X - \mathbb{E}X| \geq t\right) \leq \frac{1}{t^2}\mathbb{E}|X - \mathbb{E}X|^2 = \frac{1}{t^2}\operatorname{Var}(X). \tag{6.3}$$

Another general and helpful inequality is about convex functions. Recall that a function $f\colon \mathbb{R} \to \mathbb{R}$ 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 $\lambda_1, \ldots, \lambda_n \geq$ such that $\sum_{i=1}^n \lambda_i = 1$ and every $x_1, \ldots, x_n \in \mathbb{R}$. The weights λ_i can of course be interpreted in probabilistic terms: if X is a random variable taking the value x_i with probability λ_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 $f\colon \mathbb{R} \to \mathbb{R}$ 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

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

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 $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function and consider the family of linear functions with rational coefficients which are *below* f,

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

Then

$$f(x) = \sup_{\ell \in \mathcal{A}} \ell(x), \qquad x \in \mathbb{R}.$$

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*

$$\mathbb{E}|XY| \leq \left(\mathbb{E}|X|^p\right)^{1/p} \left(\mathbb{E}|Y|^q\right)^{1/q}.$$

In particular, when $p = q = 2$, this gives the Cauchy-Schwarz inequality

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}|X|^2}\sqrt{\mathbb{E}|Y|^2}.$$

Proof 1. We can assume without loss of generality that $\mathbb{E}|X|^p$ and $\mathbb{E}|Y|^q$ 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 $x = \frac{|X|^p}{(\mathbb{E}|X|^p)^{1/p}}$, $y = \frac{|Y|^q}{(\mathbb{E}|Y|^q)^{1/q}}$, taking the expectation and simplifying yields the desired inequality. □

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

$$\big(\mathbb{E}|X||Y|\big)^p = \big(\tilde{\mathbb{E}}|X||Y|^{1-q}\big)^p \leq \tilde{\mathbb{E}}|X|^p|Y|^{(1-q)p} = \mathbb{E}|X|^p|Y|^{(1-q)p+q} = \mathbb{E}|X|^p = 1.$$

□

6.2 L_p-spaces

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

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

which is called the L_p space (on Ω). 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

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

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

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

and

$$\|X\|_\infty = \operatorname{ess\,sup} X = \inf\{M \geq 0,\ |X| \leq M \text{ a.s.}\}$$

(the essential supremum of X) with the usual convention that $\inf \varnothing = +\infty$. 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

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

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

$$\mathbb{E}|XY| \leq \|X\|_p \|Y\|_q. \tag{6.4}$$

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*

$$\|X\|_p = \sup\{\mathbb{E}XY,\ Y \textit{ is a random variable with } \mathbb{E}|Y|^q \leq 1\}, \tag{6.5}$$

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

Proof. To see that the supremum does not exceed the pth moment, simply apply Theorem 6.3. To see the opposite inequality, consider $Y = \mathrm{sgn}(X)|X|^{p-1}\|X\|_p^{-p/q}$. Then $\mathbb{E}XY = \|X\|_p$, 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 pth moment, the so-called Minkowski's inequality. □

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),

$$\|X+Y\|_p = \sup\{\mathbb{E}(X+Y)Z,\ \mathbb{E}|Z|^q \leq 1\}.$$

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. $\mathrm{Ber}(\alpha)$). Let us derive its analogue. Observe that for $0 < p < 1$ and every real numbers x, y, we have

$$|x+y|^p \leq |x|^p + |y|^p. \tag{6.6}$$

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

$$\mathbb{E}|X+Y|^p \leq \mathbb{E}|X|^p + \mathbb{E}|Y|^p, \qquad p \in (0,1]. \tag{6.7}$$

In other words,

$$\|X+Y\|_p^p \leq \|X\|_p^p + \|Y\|_p^p, \qquad p \in (0,1]. \tag{6.8}$$

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 form 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 $1 \le p < \infty$. By the Cauchy condition, there is a subsequence n_k such that

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

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 $\|Y_k\|_p \le 1$, by Fatou's lemma,

$$\mathbb{E}Y^p = \mathbb{E}\liminf Y_k^p \le \liminf \mathbb{E}Y_k^p \le 1,$$

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

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

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

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

This finishes the argument.

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

$$\begin{aligned} A_k &= \{|X_k| > \|X_k\|_\infty\} \\ B_{n,m} &= \{|X_m - X_n| > \|X_m - X_n\|_\infty\}. \end{aligned}$$

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). $\square$

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

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

Then

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

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

$$\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2(\|X\|_2^2 + \|Y\|_2^2). \tag{6.9}$$

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 $\|Y_n - X\|_2 \to d$. By the parallelogram law,

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

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,

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

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 $\left\langle X - \tilde{Y}, \tilde{Y} - Y\right\rangle = 0$. Since also $\left\langle X - Y, \tilde{Y} - Y\right\rangle = 0$, we get $\left\langle \tilde{Y} - Y, \tilde{Y} - Y\right\rangle = 0$, so $\tilde{Y} = Y$ a.s. □

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 $X_n \xrightarrow[n\to\infty]{a.s.} X$

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

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

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

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

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

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

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, $X_n \xrightarrow[n\to\infty]{a.s.} 0$ 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 $\bigcap_{n\geq N}\left\{|X_n - X| < \frac{1}{l}\right\} \subset \left\{|X_N - X| < \frac{1}{l}\right\}$, we have

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

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

b) $X_n \overset{a.s}{\nrightarrow} 0$ 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 $X_n \xrightarrow[n\to\infty]{a.s} X$ 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) $X_n \xrightarrow[n\to\infty]{L_p} 0$ because $\mathbb{E}|X_n|^p = 2^{-kp}$ when $2^k \leq n < 2^{k+1}$, which goes to 0 as n goes to ∞.

$\square$

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),

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

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

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

b) $X_n \overset{L_p}{\nrightarrow} 0$ because $\mathbb{E}|X_n|^p = n\frac{1}{n} = 1$; 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) $X_n \xrightarrow[n\to\infty]{a.s.} 0$ 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.

$$\begin{tikzcd}
\text{a.s.} \arrow[rr, Rightarrow, shift left, "/" marking] \arrow[dr, Rightarrow, shift left] & & \text{in } L_p \arrow[ll, Rightarrow, shift left, "/" marking] \arrow[dl, Rightarrow, shift left] \\
& \text{in } \mathbb{P} \arrow[ul, Rightarrow, shift left, "/" marking] \arrow[ur, Rightarrow, shift left, "/" marking] &
\end{tikzcd}$$

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 ε, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$, then we can find an index n_1 such that $\mathbb{P}\left(|X_{n_1} - X| > 2^{-1}\right) < 2^{-1}$. By the same logic, we can find an index $n_2 > n_1$ such that $\mathbb{P}\left(|X_{n_2} - X| > 2^{-2}\right) < 2^{-2}$, etc. We get a subsequence $(X_{n_k})_k$ such that $\mathbb{P}\left(|X_{n_k} - X| > 2^{-k}\right) < 2^{-k}$ for every k. Since the series $\sum_{k=1}^{\infty} \mathbb{P}\left(|X_{n_k} - X| > 2^{-k}\right)$ 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 $X_{n_k} \xrightarrow[k\to\infty]{} X$. $\square$

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 $\mathbb{P}(|X_n - X| > \varepsilon) \nrightarrow 0$. 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 $\mathbb{P}\left(|X_{n_{k_l}} - X| > \varepsilon\right) \to 0$. This contradiction finishes the proof. $\square$

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 $\frac{1}{6}n - \sqrt{n}$ and $\frac{1}{6}n + \sqrt{n}$ 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 $p \geq 1$, we have $c_1\sqrt{p} \leq \|g\|_p \leq c_2\sqrt{p}$ 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 $\|X\|_p \xrightarrow[p\to\infty]{} \|X\|_\infty$.

6. If $\mathbb{E}|X|^{p_0} < \infty$ for some $p_0 > 0$, then $\mathbb{E}\log_+|X| < \infty$ and
$$(\mathbb{E}|X|^p)^{1/p} \xrightarrow[p\to 0+]{} e^{\mathbb{E}\log|X|}$$
($\log_+ x = \max\{\log x, 0\}$, $\log_- x = \max\{-\log x, 0\}$, we set $\mathbb{E}\log|X| = \mathbb{E}\log_+|X| - \mathbb{E}\log_-|X| \in [-\infty, \infty)$ and use the convention that $e^{-\infty} = 0$). Thus it makes sense to define the 0th moment as $\|X\|_0 = e^{\mathbb{E}\log|X|}$.

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

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

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

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
$$\mathbb{E}|XY| \geq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}.$$

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

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

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

15. Let X be a random variable such that $\mathbb{E}|X|^p < \infty$ for every $p > 0$. Show that the function $p \mapsto \log \|X\|_{1/p}$ is convex on $(0, \infty)$.

16. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent random signs, that is $\mathbb{P}(\varepsilon_i = -1) = \frac{1}{2} = \mathbb{P}(\varepsilon_i = 1)$, $i \leq n$. 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
$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i\varepsilon_i\right| > \frac{1}{2}\sqrt{\sum_{i=1}^n a_i^2}\right) \geq c.$$

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

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

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

c) Let $S = \sum a_i\varepsilon_i$. Show that for every $t, \lambda > 0$, we have $\mathbb{P}(S > t) \leq e^{-\lambda t}\mathbb{E}e^{\lambda t}$.

d) Optimising over λ conclude that $\mathbb{P}(S > t) \le e^{-t^2/(2\sum a_i^2)}$.

e) Using symmetry, conclude that $\mathbb{P}(|S| > t) \le 2e^{-t^2/(2\sum a_i^2)}$.

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

21. *Hoeffding's inequality*: for independent random variables $X_1, \ldots, X_n$ such that $a_i \le X_i \le b_i$ a.s. for some $a_i < b_i$, $i \le n$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\right| > t\right) \le 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}, \qquad t \ge 0.$$

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

$$A_p\left(\sum_{i=1}^n a_i^2\right)^{1/2} \le \left(\mathbb{E}\left|\sum_{i=1}^n a_i\varepsilon_i\right|^p\right)^{1/p} \le B_p\left(\sum_{i=1}^n a_i^2\right)^{1/2},$$

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

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\varepsilon_1 + \ldots + \varepsilon_n}{\sqrt{2n\log n}} \le 1\right) = 1.$$

24. Let X be an integrable random variable and define

$$X_n = \begin{cases} -n, & X < -n \\ X, & |X| \le n \\ n, & X > n. \end{cases}$$

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

25. Show that if $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and $X_n \xrightarrow[n\to\infty]{\mathbb{P}} Y$, 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 $\frac{1}{n}\max_{k\le n}|X_k|$ converges to 0 in probability.

27. Show that if $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} Y$, then $X_nY_n \xrightarrow[n\to\infty]{\mathbb{P}} XY$.

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

29. Prove that a sequence of random variables X_n converges in probability if and only if for every $\varepsilon > 0$, $\lim_{n,m\to\infty}\mathbb{P}(|X_n - X_m| > \varepsilon) = 1$ (the Cauchy condition).

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

31. Let $X_1, X_2, \ldots$ be independent random variables, $X_n \sim \text{Poiss}(1/n)$. Does the sequence X_n converge a.s., in L_1, in L_2, in probability?

32. Show that if for every $\delta > 0$ we have $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \delta) < \infty$, then $X_n \xrightarrow[n\to\infty]{a.s.} X$.

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

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

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

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

Show that this a metric on V, (V, ρ) is complete and $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ if and only if $\rho(X_n, X) \to 0$.

36. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a discrete probability space. Show that for every sequence of random variables (X_n) on this space, $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ if and only if $X_n \xrightarrow[n\to\infty]{a.s.} X$.

37. Show that in general almost sure convergence is *not* metrisable.

38. *Weierstrass theorem.* Let $f\colon [0,1] \to \mathbb{R}$ be a continuous function. For $x \in [0,1]$ and an integer $n \geq 1$, let $S_{n,x}$ be a binomial random variable with parameters n and x. Let

$$Q_n(x) = \mathbb{E} f\left(\frac{S_{n,x}}{n}\right).$$

(a) Show that Q is a polynomial of degree n (in x) (Bernstein's polynomial of f).

(b) Using that f is bounded and uniformly continuous, combined with Chebyshev's inequality, show that for every $\varepsilon > 0$, there is n_0 such that for all $n \geq n_0$ and $x \in [0,1]$, we have

$$\mathbb{E}\left| f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| < \varepsilon.$$

(c) Conclude that for every $\varepsilon > 0$, there is a polynomial Q such that $\sup_{x\in[0,1]} |f(x) - Q(x)| < \varepsilon$.

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 $\frac{X_1+\ldots+X_n}{n} \xrightarrow[n\to\infty]{} \mathbb{E}X_1$. 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 $\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$. Let $\bar{S}_n = \frac{\varepsilon_1+\ldots+\varepsilon_n}{n}$. By Bernstein's inequality (Exercise 6.19), $\mathbb{P}\left(|\bar{S}_n| > t\right) \leq 2e^{-nt^2/2}$, so the series $\sum_{n=1}^\infty \mathbb{P}\left(|\bar{S}_n| > t\right)$ converges, so $\bar{S}_n \xrightarrow[n\to\infty]{a.s.} 0 = \mathbb{E}\varepsilon_1$ (check!). In other words, the sequence (ε_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*

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

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

$$\frac{S_n}{n} - \mathbb{E}\frac{S_n}{n} \xrightarrow[n\to\infty]{L_2} 0.$$

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

Proof. We have

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

Since

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

when the X_i are uncorrelated with bounded variance, we have

$$\frac{1}{n^2}\operatorname{Var}(X_1+\ldots+X_n) \le \frac{Mn}{n^2} = \frac{M}{n}$$

which goes to 0 as $n \to \infty$. $\square$

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

$$\frac{X_1^2+\ldots+X_n^2}{n} - \mathbb{E}X_1^2 \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

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 $k \ge 1$, let

$$\tau_k = \inf\{m \ge 1,\ |\{X_1, \ldots, X_m\}| = k\}.$$

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 τ_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

$$T_k = \tau_k - \tau_{k-1}, \qquad k \ge 1,$$

which is time we wait to get a coupon of a next type after we have collected $k-1$ different coupons. We have,

$$\mathbb{P}(T_k = l) = \left(\frac{k-1}{n}\right)^{l-1}\left(1 - \frac{k-1}{n}\right), \qquad l = 1, 2, \ldots,$$

that is

$$T_k \sim \operatorname{Geom}\left(1 - \frac{k-1}{n}\right)$$

and $T_1, \ldots, T_n$ are independent. Plainly,

$$\tau_n = T_1 + \ldots + T_n.$$

Thus

$$\mathbb{E}\tau_n = \sum_{k=1}^n \mathbb{E}\tau_k = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-1} = n\sum_{k=1}^n \frac{1}{n-k+1} = n\sum_{k=1}^n \frac{1}{k},$$

so for large n, we have $\mathbb{E}\tau_n \sim n\log n$. Moreover, thanks to independence,

$$\mathrm{Var}(\tau_n) = \sum_{k=1}^n \mathrm{Var}(T_k) \leq \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-2} = n^2\sum_{k=1}^n \frac{1}{k^2} < 2n^2.$$

If we let

$$t_n = \frac{\tau_n}{n\log n},$$

we obtain

$$\begin{aligned}\mathbb{E}(t_n - 1)^2 = \frac{1}{n^2\log^2 n}\mathbb{E}(\tau_n - n\log n)^2 &\leq \frac{2}{n^2\log^2 n}\Big(\mathbb{E}(\tau_n - \mathbb{E}\tau_n)^2 + (\mathbb{E}\tau_n - n\log n)^2\Big) \\ &\leq \frac{4}{\log^2 n} + \frac{2}{\log^2 n}.\end{aligned}$$

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*

$$t\mathbb{P}\left(|X_1| > t\right) \xrightarrow[t\to\infty]{} 0, \tag{7.1}$$

then

$$\frac{X_1 + \ldots + X_n}{n} - \mu_n \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{7.2}$$

where $\mu_n = \mathbb{E}X_1 \,\boldsymbol{1}_{\{|X_1|\leq n\}}$.

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

$$\frac{X_1 + \cdots + X_n}{n} - a_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

(see exercises).

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

7.8 Lemma. *Let $\{X_{n,k}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of random variables such that for every n, $X_{n,1}, \ldots, X_{n,n}$ are independent (i.e. they are independent within each row). Let (b_n) be a sequence of positive numbers such that $b_n \to \infty$. Let*

$$\tilde{X}_{n,k} = X_{n,k}\,\boldsymbol{1}_{\{|X_{n,k}|\leq b_n\}}$$

and

$$\tilde{S}_n = \sum_{k=1}^{n} \tilde{X}_{n,k}.$$

If the following two conditions are satisfied

(i) $\sum_{k=1}^n \mathbb{P}\left(|X_{n,k}| > b_n\right) \xrightarrow[n\to\infty]{} 0,$

(ii) $b_n^{-2} \sum_{k=1}^n \mathbb{E}\tilde{X}_{n,k}^2 \xrightarrow[n\to\infty]{} 0,$

then

$$\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

where $S_n = \sum_{k=1}^n X_{n,k}$.

Proof. Fix $\varepsilon > 0$. First note that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n}\right| > \varepsilon\right) &= \mathbb{P}\left(\left|\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n}\right| > \varepsilon,\ S_n \neq \tilde{S}_n\right) \\
&\quad + \mathbb{P}\left(\left|\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n}\right| > \varepsilon,\ S_n = \tilde{S}_n\right) \\
&\leq \mathbb{P}\left(S_n \neq \tilde{S}_n\right) + \mathbb{P}\left(\left|\frac{\tilde{S}_n - \mathbb{E}\tilde{S}_n}{b_n}\right| > \varepsilon\right).
\end{aligned}$$

We show that each of the two terms on the right hand side goes to 0 as n goes to ∞. For the first term, since $S_n \neq \tilde{S}_n$ implies that for some k, $X_{n,k} \neq \tilde{X}_{n,k}$ which in turn implies that $|X_{n,k}| > b_n$, by the union bound, we have,

$$\mathbb{P}\left(S_n \neq \tilde{S}_n\right) \leq \sum_{k=1}^n \mathbb{P}\left(X_{n,k} \neq \tilde{X}_{n,k}\right) \leq \sum_{k=1}^n \mathbb{P}\left(|X_{n,k}| > b_n\right) \xrightarrow[n\to\infty]{} 0,$$

by (i). It remains to handle the second term. By Chebyshev's inequality, the independence of the $\tilde{X}_{n,k}$ and a simple bound $\mathrm{Var}(Y) \leq \mathbb{E}Y^2$, we get

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{\tilde{S}_n - \mathbb{E}\tilde{S}_n}{b_n}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left|\frac{\tilde{S}_n - \mathbb{E}\tilde{S}_n}{b_n}\right|^2 &= \frac{1}{\varepsilon^2 b_n^2}\mathrm{Var}(\tilde{S}_n) = \frac{1}{\varepsilon^2 b_n^2}\sum_{k=1}^n \mathrm{Var}(\tilde{X}_{n,k}) \\
&\leq \frac{1}{\varepsilon^2 b_n^2}\sum_{k=1}^n \mathbb{E}\tilde{X}_{n,k}^2.
\end{aligned}$$

The right hand side goes to 0 as n goes to ∞ thanks to (ii). This finishes the proof. $\square$

Proof of Theorem 7.6. We use Lemma 7.8 with $X_{n,k} = X_k$ and $b_n = n$. Then

$$\tilde{X}_{n,k} = X_k \mathbf{1}_{|X_k| \leq n}$$

which has the same law as $X_1 \mathbf{1}_{|X_1| \leq n}$. It suffices to check (i) and (ii) of Lemma 7.8 because its assertion is exactly (7.2), due to the fact that

$$\mathbb{E}\tilde{S}_n = \sum_{k=1}^n \mathbb{E}X_k \mathbf{1}_{|X_k| \leq n} = n\mu_n.$$

For (i), we simply have

$$\sum_{k=1}^{n} \mathbb{P}\left(|X_{n,k}| > b_n\right) = \sum_{k=1}^{n} \mathbb{P}\left(|X_k| > n\right) = n\mathbb{P}\left(|X_1| > n\right) \xrightarrow[n\to\infty]{} 0,$$

by (7.1). For (ii), we have

$$b_n^{-2} \sum_{k=1}^{n} \mathbb{E}\tilde{X}_{n,k}^2 = \frac{1}{n^2} \sum_{k=1}^{n} \mathbb{E}X_k^2 \,\mathbf{1}_{|X_k| \le n} = \frac{1}{n} \mathbb{E}X_1^2 \,\mathbf{1}_{|X_1| \le n}.$$

We compute

$$\mathbb{E}X_1^2 \,\mathbf{1}_{|X_1| \le n} = \mathbb{E} \int_0^\infty 2t \,\mathbf{1}_{\{t \le |X_1|,\ |X_1| \le n\}} \,\mathrm{d}t \le \int_0^n 2t\mathbb{P}\left(|X_1| > t\right) \mathrm{d}t.$$

Let $f(t) = t\mathbb{P}\left(|X_1| > t\right)$. It thus remains to show that

$$\frac{1}{n} \int_0^n f(t)\mathrm{d}t \xrightarrow[n\to\infty]{} 0.$$

This is a consequence of two properties of f: $f(t) \le t$ and $f(t) \xrightarrow[t\to\infty]{} 0$ (by (7.1)). These in turn imply that $M = \sup f < \infty$. Thus the following standard Cesáro-type lemma finishes the proof. $\square$

7.9 Lemma. *Let* $f \colon [0, +\infty) \to [0, +\infty)$ *be a bounded function with* $f(t) \xrightarrow[t\to\infty]{} 0$. *Then*

$$\frac{1}{n} \int_0^n f(t)\mathrm{d}t \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $M = \sup f$. We fix $\varepsilon > 0$ and choose L such that $f(t) < \varepsilon$ for all $t > L$. Then,

$$\frac{1}{n} \int_0^n f(t)\mathrm{d}t = \frac{1}{n}\left(\int_0^L f + \int_L^n f\right) \le \frac{1}{n}\big(LM + (n-L)\varepsilon\big) < \frac{LM}{n} + \varepsilon < 2\varepsilon,$$

provided that $n > \frac{LM}{\mathbb{E}}$. This finishes the proof. $\square$

7.10 Remark. The same argument gives a similar result for sequences: if for a sequence (a_n) of real numbers we have $a_n \to a$, then

$$\frac{a_1 + \cdots + a_n}{n} \to a.$$

7.11 Remark. Our weak law of large numbers – Theorem 7.6 – in particular gives the following: if $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}|X_1| < \infty$, then

$$\frac{S_n}{n} - \mathbb{E}X_1 \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

(which will be strengthen to a.s. convergence in the next section, which is the content of the strong law of large numbers). Indeed,

$$t\mathbb{P}\left(|X_1| > t\right) = \mathbb{E}(t\,\mathbf{1}_{|X_1| > t}) \le \mathbb{E}(|X_1|\,\mathbf{1}_{|X_1| > t}) \xrightarrow[t\to)\infty]{} 0$$

by Lebesgue's dominated convergence theorem: pointwise $|X_1|\,\mathbf{1}_{|X_1|>t} \xrightarrow[t\to\infty]{} 0$ and $|X_1|$ is an integrable majorant (see also Exercise 4.4). Moreover, similarly,

$$\mu_n = \mathbb{E}X_1\,\mathbf{1}_{|X_1|\le n} \xrightarrow[n\to\infty]{} \mathbb{E}X_1,$$

which combined with the conclusion of Theorem 7.6,

$$\frac{S_n}{n} - \mu_n \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

gives

$$\frac{S_n}{n} - \mathbb{E}X_1 \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

7.12 Remark. Recall Exercise 4.5: if $\lim_{t\to\infty} t^p\mathbb{P}\left(|X_1| > t\right) = 0$, then for every $0 < \delta < 1$, we have $\mathbb{E}|X_1|^{1-\delta} < \infty$. This shows that assumption (7.1) of Theorem 7.6 "almost" implies that X_1 is integrable. An example of a random variable with $\mathbb{P}\left(|X_1| > t\right) = \frac{1}{t\log t}$, $t > e$, shows that the weak law still holds ((7.1) is fulfilled), even though the expectation does not exist. Note that if X_1 is a standard Cauchy random variable, then for $t > 1$, we have

$$t\mathbb{P}\left(|X_1| > t\right) = 2t\int_t^\infty \frac{\mathrm{d}x}{\pi(1+x^2)} \ge 2t\int_t^\infty \frac{\mathrm{d}x}{2\pi x^2} = \frac{1}{\pi},$$

so (7.1) does not hold. As discussed in Example 7.1, in this case no reasonable law-of-large-numbers-type convergence should hold.

7.13 Example. We shall describe the so-called St. Petersburg paradox. Let $X_1, X_2, \ldots$ be i.i.d. random variables with the following discrete distribution

$$\mathbb{P}\left(X_1 = 2^k\right) = 2^{-k}, \qquad k = 1, 2, \ldots$$

Such distribution models this simple casino game: you sit down and they toss a coin until the first head shows up, which finishes the game and your pay is 2^k dollars, where k is the number of tosses. The expected value of your payout is thus

$$\mathbb{E}X_1 = \frac{1}{2}\cdot 2 + \frac{1}{2^2}\cdot 2^2 + \cdots = +\infty.$$

How much should the casino charge for this game? Is there any fair charge? (Obviously, they cannot charge "$+\infty$".) Suppose you want to play n games, n is large. Your payout after n games is

$$S_n = X_1 + \cdots + X_n.$$

A fair charge per game should be a "typical value" of S_n/n. we cannot apply Theorem 7.6. Instead, we can estimate it using Lemma 7.8 which gives us extra flexibility in the choice of the normalising constant b_n. We want to choose b_n as small as possible with conditions (i) and (ii) of the lemma being satisfied. Since

$$\mathbb{P}\left(X_1 \ge 2^m\right) = \sum_{j=m}^{\infty} 2^{-j} = 2^{-m+1},$$

the quantity in (i) for $b_n = 2^{m_n}$ equals

$$n\mathbb{P}\left(X_1 \geq b_n\right) = n \cdot 2^{-m_n+1}.$$

If we thus choose $m_n = \log_2 n + \omega_n$ with $0 < \omega_n \to \infty$ slowly and such that m_n is an integer, we get that $n\mathbb{P}\left(X_1 > b_n\right) \to 0$. Moreover, for the quantity in (ii), since $b_n = n2^{\omega_n}$, we have

$$\frac{n}{b_n^2}\mathbb{E}X_1^2\,\mathbf{1}_{X_1 \leq b_n} = \frac{n}{b_n^2}\mathbb{E}X_1^2\,\mathbf{1}_{X_1 \leq 2^{m_n}} = \frac{n}{b_n^2}\sum_{j \leq m_n} 2^{2j} \cdot 2^{-j} < \frac{n}{b_n^2}2^{m_n+1} = \frac{2n}{b_n} = \frac{2}{2^{\omega_n}} \to 0.$$

Thus, by Lemma 7.8,

$$\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

It remains to find $\mathbb{E}\tilde{S}_n$. We have,

$$\mathbb{E}\tilde{S}_n = n\mathbb{E}X_1\,\mathbf{1}_{X_1 \leq b_n} = n\sum_{j \leq m_n} 2^j \cdot 2^{-j} = nm_n.$$

Consequently, choosing ω_n to be asymptotically $\log_2\log_2 n$, that is such that $\frac{\omega_n}{\log_2\log_2 n} \to 1$, we get

$$\frac{S_n - \mathbb{E}\tilde{S}_n}{b_n} = \frac{S_n - n\lfloor\log_2 n\rfloor}{n \cdot 2^{\omega_n}} = \frac{S_n}{n \cdot 2^{\omega_n}} - \frac{\lfloor\log_2 n\rfloor}{2^{\omega_n}},$$

thus,

$$\frac{S_n}{n\log_2 n} \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

As a result, S_n/n is typically like $\log_2 n$, so a fair charge for playing n games should be $\log_2 n$ per game.

7.2 Strong law of large numbers

Strong laws of large numbers concern a.s. convergence. The following simple lemma turns out to be quite useful in such situations (see Exercises 6.31 and 6.32).

7.14 Lemma. *Let $X_1, X_2, \ldots$ be a sequence of random variables such that for every ε, we have*

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|X_n| > \varepsilon\right) < \infty. \tag{7.3}$$

Then,

$$X_n \xrightarrow[n\to\infty]{a.s.} 0.$$

This also holds if for some sequence $\varepsilon_1, \varepsilon_2, \ldots$ of positive numbers convergent to 0, we have

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|X_n| > \varepsilon_n\right) < \infty. \tag{7.4}$$

Proof. From (7.3), by the first Borel-Cantelli lemma, we get

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(|X_n| > \varepsilon \text{ for infinitely many } n\right) = 0,$$

or, equivalently,

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(|X_n| \leq \varepsilon \text{ eventually}\right) = 1.$$

Since the intersection of countably many certain events is a certain event, we get

$$\mathbb{P}\left(\forall k = 1, 2, \ldots \ |X_n| \leq \frac{1}{k} \text{ eventually}\right) = 1.$$

By the definition of a limit, this implies that $\mathbb{P}(X_n \to 0) = 1$. Similarly, by (7.4), we get

$$\mathbb{P}\left(|X_n| \leq \varepsilon_n \text{ eventually}\right) = 1.$$

Since $e_n \to 0$, by the sandwich theorem, $X_n \to 0$ with probability 1. □

As a warm-up we show the strong law of large numbers under the generous assumption of a finite 4th moment. Even though we shall not need this result to prove it under optimal assumptions, the technique employed here of analysing high enough moments is quite useful and important.

7.15 Theorem. *Let $X_1, X_2, \ldots$ be independent random variables such that for all i, $\mathbb{E}|X_i|^4 \leq C$ and $\mathbb{E}X_i = \mu$ for some constants $C > 0$ and $\mu \in \mathbb{R}$. Then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

Proof. Without loss of generality, we can assume that $\mu = 0$ (otherwise, we consider $X_i - \mu$; for the assumption, we can use the triangle inequality, $\|X_i - \mu\|_4 \leq \|X_i\|_4 + \mu$). We have,

$$\mathbb{E}(X_1 + \cdots + X_n)^4 = \sum_{i,j,k,l=1}^{n} \mathbb{E}X_i X_j X_j X_k.$$

There are 5 types of terms: $\mathbb{E}X_i^4$, $\mathbb{E}X_i^2 X_j^2$, $\mathbb{E}X_i X_j X_k^2$, $\mathbb{E}X_i X_j^3$, $\mathbb{E}X_i X_j X_k X_l$ with i, j, k, l distinct here. By independence and $\mu = 0$, the last 3 types vanish, thus

$$\mathbb{E}(X_1 + \cdots + X_n)^4 = \sum_{i=1}^{n} \mathbb{E}X_i^4 + 3\sum_{i\neq j} \mathbb{E}X_i^2 \mathbb{E}X_j^2 \leq nC + 3n^2 C \leq 4Cn^2,$$

where we use in the estimate that $\mathbb{E}X_i^2 = \|X_i\|_2^2 \leq \|X_i\|_4^2 \leq C^{1/2}$ (by the monotonicity of moments – Example 6.4). Consequently, for $\varepsilon > 0$, by Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^4} \frac{\mathbb{E}(X_1 + \cdots + X_n)^4}{n^4} \leq \frac{1}{\varepsilon^4} \frac{4C}{n^2}.$$

Lemma 7.14 finishes the proof. □

Our major goal here is to show Etemadi's strong law of large numbers which assumes only pairwise independence.

7.16 Theorem (Etemadi)**.** *Let $X_1, X_2, \ldots$ be pairwise independent identically distributed random variables such that $\mathbb{E}|X_1| < \infty$. Let $\mu = \mathbb{E}X_1$. Then,*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

Proof. Since $X_n = X_n^+ - X_n^-$ and both the positive part X_n^+ and the negative part X_n^- of X_n satisfy the same assumptions as X_n, we can assume that $X_n \geq 0$ for every n.

As we have seen in Theorem 7.15, it is useful to be able to control higher moments to prove a.s. convergence. Since here we only assume existence of the first moment, we employ the technique of truncating (as in the weak law – Theorem 7.6). We divide the whole proof into several steps.

Step I (truncation). Let

$$\tilde{X}_k = X_k \mathbf{1}_{|X_k| \leq k}, \qquad k \geq 1$$

and

$$S_n = X_1 + \cdots + X_n, \qquad \tilde{S}_n = \tilde{X}_1 + \cdots + \tilde{X}_n.$$

We claim that for $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu$, it suffices to show that

$$\frac{\tilde{S}_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu. \tag{7.5}$$

We have,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(X_k \neq \tilde{X}_k\right) = \sum_{k=1}^{\infty} \mathbb{P}\left(|X_k| \geq k\right) = \sum_{k=1}^{\infty} \mathbb{P}\left(|X_1| \geq k\right) \leq \int_0^{\infty} \mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t$$
$$= \mathbb{E}|X_1| < \infty.$$

Thus, by the first Borel-Cantelli lemma,

$$\mathbb{P}\left(X_n = \tilde{X}_n \text{ eventually}\right) = 1.$$

Since on the event "$X_n = \tilde{X}_n$ eventually", we have

$$\left| \frac{X_1 + \cdots + X_n}{n} - \frac{\tilde{X}_1 + \cdots + \tilde{X}_n}{n} \right| \leq \frac{R}{n}$$

for some (random) R, our claim about (7.5) follows.

Step II (variance bounds). Here we show that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \operatorname{Var}(\tilde{X}_k) < \infty. \tag{7.6}$$

First note that

$$
\begin{aligned}
\operatorname{Var}(\tilde{X}_k) \leq \mathbb{E}\tilde{X}_k^2 &= \int_0^\infty 2t\mathbb{P}\left(|\tilde{X}_k| \geq t\right) \mathrm{d}t \\
&= \int_0^k 2t\mathbb{P}\left(|\tilde{X}_k| \geq t\right) \mathrm{d}t + \int_k^\infty 2t\mathbb{P}\left(|\tilde{X}_k| \geq t\right) \mathrm{d}t \\
&\leq \int_0^k 2t\mathbb{P}\left(|X_k| \geq t\right) \mathrm{d}t \\
&= \int_0^k 2t\mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sum_{k=1}^\infty \frac{1}{k^2} \operatorname{Var}(\tilde{X}_k) &\leq \sum_{k=1}^\infty \frac{1}{k^2} \int_0^k 2t\mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t \\
&= \sum_{k=1}^\infty \int_0^\infty \frac{1}{k^2} \mathbf{1}_{t<k} \cdot 2t\mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t \\
&= \int_0^\infty \left(\sum_{k>t} \frac{1}{k^2}\right) \cdot 2t\mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t.
\end{aligned}
$$

It is an elementary exercise to show that $\sum_{k>t} \frac{1}{k^2} \leq \frac{2}{t}$. Then the right hand side get upper-bounded by $4\int_0^\infty \mathbb{P}\left(|X_1| \geq t\right) \mathrm{d}t = 4\mathbb{E}|X_1|$, so (7.6) follows.

Step III (convergence on a subsequence). Fix $\alpha > 1$. Let $k_n = \lfloor \alpha^n \rfloor$. Our goal in this step is to show that

$$
\frac{\tilde{S}_{k_n}}{k_n} \xrightarrow[n\to\infty]{a.s.} \mu,
$$

that is (7.5) holds on the subsequence k_n. In the next step, thanks to the monotonicity of S_n (recall we assume that the X_n are nonnegative), we will extend this from the subsequence k_n to the convergence on all the terms.

Fix $\varepsilon > 0$. We have,

$$
\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{\tilde{S}_{k_n} - \mathbb{E}\tilde{S}_{k_n}}{k_n}\right| > \varepsilon\right) \leq \varepsilon^{-2} \sum_{n=1}^\infty \frac{\operatorname{Var}(\tilde{S}_{k_n})}{k_n^2} = \varepsilon^{-2} \sum_{n=1}^\infty \frac{1}{k_n^2} \sum_{j=1}^{k_n} \operatorname{Var}(\tilde{X}_j),
$$

where in the last equality we use pairwise independence. Changing the order of summation gives,

$$
\varepsilon^{-2} \sum_{n=1}^\infty \frac{1}{k_n^2} \sum_{j=1}^{k_n} \operatorname{Var}(\tilde{X}_j) = \varepsilon^{-2} \sum_{j=1}^\infty \operatorname{Var}(\tilde{X}_j) \sum_{n: k_n \geq j} \frac{1}{k_n^2}.
$$

By a simple estimate $\alpha^n \geq k_n \geq \frac{\alpha^n}{2}$,

$$
\sum_{n: k_n > j} \frac{1}{k_n^2} \leq \sum_{n: \alpha^n \geq j} \frac{4}{\alpha^{2n}} \leq \frac{4}{j^2} \frac{1}{1-\alpha^{-2}},
$$

and, as a result,

$$
\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{\tilde{S}_{k_n} - \mathbb{E}\tilde{S}_{k_n}}{k_n}\right| > \varepsilon\right) \leq \frac{4}{\varepsilon^2(1-\alpha^{-2})} \sum_{j=1}^\infty \frac{1}{j^2} \operatorname{Var}(\tilde{X}_j) < \infty,
$$

by (7.6). Therefore, by virtue of Lemma 7.14,

$$\frac{\tilde{S}_{k_n} - \mathbb{E}\tilde{S}_{k_n}}{k_n} \xrightarrow[n\to\infty]{a.s.} 0.$$

Moreover,

$$\frac{\mathbb{E}\tilde{S}_{k_n}}{k_n} = \frac{\mathbb{E}\tilde{X}_1 + \dots + \mathbb{E}\tilde{X}_{k_n}}{k_n} = \frac{\mathbb{E}X_1 + \dots + \mathbb{E}X_{k_n}}{k_n} - \frac{\mathbb{E}X_1 \mathbf{1}_{X_1 > 1} + \dots + \mathbb{E}X_{k_n} \mathbf{1}_{X_{k_n} > k_n}}{k_n}$$
$$= \mu - \frac{\mathbb{E}X_1 \mathbf{1}_{X_1 > 1} + \dots + \mathbb{E}X_1 \mathbf{1}_{X_1 > k_n}}{k_n} \xrightarrow[n\to\infty]{} \mu$$

because $\mathbb{E}X_1 \mathbf{1}_{X_1 > n} \to 0$ as $n \to 0$ (recall the Cesàro-type lemma – Remark 7.10).

Step IV (convergence on all terms). For every positive integer m, there is n such that $k_n \le m < k_{n+1}$ and then

$$\frac{\tilde{S}_{k_n}}{k_{n+1}} \le \frac{\tilde{S}_m}{m} \le \frac{\tilde{S}_{k_{n+1}}}{k_n}$$

(because $\tilde{X}_j \ge 0$ for every j). Thus, for every $\alpha > 1$, with probability 1,

$$\mu\alpha^{-1} = \liminf_{n\to\infty} \frac{\tilde{S}_{k_n}}{k_{n+1}} \le \liminf_{m\to\infty} \frac{\tilde{S}_m}{m} \le \limsup_{m\to\infty} \frac{\tilde{S}_m}{m} \le \limsup_{n\to\infty} \frac{\tilde{S}_{k_{n+1}}}{k_n} = \mu\alpha.$$

Taking, say $\alpha = 1 + \frac{1}{l}$ and letting $l \to \infty$, we get that with probability 1,

$$\mu \le \liminf_{m\to\infty} \frac{\tilde{S}_m}{m} \le \limsup_{m\to\infty} \frac{\tilde{S}_m}{m} \le \mu.$$

This shows (7.5) and finishes the proof. □

7.17 Remark. The assumption of integrability in the strong law of large numbers, Theorem 7.16, is necessary: *if $X_1, X_2, \dots$ are i.i.d. random variables such that there is a constant $c \in \mathbb{R}$ for which $\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1+\dots+X_n}{n} = c\right) > 0$, then $\mathbb{E}|X_1| < \infty$ and $c = \mathbb{E}X_1$.* We leave the proof as an exercise.

7.18 Remark. If we know that the expectation is infinite, the strong law in some sense still holds, but the limit is also infinite: *if $X_1, X_2, \dots$ are i.i.d. random variables such that one of the expectations $\mathbb{E}X_1^+, \mathbb{E}X_1^-$ is $+\infty$ and the other one is finite, then $\limsup_{n\to\infty} \left|\frac{X_1+\dots+X_n}{n}\right| = +\infty$ a.s.* We leave the proof as an exercise.

7.19 Remark. Let $(\Omega_0, \mathcal{F}_0, \mathbb{P}_0)$ be the infinite product of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Fix an event $A \in \mathcal{F}$. By the strong law of large numbers, for $\mathbb{P}_0$-a.e. point $\omega = (\omega_1, \omega_2, \dots) \in \Omega_0$, we have

$$\frac{\mathbf{1}_A(\omega_1) + \dots + \mathbf{1}_A(\omega_n)}{n} \xrightarrow[n\to\infty]{} \mathbb{E}\,\mathbf{1}_A = \mathbb{P}\left(A\right).$$

In other words, the probability of A is the limit of its frequency as the number of trials goes to ∞. This justifies the so-called *frequential* definition of probability.

We refer to the exercises for additional extensions and applications of the strong law of large numbers.

7.3 Exercises

1. Let X be a random variable and let X' be its independent copy. Show that for every $t \geq 0$,

$$\mathbb{P}\left(|X - X'| > 2t\right) \leq 2\mathbb{P}\left(|X| > t\right).$$

Moreover, if $a \geq 0$ is chosen such that $\mathbb{P}\left(X \leq a\right) \geq u$ and $\mathbb{P}\left(X \geq -a\right) \geq u$ for some $u \in [0, 1]$, then

$$\mathbb{P}\left(|X - X'| > t\right) \geq u\mathbb{P}\left(|X| > t + a\right).$$

In particular, if m is a median of X, then

$$\mathbb{P}\left(|X - X'| > t\right) \geq \frac{1}{2}\mathbb{P}\left(|X - m| > t\right).$$

2. Let $X_1, \ldots, X_n$ be independent symmetric random variables, that is X_i has the same distribution as $-X_i$. Then for every $t \geq 0$,

$$\mathbb{P}\left(|X_1 + \ldots + X_n| > t\right) \geq \frac{1}{2}\mathbb{P}\left(\max_{j \leq n} |X_j| > t\right).$$

In particular, if the X_i are identically distributed, then

$$\mathbb{P}\left(|X_1 + \ldots + X_n| > t\right) \geq \frac{1}{2}\big[1 - \exp\left\{-n\mathbb{P}\left(|X_1| > t\right)\right\}\big].$$

3. Using the symmetrisation from Exercise 7.1 and inequalities from Exercise 7.2, justify Remark 7.7.

4. Let $X_1, X_2, \ldots$ be independent random variables with

$$\mathbb{P}\left(X_n = n + 1\right) = \mathbb{P}\left(X_n = -(n + 1)\right) = \frac{1}{2(n + 1)\log(n + 1)},$$
$$\mathbb{P}\left(X_n = 0\right) = 1 - \frac{1}{(n + 1)\log(n + 1)}, \qquad n \geq 1.$$

Show that $(X_n)_{n=1}^{\infty}$ satisfies the weak law of large numbers, that is $\frac{X_1 + \ldots + X_n}{n}$ converges in probability. Show that $\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| > n\right) = \infty$ and conclude that $(X_n)_{n=1}^{\infty}$ does not satisfy the strong law of large numbers.

5. Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables with distribution function F. Define the sequence of empirical distribution functions by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

(which are random). Show that for every $x \in \mathbb{R}$, we have $\mathbb{P}\left(F_n(x) \to F(x)\right) = 1$.

6. Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathbb{P}\left(|X_i| < 1\right) = 1$. Show that $X_1 X_2 \cdot \ldots \cdot X_n$ converges to 0 a.s. and in L_1.

7. Let $X_1, X_2, \ldots$ be i.i.d. exponential random variables with parameter λ. Show that

a) $Y_n = \frac{X_1+\cdots+X_n}{X_1^2+\cdots+X_n^2}$

b) $Z_n = \frac{X_1X_2+X_2X_3+\cdots+X_nX_{n+1}}{n}$

converge a.s. and find their limits.

8. Let $X_1, X_2, \ldots$ be i.i.d. random variables with density g which is positive. Show that for every continuous function f such that $\int_{\mathbb{R}} |f| < \infty$, we have $\frac{1}{n}\sum_{i=1}^n \frac{f(X_i)}{g(X_i)} \xrightarrow[n\to\infty]{a.s.} \int_{\mathbb{R}} f$. (This provides a method of numerical integration.)

9. Let f be a continuous function on $[0,1]$ taking values in $[0,1]$. Let $X_1, Y_1, X_2, Y_2, \ldots$ be independent random variables uniformly distributed on $[0,1]$. Let $Z_i = \mathbf{1}_{\{f(X_i)>Y_i\}}$. Show that $\frac{1}{n}\sum_{i=1}^n Z_i$ converges almost surely to $\int_0^1 f$.

10. Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = -1)$ with $\frac{1}{2} < p < 1$. Let $S_n = X_1 + \ldots + X_n$ (a random walk with a drift to the right). Show that $S_n \xrightarrow[n\to\infty]{a.s.} \infty$.

11. Find
$$\lim_{n\to\infty} \int_0^1 \cdots \int_0^1 \frac{x_1^3 + \ldots + x_n^2}{x_1 + \ldots + x_n} \mathrm{d}x_1 \ldots \mathrm{d}x_n$$
(or show the limit does not exist).

12. Find
$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \int_0^1 \cdots \int_0^1 \sqrt{x_1^2 + \ldots + x_n^2} \mathrm{d}x_1 \ldots \mathrm{d}x_n$$
(or show the limit does not exist).

13. Suppose that f is a continuous function on $[0,1]$. Find
$$\lim_{n\to\infty} \int_0^1 \cdots \int_0^1 f(\sqrt[n]{x_1 \cdot \ldots \cdot x_n}) \mathrm{d}x_1 \ldots \mathrm{d}x_n$$
(or show the limit does not exist).

14. Let $X_1, X_2, \ldots$ be i.i.d. random variables uniform on $[-1,1]$. Does the sequence
$$\frac{X_1 + X_2^2 + \cdots + X_n^n}{n}, \qquad n = 1, 2, \ldots,$$
converge a.s.?

15. We say that a number $x \in [0,1]$ is *simply normal* in an integer base $b \geq 2$, if its sequence of digits is *uniform* in the sense that each of the digits $\{0, 1, \ldots, b-1\}$ occurs with the same density $1/b$, that is, formally, for every $d \in \{0, 1, \ldots, b-1\}$, we have
$$\lim_{n\to\infty} \frac{1}{n} |\{j \leq n,\ d_j(x) = d\}| = \frac{1}{b},$$
where $x = \sum_{k=1}^\infty \frac{d_k(x)}{b^k}$ with $d_1(x), d_2(x), \ldots \in \{0, 1, \ldots, b-1\}$. Show Borel's theorem: *almost every number $x \in [0,1]$ is simply normal in every base.* On the other hand,

it is not known whether $e-2$, $\pi-3$, or $\sqrt{2}-1$ are simply normal in any given base (although, it is of course widely believed so). It is not even known whether $\sqrt{2}$ has infinitely many 5's in its decimal expansion!

16. We say that a random variable X is *singular* if its cumulative distribution function is continuous and there is a subset A of $\mathbb{R}$ of Lebesgue measure 0 with $\mathbb{P}(X \in A) = 1$. Fix $p \in (0,1)$, $p \neq \frac{1}{2}$. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1-p$, $i = 1, 2, \ldots$. Show that $Y = \sum_{n=1}^{\infty} \frac{X_n}{2^n}$ is singular. What is the distribution of Y when $p = \frac{1}{2}$?

17. Let b be an integer, $b \geq 3$. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 0) = \frac{1}{2}$, $i = 1, 2, \ldots$. Show that $Y = \sum_{n=1}^{\infty} \frac{X_n}{b^n}$ is singular (in particular, for $b = 3$ we get the distribution from 2.22).

18. Justify Remark 7.17.

19. Justify Remark 7.18

20. Give an example of a sequence of i.i.d. random variables $X_1, X_2, \ldots$ for which $\frac{X_1+\ldots+X_n}{n}$ converges in probability but not a.s.

21. Let $f\colon [0,+\infty) \to \mathbb{R}$ be a continuous bounded function. Define its Laplace transform

$$L(t) = \int_0^{\infty} e^{-tx} f(x) \mathrm{d}x, \qquad t > 0.$$

Show that L is C^∞ and $L^{(n)}(t) = \int_0^\infty (-x)^n e^{-tx} f(x) \mathrm{d}x$. Let S_n be the sum of n i.i.d. exponential random variables with parameter t. Show that

$$\mathbb{E} f(S_n) = (-1)^{n-1} \frac{t^n L^{(n-1)}(t)}{(n-1)!}$$

and deduce the following inversion formula for the Laplace transform

$$f(y) = \lim_{n\to\infty} (-1)^{n-1} \frac{(n/y)^n L^{(n-1)}(n/y)}{(n-1)!}, \qquad y > 0.$$

8 Weak convergence

8.1 Definition and equivalences

We start with a general definition. We say that a sequence (μ_n) of probability measures on a metric space (E,d) converges **weakly** to a Borel probability measure μ on (E,d) if for every bounded continuous function $f\colon E\to\mathbb{R}$, we have

$$\int_E f\mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_E f\mathrm{d}\mu.$$

The following equivalences explain weak convergence on the level of sets. It is sometimes referred to as the Portmanteau theorem.

8.1 Theorem. *Let $\mu,\mu_1,\mu_2,\dots$ be Borel probability measures on a metric space (E,d). The following are equivalent*

(i) $\mu_n\to\mu$ *weakly,*

(ii) $\limsup\mu_n(F)\le\mu(F)$ *for all closed sets* F *in* E*,*

(iii) $\liminf\mu_n(G)\ge\mu(G)$ *for all open sets* G *in* E*,*

(iv) $\mu_n(A)\to\mu(A)$ *for all Borel sets* A *in* E *with* $\mu(\partial A)=0$*.*

Proof. (i) $\Rightarrow$ (ii): Fix $\varepsilon>0$ and let

$$g_\varepsilon(x)=\left(1-\frac{1}{\varepsilon}d(x,F)\right)_+,\qquad x\in E,$$

where as usual $d(x,F)=\inf\{d(x,y),\ y\in F\}$ is the distance from x to the set F. We also define

$$F_\varepsilon=\{x\in E,\ d(x,F)\le\varepsilon\},$$

the ε-enlargement of F. Note that for every $x\in E$,

$$\mathbf{1}_F(x)\le g_\varepsilon(x)\le\mathbf{1}_{F_\varepsilon}(x).$$

Moreover, the function g_ε is bounded and continuous (it is $1/\varepsilon$-Lipschitz). Thus,

$$\begin{aligned}\limsup\mu_n(F)&=\limsup\int_E\mathbf{1}_F\ \mathrm{d}\mu_n\\&\le\limsup\int_E g_\varepsilon\mathrm{d}\mu_n\\&=\int_E g_\varepsilon\mathrm{d}\mu\\&\le\int_E\mathbf{1}_{F_\varepsilon}\ \mathrm{d}\mu\\&=\mu(F_\varepsilon),\end{aligned}$$

where in the second equality we used (i) and the inequalities follow from the pointwise bounds $\mathbf{1}_F \leq g_\varepsilon \leq \mathbf{1}_{F_\varepsilon}$. Letting $\varepsilon \to 0$ in a decreasing way, we get $\mu(F_\varepsilon) \to \mu(F)$, by continuity of probability measures. This shows (ii).

(ii) $\Leftrightarrow$ (iii): We use complements.

(ii) & (iii) $\Rightarrow$ (iv): Let $A \subset E$ be a Borel set such that $\mu(\partial A) = 0$. Since $\mathrm{cl}(A) = A \cup \partial A$ and $\mathrm{int}(A) = A \setminus \partial A$, we get

$$\limsup \mu_n(A) \leq \limsup \mu_n(\mathrm{cl}(A)) = \mu(\mathrm{cl}(A)) = \mu(A),$$
$$\liminf \mu_n(A) \geq \liminf \mu_n(\mathrm{int}(A)) = \mu(\mathrm{int}(A)) = \mu(A).$$

These show that $\lim \mu_n(A) = \mu(A)$.

(iv) $\Rightarrow$ (ii): Let $F \subset E$ be a closed set and for $\varepsilon > 0$. We set as before $F_\varepsilon = \{x \in E,\ d(x,F) \leq \varepsilon\}$. Let $S_\varepsilon = \{x \in E,\ d(x,F) = \varepsilon\}$. Note that $\partial F_\varepsilon \subset S_\varepsilon$. Since the sets S_ε are disjoint for different ε, for only countably many ε we can have $\mu(S_\varepsilon) > 0$. Thus we can find a decreasing sequence $\varepsilon_k \to 0$ such that $\mu(S_{\varepsilon_k}) = 0$ for every k. Consequently, $\mu(\partial F_{\varepsilon_k}) = 0$ for every k. We get by (iv) that

$$\limsup \mu_n(F) \leq \limsup \mu_n(F_{\varepsilon_k}) = \mu(F_{\varepsilon_k})$$

and $\mu(F_{\varepsilon_k}) \to \mu(F)$ by continuity. These show (ii).

(iii) $\Rightarrow$ (i): Let $f\colon E \to \mathbb{R}$ be a bounded continuous function. Suppose $f \geq 0$ (otherwise, we consider $f - \inf f$). Using Fatou's lemma and (iii) (sets $\{x \in E,\ f(x) > t\}$ are open), we get

$$\begin{aligned}
\liminf \int_E f \mathrm{d}\mu_n &= \liminf \int_0^\infty \mu_n(\{x \in \mathbb{E},\ f(x) > t\}) \mathrm{d}t \\
&\geq \int_0^\infty \liminf \mu_n(\{x \in \mathbb{E},\ f(x) > t\}) \mathrm{d}t \\
&\geq \int_0^\infty \mu(\{x \in \mathbb{E},\ f(x) > t\}) \mathrm{d}t \\
&= \int_E f \mathrm{d}\mu.
\end{aligned}$$

Since this inequality holds for an arbitrary function, applying it to $-f$ and combining the two gives $\int_E f \mathrm{d}\mu_n \to \int_E f \mathrm{d}\mu$, as required. $\square$

The following equivalences show that even the smaller set of all Lipschitz functions captures the same.

8.2 Theorem. *Let $\mu, \mu_1, \mu_2, \ldots$ be Borel probability measures on a metric space (E,d). The following are equivalent* *(i)* $\mu_n \to \mu$ *weakly,*

(ii) for every bounded uniformly continuous function $f\colon E \to \mathbb{R}$*, we have*

$$\int_E f \mathrm{d}\mu_n \to \int_E f \mathrm{d}\mu,$$

(iii) for every bounded Lipschitz function $f\colon E \to \mathbb{R}$*, we have*

$$\int_E f \mathrm{d}\mu_n \to \int_E f \mathrm{d}\mu.$$

Proof. The implications (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) are clear. To show (iii) $\Rightarrow$ (i), we use condition (ii) of Theorem 8.1: let F be a closed set and for $\varepsilon > 0$, let

$$g_\varepsilon(x) = \left(1 - \frac{1}{\varepsilon} d(x, F)\right)_+, \qquad x \in E,$$

which is a bounded Lipschitz function. Therefore using (iii), we can repeat verbatim the argument "(i) $\Rightarrow$ (ii)" of the proof of Theorem 8.1 to show that $\limsup \mu_n(F) \leq \mu(F)$. $\square$

We immediately get the following corollaries.

8.3 Corollary. *Let* μ *and* ν *be Borel probability measures on a metric space* (E, d)*. If*

$$\int_E f \mathrm{d}\mu = \int_E f \mathrm{d}\nu$$

for every function $f\colon E \to \mathbb{R}$ *which is bounded and Lipschitz, then* $\mu = \nu$*.*

Proof. Letting $\mu_1 = \mu_2 = \ldots = \mu$, we get by the assumption and Theorem 8.2 (iii) that $\mu_n \to \nu$ weakly. Therefore, for every closed set F, we have $\mu(F) = \limsup \mu_n(F) \leq \nu(F)$. By symmetry, we get the reverse inequality as well, thus $\mu(F) = \nu(F)$. Since the closed sets generate the Borel σ-algebra, by Dynkin's theorem on π-λ systems, we get $\mu = \nu$. $\square$

8.4 Corollary. *Weak limits are uniquely determined, that is if* $\mu_n \to \mu$ *and* $\mu_n \to \nu$ *weakly, then* $\mu = \nu$*.*

Proof. It follows from Theorem 8.2 (iii) and the previous corollary. $\square$

Convergence in distribution of random variables

We say that a sequence of random variables (X_n) converges to a random variable X **in distribution (or in law)**, denoted $X_n \xrightarrow[n\to\infty]{d} X$, if $\mu_{X_n} \to \mu_X$ weakly, that is for every continuous bounded function $f\colon \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E} f(X_n) \to \mathbb{E} f(X).$$

Of course, this definition easily extends to random variables taking values in a metric space. Note that this notion of convergence only depends on the law of random variables involved and not on their particular realisations as functions on a probability space (in fact, they can be defined on different probability spaces). Particularly, if $X_n \xrightarrow[n\to\infty]{d} X$ and say X' is another random variable with the same distribution as X (i.e. $\mu_{X'} = \mu_X$), then we can also write $X_n \xrightarrow[n\to\infty]{d} X'$.

In the case of real-valued random variables, there is an intuitive equivalent formulation in terms of distribution functions.

8.5 Theorem. *A sequence (X_n) of random variables converges in distribution to a random variable X if and only if*

$$F_{X_n}(t) \xrightarrow[n\to\infty]{} F_X(t) \qquad \textit{for every point of continuity of } F_X. \tag{8.1}$$

Proof. ($\Rightarrow$): For parameters $t \in \mathbb{R}$ and $\varepsilon > 0$ define the continuous bounded functions

$$g_{t,\varepsilon}(x) = \begin{cases} 1, & x \leq t, \\ 1 - \frac{x-t}{\varepsilon}, & t < x \leq t + \varepsilon, \\ 0, & x > t + \varepsilon. \end{cases}$$

The idea is that these functions are continuous approximations of indicator functions. We have, $\mathbf{1}_{\{x \leq t\}} \leq g_{t,\varepsilon}(x) \leq \mathbf{1}_{\{x \leq t+\varepsilon\}}$. Consequently,

$$\begin{aligned} \limsup \mathbb{P}\left(X_n \leq t\right) = \limsup \mathbb{E}\,\mathbf{1}_{\{X_n \leq t\}} &\leq \limsup \mathbb{E} g_{t,\varepsilon}(X_n) \\ &= \mathbb{E} g_{t,\varepsilon}(X) \leq \mathbb{E}\,\mathbf{1}_{\{X \leq t+\varepsilon\}} = \mathbb{P}\left(X \leq t + \varepsilon\right). \end{aligned}$$

Letting $\varepsilon \to 0$ gives

$$\limsup F_{X_n}(t) \leq F_X(t).$$

On the other hand, since

$$\begin{aligned} \liminf \mathbb{P}\left(X_n \leq t\right) = \liminf \mathbb{E}\,\mathbf{1}_{\{X_n \leq t\}} &\geq \liminf \mathbb{E} g_{t-\varepsilon,\varepsilon}(X_n) \\ &= \mathbb{E} g_{t-\varepsilon,\varepsilon}(X) \geq \mathbb{E}\,\mathbf{1}_{\{X \leq t-\varepsilon\}} = \mathbb{P}\left(X \leq t - \varepsilon\right) \end{aligned}$$

after taking $\varepsilon \to 0$, we get

$$\liminf F_{X_n}(t) \geq F_X(t-).$$

If t is a point of continuity of F_X, $F_X(t-) = F_X(t)$ and we obtain $\lim F_{X_n}(t) = F_X(t)$, which means $X_n \xrightarrow{d} X$.

($\Leftarrow$): We first show a lemma which allows us to relate condition (8.1) to a.s. convergence.

8.6 Lemma. *If random variables $X, X_1, X_2, \ldots$ satisfy* (8.1)*, then there are random variables $Y, Y_1, Y_2, \ldots$ such that Y_n has the same distribution as X_n, Y has the same distribution as X and $Y_n \to Y$ a.s.*

Proof. Let $F_n = F_{X_n}$ be the distribution function of X_n and let $F = F_X$ be the distribution function of X. Let $\Omega = (0,1)$, $\mathcal{F}$ be the Borel subsets of $(0,1)$ and $\mathbb{P}(\cdot)$ be uniform. For every $x \in (0,1)$ define the "inverse" distribution functions

$$Y_n(x) = \sup\{y \in \mathbb{R},\ F_n(y) < x\}$$

and similarly

$$Y(x) = \sup\{y \in \mathbb{R},\ F(y) < x\}.$$

By the construction, $F_{Y_n} = F_n$ and $F_Y = F$. Note that Y_n and Y are nondecreasing right-continuous functions whose only discontinuities are jumps which happen at at most countably many points. If we let Ω_0 to be the set of points where Y is continuous, then $\mathbb{P}(\Omega_0) = 1$. Fix $x \in \Omega_0$. We claim that $Y_n(x) \to Y(x)$, which then gives $Y_n \to Y$ a.s. We have

1. $\liminf Y_n(x) \geq Y(x)$, for suppose $y < Y(x)$ is a continuity point of F; then $F(y) < x$ (since $x \in \Omega_0$), so for large n, $F_n(y) < x$ and by the definition of the supremum, $y \leq Y_n(x)$. Taking $\liminf$, we get $\liminf Y_n(x) \geq y$ for every $y < Y(x)$, so $\liminf Y_n(x) \geq Y(x)$.

2. $Y(x) \geq \limsup Y_n(x)$, for suppose $y > Y(x)$ is a continuity point of F; then $F(y) > x$, so for large n, $F_n(y) > x$ which gives $y \geq Y_n(x)$. Taking $\limsup$ finishes the argument.

□

Let Y_n and Y be as in Lemma 8.6, $Y_n \xrightarrow{a.s.} Y$. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a bounded continuous function. Since we also have $f(Y_n) \xrightarrow{a.s.} f(Y)$, so by Lebesgue's dominated convergence theorem (f is bounded),

$$\mathbb{E}g(X_n) = \mathbb{E}g(Y_n) \to \mathbb{E}g(Y) = \mathbb{E}g(X).$$

□

8.7 Example. Let ε be a symmetric random sign. Consider the sequence $(X_n)_{n=1}^{\infty} = (\varepsilon, -\varepsilon, \varepsilon, -\varepsilon, \ldots)$. Since $-\varepsilon$ has the same distribution as ε, we have $F_{X_n} = F_\varepsilon$ for every n, so $X_n \xrightarrow{d} \varepsilon$. On the other hand, the sequence (X_n) does not converge in probability, for suppose $X_n \xrightarrow{\mathbb{P}} X$ for some random variable X. Then for n, m large enough $\mathbb{P}(|X_n - X_m| > 1) \leq \mathbb{P}(|X_n - X| > 1/2) + \mathbb{P}(|X - X_m| > 1/2) \leq 1/4$. Taking n and m of different parity, we get $\mathbb{P}(|X_n - X_m| > 1) = \mathbb{P}(|2\varepsilon| > 1) = 1$, a contradiction.

8.8 Example. Let X be a random variable and consider the sequence $X_n = X + \frac{1}{n}$. For any reasonable definition of "convergence in distribution" we should have $X_n \to X$. Note that for a fixed $t \in \mathbb{R}$, we have

$$\lim F_{X_n}(t) = \lim \mathbb{P}(X_n \leq t) = \lim \mathbb{P}\left(X \leq t - \frac{1}{n}\right) = F(t-),$$

which is $F(t)$ if and only if t is a continuity point of F. This explains why in the definition we make this exclusion.

Convergence in distribution is metrisable. For two distribution functions $F, G\colon \mathbb{R} \to [0,1]$, we let

$$\rho(F,G) = \inf\{\varepsilon > 0,\ \forall x \in \mathbb{R}\ G(x-\varepsilon) - \varepsilon \leq F(x) \leq G(x+\varepsilon) + \varepsilon\},$$

which is called the **Lévy metric**. We naturally extend this definition for random variables, setting $\rho(X,Y) = \rho(F_X, F_Y)$ for random variables X, Y defined on the same probability space.

8.9 Theorem. *Let L_0 be the set of all random variables on a given probability space. Then ρ is a metric on L_0 and (L_0, ρ) is separable and complete. Moreover, $X_n \to X$ in distribution if and only if $\rho(X_n, X) \to 0$.*

We leave the proofs as exercises.

8.2 Relations to other notions of convergence and basic algebraic properties

Lemma 8.6 explains the relation between convergence in distribution and a.s. convergence. If the random variables are defined on the same probability space, then convergence in distribution is the weakest of all types of convergences we have seen.

8.10 Theorem. *Let $X, X_1, X_2, \ldots$ be random variables such that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$. Then, we also have $X_n \xrightarrow[n\to\infty]{d} X$.*

Proof. Suppose the assertion does not hold. Then, there is a bounded continuous function $f\colon \mathbb{R} \to \mathbb{R}$ and $\varepsilon > 0$ such that $|\mathbb{E}f(X_n) - \mathbb{E}f(X)| > \varepsilon$ for infinitely many n, say for $n_1 < n_2 < \ldots$. By Theorem 6.16, there is a subsequence n_{k_l} such that $X_{n_{k_l}}$ converges to X a.s., but then, by Lebesgue's dominated convergence theorem, we get a contradiction with $|\mathbb{E}f(X_{n_{k_l}}) - \mathbb{E}f(X)| > \varepsilon$. □

We record basic algebraic properties and defer their proofs to exercises.

8.11 Theorem. *Let $(X_n), (Y_n)$ be sequences of random variables such that $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{d} c$ for some random variable X and a constant $c \in \mathbb{R}$. Then*

$$X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$$

and

$$X_n Y_n \xrightarrow[n\to\infty]{d} cX.$$

8.12 Remark. There are examples showing that if the sequence (Y_n) converges to a (non-constant) random variable Y, then it is not true that $X_n + Y_n \xrightarrow[n\to\infty]{d} X + Y$, $X_nY_n \xrightarrow[n\to\infty]{d} XY$.

We finish with a simple fact concerning images of convergent sequences under "essentially" continuous maps.

8.13 Theorem. *Let $h\colon E \to E'$ be a Borel function between two metric spaces and let D_h be the set of its discontinuity points. If $X, X_1, X_2, \ldots$ are E-valued random variables such that $X_n \xrightarrow[n\to\infty]{d} X$ and $\mathbb{P}(X \in D_h) = 0$, then $h(X_n) \xrightarrow[n\to\infty]{d} h(X)$.*

Proof. We shall use condition (ii) of Theorem 8.1. Let F be a closed set in E'. We have,

$$\begin{aligned}\limsup \mathbb{P}(h(X_n) \in F) = \limsup \mathbb{P}\left(X_n \in h^{-1}(F)\right) &\leq \limsup \mathbb{P}\left(X_n \in \mathrm{cl}(h^{-1}(F))\right) \\ &\leq \mathbb{P}\left(X \in \mathrm{cl}(h^{-1}(F))\right),\end{aligned}$$

where the last inequality follows because $X_n \xrightarrow[n\to\infty]{d} X$. Since $\mathrm{cl}(h^{-1}(F)) \subset h^{-1}(F) \cup D_h$ and $\mathbb{P}(X \in D_h) = 0$, the right hand side equals $\mathbb{P}\left(X \in h^{-1}(F)\right) = \mathbb{P}(h(X) \in F)$, showing that $h(X_n) \xrightarrow[n\to\infty]{d} h(X)$. □

8.3 Compactness

Being able to extract convergent subsequences often helps. For real-valued random variables, we can work with their distribution functions to establish weak convergence. Since distribution functions are bounded and monotone, extracting convergent subsequences is always possible, as stated in the next theorem.

8.14 Theorem (Helly's selection theorem)**.** *If $(F_n)_n$ is a sequence of distribution functions, then there is a subsequence $(F_{n_k})_k$ and a right-continuous nondecreasing function $F : \mathbb{R} \to [0,1]$ such that $F_{n_k}(t) \xrightarrow[k\to\infty]{} F(t)$ for every point t of continuity of F.*

8.15 Remark. In general, F may not be a distribution function – it may happen that $F(\infty) < 1$ or $F(-\infty) > 0$.

Proof. To construct the desired subsequence we use a standard diagonal argument. Let $q_1, q_2, \ldots$ be a sequence of all rationals. Since the sequence $F_n(q_1)$ is bounded, it has a convergent subsequence, say $F_{n_k^{(1)}}(q_1)$ converges to $G(q_1)$. Then we look at the sequence $F_{n_k^{(1)}}(q_2)$ which is bounded, so it has a convergent subsequence, say $F_{n_k^{(2)}}(q_2)$ converges to $G(q_2)$, etc. We obtain subsequences $(n_k^{(l)})$ such that $(n_k^{(l+1)})$ is a subsequence of $(n_k^{(l)})$ and $F_{n_k^{(l)}}(q_l)$ converges to $G(q_l)$. Choose the diagonal subsequence $n_k = n_k^{(k)}$. Then $F_{n_k^{(k)}}(q_l)$ converges to $G(q_l)$ for every l. The function $G : \mathbb{Q} \to [0,1]$ obtained as the limit is nondecreasing. We extend it to the nondecreasing function $F : \mathbb{R} \to [0,1]$ by

$$F(x) = \inf\{G(q),\ q \in \mathbb{Q}, q > x\}, \qquad x \notin \mathbb{Q}.$$

The function F, as monotone, satisfies $F(x-) \le F(x) \le F(x+)$ for every x. At the points x, where F is not right-continuous, we modify it and set $F(x) = F(x+)$ (there are at most countably many such points).

It remains to check that F_{n_k} converges to F at its points of continuity. Let x be such a point and let q, r be rationals such that $q < x < r$. Then

$$\begin{aligned} F(q) = G(q) = \liminf_k F_{n_k}(q) &\le \liminf_k F_{n_k}(x) \\ &\le \limsup_k F_{n_k}(x) \le \limsup_k F_{n_k}(r) = G(r) = F(r). \end{aligned}$$

Letting $q, r \to x$, we get $F(q), F(r) \to F(x)$, so $\liminf_k F_{n_k}(x) = \limsup_k F_{n_k}(x) = F(x)$. $\square$

To capture when the limiting function is a distribution function of a random variable, we need the notion of tightness. A sequence (X_n) of random variables is **tight** if for every $\varepsilon > 0$, there is $M > 0$ such that $\mathbb{P}(|X_n| \le M) > 1 - \varepsilon$ for every n.

8.16 Remark. If there is $\delta > 0$ such that $C = \sup_n \mathbb{E}|X_n|^\delta < \infty$, then the sequence (X_n) is tight. Indeed, by Chebyshev's inequality,

$$\mathbb{P}(|X_n| > M) \le M^{-\delta}\mathbb{E}|X_n|^\delta \le \frac{C}{M^\delta}$$

which is less than ε for M large enough.

The main result of this section is the following compactness type result. It gives a necessary and sufficient condition for existence of convergent subsequences in distribution in terms of tightness.

8.17 Theorem. *A sequence of random variables (X_n) is tight if and only every subsequence $(X_{n_k})_k$ has a subsequent $(X_{n_{k_l}})_l$ which converges in distribution to some random variable.*

Proof. Let F_n be the distribution function of X_n.

($\Rightarrow$) By Helly's theorem applied to $(F_{n_k})_k$, there is a subsequence $(F_{n_{k_l}})_l$ which converges to a right-continuous nondecreasing function $F : \mathbb{R} \to [0,1]$ pointwise at the points of continuity of F. It remains to check that F is a distribution function, that is $F(-\infty) = 0$ and $F(+\infty) = 1$. By tightness, there is $M > 0$ such that $F_n(M) - F_n(-M) > 1 - \varepsilon$, for every n and we can further arrange that $-M$ and M are points of continuity of F. Taking $n = n_{k_l}$ and letting $l \to \infty$, we get $F(M) - F(-M) \ge 1 - \varepsilon$. Since ε is arbitrary and F is monotone, this yields $F(-\infty) = 0$ and $F(+\infty) = 1$.

($\Leftarrow$) If (X_n) is not tight, there is $\varepsilon > 0$ and an increasing sequence of indices n_k such that $\mathbb{P}(|X_{n_k}| \le k) \le 1 - \varepsilon$ for every k. By the assumption, $X_{n_{k_l}} \xrightarrow[l\to\infty]{d} X$. Let $x < 0 < y$ be points of continuity of F_X. Then

$$F_X(y) - F_X(x) = \lim_l (F_{n_{k_l}}(y) - F_{n_{k_l}}(x)) \le \limsup_l (F_{n_{k_l}}(k_l) - F_{n_{k_l}}(-k_l)) \le 1 - \varepsilon.$$

Taking $x \to -\infty$ and $y \to \infty$ gives $1 \leq 1 - \varepsilon$, a contradiction. $\square$

8.18 Remark. Theorem 8.17 can be greatly generalised to the setting of arbitrary separable complete metric spaces (the so-called Polish spaces), which is called Prokhorov's theorem. As we have seen, in the real-valued case, we take huge advantage of CDFs. In general, the proof is much more complicated. It can be also quite easily deduced from the Banach-Alaoglu's concerning compactness of weak-$*$ convergence (which can be identified with the notion of weak convergence).

8.4 Exercises

1. Give an example of Borel probability measures $\mu, \mu_1, \mu_2, \ldots$ on $\mathbb{R}$ and a Borel set B in $\mathbb{R}$ such that $\mu_n \to \mu$ weakly and $\mu_n(B) \not\to \mu(B)$.

2. Let $X_1, X_2, \ldots$ be random variables such that $\mathbb{P}\left(X_n = \frac{k}{n}\right) = \frac{1}{n}$, $k = 1, \ldots, n$, $n = 1, 2, \ldots$. Does the sequence (X_n) converge in distribution? If yes, find the limiting distribution.

3. Let $U_1, U_2, \ldots$ be i.i.d. random variables uniformly distributed on $[0, 1]$. Let $X_n = \min\{U_1, \ldots, U_n\}$. Show that nX_n converges in distribution to an exponential random variable with parameter one.

4. Suppose that $X, X_1, X_2, \ldots$ are nonnegative integer-valued random variables. Show that $X_n \xrightarrow[n\to\infty]{d} X$, if and only if $\mathbb{P}(X_n = k) \xrightarrow[n\to\infty]{} \mathbb{P}(X = k)$, for every $k = 0, 1, 2, \ldots$.

5. For $p \in [0, 1]$, let X_p be a Geometric random variable with parameter p. Show that the sequence $(\frac{1}{n} X_{1/n})$ converges in distribution to an exponential random variable with parameter 1.

6. Suppose that a sequence of random variables converges in distribution to a constant. Then it also converges in probability.

7. Prove Theorem 8.11 and Remark 8.12.

8. Prove Scheffé's lemma: If $X_1, X_2, \ldots$ is a sequence of continuous random variables with densities $f_1, f_2, \ldots$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for every $x \in \mathbb{R}$ for some probability density f, then $\int_{\mathbb{R}} |f - f_n| \xrightarrow[n\to\infty]{} 0$. Conclude that then $X_n \xrightarrow{d} X$ for a random variable X with density f (in other words, pointwise convergence of densities implies convergence in distribution). Considering $f_n(x) = (1 + \cos(2\pi n x))\,\mathbf{1}_{[0,1]}(x)$, show that the converse statement does not hold.

9. Let $X_1, X_2, \ldots$ be i.i.d. random variables uniform on $\{1, 2, \ldots, n\}$. Let

$$N_n = \min\{l \geq 2,\ X_k = X_l \text{ for some } k < l\}.$$

   Show that $\mathbb{P}(N_n > k) = \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right)$, for every integer $k \geq 1$ (the birthday problem). For every $t \geq 0$ show that $\lim_{n\to\infty} \mathbb{P}\left(\frac{N_n}{\sqrt{n}} > t\right) = e^{-t^2/2}$ and show that the sequence $(\frac{N_n}{\sqrt{n}})$ converges in distribution to a random variable with density function $x e^{-x^2/2}\,\mathbf{1}_{\{x \geq 0\}}$.

10. Let $X_1, X_2, \ldots$ be i.i.d. exponential random variables with parameter 1. Let $M_n = \max\{X_1, \ldots, X_n\}$. Show that $M_n - \log n$ converges in distribution to a random variable with the distribution function $e^{-e^{-x}}$, $x \in \mathbb{R}$.

11. Let $U_1, \ldots, U_{2n+1}$ be i.i.d. random variables uniform on $[0,1]$. Order them in a nondecreasing way and call the $n+1$ term (the middle one) M_n. Show that M_n has density $(2n+1)\binom{2n}{n}x^n(1-x)^n \mathbf{1}_{[0,1]}(x)$. Find $\mathbb{E}M_n$ and $\mathrm{Var}(M_n)$. Show that $\sqrt{8n}\left(M_n - \frac{1}{2}\right)$ converges in distribution to a standard Gaussian random variable.

12. Show that for positive t, $\int_t^\infty e^{-x^2/2}\mathrm{d}x \leq \frac{1}{t}e^{-t^2/2}$ and $\int_t^\infty e^{-x^2/2}\mathrm{d}x \geq \frac{t}{t^2+1}e^{-t^2/2}$. Conclude that for a standard Gaussian random variable Z and positive t,

$$\frac{1}{\sqrt{2\pi}}\frac{t}{t^2+1}e^{-t^2/2} \leq \mathbb{P}\left(Z > t\right) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-t^2/2}$$

and

$$\lim_{t\to\infty}\frac{\mathbb{P}\left(Z > t\right)}{\frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-t^2/2}} = 1.$$

13. Let $X_1, X_2, \ldots$ be i.i.d. standard Gaussian random variables. For $n = 2, 3, \ldots$ let b_n be such that $\mathbb{P}\left(X_1 > b_n\right) = \frac{1}{n}$. Show that $\lim_{n\to\infty}\frac{b_n}{\sqrt{2\log n}} = 1$. Let $M_n = \max\{X_1, \ldots, X_n\}$. Show that $b_n(M_n - b_n)$ converges in distribution to a random variable with the distribution function $e^{-e^{-x}}$, $x \in \mathbb{R}$.

    Hint: Using Exercise 8.12, first show that for every $a \in \mathbb{R}$, $\lim_{t\to\infty}\frac{\mathbb{P}\left(X_1 > t + \frac{a}{t}\right)}{\mathbb{P}(X_1 > t)} = e^{-a}$.

14. Assume that $X_1, X_2, \ldots$ are i.i.d. standard Gaussian random variables. Define $M_n = \max\{X_1, \ldots, X_n\}$. Show that $\frac{M_n}{\sqrt{2\log n}} \xrightarrow{\mathbb{P}} 1$.

15. Let (X_n) and (Y_n) be two sequences of random variables such that $X_n \xrightarrow[n\to\infty]{d} 0$ and $X_nY_n \xrightarrow[n\to\infty]{d} Z$ for some random variable Z. Prove that for a function f differentiable at 0, we have $(f(X_n) - f(0))Y_n \xrightarrow[n\to\infty]{d} f'(0)Z$.

16. Prove that if $X_n \xrightarrow[n\to\infty]{d} X$, then $\mathbb{E}|X| \leq \liminf \mathbb{E}|X_n|$.

17. Prove that if $X_n \xrightarrow[n\to\infty]{d} X$ and $\sup_n \mathbb{E}|X_n|^{p+\varepsilon} < \infty$ for some $p, \varepsilon > 0$, then $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|X_n|^p \to \mathbb{E}|X|^p$ and $\mathbb{E}X_n^p \to \mathbb{E}X^p$.

18. Suppose that $X_1, X_2, \ldots$ are nonnegative random variables such that for some $0 < \alpha < \beta$, we have $\mathbb{E}X_n^\alpha \to 1$ and $\mathbb{E}X_n^\beta \to 1$ as $n \to \infty$. Then $X_n \to 1$ in probability.

19. Prove Theorem 8.9.

20. Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}^n$ uniformly distributed on the sphere $\{x \in \mathbb{R}^n, x_1^2 + \ldots + x_n^2 = n\}$. Show that X_1 converges in distribution to a standard Gaussian random variable.

9 Characteristic functions

9.1 Definition and basic properties

The **characteristic function** of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ defined as

$$\phi_X(t) = \mathbb{E}e^{itX}, \qquad t \in \mathbb{R}.$$

(For complex valued random variables, say $Z = X + iY$, we of course define $\mathbb{E}Z = \mathbb{E}X + i\mathbb{E}Y$ provided that $\mathbb{E}|Z| < \infty$.) Since $e^{ix} = \cos x + i \sin x$, $x \in \mathbb{R}$ is a complex number of modulus 1, e^{itX} is a bounded random variable hence its expectation exists, so ϕ_X is well-defined on $\mathbb{R}$. We also use the notation $\phi_\mu(t)$ to denote the characteristic function of a Borel probability measure μ on $\mathbb{R}$ (i.e., of a random variable with law μ),

$$\phi_\mu(t) = \int_{\mathbb{R}} e^{itx} \mathrm{d}\mu(x), \qquad t \in \mathbb{R}.$$

9.1 Example. For a symmetric random sign ε,

$$\phi_\varepsilon(t) = \mathbb{E}e^{it\varepsilon} = \frac{e^{it} + e^{-it}}{2} = \cos t.$$

9.2 Example. For an exponential random variable X with parameter λ,

$$\phi_X(t) = \mathbb{E}e^{itX} = \int_{-\infty}^{\infty} e^{itx} f_X(x) \mathrm{d}x = \int_0^\infty \lambda e^{(it-\lambda)x} \mathrm{d}x = \lambda \frac{e^{-\lambda x} e^{itx}}{it - \lambda} \Bigg|_0^\infty = \frac{\lambda}{\lambda - it}$$

(when taking the limit $x \to \infty$, we use that e^{itx} is bounded).

We gather several basic properties in the following theorem.

9.3 Theorem. *Let X be a random variable with characteristic function ϕ_X. Then*

(i) $|\phi_X(t)| \le 1$, $t \in \mathbb{R}$,

(ii) $\phi_X(0) = 1$,

(iii) $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$,

(iv) ϕ_X *is uniformly continuous,*

(v) if $\mathbb{E}|X|^n < \infty$ *for some positive integer* n, *then the* n*th derivative* $\phi_X^{(n)}$ *exists, equals* $\phi_X^{(n)}(t) = i^n \mathbb{E}X^n e^{itX}$ *and is uniformly continuous.*

Proof. (i), (ii), (iii): These are easy to directly verify, $|\phi_X(t)| = |\mathbb{E}e^{itX}| \le \mathbb{E}|e^{itX}| = 1$ and $\phi_X(0) = \mathbb{E}e^{i \cdot 0 \cdot X} = 1$ and $\phi_{aX+b}(t) = \mathbb{E}e^{it(aX+b)} = e^{itb}\phi_X(at)$.

(iv): For every $t, h \in \mathbb{R}$,

$$|\phi_X(t+h) - \phi_X(t)| = |\mathbb{E}e^{itX}(e^{ihX} - 1)| \le \mathbb{E}|e^{ihX} - 1| \xrightarrow[h \to 0]{} 0$$

where the limit is justified by Lebesgue's dominated convergence theorem ($|e^{ihX}-1| \to 0$ pointwise and the sequence is bounded by 2). This implies the continuity of ϕ_X at t. The continuity is uniform because the bound does not depend on t.

(v): Fix n such that $\mathbb{E}|X|^n < \infty$. First, we inductively show that for $0 \leq k \leq n$,

$$\phi_X^{(k)}(t) = \mathbb{E}(iX)^k e^{itX}.$$

This is clear for $k=0$ and for $k<n$, inductively, we have

$$\begin{aligned}\phi_X^{(k+1)}(t) = \lim_{h\to 0} \frac{\phi_X^{(k)}(t+h) - \phi_X^{(k)}(t)}{h} &= \lim_{h\to 0} \mathbb{E}\left[(iX)^k e^{itX}\frac{e^{ihX}-1}{h}\right] \\ &= \mathbb{E}\left[(iX)^k e^{itX}\lim_{h\to 0}\frac{e^{ihX}-1}{h}\right].\end{aligned}$$

The last equality is justified by Lebesgue's dominated convergence theorem because

$$\left|(iX)^k e^{itX}\frac{e^{ihX}-1}{h}\right| \leq |X|^k|X| = |X|^{k+1}$$

and by the assumption $\mathbb{E}|X|^{k+1} < \infty$; we also used that for $t \in \mathbb{R}$, $|e^{it}-1| \leq |t|$ which can be justified as follows

$$|e^{it}-1| = \left|\frac{1}{i}\int_0^t e^{ix}\mathrm{d}x\right| \leq \int_0^t |e^{ix}|\mathrm{d}x = t$$

when $t \geq 0$ and similarly for $t < 0$. Finally, $\lim_{h\to 0}\frac{e^{ihX}-1}{h} = iX$ which finishes the inductive argument. Having the formula, uniform continuity follows as in (iii). □

9.4 Example. Let X be a standard Gaussian random variable. We have,

$$\phi_X(t) = \int_{\mathbb{R}} e^{itx}e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} = e^{-t^2/2}\int_{\mathbb{R}} e^{-(x-it)^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} = e^{-t^2/2},$$

where the last step would need proper justification (e.g., integrating along an appropriate contour and using $\int_{\mathbb{R}} e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}}$). Instead, we use Theorem 9.3 (iv),

$$\phi_X'(t) = i\mathbb{E}Xe^{itX} = -\mathbb{E}X\sin(tX) + i\mathbb{E}X\cos(tX).$$

Since X is symmetric and cos is even, $\mathbb{E}X\cos(tX) = 0$ and integrating by parts,

$$\begin{aligned}\phi_X'(t) = -\mathbb{E}X\sin(tX) = -\int x\sin(tx)e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} &= \int \sin(tx)(e^{-x^2/2})'\frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= -t\int\cos(tx)e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}}\end{aligned}$$

which is $-t\mathbb{E}\cos(tX) = -t\mathbb{E}e^{itX} = -t\phi_X(t)$ (by the symmetry of X, again, $\mathbb{E}\sin(tX) = 0$), so $\phi_X'(t) = -t\phi_X(t)$. That is, $\phi_X'(t) = -t\phi_X(t)$, equivalently, $(e^{t^2/2}\phi_X(t))' = 0$ which finally gives $e^{t^2/2}\phi_X(t) = \phi_X(0) = 1$.

If $Y \sim N(\mu,\sigma^2)$, then $Y = \mu + \sigma X$ and we thus get

$$\phi_Y(t) = \mathbb{E}e^{it(\mu+\sigma X)} = e^{it\mu}\mathbb{E}e^{i(t\sigma)X} = e^{it\mu - \sigma^2t^2/2}. \tag{9.1}$$

Note a simple but very powerful observation involving independence.

9.5 Theorem. *If X and Y are independent random variables, then*

$$\phi_{X+Y} = \phi_X \cdot \phi_Y.$$

Proof. Clearly, $\mathbb{E}e^{it(X+Y)} = \mathbb{E}e^{itX}e^{itY} = \mathbb{E}e^{itX}\mathbb{E}e^{itY}$. $\square$

9.2 Inversion formulae

One of the crucial properties of characteristic functions is that they determine the distribution uniquely. En route to proving that, we establish an inversion formula, quite standard in Fourier analysis. We first need a lemma.

9.6 Lemma. *For two independent random variables X and Y and every $t \in \mathbb{R}$, we have*

$$\mathbb{E}e^{-itY}\phi_X(Y) = \mathbb{E}\phi_Y(X-t).$$

Proof. Changing the order of taking expectation, we have

$$\mathbb{E}_Y e^{-itY}\phi_X(Y) = \mathbb{E}_Y e^{-itY}\mathbb{E}_X e^{iYX} = \mathbb{E}_{X,Y} e^{iY(X-t)} = \mathbb{E}_X\mathbb{E}_Y e^{iY(X-t)} = \mathbb{E}_X\phi_Y(X-t).$$

$\square$

9.7 Theorem (Inversion formula)**.** *For a random variable X, at every point x of continuity of its distribution function F_X, we have*

$$F_X(x) = \lim_{a\to\infty}\int_{-\infty}^{x}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s\right)\mathrm{d}t.$$

Proof. Let G be a standard Gaussian random variable, independent of X. For $a > 0$, consider $X_a = X + a^{-1}G$. Since X_a converges pointwise to X as $a \to \infty$, by Lebesgue's dominated convergence theorem $\mathbb{E}g(X_a) \to \mathbb{E}g(X)$ for every bounded continuous function g, thus $X_a \xrightarrow{d} 0$ as $a \to \infty$ (Theorem 8.5). Consequently, for every continuity point x of F_X, we have

$$F_X(x) = \lim_{a\to\infty} F_{X_a}(x).$$

Let us find the distribution function of X_a. We have,

$$\begin{aligned} F_{X_a}(x) = \mathbb{P}\left(X + a^{-1}G \le x\right) = \mathbb{E}_{X,G}\,\mathbf{1}_{\{X+a^{-1}G\le x\}} &= \mathbb{E}_X\mathbb{E}_G\,\mathbf{1}_{\{X+a^{-1}G\le x\}} \\ &= \mathbb{E}_X\mathbb{P}\left(X + a^{-1}G \le x\right). \end{aligned}$$

For any $y \in \mathbb{R}$, the density of $y + a^{-1}G$ at t is $\frac{a}{\sqrt{2\pi}}e^{-a^2(t-y)^2/2}$, thus

$$F_{X_a}(x) = \mathbb{E}_X\int_{-\infty}^{x}\frac{a}{\sqrt{2\pi}}e^{-a^2(t-X)^2/2}\mathrm{d}t = \int_{-\infty}^{x}\mathbb{E}_X\frac{a}{\sqrt{2\pi}}e^{-a^2(t-X)^2/2}\mathrm{d}t.$$

Note that $e^{-a^2s^2/2}$ is the characteristic function of aG at s (Example 9.4), so by Lemma 9.6,

$$\mathbb{E}_X \frac{a}{\sqrt{2\pi}} e^{-a^2(t-X)^2/2} = \frac{a}{\sqrt{2\pi}} \mathbb{E}_X \phi_{aG}(X-t) = \frac{a}{\sqrt{2\pi}} \mathbb{E} e^{-itaG} \phi_X(aG).$$

Writing this explicity using the density of aG yields

$$\begin{aligned}\frac{a}{\sqrt{2\pi}} \mathbb{E} e^{-itaG} \phi_X(aG) &= \frac{a}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi} a} \int_{-\infty}^{\infty} e^{-its} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s.\end{aligned}$$

Plugging this back,

$$F_{X_a}(x) = \int_{-\infty}^{x} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s \right) \mathrm{d}t,$$

which combined with $F_X(x) = \lim_{a\to\infty} F_{X_a}(x)$ remarked earlier finishes the proof. $\square$

Now we can prove that characteristic functions determine distribution.

9.8 Theorem. *Random variables X and Y have the same distribution (that is, $F_X = F_Y$) if and only if they have the same characteristic functions $\phi_X = \phi_Y$.*

Proof. By Theorem 9.7, $F_X(x) = F_Y(x)$ for every $x \in \mathbb{R} \setminus B$, where B is the union of the discontinuity points of F_X and the discontinuity points of F_Y. For $x \in B$, take $x_n > x$ such that $x_n \in \mathbb{R} \setminus B$ and $x_n \to x$ (it is possible since B is at most countable). Then $F_X(x_n) = F_Y(x_n)$ and by right-continuity, $F_X(x) = F_Y(x)$. $\square$

The inversion formula from Theorem 9.7 gives us several other interesting corollaries. Since the characteristic function determines distribution, it should be possible to reconstruct densities from characteristic functions.

9.9 Theorem. *If X is a random variable such that $\int_{\mathbb{R}} |\phi_X| < \infty$, then X has density f given by*

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-isx} \phi_X(s) \mathrm{d}s$$

which is bounded and uniformly continuous.

9.10 Remark. If X is a continuous random variable with density f, then clearly

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{its} f(s) \mathrm{d}s$$

The two formulae have the same form!

Proof. For two continuity points $x < y$ of F_X, we have from Theorem 9.7,

$$F_X(y) - F_X(x) = \lim_{a\to\infty} \int_{x}^{y} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist} \phi_X(s) e^{-\frac{s^2}{2a^2}} \mathrm{d}s \right) \mathrm{d}t.$$

Since $|e^{-ist}\phi_X(s)e^{-\frac{s^2}{2a^2}}| \leq |\phi_X(s)|$, that is the integrand is dominated by $|\phi_X|$ which is integrable on $[x,y] \times \mathbb{R}$, by Lebesgue's dominated convergence theorem,

$$F_X(y) - F_X(x) = \int_x^y \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ist}\phi_X(s)\mathrm{d}s \right) \mathrm{d}t$$

which gives that X has density given by the promised formula. The rest follows as for characteristic functions (recall the proof of Theorem 9.3 (iii)). □

9.11 Corollary. *If X is a continuous random variable with density f_X and characteristic function ϕ_X which is nonnegative, then $\int_{\mathbb{R}} \phi_X < \infty$ if and only if f is bounded.*

Proof. If $\int_{\mathbb{R}} \phi_X < \infty$, then by Theorem 9.9, f is bounded. Conversely, let as in the proof of Theorem 9.9, G be a standard Gaussian random variable independent of X. Then the density of $X + a^{-1}G$ at x equals

$$\int_{\mathbb{R}} f_X(x-y) f_{a^{-1}G}(y)\mathrm{d}y.$$

On the other hand, it equals $\frac{\mathrm{d}}{\mathrm{d}x} F_{X_a}(x)$ and from the last identity in the proof of Theorem 9.9, this becomes

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isx}\phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s.$$

For $x = 0$ we thus get

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s = \int_{\mathbb{R}} f_X(-y) f_{a^{-1}G}(y)\mathrm{d}y.$$

If f_X is bounded by, say M, we obtain that the right hand side is bounded by M, so

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(s)e^{-\frac{s^2}{2a^2}}\mathrm{d}s \leq M.$$

As $a \to \infty$, by Lebesgue's monotone convergence theorem, the left hand side converges to $\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X$, which proves that $\int_{\mathbb{R}} \phi_X \leq 2\pi M$. □

9.12 Example. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables uniform on $[-1,1]$. Then $X_1 + \ldots + X_n$ for $n \geq 2$ has density

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos(tx) \left(\frac{\sin t}{t} \right)^n \mathrm{d}t.$$

Indeed, note that $\phi_{X_i}(t) = \frac{\sin t}{t}$, so $\phi_{X_1+\ldots+X_n}(t) = \left(\frac{\sin t}{t}\right)^n$ which is integrable for $n \geq 2$ and the formula follows from Theorem 9.9.

We finish with two Fourier analytic identities.

9.13 Theorem (Parseval's identities)**.** *If X and Y are continuous random variables with densities f_X and f_Y, then* *(i)* $\int_{\mathbb{R}} |\phi_X|^2 < \infty$ *if and only if* $\int_{\mathbb{R}} f_X^2 < \infty$ *and then*

$$\int_{\mathbb{R}} f_X^2 = \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_X|^2,$$

(ii) if $\int_{\mathbb{R}} f_X^2 < \infty$ *and* $\int_{\mathbb{R}} f_Y^2 < \infty$*, then*

$$\int_{\mathbb{R}} f_X f_Y = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_X \overline{\phi_Y}.$$

Proof. (i) Let X' be an independent copy of X. Consider $\tilde{X} = X - X'$. We have,

$$\phi_{\tilde{X}}(t) = \phi_X(t)\phi_{-X'}(t) = \phi_X(t)\phi_{-X}(t) = \phi_X(t)\overline{\phi_X(t)} = |\phi_X(t)|^2.$$

On the other hand, $\tilde{X}$ is continuous with density given by convolution,

$$f_{\tilde{X}}(y) = (f_X \star f_{-X})(y) = \int_{\mathbb{R}} f_X(x) f_{-X}(y - x)\mathrm{d}x.$$

It can be seen from here that if $\int f_X^2 < \infty$, then by the Cauchy-Schwarz inequality, $f_{\tilde{X}}$ is bounded. Then by Corollary 9.11, $\phi_{\tilde{X}} = |\phi_X|^2$ is integrable. Conversely, if $|\phi_X|^2$ is integrable, then from Theorem 9.9 applied to $\tilde{X}$, we get

$$f_{\tilde{X}}(0) = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_{\tilde{X}} = \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_X|^2.$$

Since

$$f_{\tilde{X}}(0) = (f_X \star f_{-X})(0) = \int_{\mathbb{R}} f_X(x) f_{-X}(0 - x)\mathrm{d}x = \int_{\mathbb{R}} f_X(x) f_X(x)\mathrm{d}x = \int_{\mathbb{R}} f_X^2,$$

we get that $\int f_X^2 = \frac{1}{2\pi} \int |\phi_X|^2$. In particular, f_X^2 is integrable.

(ii) Apply (i) to the density $\frac{f_X + f_Y}{2}$. □

9.3 Relations to convergence in distribution

The second crucial property of characteristic functions is that their pointwise convergence captures convergence in distribution. To establish that, we will need to use compactness-type arguments. We start with a lemma that will help us get tightness.

9.14 Lemma. *For a random variable* X *and* $\delta > 0$,

$$\mathbb{P}\left(|X| > \frac{2}{\delta}\right) \le \frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \phi_X(t)]\mathrm{d}t.$$

Proof. Note that

$$\int_{-\delta}^{\delta} [1 - \phi_X(t)]\mathrm{d}t = \int_{-\delta}^{\delta} [1 - \mathbb{E}e^{itX}]\mathrm{d}t = 2\delta - \mathbb{E}\int_{-\delta}^{\delta} e^{itX}\mathrm{d}t = 2\delta - \mathbb{E}\frac{e^{i\delta X} - e^{-i\delta X}}{iX}$$
$$= 2\delta - 2\mathbb{E}\frac{\sin(\delta X)}{X}$$

(this incidentally shows that the a priori complex number $\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t$ is real). Thus

$$\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t = 2\mathbb{E}\left[1-\frac{\sin(\delta X)}{\delta X}\right].$$

Using $|\sin x| \le |x|$, we have $1-\frac{\sin x}{x} \ge 0$, so

$$\begin{aligned}\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t &\ge 2\mathbb{E}\left[\left(1-\frac{\sin(\delta X)}{\delta X}\right)\mathbf{1}_{\{|\delta X|>2\}}\right]\\ &= 2\mathbb{E}\left[\left(1-\frac{\sin(\delta|X|)}{\delta|X|}\right)\mathbf{1}_{\{|\delta X|>2\}}\right],\end{aligned}$$

where is the last equality we used that $\frac{\sin x}{x}$ is even. Crudely, $-\sin(\delta|X|) \ge -1$, hence

$$\begin{aligned}\frac{1}{\delta}\int_{-\delta}^{\delta}[1-\phi_X(t)]\mathrm{d}t \ge 2\mathbb{E}\left[\left(1-\frac{1}{\delta|X|}\right)\mathbf{1}_{\{|\delta X|>2\}}\right] &\ge 2\mathbb{E}\left[\frac{1}{2}\mathbf{1}_{\{|\delta X|>2\}}\right]\\ &= \mathbb{P}\left(|\delta X|>2\right).\end{aligned}$$

□

The main result about convergence in distribution is the following so-called (Lévy's) continuity theorem.

9.15 Theorem (Lévy's continuity theorem)**.** *Let (X_n) be a sequence of random variables such that for every $t \in \mathbb{R}$, $\phi_{X_n}(t) \xrightarrow[n\to\infty]{} \phi(t)$ for some function $\phi : \mathbb{R} \to \mathbb{C}$ which is continuous at $t = 0$. Then there is a random variable X such that $\phi = \phi_X$ and $X_n \xrightarrow{d} X$.*

9.16 Remark. The converse to Lévy's Theorem also holds: if $X_n \xrightarrow{d} X$, then $\phi_{X_n}(t) \to \phi_X(t)$ for every $t \in \mathbb{R}$. Indeed, since sin is continuous and bounded, $\mathbb{E}\sin(tX_n) \to \mathbb{E}\sin(tX)$ and the same for the cos function, so $\phi_{X_n}(t) = \mathbb{E}\cos(tX_n) + i\mathbb{E}\sin(tX_n) \to \phi_X(t)$.

Proof of Theorem 9.15. Since $|\phi_{X_n}(t)| \le 1$ for every t, by taking the limit, we have $|\phi(t)| \le 1$ for every t.

Step 1 (tightness). Since ϕ is continuous at 0 and $\phi(0) = \lim_n \phi_{X_n}(0) = 1$, for every $\varepsilon > 0$, there is $\delta > 0$ such that $|1-\phi(t)| < \varepsilon$ for $|t| < \delta$, so

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi(t)|\mathrm{d}t \le 2\varepsilon.$$

By Lebesgue's dominated convergence theorem,

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi_{X_n}(t)|\mathrm{d}t \xrightarrow[n\to\infty]{} \frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi(t)|\mathrm{d}t,$$

so for large n,

$$\frac{1}{\delta}\int_{-\delta}^{\delta}|1-\phi_{X_n}(t)|\mathrm{d}t < 3\varepsilon.$$

By Lemma 9.14, we obtain

$$\mathbb{P}\left(|X_n| > \frac{2}{\delta}\right) < \varepsilon.$$

This shows that the sequence (X_n) is tight. By Theorem 8.17, there is a subsequence (X_{n_k}) which converges in distribution to a random variable, say X. This is our candidate for the limit of (X_n).

Step 2 ($\phi = \phi_X$). Since $X_{n_k} \xrightarrow{d} X$, we get $\phi_{X_{n_k}} \to \phi_X$ at every point (Remark 9.16, but also $\phi_{X_{n_k}} \to \phi$ at every point, so $\phi = \phi_X$, which proves that ϕ is a characteristic function.

Step 3 ($X_n \xrightarrow{d} X$). If this is not the case, then, by the definition, there is a bounded continuous function g such that $\mathbb{E}g(X_n) \nrightarrow \mathbb{E}g(X)$. Therefore, there is $\varepsilon > 0$ and a sequence m_k such that $|\mathbb{E}g(X_{m_k}) - \mathbb{E}g(X)| > \varepsilon$. Since (X_n) is tight, using Theorem 8.17 again, there is a convergent subsequence $X_{m_{k_l}}$ to some random variable, say X'. As in Step 2, $\phi_{X'} = \phi = \phi_X$, so X' has the same distribution as X (Theorem 9.8) and $|\mathbb{E}g(X_{m_{k_l}}) - \mathbb{E}g(X')| = |\mathbb{E}g(X_{m_{k_l}}) - \mathbb{E}g(X)| > \varepsilon$ contradicts that $X_{m_{k_l}} \xrightarrow{d} X'$. □

9.17 Example. In Levy's theorem the continuity assumption is necessary. Let G be a standard Gaussian random variable and consider the sequence $X_n = nG$. We have $\phi_{X_n}(t) = \phi_{nG}(t) = \phi_G(nt) = e^{-n^2t^2/2}$, so

$$\phi_{X_n}(t) \to \begin{cases} 0, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

The limiting function is discontinuous at 0. The sequence X_n does not converge in distribution because $F_{X_n}(t) = \mathbb{P}(G \le t/n) \to \mathbb{P}(G \le 0) = 1/2$, but the limit is not a distribution function (an alternative argument: by Remark 9.16, if $X_n \xrightarrow{d} X$, then ϕ_{X_n} would converge to a characteristic function which is continuous).

9.4 Exercises

1. In the proof of Theorem 9.5 we implicitly used: if X, Y are *complex*-valued integrable random variables which are independent, then $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$. Fill out this gap.

Table 1: Characteristic functions of common discrete distributions (see Section 5.1).

Distribution μ	Characteristic function $\phi_\mu(t)$
Dirac delta δ_a	e^{ita}
Ber(p)	$1 - p + pe^{it}$
Bin(n, p)	$(1 - p + pe^{it})^n$
Poiss(λ)	$\exp\{\lambda(e^{it} - 1)\}$
Geom(p)	$\frac{pe^{it}}{1-(1-p)e^{it}}$

Table 2: Characteristic functions of common continuous distributions (see Section 5.1).

Distribution μ	Density function $f_\mu(x)$	Characteristic function $\phi_\mu(t)$
Unif$([0, a])$	$\frac{1}{a}\mathbf{1}_{[0,a]}(x)$	$\frac{e^{iat}-1}{iat}$
Exp(λ)	$\lambda e^{-\lambda x}\mathbf{1}_{(0,\infty)}(x)$	$\frac{\lambda}{\lambda - e^{it}}$
Sym-Exp	$\frac{1}{2}e^{-\lvert x\rvert}$	$\frac{1}{1+t^2}$
Cauchy	$\frac{1}{\pi(1+x^2)}$	$e^{-\lvert t\rvert}$
Gamma(β, λ)	$\frac{\lambda^\beta}{\Gamma(\beta)}x^{\beta-1}e^{-\lambda x}\mathbf{1}_{(0,\infty)}(x)$	$\left(1 - \frac{it}{\lambda}\right)^{-\beta}$
$N(a, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}$	$e^{iat - \sigma^2 t^2/2}$
Hyperbolic cosine	$\frac{1}{\pi\cosh x}$	$\frac{1}{\cosh(\pi t/2)}$

2. Justify Table 1.

3. Justify Table 2.

4. Decide whether the following functions are the characteristic functions of some distributions. If yes, describe the corresponding distribution.

   a) $\cos t$,

   b) $\cos^2 t$,

   c) $\frac{1}{4}(1 + e^{it})^2$,

   d) $\frac{1+\cos t}{2}$,

   e) $(2 - e^{it})^{-1}$.

5. Suppose $\phi_1, \ldots, \phi_n$ are characteristic functions of some distributions. Let nonnegative numbers $p_1, \ldots, p_n$ be such that $\sum_{j=1}^n p_j = 1$. Show that $\sum_{j=1}^n p_j \phi_j$ is also a characteristic function.

6. Let $X_1, X_2, \ldots$ be i.i.d. with characteristic function ϕ. Let N be a Poisson random variable with parameter $\lambda > 0$, independent of the X_j. Find the characteristic function of $Y = \sum_{j=1}^N X_j$ (we adopt the convention that $\sum_{j=1}^0 X_j = 0$).

7. Suppose ϕ is the characteristic function of some random variable X. Decide whether the following functions are always characteristic functions

   a) ϕ^2,

   b) $\mathrm{Re}\phi$,

   c) $|\phi|^2$,

   d) $|\phi|$.

8. Show that if $\phi_X''(0)$ exists, then $\mathbb{E}X^2 < \infty$.

9. Let X be a random variable with density $f(x) = \frac{C}{(1+x^2)\log(e+x^2)}$, $x \in \mathbb{R}$. Show that $\phi_X'(0)$ exits, but $\mathbb{E}|X| = +\infty$.

10. Let X be a random variable such that $\phi_X(t) = 1 - ct^2 + o(t^2)$ as $t \to 0$ for some constant $c \in \mathbb{R}$. Then $\mathbb{E}X = 0$, $\mathbb{E}X^2 = 2c$. In particular, if $\phi_X(t) = 1 + o(t^2)$, then $X = 0$ a.s. As a corollary, $\phi(t) = e^{-|t|^\alpha}$ is *not* a characteristic function for any $\alpha > 2$.

11. For a sequence (X_n) of random variables, $X_n \xrightarrow{d} 0$ if and only if there is $\delta > 0$ such that $\phi_{X_n}(t) \to 1$ for every $t \in (-\delta, \delta)$.

12. For an integer-valued random variable X and an integer k, we have
$$\mathbb{P}(X = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi_X(t) \mathrm{d}t.$$
(This is Lemma 11.8.)

13. If X is a continuous random variable, then $\phi_X(t) \to 0$ as $t \to \infty$.

14. These help show that certain important functions are characteristic functions:

    a) Show that $\phi(t) = 2\frac{1-\cos t}{t^2}$ is the characteristic function of the triangular distribution with density $(1 - |x|)\,\mathbf{1}_{[-1,1]}(x)$ (the distribution of the sum of two i.i.d. Unif$[-\frac{1}{2}, \frac{1}{2}]$ random variables).

    b) Using the inverse formula (Theorem 9.9), show that $(1 - \frac{|t|}{a})\,\mathbf{1}_{[-a,a]}(t)$ is a characteristic function.

c) Show that if a function $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfies: $\phi(0) = 1$, ϕ is even, ϕ is piece-wise linear and ϕ is nonincreasing and convex on $[0, +\infty$, then ϕ is a characteristic function.

15. Using Exercises 9.5 and 9.14, show the so-called *Pólya's criterion:* every function ϕ with $\phi(0) = 1$ which is even, nonincreasing and convex on $[0, +\infty)$ is a characteristic function. In particular, $e^{-|t|^\alpha}$ is a characteristic function for $\alpha \in (0, 1]$.

16. Let $0 < \alpha < 2$ and $\psi(t) = 1 - (1 - \cos t)^{\alpha/2}$. Using the binomial series argue that
$$\psi(t) = \sum_{n=1}^{\infty} p_n (\cos t)^n$$
for nonnegative $p_1, p_2, \ldots$ with $\sum_{n=1}^{\infty} p_n = 1$. Show that
$$e^{-|t|^\alpha} = \lim_{n\to\infty} \left[\psi(\sqrt{2}tn^{-1/\alpha})\right]^n$$
and conclude that $e^{-|t|^\alpha}$ is a characteristic function.

17. Let X_n be a Poisson random variable with parameter $\lambda_n > 0$. If $\lambda_n \to \lambda$ for some $\lambda > 0$, then $X_n \xrightarrow{d} X$ for a Poisson random variable X with parameter λ.

18. Suppose a sequence of random variables (X_n) converges in law to a random variable X. Suppose sequences of reals (a_n) and (b_n) converge to a and b, respectively. Show that $a_n X_n + b_n \xrightarrow{d} aX + b$.

19. Let (X_n) be a sequence of random variables with $\mathbb{P}\left(X_n = \frac{k}{n}\right) = \frac{1}{n^2}$, $k = 1, \ldots, n^2$. Does it converge in distribution?

20. Prove that $\phi(t) = \frac{2}{1+e^{t^2}}$ is *not* a characteristic function.

21. Let $X_1, \ldots, X_n$ be i.i.d. Cauchy random variables, that is with density $\frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$. Show that for every reals $a_1, \ldots, a_n$, the weighted sum $\sum_{j=1}^n a_j X_j$ has the same distribution as $(\sum_{j=1}^n |a_j|)X_1$.

22. For a random variable X and $a \in \mathbb{R}$, we have
$$\mathbb{P}(X = a) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-iat} \phi_X(t) \mathrm{d}t.$$

23. Let X and Y be i.i.d. random variables. Then
$$\mathbb{P}(X = Y) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |\phi_X(t)|^2 \mathrm{d}t.$$
On the other hand,
$$\mathbb{P}(X = Y) = \sum_{x\in\mathbb{R}} \mathbb{P}(X = x)^2$$
(of course the set $\{x,\ \mathbb{P}(X = x) > 0\}$ is at most countable).

24. Show that a random variable X has no atoms if and only if

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}|\phi_X(t)|^2\mathrm{d}t = 0.$$

In particular, if $\phi_X(t) \to 0$ as $t \to \infty$, then X has no atoms. The converse is not true (devil's staircase).

25. Let $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d. symmetric random signs. For every $a = (a_1, a_2, \dots) \in \ell_2$, i.e. $\sum a_n^2 < \infty$, the series $\sum_{n=1}^{\infty} a_n\varepsilon_n$ converges in probability, hence in distribution (hint: Cauchy's condition from Exercise 5.28). Show that the distribution of $\sum_{n=1}^{\infty}\varepsilon_n 2^{-n}$ is uniform on $[-1, 1]$.

26. Show that for a random variable X the following are equivalent

    (a) X is symmetric, that is X and $-X$ have the same distribution

    (b) X and εX have the same distribution, where ε is an independent random sign

    (c) X and $\varepsilon|X|$ have the same distribution, where ε is an independent random sign

    (d) the characteristic function of X is real valued.

27. For an integrable random variable X,

$$\mathbb{E}|X| = \frac{2}{\pi}\int_0^{\infty}\frac{1-\mathrm{Re}\phi_X(t)}{t^2}\mathrm{d}t.$$

28. Prove *Shepp's inequality*: if X, Y are i.i.d. integrable random variables, then

$$\mathbb{E}|X - Y| \le \mathbb{E}|X + Y|.$$

29. Find an example of two *different* distributions whose characteristic functions agree on $[-1, 1]$.

    Hint: Consider $\phi(t) = (1 - |t|)_+$ and $\psi(t)$ defined to be equal to $\phi(t)$ on $[-1, 1]$ and 2-periodic.

30. Find an example of 3 random variables X, Y, Z which are independent such that Y and Z do *not* have the same distribution, but $X + Y$ and $X + Z$ do.

31. Show that if random variables X, Y are independent and $X + Y$ has the same distribution as X, then $Y = 0$ a.s.

32. *Cramér's decomposition theorem.* If ξ is a Gaussian random variable and $\xi = X + Y$ for some independent random variables X and Y, then X and Y are also Gaussian (point masses are assumed to be Gaussian).

    Here is a possible, very analytic approach. A function $f\colon \mathbb{C} \to \mathbb{C}$ is entire if it is holomorphic on $\mathbb{C}$. The order of the entire function f is the infimum of $\rho > 0$

such that $|f(z)| = O(\exp(|z|^\rho))$ as $z \to \infty$. It is a consequence of Hadamard's factorisation theorem that an entire function f of order ρ without any zeros is of the form $f(z) = e^{g(z)}$, where g is a polynomial of degree at most ρ. For instance, $f(z) = e^{z^2}$ is entire of order 2.

(a) Using Hadamard's factorisation theorem show: if for some random variable X there is $a > 0$ such that $\mathbb{E}e^{aX^2} < \infty$, then ϕ_X extends to $\mathbb{C}$, is entire of order at most 2 and if additionally $\phi_X(z) \neq 0$ for all $z \in \mathbb{C}$, then X is Gaussian.

(b) Show that if $X + Y$ is Gaussian for some independent random variables X, Y, then $\mathbb{E}e^{\delta(X+Y)^2} < \infty$ for some $\delta > 0$.

(c) Deduce from independence that $\mathbb{E}e^{\delta(X+c)^2} < \infty$ for some $c \in \mathbb{R}$ and from (a) that X is Gaussian.

10 Central limit theorem

Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}|X_1|^2 < \infty$. By the strong law of large numbers,

$$Y_n = \frac{X_1 + \ldots + X_n}{n} - \mathbb{E}X_1$$

converges to 0 a.s. By our assumption, we can compute

$$\mathrm{Var}(Y_n) = \frac{\mathrm{Var}(X_1 + \ldots + X_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\mathrm{Var}(X_1)}{n},$$

so Y_n concentrates around its expectation, which is 0 and in a sense it is not surprising that Y_n goes to 0. What happens if we zoom in, that is rescale appropriately so that the variance of Y_n is fixed, that is when fluctuations of Y_n have a fixed size, as opposed to decaying like $1/n$ as earlier? Consider

$$Z_n = \frac{Y_n}{\sqrt{\mathrm{Var}(Y_n)}} = \sqrt{n}\frac{1}{\sqrt{\mathrm{Var}(X_1)}}\left(\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}X_1\right)$$

which has variance 1 for all n. What "limit distribution" does Z_n have as $n \to \infty$ (if any)? This is addressed by the central limit theorem which says that the weak limit exists and is Gaussian! (If it exists and is universal, that is the same for all i.i.d. sequences, then it has to be Gaussian because when the X_i are standard Gaussian, Z_n is also standard Gaussian.) To establish weak convergence, we shall use characteristic functions.

10.1 Auxiliary elementary lemmas

To handle the convergence of characteristic functions, we shall need several elementary estimates for complex numbers.

10.1 Lemma. *If $z_1, \dots, z_n$ and $w_1, \dots, w_n$ are complex numbers all with modulus at most θ, then*

$$\left|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j\right| \le \theta^{n-1}\sum_{j=1}^n |z_j - w_j|.$$

Proof. We proceed by induction on n. For $n = 1$, we have equality. For $n > 1$, we have

$$\begin{aligned}
\left|\prod_{j=1}^n z_j - \prod_{j=1}^n w_j\right| &= \left|z_1\prod_{j=2}^n z_j - w_1\prod_{j=2}^n w_j\right| \\
&\le \left|z_1\prod_{j=2}^n z_j - z_1\prod_{j=2}^n w_j\right| + \left|z_1\prod_{j=2}^n w_j - w_1\prod_{j=2}^n w_j\right| \\
&= |z_1|\left|\prod_{j=2}^n z_j - \prod_{j=2}^n w_j\right| + \left|\prod_{j=2}^n w_j\right||z_1 - w_1| \\
&\le \theta\left|\prod_{j=2}^n z_j - \prod_{j=2}^n w_j\right| + \theta^{n-1}|z_1 - w_1|
\end{aligned}$$

and the inductive assumption allows to finish the proof. $\square$

10.2 Lemma. *For a complex number z with $|z| \le 1$, we have*

$$|e^z - (1+z)| \le |z|^2.$$

Proof. Using the power series expansion of e^z, we get

$$\begin{aligned}|e^z - (1+z)| = \left|\frac{z^2}{2!} + \frac{z^3}{3!} + \dots\right| \le |z|^2\left(\frac{1}{2!} + \frac{|z|}{3!} + \dots\right) &\le |z|^2\left(\frac{1}{2!} + \frac{1}{3!} + \dots\right) \\ &= |z|^2(e-2).\end{aligned}$$

$\square$

10.3 Lemma. *If (z_n) is a sequence of complex numbers such that $z_n \to z$ for some $z \in \mathbb{C}$, then*

$$\left(1 + \frac{z_n}{n}\right)^n \to e^z.$$

Proof. Fix $c > |z|$. Then eventually, $|z_n| < c$ and consequently, $\left|1 + \frac{z_n}{n}\right| \le 1 + \frac{c}{n} \le e^{c/n}$ and $|e^{z_n/n}| = e^{\mathrm{Re}(z_n)/n} \le e^{c/n}$, so applying Lemma 10.1 with $\theta = e^{c/n}$, for large n,

$$\left|\left(1+\frac{z_n}{n}\right)^n - e^{z_n}\right| = \left|\prod_{j=1}^n \left(1+\frac{z_n}{n}\right) - \prod_{j=1}^n e^{z_n/n}\right| \le \left(e^{c/n}\right)^{n-1} n \left|1 + \frac{z_n}{n} - e^{z_n/n}\right|.$$

Clearly eventually, $|z_n/n| \le 1$, so by Lemma 10.2,

$$\left|\left(1+\frac{z_n}{n}\right)^n - e^{z_n}\right| \le \left(e^{c/n}\right)^{n-1} n \left|\frac{z_n}{n}\right|^2 \le e^c\frac{c^2}{n}.$$

It remains to use continuity, that is that $e^{z_n} \to e^z$. $\square$

10.4 Lemma. *For every real number t and $n = 0, 1, 2, \dots$, we have*

$$\left|e^{it} - \sum_{k=0}^n \frac{(it)^k}{k!}\right| \le \frac{|t|^{n+1}}{(n+1)!}.$$

Proof. We proceed by induction on n. For $n = 0$, we have

$$|e^{it} - 1| = \left|\int_0^t e^{is}\mathrm{d}s\right| \le |t|.$$

Suppose the assertion holds for $n \ge 0$. Then, we have

$$\left|e^{it} - \sum_{k=0}^{n+1}\frac{(it)^k}{k!}\right| = \left|\int_0^t \left(e^{is} - \sum_{k=0}^n \frac{(is)^k}{k!}\right)\mathrm{d}s\right| \le \int_0^t \frac{|s|^{n+1}}{(n+1)!}\mathrm{d}s = \frac{|t|^{n+2}}{(n+2)!}.$$

$\square$

10.2 Vanilla Central Limit Theorem

10.5 Theorem. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^2 < \infty$. Then the sequence (Z_n) of normalised sums*

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\operatorname{Var}(X_1)}}$$

converges in distribution to a standard Gaussian random variable.

Proof. Let $\bar{X}_i = \frac{X_i - \mathbb{E}X_i}{\operatorname{Var}(X_1)}$. Then $\mathbb{E}\bar{X}_i = 0$, $\mathbb{E}|\bar{X}_i|^2 = 1$,

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\operatorname{Var}(X_1)}} = \frac{\bar{X}_1 + \ldots + \bar{X}_n}{\sqrt{n}}$$

and by independence

$$\phi_{Z_n}(t) = \phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right) \ldots \phi_{\bar{X}_n}\left(\frac{t}{\sqrt{n}}\right) = \left[\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

We investigate pointwise convergence of ϕ_{Z_n}. By Theorem 9.3 (v), $\phi_{\bar{X}_1}$ is twice continuously differentiable and we can compute that $\phi'_{\bar{X}_1}(0) = i\mathbb{E}\bar{X}_1 = 0$ and $\phi''_{\bar{X}_1}(0) = i^2\mathbb{E}\bar{X}_1^2 = -1$. Thus by Taylor's formula with Lagrange's remainder

$$\begin{aligned}
\phi_{\bar{X}_1}(t) &= \phi_{\bar{X}_1}(0) + t\phi'_{\bar{X}_1}(0) + \frac{t^2}{2}\phi''_{\bar{X}_1}(\xi_t) \\
&= 1 + t\phi'_{\bar{X}_1}(0) + \frac{t^2}{2}\phi''_{\bar{X}_1}(0) + t^2R(t) \\
&= 1 - \frac{t^2}{2} + t^2R(t),
\end{aligned}$$

for some ξ_t between 0 and t and $R(t) = \frac{1}{2}(\phi''_{\bar{X}_1}(\xi_t) - \phi''_{\bar{X}_1}(0))$. By the continuity of $\phi''_{\bar{X}_1}$ (at 0), $R(t) \xrightarrow[t\to 0]{} 0$. Note that $R(t)$ may be complex. By Lemma 10.3, for every $t \in \mathbb{R}$,

$$\phi_{Z_n}(t) = \left[\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 - \frac{t^2}{2n} + \frac{t^2}{n}R(t)\right]^n \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

By Theorem 9.15, Z_n converges in distribution to a random variable whose characteristic function is $e^{-t^2/2}$, that is a standard Gaussian random variable. □

We have two remarks. The first one shows that other notions of convergence are too strong to capture the limit behaviour of sequences of sums of i.i.d. random variables, normalised to have a fixed variance. The second one shows that the finite variance is really a necessary assumption to make for the weak limit to exist. We defer their proofs to exercises.

10.6 Remark. Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with mean 0 and finite variance. Then by the vanilla central limit theorem and Kolmogorov's $0-1$ law,

$$\limsup_{n\to\infty} \frac{X_1 + \cdots + X_n}{\sqrt{n}} = +\infty \quad \text{a.s.}$$

Moreover, the sequence $(\frac{X_1+\cdots+X_n}{\sqrt{n}})$, or any of its subsequences, does *not* converge in probability.

10.7 Remark. Suppose $X_1, X_2, \dots$ are i.i.d. random variables such that the sequence $(\frac{X_1+\cdots+X_n}{\sqrt{n}})$ converges in distribution. Then $\mathbb{E}X_1^2 < \infty$.

In Appendices F and G, we present two other completely different proofs of the vanilla central limit theorem.

10.3 Lindeberg's Central Limit Theorem

The assumption of identical distribution of the summands in the vanilla central limit theorem can be weakened. The so-called Lindeberg condition is an *almost* optimal condition under which the central limit theorem holds.

10.8 Theorem. *Let $\{X_{n,}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of random variables with $\mathbb{E}X_{n,k}^2 < \infty$ such that for every $n \geq 1$, the variables $X_{n,1}, \dots, X_{n,n}$ are independent. Let*

$$\bar{X}_{n,k} = \frac{X_{n,k} - \mathbb{E}X_{n,k}}{\sqrt{\sum_{k=1}^n \mathrm{Var}(X_{n,k})}}$$

and for $\varepsilon > 0$, set

$$L_n(\varepsilon) = \sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}^2 \, \boldsymbol{1}_{\{|X_{n,k}|>\varepsilon\}} \, .$$

If the following Lindeberg condition holds:

$$\text{for every } \varepsilon > 0, \quad L_n(\varepsilon) \xrightarrow[n\to\infty]{} 0, \tag{10.1}$$

then

$$\sum_{k=1}^n \bar{X}_{n,k} \xrightarrow[n\to\infty]{d} Z,$$

where Z is a standard Gaussian random variable.

10.9 Remark. Condition (10.1) implies that

$$\max_{1\leq k\leq n} \mathrm{Var}(\bar{X}_{n,k}) \xrightarrow[n\to\infty]{} 0 \tag{10.2}$$

(individual contributions of the summans in Z_n are small). Indeed, for every $\varepsilon > 0$, we have

$$\mathrm{Var}(\bar{X}_{n,k}) \leq \mathbb{E}\bar{X}_{n,k}^2 \leq \mathbb{E}\bar{X}_{n,k}^2 \, \mathbf{1}_{\{|\bar{X}_{n,k}|>\varepsilon\}} + \varepsilon \leq L_n(\varepsilon) + \varepsilon.$$

Proof of Theorem 10.8. Denote $Z_n = \sum_{k=1}^n \bar{X}_{n,k}$ and $\sigma_{n,k} = \sqrt{\mathrm{Var}(\bar{X}_{n,k})}$. By the definition of $\bar{X}_{n,k}$, for every $n \geq 1$, we have

$$\sum_{k=1}^n \sigma_{n,k}^2 = 1. \tag{10.3}$$

To show $Z_n \xrightarrow[n\to\infty]{d} Z$, in view of Lévy's continuity theorem (Theorem 9.15), it is enough to show that for every $t \in \mathbb{R}$, we have

$$\phi_{Z_n}(t) \xrightarrow[n\to\infty]{} e^{-t^2/2}.$$

By independence, (10.3) and Lemma 10.1,

$$\left|\phi_{Z_n}(t) - e^{-t^2/2}\right| = \left|\prod_{k=1}^n \phi_{\bar{X}_{n,k}}(t) - \prod_{k=1}^n e^{-\sigma_{n,k}^2 t^2/2}\right| \le \sum_{k=1}^n \left|\phi_{\bar{X}_{n,k}}(t) - e^{-\sigma_{n,k}^2 t^2/2}\right|.$$

Denoting

$$s_n(t) = \sum_{k=1}^n \left|\mathbb{E}e^{it\bar{X}_{n,k}} - 1 + \frac{1}{2}\sigma_{n,k}^2 t^2\right| = \sum_{k=1}^n \left|\mathbb{E}\left[e^{it\bar{X}_{n,k}} - 1 - it\bar{X}_{n,k} + \frac{1}{2}t^2\bar{X}_{n,k}^2\right]\right|$$

and

$$r_n(t) = \sum_{k=1}^n \left|1 - \frac{1}{2}\sigma_{k,n}^2 t^2 - e^{-\sigma_{n,k}^2 t^2/2}\right|,$$

by the triangle inequality we thus have

$$\left|\phi_{Z_n}(t) - e^{-t^2/2}\right| \le s_n(t) + r_n(t).$$

Fix $\varepsilon > 0$. Splitting the expectation in $s_n(t)$ into two: on $\{|\bar{X}_{n,k}| \le \varepsilon\}$ and $\{|\bar{X}_{n,k}| > \varepsilon\}$, we get

$$s_n(t) \le s_n^{(1)}(t) + s_n^{(2)}(t)$$

with

$$s_n^{(1)}(t) = \sum_{k=1}^n \left|\mathbb{E}\left[e^{it\bar{X}_{n,k}} - 1 - it\bar{X}_{n,k} + \frac{1}{2}t^2\bar{X}_{n,k}^2\right]\mathbf{1}_{\{|\bar{X}_{n,k}|\le\varepsilon\}}\right|,$$

$$s_n^{(2)}(t) = \sum_{k=1}^n \left|\mathbb{E}\left[e^{it\bar{X}_{n,k}} - 1 - it\bar{X}_{n,k} + \frac{1}{2}t^2\bar{X}_{n,k}^2\right]\mathbf{1}_{\{|\bar{X}_{n,k}|>\varepsilon\}}\right|.$$

Thanks to Lemma 10.4 (the case $n = 3$),

$$s_n^{(1)}(t) \le \sum_{k=1}^n \mathbb{E}\frac{|t|^3|\bar{X}_{n,k}|^3}{6}\mathbf{1}_{\{|\bar{X}_{n,k}|\le\varepsilon\}} \le \frac{|t|^3\varepsilon}{6}\sum_{k=1}^n \mathbb{E}|\bar{X}_{n,k}|^2 = \frac{|t|^3\varepsilon}{6}.$$

Thanks to Lemma 10.4 (the case $n = 2$) and the triangle inequality,

$$s_n^{(2)}(t) \le \sum_{k=1}^n \mathbb{E}\left(\frac{t^2\bar{X}_{n,k}^2}{2} + \frac{t^2\bar{X}_{n,k}^2}{2}\right)\mathbf{1}_{\{|\bar{X}_{n,k}|>\varepsilon\}} = t^2 L_n(\varepsilon).$$

To bound $r_n(t)$, note that thanks to (10.2), for large enough n and every $k \le n$, we have $\sigma_{n,k}^2 t^2 \le 1$. Thus, thanks to Lemma 10.2,

$$r_n(t) \le \sum_{k=1}^n \left(\frac{1}{4}\sigma_{n,k}^2 t^2\right)^2 \le \frac{t^4}{4}\max_{k\le n}\sigma_{n,k}^2 \sum_{k=1}^n \sigma_{n,k}^2 = \frac{t^4}{4}\max_{k\le n}\sigma_{n,k}^2.$$

Putting the bounds on $s_n(t)$ and $r_n(t)$ together yields

$$\left|\phi_{Z_n}(t) - e^{-t^2/2}\right| \le \frac{|t|^3\varepsilon}{6} + t^2 L_n(\varepsilon) + \frac{t^4}{4}\max_{k\le n}\sigma_{n,k}^2$$

which by (10.1) and (10.2) gives $\phi_{Z_n}(t) \to e^{-t^2/2}$, as desired. $\square$

10.10 Remark. Of course Lindeberg's central limit theorem implies the vanilla one. Indeed, if $X_1, X_2, \ldots$ are i.i.d. square integrable random variables, say with mean 0 and variance 1, then setting $X_{n,k} = X_k/\sqrt{n}$, we check that the Lindeberg condition (10.1) holds,

$$L_n(\varepsilon) = n\mathbb{E}\frac{X_1^2}{n}\mathbf{1}_{\{|X_1|>\varepsilon\sqrt{n}\}} = \mathbb{E}X_1^2\,\mathbf{1}_{\{|X_1|>\varepsilon\sqrt{n}\}}$$

which goes to 0 as n goes to ∞ by Lebesgue's dominated convergence theorem.

10.11 Remark. The Lindeberg condition (10.1) roughly says that the contribution of each $X_{n,k}$, $k = 1, \ldots, n$ to the sum should be "equal and small". This is not a necessary condition for the central limit theorem to hold as the following simple example shows. Let $X_{n,1} \sim N(0, \frac{1}{2})$ and $X_{n,k} \sim N(0, \sigma_{n,k}^2)$, $k = 2, \ldots, n$ with arbitrary $\sigma_{n,k}$ satisfying $\sum_{k=2}^n \sigma_{n,k}^2 = \frac{1}{2}$. Then, trivially, $Z_n \sim N(0, 1)$, but (10.1) does not hold (because even its consequence (10.2) fails). Excluding such situations where one of the summands dominates, it turns out that the Lindeberg condition is necessary for the central limit theorem to hold, as shown by Feller (we defer its proof to Appendix H).

10.12 Theorem. *Let $\{X_{n,}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of random variables with $\mathbb{E}X_{n,k}^2 < \infty$ such that for every $n \geq 1$, the variables $X_{n,1}, \ldots, X_{n,n}$ are independent. Let $\bar{X}_{n,k}$ be the normalised sequence and $Z_n = \sum_{k=1}^n \bar{X}_{n,k}$, as in Theorem* (10.8)*. Assume* (10.2)*. If the sequence $(Z_n)_n$ converges in distribution to a standard Gaussian random variable, then* (10.1) *holds.*

10.4 Multidimensional case

Let $X = (X_1, \ldots, X_d)$ be a random vector in $\mathbb{R}^d$. We define its **characteristic function** $\phi_X : \mathbb{R}^d \to \mathbb{C}$ by

$$\phi_X(t) = \mathbb{E}e^{i\langle t,X\rangle}, \qquad t \in \mathbb{R}^d.$$

10.13 Example. Let $X = (X_1, \ldots, X_d)$ be a standard Gaussian random vector in $\mathbb{R}^d$. Then by the independence of its components,

$$\phi_X(t) = \prod_{j=1}^d \phi_{X_j}(t_j) e^{-\frac{1}{2}\sum_{j=1}^d t_j^2}, \qquad t \in \mathbb{R}^d.$$

Let $Y = AX + b$ be an arbitrary Gaussian random vector in $\mathbb{R}^m$, where A is an $m \times n$ matrix and b is a vector in $\mathbb{R}^m$. Then

$$\phi_Y(t) = \mathbb{E}e^{i\langle AX+b,t\rangle} = e^{i\langle b,t\rangle}\mathbb{E}e^{i\left\langle X,A^\top t\right\rangle} = e^{i\langle b,t\rangle}e^{-|AX|^2/2} = e^{i\langle b,t\rangle}e^{-\left\langle AA^\top t,t\right\rangle/2},$$

thus, in general, we have

$$\phi_Y(t) = e^{i\langle b,t\rangle - \langle Qt,t\rangle/2}, \qquad Y \sim N(b, Q). \tag{10.4}$$

Working with the multidimensional CDFs in $\mathbb{R}^d$, we can prove a tightness result analogous to Theorem 8.17: *a sequence $(\mu_n)_n$ of Borel probability measures on $\mathbb{R}^d$ is tight if and only if it its every subsequence has a weakly convergent subsequence.* We also have the inversion formula analogous to the one from Theorem 9.7,

$$F_X(x_1,\dots,x_n) = \lim_{a\to\infty}\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\left(\frac{1}{2\pi}\int_{\mathbb{R}^d} e^{-i\langle s,t\rangle}\phi_X(s)e^{-\frac{|s|^2}{2a^2}}\mathrm{d}s\right)\mathrm{d}t.$$

These allow to establish Lévy's continuity theorem in $\mathbb{R}^d$ in a similar fashion and, consequently, the central limit theorem in $\mathbb{R}^d$.

10.14 Theorem. *Let (X_n) be a sequence of random vectors in $\mathbb{R}^d$ such that for every $t\in\mathbb{R}^d$, $\phi_{X_n}(t)\xrightarrow[n\to\infty]{}\phi(t)$ for some function $\phi:\mathbb{R}^d\to\mathbb{C}$ which is continuous at $t=0$. Then there is a random vector X in $\mathbb{R}^d$ such that $\phi=\phi_X$ and $X_n\xrightarrow{d}X$.*

Proof. Let $e_1,\dots,e_d$ be the standard basis vectors in $\mathbb{R}^d$. Fix $j\le d$ and $s\in\mathbb{R}$. By the assumption, $\phi_{\langle X_n,e_j\rangle}(s)=\phi_{X_n}(e_js)\to\phi_X(e_js)=\phi_{\langle X,e_j\rangle}(s)$. Consequently, by the 1-dimensional version of Lévy's theorem (Theorem 9.15), we get the tightness of the sequence $(\langle X_n,e_j\rangle)_n$ of each component of X_n, thus of the sequence (X_n). Having the tightness of $(X_n)_n$, we proceed exactly as in the proof of the 1-dimensional case. □

10.15 Theorem (Vanilla version of CLT in $\mathbb{R}^d$)**.** *Let $X_1,X_2,\dots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}X_i^2<\infty$ for each i, so that the covariance matrix $Q=\mathrm{Cov}(X)$ is well-defined. Then*

$$Z_n=\frac{X_1+\ldots+X_n-n\mathbb{E}X_1}{\sqrt{n}}\xrightarrow[n\to\infty]{}Z_Q,$$

where Z_Q is a Gaussian random vector in $\mathbb{R}^d$ with mean 0 and covariance matrix Q.

Proof. By Lévy's continuity theorem, it is enough to show that for every $t\in\mathbb{R}^d$, we have

$$\phi_{Z_n}(t)\xrightarrow[n\to\infty]{}e^{-\langle Qt,t\rangle/2}.$$

(recall (10.4)). Since $\mathbb{E}\langle t,X_1-\mathbb{E}X_1\rangle^2=\langle Qt,t\rangle$ and $\phi_{Z_n}(t)=\phi_{\langle Z_n,t\rangle}(1)$, this follows from the 1-dimensional vanilla CLT (Theorem 10.5) applied to the sequence $(\langle t,Z_n\rangle)_{n=1}^\infty$. □

10.5 Poisson limit theorem

The following result, sometimes called the law of rare events, explains how the Poisson distribution arises as a limit of the binomial distribution when the expected number of successes converges to a constant as the number of Bernoulli trials goes to infinity.

10.16 Theorem (Vanilla Poisson limit theorem)**.** *Let a sequence of numbers $p_n\in[0,1]$ be such that $np_n\xrightarrow[n\to\infty]{}\lambda$ for some $\lambda>0$. Let S_n be a binomial random variable with parameters p_n and n. Then $S_n\xrightarrow{d}X$, where X is a Poisson random variable with parameter λ.*

Proof. For nonnegative integer-valued random variables convergence in distribution is equivalent to the pointwise convergence of the probability mass functions (Exercise 8.4). Thus, $S_n \xrightarrow{d} X$ if and only if $\mathbb{P}(S_n = k) \xrightarrow[n\to\infty]{} \mathbb{P}(X = k)$, for every integer $k \geq 0$. Fix then such k and note that as $n \to \infty$, we have

$$\begin{aligned}\mathbb{P}(S_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} &= \frac{n(n-1)\dots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\ &= \frac{1+O(n^{-1})}{k!}(np_n)^k (1-p_n)^{n-k}.\end{aligned}$$

By the assumption, $np_n \to \lambda$. In particular, $p_n \to 0$. Consequently, $(1-p_n)^{-k} \to 1$ and $(1-p_n)^n \to e^{-\lambda}$, so

$$\mathbb{P}(S_n = k) \xrightarrow[n\to\infty]{} \frac{1}{k!}\lambda^k e^{-\lambda} = \mathbb{P}(X = k).$$

□

There is a generalisation to triangular arrays of Bernoulli random variables satisfying two assumptions: (i) the means stabilise in the limit and (ii) each random variable has a small contribution.

10.17 Theorem (Poisson limit theorem)**.** *Let $\{X_{n,}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of Bernoulli random variables such that for every $n \geq 1$, the variables $X_{n,1}, \dots, X_{n,n}$ are independent. If they satisfy the following two conditions*

(i) $\mathbb{E}\sum_{k=1}^n X_{n,k} \xrightarrow[n\to\infty]{} \lambda$ *for some* $\lambda \in (0,\infty)$,

(ii) $\max_{1\leq k\leq n} \mathbb{E}X_{n,k} \xrightarrow[n\to\infty]{} 0$,

then

$$X_{n,1} + \dots + X_{n,n} \xrightarrow[n\to\infty]{d} Z,$$

where Z is a Poisson random variable with parameter λ.

For now we show a Fourier-analytic proof. In the next chapter, we include another proof, based on the total variation distance, which gives some quantitative bounds on the rate of convergence and moreover is quite simple.

1st proof of Theorem 10.17. Let $S_n = X_{n,1} + \dots + X_{n,n}$ and let $p_{n,k} = \mathbb{E}X_{n,k}$ be the parameter of the Bernoulli distribution of $X_{n,k}$. We have

$$\phi_{S_n}(t) = \prod_{k=1}^n \phi_{X_{n,k}}(t) = \prod_{k=1}^n (1 + p_{n,k}(e^{it} - 1))$$

and

$$\phi_Z(t) = e^{\lambda(e^{it}-1)}.$$

Fix $t \in \mathbb{R}$. Assumption (i) is that $\sum_{k=1}^n p_{n,k} \xrightarrow[n\to\infty]{} \lambda$, thus

$$\prod_{k=1}^n e^{p_{n,k}(e^{it}-1)} = e^{(\sum_{k=1}^n p_{n,k})(e^{it}-1)} \xrightarrow[n\to\infty]{} e^{\lambda(e^{it}-1)} = \phi_Z(t)$$

and it suffices to show that

$$u_n = \prod_{k=1}^n (1 + p_{n,k}(e^{it} - 1)) - \prod_{k=1}^n e^{p_{n,k}(e^{it}-1)} \xrightarrow[n\to\infty]{} 0.$$

For $0 \le p \le 1$, we have

$$|e^{p(e^{it}-1)}| = e^{p\mathrm{Re}(e^{it}-1)} = e^{p(\cos t - 1)} \le 1$$

and

$$|1 + p(e^{it} - 1)| = |1 - p + pe^{it}| \le 1 - p + p|e^{it}| = 1,$$

so by Lemma 10.1 with $\theta = 1$, we get

$$|u_n| = \left| \prod_{k=1}^n (1 + p_{n,k}(e^{it} - 1)) - \prod_{k=1}^n e^{p_{n,k}(e^{it}-1)} \right| \le \sum_{k=1}^n \left| 1 + p_{n,k}(e^{it} - 1) - e^{p_{n,k}(e^{it}-1)} \right|.$$

Thanks to assumption (ii), for large enough n, we can use Lemma 10.2 applied to $z = p_{n,k}(e^{it} - 1)$ and thus get

$$|u_n| \le \sum_{k=1}^n p_{n,k}^2 |e^{it} - 1|^2 \le 4 \sum_{k=1}^n p_{n,k}^2 \le 4 \left(\max_{1\le k\le n} p_{n,k} \right) \sum_{k=1}^n p_{n,k} \xrightarrow[n\to\infty]{} 0.$$

□

10.6 Exercises

1. Let S be the number of ones when throwing a fair die 18000 times. Using the central limit theorem, find a Gaussian approximation to $\mathbb{P}(2950 < S < 3050)$.

2. Let G be a standard Gaussian random vector in $\mathbb{R}^n$. Let $\|G\| = \sqrt{G_1^2 + \ldots + G_n^2}$ be its magnitude. Let $a_n = \mathbb{P}(\sqrt{n} - 1 \leq \|G\| \leq \sqrt{n} + 1)$. Find $a = \lim_{n\to\infty} a_n$.

3. For $\lambda > 0$, let X_λ be a Poisson random variable with parameter λ. Let (λ_n) be a sequence of positive numbers with $\lambda_n \to \infty$ as $n \to \infty$. Find the weak limit of the sequence $(\frac{X_{\lambda_n} - \lambda_n}{\sqrt{\lambda_n}})$.

4. Show that $e^{-n} \sum_{k=1}^n \frac{n^k}{k!} \xrightarrow[n\to\infty]{} \frac{1}{2}$.

   Hint: Poiss(n) random variable is a sum of n i.i.d. Poiss(1) random variables.

5. Let X_n be a Poisson random variable with parameter n. Let Z be a standard Gaussian random variable. Show that

   (i) $\mathbb{E}\left(\frac{X_n - n}{\sqrt{n}}\right)_- = e^{-n} \frac{n^{n+1/2}}{n!}$,

   (ii) $\left(\frac{X_n - n}{\sqrt{n}}\right)_- \xrightarrow[n\to\infty]{d} Z_-$,

   (iii) $\mathbb{E}\left(\frac{X_n - n}{\sqrt{n}}\right)_- \xrightarrow[n\to\infty]{} \mathbb{E}Z_-$,

   (iv) conclude Stirling's formula, $\frac{n!}{e^{-n} n^{n+1/2}} \xrightarrow[n\to\infty]{} \sqrt{2\pi}$.

   Here, as usual, $X_- = \max\{-X, 0\}$ denotes the negative part of X.

6. Suppose that a random variable X with variance one has the following property: $\frac{X+X'}{\sqrt{2}}$ has the same distribution as X, where X' is an independent copy of X. Show that $X \sim N(0,1)$.

7. Suppose that a random vector $X = (X_1, X_2)$ in $\mathbb{R}^2$ is such that $\mathbb{E}X_1^2, \mathbb{E}X_2^2 < \infty$, X is rotationally invariant (UX has the same distribution as X for every 2×2 orthogonal matrix U) and X has independent components, that is X_1 and X_2 are independent. Show that

   (a) X_1 has the same distribution as X_2 and is symmetric (that is has the same distribution as $-X_1$)

   (b) X_1 has the same distribution as $\frac{X_1+X_2}{\sqrt{2}}$

   (c) Using Exercise 10.6, deduce that $X \sim N(0, \sigma^2 I_{2\times 2})$.

   Generalise this characterisation of multiples of standard Gaussian vectors to $\mathbb{R}^n$.

8. A roulette wheel has slots numbered 1–36 (18 red and 18 black) and two slots numbered 0 and 00 that are painted green. You can bet \$1 that the ball will land in a

red (or black) slot and win \$1 if it does. What is the expected value of your winnings after 361 spins of the wheel and what is approximately the probability that it will be positive?

9. A biased coin showing heads with probability p is thrown 2500 times. Using the Poisson or central limit theorem, find approximately the probability of getting no heads when a) $p = \frac{1}{2500}$, b) $p = \frac{1}{5}$? How about the probability of getting 500 heads?

10. Prove Remark 10.6.

11. Prove Remark 10.7. Here is a suggestion how to proceed.

    (a) Considering $X_i - X_i'$ instead of X_i, where X_i' is an independent copy of X_i (independent of all the other random variables), show that we can assume the X_i are symmetric (Exercise 7.1 may be of use).

    (b) Show that for independent symmetric random variables $X_1, \dots, X_n$ and every $t, A \geq 0$, we have
    $$\mathbb{P}\left(X_1 + \cdots + X_n \geq t\right) \geq \frac{1}{2}\mathbb{P}\left(X_1 \mathbf{1}_{|X_1| \leq A} + \cdots + X_n \mathbf{1}_{|X_n| \leq A} \geq t\right).$$
    To this end, consider $S = X_1 \mathbf{1}_{|X_1| \leq A} + \cdots + X_n \mathbf{1}_{|X_n| \leq A}$, $T = X_1 \mathbf{1}_{|X_1| > A} + \cdots + X_n \mathbf{1}_{|X_n| > A}$ and use that (S, T) has the same distribution as $(\pm S, \pm T)$, by symmetry.

    (c) Apply (b) with $t = u\sqrt{n}$ and treat the right hand side with the central limit theorem to show that if $\mathbb{E}X_i^2 = +\infty$ ($\mathbb{E}X_i^2 \mathbf{1}\, |X_i| \leq A \nearrow \mathbb{E}X_i^2$ as $A \to \infty$), then $\mathbb{P}\left(X_1 + \cdots + X_n \geq u\sqrt{n}\right) > \frac{1}{5}$ for large enough n. Choose first u, then A, then n to reach a contradiction.

12. Let $X_1, X_2, \dots$ be i.i.d. random variables with mean 0 and variance 1. Show that
    $$\frac{\sqrt{n}(X_1 + \cdots + X_n)}{X_1^2 + \cdots + X_n^2} \xrightarrow[n\to\infty]{d} Z$$
    and
    $$\frac{X_1 + \cdots + X_n}{\sqrt{X_1^2 + \cdots + X_n^2}} \xrightarrow[n\to\infty]{d} Z,$$
    where Z is a standard Gaussian random variable.

13. Let $X_1, X_2, \dots$ be i.i.d. nonnegative random variables with mean 1 and variance σ^2. Show that
    $$2(\sqrt{X_1 + \cdots + X_n} - \sqrt{n}) \xrightarrow[n\to\infty]{d} Z,$$
    where Z is a Gaussian random variable with mean 0 and variance σ^2.

14. Let $a > 0$. Let $X_1, X_2, \dots$ be i.i.d. random variables such that $\mathbb{P}\left(X_k = a\right) = \mathbb{P}\left(X_k = 1/a\right) = \frac{1}{2}$. Investigate the weak convergence of the sequence $Z_n = (X_1 \cdot \ldots \cdot X_n)^{1/\sqrt{n}}$, $n \geq 1$.

15. Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean 0 and variance σ^2. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function differentiable at 0. Show that

$$\sqrt{n}\left(f\left(\frac{X_1 + \cdots + X_n}{n}\right) - f(0)\right) \xrightarrow[n\to\infty]{d} Z,$$

where Z is a Gaussian random variable with mean 0 and variance $\sigma^2 f'(0)^2$.

16. Show that the Lindeberg condition (10.1) implies that for every $\varepsilon > 0$, we have

$$\max_{1\leq k\leq n} \mathbb{P}\left(|\bar{X}_{n,k}| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

17. Under the notation and assumptions of Theorem 10.8, consider the so-called *Lyapunov condition*: there is $\delta > 0$ such that $\mathbb{E}|X_{n,k}|^{2+\delta} < \infty$ for all n, k and

$$\sum_{k=1}^{n} \mathbb{E}|\bar{X}_{n,k}|^{2+\delta} \xrightarrow[n\to\infty]{} 0.$$

Show that this implies Lindeberg's condition (10.1).

18. Let $X_1, X_2, \ldots$ be independent random variables such that for some constant $C > 0$ and all n, $|X_n| \leq C$. If $\sum_{k=1}^{n} \mathrm{Var}(X_k) = \infty$, then $\frac{X_1+\cdots+X_n - \mathbb{E}(X_1+\cdots+X_n)}{\sqrt{\sum_{k=1}^{n}\mathrm{Var}(X_k)}}$ converges to a standard Gaussian.

19. Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean 0 and variance 1. Given positive parameters α and t, find

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n^\alpha}\right| > t\right).$$

20. Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{P}(X_k = k) = \mathbb{P}(X_k = -k) = \frac{1}{2}$, $k \geq 1$. Investigate the convergence in distribution of the sequence

$$\frac{X_1 + \cdots + X_n}{\sqrt{\mathrm{Var}(X_1 + \cdots + X_n)}}, \qquad n \geq 1.$$

21. Let $U_1, U_2, \ldots$ be independent random variables with U_k being uniform on $[-a_k, a_k]$. Let $\sigma_n = \sqrt{\mathrm{Var}(X_1 + \cdots + X_n)}$. Investigate the convergence in distribution of the sequence

$$\frac{U_1 + \cdots + U_n}{\sigma_n}, \qquad n \geq 1,$$

in the following two cases

a) The sequence (a_n) is bounded and $\sigma_n \to \infty$ as $n \to \infty$.

b) $\sum a_n^2 < \infty$.

22. Show that the components X_j of a random vector $X = (X_1, \ldots, X_d)$ in $\mathbb{R}^d$ are independent if and only if $\phi_X(t) = \prod_{k=1}^{n} \phi_{X_k}(t_k)$ for every $t \in \mathbb{R}^d$.

23. Show that a random vector X in $\mathbb{R}^d$ is Gaussian with mean $b \in \mathbb{R}^d$ and covariance matrix Q if and only if for every $t \in \mathbb{R}^d$, $\langle t, X \rangle$ is a Gaussian random variable with mean $\langle t, b \rangle$ and variance $\langle Qt, t \rangle$.

24. Let T_n and Z_n be the number of inversions and cycles, respectively, in a permutation chosen uniformly at random from the set of all permutations on an n-element set. Show that

$$6\frac{S_n - n^2/4}{n^{3/2}} \xrightarrow[n\to\infty]{d} Z$$

and

$$\frac{Z_n - \log n}{(\log n)^{1/2}} \xrightarrow[n\to\infty]{d} Z,$$

where Z is a standard Gaussian random variable.

11 Quantitative versions of the limit theorem*

11.1 Berry-Esseen theorem via Stein's mehtod

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite variance. Let $Z_n = \frac{X_1+\ldots+X_n-n\mathbb{E}X_1}{\sqrt{n\operatorname{Var}(X_1)}}$ and let Z be a standard Gaussian random variable. The central limit theorem asserts that for every $t \in \mathbb{R}$,

$$\mathbb{P}\left(Z_n \le t\right) \xrightarrow[n\to\infty]{} \mathbb{P}\left(Z \le t\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\mathrm{d}x.$$

For practical purposes, we would like to know what is the error we make when we use $\mathbb{P}\left(Z \le t\right)$ as an approximation to $\mathbb{P}\left(Z_n \le t\right)$ for large n. This is possible under an additional assumption (finite third moment) and is settled in the following theorem, discovered independently by Berry and Esseen.

11.1 Theorem (Berry-Esseen theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}|X_1|^3 < \infty$. Let*

$$Z_n = \frac{X_1 + \ldots + X_n - n\mathbb{E}X_1}{\sqrt{n\operatorname{Var}(X_1)}},$$

$$\rho = \mathbb{E}\left|\frac{X_1 - \mathbb{E}X_1}{\sqrt{\operatorname{Var}(X_1)}}\right|^3$$

and let Z be a standard Gaussian random variable. There is a universal constant C such that for every $n \ge 1$ and every $t \in \mathbb{R}$, we have

$$\left|\mathbb{P}\left(Z_n \le t\right) - \mathbb{P}\left(Z \le t\right)\right| \le \frac{C\rho}{\sqrt{n}}.$$

11.2 Remark. We present a proof which will give $C = 15.2$, but this value is far from optimal. Currently, the best value is $C = 0.4774$ (estblished via Fourier analytic methods in [8]). Esseen proved a lower bound: $C \ge \frac{10+\sqrt{3}}{6\sqrt{2\pi}} = 0.4097\ldots$

11.3 Remark. The rate $1/\sqrt{n}$ of the error is optimal. Consider i.i.d. symmetric random signs $\varepsilon_1, \varepsilon_2, \ldots$ and let $Z_n = \frac{\varepsilon_1+\ldots+\varepsilon_n}{\sqrt{n}}$. For even n, by symmetry, we have

$$\mathbb{P}\left(Z_n \le 0\right) = \frac{1 + \mathbb{P}\left(\varepsilon_1 + \ldots + \varepsilon_n = 0\right)}{2} = \frac{1}{2} + \frac{1}{2}\binom{n}{n/2}\frac{1}{2^n},$$

thus, thanks to Stirling's formula,

$$\left|\mathbb{P}\left(Z_n \le 0\right) - \mathbb{P}\left(Z \le 0\right)\right| = \left|\mathbb{P}\left(Z_n \le 0\right) - \frac{1}{2}\right| = \frac{1}{2}\binom{n}{n/2}\frac{1}{2^n} \approx \frac{1}{2}\frac{\sqrt{2}}{\sqrt{\pi n}},$$

so in this case the error is of the order $1/\sqrt{n}$.

For the proof the Berry-Esseen theorem, we shall need the following elementary tail bound for the standard Gaussian distribution.

11.4 Lemma. *For $x > 0$, we have*

(i) $\int_x^\infty e^{-u^2/2}\mathrm{d}u \leq \sqrt{\frac{\pi}{2}}e^{-x^2/2}$,

(ii) $\int_x^\infty e^{-u^2/2}\mathrm{d}u \leq \frac{1}{x}e^{-x^2/2}$.

Proof. (i) let $f(x) = \sqrt{\frac{\pi}{2}}e^{-x^2/2} - \int_x^\infty e^{-u^2/2}\mathrm{d}u$. Since $f'(x) = \left(1 - x\sqrt{\frac{\pi}{2}}\right)e^{-x^2/2}$ is first positive, then negative, f first increases, then decreases. Combined with $f(0) = 0$ and $f(x) \xrightarrow[t\to\infty]{} 0$, this proves that $f(x) \geq 0$.

(ii) We have $\int_x^\infty xe^{-u^2/2}\mathrm{d}u \leq \int_x^\infty ue^{-u^2/2}\mathrm{d}u = e^{-u^2/2}$. $\square$

Proof of Theorem 11.1. For $t, x \in \mathbb{R}$ and $\lambda > 0$ define functions

$$h_t(x) = \mathbf{1}_{(-\infty,t]}(x),$$

and their continuous linear approximations

$$h_{t,\lambda}(x) = \begin{cases} 1, & x \leq t, \\ 1 - \frac{x-t}{\lambda}, & t < x \leq t+\lambda, \\ 0, & x > t+\lambda. \end{cases}$$

We will frequently use the following integral representation

$$h_{t,\lambda}(x) = \int_x^\infty \frac{1}{\lambda}\mathbf{1}_{(t,t+\lambda)}(s)\mathrm{d}s.$$

Given $\gamma \geq 1$, define the class of random variables

$$\mathscr{L}_\gamma = \{X,\ X \text{ is random variable such that } \mathbb{E}X = 0,\ EX^2 = 1,\ \mathbb{E}|X|^3 = \gamma\}$$

and for $n = 1, 2, \ldots$ define two quantities

$$B_0(\gamma, n) = \sup_{X_1,\ldots,X_n i.i.d., X_i \in \mathscr{L}_\gamma} \sup_{t\in\mathbb{R}} |\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z)|,$$

$$B(\lambda, \gamma, n) = \sup_{X_1,\ldots,X_n i.i.d., X_i \in \mathscr{L}_\gamma} \sup_{t\in\mathbb{R}} |\mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z)|.$$

Plainly, $\mathbb{P}(X \leq t) = \mathbb{E}\,\mathbf{1}_{X\leq t} = \mathbb{E}h_t(X)$, so to prove the theorem, we would like to show that

$$\frac{\sqrt{n}}{\gamma}B_0(\gamma, n) \leq C, \qquad n \geq 1, \gamma \geq 1.$$

This is clear for $n = 1$ with $C = 1$ because $|\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z)| \leq 1$, so from now on we assume $n \geq 2$ and divide the rest of the proof into several steps.

Step 1: regularisation (upper bound for B_0 in terms of B). Since $h_{t-\lambda} \leq h_t \leq h_{t,\lambda}$, we get

$$\begin{aligned} \mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) &\leq \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_t(Z) \\ &= \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z) + \mathbb{E}h_{t,\lambda}(Z) - \mathbb{E}h_t(Z) \\ &\leq \mathbb{E}h_{t,\lambda}(Z_n) - \mathbb{E}h_{t,\lambda}(Z) + \mathbb{E}h_{t+\lambda}(Z) - \mathbb{E}h_t(Z). \end{aligned}$$

Observe that the first difference is upper bounded by $B(\lambda, \gamma, n)$ by its definition. The second difference is

$$\mathbb{P}\left(t < Z \leq t+\lambda\right) = \int_t^{t+\lambda} e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} \leq \int_t^{t+\lambda} \frac{\mathrm{d}x}{\sqrt{2\pi}} = \frac{\lambda}{\sqrt{2\pi}}.$$

Altogether,

$$\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) \leq B(\lambda, \gamma, n) + \frac{\lambda}{\sqrt{2\pi}}.$$

Similarly,

$$\mathbb{E}h_t(Z_n) - \mathbb{E}h_t(Z) \geq -B(\lambda, \gamma, n) - \frac{\lambda}{\sqrt{2\pi}}.$$

Thus

$$B_0(\gamma, n) \leq B(\lambda, \gamma, n) + \frac{\lambda}{\sqrt{2\pi}}.$$

Step 2: Stein's method ("encoding" $\mathbb{E}h(Z)$ into a function). Fix $t \in \mathbb{R}$, $\lambda > 0$ and set $h = h_{t,\lambda}$. Our goal is to upper bound B, so to upper bound $\mathbb{E}h(Z_n) - \mathbb{E}h(Z)$. The heart of Stein's method is to rewrite this in terms of Z_n only. Let

$$f(x) = e^{x^2/2} \int_{-\infty}^x [h(u) - \mathbb{E}h(Z)] e^{-u^2/2} \mathrm{d}u.$$

Then

$$f'(x) - xf(x) = h(x) - \mathbb{E}h(Z),$$

so

$$\mathbb{E}h(Z_n) - \mathbb{E}h(Z) = \mathbb{E}\big[f'(Z_n) - Z_n f(Z_n)\big]. \tag{11.1}$$

Step 3: Estimates for f and f'. For every $x \in \mathbb{R}$, we have

$$|f(x) \leq \sqrt{\frac{\pi}{2}}, \qquad |xf(x)| \leq 1, \qquad |f'(x)| \leq 2 \tag{11.2}$$

and for every $x, y \in \mathbb{R}$, we have

$$|f'(x+y) - f'(x)| \leq |y| \left(\sqrt{\frac{\pi}{2}} + 2|x| + \frac{1}{\lambda} \int_0^1 \mathbf{1}_{(t,t+\lambda)}(x+vy) \mathrm{d}v \right). \tag{11.3}$$

Indeed, since h takes values in $[0,1]$, we have $|h(u) - h(v)| \leq 1$ for any u and v, so for $x < 0$,

$$|f(x)| \leq e^{x^2/2} \int_{-\infty}^x |h(u) - \mathbb{E}h(Z)| e^{-u^2/2} \mathrm{d}u \leq e^{x^2/2} \int_{-\infty}^x e^{-u^2/2} \mathrm{d}u = e^{x^2/2} \int_{-x}^\infty e^{-u^2/2} \mathrm{d}u,$$

which by Lemma 11.4 (i) is upper bounded by $\sqrt{\frac{\pi}{2}}$. For $x > 0$, notice that $\int_{-\infty}^\infty [h(u) - \mathbb{E}h(Z)] e^{-u^2/2} \frac{\mathrm{d}u}{\sqrt{2\pi}} = 0$, so

$$f(x) = -e^{x^2/2} \int_x^\infty [h(u) - \mathbb{E}h(Z)] e^{-u^2/2} \mathrm{d}u$$

and as above we get the bound $|f(x)| \le \sqrt{\frac{\pi}{2}}$. To bound $xf(x)$ we proceed the same way but use Lemma 11.4 (ii). Finally, since $f'(x) = xf(x) + h(x) - \mathbb{E}h(Z)$ (Step 2), we get

$$|f'(x)| \le |xf(x)| + |h(x) - \mathbb{E}h(Z)| \le 1 + 1 = 2.$$

This establishes (11.2). To prove (11.3), we use the formula for f' from Step 2 and write

$$\begin{aligned}
|f'(x+y) - f'(x)| &= |(x+y)f(x+y) + h(x+y) - xf(x) - h(x)| \\
&= |yf(x+y) + x(f(x+y) - f(x)) + h(x+y) - h(x)| \\
&\le |y|\sqrt{\frac{\pi}{2}} + 2|x||y| + |h(x+y) - h(x)|,
\end{aligned}$$

where in the last inequality we used the mean value theorem writing $f(x+y) - f(x) = f'(\xi)y$ and then estimating $|f'(\xi)| \le 2$. Finally, by the integral representation for h,

$$|h(x+y) - h(x)| = \left|\frac{1}{\lambda}\int_x^{x+y} \mathbf{1}_{(t,t+\lambda)}(u)\mathrm{d}u\right| = \left|\frac{y}{\lambda}\int_0^1 \mathbf{1}_{(t,t+\lambda)}(x+vy)\mathrm{d}v\right|$$

which after plugging back in the previous inequality finishes the proof of (11.3).

Step 4: Estimates for $B(\lambda,\gamma,n)$ via (11.1). To estimate $B(\lambda,\gamma,n)$, we need to upper bound $\mathbb{E}h(Z_n) - \mathbb{E}h(Z) = \mathbb{E}[f'(Z_n) - Z_nf(Z_n)]$ (recall (11.1) from Step 2). Here we exploit that $Z_n = \frac{X_1+\ldots+X_n}{\sqrt{n}}$ is a sum of i.i.d. random variables. Since the X_i have the same distribution, by linearity,

$$\mathbb{E}Z_nf(Z_n) = \mathbb{E}\frac{\sum X_i}{\sqrt{n}}f(Z_n) = \sqrt{n}\mathbb{E}X_nf(Z_n).$$

Note also that $Z_n = \sqrt{\frac{n-1}{n}}Z_{n-1} + \frac{X_n}{\sqrt{n}}$ and thus

$$\begin{aligned}
\mathbb{E}[f'(Z_n) - Z_nf(Z_n)] &= \mathbb{E}[f'(Z_n) - \sqrt{n}X_nf(Z_n)] \\
&= \mathbb{E}\Big[f'(Z_n) - \sqrt{n}X_n\int_0^1 \frac{\mathrm{d}}{\mathrm{d}u}f\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right)\mathrm{d}u \\
&\qquad\qquad - \sqrt{n}X_nf\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\Big]
\end{aligned}$$

By independence and $\mathbb{E}X_n = 0$ the last term vanishes and after computing the derivative we get

$$\begin{aligned}
\mathbb{E}[f'(Z_n) - Z_nf(Z_n)] &= \mathbb{E}\left[f'(Z_n) - X_n^2\int_0^1 f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right)\mathrm{d}u\right] \\
&= \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\right] \\
&\quad + \mathbb{E}\Big[- X_n^2\int_0^1 \Big\{f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1} + u\frac{X_n}{\sqrt{n}}\right) \\
&\qquad\qquad - f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\Big\}\mathrm{d}u\Big]
\end{aligned}$$

where in the last equality we used independence and $\mathbb{E}X_n^2 = 1$. We bound the two terms separately.

Step 4.1: First term. Using $Z_n = \sqrt{\frac{n-1}{n}} Z_{n-1} + \frac{X_n}{\sqrt{n}}$ and (11.3),

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \mathbb{E}\left| \frac{X_n}{\sqrt{n}} \right| \left(\sqrt{\frac{\pi}{2}} + 2\sqrt{\frac{n-1}{n}} |Z_{n-1}| + \frac{1}{\lambda} \int_0^1 \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u\frac{X_n}{\sqrt{n}} \right) \mathrm{d}u \right)
\end{aligned}$$

Since $\mathbb{E}|X_n| \leq \sqrt{\mathbb{E}|X_n|^2} = 1$ and similarly $\mathbb{E}|Z_{n-1}| \leq 1$, as well as trivially $\sqrt{\frac{n-1}{n}} \leq 1$, we get

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} + 2\frac{1}{\sqrt{n}} + \frac{1}{\lambda\sqrt{n}} \mathbb{E}_{X_n}\left[|X_n| \int_0^1 \mathbb{E}_{Z_{n-1}} \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u\frac{X_n}{\sqrt{n}} \right) \mathrm{d}u \right],
\end{aligned}$$

where in the last term we used the independence of X_n and Z_{n-1}. Note that

$$\begin{aligned}
&\mathbb{E}_{Z_{n-1}} \mathbf{1}_{(t,t+\lambda)} \left(\sqrt{\frac{n-1}{n}} Z_{n-1} + u\frac{X_n}{\sqrt{n}} \right) \\
&= \mathbb{P}_{Z_{n-1}} \left(\left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}} < Z_{n-1} < \left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}} + \lambda\sqrt{\frac{n}{n-1}} \right),
\end{aligned}$$

Denoting $a = \left(t - u\frac{X_n}{\sqrt{n}} \right) \sqrt{\frac{n}{n-1}}$ and estimating $\frac{n}{n-1} \leq 2$, we get that this probability is upper bounded by

$$\mathbb{P}\left(a < Z_{n-1} < a + \lambda\sqrt{2} \right)$$

which we rewrite in order to upper bound it in terms of B_0,

$$\begin{aligned}
\mathbb{P}\left(a < Z_{n-1} < a + \lambda\sqrt{2} \right) &= \mathbb{P}\left(Z_{n-1} < a + \lambda\sqrt{2} \right) - \mathbb{P}\left(Z < a + \lambda\sqrt{2} \right) \\
&\quad + \mathbb{P}\left(Z \leq a \right) - \mathbb{P}\left(Z_{n-1} \leq a \right) + \mathbb{P}\left(a \leq Z \leq a + \lambda\sqrt{2} \right) \\
&\leq 2B_0(\gamma, n-1) + \frac{\lambda\sqrt{2}}{\sqrt{2\pi}},
\end{aligned}$$

where the last term was crudely bounded using the maximum of standard Gaussian density. Plugging this back yields

$$\begin{aligned}
&\left| \mathbb{E}\left[f'(Z_n) - f'\left(\sqrt{\frac{n-1}{n}} Z_{n-1} \right) \right] \right| \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\frac{\pi}{2}} + 2\frac{1}{\sqrt{n}} + \frac{1}{\lambda\sqrt{n}} \mathbb{E}\left[|X_n| \left(2B_0(\gamma, n-1) + \frac{\lambda}{\sqrt{\pi}} \right) \right] \\
&\leq \frac{1}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{2B_0(\gamma, n-1)}{\lambda} + \frac{1}{\sqrt{\pi}} \right) \\
&\leq \frac{\gamma}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{2B_0(\gamma, n-1)}{\lambda} + \frac{1}{\sqrt{\pi}} \right).
\end{aligned}$$

Step 4.2: Second term. Using again (11.3) and independence,

$$\begin{aligned}
&\left|\mathbb{E}\left[-X_n^2\int_0^1\left\{f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+u\frac{X_n}{\sqrt{n}}\right)-f'\left(\sqrt{\frac{n-1}{n}}Z_{n-1}\right)\right\}\mathrm{d}u\right]\right|\\
&\le \mathbb{E}X_n^2\frac{|X_n|}{\sqrt{n}}\int_0^1 u\Bigg(\sqrt{\frac{\pi}{2}}+2\sqrt{\frac{n-1}{n}}|Z_{n-1}|\\
&\qquad\qquad +\frac{1}{\lambda}\int_0^1\mathbf{1}_{(t,t+\lambda)}\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+uv\frac{X_n}{\sqrt{n}}\right)\mathrm{d}v\Bigg)\mathrm{d}u\\
&\le \mathbb{E}\frac{|X_n|^3}{\sqrt{n}}\int_0^1 u\Bigg(\sqrt{\frac{\pi}{2}}+2\mathbb{E}_{Z_{n-1}}|Z_{n-1}|\\
&\qquad\qquad +\frac{1}{\lambda}\int_0^1\mathbb{E}_{Z_{n-1}}\mathbf{1}_{(t,t+\lambda)}\left(\sqrt{\frac{n-1}{n}}Z_{n-1}+uv\frac{X_n}{\sqrt{n}}\right)\mathrm{d}v\Bigg)\mathrm{d}u\\
&\le \mathbb{E}\frac{|X_n|^3}{\sqrt{n}}\int_0^1 u\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\lambda}\left(2B_0(\gamma,n-1)+\frac{\lambda}{\sqrt{\pi}}\right)\right)\mathrm{d}u\\
&=\frac{\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).
\end{aligned}$$

Putting Steps 4.1 and 4.2 together yields

$$|\mathbb{E}[f'(Z_n)-Z_nf(Z_n)]|\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).$$

By Step 2, this gives

$$B(\lambda,\gamma,n)\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right).$$

Step 5: Optimisation of parameters and end of proof. The previous inequality and Step 1 yield

$$\begin{aligned}
B_0(\gamma,n)&\le\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{2B_0(\gamma,n-1)}{\lambda}+\frac{1}{\sqrt{\pi}}\right)+\frac{\lambda}{\sqrt{2\pi}}\\
&=\frac{3\gamma}{2\sqrt{n}}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{1}{\lambda}\frac{3\gamma B_0(\gamma,n-1)}{\sqrt{n}}+\frac{\lambda}{\sqrt{2\pi}}.
\end{aligned}$$

Set $\lambda=\alpha\frac{\gamma}{\sqrt{n}}$, $\alpha>0$ and multiply both sides by $\frac{\sqrt{n}}{\gamma}$ to get

$$B_0(\gamma,n)\frac{\sqrt{n}}{\gamma}\le\frac{3}{2}\left(\sqrt{\frac{\pi}{2}}+2+\frac{1}{\sqrt{\pi}}\right)+\frac{3}{\alpha}B_0(\gamma,n-1)\frac{\sqrt{n}}{\gamma}+\frac{\alpha}{\sqrt{2\pi}}.$$

Let

$$B=\sup_{\gamma\ge1,n\ge2}B_0(\gamma,n)\frac{\sqrt{n}}{\gamma}.$$

For $n\ge2$, we have

$$B_0(\gamma,n-1)\frac{\sqrt{n}}{\gamma}=B_0(\gamma,n-1)\frac{\sqrt{n-1}}{\gamma}\sqrt{\frac{n}{n-1}}\le\max\left\{\sqrt{2},B\sqrt{\frac{3}{2}}\right\}$$

(recall that trivially $B_0(\gamma,1)\frac{1}{\gamma} \le 1$). If $B > \frac{2}{\sqrt{3}}$, we thus obtain

$$B \le \frac{3}{2}\left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}}\right) + \frac{3}{\alpha}B\sqrt{\frac{3}{2}} + \frac{\alpha}{\sqrt{2\pi}}.$$

For $\alpha > 3\sqrt{\frac{3}{2}}$ this gives

$$B \le \frac{\alpha}{\alpha - 3\sqrt{\frac{3}{2}}}\frac{3}{2}\left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}}\right) + \frac{\alpha^2}{\alpha - 3\sqrt{\frac{3}{2}}}\frac{1}{\sqrt{2\pi}}.$$

The choice of α which equates the two terms on the right hand side gives

$$B < 15.4.$$

Optimising over α (which requires more computations) gives a slightly better estimate

$$B < 15.2.$$

□

11.5 Remark. The proof presented here is from [2]. The heart of the argument is based on Stein's method (Step 2), introduced by Charles Stein, who developed this influential technique for teaching purposes of the central limit theorem for his course in statistics.

11.6 Example. Let us apply the Berry-Esseen theorem to i.i.d. Bernoulli random variables $X_1, \ldots, X_n$ with parameter $0 < p < 1$. We have $\mathbb{E}X_i = p$, $\mathrm{Var}(X_i) = p(1-p)$ and we obtain for every real t and every integer $n \ge 1$

$$\left|\mathbb{P}\left(\frac{X_1 + \ldots + X_n - np}{\sqrt{np(1-p)}} \le t\right) - \mathbb{P}\left(Z \le t\right)\right| \le C\frac{\rho}{\sqrt{n}},$$

where

$$\rho = \mathbb{E}\left|\frac{X_1 - p}{\sqrt{p(1-p)}}\right|^3 = \frac{p(1-p)^3 + (1-p)p^3}{\sqrt{p(1-p)}^3} = \frac{1-2p(1-p)}{\sqrt{p(1-p)}}.$$

In particular, when np is of the constant order for large n, the Berry-Esseen theorem is not useful at all because the bound of the error, $C\frac{\rho}{\sqrt{n}}$ is of the order $\frac{C}{\sqrt{np(1-p)}}$ which is constant. This might suggest that the Gaussian approximation is not valid in this case, which is in fact true in view of the Poisson limit theorem (Theorem 10.16).

11.2 Local central limit theorem

In applications we often need to address the following: suppose $X_1, X_2, \ldots$ are i.i.d. discrete, say integer-valued random variables and we would like to know for large n what is the approximate value of $\mathbb{P}\left(X_1 + \ldots + X_n = x_n\right)$ for some $x_n \in \mathbb{Z}$. If $\mathbb{E}X_i^2 < \infty$,

$\mu = \mathbb{E}X_1$, $\sigma^2 = \mathrm{Var}(X_1)$ and $\frac{x_n - n\mu}{\sqrt{n}} \approx y$ is of constant order for large n, by the central limit theorem,

$$
\begin{aligned}
&\mathbb{P}\left(X_1 + \ldots + X_n = x_n\right) \\
&= \mathbb{P}\left(x_n - \frac{1}{2} < X_1 + \ldots + X_n < x_n + \frac{1}{2}\right) \\
&= \mathbb{P}\left(\frac{x_n - n\mu}{\sqrt{n}} - \frac{1}{2\sqrt{n}} < \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}} < \frac{x_n - n\mu}{\sqrt{n}} + \frac{1}{2\sqrt{n}}\right) \\
&\approx \frac{1}{\sqrt{2\pi}\sigma} \int_{y - \frac{1}{2\sqrt{n}}}^{y + \frac{1}{2\sqrt{n}}} e^{-\frac{t^2}{2\sigma^2}} \mathrm{d}t \\
&\approx \frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}},
\end{aligned}
$$

obtaining the approximation for $\mathbb{P}\left(X_1 + \ldots + X_n = x_n\right)$ by the Gaussian density. To control the error in this approximation, we cannot simply use the Berry-Esseen theorem here because its error bound $O(\frac{1}{\sqrt{n}})$ is of the same order as the value of our approximation $\frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2}$. The local central limit theorem addresses this deficiency. We only discuss the discrete case. There are also versions which give approximations to densities of sums of i.i.d. continuous random variables.

We shall use the common notation $a + b\mathbb{Z}$ for the set $\{a + bx,\ x \in \mathbb{Z}\}$.

11.7 Theorem (Local central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. integer-valued random variables such that $\mathbb{E}X_1^2 < \infty$. Suppose X_i is not supported on any proper subprogression of $\mathbb{Z}$, that is there are no $r > 1$, $a \in \mathbb{R}$ such that $\mathbb{P}\left(X_i \in a + r\mathbb{Z}\right) = 1$. Denote $\mu = \mathbb{E}X_1$, $\sigma = \sqrt{\mathrm{Var}(X_1)}$ and*

$$p_n(x) = \mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}} = x\right), \qquad x \in \frac{\mathbb{Z} - n\mu}{\sqrt{n}}.$$

Then

$$\sup_{x \in \frac{\mathbb{Z} - n\mu}{\sqrt{n}}} \left| \sqrt{n} p_n(x) - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \right| \xrightarrow[n \to \infty]{} 0.$$

11.8 Lemma. *For an integer-valued random variable X and an integer k, we have*

$$\mathbb{P}\left(X = k\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi_X(t) \mathrm{d}t.$$

Proof. Note that for two integers k and l, we have

$$\mathbf{1}_{\{l = k\}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it(l-k)} \mathrm{d}t.$$

Thus

$$
\begin{aligned}
\mathbb{P}\left(X = k\right) = \mathbb{E}\,\mathbf{1}_{\{X = k\}} = \mathbb{E}\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it(X-k)} \mathrm{d}t &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \mathbb{E}e^{itX} \mathrm{d}t \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi_X(t) \mathrm{d}t.
\end{aligned}
$$

□

Proof of Theorem 11.7. Applying Lemma 11.8 to $X_1 + \ldots + X_n$ and changing the variables yields

$$\begin{aligned} p_n(x) &= \mathbb{P}\left(X_1 + \ldots + X_n = x\sqrt{n} + n\mu\right) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-it(x\sqrt{n}+n\mu)}\phi_{X_1+\ldots+X_n}(t)\mathrm{d}t \\ &= \frac{1}{\sqrt{n}}\frac{1}{2\pi}\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{-itx}\left[e^{-i\frac{t}{\sqrt{n}}\mu}\phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \mathrm{d}t. \end{aligned}$$

Using that the characteristic function of a centred Gaussian random variable with variance $1/\sigma^2$ is $e^{-\frac{x^2}{2\sigma^2}}$, we have

$$e^{-\frac{x^2}{2\sigma^2}} = \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-t^2\sigma^2/2}\mathrm{d}t,$$

which gives (by symmetry, we can write e^{-itx} instead of e^{itx})

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2\sigma^2/2}\mathrm{d}t.$$

Therefore,

$$\begin{aligned} \left|\sqrt{n}p_n(x) - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}\right| \le \frac{1}{2\pi}\int_{|t|\le\pi\sqrt{n}} &\left|\phi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2\sigma^2/2}\right|\mathrm{d}t \\ &+ \frac{1}{2\pi}\int_{|t|\ge\pi\sqrt{n}} e^{-t^2\sigma^2/2}\mathrm{d}t. \end{aligned}$$

Since the right hand side does not depend on x, we need to show that it converges to 0 as $n \to \infty$. The second integral clearly does. To deal with the first integral, we change the variables

$$\int_{|t|\le\pi\sqrt{n}}\left|\phi_{X_1-\mu}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2\sigma^2/2}\right|\mathrm{d}t = \frac{1}{\sigma}\int_{|t|\le\pi\sigma\sqrt{n}}\left|\phi_{\frac{X_1-\mu}{\sigma}}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t,$$

let $\bar{X}_1 = \frac{X_1-\mu}{\sigma}$ (which has mean 0 and variance 1) and break it into two pieces

$$\int_{|t|\le\varepsilon\sqrt{n}}\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t + \int_{\varepsilon\sqrt{n}\le|t|\le\pi\sigma\sqrt{n}}\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right|\mathrm{d}t. \quad (11.4)$$

Recall from the proof of the central limit theorem that

$$\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n \to e^{-t^2/2}$$

and by Taylor's formula,

$$\phi_{\bar{X}_1}(t) = 1 - \frac{t^2}{2} + t^2R(t),$$

for some (complex-valued) function R such that $R(t) \to 0$ as $t \to 0$. Choose $\varepsilon < 1$ such that $|R(t)| < \frac{1}{4}$ for all $|t| < \varepsilon$. Then for $|t| \le \varepsilon\sqrt{n}$,

$$\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)\right| \le \left|1 - \frac{t^2}{2n}\right| + \frac{t^2}{4n} = 1 - \frac{t^2}{4n} \le e^{-\frac{t^2}{4n}},$$

so

$$\left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right| \le e^{-t^2/4} + e^{-t^2/2}.$$

By Lebesgue's dominated convergence theorem, the first piece in (11.4) converges to 0 as $n \to \infty$. Finally, to handle the second piece, we claim that: $|\phi_{\bar{X}_1}(t)| < c_\varepsilon$ for all $\varepsilon \le |t| \le \pi\sigma$ for some constant $c_\varepsilon < 1$. This suffices because then

$$\int_{\varepsilon\sqrt{n}\le|t|\le\pi\sigma\sqrt{n}} \left|\phi_{\bar{X}_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-t^2/2}\right| \mathrm{d}t \le \int_{\varepsilon\sqrt{n}\le|t|\le\pi\sigma\sqrt{n}} (c_\varepsilon^n + e^{-t^2/2})\mathrm{d}t$$

and the right hand side clearly goes to 0 as $n \to \infty$. Now we use that X_1 is integer-valued, not concentrated on any proper subprogreesion to show the claim. Since X_1 is integer-valued, ϕ_{X_1} is 2π-periodic and in particular $\phi_{X_1}(2\pi) = 1$. Moreover, $|\phi_{X_1}(t)| < 1$ for all $0 < t < 2\pi$. Otherwise, if $|\phi_{X_1}(t_0)| = 1$ for some $0 < t_0 < 2\pi$, then $e^{it_0X_1}$ is constant, say equal to e^{ia}. Consequently, $X_1 \in \frac{a}{t_0} + \frac{2\pi}{t_0}\mathbb{Z}$, which contradicts the assumption. By periodicity and continuity, there is $c_\varepsilon < 1$ such that $|\phi_{X_1}(t)| < c_\varepsilon$ for all $\varepsilon < |t| \le \pi$. Since $\phi_{\bar{X}_1}(t) = e^{-i\frac{\mu}{\sigma}}\phi_{X_1}(\frac{t}{\sigma})$, the claim follows. □

Of course, in the proof it was not important that the X_i are integer-valued because by rescaling we could assume that they take values in $a + r\mathbb{Z}$ for some $a, r \in \mathbb{R}$. Such random variables are said to have a **lattice distribution**. We finish this section by summarising periodicity properties of their characteristic functions, which played a crucial role in the proof of the local central limit theorem.

11.9 Lemma. *For a random variable X with characteristic function ϕ_X the following are equivalent*

(i) $\phi_X(s) = 1$ for some $s \neq 0$,

(ii) $\mathbb{P}\left(X \in \frac{2\pi}{s}\mathbb{Z}\right) = 1$,

(iii) ϕ_X is $|s|$ periodic.

Proof. (i) ⇒ (ii): Since $1 = \phi_X(s) = \mathbb{E}\cos(sX) + i\mathbb{E}\sin(sX)$, we have $0 = \mathbb{E}(1-\cos(sX))$. Since $1 - \cos(sX)$ is a nonnegative random variable whose expectation is 0, we have $\mathbb{P}(\cos(sX) = 1) = 1$ (see Theorem E.2 (c)), equivalently $\mathbb{P}(sX \in 2\pi\mathbb{Z}) = 1$.

(ii) ⇒ (iii): We have

$$\phi_X(t + 2\pi|s|) = \mathbb{E}e^{i(t+|s|)X} = \sum_{k\in\mathbb{Z}} e^{i(t+|s|)\frac{2\pi}{|s|}k}\mathbb{P}\left(X = \frac{2\pi}{|s|}k\right) = \sum_{k\in\mathbb{Z}} e^{it\frac{2\pi}{|s|}k}\mathbb{P}\left(X = \frac{2\pi}{|s|}k\right)$$
$$= \phi_X(t).$$

(iii) ⇒ (i): Plainly, $\phi_X(s) = \phi_X(0) = 1$. □

11.10 Lemma. *Let X be a random variable with characteristic function ϕ_X. There are only 3 possibilities*

(i) $|\phi_X(t)| < 1$ *for every* $t \neq 0$,

(ii) $|\phi_X(s)| = 1$ *for some* $s > 0$ *and* $|\phi_X(t)| < 1$ *for all* $0 < t < s$ *and then* ϕ_X *is* s*-periodic and* $X \in a + \frac{2\pi}{s}\mathbb{Z}$ *a.s. for some* $a \in \mathbb{R}$,

(iii) $|\phi_X(t)| = 1$ *for every* $t \in \mathbb{R}$ *and then we have that* $\phi_X(t) = e^{ita}$ *for some* $a \in \mathbb{R}$, *that is* $X = a$ *a.s.*

If (ii) holds, X has a lattice distribution and since $|\phi_X(t)| < 1$ for all $0 < t < s$, by Lemma 11.9, s is the largest $r > 0$ such that $\mathbb{P}(X \in a + r\mathbb{Z}) = 1$. We sometimes call s the **span** of the distribution of X.

Proof. Let us first explain the implication in (ii). Suppose $|\phi_X(s)| = 1$ for some $s > 0$. Then $\phi_X(s) = e^{ia}$ for some $a \in \mathbb{R}$. Since $1 = e^{-ia}\phi_X(s) = \phi_{X-a}(s)$, by Lemma 11.9 applied to $X - a$, we get that $X - a \in \frac{2\pi}{s}\mathbb{Z}$ a.s. and ϕ_{X-a} is s-periodic, so $\phi_X = e^{ia}\phi_{X-a}$ is s-periodic.

To prove the trichotomy, suppose (i) and (ii) do not hold. Then there is a positive sequence $t_n \to 0$ such that $|\phi_X(t_n)| = 1$. Consequently, by what we just proved, there are $a_n \in \mathbb{R}$ such that $X \in a_n + \frac{2\pi}{t_n}\mathbb{Z}$ a.s. and ϕ_X is t_n-periodic. Without loss of generality, we can pick $a_n \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n}]$. Since $t_n \to 0$, we have $\mathbb{P}\left(X \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n})\right) \to 1$, which combined with $X \in a_n + \frac{2\pi}{t_n}\mathbb{Z}$ and $a_n \in (-\frac{\pi}{t_n}, \frac{\pi}{t_n}]$ gives $\mathbb{P}(X = a_n) \to 1$. Consequently, there is n_0 such that for all $n \geq n_0$, $\mathbb{P}(X = a_n) > 3/4$, but then all a_n, $n \geq n_0$ have to be equal, say $a_n = a$ and $\mathbb{P}(X = a_n) \to 1$ finally gives $\mathbb{P}(X = a) = 1$. Then $\phi_X(t) = e^{ita}$, consequently (iii) holds. $\square$

11.3 Poisson limit theorem

For probability measures μ and ν supported on $\mathbb{Z}$, the **total variation distance** between μ and ν is

$$\|\mu - \nu\|_{TV} = \sup_{A \subset \mathbb{Z}} |\mu(A) - \nu(A)|,$$

where the supremum is over all subsets A of $\mathbb{Z}$.

Our goal is to prove the following quantitative version of Theorem 10.17.

11.11 Theorem. *Let $\{X_{n,}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of Bernoulli random variables, $X_{n,k} \sim Ber(p_{n,k})$, such that for every $n \geq 1$, the variables $X_{n,1}, \dots, X_{n,n}$ are independent. Let $S_n = X_{n,1} + \dots + X_{n,n}$ and let Z_n be a Poisson random variable with parameter $\mathbb{E}S_n = \sum_{k=1}^n p_{n,k}$. Then*

$$\|\mu_n - \nu_n\|_{TV} \leq \sum_{k=1}^n p_{n,k}^2,$$

where μ_n is the law of S_n and ν_n is the law of Z_n.

11.12 Remark. This theorem quantitatively shows when and how close the sum of Bernoullis S_n is to a Poisson random variable with parameter $\lambda_n = \mathbb{E}S_n$. For instance, in the case of Theorem 10.16, when $p_{n,k} = p_n$, we see that the distribution of S_n is close to Poiss(λ_n), as long as $\sum_{k=1}^n p_{n,k}^2 = np_n^2 = \frac{\lambda_n^2}{n} \to 0$; in particular, if $\lambda_n \to \infty$ with $\lambda_n = o(\sqrt{n})$.

En route to proving Theorem 11.11, we need to develop a few facts related to the total variation distance.

There is a convenient explicit expression for this distance in terms of the ℓ_1 norm of the sequence of differences of atoms.

11.13 Theorem. *For two probability measures μ, ν on $\mathbb{Z}$, we have*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\sum_{x\in\mathbb{Z}} |\mu(\{x\}) - \nu(\{x\})|.$$

Proof. For a subset A of $\mathbb{Z}$, by the triangle inequality, we have

$$\begin{aligned}
2|\mu(A) - \nu(A)| &= |\mu(A) - \nu(A)| + |\mu(A^c) - \nu(A^c)| \\
&= \left|\sum_{x\in A} \Big(\mu(\{x\}) - \nu(\{x\})\Big)\right| + \left|\sum_{x\in A^c} \Big(\mu(\{x\}) - \nu(\{x\})\Big)\right| \\
&\leq \sum_{x\in\mathbb{Z}} |\mu(\{x\}) - \nu(\{x\})|
\end{aligned}$$

with equality for $A = \{x \in \mathbb{Z},\ \mu(\{x\}) \geq \nu(\{x\})\}$. □

As a corollary, we easily see that the total variation distance is a metric. Convergence in the total variation distance is equivalent to pointwise convergence on atoms (this can be seen as an analogy in this discrete setup to Scheffé's lemma from Exercise 8.8; the proofs are of course identical).

11.14 Theorem. *Let $(\mu_n)_n$ be a sequence of probability measures on $\mathbb{Z}$. For a probability measure μ on $\mathbb{Z}$, we have*

$$\|\mu_n - \mu\|_{TV} \xrightarrow[n\to\infty]{} 0 \qquad \textit{if and only if} \qquad \forall x \in \mathbb{Z}\ \mu_n(\{x\}) \xrightarrow[n\to\infty]{} \mu(\{x\}).$$

Proof. To ease the notation, let $p_{n,x} = \mu_n(\{x\})$ and $p_x = \mu(\{x\})$. In view of Theorem 11.13, we want to show that

$$\sum_{x\in\mathbb{Z}} |p_{n,x} - p_x| \xrightarrow[n\to\infty]{} 0 \qquad \text{if and only if} \qquad \forall x \in \mathbb{Z}\ \ p_{n,x} \xrightarrow[n\to\infty]{} p_x.$$

Implication "⇒" is clear. For the other one, note that

$$|p_x - p_{n,x}| = (p_x - p_{n,x})_+ + (p_x - p_{n,x})_-.$$

Since

$$0 = 1 - 1 = \sum_{x\in\mathbb{Z}}(p_x - p_{n,x}) = \sum_{x\in\mathbb{Z}}(p_x - p_{n,x})_+ - \sum_{x\in\mathbb{Z}}(p_x - p_{n,x})_-,$$

we get

$$\sum_{x\in\mathbb{Z}}|p_x - p_{n,x}| = 2\sum_{x\in\mathbb{Z}}(p_x - p_{n,x})_+.$$

Moreover, $(p_x - p_{n,x})_+ \le p_x$, so Lebesgue's dominated convergence theorem finishes the argument. □

This in turn easily leads to the conclusion that the weak convergence is equivalent to convergence in total variation distance.

11.15 Corollary. *Let $(\mu_n)_n$ be a sequence of probability measures on $\mathbb{Z}$. For a probability measure μ on $\mathbb{Z}$, we have*

$$\|\mu_n - \mu\|_{TV} \xrightarrow[n\to\infty]{} 0 \qquad \text{if and only if} \qquad \mu_n \xrightarrow[n\to\infty]{d} \mu.$$

Proof. "⇒": Fix a continuous bounded function $f\colon \mathbb{R}\to\mathbb{R}$ with $M = \sup|f|$. Then,

$$\begin{aligned}\left|\int f\mathrm{d}\mu_n - \int f\mathrm{d}\mu\right| = \left|\sum_{x\in\mathbb{Z}} f(x)(\mu(\{x\}) - \nu(\{x\}))\right| &\le M\sum_{x\in\mathbb{Z}}|\mu(\{x\}) - \nu(\{x\})| \\ &= 2M\|\mu_n - \mu\|_{TV}.\end{aligned}$$

"⇐": Fix $x\in\mathbb{Z}$. Let $f\colon \mathbb{R}\to\mathbb{R}$ be continuous bounded such that $f(t) = 0$ for all $t\in\mathbb{Z}\setminus\{x\}$ and $f(x) = 1$. Then

$$\left|\int f\mathrm{d}\mu_n - \int f\mathrm{d}\mu\right| = |\mu_n(\{x\}) - \mu(\{x\})|$$

and we conclude by Theorem 11.14. □

Of course, the definition of $\|\cdot\|_{TV}$ and the above results extend verbatim from $\mathbb{Z}$ to any countable set.

To prove a quantitative version of the Poisson limit theorem, we need three lemmas.

11.16 Lemma. *Let $\mu_1, \mu_2, \nu_1, \nu_2$ be probability measures on $\mathbb{Z}$. Then*

$$\|\mu_1\otimes\mu_2 - \nu_1\otimes\nu_2\|_{TV} \le \|\mu_1 - \nu_1\|_{TV} + \|\mu_2 - \nu_2\|_{TV}.$$

Proof. By Theorem 11.13 and the triangle inequality, we have

$$\begin{aligned}2\|\mu_1\otimes\mu_2 - \nu_1\otimes\nu_2\|_{TV} &= \sum_{x,y\in\mathbb{Z}}|\mu_1(\{x\})\mu_2(\{y\}) - \nu_1(\{x\})\nu_2(\{y\})| \\ &\le \sum_{x,y\in\mathbb{Z}}|\mu_1(\{x\})\mu_2(\{y\}) - \nu_1(\{x\})\mu_2(\{y\})| + \sum_{x,y\in Z}|\nu_1(\{x\})\mu_2(\{y\}) - \nu_1(\{x\})\nu_2(\{y\})| \\ &= \sum_{y\in\mathbb{Z}}\mu_2(\{y\})\sum_{x\in\mathbb{Z}}|\mu_1(\{x\}) - \nu_1(\{x\})| + \sum_{x\in\mathbb{Z}}\nu_1(\{x\})\sum_{y\in\mathbb{Z}}|\mu_2(\{y\}) - \nu_2(\{y\})| \\ &= 2\|\mu_1 - \nu_1\|_{TV} + 2\|\mu_2 - \nu_2\|_{TV}.\end{aligned}$$

□

Recall that the convolution of two measures μ and ν on $\mathbb{Z}$ is defined as a measure

$$(\mu * \nu)(A) = \sum_{y\in\mathbb{Z}} \mu(\{A-y\})\nu(\{y\}), \qquad A \subset \mathbb{Z}.$$

It is the distribution of the sum $X+Y$ of two independent $\mathbb{Z}$-valued random variables X, Y with law μ, ν respectively.

11.17 Lemma. *Let $\mu_1, \mu_2, \nu_1, \nu_2$ be probability measures on $\mathbb{Z}$. Then*

$$\|\mu_1 * \mu_2 - \nu_1 * \nu_2\|_{TV} \leq \|\mu_1 - \nu_1\|_{TV} + \|\mu_2 - \nu_2\|_{TV},$$

Proof. By Theorem 11.13 and the triangle inequality, we have

$$\begin{aligned}
2\|\mu_1 * \mu_2 - \nu_1 * \nu_2\|_{TV} &= \sum_{x\in\mathbb{Z}} |(\mu_1 * \mu_2)(\{x\}) - (\nu_1 * \nu_2)(\{x\})| \\
&= \sum_{x\in\mathbb{Z}} \left| \sum_{y\in\mathbb{Z}} \mu_1(\{x-y\})\mu_2(\{y\}) - \sum_{y\in\mathbb{Z}} \nu_1(\{x-y\})\nu_2(\{y\}) \right| \\
&\leq \sum_{x\in\mathbb{Z}} \sum_{y\in\mathbb{Z}} |\mu_1(\{x-y\})\mu_2(\{y\}) - \nu_1(\{x-y\})\nu_2(\{y\})| \\
&= \sum_{x\in\mathbb{Z}} \sum_{y\in\mathbb{Z}} |\mu_1(\{x\})\mu_2(\{y\}) - \nu_1(\{x\})\nu_2(\{y\})| \\
&= 2\|\mu_1 - \nu_1\|_{TV} + \|\mu_2 - \nu_2\|_{TV}.
\end{aligned}$$

□

11.18 Lemma. *Let $p \in [0,1]$. Let μ be the Bernoulli distribution with parameter p and let ν be the Poisson distribution with parameter p. Then*

$$\|\mu - \nu\|_{TV} \leq p^2.$$

Proof. By Theorem 11.13, we have

$$\begin{aligned}
2\|\mu - \nu\|_{TV} &= |\mu(\{0\}) - \nu(\{0\})| + |\mu(\{1\}) - \nu(\{1\})| + \sum_{k\geq 2} \nu(\{k\}) \\
&= |1 - p - e^{-p}| + |p - pe^{-p}| + 1 - e^{-p}(1+p) \\
&= e^{-p} - 1 + p + p - pe^{-p} + 1 - e^{-p}(1+p) = 2p(1-e^{-p}) \leq 2p^2
\end{aligned}$$

(we use $e^{-p} \geq 1-p$). □

Proof of Theorem 11.11. Let μ_n be the distribution of $S_n = X_{n,1} + \cdots + X_{n,n}$ and let $\mu_{n,k}$ be the distribution of $X_{n,k}$ which is Bernoulli with parameter $p_{n,k}$. Then

$$\mu_n = \mu_{n,1} * \cdots * \mu_{n,n}.$$

Let ν_n be the Poisson distribution with parameter $\lambda_n = \mathbb{E}S_n = \sum_{k=1}^n p_{n,k}$ and let $\nu_{n,k}$ be the Poisson distribution with parameter $p_{n,k}$. Since sums of independent Poisson random variables are Poisson, we have

$$\nu_n = \nu_{n,1} * \cdots * \nu_{n,n}.$$

Thus, by Lemmas 11.17, 11.16 and 11.18,

$$\|\mu_n - \nu_n\|_{TV} \leq \sum_{k=1}^n p_{n,k}^2. \tag{11.5}$$

□

11.19 Remark. We can quickly deduce Theorem 10.17 from (11.5). Let ν be the Poisson distribution with parameter λ. As we saw in the proof of Theorem 10.17, the right hand side of (11.5) goes to 0 as $n \to \infty$. Moreover, by Theorem 11.14, $\|\nu_n - \nu\|_{TV} \to 0$ because $\lambda_n \to \lambda$. Thus, $\|\mu_n - \nu\|_{TV} \to 0$, as desired.

11.20 Remark. When $p_{n,k} = \frac{1}{n}$, we have $\lambda_n = \lambda = 1$ and from the second proof,

$$\|\mu_n - \nu\|_{TV} \leq \frac{1}{n}.$$

On the other hand, since $1 - x \geq e^{-x-x^2}$ and $1 - e^{-x} \geq \frac{x}{2}$ for $x \in [0, \frac{1}{2}]$, we have for $n \geq 2$,

$$|\mu_n(\{0\}) - \nu(\{0\})| = e^{-1} - \left(1 - \frac{1}{n}\right)^n \geq e^{-1} - e^{-1-1/n} = e^{-1}\left(1 - e^{-1/n}\right) \geq \frac{1}{2en},$$

which shows that the rate $\frac{1}{n}$ is optimal here.

11.4 Exercises

1. Using the Berry-Esseen theorem, how can you bound the error you make in your approximation in Exercise 10.1?

2. In Exercise 10.2, using the Berry-Esseen theorem, show additionally that $|a_n - a| \leq \frac{15}{\sqrt{n}}$ for all $n \geq 1$.

3. For two probability measures on $\mathbb{Z}$, $\|\mu - \nu\|_{TV} \leq \delta$ if and only if there are random variables X, Y with distributions μ, ν, respectively such that $\mathbb{P}\left(X \neq Y\right) \leq \frac{\delta}{2}$.

12 Conditional expectation

We begin with a motivating example. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose we have two discrete random variables

$$X \colon \Omega \to \{x_1, \ldots, x_m\},$$
$$Z \colon \Omega \to \{z_1, \ldots, z_n\},$$

taking finitely many values. Recall the conditional probabilities

$$\mathbb{P}\left(X = x_i \mid Z = z_j\right) = \frac{\mathbb{P}\left(X = x_i,\ Z = z_j\right)}{\mathbb{P}\left(Z = z_j\right)}$$

and conditional expectations

$$y_j = \mathbb{E}(X|Z = z_j) = \sum x_i \mathbb{P}\left(X = x_i | Z = z_j\right).$$

In this simple situation, the conditional expectation of X given Z is a random variable $Y = \mathbb{E}(X|Z)$ defined as

$$Y(\omega) = y_j = \mathbb{E}(X|Z = z_j) \qquad \text{on } \{Z = z_j\}.$$

In other words, "knowing" the value of Z amounts to the partitioning $\Omega = \bigcup \{Z = z_j\}$ and Y is set to be constant y_j on the *atoms* $\{Z = z_j\}$ of this partitioning. We point out two features of Y which will be central in the general definition. Let $\mathcal{G} = \sigma(Z)$ be the σ-algebra generated by Z. Here $\mathcal{G}$ is generated by the partitioning $\{Z = z_j\}_j$. Since Y is constant on the atoms of $\mathcal{G}$,

(1) Y is $\mathcal{G}$-measurable.

Note also that

$$\begin{aligned} \mathbb{E} Y \mathbf{1}_{\{Z = z_j\}} = y_j \mathbb{P}\left(Z = z_j\right) &= \sum x_i \mathbb{P}\left(X = x_i | Z = z_j\right) \mathbb{P}\left(Z = z_j\right) \\ &= \sum x_i \mathbb{P}\left(X = x_i,\ Z = z_j\right) = \mathbb{E} X \mathbf{1}_{\{Z = z_j\}}. \end{aligned}$$

Since every element in $\mathcal{G}$ is a union of the atoms $\{Z = z_j\}$, by linearity, we thus have

(2) $\forall G \in \mathcal{G} \quad \mathbb{E} Y \mathbf{1}_G = \mathbb{E} X \mathbf{1}_G$.

12.1 Construction

The following theorem due to Kolmogorov (1933) gives a way to define conditional expectations.

12.1 Theorem. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}|X| < \infty$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. There is a random variable Y with $\mathbb{E}|Y| < \infty$ such that*

(1) Y is $\mathcal{G}$-measurable,

(2) $\forall G \in \mathcal{G} \quad \mathbb{E}Y\,\mathbf{1}_G = \mathbb{E}X\,\mathbf{1}_G$.

Moreover, if $\tilde{Y}$ is another random variable with $\mathbb{E}|\tilde{Y}| < \infty$ satisfying (1) and (2), then $\tilde{Y} = Y$ a.s.

This random variable Y is called (a version) of the **conditional expectation of X given** $\mathcal{G}$, denoted $\mathbb{E}(X|\mathcal{G})$. We then write $Y = \mathbb{E}(X|\mathcal{G})$. For a random variable Z, the conditional expectation of X given Z, $\mathbb{E}(X|Z)$, is defined as $\mathbb{E}(X|\sigma(Z))$. More generally, we define $\mathbb{E}(X|Z_1, Z_2, \ldots) = \mathbb{E}(X|\sigma(Z_1, Z_2, \ldots))$. Of course, we also adopt the natural definition of the **conditional probability** of an event A given $\mathcal{G}$, $\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(\mathbf{1}_A\,|\mathcal{G})$ (which is a random variable!).

The intuitive meaning of conditional expectations:

a) $\mathbb{E}(X|A)$, that is given an even A: an experiment has been performed and all we know is whether $\omega \in A$ or not and we recalculate the expectation according to $\mathbb{P}(\cdot|A)$

b) $\mathbb{E}(X|Z)$, that is given a discrete random variable Z: an experiment has been performed and all we know about ω is in which set $\{Z = z_j\}$ it is, so $\mathbb{E}(X|Z)(\omega)$ is still a random quantity, but constant on these sets

c) $\mathbb{E}(X|\mathcal{G})$, that is given a sub-$\sigma$-algebra $\mathcal{G}$: an experiment has been performed and all we know about ω is $\{Z(\omega) : Z \text{ is } \mathcal{G}\text{-measurable}\}$, so $\mathbb{E}(X|\mathcal{G})(\omega)$ is a random quantity being the average of X given this information.

The richer $\mathcal{G}$ is, the more we "know" about X, so that $\mathbb{E}(X|\mathcal{G})$ more accurately describes X. In the two extreme cases:

a) if $\mathcal{G} = \{\varnothing, \Omega\}$ (a trivial σ-algebra, so no knowledge), then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$ (constant)

b) if $\mathcal{G} = \mathcal{F}$ (full knowledge), then $\mathbb{E}(X|\mathcal{G}) = X$.

These can be simply verified by checking that the claimed variables $\mathbb{E}X$ and X respectively satisfy (1) and (2).

Proof of Theorem 12.1. We break the proof into several steps.

Step 1: uniqueness. Let Y, $\tilde{Y}$ be two integrable $\mathcal{G}$-measurable versions of $\mathbb{E}(X|\mathcal{G})$, that is satisfying (1) and (2). Since $\{Y - \tilde{Y} > \frac{1}{n}\} \in \mathcal{G}$, we get from (2),

$$\mathbb{E}Y\,\mathbf{1}_{\{Y-\tilde{Y}>\frac{1}{n}\}} = \mathbb{E}X\,\mathbf{1}_{\{Y-\tilde{Y}>\frac{1}{n}\}} = \mathbb{E}\tilde{Y}\,\mathbf{1}_{\{Y-\tilde{Y}>\frac{1}{n}\}},$$

or,

$$0 = \mathbb{E}(Y - \tilde{Y})\,\mathbf{1}_{\{Y-\tilde{Y}>\frac{1}{n}\}} \geq \frac{1}{n}\mathbb{P}\left(Y - \tilde{Y} > \frac{1}{n}\right),$$

which shows that $\mathbb{P}\left(Y - \tilde{Y} > \frac{1}{n}\right) = 0$ and consequently, taking the union of these events over n, $\mathbb{P}\left(Y - \tilde{Y} > 0\right) = 0$. Swapping the roles of Y and $\tilde{Y}$, we also get that $\mathbb{P}\left(\tilde{Y} - Y > 0\right) = 0$, so $\mathbb{P}\left(Y \neq \tilde{Y}\right) = 0$.

Step 2: existence for $X \in L_2$. Let $X \in L_2(\Omega, \mathcal{F}, \mathbb{P})$ and consider the subspace

$$\mathcal{H} = L_2(\Omega, \mathcal{G}, \mathbb{P}) \subset L_2(\Omega, \mathcal{F}, \mathbb{P})$$

which is complete (as being an L_2 space – see Theorem 6.10), so by the existence of the orthogonal projection, Theorem 6.11, there is a random variable $Y \in \mathcal{H}$ which is closest to X, that is it satisfies

$$\mathbb{E}(X - Y)^2 = \inf\{\mathbb{E}(X - W)^2,\ W \in \mathcal{H}\},$$

or, equivalently,

$$X - Y \perp \mathcal{H}, \text{ that is } \forall W \in \mathcal{G}\ \mathbb{E}(X - Y)W = 0.$$

We claim that this Y satisfies (1) and (2). Since $Y \in \mathcal{H}$, it is $\mathcal{G}$ measurable. For $G \in \mathcal{G}$, letting $W = \mathbf{1}_G \in \mathcal{H}$, we get $\mathbb{E}(X - Y)W = 0$, that is (2).

Step 3: basic and key properties of $\mathbb{E}(X|\mathcal{G})$ for $X \in L_2$.

(i) linearity: $\mathbb{E}(a_1X_1 + a_2X_2|\mathcal{G}) = a_1\mathbb{E}(X_1|\mathcal{G}) + a_2\mathbb{E}(X_2|\mathcal{G})$, $a_1, a_2 \in \mathbb{R}$, $X_1, X_2 \in L_2$

(ii) monotonicity: if $X \in L_2$, $X \geq 0$ a.s., then $\mathbb{E}(X|\mathcal{G}) \geq 0$ a.s.

(iii) if $X_1, X_2 \in L_2$, $X_1 \geq X_2$ a.s., then $\mathbb{E}(X_1|\mathcal{G}) \geq \mathbb{E}(X_2|\mathcal{G})$ a.s.

Property (i) is clear because the orthogonal projection is a linear map. Property (ii) follows by an argument identical to the one from Step 1. Property (iii) follows from (i) and (ii).

Step 3: existence of $\mathbb{E}(X|\mathcal{G})$ for $X \in L_1$, $X \geq 0$. Let $X_n = \min\{X, n\}$. Then each X_n is bounded, in particular $X_n \in L_2$ and $X_n \nearrow X$ as $n \to \infty$. Let Y_n be a version of $\mathbb{E}(X_n|\mathcal{G})$ (constructed in Step 2). By Property (iii), Y_n is a monotone sequence. Let $Y = \lim Y_n$. As a limit of $\mathcal{G}$-measurable functions, Y is $\mathcal{G}$-measurable. By Lebesgue's monotone convergence theorem, for $G \in \mathcal{G}$,

$$\mathbb{E}Y\,\mathbf{1}_G = \mathbb{E}(\lim Y_n)\,\mathbf{1}_G = \lim \mathbb{E}Y_n\,\mathbf{1}_G$$

but since Y_n is a version of $\mathbb{E}(X_n|\mathcal{G})$, by (2), we have $\mathbb{E}Y_n\,\mathbf{1}_G = \mathbb{E}X_n\,\mathbf{1}_G$. Thus,

$$\lim \mathbb{E}Y_n\,\mathbf{1}_G = \lim \mathbb{E}X_n\,\mathbf{1}_G = \mathbb{E}(\lim X_n)\,\mathbf{1}_G = \mathbb{E}X\,\mathbf{1}_G,$$

so $\mathbb{E}Y\,\mathbf{1}_G = \mathbb{E}X\,\mathbf{1}_G$. In particular, $\mathbb{E}Y = \mathbb{E}X < \infty$. These show that Y has the desired properties.

Step 4: existence of $\mathbb{E}(X|\mathcal{G})$ for arbitrary $X \in L_1$. We decompose $X = X^+ - X^-$ and set $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X^+|\mathcal{G}) - \mathbb{E}(X^-|\mathcal{G})$. □

12.2 Important properties

12.2 Theorem. *All random variables are assumed to be integrable, $\mathcal{G}$ is sub-σ-algebra.*

(a) If $Y = \mathbb{E}(X|\mathcal{G})$, then $\mathbb{E}Y = \mathbb{E}X$.

(b) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$.

(c) Linearity: $\mathbb{E}(a_1X_1+a_2X_2|\mathcal{G}) = a_1\mathbb{E}(X_1|\mathcal{G})+a_2\mathbb{E}(X_2|\mathcal{G})$ a.s., $a_1, a_2 \in \mathbb{R}$ (understood as: if Y_i is a version of $\mathbb{E}(X_i|\mathcal{G})$, then $a_1Y_1+a_2Y_2$ is a version of $\mathbb{E}(a_1X_1+a_2X_2|\mathcal{G})$).

(d) Positivity: if $X \geq 0$ a.s., then $\mathbb{E}(X|\mathcal{G}) \geq 0$ a.s. If $X_1 \geq X_2$, then $\mathbb{E}(X_1|\mathcal{G}) \geq \mathbb{E}(X_2|\mathcal{G})$ a.s.

(e) Lebesgue's monotone convergence theorem: if $0 \leq X_n \nearrow X$ a.s., then

$$\mathbb{E}(X_n|\mathcal{G}) \nearrow \mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

(f) Fatou's lemma: if $X_n \geq 0$, then

$$\mathbb{E}(\liminf X_n|\mathcal{G}) \leq \liminf \mathbb{E}(X_n|\mathcal{G}) \quad \text{a.s.}$$

(g) Lebesgue's dominated convergence theorem: if $\forall n\ |X_n| \leq V$ for some $V \in L_1$ and $X_n \to X$, then

$$\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

(h) Jensen's inequality: if $f\colon \mathbb{R} \to \mathbb{R}$ is convex and $f(X) \in L_1$, then

$$\mathbb{E}(f(X)|\mathcal{G}) \geq f(\mathbb{E}(X|\mathcal{G})) \text{ a.s.}$$

(i) Tower property: if $\mathcal{H} \subset \mathcal{G}$ is a sub-σ-algebra, then

$$\mathbb{E}\big(\mathbb{E}(X|\mathcal{G})|\mathcal{H}\big) = \mathbb{E}(X|\mathcal{H}) = \mathbb{E}\big(\mathbb{E}(X|\mathcal{H})|\mathcal{G}\big) \text{ a.s.}$$

(j) "Taking out what is known": if Z is a $\mathcal{G}$-measurable bounded random variable, then

$$\mathbb{E}(ZX|\mathcal{G}) = Z\mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

(k) "Role of independence": if $\mathcal{H}$ is a σ-algebra independent of $\sigma(X, \mathcal{G})$, then

$$\mathbb{E}(X|\sigma(\mathcal{G}, \mathcal{H})) = \mathbb{E}(X|\mathcal{G}) \text{ a.s.}$$

In particular, if X is independent of $\mathcal{H}$, for $\mathcal{G} = \{\varnothing, \Omega\}$, we get

$$\mathbb{E}(X|\mathcal{H}) = \mathbb{E}X \text{ a.s.}$$

Proof. (a) We use (2) with $G = \Omega$ to get $\mathbb{E}X = \mathbb{E}X\,\mathbf{1}_\Omega = \mathbb{E}Y\,\mathbf{1}_\Omega = \mathbb{E}Y$.

(b) Obvious because $Y = X$ clearly satisfies (1) and (2).

(c) It follows from the linearity of $\mathbb{E}(\cdot)\,\mathbf{1}_G$ for every fixed $G \in \mathcal{G}$.

(d) It is done in a similar way as in Step 1 of the proof of Theorem 12.1.

(e) It is done is a similar way as in Step 3 of the proof of Theorem 12.1. We let $Y_n = \mathbb{E}(Y_n|\mathcal{G})$, use monotonicity to put $Y = \lim Y_n$ and argue through the usual Lebesgue's monotone convergence theorem.

(f) Use (e) to deduce it as in the unconditional case.

(g) Use (f) to deduce it as in the unconditional case (see Appendix E).

(h) It is done as in the unconditional case in the proof of Theorem 6.2. We write $f(x) = \sup_{\ell \in \mathcal{L}} \ell(x)$ for a countable family of linear functions $\mathcal{L}$. Fix $\ell \in \mathcal{L}$. Since $f(X) \geq \ell(X)$, by monotonicity and linearity, $\mathbb{E}(f(X)|\mathcal{G}) \geq \mathbb{E}(\ell(X)|\mathcal{G}) = \ell(\mathbb{E}(X|\mathcal{G}))$ a.s. Since $\mathcal{L}$ is countable, this also holds for all $\ell \in \mathcal{L}$ a.s. Taking the supremum over ℓ finishes the argument.

(i) Let $Y = \mathbb{E}(X|\mathcal{H})$. It is $\mathcal{H}$-measurable, so also $\mathcal{G}$-measurable and by (b),

$$\mathbb{E}\big(\mathbb{E}(X|\mathcal{H})|\mathcal{G}\big) = \mathbb{E}(Y|\mathcal{G}) = Y.$$

To check that

$$\mathbb{E}\big(\mathbb{E}(X|\mathcal{G})|\mathcal{H}\big) = Y,$$

it suffices to show that for every $H \in \mathcal{H}$, $\mathbb{E}\mathbb{E}(X|\mathcal{G})\,\mathbf{1}_H = \mathbb{E}Y\,\mathbf{1}_H$. Since $H \in \mathcal{G}$, by the definition of $\mathbb{E}(X|\mathcal{G})$, the left hand side is $\mathbb{E}X\,\mathbf{1}_H$. Since $H \in \mathcal{H}$, by the definition of $\mathbb{E}(X|\mathcal{H})$, the right hand side is also $\mathbb{E}X\,\mathbf{1}_H$.

(j) Thanks to linearity, without loss of generality we can assume that $X \geq 0$. Then the standard argument of complicating Z works because we can use monotone convergence (first we check the claim for $Z = \mathbf{1}_G$, $G \in \mathcal{G}$, then by linearity we have it for simple Z and then for all $\mathcal{G}$-measurable bounded Z).

(k) Thanks to linearity, without loss of generality we can assume that $X \geq 0$. Let $Y = \mathbb{E}(X|\mathcal{G})$. We want to show that for every $A \in \sigma(\mathcal{G}, \mathcal{H})$, $\mathbb{E}Y\,\mathbf{1}_A = \mathbb{E}X\,\mathbf{1}_A$. By Dynkin's theorem, it suffices to this for all A in a π-system generating $\sigma(\mathcal{G}, \mathcal{H})$, so for every $A = G \cap H$ with $G \in \mathcal{G}$ and $H \in \mathcal{H}$. Then, we have

$$\mathbb{E}Y\,\mathbf{1}_{G\cap H} = \mathbb{E}(Y\,\mathbf{1}_G)\,\mathbf{1}_H = \mathbb{E}Y\,\mathbf{1}_G\,\mathbb{E}\,\mathbf{1}_H$$

because $Y\,\mathbf{1}_G$ is $\mathcal{G}$-measurable, hence independent of $\mathcal{H}$. By the definition of $\mathbb{E}(X|\mathcal{G})$, $\mathbb{E}Y\,\mathbf{1}_G = \mathbb{E}X\,\mathbf{1}_G$. Since $X\,\mathbf{1}_G$ and $\mathbf{1}_H$ are independent, we conclude that

$$\mathbb{E}Y\,\mathbf{1}_{G\cap H} = \mathbb{E}X\,\mathbf{1}_G\,\mathbb{E}\,\mathbf{1}_H = \mathbb{E}X\,\mathbf{1}_{G\cap H}\,.$$

□

12.3 Basic examples

12.3 Example. Let (X, Z) be a continuous random vector in $\mathbb{R}^2$ with density f. Then

$$f_X(x) = \int_{\mathbb{R}} f(x,z)\mathrm{d}z \qquad \text{is the density of } X,$$
$$f_Z(z) = \int_{\mathbb{R}} f(x,z)\mathrm{d}x \qquad \text{is the density of } Z.$$

Recall we define the conditional density $f_{X|Z}$ of X given Z as

$$f_{X|Z}(x|z) = \begin{cases} \frac{f(x,z)}{f_Z(z)}, & \text{if } f_Z(z) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Let $h\colon \mathbb{R} \to \mathbb{R}$ be a Borel function such that $\mathbb{E}|h(X)| < \infty$. Let

$$g(z) = \int_{\mathbb{R}} h(x) f_{X|Z}(x|z)\mathrm{d}x.$$

Claim. $Y = g(Z)$ is a version of $\mathbb{E}(h(X)|Z)$.

Proof. Clearly Y is Z-measurable (i.e. $\sigma(Z)$-measurable), so it suffices to check that for every $A \in \sigma(Z)$, $\mathbb{E}g(Z)\,\mathbf{1}_A = \mathbb{E}h(X)\,\mathbf{1}_A$. We have $A = \{Z \in B\}$ for some Borel set B. Then, by the definition of g,

$$\begin{aligned}\mathbb{E}g(Z)\,\mathbf{1}_A = \mathbb{E}g(Z)\,\mathbf{1}_{Z\in B} = \int g(z)\,\mathbf{1}_B(z) f_Z(z)\mathrm{d}z &= \iint h(x)\,\mathbf{1}_B(z) f(x,z)\mathrm{d}x\mathrm{d}z \\ &= \mathbb{E}h(X)\,\mathbf{1}_{Z\in B}.\end{aligned}$$

□

12.4 Example. Let $X_1, \dots, X_n$ be independent random variables. Let $h\colon \mathbb{R}^n \to \mathbb{R}$ be a bounded Borel function. Then

$$\mathbb{E}(h(X_1, \dots, X_n)|X_1) = g(X_1),$$

where $g(x) = \mathbb{E}h(x, X_2, \dots, X_n)$, $x \in \mathbb{R}$. Clearly $g(X_1)$ is X_1-measurable. That condition (2) holds, follows from Fubini's theorem.

12.5 Example. Let $X_1, \dots, X_n$ be i.i.d. integrable random variables and let $S_n = X_1 + \dots + X_n$. Let $\mathcal{G}_n = \sigma(S_n, X_{n+1}, X_{n+2}, \dots)$. By the "Role of independence" property (point (k) of Theorem 12.2), we have

$$\mathbb{E}(X_1|\mathcal{G}_n) = \mathbb{E}(X_1|S_n).$$

To find the latter, we use symmetry, which gives

$$Y = \mathbb{E}(X_1|S_n) = \mathbb{E}(X_2|S_n) = \ldots = \mathbb{E}(X_n|S_n).$$

By linearity, $nY = \mathbb{E}(S_n|S_n) = S_n$, thus $Y = \frac{S_n}{n}$, that is

$$\mathbb{E}(X_1|S_n) = \frac{S_n}{n}.$$

We finish by remarking that in all of these examples $\mathbb{E}(X|Z)$ is of the form $g(Z)$ for some Borel function g. This holds in general and immediately follows from the following lemma.

12.6 Lemma. *If Z is a random vector in $\mathbb{R}^n$ and X is a Z-measurable random variable, then $X = g(Z)$ for some Borel function $g\colon \mathbb{R}^n \to \mathbb{R}$.*

A proof by a standard complication of X is left as an exercise. Thus we immediately get the following general observation about conditional expectations.

12.7 Theorem. *Let Z be a random vector in $\mathbb{R}^n$ and let X be an integrable random variable. Then there is a Borel function $g\colon \mathbb{R}^n \to \mathbb{R}$ such that $g(Z)$ is a version of $\mathbb{E}(X|Z)$.*

This in particular gives a way around defining conditioning on events of probability 0. We set $\mathbb{E}(X|Z = z) = g(z)$, $z \in \mathbb{R}^n$, where g is the function provided by Theorem 12.7.

12.4 Exercises

1. We toss a fair coin 10 times. Let X be the number of heads altogether and Y in the first 4 tosses. Find $\mathbb{E}(X|Y)$ and $\mathbb{E}(Y|X)$.

2. Give an example of random variables X and Y which are *not* independent, but $\mathbb{E}(X|Y) = \mathbb{E}X$.

3. Let (X,Y) be a centred Gaussian random vector in $\mathbb{R}^2$. Show that $\mathbb{E}(X|Y) = \frac{\mathbb{E}XY}{\mathbb{E}Y^2}Y$.

4. Let $\rho \in (-1,1)$ and let (U,V) be a random vector in $\mathbb{R}^2$ with density
$$f(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right\}, \qquad (u,v) \in \mathbb{R}^2.$$
Find $\mathbb{E}(U|V)$.

5. Let $X_1, \ldots, X_n$ be i.i.d. random variables uniform on $[0,1]$. Find the conditional expectation $\mathbb{E}(X_1|\max\{X_1,\ldots,X_n\})$.

6. Let X be a nonnegative integrable random variable and let $\mathcal{G}$ be a sub-σ-algebra. Show that

   a) $\mathbb{E}(X|\mathcal{G}) = \int_0^\infty \mathbb{P}(X > t|\mathcal{G})\,\mathrm{d}t$,

   b) $\mathbb{P}(X > t|\mathcal{G}) \leq t^{-p}\mathbb{E}(X^p|\mathcal{G})$, $p, t > 0$, provided that $X \in L_p$.

7. Let X and Y be independent random variables uniform on $[-1,1]$. Find $\mathbb{E}(X|X^2+Y^2)$ and $\mathbb{E}(X^2|X+Y)$.

8. Suppose X, Y are integrable random variables such that $\mathbb{E}(X|Y) = Y$ a.s. and $\mathbb{E}(Y|X) = X$ a.s. Then $X = Y$ a.s.

9. Prove Lemma 12.6

13 Martingales I

13.1 Definitions and basic examples

By a **process** $X = (X_n)_{n\geq 0} = (X_0, X_1, X_2, \ldots)$ we just mean a sequence of random variables $X_0, X_1, \ldots$ (index n is thought of as time which is discrete here). A **filtration** $\{\mathcal{F}_n\}_{n\geq 0}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a nondecreasing sequence of sub-σ-algebras of $\mathcal{F}$,

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}.$$

We set

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{n\geq 0} \mathcal{F}_n\right)$$

which is also a sub-σ-algebra of $\mathcal{F}$. Sometimes $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n\geq 0}, \mathbb{P})$ is then referred to as a filtered probability space.

Intuitively, $\mathcal{F}_n$ carries information available at time n.

Given a process $X = (X_n)_{n\geq 0}$, its **natural filtration** is given by

$$\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n), \qquad n \geq 0.$$

The process X is called **adapted** (to $\{\mathcal{F}_n\}$) if for every $n \geq 0$, X_n is $\mathcal{F}_n$-measurable. It is called **predictable** or **previsible** if for every $n \geq 1$, X_n is $\mathcal{F}_{n-1}$-measurable and X_0 is constant. Intuitively, if X is adapted, then at time n we know $X_n(\omega)$ (and $X_0(\omega), \ldots, X_{n-1}(\omega)$); if it is predictable, at time n we additionally know $X_{n+1}(\omega)$.

A process $X = (X_n)_{n\geq 0}$ is a **martingale** (with respect to a filtration $\{\mathcal{F}\}_{n\geq 0}$) if

(i) X is adapted,

(ii) $\mathbb{E}|X_n| < \infty$, for every $n \geq 0$,

(iii) $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$ a.s., for every $n \geq 1$.

It is called a **supermartingale** (respectively **submartingale**) if (iii) is replaced by: $\mathbb{E}(X_n|\mathcal{F}_{n-1}) \leq X_{n-1}$ (resp. $\geq X_{n-1}$) a.s., for every $n \geq 1$.

Submartingales correspond to subharmonic functions (f on $\mathbb{R}^n$ is subharmonic if and only if $f(B)$ is a local submartingale, where B is a standard Brownian motion on $\mathbb{R}^n$).

13.1 Remark. If X is a supermartingale, then

$$\mathbb{E}X_n = \mathbb{E}(\mathbb{E}(X_n|\mathcal{F}_{n-1})) \leq \mathbb{E}X_{n-1},$$

so its averages do not increase over time. For martingales, the averages are of course constant.

The word ‘martingale’ origins from the Spanish word ‘almátaga’ = fastening. In *everyday* use, a martingale refers to a horse gear item which is a strap attached to the reins of horse used to prevent the horse from raising its head too high (P. Halmos allegedly sent J. Doob such a martingale as a “gift”; J. Hammersley, after his lecture in 1965 was made aware of this meaning of the word martingale, thought that the mathematical term originated from its equestrian meaning and started calling his martingale process, “harness process”). As a mathematical term, it was first used in French by J. Ville in his thesis in 1939, where initially he writes “game system or martingale” and then just continues with “martingale”. It was J. Doob, the godfather of martingale theory, who after reviewing Ville’s thesis, coined and used the term ‘martingale’.

13.2 Example. Let $X_1, X_2, \ldots$ be independent integrable random variables with mean 0, $\mathbb{E}X_k = 0$ for every $k \geq 1$. Let

$$\begin{aligned} &S_0 = 0, \\ &S_n = X_1 + \cdots + X_n, \qquad n \geq 1 \end{aligned}$$

and

$$\begin{aligned} &\mathcal{F}_0 = \{\varnothing, \Omega\}, \\ &\mathcal{F}_n = \sigma(X_1, \ldots, X_n), \qquad n \geq 1. \end{aligned}$$

Then $(S_n)_{n\geq 0}$ is a martingale with respect to the natural filtration $\{\mathcal{F}_n\}$. Indeed, because $S_n = S_{n-1} + X_n$ and S_{n-1} is $\mathcal{F}_{n-1}$ measurable, whereas X_n is independent of $\mathcal{F}_{n-1}$, we have

$$\mathbb{E}(S_n|\mathcal{F}_{n-1}) = \mathbb{E}(S_{n-1}|\mathcal{F}_{n-1}) + \mathbb{E}(X_n|\mathcal{F}_{n-1}) = S_{n-1} + \mathbb{E}X_n = S_{n-1}.$$

13.3 Example. Let $X_1, X_2, \ldots$ be independent nonnegative integrable random variables with $\mathbb{E}X_k = 1$ for every $k \geq 1$. Let

$$\begin{aligned} &M_0 = 1, \\ &M_n = X_1 \cdot \ldots \cdot X_n, \qquad n \geq 1 \end{aligned}$$

and

$$\begin{aligned} &\mathcal{F}_0 = \{\varnothing, \Omega\}, \\ &\mathcal{F}_n = \sigma(X_1, \ldots, X_n), \qquad n \geq 1. \end{aligned}$$

Then $(M_n)_{n\geq 0}$ is a martingale with respect to the natural filtration $\{\mathcal{F}_n\}$, thanks to the “taking out what is known property”.

13.4 Example. Let X be an integrable random variable and let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration. Let

$$X_n = \mathbb{E}(X|\mathcal{F}_n), \qquad n \geq 0.$$

Then $(X_n)_{n\geq 0}$ is a martingale (with respect to $\{\mathcal{F}_n\}$), by the tower property.

Intuitively, martingales model fair gambling games. If X_n is your capital at time n, then $X_n - X_{n-1}$ are your net winnings in game n. Note that (X_n) is a martingale if and only if $\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) = 0$, that is the game is fair (with all the knowledge at time $n-1$, the net winnings for game n are 0 on average).

13.2 Martingale transforms and stopping times

Let $X = (X_n)_{n\geq 0}$ be an adapted process and let $H = (H_n)_{n\geq 0}$ be a predictable process. Define

$$\begin{aligned} Y_0 &= 0, \\ Y_n &= \sum_{k=1}^{n} H_k(X_k - X_{k-1}), \qquad n \geq 1. \end{aligned}$$

The process $Y = (Y_n)_{n\geq 0}$, denoted $Y = H \bullet X$, is called the **martingale transform** of X by H. (It is a discrete analogue of the stochastic integral $\int H \mathrm{d}X$).

Intuitively, thinking of H_n as your stake on game n, your total winnings at time n are $Y_n = (H \bullet X)_n$.

The fundamental lemmasays that martingale transforms of martingales are of course martingales.

13.5 Lemma. *Let $H = (H_n)$ be a bounded predictable process, that is for some constant K, $|H_n| \leq K$ a.s. for every n. Let $X = (X_n)$ be an adaptable process.*

(i) If H is nonnegative, that is for every n, $H_n \geq 0$ and X is a supermartingale, then $H \bullet X$ is a supermartingale.

(ii) If X is a martingale, then $H \bullet X$ is a martingale.

Proof. (i): Let $Y = H \bullet X$. Since H is bounded, Y is in L_1. Moreover, $Y_n - Y_{n-1} = H_n(X_n - X_{n-1})$ and since H_n is $\mathcal{F}_{n-1}$-measurable (as being predictable), we have

$$\mathbb{E}(Y_n - Y_{n-1}|\mathcal{F}_{n-1}) = \mathbb{E}(H_n(X_n - X_{n-1})|\mathcal{F}_{n-1}) = H_n\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) \leq 0$$

because $H_n \geq 0$ and $\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1})$ (X is a supermartingale). The proof of (ii) is the same. □

A function $\tau\colon \Omega \to \{0, 1, 2, \ldots\} \cup \{+\infty\}$ is a **stopping time** if for every $n \geq 0$, $\{\tau \leq n\} \in \mathcal{F}_n$ (equivalently, $\{\tau = n\} \in \mathcal{F}_n$). Intuitively, τ tells you when to stop playing.

13.6 Example. Let $X = (X_n)$ be an adapted process and let $B \in \mathcal{B}(\mathbb{R})$ be a Borel set. The time of the first entry of X into B,

$$\tau = \inf\{n \geq 0,\ X_n \in B\}$$

is a stopping time. Indeed,

$$\{\tau \leq n\} = \bigcup_{k \leq n} \{X_k \in B\} \in \sigma\left(\bigcup_{k \leq n} \mathcal{F}_k\right) \subset \mathcal{F}_n.$$

On the other hand, in general, $\eta = \sup\{n \leq 10,\ X_n \in B\}$ is *not* a stopping time (why?).

We shall often use notation

$$a \wedge b = \min\{a, b\}, \qquad a \vee b = \max\{a, b\}, \qquad a, b \in \mathbb{R}.$$

13.7 Example. If σ, τ are stopping times, then $\sigma \wedge \tau$, $\sigma \vee \tau$, $\sigma + \tau$ are also stopping times.

For a process $X = (X_n)_{n \geq 0}$, we set $X^\tau = (X_{\tau \wedge n})_{n \geq 0}$ which is called the **stopped process**.

The following, often called the optional sampling (or stopping) lemma, says that the stopped process of a supermartingale is a supermartingale (without any extra assumptions).

13.8 Lemma (Doob's optional sampling lemma)**.** *If X is a supermartingale and τ is a stopping time, then the stopped process X^τ is a supermartingale.*

Proof. Let $H_n = \mathbf{1}_{\tau \geq n}$ (we bet 1 until we quit the game and then bet 0). This process takes values in $\{0, 1\}$, so it is in particular nonnegative and bounded. To check that it is predictable, it thus suffices to check that $\{H_n = 0\}$ is in $\mathcal{F}_{n-1}$ which is clear because $\{H_n = 0\} = \{\tau \leq n - 1\}$. Finally,

$$\begin{aligned}(H \bullet X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1}) &= \sum_{1 \leq k \leq n} \mathbf{1}_{k \leq \tau}(X_k - X_{k-1}) \\ &= \sum_{1 \leq k \leq \tau \wedge n} (X_k - X_{k-1}) \\ &= X_{\tau \wedge n} - X_0,\end{aligned}$$

so $H \bullet X = X^\tau - X_0$, the stopped process is a martingale transform. We are done by Lemma 13.5. $\square$

13.9 Example. Let $X = (X_n)_{n \geq 0}$ be a supermartingale and let τ be a stopping time. Then, for every $n \geq 0$,

$$\mathbb{E}X_{\tau \wedge n} \leq \mathbb{E}X_0$$

and

$$\mathbb{E}X_{\tau\wedge n} \geq \mathbb{E}X_n.$$

The first inequality follows because X^τ is a supermartingale, so in particular,

$$\mathbb{E}X_{\tau\wedge n} = \mathbb{E}X_n^\tau \leq \mathbb{E}X_0^\tau = \mathbb{E}X_{\tau\wedge 0} = \mathbb{E}X_0.$$

The second inequality follows because $\{\tau = k\} \in \mathcal{F}_k$, so for every $k \leq n$, $\mathbb{E}X_k \mathbf{1}_{\{\tau=k\}} \geq \mathbb{E}X_n \mathbf{1}_{\{\tau=k\}}$ (since $X_k \geq \mathbb{E}(X_n|\mathcal{F}_k)$), hence

$$\mathbb{E}X_{\{\tau\wedge n\}} = \sum_{k=0}^{n} \mathbb{E}X_k \mathbf{1}_{\{\tau=k\}} + \mathbb{E}X_n \mathbf{1}_{\{\tau>n\}} \geq \sum_{k=0}^{n} \mathbb{E}X_n \mathbf{1}_{\{\tau=k\}} + \mathbb{E}X_n \mathbf{1}_{\{\tau>n\}} = \mathbb{E}X_n.$$

13.10 Lemma (Doob's optional sampling lemma – continuation)**.** *Let X be a supermartingale and let τ be a stopping time. Then*

$$X_\tau \in L_1 \qquad \text{and} \qquad \mathbb{E}X_\tau \leq \mathbb{E}X_0,$$

if one of the following conditions holds

(i) τ is bounded,

(ii) X is bounded (say, $|X_n| \leq K$ for every $n \geq 0$) and $\tau < \infty$ a.s.

(iii) $\mathbb{E}\tau < \infty$ and X has bounded increments: $|X_n - X_{n-1}| \leq K$ for every $n \geq 1$,

(iv) X is nonnegative and $\tau < \infty$ a.s.

Moreover, if X is martingale and one of the conditions (i)-(iii) holds, then

$$X_\tau \in L_1 \qquad \text{and} \qquad \mathbb{E}X_\tau = \mathbb{E}X_0.$$

Proof. Let X be a supermartingale. We have

$$|X_{\tau\wedge n} - X_0| \leq \sum_{1\leq k\leq \tau\wedge n} |X_k - X_{k-1}|.$$

If (i) holds, say $\tau \leq T$ a.s. for some positive integer T, then applying the above inequality to $n = T$ gives $|X_\tau| \leq |X_0| + \sum_{k\leq T} |X_k - X_{k-1}|$ showing that $X_\tau \in L_1$. If (ii) holds, we trivially have $|X_\tau| \leq K$, so $X_\tau \in L_1$. If (iii) holds, we get $|X_{\tau\wedge n}| \leq |X_0| + \sum_{1\leq k\leq\tau} |X_k - X_{k-1}| \leq |X_0| + \tau K$ which is in L_1, so taking the expectation, letting $n \to \infty$ and using Lebesgue's dominated convergence theorem shows that $\mathbb{E}|X_\tau| \leq \mathbb{E}|X_0| + K\mathbb{E}\tau$. If (iv) holds, since X is assumed to be nonnegative, $X_\tau \in L_1$ follows from the inequality $\mathbb{E}X_\tau \leq \mathbb{E}X_0$ argued below.

Now we show $\mathbb{E}X_\tau \leq \mathbb{E}X_0$. From Lemma 13.8 we know that $X_{\tau\wedge n} \in L_1$ and $\mathbb{E}X_{\tau\wedge n} \leq \mathbb{E}X_0$. For (i), say $\tau \leq N$ a.s., simply take $n = N$. For (ii), take $n \to \infty$ and use Lebesgue's dominated convergence theorem. For (iii), as noted earlier, $X_{\tau\wedge n}$

is dominated by $|X_0| + K\tau$, so we can use Lebesgue's dominated convergence theorem again. For (iv), we use Fatou's lemma,

$$\mathbb{E}X_\tau = \mathbb{E}\liminf X_{\tau\wedge n} \le \liminf \mathbb{E}X_{\tau\wedge n} \le \mathbb{E}X_0.$$

Finally, if X is a martingale, in each of the cases (i)-(iii), we use the previous part for X and $-X$. $\square$

13.11 Example. Let X be a simple random walk on $\mathbb{Z}$, that is $X_0 = 0$, $X_n = \varepsilon_1 + \cdots + \varepsilon_n$, where $\varepsilon_1, \varepsilon_2, \ldots$ are i.i.d. symmetric random signs, so X is a martingale (with respect to the natural filtration). Let $\tau = \inf\{n \ge 1,\ X_n = 1\}$ be the first moment of visiting 1. It is known that $\mathbb{P}(\tau < \infty) = 1$. However, $\mathbb{E}X_\tau = \mathbb{E}1 = 1$ and $\mathbb{E}X_0 = 0$, so in this case $\mathbb{E}X_\tau \neq \mathbb{E}X_0$. Since X has bounded increments, in view of Lemma 13.10 (iii), we have $\mathbb{E}\tau = +\infty$.

We shall now determine the distribution of τ using a martingale argument. Fix $\lambda > 0$ and let $M_n = e^{\lambda X_n}/(\mathbb{E}e^{\lambda\varepsilon_1})^n$. By Example 13.3, M is a martingale. We have

$$\mathbb{E}e^{\lambda\varepsilon_1} = \frac{e^\lambda + e^{-\lambda}}{2} = \cosh\lambda,$$

so

$$M_n = (\cosh\lambda)^{-n} e^{\lambda X_n}.$$

By Doob's optional sampling lemma,

$$\mathbb{E}[(\cosh\lambda)^{-\tau\wedge n} e^{\lambda X_{\tau\wedge n}}] = \mathbb{E}M_{\tau\wedge n} = 1.$$

Since $\lambda > 0$ and $X_{\tau\wedge n} \le 1$, we have that $e^{\lambda X_{\tau\wedge n}}$ is bounded by e^λ. Clearly, we have $(\cosh\lambda)^{-\tau\wedge n} \le 1$, so by Lebesgue's dominated convergence theorem, letting $n \to \infty$ yields

$$\mathbb{E}(\cosh\lambda)^{-\tau} e^\lambda = 1$$

because

$$(\cosh\lambda)^{-\tau\wedge n} e^{\lambda X_{\tau\wedge n}} \to \begin{cases} (\cosh\lambda)^{-\tau} e^\lambda, & \text{on } \{\tau < \infty\}, \\ 0, & \text{on } \{\tau = \infty\} \end{cases}$$

($X_{\tau\wedge n}$ stays bounded by 1 and if $\tau < \infty$ clearly converges to X_τ). Letting $\lambda \to 0+$ in a decreasing way, we have $(\cosh\lambda)^{-\tau} \to \mathbf{1}_{\{\tau<\infty\}}$, thus $\mathbb{P}(\tau < \infty) = \mathbb{E}\,\mathbf{1}_{\{\tau<\infty\}} = 1$. Finally, letting $x = (\cosh\lambda)^{-1}$ and using $e^{-\lambda} = \frac{1-\sqrt{1-x^2}}{x}$, we find the generating function of τ,

$$\mathbb{E}x^\tau = e^{-\lambda} = \frac{1-\sqrt{1-x^2}}{x} = \sum_{k=1}^{\infty} (-1)^{k+1}\binom{1/2}{k} x^{2k-1},$$

hence the distribution of τ, $\mathbb{P}(\tau = 2k-1) = (-1)^{k+1}\binom{1/2}{k}$, $k \ge 1$.

13.3 Convergence theorem

The main martingale convergence theorem is due to Doob.

13.12 Theorem (Doob's "forward" convergence theorem)**.** *Let X be a supermartingale bounded in L_1, that is for some constant K, $\mathbb{E}|X_n| \leq K$ for all n. Then there is an integrable random variable X_∞ such that $X_n \xrightarrow[n\to\infty]{a.s.} X_\infty$.*

13.13 Corollary. *If X is a nonnegative supermartingale, then there is an integrable random variable X_∞ such that $X_n \xrightarrow[n\to\infty]{a.s.} X_\infty$.*

Proof. We have $\mathbb{E}|X_n| = \mathbb{E}X_n \leq \mathbb{E}X_0$ (Remark 13.1), so X is bounded in L_1. □

13.14 Remark. Since each X_n is $\mathcal{F}_\infty$ measurable, so is X_∞. Suppose X is nonnegative. Then, for a fixed index m, by Fatou's lemma,

$$\mathbb{E}(X_\infty|\mathcal{F}_m) = \mathbb{E}(\liminf X_n|\mathcal{F}_m) \leq \liminf \mathbb{E}(X_n|\mathcal{F}_m) \leq \mathbb{E}X_m,$$

so the extended sequence $(X_0, X_1, X_2, \ldots, X_\infty)$ satisfies the supermartingale property (with respect to $\{\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_\infty\}$). In particular,

$$\mathbb{E}X_\infty \leq \ldots \leq \mathbb{E}X_1 \leq \mathbb{E}X_0.$$

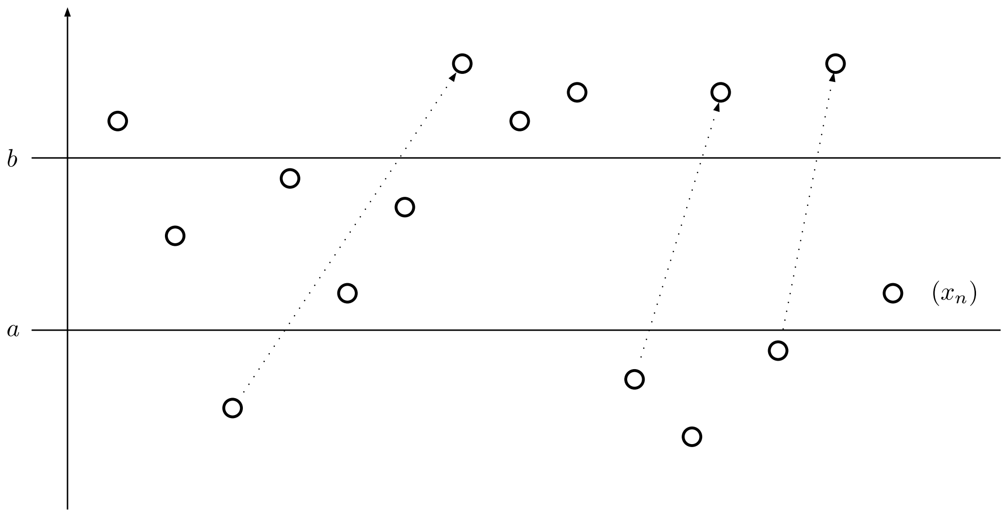

Figure 1: The number of upcrossings is 3.

The proof (undoubtedly from the book) of Doob's convergence theorem relies on the so-called upcrossings. For $a < b$, a sequence $(x_n)_{n\geq 0}$ and an index $N \geq 0$, we define the **number of upcrossings by time** N as

$$\begin{aligned} U_N(a,b) = \text{largest } k \geq 1 \text{ such that there are indices} \\ 0 \leq s_1 < t_1 < s_2 < t_2 < \ldots < s_k < t_k \leq N \\ \text{with } x_{s_i} < a \text{ and } x_{t_i} > b \text{ for each } i = 1, \ldots, k. \end{aligned}$$

See Figure 1.

13.15 Lemma (Doob's upcrossing inequality)**.** *Let X be a supermartingale, let $a < b$, $N \geq 0$. Let $U_N(a,b)$ be the number of uprossings by time N of $(X_n)_{n\geq 0}$. Then*

$$(b-a)\mathbb{E}U_N(a,b) \leq \mathbb{E}(X_N - a)_-.$$

Proof. Define

$$\begin{aligned} H_1 &= \mathbf{1}_{\{X_0 < a\}}, \\ H_{n+1} &= \mathbf{1}_{\{H_n = 1\}}\,\mathbf{1}_{\{X_n \leq b\}} + \mathbf{1}_{\{H_n = 0\}}\,\mathbf{1}_{\{X_n < a\}}, \qquad n \geq 1. \end{aligned}$$

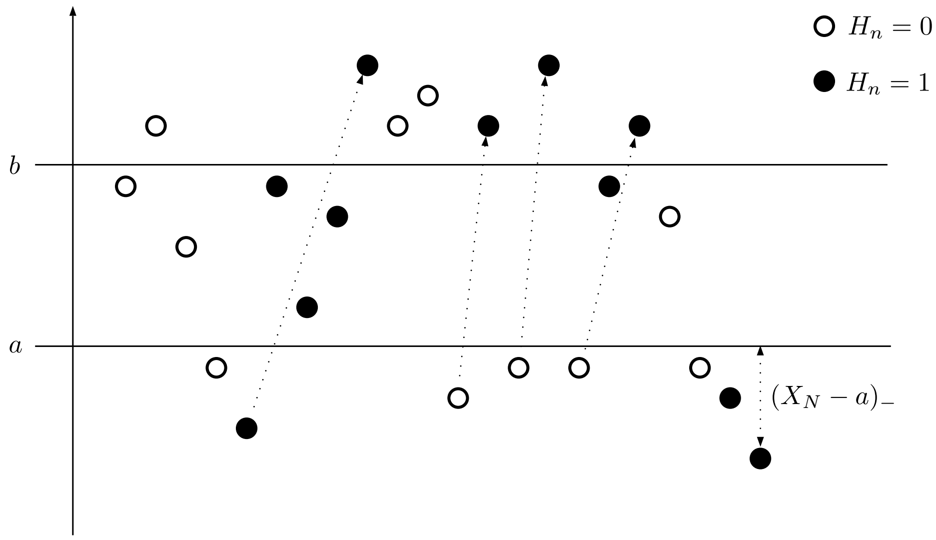

Figure 2: The process Y increases along upcrossings and $(X_N - a)_-$ bounds the loss in the last interval of play $\{n,\ H_n = 1\}$.

This is a predictable process taking values in $\{0, 1\}$. Informally, $H = 1$ on the ways to upcross, or as a gambling strategy: "wait until X gets below a, play stake 1 until X gets above b, repeat".

Let $Y = H \bullet X$ be the martingale transform of X by H. The crucial observation is that

$$Y_N \geq (b-a)U_N(a,b) - (X_N - a)_-.$$

Explanation: every upcrossing increases Y by at least $(b-a)$ and the nonpositive term $-(X_N - a)_-$ offsets the potential loss incurred in the last interval of play (see Figure 2).

Taking the expectation and using that $\mathbb{E}Y_N \leq \mathbb{E}Y_0 = 0$ (Y is a supermartingale, see Lemma 13.5) finishes the proof. $\square$

Proof of Theorem 13.12. Consider the event

$$
\begin{aligned}
\mathcal{E} &= \{\omega,\ X_n(\omega) \text{ does not converge to a limit in } [-\infty, \infty]\} \\
&= \{\liminf X_n < \limsup X_n\} \\
&= \bigcup_{\substack{a<b \\ a,b\in\mathbb{Q}}} \{\liminf X_n < a < b < \limsup X_n\}.
\end{aligned}
$$

Fix $a < b$. Note that by the definition of lim inf and lim sup,

$$
\{\liminf X_n < a < b < \limsup X_n\} \subset \{U_\infty(a, b) = \infty\},
$$

where $U_\infty(a, b)$ is defined (pointwise) as the limit $\lim_{N\to\infty} U_N(a, b) \in [0, +\infty]$ (which exists by the monotonicity of $U_N(a, b)$). By Lemma 13.15,

$$
(b - a)\mathbb{E}U_N(a, b) \leq \mathbb{E}(X_N - a)_- \leq \mathbb{E}|X_N| + |a| \leq K + |a|,
$$

so in particular, letting $N \to \infty$, we get (by Lebesgue's monotone convergence theorem) that $\mathbb{E}U_\infty(a, b) < \infty$ and thus $\mathbb{P}\left(U_\infty(a, b) = \infty\right) = 0$. As a result, $\mathbb{P}\left(\mathcal{E}\right) = 0$. On $\mathcal{E}^c$, we can define

$$
X_\infty = \lim X_n \in [-\infty, \infty].
$$

By Fatou's lemma,

$$
\mathbb{E}|X_\infty| = \mathbb{E} \liminf |X_n| \leq \liminf \mathbb{E}|X_n| \leq K,
$$

so X_∞ is integrable (and thus $X_\infty \in (-\infty, \infty)$). $\square$

13.4 Exercises

1. Verify the claim made in Example 13.3.

2. Verify the claim made in Example 13.4.

3. Let τ, σ be stopping times (relative to $(\Omega, \{\mathcal{F}_n\}, \mathcal{F}, \mathbb{P})$). Prove that $\tau \wedge \sigma$, $\tau \vee \sigma$ and $\tau + \sigma$ are also stopping times. Are $\tau + 1$, $\tau - 1$ stopping times as well?

   Here and throughout: $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$, $a, b \in \mathbb{R}$.

4. Let $X = (X_n)_{n \geq 0}$ be an adaptable process (to a filtration $\{\mathcal{F}_n\}_{n \geq 0}$). Let B be a Borel subset of $\mathbb{R}$. Define

$$\tau_1 = \inf\{n : \ X_n \in B\} \qquad \text{the first visit in } B,$$

$$\tau_k = \inf\{n > \tau_{k-1} : \ X_n \in B\} \qquad \text{the } k\text{th visit in } B,\ k \geq 2.$$

   We know that τ_1 is a stopping time. Show that each τ_k is also a stopping time.

5. Let $X = (X_n)_{n \geq 0}$ be a martingale and let $f \colon \mathbb{R} \to \mathbb{R}$ be a convex function such that $\mathbb{E}|f(X_n)| < \infty$ for every $n \geq 0$. Show that $(f(X_n))_{n \geq 0}$ is a submartingale.

6. Let $X_0, X_1, \ldots$ be i.i.d. square-integrable random variables with mean 0. Let $Y_0 = 0$ and $Y_n = X_0X_1 + X_1X_2 + \cdots + X_{n-1}X_n$, $n \geq 1$. Show that $(Z_n)_{n \geq 0}$ is a martingale relative to the natural filtration $\{\mathcal{F}_n\}_{n \geq 0}$ of $(X_n)_{n \geq 0}$.

7. Let $X = (X_n)_{n \geq 0}$ be an integrable adapted process (to a filtration $\{\mathcal{F}_n\}_{n \geq 0}$). Show that X is a martingale if and only if for every bounded stopping time τ, we have $\mathbb{E}X_\tau = \mathbb{E}X_0$.

   Hint: First show that $\mathbb{E}X_n = \mathbb{E}X_0$ for every $n \geq 0$. Then, given $n \geq 0$ and $A \in \mathcal{F}_n$, consider $\tau = n\,\mathbf{1}_A + (n+1)\,\mathbf{1}_{A^c}$.

8. Let X be a nonnegative supermartingale and let τ be a stopping time. Show that

$$\mathbb{E}X_\tau \mathbf{1}_{\tau < \infty} \leq \mathbb{E}X_0.$$

   Deduce that $\mathbb{P}\left(\sup_n X_n \geq t\right) \leq \frac{\mathbb{E}X_0}{t}$, for $t > 0$.

9. *Pólya's urn.* At time 0, an urn contains 1 black ball and 1 white ball. At each time $n = 1, 2, 3, \ldots$ a ball is chosen at random from the urn and is replaced together with a new ball of the same colour. Just after time n, there are therefore $n + 2$ balls in the urn, of which $B_n + 1$ are black, where B_n denotes the number of black balls chosen by time n, $B_0 = 0$. Let $M_n = \frac{B_n + 1}{n+2}$, $n \geq 0$. Prove that

   (a) M is a martingale (relative to the natural filtration $\mathcal{F}_n = \sigma(B_0, \ldots, B_n)$),

   (b) $\mathbb{P}\left(B_n = k\right) = \frac{1}{n+1}$, for $0 \leq k \leq n$,

(c) M_n converges a.s., say to M_∞

(d) M_∞ is uniform on $[0,1]$,

(e) $X_n = \frac{(n+1)!}{B_n!(n-B_n)!}\theta^{B_n}(1-\theta)^{n-B_n}$ is a martingale, where $0<\theta<1$ is fixed.

10. *Bellman's Optimality Principle.* Let $\frac{1}{2} < p < 1$. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = -1)$. Your winnings per unit stake on game n are X_n. Your stake H_n on game n must satisfy $H_n \in (0, Y_{n-1})$, where Y_n is your fortune at time n, Y_0 is a positive constant and $Y_n = Y_{n-1} + H_n X_n$. Show that if H is a predictable process, then $(\log Y_n - n\alpha)_{n\geq 0}$ is a supermartingale, where $\alpha = p\log p + (1-p)\log(1-p) + \log 2$. Deduce that $\mathbb{E}\log(Y_n/Y_0) \leq n\alpha$. Find the best strategy H, that is the one that gives a martingale, hence equality.

11. Suppose that τ is a stopping time (relative to $\{\mathcal{F}_n\}_{n\geq 0}$) such that for some $N \geq 1$ and $\varepsilon > 0$, we have for every n,

$$\mathbb{P}(\tau \leq n + N | \mathcal{F}_n) > \varepsilon \text{ a.s.}$$

Show that for $k = 1, 2, \ldots$, we have $\mathbb{P}(\tau > kN) \leq (1-\varepsilon)^k$ and deduce that $\mathbb{E}\tau < \infty$.

12. *ABRACADABRA.* At each of times $1, 2, 3, \ldots$, a monkey types a capital letter at random, the sequence of letters typed forming an i.i.d. sequence of random variables each chosen uniformly from the 26 capital letters. Let τ be the first time by which the monkey has produced the consecutive sequence 'ABRACADABRA'. Using martingale theory, show that

$$\mathbb{E}\tau = 26^{11} + 26^4 + 26.$$

13. *Gambler's ruin.* Let $0 < p < 1$, $p \neq \frac{1}{2}$. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = -1)$ for every n. Let a and b be integers with $0 < a < b$. Let $S_0 = a$, $S_n = a + X_1 + \ldots + X_n$, $n \geq 1$ and $\tau = \inf\{n \geq 0 : \; S_n \in \{0, b\}\}$. Show that $\mathbb{E}\tau < \infty$. Let $X_n = S_n - n(2p-1)$ and $Y_n = \left(\frac{1-p}{p}\right)^{S_n}$, $n \geq 0$. Show that X and Y are martingales. Deduce the values of $\mathbb{P}(S_\tau = 0)$ and $\mathbb{E}S_\tau$. What about the symmetric case $p = \frac{1}{2}$?

14. *Wald's identity.* Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration and let X be an integrable adaptable process such that the X_k have the same distribution and X_{k+1} is independent of $\mathcal{F}_k$ for every $k \geq 0$. Let $S_n = X_1 + \ldots + X_n$. Let τ be a stopping time with $\mathbb{E}\tau < \infty$. Show that

$$\mathbb{E}(X_1 + \cdots + X_\tau) = \mathbb{E}\tau \cdot \mathbb{E}X_1.$$

If additionally $\mathbb{E}X_k^2 < \infty$, then

$$\mathbb{E}(S_\tau - \tau\mathbb{E}X_1)^2 = (\mathbb{E}\tau)\operatorname{Var}(X_1).$$

15. *Baby version of Kakutani's theorem.* Let $X_1, X_2, \ldots$ be i.i.d. nonnegative random variables with mean 1 such that $\mathbb{P}(X_i = 1) < 1$. Let $M_n = X_1 \cdot \ldots \cdot X_n$, $n \geq 1$. Show that M_n converges a.s. but *not* in L_1.

   Hint. Use Corollary 13.13 (M is a nonnegative martingale). To tackle convergence in L_1, first using a Cauchy condition in L_1, show that $M_n \to M_\infty$ in L_1 implies that $\lim_n M_n = 1$ a.s. Then consider a.s. convergence of $\tilde{M}_n = \frac{\sqrt{X_1} \cdot \ldots \cdot \sqrt{X_n}}{(\mathbb{E}\sqrt{X_1})^n}$ using again a martingale argument.

14 Martingales II

14.1 L_2 martingales

Whenever possible, one of the easiest ways to prove that a martingale is bounded in L_1 is to show that it is bounded in L_2 and use $\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{1/2}$. For martingales, verifying L_2-type conditions can be done very efficiently. Moreover, bounded L_2 martingales admit better convergence results and have nice applications to classical topics such as series of independent random variables.

We begin with a basic condition saying that a martingale is bounded in L_2 if and only if the series of the squares of the L_2-norms of its increments is convergent.

14.1 Theorem. *Let $M = (M_n)_{n \geq 0}$ be a square-integrable martingale, that is $\mathbb{E}M_n^2 < \infty$ for every $n \geq 0$. Then*

$$M \text{ is bounded in } L_2, \text{ that is } \exists K > 0 \ \forall n \geq 0 \ \mathbb{E}M_n^2 \leq K \tag{14.1}$$

if and only if

$$\sum_{k \geq 1} \mathbb{E}(M_k - M_{k-1})^2 < \infty. \tag{14.2}$$

Moreover, when this holds, there exists a random variable M_∞ which is in L_2 such that

$$M_n \xrightarrow[n \to \infty]{} M_\infty \text{ a.s. and in } L_2.$$

Proof. For $i < j$, $\mathbb{E}(M_j | \mathcal{F}_i) = M_i$, that is $\mathbb{E}(M_j - M_i)\mathbf{1}_A = 0$ for every $A \in \mathcal{F}_i$, or, in other words, $M_j - M_i$ is orthogonal to the subspace $L_2(\mathcal{F}_i)$. In particular, if $k < l \leq m < n$, then

$$\mathbb{E}(M_l - M_k)(M_n - M_m) = 0.$$

Consequently, writing M as the sum of its increments,

$$M_n = M_0 + \sum_{k=1}^{n} (M_k - M_{k-1}),$$

we find that all the terms are orthogonal, so in particular,

$$\mathbb{E}M_n^2 = \mathbb{E}M_0^2 + \sum_{k=1}^{n} \mathbb{E}(M_k - M_{k-1})^2.$$

This identity explains the equivalence of (14.1) and (14.2).

Suppose these conditions hold. By Doob's convergence theorem for martingales bounded in L_1 (Theorem 13.12), there is an integrable random variable M_∞ such that $M_n \to M_\infty$ a.s. To see that M_∞ is in fact square-integrable and the convergence is also in L_2, first note that, by orthogonality, for $n, r \geq 0$, we have

$$\mathbb{E}(M_{n+r} - M_n)^2 = \sum_{k=n+1}^{n+r} \mathbb{E}(M_k - M_{k-1})^2,$$

so letting $r \to \infty$ and using Fatou's lemma, we conclude

$$\mathbb{E}(M_\infty - M_n)^2 \leq \sum_{k \geq n+1} \mathbb{E}(M_k - M_{k-1})^2 < \infty,$$

hence $M_\infty \in L_2$. Moreover, since tails of convergent series go to 0, this bound also shows that

$$\mathbb{E}(M_\infty - M_n)^2 \xrightarrow[n\to\infty]{} 0,$$

that is $M_n \to M_\infty$ in L_2. □

We introduce Doob's decomposition, which will be particularly fruitful to study L_2 martingales.

14.2 Theorem (Doob's decomposition)**.** *Let $X = (X_n)_{n \geq 0}$ be an adapted, integrable process. It admits the decomposition*

$$X_n = X_0 + M_n + A_n, \quad n \geq 0,$$

where $M = (M_n)_{n \geq 0}$ is a martingale with $M_0 = 0$ and $A = (A_n)_{n \geq 0}$ is a predictable process with $A_0 = 0$. Moreover, if $X = X_0 + \tilde{M} + \tilde{A}$ is another such decomposition, then

$$\mathbb{P}\left(\forall n \geq 0 \; M_n = \tilde{M}_n, A_n = \tilde{A}_n\right) = 1.$$

Moreover, X is a supermartingale if and only if the process A is nonincreasing, that is $A_n \geq A_{n+1}$ for all $n \geq 0$ a.s.

Proof. Let

$$\begin{aligned} A_0 &= 0, \\ A_n &= \sum_{k=1}^{n} \mathbb{E}(X_k - X_{k-1} | \mathcal{F}_{k-1}), \quad n \geq 1 \\ M_n &= X_n - X_0 - A_n, \quad n \geq 0. \end{aligned}$$

Plainly, A satisfies the desired properties. Moreover, M is integrable and adaptable and it satisfies the martingale property because

$$\begin{aligned} M_{n+1} - M_n = X_{n+1} - X_n - (A_{n+1} - A_n) &= X_{n+1} - X_n - \mathbb{E}(X_{n+1} - X_n | \mathcal{F}_n) \\ &= X_{n+1} - \mathbb{E}(X_{n+1} | \mathcal{F}_n), \end{aligned}$$

so

$$\mathbb{E}(M_{n+1} - M_n | \mathcal{F}_n) = \mathbb{E}(X_{n+1} | \mathcal{F}_n) - \mathbb{E}(X_{n+1} | \mathcal{F}_n) = 0.$$

Moreover, if $X = X_0 + \tilde{M} + \tilde{A}$ for a martigale $\tilde{M}$ and a predictable process $\tilde{A}$, then for every $n \geq 1$,

$$\begin{aligned} \mathbb{E}(X_n - X_{n-1} | \mathcal{F}_{n-1}) &= \mathbb{E}(\tilde{M}_n - \tilde{M}_{n-1} | \mathcal{F}_{n-1}) + \mathbb{E}(\tilde{A}_n - \tilde{A}_{n-1} | \mathcal{F}_{n-1}) \\ &= 0 + \tilde{A}_n - \tilde{A}_{n-1} = \tilde{A}_n - \tilde{A}_{n-1}, \end{aligned}$$

which gives that $\tilde{A} = A$ a.s. and consequently $\tilde{M} = M$ a.s. Finally, by the identity,

$$\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) = A_n - A_{n-1},$$

we have that $\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) \leq 0$ if and only if $A_n \leq A_{n-1}$. $\square$

We note that Doob's decomposition of the stopped process comes from stopping Doob's decomposition processes.

14.3 Lemma. *Let X be an adapted, integrable process with Doob's decomposition $X = X_0 + M + A$ with M being a martingale and A being a predictable process. Let τ be a stopping time. Then*

$$X^\tau = X_0 + M^\tau + A^\tau$$

is Doob's decomposition of the stopped process X^τ.

Proof. By Doob's optional sampling lemma, we know that M^τ is a martingale, so it suffices to show that A^τ is predictable. This can be seen from the identity

$$A_n^\tau = A_{\tau\wedge n} = A_n \mathbf{1}_{\tau \geq n} + \sum_{k<n} A_k \mathbf{1}_{\tau = k}$$

because $\{\tau \geq n\}$ is in $\mathcal{F}_{n-1}$. $\square$

Let M be a martingale in L_2, that is $\mathbb{E}M_n^2 < \infty$ for every $n \geq 0$. We define its quadratic variation process (the angle-brackets process), denoted $\langle M \rangle = (\langle M \rangle_n)_{n\geq 0}$ as

$$\langle M \rangle_n = \sum_{k=1}^n \mathbb{E}(M_k^2 - M_{k-1}^2|\mathcal{F}_{k-1}),$$

that is $\langle M \rangle = A$, where

$$M^2 = M_0^2 + N + A$$

is Doob's decomposition of M^2 into a martingale N and a predictable process A.

14.4 Remark. Since $\mathbb{E}(M_k M_{k-1}|\mathcal{F}_{k-1}) = M_{k-1}\mathbb{E}(M_k|\mathcal{F}_{k-1}) = M_{k-1}^2$, we have

$$\langle M \rangle_n = \sum_{k=1}^n \mathbb{E}((M_k - M_{k-1})^2|\mathcal{F}_{k-1}).$$

In particular, $\langle M \rangle$ is a nonnegative process. Moreover,

$$\langle M \rangle_n - \langle M \rangle_{n-1} = \mathbb{E}(M_n^2 - M_{n-1}^2|\mathcal{F}_{n-1}) = \mathbb{E}((M_n - M_{n-1})^2|\mathcal{F}_{n-1}) \geq 0,$$

hence $\langle M \rangle = A$ is nondecreasing (which also follows from the fact that M^2 is a submartingale – by Jensen's inequality, see Exercise 13.5). Consequently, we can define (point-wise)

$$\langle M \rangle_\infty = \lim_{n\to\infty} \langle M \rangle_n \in [0, +\infty].$$

14.5 Example. Let $X_1, X_2, \ldots$ be independent random variables which are in L_2. Let $M_0 = 0$ and $M_n = X_1 + \cdots + X_n - (\mathbb{E}X_1 + \cdots + \mathbb{E}X_n)$. We know this is a martingale (Example 13.2). We have,

$$\langle M \rangle_n = \sum_{k=1}^n \mathrm{Var}(X_k)$$

(exercise). Thus, $M_n^2 - \mathrm{Var}(M_n)$ is a martingale.

14.6 Remark. From Doob's decomposition, $\mathbb{E}M_n^2 = \mathbb{E}M_0^2 + \mathbb{E}N_n + \mathbb{E}\langle M \rangle_n = \mathbb{E}M_0^2 + \mathbb{E}\langle M \rangle_n$, thus

$$M \text{ is bounded in } L_2 \text{ if and only if } \mathbb{E}\langle M \rangle_\infty < \infty. \tag{14.3}$$

In view of Theorem 14.1 and the above remark, if $\mathbb{E}\langle M \rangle_\infty < \infty$, then M_n converges a.s. and in L_2. The next theorem refines that and describes very precisely convergence of L_2 martingales in a general situation.

14.7 Theorem. *Let M be a martingale with $M_0 = 0$ which is in L_2, that is $\mathbb{E}M_n^2 < \infty$ for every n.*

(i) On the event $\{\langle M \rangle_\infty < \infty\}$, we have "$\lim_n M_n$ exists and is finite".

(ii) If M has additionally bounded increments, that is there is a constant $K > 0$ such that $|M_n - M_{n-1}| \leq K$ for all $n \geq 0$ a.s., then the converse holds: on the even $\{\lim_n M_n$ exists and is finite$\}$, we have $\langle M \rangle_\infty < \infty$.

(iii) On the event $\{\langle M \rangle_\infty = \infty\}$, we have $\frac{M_n}{\langle M \rangle_n} \xrightarrow[n\to\infty]{} 0$.

(All the inclusions hold modulo sets of measure 0).

14.8 Remark. Part (iii) can be thought of a strong law of large numbers for L_2 martingales. To illustrate this point, let $X_1, X_2, \ldots$ be i.i.d. random variables which are in L_2. We consider the sum martingale $M_n = X_1 + \cdots + X_n - n\mathbb{E}X_1$, for which $\langle M \rangle_n = n\,\mathrm{Var}(X_1)$ (see Example 14.5). Thus, $\{\langle M \rangle_\infty = \infty\} = \Omega$, so, a.s.,

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}X_1}{n\,\mathrm{Var}\,X_1} = \frac{M_n}{\langle M \rangle_n} \to 0,$$

equivalently,

$$\frac{X_1 + \cdots + X_n}{n} \to \mathbb{E}X_1 \quad a.s.$$

Proof. (i): For $l = 1, 2, \ldots$, we define

$$\tau_l = \inf\{n \geq 0 : \langle M \rangle_{n+1} > l\}.$$

Since $\langle M \rangle$ is predictable, this is a stopping time. Moreover, by its definition,

$$\langle M \rangle_{\tau_l} \leq l.$$

Thus, for the stopped process M^{τ_l},

$$\langle M^{\tau_l}\rangle_\infty = \langle M\rangle_\infty^{\tau_l} = \langle M\rangle_{\tau_l} \le l$$

(in the first equality we use Lemma 14.3). In view of (14.1), the stopped process M^{τ_l} is a bounded in L_2 martingale, so $\lim_n M_n^{\tau_l}$ exists (a.s.) and is in L_2. Therefore,

$$\{\langle M\rangle_\infty < \infty\} = \bigcup_{l\ge 1}\{\tau_l = \infty\} \subset \{\lim_n M_n \text{ exists and is finite}\}$$

because on $\{\tau_l = \infty\}$, we have $M_n^{\tau_l} = M_n$. This shows (a).

(ii): For $l = 1, 2, \ldots$, we define

$$\sigma_l = \inf\{n \ge 0 : \ |M_n| > l\}.$$

This is a stopping time. Since M has bounded increments, the stopped process is bounded,

$$|M_n^{\sigma_l}| = |M_{\sigma_l\wedge n}| \le |M_{(\sigma_l\wedge n)-1}| + K \le l + K.$$

Thus M^{σ_l} is bounded in L_2. Therefore $\mathbb{E}\langle M^{\sigma_l}\rangle_\infty < \infty$ and

$$\mathbb{E}\langle M\rangle_{\sigma_l} = \mathbb{E}\langle M\rangle_\infty^{\sigma_l} = \mathbb{E}\langle M^{\sigma_l}\rangle_\infty < \infty.$$

In particular, on $\{\sigma_l = \infty\}$, we have $\langle M\rangle_{\sigma_l} = \langle M\rangle_\infty$ and thus $\langle M\rangle_\infty < \infty$ a.s. on $\{\sigma_l = \infty\}$. Since convergent sequences are bounded, by the definition of σ_l, this finishes the proof of (b),

$$\{\lim M_n \text{ exists and is finite}\} \subset \{\sup_n |M_n| < \infty\} = \sum_{l\ge 1}\{\sigma_l = \infty\} \subset \{\langle M\rangle_\infty < \infty\}.$$

(iii): We define the process Y with $Y_0 = 0$ and

$$Y_n = \sum_{k=1}^n \frac{1}{1+\langle M\rangle_k}(M_k - M_{k-1}), \quad n \ge 1,$$

which is the martingale transform of M by the predictable process $\frac{1}{1+\langle M\rangle}$ (which is bounded because $\langle M\rangle$ is nonnegative). Thus Y is a martingale (in L_2) and since $\langle M\rangle$ is nondecreasing, we have,

$$\begin{aligned}\mathbb{E}((Y_n - Y_{n-1})^2|\mathcal{F}_{n-1}) &= \frac{\langle M\rangle_n - \langle M\rangle_{n-1}}{(1+\langle M\rangle_n)^2}\\ &\le \frac{\langle M\rangle_n - \langle M\rangle_{n-1}}{(1+\langle M\rangle_n)(1+\langle M\rangle_{n-1})}\\ &= \frac{1}{1+\langle M\rangle_{n-1}} - \frac{1}{1+\langle M\rangle_n}.\end{aligned}$$

Therefore,

$$\mathbb{E}Y_n^2 = \sum_{k=1}^n \mathbb{E}(Y_k - Y_{k-1})^2 = \mathbb{E}\left(1 - \frac{1}{1+\langle M\rangle_n}\right) \le 1,$$

so Y is bounded in L_2 and, consequently, $\lim Y_n$ exists and is finite a.s. On the event $\{\langle M\rangle_n \nearrow \infty\} = \{\langle M\rangle_\infty = \infty\}$, we get $\frac{M_n}{\langle M\rangle_n} \to 0$, by the following standard Cesáro-type lemma. □

14.9 Lemma (Kronecker)**.** *Let (b_n) be a sequence of positive real numbers with $b_n \nearrow \infty$. Let (x_n) be a sequence of real numbers. Then,*

$$\textit{if } \sum \frac{x_n}{b_n} \textit{ converges, then } \frac{x_1 + \cdots + x_n}{b_n} \to 0.$$

Proof. Let $s_n = \sum_{k=1}^n \frac{x_k}{b_k}$. Then $s_1 = \frac{x_1}{b_1}$, $s_n - s_{n-1} = \frac{x_n}{b_n}$, $n \geq 2$, so

$$\begin{aligned}\frac{x_1 + \ldots + x_n}{b_n} &= \frac{b_1 s_1 + b_2(s_2 - s_1) + b_3(s_3 - s_2) + \ldots + b_n(s_n - s_{n-1})}{b_n} \\ &= \frac{(b_1 - b_2)s_1 + (b_2 - b_3)s_2 + \cdots + (b_{n-1} - b_n)s_{n-1} + b_n s_n}{b_n} \\ &= \frac{(b_1 - b_2)(s_1 - s_n) + (b_2 - b_3)(s_2 - s_n) + \cdots + (b_{n-1} - b_n)(s_{n-1} - s_n)}{b_n} \\ &\quad + \frac{b_1 s_n}{b_n}.\end{aligned}$$

Fix $\varepsilon > 0$. Since (s_n) is a convergent sequence, it is bounded, say $|s_n| \leq M$ for every n, and by the Cauchy criterion, there is N such that for $n, m > N$, we have $|s_n - s_m| < \varepsilon$. Consequently, for $n > N$,

$$\begin{aligned}&\left|\frac{x_1 + \ldots + x_n}{b_n}\right| \\ &\leq \frac{(b_2 - b_1)2M + \cdots + (b_N - b_{N-1})2M + (b_{N+1} - b_N)\varepsilon + \cdots + (b_n - b_{n-1})\varepsilon}{b_n} + \frac{b_1 M}{b_n} \\ &= 2M\frac{b_N - b_1}{b_n} + \frac{b_n - b_N}{b_n}\varepsilon + \frac{b_1 M}{b_n} \leq 2M\frac{b_N - b_1}{b_n} + \varepsilon + \frac{b_1 M}{b_n}\end{aligned}$$

which is less than, say 2ε for n large enough. □

14.2 Uniformly integrable martingales

We say that a family of random variables $\{X_t\}_{t\in T}$ is **uniformly integrable** if for every $\varepsilon > 0$, there is $K > 0$ such that for all $t \in T$, we have $\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} \leq \varepsilon$. We refer to Appendix I for basic results. We recall one: for $p > 0$, we have

$$X_n \xrightarrow[n\to\infty]{L_p} X \quad \text{if and only if} \quad X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \text{ and } \{|X_n|^p\} \text{ is uniformly integrable} \tag{14.4}$$

(see Theorem I.6

In the context of martingales, the following construction of a uniformly integrable family is rather important.

14.10 Lemma. *If X is an integrable random variable and $\{\mathcal{F}_t\}_{t\in T}$ is a family of sub-σ-algebras, then the family $\{X_t = \mathbb{E}(X|\mathcal{F}_t)\}_{t\in T}$ is uniformly integrable.*

Proof. Fix $\varepsilon > 0$. We choose $\delta > 0$ such that for every event A with $\mathbb{P}(A) < \delta$, we have $\mathbb{E}|X|\,\mathbf{1}_A < \varepsilon$ (see Remark I.5). First note that, by Jensen's inequality,

$$|X_t| = |\mathbb{E}(X|\mathcal{F}_t)| \leq \mathbb{E}(|X|\,|\mathcal{F}_t)$$

and, consequently,

$$\mathbb{E}|X_t| \leq \mathbb{E}|X|.$$

For the event $A = \{|X_t| > K\}$, if K is large enough, we have, by Markov's inequality,

$$\mathbb{P}(A) \leq \frac{1}{K}\mathbb{E}|X_t| \leq \frac{1}{K}\mathbb{E}|X| < \delta,$$

so, using that $A \in \mathcal{F}_t$,

$$\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} = \mathbb{E}|X_t|\,\mathbf{1}_A \leq \mathbb{E}((\mathbb{E}(|X|\,|\mathcal{F}_t)\,\mathbf{1}_A) = \mathbb{E}(\mathbb{E}(|X|\,\mathbf{1}_A\,|\mathcal{F}_t) = \mathbb{E}|X|\,\mathbf{1}_A < \varepsilon,$$

which is the definition of uniform integrability. □

Uniformly integrable martingales are bounded in L_1 (see (i) of Theorem I.4). Thus such martingales converge a.s. Moreover, the following basic result gives L_1 convergence and says that such martingales are of the tower form (see Example 13.4).

14.11 Theorem. *Let M be an uniformly integrable martingale. Then there is an integrable random variable M_∞ such that $M_n \to M_\infty$ a.s. and in L_1. Moreover,*

$$M_n = \mathbb{E}(M_\infty|\mathcal{F}_n), \qquad n \geq 0.$$

Proof. As we said, the existence of $M_\infty \in L_1$ such that $M_n \to M_\infty$ a.s. follows from Doob's convergence theorem (Theorem 13.12). Since $\{M_n\}$ is uniformly integrable and converges in probability to M_∞, by (14.4), $M_n \to M_\infty$ also in L_1. It remains to show that $M_n = \mathbb{E}(M_\infty|\mathcal{F}_n)$ for every n. To this end, we fix $n \geq 0$, fix an event $A \in \mathcal{F}_n$ and argue that $\mathbb{E}M_n\,\mathbf{1}_A = \mathbb{E}M_\infty\,\mathbf{1}_A$. By the martingale property, for every $r \geq n$, we have $\mathbb{E}M_r\,\mathbf{1}_A = \mathbb{E}M_n\,\mathbf{1}_A$. Moreover,

$$|\mathbb{E}M_r\,\mathbf{1}_A - \mathbb{E}M_\infty\,\mathbf{1}_A| \leq \mathbb{E}|M_r - M_\infty| \xrightarrow[r\to\infty]{} 0,$$

so $\mathbb{E}M_n\,\mathbf{1}_A = \mathbb{E}M_r\,\mathbf{1}_A \to \mathbb{E}M_\infty\,\mathbf{1}_A$ as $r \to \infty$ and thus, $\mathbb{E}M_n\,\mathbf{1}_A = \mathbb{E}M_\infty\,\mathbf{1}_A$, as desired. □

For tower-type martingales, we have two refinements.

14.12 Theorem (Lévy's "upward" convergence theorem)**.** *Let X be an integrable random variable and let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration, $\mathcal{F}_\infty = \sigma\left(\bigcup_{n\geq 0}\mathcal{F}_n\right)$. Let $X_n = \mathbb{E}(X|\mathcal{F}_n)$, $n \geq 0$. We have,*

$$X_n \xrightarrow[n\to\infty]{} \mathbb{E}(X|\mathcal{F}_\infty) \quad \textit{a.s. and in } L_1.$$

Proof. By Lemma 14.10, the martingale $(X_n)_{n\geq 0}$ is uniformly integrable, so by Theorem 14.11, there is $X_\infty \in L_1$ such that $X_n \to X_\infty$ a.s. and in L_1. It remains to show that $X_\infty = \mathbb{E}(X|\mathcal{F}_\infty)$. Since X_∞ is $\mathcal{F}_\infty$-measurable, it suffices to argue that for every event $A \in \mathcal{F}_\infty$, we have $\mathbb{E}X_\infty \mathbf{1}_A = \mathbb{E}X \mathbf{1}_A$. All events satisfying this form a λ-system. By Dynkin's theorem, it thus suffices to show that it contains the π-system $\bigcup_{n\geq 0}\mathcal{F}_n$. Suppose $A \in \mathcal{F}_n$ for some n. Then,

$$\mathbb{E}X \mathbf{1}_A = \mathbb{E}X_n \mathbf{1}_A = \mathbb{E}X_\infty \mathbf{1}_A,$$

where the first equality holds because $X_n = \mathbb{E}(X|\mathcal{F}_n)$ and the second one holds because $X_n = \mathbb{E}(X_\infty|\mathcal{F}_n)$, as provided by Theorem 14.11. This finishes the proof. $\square$

14.13 Theorem (Lévy's "downward" convergence theorem)**.** *Let $\{\mathcal{G}_{-n}, n = 1, 2, \ldots\}$ be sub-σ-algebras such that*

$$\mathcal{G}_{-1} \supset \mathcal{G}_{-2} \supset \cdots \supset \mathcal{G}_{-\infty}$$

with $\mathcal{G}_{-\infty} = \bigcap_{n\geq 1}\mathcal{G}_{-n}$. Let X be an integrable random variable and $X_{-n} = \mathbb{E}(X|\mathcal{G}_{-n})$, $n \geq 1$. Then

$$X_{-n} \xrightarrow[n\to\infty]{} \mathbb{E}(X|\mathcal{G}_{-\infty}) \quad \text{a.s. and in } L_1.$$

Proof. We consider the martingale $(X_{-N}, X_{-N+1}, \ldots, X_{-1})$. By Doob's upcrossing inequality from Lemma 13.15,

$$(b-a)\mathbb{E}U_N(a,b) \leq \mathbb{E}(X_{-1}-a)_- < \infty,$$

so as in the proof of Doob's convergence theorem, $X_{-\infty} = \lim_n X_{-n}$ exists and is finite a.s. By the uniform integrability of $\{X_{-n}\}_{n\geq 1}$, we get that also $X_n \to X_{-\infty}$ in L_1. As in the proof of Lévy's upward convergence theorem, $X_{-\infty} = \mathbb{E}(X|\mathcal{G}_{-\infty})$. $\square$

14.3 Maximal inequalities

Maximal inequalities concern tail and moment bounds for the maximum $\max_{k\leq n} X_k$ of a process X.

14.14 Theorem (Doob's maximal inequality)**.** *Let $X = (X_n)_{n\geq 0}$ be a submartingale. Then for every $t > 0$, we have*

$$\mathbb{P}\left(\max_{k\leq n} X_k \geq t\right) \leq \frac{1}{t}\mathbb{E}X_n \mathbf{1}_{\{\max_{k\leq n} X_k \geq t\}} \leq \frac{1}{t}\mathbb{E}X_n^+, \tag{14.5}$$

$$\mathbb{P}\left(\min_{k\leq n} X_k \leq -t\right) \leq \frac{1}{t}(\mathbb{E}X_n \mathbf{1}_{\{\min_{k\leq n} X_k \leq -t\}} - \mathbb{E}X_0) \leq \frac{1}{t}(\mathbb{E}X_n^+ - \mathbb{E}X_0), \tag{14.6}$$

$$\mathbb{P}\left(\max_{k\leq n} |X_k| \geq t\right) \leq \frac{1}{t}(2\mathbb{E}X_n^+ - \mathbb{E}X_0). \tag{14.7}$$

Proof. For the first inequality, consider the stopping time $\tau = \inf\{n \geq 0 : \; X_n \geq t\}$. By Example 13.9,

$$\mathbb{E}X_n \geq \mathbb{E}X_{\tau\wedge n}$$

and

$$\mathbb{E}X_{\tau\wedge n} = \mathbb{E}X_\tau \mathbf{1}_{\{\tau\leq n\}} + \mathbb{E}X_n \mathbf{1}_{\{\tau>n\}} \geq t\mathbb{P}\left(\tau \leq n\right) + \mathbb{E}X_n \mathbf{1}_{\{\tau>n\}},$$

hence

$$t\mathbb{P}\left(\tau \leq n\right) \leq \mathbb{E}X_n \mathbf{1}_{\tau\leq n},$$

so $\{\tau \leq n\} = \{\max_{k\leq n} X_k \geq t\}$ finishes the argument.

For the second inequality, we consider the stopping time $\sigma = \inf\{n \geq 0 : \; X_n \leq -t\}$, use $\mathbb{E}X_0 \leq \mathbb{E}X_{\tau\wedge n}$ and proceed analogously (we leave the details as an exercise).

The third inequality follows from $\max_k |X_k| = \max\{\max_k X_k, \max_k(-X_k)\}$, the union bound and applying the previous two inequalities (we leave the details as an exercise). □

14.15 Corollary. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be convex and let X be a martingale such that $\mathbb{E}|f(X_n)| < \infty$ for every n. Then for every $t > 0$, we have*

$$\mathbb{P}\left(\max_{k\leq n} f(X_k) \geq t\right) \leq \frac{1}{t}\mathbb{E}|f(X_n)|.$$

Proof. It follows from (14.5) applied to the submartingale $(f(X_n))$ (see also Exercise 13.5). □

14.16 Corollary (Kolmogorov's maximal inequality)**.** *Let $X_1, \ldots, X_n$ be independent square-integrable random variables, each with mean* 0 *and let $S_k = X_1 + \ldots + X_k$, $1 \leq k \leq n$. Then*

$$\mathbb{P}\left(\max_{k\leq n} |S_k| \geq t\right) \leq \frac{1}{t^2}\mathbb{E}S_n^2, \qquad t > 0.$$

Proof. We have,

$$\mathbb{P}\left(\max_{k\leq n} |S_k| \geq t\right) = \mathbb{P}\left(\max_{k\leq n} |S_k|^2 \geq t^2\right)$$

and the result follows from the previous corollary applied to the sum martingale ($S_0 = 0, S_1, \ldots, S_n$) and convex function $f(x) = x^2$. □

14.17 Theorem. *(Doob's maximal inequality in L_p) Let $p > 1$ and let $X = (X_n)_{n\geq 0}$ be a nonnegative submartingale. Then*

$$\left(\mathbb{E}\max_{k\leq n} X_k^p\right)^{1/p} \leq \frac{p}{p-1}(\mathbb{E}X_n^p)^{1/p}.$$

Proof. If the right hand side is $+\infty$, there is nothing to prove. Suppose that it is finite. Then, by a trivial bound,

$$\mathbb{E}\max_{k\leq n} X_k^p \leq \mathbb{E}\sum_{k\leq n} X_k^p = \sum_{k=1}^{n} \mathbb{E}X_k^p \leq n\mathbb{E}X_n^p,$$

(we use $\mathbb{E}X_k^p \le \mathbb{E}(\mathbb{E}(X_{k+1}|\mathcal{F}_k))^p \le \mathbb{E}(\mathbb{E}(X_{k+1}^p|\mathcal{F}_k)) = \mathbb{E}X_{k+1}^p$), we conclude that the left hand side is also finite. By (14.5), we have

$$\mathbb{E}\max_{k\le n} X_k^p = \int_0^\infty pt^{p-1}\mathbb{P}\left(\max_{k\le n} X_k > t\right)\mathrm{d}t \le \int_0^\infty pt^{p-2}\mathbb{E}X_n\,\mathbf{1}_{\{\max_{k\le n} X_k \ge t\}}\,\mathrm{d}t.$$

By Fubini's theorem (all the terms are nonnegative), we have

$$\begin{aligned}\int_0^\infty pt^{p-2}\mathbb{E}X_n\,\mathbf{1}_{\{\max_{k\le n} X_k \ge t\}}\,\mathrm{d}t &= p\mathbb{E}X_n\int_0^{\max_{k\le n} X_k} t^{p-2}\mathrm{d}t \\ &= \frac{p}{p-1}\mathbb{E}X_n(\max_{k\le n} X_k)^{p-1} \\ &\le \frac{p}{p-1}(\mathbb{E}X_n^p)^{1/p}(\mathbb{E}\max_{k\le n} X_k^p)^{1-1/p},\end{aligned}$$

where in the last estimate we use Hölder's inequality. Thus

$$\mathbb{E}\max_{k\le n} X_k^p \le \frac{p}{p-1}(\mathbb{E}X_n^p)^{1/p}(\mathbb{E}\max_{k\le n} X_k^p)^{1-1/p},$$

so dividing by $(\mathbb{E}\max_{k\le n} X_k^p)^{1-1/p}$ finishes the proof (it is finite; if it is 0, the inequality is trivial). □

14.4 Martingales bounded in L_p, $p > 1$

14.18 Theorem. *Let $p > 1$.*

(i) If $X = (X_n)_{n\ge 0}$ is a nonnegative submartingale bounded in L_p, then there is a random variable $X_\infty \in L_p$ such that $X_n \to X_\infty$ a.s. and in L_p. Moreover, $\|X_n\|_p \nearrow \|X_\infty\|$.

(ii) If $M = (M_n)_{n\ge 0}$ is a martingale bounded in L_p, then there is a random variable $M_\infty \in L_p$ such that $M_n \to M_\infty$ a.s. and in L_p.

Proof. Let $X_* = \sup_{k\ge 0} X_k$. From Doob's maximal inequality in L_p, Theorem 14.17, and Lebesgue's monotone convergence theorem, we obtain

$$\|X_*\|_p \le \frac{p}{p-1}\sup_{k\ge 0}\|X_k\|_p$$

and the right hand side is finite by the assumption. Since $-X$ is a supermartingale, which is bounded in L_1, thanks to Doob's convergence theorem, there exists a random variable $X_\infty \in L_1$ such that $X_n \to X_\infty$ a.s. By the triangle inequality,

$$|X_n - X_\infty|^p \le (2X_*)^p.$$

Consequently, $X_\infty \in L_p$ and by Lebesgue's dominated convergence theorem, from the pointwise convergence we can conclude that $X_n \to X_\infty$ in L_p. Finally, as we saw in the proof of Doob's maximal inequality in L_p, $\|X_n\|_p$ is a nondecreasing sequence, so $\|X_n\|_p \nearrow \|X_\infty\|_p$. This finishes the proof of (i).

To prove (ii), we apply (i) to $X = |M|$. The existence of $M_\infty \in L_1$ such that $M_n \to M_\infty$ a.s. is guaranteed by Doob's convergence theorem. Since

$$|M_n - M_\infty|^p \leq (|M_n| + |M_\infty|)^p \leq (2X_*)^p,$$

we get the convergence of M_n to M_∞ in L_p, as before, thanks to Lebesgue's dominated convergence theorem. $\square$

14.5 Exercises

1. Find the quadratic variation process of the sum martingale (see Example 14.5).

2. Give an example of a family of random variables $\{X_n\}_{n=1,2,\ldots}$ which is uniformly integrable and $\mathbb{E}\sup_n |X_n| = \infty$.

3. Prove inequality (14.6) and then deduce inequality (14.7).

4. Show that there is no positive finite constant C such that for every nonnegative submartingle $(X_n)_{n\geq 0}$, we have

$$\mathbb{E}\max_{k\leq n} X_k \leq C\mathbb{E}X_n.$$

5. Show that for every nonnegative submartingle $(X_n)_{n\geq 0}$, we have

$$\mathbb{E}\max_{k\leq n} X_k \leq \frac{e}{e-1}\left(1+\mathbb{E}X_n \log_+ X_n\right),$$

   where $\log_+ x = \max\{\log x, 0\}$, $x \geq 0$.

   Hint. For $a \geq 0$, $b > 0$, we have $a\log b \leq a\log_+ a + \frac{b}{e}$.

6. Prove that the constant $\frac{p}{p-1}$ in Doob's maximal inequality from Theorem 14.17 is optimal.

7. *Azuma's inequality.* Let $(M_n)_{n\geq 0}$ be a martingale with $M_0 = 0$ and $|M_k - M_{k-1}| \leq a_k$ for every $k \geq 1$ for some positive constants $a_1, a_2, \ldots$. Then for every $n \geq 1$ and $t > 0$,

$$\mathbb{P}\left(\max_{k\leq n} M_k\right) \leq \exp\left\{-\frac{t^2}{2\sum_{k=1}^n a_k^2}\right\}.$$

   Hint. Follow the proof of Bernstein's inequality (Exercise 6.19). Using convexity, show that for a random variable X with $|X| \leq a$ for some $a > 0$ and $\mathbb{E}X = 0$, we have $\mathbb{E}e^{tX} \leq \cosh(ta)$, $t \in \mathbb{R}$.

15 Applications of martingale theory

15.1 Series of independent random variables

We begin with applications of the L_2 theory to series of independent random variables. The main classical result is Kolmogorov's three-series test which gives necessary and sufficient conditions for a series $\sum X_i$ of independent random variables to be convergent a.s. Combined with Kronecker's lemma, it leads to the strong law of numbers. An advantage of this approach is that it gives a way of obtaining rates of convergence. First, we need a lemma which is a direct consequence of the convergence result for L_2 bounded martingales. Historically, it was established by means of Kolmogorov's maximal inequality.

15.1 Lemma. *Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}X_k^2 < \infty$ and $\mathbb{E}X_k = 0$ for every k. If $\sum_{n=1}^\infty \mathrm{Var}(X_n) < \infty$, then $\sum_{n=1}^\infty X_n$ converges a.s.*

Proof. Of course, we consider the sum martingale $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Since

$$\sup_n \mathbb{E}S_n^2 = \sum_{k\geq 1} \mathrm{Var}(X_k) < \infty,$$

the martingale (S_n) is bounded in L_2, so S_n converges a.s. (and in L_2), by Theorem 14.1. □

15.2 Remark. Alternatively, we can say that since $\langle S\rangle_\infty = \sum_{k=1}^\infty < \infty$, we get the assertion by (i) of Theorem 14.7.

15.3 Remark. If the variables X_n are all bounded, that is there is a constant $K > 0$ such that $|X_n| \leq K$ for every n, then (S_n) has bounded increments and the converse holds: if $\sum_{n=1}^\infty X_n$ converges, then $\sum_{n=1}^\infty \mathrm{Var}(X_n) < \infty$. This follows immediately from (ii) of Theorem 14.7. We shall strengthen it soon by removing the assumption of mean 0.

15.4 Theorem (Kolmogorov's three-series test)**.** *Let $X_1, X_2, \ldots$ be independent random variables. Then $\sum X_n$ converges a.s. if for some $K > 0$ (equivalently, every $K > 0$) the following three conditions hold*

(i) $\sum \mathbb{P}(|X_n| > K) < \infty$,

(ii) $\sum \mathbb{E}X_n \mathbf{1}_{\{|X_n|\leq K\}}$ *converges,*

(iii) $\sum \mathrm{Var}(X_n \mathbf{1}_{\{|X_n|\leq K\}}) < \infty$.

Conversely, if $\sum X_n$ converges a.s., then (i)-(iii) hold for every $K > 0$.

Proof of sufficiency. Suppose that for some positive K conditions (i), (ii) and (iii) hold. Let $\tilde{X}_n = X_n \mathbf{1}_{\{|X_n|\leq K\}}$. First note that, thanks to (i),

$$\sum \mathbb{P}\left(X_n \neq \tilde{X}_n\right) = \sum \mathbb{P}\left(|X_n| > K\right) < \infty,$$

so by the first Borel-Cantelli lemma,

$$\mathbb{P}\left(X_n = \tilde{X}_n \text{ for all but finitely many } n\text{'s}\right) = 1,$$

so it is enough to show that $\sum \tilde{X}_n$ converges a.s. Thanks to (ii), it is enough to show that $\sum(\tilde{X}_n - \mathbb{E}\tilde{X}_n)$ converges a.s. which in view of Lemma 15.1 follows from (iii). $\square$

Proof of necessity. Fix $K > 0$ and suppose that $\sum X_n$ converges a.s. In particular, $X_n \to 0$ a.s., so $\mathbb{P}\left(|X_n| > K \text{ for infinitely many } n\text{'s}\right) = 0$, so by the second Borel-Cantelli lemma, $\sum \mathbb{P}\left(|X_n| > K\right) < \infty$, that is (i) holds. As in the proof of sufficiency, this in turn gives that $X_n = \tilde{X}_n$ eventually, a.s., so we also know that $\sum \tilde{X}_n$ converges a.s. The following lemma applied to the sequence $(\tilde{X}_n)$ finishes the proof. $\square$

15.5 Lemma. *Let (X_n) be a sequence of independent random variables bounded by some positive constant K, that is $|X_n| \leq K$ for every n. If $\sum X_n$ converges, then $\sum \mathbb{E}X_n$ and $\sum \mathrm{Var}(X_n)$ converge.*

Proof. We shall use characteristic functions. Let $Y_n = X_n - \mathbb{E}X_n$. First note that, $|Y_n| \leq 2K$ and, plainly, $\phi_{Y_n}(t) = e^{-it\mathbb{E}X_n}\phi_{X_n}(t)$, so

$$|\phi_{Y_n}(t)| = |\phi_{Y_n}(t)|.$$

Denote $\sigma_n^2 = \mathrm{Var}(X_n) = \mathbb{E}Y_n^2$. By Lemma 10.4,

$$\begin{aligned}\left|\phi_{Y_n}(t) - (1 - \frac{1}{2}\sigma_n^2 t^2)\right| &= \left|\mathbb{E}\left[e^{itY_n} - (1 + itY_n - \frac{1}{2}t^2 Y_n^2\right]\right| \\ &\leq \mathbb{E}\frac{|t|^3|Y_n|^3}{6} \leq \frac{|t|^3(2K)\mathbb{E}Y_n^2}{6} = \frac{|t|K}{3}\sigma_n^2 t^2.\end{aligned}$$

Consequently, for all $|t| < \frac{3}{4K}$, we have

$$|\phi_{Y_n}(t)| \leq 1 - \frac{1}{2}\sigma_n^2 t^2 + \frac{1}{4}\sigma_n^2 t^2 = 1 - \frac{1}{4}\sigma_n^2 t^2.$$

Since $S_n = X_1 + \dots + X_n$ converges a.s., say to S, we have $\phi_{S_n}(t) \to \phi_S(t)$ for every t. By continuity, $|\phi_S(t)| > \frac{1}{2}$ for all t sufficiently small. Fix one such positive t with $t < \frac{3}{4K}$. Then, for all n large enough,

$$\frac{1}{4} < |\phi_{S_n}(t)| = \prod_{k=1}^{n} |\phi_{X_k}(t)| = \prod_{k=1}^{n} |\phi_{Y_k}(t)| \leq e^{-\frac{1}{4}t^2 \sum_{k=1}^{n} \sigma_k^2},$$

which gives $\sum_{k=1}^{\infty} \sigma_n^2 < \infty$, as desired. Finally, by Lemma 15.1 applied to Y_n, we get that the series $\sum Y_n = \sum(X_n - \mathbb{E}X_n)$ converges a.s., which together with $\sum X_n$ being convergent a.s., gives that $\sum \mathbb{E}X_n$ converges. $\square$

We refer to the exercises for applications of the basic lemma 15.1 combined with Kronecker's Lemma 14.9 allowing to obtain strong laws of large numbers with rates of convergence.

15.2 Kolmogorov's $0-1$ law and strong law of large numbers

Here we present applications of Lévy's convergence theorems.

Martingale proof of Theorem 3.17 (Kolmogorov's $0-1$ *law).* Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let $A \in \mathcal{T}$ and $X = \mathbf{1}_A$. Clearly X is independent of $\mathcal{F}_n$ for every n and X is $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$-measurable. Thus, $\mathbb{E}(X|\mathcal{F}_n) = \mathbb{E}X = \mathbb{P}(A)$ and $\mathbb{E}(X|\mathcal{F}_\infty) = X$. By Theorem 14.12,

$$\mathbb{E}(X|\mathcal{F}_n) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X|\mathcal{F}_\infty),$$

so $\mathbb{P}(A) \to X$ a.s. and because $X \in \{0,1\}$, this gives $\mathbb{P}(A) \in \{0,1\}$. $\square$

Strong law of large numbers: martingale proof. Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}|X_1| < \infty$. Then

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{} \mathbb{E}X_1 \quad \text{a.s. and in } L_1$$

(compare this to Etemadi's strong law, Theorem 7.16, where we only assume pairwise independence but only conclude a.s. convergence).

For the proof, let $S_n = X_1 + \cdots + X_n$, $\mathcal{G}_{-n} = \sigma(S_n, S_{n+1}, \ldots)$, $n \geq 1$ and as in Lévy's downward convergence theorem, $\mathcal{G}_{-\infty} = \bigcap_n \mathcal{G}_{-n}$. By Example 12.5,

$$\mathbb{E}(X_1|\mathcal{G}_{-n}) = \frac{S_n}{n}.$$

Thus, by Lévy's theorem (Theorem 14.13), there is an integrable random variable Y such that

$$\frac{S_n}{n} \to Y \quad \text{a.s. and in } L_1.$$

Observe that for every fixed m, $Y = \lim_n \frac{S_n}{n} = \lim_n \frac{X_{m+1}+\cdots+X_n}{n}$. As a result, Y is $\sigma(X_{m+1}, X_{m+2}, \ldots)$-measurable, thus it is $\mathcal{T}$-measurable, where $\mathcal{T}$ is the tail σ-algebra. By Kolmogorov's $0-1$ law, Y is therefore constant a.s., say $Y = c$ a.s. for some $c \in \mathbb{R}$. By the L_1 convergence, $\mathbb{E}X_1 = \lim_n \mathbb{E}\frac{S_n}{n} = \mathbb{E}Y = c$. $\square$

15.3 Kakutani's theorem

The next result is obtained as a basic application of Doob's convergence theorem to product martingales and Doob's maximal inequality in L_2.

15.6 Theorem (Kakutani's theorem on product martingales)**.** *Let* $X_1, X_2, \ldots$ *be independent nonnegative random variables, each one with mean* 1*. Let* $M_0 = 1$ *and*

$M_n = X_1 \cdot \ldots \cdot X_n$. *Then there is an integrable random variable* M_∞ *such that* $M_n \to M_\infty$ *a.s. and the following conditions are equivalent*

(i) $\mathbb{E}M_\infty = 1$

(ii) $M_n \to M_\infty$ *in* L_1

(iii) $M = (M_n)_{n\geq 0}$ *is uniformly integrable*

(iv) $\prod_{k=1}^\infty \mathbb{E}X_k^{1/2} > 0$

(v) $\sum_{k=1}^\infty (1 - \mathbb{E}X_k^{1/2}) < \infty$.

If they fail to hold, then $M_\infty = 0$ *a.s.*

Proof. Here is how we proceed: (ii) $\Leftrightarrow$ (iii), (iv) $\Leftrightarrow$ (v), (iv) $\Rightarrow$ (i)-(iii), NOT(iv) $\Rightarrow$ $M_\infty = 0$ a.s. and as a result NOT(i),NOT(ii). These suffice.

Since M is a nonnegative martingale (see Example 13.3), the existence of $M_\infty \in L_1$ with $M_n \to M_\infty$ a.s. immediately follows from Doob's convergence theorem.

Note that (ii) and (iii) are equivalent because of the characterisation of L_p convergence in terms of uniform integrability, Theorem I.6.

Let $a_k = \mathbb{E}X_k^{1/2}$ which is positive (because $\mathbb{E}X_k = 1$). By Jensen's inequality, $\mathbb{E}X_k^{1/2} \leq (\mathbb{E}X_k)^{1/2} = 1$, so in fact $a_k \in (0, 1]$. Then the equivalence of (iv) and (v), $\prod a_k > 0 \Leftrightarrow \sum(1 - a_k) < \infty$, is a straightforward consequence of the inequalities $1 - x \leq e^{-x}$, $x \in \mathbb{R}$ and, say $1 - x \geq e^{-2x}$, $x \in [0, \frac{1}{2}]$.

Suppose (iv) holds. Consider

$$
\begin{aligned}
Y_0 &= 1, \\
Y_n &= \frac{X_1^{1/2}}{a_1} \cdot \ldots \cdot \frac{X_n^{1/2}}{a_n}.
\end{aligned}
$$

This is a nonnegative martingale, bounded in L_2 because

$$
\mathbb{E}Y_n^2 = \frac{1}{\prod_{k=1}^n a_k^2} \leq \frac{1}{\prod_{k=1}^\infty a_k^2} < \infty.
$$

Note that $M_n = Y_n^2(a_1 \cdot \ldots \cdot a_n)^2 \leq Y_n^2$. Therefore, by Doob's maximal L_2 inequality,

$$
\mathbb{E}\sup_{n\geq 1} M_n \leq \mathbb{E}\sup_{n\geq 1} Y_n^2 \leq 4 \sup_{n\geq 1} \mathbb{E}Y_n^2 < \infty.
$$

Letting $M_* = \sup_{n\geq 1} M_n$, which is in L_1 by the above, we have $M_n \leq M_*$ showing that (M_n) is uniformly integrable (Lemma I.1). Thus (iii) holds, hence (ii), too. Of course, (ii) and $\mathbb{E}M_n = 1$ implies (i).

Suppose (iv) does not hold, that is $\prod_{k=1}^n a_k \to 0$. Then, since $Y_n \to Y_\infty$ a.s. for some integrable random variable Y_∞ (Y is a nonnegative martingale!), we have

$$
M_n = Y_n^2 \left(\prod_{k=1}^n a_k\right)^2 \xrightarrow[n\to\infty]{a.s.} 0
$$

that is $M_\infty = 0$ a.s. Consequently, neither (i) nor (ii) do hold. $\square$

In particular, when the X_i are i.i.d., unless they are constant a.s., the product $X_1 \cdot \ldots \cdot X_n$ converges a.s. but not in L_1, to $M_\infty = 0$ a.s. (see also Exercise 13.15).

15.4 The law of the iterated logarithm for Gaussians

15.7 Theorem. *Let $X_1, X_2, \ldots$ be i.i.d. standard Gaussian random variables and set $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Then*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1 \quad a.s.$$

and

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1 \quad a.s.$$

Proof. Let

$$h(x) = \sqrt{2x\log\log x}, \qquad x > e.$$

The statement about lim inf follows from the one about lim sup by symmetry ($-S_n$ has the same distribution as S_n). To prove the latter, we split the argument into two parts.

Upper bound. First we exploit Doob's maximal inequality, to get an exponential bound on tail probabilities for the maximum. Fix $\lambda > 0$. We have,

$$\mathbb{E}e^{\lambda S_n} = e^{\frac{1}{2}\lambda^2 n}$$

(because $S_n \sim N(0,n)$). Moreover, $(e^{\lambda S_n})_{n\geq 1}$ is a submartingale because $x \mapsto e^x$ is convex. By Doob's maximal inequality 14.5, for every t, we have

$$\mathbb{P}\left(\max_{k\leq n} S_k \geq t\right) = \mathbb{P}\left(\max_{k\leq n} e^{\lambda S_k} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\mathbb{E}e^{\lambda S_n} = e^{-\lambda t + \frac{1}{2}\lambda^2 n}$$

and optimising over λ yields

$$\mathbb{P}\left(\max_{k\leq n} S_k \geq t\right) \leq e^{-\frac{t^2}{2n}}.$$

Fix $\alpha > 1$ and let $a_n = \alpha h(\alpha^{n-1})$. Since

$$\begin{aligned}\mathbb{P}\left(\max_{k\leq \alpha^n} S_k \geq a_n\right) \leq e^{-\frac{a_n^2}{2\alpha^n}} &= \exp\left\{-\frac{\alpha^2 \cdot 2\alpha^{n-1}\log\log\alpha^{n-1}}{2\alpha^n}\right\} \\ &= \exp\left\{-\alpha\log\log\alpha^{n-1}\right\} \\ &= (n-1)^{-\alpha}e^{-\alpha\log\log\alpha},\end{aligned}$$

the series $\sum \mathbb{P}\left(\max_{k\leq\alpha^n} S_k \geq a_n\right)$ converges and by the first Borel-Cantelli lemma, a.s., $\max_{k\leq\alpha^n} S_k \geq a_n$ holds only for finitely many n's. Consequently, the event

$$\{\max_{k\leq\alpha^n} S_k < a_n \quad \text{for all but finitely many } n\text{'s}\}$$

has probability 1. This invent is contained in the event

$$U_\alpha = \{\exists n_0 \ \forall n \geq n_0 \ \forall \alpha^{n-1} \leq k \leq \alpha^n \quad S_k \leq \alpha h(k)\}$$

because of the monotonicity of $h(x)$, so U_α has probability 1. On the event U_α,

$$\limsup_{n\to\infty} \frac{S_k}{h(k)} \leq \alpha,$$

thus on the event $\bigcap_{l=2}^{\infty} U_{1+1/l}$, which also has probability 1,

$$\limsup_{n\to\infty} \frac{S_k}{h(k)} \leq 1,$$

hence

$$\limsup_{n\to\infty} \frac{S_k}{h(k)} \leq 1 \quad \text{a.s.}$$

Lower bound. We shall need the following elementary estimate on Gaussian tails (cf. Lemma 11.4).

Claim. If g is a standard Gaussian random variable, then for $t > 0$,

$$\mathbb{P}\left(g > t\right) \geq \frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2} e^{-t^2/2}.$$

Indeed,

$$\begin{aligned}
\mathbb{P}\left(g > t\right) = \int_t^\infty e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} &\geq \int_t^\infty \frac{1+x^{-2}}{1+t^{-2}} e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}} \\
&= \frac{1}{1+t^{-2}} \int_t^\infty \left(-\frac{1}{x} e^{-x^2/2}\right)' \frac{\mathrm{d}x}{\sqrt{2\pi}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2} e^{-t^2/2}.
\end{aligned}$$

We fix $\varepsilon \in (0,1)$ and an integer $N > 1$. We consider the events

$$A_n = \{S_{N^{n+1}} - S_{N^n} > (1-\varepsilon) h(N^{n+1} - N^n)\}.$$

Since $S_{N^{n+1}} - S_{N^n}$ has the same distribution as $\sqrt{N^{n+1} - N^n} g$, where g is standard Gaussian, by the claim, we get

$$\begin{aligned}
\mathbb{P}\left(A_n\right) &= \mathbb{P}\left(g > (1-\varepsilon)\sqrt{2\log\log(N^{n+1} - N^n)}\right) \\
&\geq \frac{1}{\sqrt{2\pi}} \frac{(1-\varepsilon)\sqrt{2\log\log(N^{n+1} - N^n)}}{1 + 2(1-\varepsilon)^2 \log\log(N^{n+1} - N^n)} \left(\log N^n(N-1)\right)^{-(1-\varepsilon)^2} \\
&= \Omega(n^{-1})
\end{aligned}$$

provided N is large enough. This gives $\sum \mathbb{P}\left(A_n\right) = \infty$ and of course, the events A_n are independent, so by the second Borel-Cantelli lemma, infinitely many A_n occur with probability 1. In other words,

$$\mathbb{P}\left(S_{N^{n+1}} > (1-\varepsilon) h(N^{n+1} - N^n) + S_{N^n}, \quad \text{for infinitely many } n\text{'s}\right) = 1.$$

By the upper bound, a.s., $S_{N^n} > -2h(N^n)$ eventually, so the events

$$V_{N,\varepsilon} = \{S_{N^{n+1}} > (1-\varepsilon)h(N^{n+1}-N^n) - 2h(N^n), \quad \text{for infinitely many } n\text{'s}\}$$

have probability 1. On $V_{N,\varepsilon}$, we have

$$\begin{aligned}\limsup \frac{S_n}{h(n)} \geq \limsup \frac{S_{N^{n+1}}}{h(N^{n+1})} &\geq \limsup \left((1-\varepsilon)\frac{h(N^{n+1}-N^n)}{h(N^{n+1})} - 2\frac{h(N^n)}{h(N^{n+1})}\right) \\ &= \limsup \left((1-\varepsilon)\sqrt{\frac{2N^n(N-1)\log\log N^n(N-1)}{2N^{n+1}\log\log N^{n+1}}}\right. \\ &\qquad \left. - 2\sqrt{\frac{2N^n \log\log N^n}{2N^{n+1}\log\log N^{n+1}}}\right) \\ &= (1-\varepsilon)\sqrt{\frac{N-1}{N}} - \frac{2}{\sqrt{N}}.\end{aligned}$$

Therefore, on $\bigcap_{N=2}^{\infty}\bigcap_{l=2}^{\infty} V_{N,1/l}$, we have

$$\limsup \frac{S_n}{h(n)} \geq 1.$$

□

15.8 Remark. We have,

$$\frac{S_n}{\sqrt{2n\log\log n}} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

This is very simple: since S_n has the same distribution as $\sqrt{n}g$, where g is standard Gaussian, we have

$$\mathbb{P}\left(\left|\frac{S_n}{\sqrt{2n\log\log n}}\right| > \varepsilon\right) = \mathbb{P}\left(|g| > \varepsilon\sqrt{\log\log n}\right) \to 0.$$

15.9 Remark. Theorem 15.7 can be substantially generalised: it holds for an arbitrary sequence of i.i.d. random variables with mean 0 and variance 1 (the Hartman-Wintner law of the iterated logarithm). There are several proofs (see e.g. [1] for an elementary proof, or [3] for a modern proof using Brownian motion and Donsker's theorem).

15.5 The Radon-Nikodym theorem

Let μ and ν be finite measures on $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** with respect to μ if for every $A \in \mathcal{F}$ with $\mu(A) = 0$, we also have $\nu(A) = 0$. This is sometimes denoted $\nu \ll \mu$. The Radon-Nikodym theorem implies that then ν has a density with respect to μ, that is there is a measurable function $g\colon \Omega \to [0, +\infty)$ such that for every measurable set A,

$$\nu(A) = \int_A g \mathrm{d}\mu.$$

This function is sometimes called the Radon-Nikodym derivative, denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. Clearly, the converse holds as well. We shall present a martingale proof of the Radon-Nikodym theorem.

15.10 Theorem (Radon-Nikodym)**.** *Let μ and ν be finite measures on $(\Omega, \mathcal{F})$. There is an $\mathcal{F}$-measurable function $g\colon \Omega \to [0, +\infty)$ and a set $S \in \mathcal{F}$ such that $\mu(S) = 0$ and*

$$\nu(A) = \nu(A \cap S) + \int_A g \mathrm{d}\mu, \qquad A \in \mathcal{F}.$$

Moreover, g is unique up to sets of μ-measure 0 and S is unique up to sets of $(\mu + \nu)$-measure 0.

15.11 Remark. In particular, if $\nu \ll \mu$, then $\nu(S) = 0$, so

$$\nu(A) = \nu(A \cap S) + \int_A g \mathrm{d}\mu, \qquad A \in \mathcal{F}.$$

15.12 Remark. Considering μ and ν on pieces where they are finite, the theorem extends to the case when μ and ν are σ-finite.

To "construct" g, we will work with sequences in $L_1(\Omega, \mathcal{F}, \mu + \nu)$ indexed by finite sub-σ-algebras of $\mathcal{F}$, so we need to extend a bit notions of convergence to such sequences. The completeness of L_1 will play a crucial role.

Let (E, d) be a metric space. Let T be a directed set, that is a partially ordered set by a relation $\preccurlyeq$ (reflexive, antisymmetric and transitive) with the property that every two elements of T have an upper bound, that is for every $s, t \in T$, there is $u \in T$ with $s \preccurlyeq u$ and $t \preccurlyeq u$. We say that a sequence $(a_t)_{t \in T}$ in E indexed by T converges to $a \in E$ if

$$\forall \varepsilon > 0 \;\; \exists t_0 \in T \;\; \forall t \in T \;\; t \succcurlyeq t_0 \;\Rightarrow\; d(a_t, a) < \varepsilon. \tag{15.1}$$

We say that the sequence $(a_t)_{t \in T}$ satisfies the Cauchy condition (or, simply, is Cauchy) if

$$\forall \varepsilon > 0 \;\; \exists t_0 \in T \;\; \forall t \in T \;\; t \succcurlyeq t_0 \;\Rightarrow\; d(a_t, a_{t_0}) < \varepsilon. \tag{15.2}$$

15.13 Lemma. *Let (E, d) be a complete metric space and let $(T, \preccurlyeq)$ be a directed set.*

(i) If $(a_t)_{t \in T}$ is a sequence in E such that for every nondecreasing sequence of indices $t_1 \preccurlyeq t_2 \preccurlyeq \ldots$, the sequence $(a_{t_n})_{n \geq 1}$ converges, then $(a_t)_{t \in T}$ is Cauchy.

(ii) If $(a_t)_{t \in T}$ is a Cauchy sequence in E, then it converges to some $a \in E$ and there exists a nondecreasing sequence of indices $t_1 \preccurlyeq t_2 \preccurlyeq \ldots$ such that $a_{t_n} \to a$ in E.

Proof. (i): If (a_t) does not satisfy the Cauchy condition, then there is $\varepsilon > 0$ such that for every $t_0 \in T$, there is $t \in T$ with $t \succcurlyeq t_0$ and $d(a_t, a_{t_0}) \geq \varepsilon$. Choose $t_1 \in T$ arbitrarily. Given t_n, define t_{n+1} as the index t given by the previous condition applied to $t_0 = t_n$. We obtain the sequence $t_1 \preccurlyeq t_2 \preccurlyeq \ldots$ with $d(a_{t_n}, a_{t_{n+1}}) \geq \varepsilon$, so $(a_{t_n})_n$ does not converge in E, a contradiction.

(ii): For $n = 1, 2, \ldots$, we apply the Cauchy condition with $\varepsilon = \frac{1}{n}$ and set t_n' to be the index t_0 provided in (15.2). Then we define the sequence t_n recursively, $t_1 = t_1'$ and given t_n, we choose t_{n+1} as any common upper bound of t_n and t'_{n+1}. This way we obtain the sequence of indices $t_1 \preccurlyeq t_2 \preccurlyeq \dots$ such that for every $n \geq$ and every $t \in T$ with $t \succcurlyeq t_n$, we have

$$d(a_{t_n}, a_t) \leq d(a_{t_n}, a_{t'_n}) + d(a_{t'_n}, a_t) < \frac{2}{n}.$$

This shows that the sequence $(a_{t_n})_n$ is Cauchy. Since E is complete, it converges to, say $a \in E$. It remains to show that the whole sequence $(a_t)_{t\in T}$ also converges to a in the sense of (15.1). Fix $\varepsilon > 0$. Let n be such that $\varepsilon > \frac{1}{n}$ and $d(a_{t_n}, a) < \varepsilon$. Choose $t_0 = t_n$. Then for every $t \in T$ with $t \succcurlyeq t_0$, we have

$$d(a_t, a) \leq d(a_t, a_{t_n}) + d(a_{t_n}, a) < \frac{2}{n} + \varepsilon < 3\varepsilon.$$

Thus (15.1) holds and the proof is finished. □

Proof of Theorem 15.10. Let $c = \mu(\Omega) + \nu(\Omega)$. If $c = 0$, there is nothing to prove, so we assume $c > 0$. Let $\mathbb{P} = \frac{1}{c}(\mu + \nu)$, so that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. We shall denote the integral against $\mathbb{P}$ by $\mathbb{E}$, that is

$$\mathbb{E}f = \int_\Omega f \frac{\mathrm{d}(\mu + \nu)}{c}, \qquad f\colon \Omega \to \mathbb{R},\ f \text{ is measurable.}$$

Step I (martingale argument). We define

$$\begin{aligned} T = \{\mathcal{G} \subset \mathcal{F} : \mathcal{G} &\text{ is a finite sub-}\sigma\text{-algebra, i.e.} \\ &\mathcal{G} = \sigma(A_1, \dots, A_n) \text{ for some } A_1, \dots, A_n \in \mathcal{F}\} \end{aligned}$$

Equipped with the inclusion relation $\subset$, this is a directed set (a common upper bound for $\mathcal{G}_1, \mathcal{G}_2 \in T$ is simply $\sigma(\mathcal{G}_1, \mathcal{G}_2)$). We set

$$E = L_1(\Omega, \mathcal{F}, \mathbb{P})$$

which is a complete metric space (see Theorem 6.10). For $\mathcal{G} \in T$ generated by atoms $A_1, \dots, A_n$ (meaning that $\Omega = A_1 \cup \dots \cup A_n$ is a disjoint partition and every set in $\mathcal{G}$ is of the form $\bigcup_{i\in I} A_i$ for a subset I of $\{1, \dots, n\}$), we define

$$X_\mathcal{G}(\omega) = \begin{cases} \frac{\nu(A_j)}{\mathbb{P}(A_j)}, & \text{for } \omega \in A_j,\ \text{if } \mathbb{P}(A_j) > 0 \\ 0, & \text{for } \omega \in A_j,\ \text{if } \mathbb{P}(A_j) = 0 \end{cases}.$$

Note that

(a) $0 \leq X_\mathcal{G} \leq c$,

(b) $X_\mathcal{G}$ is the density of ν with respect to $\mathbb{P}$ on $\mathcal{G}$, that is

$$\nu(A) = \mathbb{E}X_\mathcal{G}\, \mathbf{1}_A, \qquad \text{for every } A \in \mathcal{G},$$

(c) For every sequence $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$, $\mathcal{G}_n \in T$, $n = 1, 2, \ldots$, letting $X_n = X_{\mathcal{G}_n}$, we have

$$(X_n)_{n\geq 1} \quad \text{is a martingale on } (\Omega, \mathcal{F}, \{\mathcal{G}_n\}_{n\geq 1}, \mathbb{P}).$$

This is because for every $n \geq 1$ and every $A \in \mathcal{G}_n$, thanks to (a), we have

$$\mathbb{E}X_{n+1}\,\mathbf{1}_A = \nu(A) = \mathbb{E}X_n\,\mathbf{1}_A,$$

or, equivalently, $\mathbb{E}(X_{n+1}|\mathcal{G}_n) = X_n$.

Moreover, since (X_n) is bounded (by virtue of (a)), (X_n) is a uniformly integrable martingale (Lemma I.1), so it converges a.s. and in L_1.

By (c) and Lemma 15.13, the sequence $(X_{\mathcal{G}})_{\mathcal{G}\in T}$ converges in L_1 (in the sense of (15.1)) to a random variable $X \in L_1$. Moreover, there is a sequence $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$ in T such that $X_{\mathcal{G}_n} \to X$ in L_1. As a uniformly integrable martingale, this sequence $(X_{\mathcal{G}_n})$ also converges a.s. and in L_1 to some L_1 random variable, say X_∞. By the uniqueness of limits in L_1, $X_\infty = X$ and by the a.s. convergence, $X \in [0, c]$ a.s.

Step II (limit argument). The idea is of course that because each $X_{\mathcal{G}}$ is the density of ν with respect to $\mathbb{P}$ on $\mathcal{G}$, the random variable X constructed in the previous step as the limit of $X_{\mathcal{G}}$, should be the density of ν with respect to $\mathbb{P}$ on the whole $\mathcal{F}$.

Formally, fix $\varepsilon > 0$. Since $X_{\mathcal{G}} \to X$ (in the sense of (15.1)), there is $\mathcal{K} \in T$ such that for every $\mathcal{G} \supset \mathcal{K}$, we have

$$\mathbb{E}|X_{\mathcal{G}} - X| < \varepsilon.$$

Fix $A \in \mathcal{F}$ and let $\mathcal{G} = \sigma(\mathcal{K}, A)$. Since $\mathcal{G} \supset \mathcal{K}$, we have

$$|\mathbb{E}X_{\mathcal{G}}\,\mathbf{1}_A - \mathbb{E}X\,\mathbf{1}_A| \leq \mathbb{E}|X_{\mathcal{G}} - X|\,\mathbf{1}_A < \varepsilon.$$

By (b), $\mathbb{E}X_{\mathcal{G}}\,\mathbf{1}_A = \nu(A)$, so we obtain

$$|\nu(A) - \mathbb{E}X\,\mathbf{1}_A| < \varepsilon,$$

hence

$$\nu(A) = \mathbb{E}X\,\mathbf{1}_A$$

(because ε is arbitrary). In view of $\mathbb{E}X\,\mathbf{1}_A = \frac{1}{c}\int_A X\mathrm{d}\mu + \frac{1}{c}\int_A X\mathrm{d}\nu$, we equivalently have

$$\int_A (c - X)\mathrm{d}\nu = \int_A X\mathrm{d}\mu,$$

for every $A \in \mathcal{F}$. By a standard argument of complicating measurable functions, we also obtain from this that

$$\int_A (c - X)f\mathrm{d}\nu = \int_A Xf\mathrm{d}\mu, \tag{15.3}$$

for every $A \in \mathcal{F}$ and every $\mathcal{F}$-measurable function $f\colon \Omega \to \mathbb{R}$.

Step III (derivation of density). It remains to define S and g. We set

$$S = \{X = c\}.$$

Choosing $A = S$, $f = 1$ in (15.3) yields

$$0 = \int_S (c - X)\mathrm{d}\nu = \int_S X\mathrm{d}\mu = \int_S c\mathrm{d}\mu = c\mu(S),$$

hence $\mu(S) = 0$. Recall that $X \in [0, c]$ $\mathbb{P}$-a.s., so in particular we have, $\mu(\{X > c\}) = \nu(\{X > c\}) = 0$. Applying now (15.3) with

$$f(\omega) = \begin{cases} \frac{1}{c - X(\omega)}, & \omega \in S^c, \\ 0, & \omega \in S, \end{cases}$$

we get

$$\nu(A \cap S^c) = \int_A (c - X) f\mathrm{d}\nu = \int_A X f\mathrm{d}\mu = \int_{A \cap S^c} \frac{X}{c - X}\mathrm{d}\mu.$$

Therefore, we define

$$g = \begin{cases} \frac{X}{c - X}, & \text{on } S^c, \\ 0, & \text{on } S, \end{cases}$$

and the previous identity becomes

$$\nu(A \cap S^c) = \int_A g\mathrm{d}\mu.$$

Finally,

$$\nu(A) = \nu(A \cap S) + \nu(A \cap S^c) = \nu(A \cap S) + \int_A g\mathrm{d}\mu,$$

as desired.

Step IV (uniqueness). Suppose we have another set $\tilde{S}$ and function $\tilde{g}$ satisfying the required properties. Then, for every $A \in \mathcal{F}$,

$$\nu(A \cap S) + \int_A g\mathrm{d}\mu = \nu(A \cap \tilde{S}) + \int_A \tilde{g}\mathrm{d}\mu.$$

Taking $A = S$ gives $\nu(S) = \nu(S \cap \tilde{S})$ (because $\int_S g\mathrm{d}\mu = 0 = \int_S \tilde{g}\mathrm{d}\mu$, as $\mu(S) = 0$). By symmetry, $\nu(\tilde{S}) = \nu(S \cap \tilde{S})$, hence $\nu(S \triangle \tilde{S}) = 0$. Thus S is unique up to sets of $(\mu + \nu)$-measure 0. In particular, now we know that $\nu(A \cap S) = \nu(A \cap \tilde{S})$, so

$$\int_A g\mathrm{d}\mu = \int_A \tilde{g}\mathrm{d}\mu,$$

for every $A \in \mathcal{F}$. As a result, $g = \tilde{g}$ μ-a.e. $\square$

15.6 Exercises

1. Let $\varepsilon_1, \varepsilon_2, \ldots$ be i.i.d. symmetric random signs, that is $\mathbb{P}\left(\varepsilon_k = 1\right) = \frac{1}{2} = \mathbb{P}\left(\varepsilon_k = -1\right)$ for every k. Then $\sum_{n=1}^{\infty} a_n \varepsilon_n$ converges a.s. if and only if $\sum_{n=1}^{\infty} a_n^2 < \infty$.

2. Let $X_1, X_2, \ldots$ be i.i.d. exponential random variables with parameter 1. Let $(a_n)_{n \geq 1}$ be a sequence of nonnegative numbers. Show that $\sum a_n X_n$ converges a.s. if and only if $\sum a_n < \infty$.

3. If $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}X_1^2 < \infty$ and $\mathbb{E}X_1 = 0$, then for every $\varepsilon > 0$, we have
$$\frac{X_1 + \cdots + X_n}{n^{1/2} \log^{1/2+\varepsilon} n} \xrightarrow[n\to\infty]{a.s.} 0.$$
In other words, $\frac{X_1+\cdots+X_n}{n} = o\left(\frac{\log^{1/2+\varepsilon} n}{n^{1/2}}\right)$ a.s., giving the rate of convergence in the strong law large numbers under the assumption that $\mathbb{E}X_1^2 < \infty$.

4. Prove Marcinkiewicz's theorem: *if $p \in (0,2)$ and $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}|X_1|^p < \infty$, then*
$$\frac{X_1 + \cdots + X_n - n\mu}{n^{1/p}} \xrightarrow[n\to\infty]{a.s.} 0,$$
where $\mu = 0$ for $p \in (0,1)$ and $\mu = \mathbb{E}X_1$ for $p \in [1,2)$.

   In other words, for $p \in (1,2)$, $\frac{X_1+\cdots+X_n}{n} - \mathbb{E}X_1 = o(n^{1/p-1})$ a.s. which is the strong law of large numbers with the rate of convergence.

5. Let μ, ν be two finite measures on $(\Omega, \mathcal{F})$. Show that ν is absolutely continuous with respect to μ if and only if for every $\varepsilon > 0$, there is $\delta > 0$ such that for every $A \in \mathcal{F}$ with $\mu(A) < \delta$, we have $\nu(A) < \varepsilon$.

16 Large deviations

The weak law of large numbers tells us that for every $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{S_n}{n} - \mu > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

where $S_n = X_1 + \ldots + X_n$, $X_1, X_2, \ldots$ are i.i.d. integrable random variables and $\mu = \mathbb{E}X_1$. How fast do these probabilities converge to 0?

Large deviations theory answers this question and establishes, under additional assumptions, precise exponential rates of convergence. Cramér's theorem provides a convex function $I\colon (\mu, +\infty) \to (0, +\infty)$ (determined by the distribution of X_1) such that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} > a\right) = -I(a), \qquad a > \mu,$$

so that, roughly, $\mathbb{P}\left(\frac{S_n}{n} > a\right) \approx e^{-nI(a)}$.

We stress out that as opposed to the c e n t r a l limit theorem, which identifies the limiting behaviour of the probabilities for the bulk (centre), that is a narrow window of width $O(\frac{1}{\sqrt{n}})$ around the mean,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{n} - \mu \in \left(\frac{\sigma a}{\sqrt{n}}, \frac{\sigma b}{\sqrt{n}}\right)\right) = \int_a^b e^{-x^2/2} \frac{\mathrm{d}x}{\sqrt{2\pi}}, \qquad a < b,$$

l a r g e deviations treat the limiting behaviour of the probabilities for the tail, a constant away from the mean,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{S_n}{n} - \mu > \varepsilon\right) = -I(\mu + \varepsilon), \qquad \varepsilon > 0.$$

We end this introduction with a simple lemma showing that as a consequence of independence, such limit always exists. The ultimate goal would be to determine its value.

16.1 Lemma. *Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables. For every $a \in \mathbb{R}$, the limit*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) \in [-\infty, 0]$$

exists. It equals $-\infty$ if and only $\mathbb{P}\left(X_1 \geq a\right) = 0$.

Proof. Fix a. Let $b_n = \log \mathbb{P}\left(S_n \geq na\right) \in [-\infty, 0]$, $n \geq 1$. Note that for $n \geq m \geq 1$,

$$b_{n+m} \geq \log \mathbb{P}\left(S_n - S_m \geq (n-m)a,\ S_m \geq ma\right),$$

so, thanks to independence,

$$b_{n+m} \geq b_n + b_m,$$

that is the sequence (b_n) is subadditive and hence $\lim_{n\to\infty} \frac{b_n}{n}$ exists and it is equal to $\sup_{m\geq 1} \frac{b_m}{m} \in [-\infty, 0]$, as explained in the next basic lemma. The last part is left as an exercises. $\square$

16.2 Lemma. *Let* $(b_n)_{n\geq 1}$ *be a sequence of real numbers which is subadditive, that is for every* $m, n \geq 1$,

$$b_{m+n} \geq b_m + b_n.$$

Then $\lim_{n\to\infty} \frac{b_n}{n}$ *exits and equals* $\sup_{n\geq 1} \frac{b_n}{n} \in (-\infty, +\infty]$.

Proof. Let $a = \sup_{n\geq 1} \frac{b_n}{n}$. Plainly, $\limsup \frac{b_n}{n} \leq a$. To show that $\liminf \frac{b_n}{n} \geq a$, it is enough to prove that for every m, we have $\liminf \frac{b_n}{n} \geq \frac{b_m}{m}$. Fix $m \geq 1$. For n, we write $n = km + r$ with $k \geq 0$ and $r \in \{0, 1, \dots, m-1\}$. Iterating the assumption yields $b_n \geq kb_m + b_r$ ($b_0 = 0$), thus, dividing by $n = km + r$,

$$\frac{b_n}{n} \geq \frac{km}{km+r}\frac{b_m}{m} + \frac{b_r}{km+r}.$$

As $n \to \infty$, also $k \to \infty$, so taking $\liminf$ and using that $|b_r| \leq \max_{1\leq j<m} |b_j|$ is bounded gives the desired claim. □

16.1 Moment generating functions

For a random variable X, we define its **moment generating function** $\psi\colon \mathbb{R} \to (0, +\infty)$ as

$$\psi(\lambda) = \mathbb{E}e^{\lambda X}.$$

Note that

$$\psi(0) = 1.$$

16.3 Lemma. *Function* $\log\psi$ *is convex (in other words,* ψ *is log-convex).*

Proof. It is evident from the fact that sums of log-convex functions are log-convex (by Hölder's inequality) and $\lambda \mapsto e^{\lambda X}$ is log-affine. Alternatively, it can be seen by applying Hölder's inequality directly (with weights $1/p = t$, $1/q = 1-t$),

$$\begin{aligned}\log\psi(t\lambda_1 + (1-t)\lambda_2) = \log\mathbb{E}e^{t\lambda_1 X}e^{(1-t)\lambda_2 X} &\leq \log(\mathbb{E}e^{\lambda_1 X})^t(\mathbb{E}e^{\lambda_2 X})^{1-t} \\ &= t\log\psi(\lambda_1) + (1-t)\log\psi(\lambda_2),\end{aligned}$$

for every $\lambda_1, \lambda_2 \in \mathbb{R}$ and $t \in (0,1)$. □

16.4 Corollary. *If for some* $\lambda_1 < \lambda_2$, $\psi(\lambda_1), \psi(\lambda_2) < \infty$, *then* $\psi(\lambda) < \infty$ *for all* $\lambda \in [\lambda_1, \lambda_2]$.

In view of this corollary, it makes sense to define

$$\lambda_- = \inf\{\lambda \in \mathbb{R} :\ \mathbb{E}e^{\lambda X} < \infty\}, \qquad \lambda_+ = \sup\{\lambda \in \mathbb{R} :\ \mathbb{E}e^{\lambda X} < \infty\} \tag{16.1}$$

and then (λ_-, λ_+) is the largest open interval where $\psi < \infty$. Since $\psi(0) = 1$, of course

$$\lambda_- \leq 0 \leq \lambda_+.$$

16.5 Lemma. *Suppose $\psi(\lambda) = \mathbb{E}e^{\lambda X}$ is the moment generating function of a random variable X.*

(i) If $\psi < \infty$ on $(-\delta, \delta)$ for some $\delta > 0$, then $\mathbb{E}|X|^k < \infty$ and $\psi^{(k)}(0) = \mathbb{E}X^k$, for every integer $k \geq 0$.

(ii) If $\psi < \infty$ on $(\lambda_0 - \delta, \lambda_0 + \delta)$ for some $\lambda_0 \in \mathbb{R}$ and $\delta > 0$, then $\mathbb{E}|X|^k e^{\lambda_0 X} < \infty$ and $\psi^{(k)}(\lambda_0) = \mathbb{E}X^k e^{\lambda_0 X}$, for every integer $k \geq 0$.

Proof. (i): Note that $e^{|x|} \leq e^x + e^{-x}$, $x \in \mathbb{R}$. In particular, for $-\delta < h < \delta$, $k \geq 0$,

$$\frac{|h|^k |X|^k}{k!} \leq e^{|hX|} \leq e^{-hX} + e^{hX}$$

and the right hand side is integrable because $\psi(-h), \psi(h) < \infty$. Thus $\mathbb{E}|X|^k < \infty$. Moreover,

$$\left| \sum_{k=0}^{\infty} \frac{(hX)^k}{k!} \right| \leq \sum_{k=0}^{\infty} \frac{|h|^k |X|^k}{k!} = e^{|hX|},$$

so $\sum_{k=0}^{\infty} \frac{(hX)^k}{k!}$ is integrable and by Fubini's theorem,

$$\psi(h) = \mathbb{E} \sum_{k=0}^{\infty} \frac{(hX)^k}{k!} = \sum_{k=0}^{\infty} h^k \frac{EX^k}{k!}.$$

Consequently, ψ is C^∞ on $(-\delta, \delta)$ (as a convergent power series) and $\psi^{(k)}(0) = \mathbb{E}X^k$.

(ii): We shall deduce this part from (i) using the so-called "exponential tilting" of measure, one of the key ideas of large deviations. Since $\psi(\lambda_0) < \infty$, we can define a new random variable Y which is absolutely continuous with respect to X with density $\frac{e^{\lambda_0 X}}{\psi(\lambda_0)}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, that is

$$\mathbb{P}(Y \in A) = \mathbb{E} \frac{e^{\lambda_0 X}}{\psi(\lambda_0)} \mathbf{1}_{X \in A},$$

or, equivalently,

$$\mathrm{d}\mu_Y(x) = \frac{e^{\lambda_0 x}}{\psi(\lambda_0)} \mathrm{d}\mu_X(x).$$

Then,

$$\mathbb{E}f(Y) = \mathbb{E}f(X) \frac{e^{\lambda_0 X}}{\psi(\lambda_0)},$$

for every Borel function f (for which the right hand side exists). In particular, for the moment generating function ψ_Y of Y, we get

$$\psi_Y(\lambda) = \mathbb{E}e^{\lambda Y} = \mathbb{E} \frac{e^{\lambda_0 X}}{\psi(\lambda_0)} e^{\lambda X} = \frac{\psi(\lambda + \lambda_0)}{\psi(\lambda_0)}$$

which is finite on $(-\delta, \delta)$. Applying (i) to Y, we thus get

$$E|Y|^k = \frac{\mathbb{E}|X|^k e^{\lambda_0 X}}{\psi(\lambda_0)} < \infty$$

and

$$\frac{\psi^{(k)}(\lambda_0)}{\psi(\lambda_0)} = \psi_Y^{(k)}(0) = \mathbb{E}Y^k = \mathbb{E}X^k \frac{e^{\lambda_0 X}}{\psi(\lambda_0)},$$

which gives $\psi^{(k)}(\lambda_0) = \mathbb{E}X^k e^{\lambda_0 X}$, as desired. □

16.6 Corollary. *Function ψ is C^∞ on (λ_-, λ_+).*

When $\lambda_- = \lambda_+ = 0$, we will not be able to establish exponential rates of convergence to 0 for probabilities $\mathbb{P}\left(\frac{S_n}{n} > \mu + \varepsilon\right)$ (for a reason – the rates are slower – see exercises). We shall thus make the minimal assumption that at least one of λ_-, λ_+ is nonzero. Since we can always consider $-X$ instead of X, we shall focus on the case when

$$\lambda_+ = \sup\{\lambda \in \mathbb{R} : \ \mathbb{E}e^{\lambda X} < \infty\} > 0. \tag{16.2}$$

As it will turn out, this assumption allows to control upper tails: $\mathbb{P}(S_n \geq an)$ for $a > \mu$.

In view of Corollary 16.6, the next two lemmas concerning continuity of ψ and its derivatives at 0 are trivial when (λ_-, λ_+) contains 0, but since we want to work under the minimal assumption (16.2), it requires some care and extra work when $\lambda_- = 0$.

16.7 Remark. If (16.2) holds, then $\mathbb{E}X_+^k < \infty$, for every $k \geq 1$. In particular,

$$\mathbb{E}X \in [-\infty, +\infty).$$

Proof. Fix $0 < \lambda_0 < \lambda_+$. Since $\frac{(\lambda_0 X_+)^k}{k!} \leq e^{\lambda_0 X_+}$, we have $\mathbb{E}X_+^k < \infty$. $\square$

16.8 Lemma. *Suppose ψ is the moment generating function of a random variable X and* (16.2) *holds. Then*

(i) ψ is continuous at 0*,*

(ii) $\lim_{\lambda \to 0+} \psi^{(k)}(\lambda) = \mathbb{E}X^k$*, for every $k \geq 1$.*

Proof. (i): We need to show that $\lim_{\lambda \to 0+} \psi(\lambda) = \psi(0)$, that is $\lim_{\lambda \to 0+} \mathbb{E}e^{\lambda X} = 1$, which will of course follow if we can change the order of taking the limit and expectation. Fix $0 < \lambda_0 < \lambda_+$ and note that

$$e^{\lambda X} \leq 1 + e^{\lambda_0 X}, \qquad 0 < \lambda < \lambda_0$$

(this holds because if $X < 0$, then $e^{\lambda X} \leq 1$ and if $X \geq 0$, then $e^{\lambda X} \leq e^{\lambda_0 X}$, by monotonicity). Lebesgue's dominated convergence theorem finishes the argument.

(ii): Following the same argument, we want to dominate $X^k e^{\lambda X}$ for all $0 < \lambda < \lambda_0$ by an integrable random variable. We write $X^k = X_+^k - X_-^k$. For small enough $\varepsilon > 0$,

$$\frac{\varepsilon^k X_+^k}{k!} \leq e^{\varepsilon X_+} \leq 1 + e^{\varepsilon X},$$

so that $X_+^k e^{\lambda X}$ is dominated by $(1 + e^{\varepsilon X})(1 + e^{\lambda_0 X})$ which is integrable provided that $\lambda_0 + \varepsilon < \lambda_+$. Hence,

$$\lim_{\lambda \to 0+} \mathbb{E}X_+^k e^{\lambda X} = \mathbb{E}X_+^k.$$

Now we analyse $X_-^k e^{\lambda X}$. If $\mathbb{E}X_-^k = +\infty$, Fatou's lemma gives

$$\liminf_{\lambda\to 0+} \mathbb{E}X_-^k e^{\lambda X} \geq \mathbb{E}\liminf_{\lambda\to 0+} X_-^k e^{\lambda X} = \mathbb{E}X_-^k = +\infty.$$

As a result, in this case,

$$\lim_{\lambda\to 0+} \psi^{(k)}(\lambda) = \lim_{\lambda\to 0+} \mathbb{E}X^k e^{\lambda X} = -\infty = \mathbb{E}X^k.$$

If $\mathbb{E}X_-^k < \infty$, then since

$$X_-^k e^{\lambda X} \leq X_-^k(1 + e^{\lambda_0 X}) = X_-^k + X_-^k e^{\lambda_0 X}$$

as well as

$$X_-^k e^{\lambda_0 X} \leq \frac{k!}{\varepsilon^k} e^{\varepsilon X_-} e^{\lambda_0 X} = \frac{k!}{\varepsilon^k} e^{\varepsilon X_+ + (\lambda_0 - \varepsilon)X} \leq \frac{k!}{\varepsilon^k}(1 + e^{\varepsilon X})e^{(\lambda_0-\varepsilon)X},$$

we can conclude by Lebesgue's dominated convergence theorem. $\square$

16.9 Remark. A similar argument shows that $\psi(\lambda) \to \psi(\lambda_+)$ as $\lambda \to \lambda_+-$: we write $e^{\lambda X} = e^{\lambda X}\,\mathbf{1}_{\{X\geq 0\}} + e^{\lambda X}\,\mathbf{1}_{\{X<0\}}$ and use Lebesgue's monotone convergence theorem for the first term and Lebesgue's dominated convergence theorem for the second one.

We close this section with a relationship between moment generating functions and tilted measures.

16.10 Lemma. *Let X be a random variable with moment generating function ψ and let $\lambda_\pm$ be given by* (16.1). *For $\lambda \in (\lambda_-, \lambda_+)$, let μ_λ be the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by*

$$\mu_\lambda(A) = \int_A \frac{e^{\lambda x}}{\psi(\lambda)}\mathrm{d}\mu_X(x), \qquad A \in \mathcal{B}(\mathbb{R}),$$

where μ_X is the distribution of X. Let Y_λ be a random variable with distribution μ_λ. Then

$$(\log\psi)'(\lambda) = \mathbb{E}Y_\lambda,$$
$$(\log\psi)''(\lambda) = \mathrm{Var}(Y_\lambda).$$

Proof. By the definition of Y_λ,

$$\mathbb{E}f(Y_\lambda) = \mathbb{E}f(X)\frac{e^{\lambda X}}{\psi(\lambda)},$$

for every measurable function f for which the right hand side exists. Thus, by Lemma 16.5,

$$(\log\psi)'(\lambda) = \frac{\psi'(\lambda)}{\psi(\lambda)} = \frac{\mathbb{E}Xe^{\lambda X}}{\psi(\lambda)} = \mathbb{E}Y_\lambda$$

and

$$(\log\psi)''(\lambda) = \frac{\psi''(\lambda)}{\psi(\lambda)} - \left(\frac{\psi'(\lambda)}{\psi(\lambda)}\right)^2 = \frac{\mathbb{E}X^2e^{\lambda X}}{\psi(\lambda)} - (\mathbb{E}Y_\lambda)^2 = \mathbb{E}Y_\lambda^2 - (\mathbb{E}Y_\lambda)^2 = \mathrm{Var}(Y_\lambda).$$

$\square$

16.2 Upper bounds: Chernoff's inequality

16.11 Lemma (Chernoff's bound)**.** *Let X be a random variable with moment generating function ψ and let $a \in \mathbb{R}$. Then*

$$\mathbb{P}(X \geq a) \leq e^{-(a\lambda - \log\psi(\lambda))}, \qquad \lambda \geq 0.$$

Proof. The assertion follows from exponential Chebyshev's inequality, since for $\lambda > 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda a}\right) \leq e^{-\lambda a}\mathbb{E}e^{\lambda X} = e^{-(\lambda a - \log\psi(\lambda)}.$$

□

16.12 Corollary. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with moment generating function ψ, $S_n = X_1 + \ldots + X_n$. For $a \in \mathbb{R}$,*

$$\mathbb{P}(S_n \geq an) \leq \exp\left\{-n\sup_{\lambda>0}\{\lambda a - \log\psi(\lambda)\}\right\}.$$

If (16.2) *holds, then for every $a > \mu = \mathbb{E}X_1 \in [-\infty, +\infty)$, we have*

$$\sup_{\lambda>0}\{\lambda a - \log\psi(\lambda)\} > 0,$$

that is the above upper bound is meaningful (and is exponentially small in n).

Proof. The upper bound on $\mathbb{P}(S_n \geq an)$ follows from Chernoff's bound applied to $X = S_n$. Independence yields, $\psi_{S_n}(\lambda) = \psi(\lambda)^n$ and the supremum appears because the bound holds for all $\lambda > 0$.

If (16.2) holds, then by the intermediate value theorem applied to $\log\psi$ on $(0, \lambda)$, we have

$$a\lambda - \log\psi(\lambda) = a\lambda - (\log\psi(\lambda) - \log\psi(0)) = \lambda \cdot [a - (\log\psi)'(\theta)],$$

for some $\theta \in (0, \lambda)$. Since $(\log\psi)'(\theta) = \frac{\psi'(\theta)}{\psi(\theta)} \to \mathbb{E}X = \mu$ as $\lambda \to 0$ (Lemma 16.8), for small λ, the expression in the square bracket is close to $a - \mu > 0$, which shows that the supremum $\sup_{\lambda>0}\{\lambda a - \log\psi(\lambda)\}$ is positive, as desired. □

The above upper bound motivates the following definition: the **rate function** $I\colon \mathbb{R} \to [0, +\infty]$ of a random variable X with moment generating function ψ is defined as

$$I(a) = \sup_{\lambda\in\mathbb{R}}\{\lambda a - \log\psi(\lambda)\}, \qquad a \in \mathbb{R}.$$

(It is the Legendre transform of the log-moment generating function $\log\psi$.) As a pointwise supremum of linear functions, I is a convex function.

16.13 Example. For the common distributions, in some cases, the rate function can be written down explicitly (see exercises).

1) Standard Gaussian distribution, $X \sim N(0,1)$. We have

$$\psi(\lambda) = \mathbb{E}e^{\lambda X} = e^{\lambda^2/2}$$

Then, given $a \in \mathbb{R}$, $\lambda a - \log \psi(\lambda)$ is maximised over $\lambda \in \mathbb{R}$ when

$$a = (\log \psi)'(\lambda) = \lambda,$$

which gives

$$I(a) = a^2 - \log \psi(a) = \frac{a^2}{2}.$$

2) Standard exponential distribution, $X \sim \mathrm{Exp}(1)$. We have

$$\psi(\lambda) = \begin{cases} \frac{1}{1-\lambda}, & \lambda < 1, \\ +\infty, & \lambda \geq 1 \end{cases}$$

and

$$I(a) = \begin{cases} a - 1 - \log a, & a > 0, \\ +\infty, & a \leq 0. \end{cases}$$

3) Bernoulli distribution, $X \sim \mathrm{Ber}(p)$. We have

$$\psi(\lambda) = 1 - p + pe^{\lambda}$$

and

$$I(a) = \begin{cases} a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}, & 0 < a < 1, \\ -\log(1-p), & a = 0, \\ -\log p, & a = 1, \\ +\infty, & a < 0 \text{ or } a > 1. \end{cases}$$

4) Poisson distribution, $X \sim \mathrm{Poiss}(\mu)$. We have

$$\psi(\lambda) = \exp\{-\mu + \mu e^{\lambda}\}$$

and

$$I(a) = \begin{cases} +\infty, & a < 0, \\ \mu, & a = 0, \\ a \log \frac{a}{e\mu} + \mu, & a > 0. \end{cases}$$

16.14 Example. Let X be a random variable with density $g(x) = Cx^{-3}e^{-x}\,\mathbf{1}_{[1,+\infty)}(x)$, where C is a normalising constant. We have

$$\psi(\lambda) = C \int_1^{\infty} x^{-} e^{\lambda x - x} \mathrm{d}x < \infty$$

if and only if $\lambda \leq 1$, so $\lambda_- = -\infty$, $\lambda_+ = 1$. Note that as $\lambda \to 1-$,

$$\frac{\psi'(\lambda)}{\psi(\lambda)} \nearrow \frac{\psi'(1)}{\psi(1)} = \frac{\int_1^\infty x^{-2}\mathrm{d}x}{\int_1^\infty x^{-3}\mathrm{d}x} = 2.$$

Thus, the supremum in the definition of $I(a)$ is attained if and only if $a \leq 2$, the case when the equation $a = \frac{\psi'(\lambda)}{\psi(\lambda)}$ has a solution.

16.3 Cramér's theorem

16.15 Theorem. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with moment generating function ψ satisfying* (16.2) *and rate function I. Let $\mu = \mathbb{E}X_1$ and $S_n = X_1 + \ldots + X_n$, $n \geq 1$. For every $a > \mu$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) = -I(a).$$

Proof. We begin with two remarks which will help us understand better the claimed value of the limit $-I(a)$. By its definition, the rate function $I(a)$ is $\sup_{\lambda\in\mathbb{R}} f(\lambda)$ with

$$f(\lambda) = \lambda a - \log\psi(\lambda).$$

First, we remark that under our assumptions, in fact we have

$$I(a) = \sup_{0<\lambda<\lambda_+} f(\lambda). \tag{16.3}$$

Plainly $\lambda > \lambda_+$ results in $f(\lambda) = -\infty$, so those λ do not count in the supremum. Moreover, note that by Jensen's inequality,

$$f(\lambda) = \lambda a - \log \mathbb{E}e^{\lambda X_1} \leq \lambda a - \log e^{\lambda\mathbb{E}X_1} = \lambda(a-\mu).$$

For $a > \mu$ and $\lambda \leq 0$, the above is thus nonpositive. If (16.2) holds, we know that the supremum over $\lambda > 0$ is positive (Corollary 16.12). Thus $\lambda \leq 0$ can also be neglected in the supremum defining $I(a)$.

Second, thanks to Lemma 16.10,

$$f''(\lambda) = -(\log\psi)''(\lambda) = -\operatorname{Var}(Y_\lambda)$$

which is strictly negative (unless Y_λ, equivalently X is a point mass, in which case there is nothing to do). Thus, f' is strictly increasing and f is strictly concave on $(0, \lambda_+)$. In particular, if f attains its supremum, it is unique, attained at $\lambda = \lambda_a \in (0, \lambda_+)$ which is a unique solution of the equation

$$f'(\lambda) = 0, \qquad \text{that is} \qquad a = (\log\psi)'(\lambda) = \frac{\psi'(\lambda)}{\psi(\lambda)}.$$

We now break the proof into two parts.

Upper bound. By Chernoff's bound,

$$\limsup \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) \leq -\sup_{\lambda>0}\{\lambda a - \log \psi(\lambda)\} = -I(a),$$

where the last equality is justified by (16.3).

Lower bound. Fix $a > \mu$. It remains to show that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) \geq -I(a) \tag{16.4}$$

with the right hand side given by (16.3).

Claim. If there is $\lambda_a \in (0, \lambda_+)$ such that $a = \frac{\psi'(\lambda_a)}{\psi(\lambda_a)}$, then lower bound (16.4) holds.

To prove the claim, fix $\lambda_a < \lambda < \lambda_+$. Let $Y_1, \ldots, Y_n$ be i.i.d. copies of the tilted random variable Y_λ from Lemma 16.10. Note that then, thanks to independence, the vector $(Y_1, \ldots, Y_n)$ has density $\frac{e^{\lambda(X_1+\ldots+X_n)}}{\psi(\lambda)^n}$ with respect to $(X_1, \ldots, X_n)$ and we have

$$\mathbb{E}f(Y_1, \ldots, Y_n) = \mathbb{E}f(X_1, \ldots, X_n)\frac{e^{\lambda S_n}}{\psi(\lambda)^n}.$$

for every measurable function $f \colon \mathbb{R}^n \to \mathbb{R}$ for which the right hand side exists. For $a' > a$, we thus have

$$\begin{aligned}\mathbb{P}\left(S_n \geq na\right) \geq \mathbb{P}\left(S_n \in (na, na')\right) &\geq \mathbb{E}e^{\lambda S_n}e^{-\lambda na'}\mathbf{1}_{\{S_n\in(na,na')\}}\\ &= \psi(\lambda)^n e^{-\lambda na'}\mathbb{E}\,\mathbf{1}_{\{Y_1+\cdots+Y_n\in(na,na')\}}\\ &= \psi(\lambda)^n e^{-\lambda na'}\mathbb{P}\left(Y_1 + \cdots + Y_n \in (na, na')\right).\end{aligned}$$

By the weak law of large numbers,

$$\frac{Y_1 + \cdots + Y_n - n\mathbb{E}Y_1}{n} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

so as long as $a' > \mathbb{E}Y_1 > a$, we get

$$\begin{aligned}&\mathbb{P}\left(Y_1 + \cdots + Y_n \in (na, na')\right)\\ &\qquad = \mathbb{P}\left(\frac{Y_1 + \cdots + Y_n - n\mathbb{E}Y_1}{n} \in (a - \mathbb{E}Y_1, a' - EY_1)\right) \xrightarrow[n\to\infty]{} 1\end{aligned}$$

Since $\mathbb{E}Y_1 = \frac{\psi'(\lambda)}{\psi(\lambda)} > \frac{\psi'(\lambda_a)}{\psi(\lambda_a)} = a$, given $\lambda > \lambda_a$, we thus fix a' such that $a' > \frac{\psi'(\lambda)}{\psi(\lambda)}$ and get

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) \geq -(\lambda a' - \log \psi(\lambda)).$$

Letting $\lambda \searrow \lambda_a$ and then $a' \searrow a$, we get

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq na\right) \geq -(\lambda_a a - \log \psi(\lambda_a)) = -I(a),$$

as desired. This finishes the proof of the claim.

Depending on the behaviour of ψ, the claim may or may not be applicable. To decide when it is, we consider

$$A = \operatorname{esssup} X \in (-\infty, +\infty].$$

Case 1: $A < +\infty$. Then $\lambda_+ = +\infty$. First we claim that

$$\frac{\psi'(\lambda)}{\psi(\lambda)} = \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}} \xrightarrow[\lambda\to\infty]{} A. \tag{16.5}$$

Indeed, since $X \leq A$, clearly $\frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}} \leq A$. On the other hand, for $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}Xe^{\lambda X} &\geq (A-\varepsilon)\mathbb{E}e^{\lambda X}\,\mathbf{1}_{\{X>A-\varepsilon\}} + \mathbb{E}Xe^{\lambda X}\,\mathbf{1}_{\{X\leq A-\varepsilon\}} \\
&= (A-\varepsilon)\mathbb{E}e^{\lambda X} + \mathbb{E}(X-(A-\varepsilon))e^{\lambda X}\,\mathbf{1}_{\{X\leq A-\varepsilon\}} \\
&= (A-\varepsilon)\mathbb{E}e^{\lambda X} + \frac{1}{\lambda}e^{\lambda(A-\varepsilon)}\mathbb{E}\lambda(X-(A-\varepsilon))e^{\lambda(X-(A-\varepsilon))}\,\mathbf{1}_{\{X\leq A-\varepsilon\}}.
\end{aligned}$$

Since $|ye^y\,\mathbf{1}_{y\leq 0}| \leq e^{-1}$ and

$$\frac{e^{\lambda(A-\varepsilon)}}{\mathbb{E}e^{\lambda X}} \leq \frac{e^{\lambda(A-\varepsilon)}}{\mathbb{E}e^{\lambda X}\,\mathbf{1}_{\{X>A-\varepsilon/2\}}} \leq \frac{e^{-\lambda\varepsilon/2}}{\mathbb{P}(X > A-\varepsilon/2)},$$

we obtain

$$\frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}} \geq A - \varepsilon - \frac{1}{\lambda}e^{-\lambda\varepsilon/2}\frac{1}{e\mathbb{P}(X > A-\varepsilon/2)},$$

thus

$$\liminf_{\lambda\to\infty} \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}} \geq A - \varepsilon.$$

Consequently, (16.5) holds.

In view of (16.5), if $a < A$, then there is λ_a with $\frac{\psi'(\lambda_a)}{\psi(\lambda_a)} = a$ and the claim finishes the proof in this case. If $a > A$, then trivially $\mathbb{P}(S_n \geq an) = 0$ for every n, so it remains to argue that $I(a) = +\infty$. This holds because $f'(\lambda) > a - A > 0$ for every $\lambda > 0$ (as f' is strictly decreasing), so $f(\lambda) \to \infty$ as $\lambda \to \infty$. Finally, if $a = A$, then $\mathbb{P}(S_n \geq an) = \mathbb{P}(X_1 = A)^n$, so it remains to argue that $I(A) = -\log\mathbb{P}(X_1 = A)$. On one hand, for every $\lambda > 0$, $\psi(\lambda) = \mathbb{E}e^{\lambda X_1} \geq e^{\lambda A}\mathbb{P}(X_1 = A)$, so

$$I(A) = \sup_{\lambda>0}\{\lambda A - \log\psi(\lambda)\} \leq -\log\mathbb{P}(X_1 = A).$$

On the other hand, this upper bound is attained in the limit as $\lambda \to \infty$ because by Lebesgue's dominated convergence theorem,

$$\lambda A - \psi(\lambda) = -\log\mathbb{E}e^{\lambda(X-A)} \xrightarrow[\lambda\to\infty]{} -\log\mathbb{P}(X_1 = A),$$

which finishes the whole argument in this case.

Case 2: $A = +\infty$. If $\lambda_+ = \infty$, then the proof of the lower bound in (16.5) shows that

$$\frac{\psi'(\lambda)}{\psi(\lambda)} \to \infty \quad \text{as } \lambda \to \infty,$$

so regardless of the value of a, the claim is applicable. Suppose now that $\lambda_+ < \infty$. Let

$$\alpha = \lim_{\lambda \to \lambda_+} \frac{\psi'(\lambda)}{\psi(\lambda)}.$$

Note that then, thanks to the monotonicity of $(\log \psi)' = \frac{\psi'}{\psi}$,

$$\log \psi(\lambda_+) = \int_0^{\lambda_+} (\log \psi)' \leq \lambda_+ \alpha < \infty$$

Writing $e^{\lambda X} = \varepsilon^{\lambda X} \mathbf{1}_{\{X>0\}} + e^{\lambda X} \mathbf{1}_{\{X \leq 0\}}$, by Lebesgue's monotone and dominated convergence theorems we see that in fact

$$\psi(\lambda) = \mathbb{E}e^{\lambda X_1} \to \mathbb{E}e^{\lambda_+ X_1} = \psi(\lambda_+)$$

and

$$\psi'(\lambda) = \mathbb{E}Xe^{\lambda X_1} \to \mathbb{E}Xe^{\lambda_+ X_1} = \psi'(\lambda_+)$$

as $\lambda \to \lambda_+$. It remains to consider the case when $\alpha < \infty$ and $a \geq \alpha$ (otherwise, again, the claim is applicable). We have,

$$I(a) = a\lambda_+ - \log \psi(\lambda_+), \qquad a \geq \alpha,$$

($I(a)$ is linear). Indeed, $f'(\lambda) = a - (\log \psi)'(\lambda) > a - \alpha \geq 0$, for every $\lambda < \lambda_+$, so f is strictly increasing, hence $I(a) = \sup_{0<\lambda<\lambda_+} f(\lambda) = f(\lambda_+) = a\lambda_+ - \log \psi(\lambda_+)$. We fix $a \geq \alpha$ and our goal is to show that

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq an\right) \geq -(a\lambda_+ - \log \psi(\lambda_+)).$$

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with the law given by the tilted measure μ_λ from Corollary 16.6 with $\lambda = \lambda_+$, so that $\mathbb{E}Y_1 = \alpha$. We proceed as in the proof of the claim: for $a' > a$, we have

$$\mathbb{P}\left(S_n \geq an\right) \geq \psi(\lambda_+)^n e^{-n\lambda_+ a'} \mathbb{P}\left(\sum_{k=1}^n Y_k \in (an, a'n)\right).$$

Using independence, for $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(\sum_{k=1}^n Y_k \in (an, a'n)\right) \\ &\qquad \geq \mathbb{P}\left(\sum_{k=1}^{n-1} Y_k \in ((\alpha - \varepsilon)n, (\alpha + \varepsilon)n)\right) \mathbb{P}\left(Y_n \in ((a - \alpha + \varepsilon)n, (a' - \alpha - \varepsilon)n)\right). \end{aligned}$$

By the weak law of large numbers, the first term is at least, say $\frac{1}{2}$ for large n. Choosing $a' = a + 3\varepsilon$ and using that $(a' - \alpha - \varepsilon)n = (a - \alpha + 2\varepsilon)n > (a - \alpha + \varepsilon)(n+1)$ for large n, we thus get

$$\begin{aligned} \frac{1}{n} \log \mathbb{P}\left(S_n \geq an\right) \geq &- \left((a + 3\varepsilon)\lambda_+ - \log \psi(\lambda_+)\right) \\ &+ \frac{1}{n} \log \frac{1}{2} + \frac{1}{n} \log \mathbb{P}\left(Y_1 \in (a - \alpha + \varepsilon)n, (a - \alpha + \varepsilon)(n+1)\right), \end{aligned}$$

for large n. Finally, if we had

$$\limsup \frac{1}{n} \log \mathbb{P}\left(Y_1 \in (a-\alpha+\varepsilon)n, (a-\alpha+\varepsilon)(n+1)\right) < 0,$$

then for some small enough $\delta > 0$, the series

$$\sum_{n=1}^{\infty} e^{\delta(a-\alpha+\varepsilon)(n+1)} \mathbb{P}\left(Y_1 \in (a-\alpha+\varepsilon)n, (a-\alpha+\varepsilon)(n+1)\right)$$

would converge, so

$$\mathbb{E}e^{\delta Y_1} < \infty$$

and, equivalently, $\mathbb{E}e^{(\delta+\lambda_+)X} < \infty$ contradicting the definition of λ_+. Therefore the lim sup above is zero and thus

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geq an\right) \geq -((a+3\varepsilon)\lambda_+ - \log\psi(\lambda_+))$$

for every $\varepsilon > 0$. By virtue of Lemma 16.1, this lim sup is the same as lim inf because the limit in fact exists, so we get the desired lower bound after letting $\varepsilon \to 0$. $\square$

16.4 Quantitative bounds

We present a result asserting that for sums of i.i.d. random variables, with probabilities exponentially close to 1 (in n), $\frac{S_k}{k}$ is ε-close to its mean for *all* $k \geq n$. It relies on maximal inequalities combined with the basic idea used in Chernoff's bounds. It can be thought of as a quantitative version of the strong law of large numbers (cf. Exercises 15.3 and 15.4).

16.16 Theorem. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with moment generating function ψ such that $\psi < \infty$ on $(-\delta, \delta)$ for some $\delta > 0$. Let I be the rate function of X_1. Let $\mu = \mathbb{E}X_1$. Then for every $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\sup_{k\geq n}\left|\frac{S_k}{k} - \mu\right| > \varepsilon\right) \leq 2e^{-n\min\{I(\mu-\varepsilon), I(\mu+\varepsilon)\}}.$$

Proof. Fix $\lambda > 0$ and a such that $\lambda a - \log\psi(\lambda) \geq 0$. Observe that

$$\begin{aligned}
\mathbb{P}\left(\sup_{k\geq n}\frac{S_k}{k} > a\right) &= \mathbb{P}\left(\exists k \geq n:\ S_k > ak\right) \\
&= \mathbb{P}\left(\exists k \geq n:\ e^{\lambda S_k - k\log\psi(\lambda)} > e^{k(\lambda a - \log\psi(\lambda))}\right) \\
&\leq \mathbb{P}\left(\exists k \geq n:\ e^{\lambda S_k - k\log\psi(\lambda)} > e^{n(\lambda a - \log\psi(\lambda))}\right) \\
&= \mathbb{P}\left(\sup_{k\geq n} e^{\lambda S_k - k\log\psi(\lambda)} > e^{n(\lambda a - \log\psi(\lambda))}\right).
\end{aligned}$$

Since $(e^{\lambda S_k - k\log\psi(\lambda)})_{k\geq n}$ is a martingale (as the product of independent random variables with mean 1), by Doob's maximal inequality (14.5),

$$\mathbb{P}\left(\sup_{k\geq n}\frac{S_k}{k} > a\right) \leq e^{-n(\lambda a - \log\psi(\lambda))}.$$

Using this with $a = \mu + \varepsilon$ and optimising over $\lambda > 0$ with $a\lambda - \log \psi(\lambda) > 0$ (recall (16.3)) gives

$$\mathbb{P}\left(\sup_{k\geq n} \frac{S_k}{k} > \mu + \varepsilon\right) \leq e^{-nI(\mu+\varepsilon)}.$$

To finish the proof, note that

$$\left\{\sup_{k\geq n}\left|\frac{S_k}{k} - \mu\right| > \varepsilon\right\} = \left\{\sup_{k\geq n}\left(\frac{S_k}{k} - \mu\right) > \varepsilon\right\} \cup \left\{\sup_{k\geq n}\left(\mu - \frac{S_k}{k}\right) > \varepsilon\right\}$$

so it remains to apply the above inequality to $-X_1, -X_2, \ldots$, to get

$$\mathbb{P}\left(\sup_{k\geq n}\left|\frac{S_k}{k} - \mu\right| > \varepsilon\right) \leq e^{-nI(\mu+\varepsilon)} + e^{-nI(\mu-\varepsilon)} \leq 2e^{-n\min\{I(\mu-\varepsilon), I(\mu+\varepsilon)\}}.$$

□

The last two subsections are related to large deviations by methods, rather than by the topic itself. We shall see how moment generating functions also play a key role in establishing nonasymptotic bounds in estimating the expected value of maxima of random variables as well as (large) deviation inequalities for sums of independent random variables.

16.5 Bounds on the expected maximum of random variables

We begin with a technical lemma which summarises the properties of the Legendre transform and discusses the inverse function.

16.17 Lemma. *Let $\delta > 0$ and let $h\colon [0, \delta) \to \mathbb{R}$ be a C^1 convex, nondecreasing function with $h(0) = h'(0) = 0$. We define its Legendre transform,*

$$h^*(a) = \sup_{\lambda\in(0,\delta)} \{\lambda a - h(\lambda)\}, \qquad a \geq 0.$$

Then h^* is a nonnegative convex nondecreasing function. Moreover, for every $b \geq 0$, the set $\{a \geq 0 : h^*(a) > b\}$ is nonempty and for the "generalised inverse" function of h^*,

$$(h^*)^{-1}(b) = \inf\{a \geq 0 : h^*(a) > b\},$$

we have

$$(h^*)^{-1}(b) = \inf_{\lambda\in(0,\delta)} \frac{b + h(\lambda)}{\lambda}.$$

Proof. Note that h^* is defined as a pointwise supremum of nondecreasing linear functions, hence it is nondecreasing and convex. By the assumptions $h(x) \geq 0$, $x \in [0, \delta)$, so $h^*(0) = \sup_{\lambda\in(0,\delta)} -h(\lambda) = 0$. By monotonicity, $h^*(a) \geq h^*(0) = 0$, for every $a \geq 0$. If we fix $\lambda_0 \in (0, \delta)$, then $h^*(a) \geq \lambda_0 a - h(\lambda_0)$ and the right hand side as a function of a

is unbounded. This explains why the sets $\{a \geq 0 : h^*(a) > b\}$ are nonempty. Finally,

$$\begin{aligned} h^*(a) > b \qquad &\Leftrightarrow \qquad \exists \lambda \in (0,\delta)\ \lambda a - h(\lambda) > b \\ &\Leftrightarrow \qquad \exists \lambda \in (0,\delta)\ a > \frac{b + h(\lambda)}{\lambda} \\ &\Leftrightarrow \qquad a > \inf_{\lambda \in (0,\delta)} \frac{b + h(\lambda)}{\lambda}. \end{aligned}$$

This shows that the set $\{a \geq 0 : h^*(a) > b\}$ is the half-line $(\inf_{\lambda \in (0,\delta)} \frac{b+h(\lambda)}{\lambda}, +\infty)$ and the claimed formula for $(h^*)^{-1}$ follows. $\square$

16.18 Theorem. *Let $X_1, \ldots, X_n$ be random variables such that for some $\delta > 0$ and a C^1 convex nondecreasing function $h\colon [0,\delta) \to \mathbb{R}$, we have*

$$\log \mathbb{E} e^{\lambda X_i} \leq h(\lambda), \qquad \lambda \in [0,\delta),$$

for every $i \leq n$. Then

$$\mathbb{E}\left(\max_{i \leq n} X_i\right) \leq (h^*)^{-1}(\log n),$$

where $(h^*)^{-1}$ is defined in Lemma 16.17.

Proof. Using a simple inequality $\max a_i \leq \sum a_i$ valid for nonnegative numbers a_i and Jensen's inequality, for every $\lambda \in (0,\delta)$, we have

$$\mathbb{E}\left(\max_{i \leq n} X_i\right) = \frac{1}{\lambda} \mathbb{E} \log \max_{i \leq n} e^{\lambda X_i} \leq \frac{1}{\lambda} \mathbb{E} \log\left(\sum_{i=1}^n e^{\lambda X_i}\right) \leq \frac{1}{\lambda} \log\left(\mathbb{E} \sum_{i=1}^n e^{\lambda X_i}\right).$$

By the assumption,

$$\mathbb{E} \sum_{i=1}^n e^{\lambda X_i} \leq n e^{h(\lambda)},$$

so

$$\mathbb{E}\left(\max_{i \leq n} X_i\right) \leq \frac{\log n + h(\lambda)}{\lambda}.$$

Taking the infimum over $\lambda \in (0,\delta)$, in view of the formula for $(h^*)^{-1}$ from Lemma 16.17, we get

$$\mathbb{E}\left(\max_{i \leq n} X_i\right) \leq (h^*)^{-1}(\log n).$$

$\square$

As an example, we apply this to Gaussian random variables. The above upper bound, as crude as it seems, can be matched from below giving the correct behaviour for large n in the independent case.

16.19 Theorem. *There are positive universal constants c, C such that if $X_1, X_2, \ldots$ are Gaussian random variables, each one with mean 0, we have*

$$\mathbb{E}\left(\max_{i \leq n} X_i\right) \leq C \cdot \sqrt{\max_{i \leq n} \mathrm{Var}(X_i)} \cdot \sqrt{\log n} \tag{16.6}$$

and if additionally the X_i are independent identically distributed,

$$\mathbb{E}\left(\max_{i\le n} X_i\right) \ge c\cdot\sqrt{\mathrm{Var}(X_1)}\cdot\sqrt{\log n}. \tag{16.7}$$

Proof. For the upper bound, we let $\sigma = \sqrt{\max_{i\le n}\mathrm{Var}(X_i)}$ and since

$$\mathbb{E}e^{\lambda X_i} = e^{\mathrm{Var}(X_i)\lambda^2/2} \le e^{\sigma^2\lambda^2/2}, \qquad \lambda\in\mathbb{R},$$

we simply use Theorem 16.18 with $\delta=\infty$ and $h(\lambda)=\frac{\sigma^2\lambda^2}{2}$. We have

$$\inf_{\lambda>0}\frac{\log n + h(\lambda)}{\lambda} = \inf_{\lambda>0}\left(\frac{\log n}{\lambda}+\frac{\sigma^2}{2}\lambda\right) = \sigma\sqrt{2\log n},$$

which proves (16.6) with $C=\sqrt{2}$.

For the lower bound, first of all, by homogeneity, we can assume that the X_i are standard Gaussian, that is with mean 0 and variance 1. When $n=1$, the assertion is trivial. When $n=2$, we have

$$\mathbb{E}\max\{X_1,X_2\} \ge \mathbb{E}\,\mathbf{1}_{\{X_1,X_2>1\}} = \mathbb{P}\left(X_1>1\right)^2.$$

When dealing with $n\ge 2$, because of this and $\mathbb{E}\max_{i\le n}X_i \ge \mathbb{E}\max_{i\le 2}X_i$, in what follows we can assume that n is large enough. Exploiting symmetry reduces our task to estimating the maximum of absolute values. Indeed,

$$\mathbb{E}\max_{i\le n}|X_i| \le \mathbb{E}|X_1| + \mathbb{E}\max_{i\le n}|X_i-X_1| \le \mathbb{E}|X_1| + \mathbb{E}\max_{i,j\le n}|X_i-X_j|,$$

but $\max_{i,j\le n}|X_i-X_j| = \max_{i,j\le n}(X_i-X_j)$ (pointwise), so

$$\mathbb{E}\max_{i,j\le n}|X_i-X_j| = \mathbb{E}\max_{i,j\le n}(X_i-X_j) \le \mathbb{E}\max_{i\le n}X_i + \mathbb{E}\max_{j\le n}(-X_j) = 2\mathbb{E}\max_{i\le n}X_i,$$

where the last equality holds because of the symmetry of the X_i. Altogether,

$$\mathbb{E}\max_{i\le n}X_i \ge \frac{1}{2}\left(\sqrt{\frac{2}{\pi}} + \mathbb{E}\max_{i\le n}|X_i|\right).$$

It suffices to show that

$$\mathbb{E}\max_{i\le n}|X_i| \ge c\sqrt{\log n}.$$

Using independence and monotonicity, we obtain for any $a>0$,

$$\begin{aligned}\mathbb{E}\max_{i\le n}|X_i| = \int_0^\infty \mathbb{P}\left(\max_{i\le n}|X_i|>t\right)\mathrm{d}t &= \int_0^\infty \left[1-\mathbb{P}\left(|X_1|\le t\right)^n\right]\mathrm{d}t \\ &\ge \int_0^a \left[1-\mathbb{P}\left(|X_1|\le t\right)^n\right]\mathrm{d}t \\ &\ge a\left[1-\mathbb{P}\left(|X_1|\le a\right)^n\right].\end{aligned}$$

We choose $a=\sqrt{\log n}$. To estimate the probability $\mathbb{P}\left(|X_1|\le a\right)$, we can use the precise Gaussian tail bound (the claim from the proof of Theorem 15.7), or just for simplicity, crudely,

$$\mathbb{P}\left(|X_1|>a\right) = \frac{1}{\sqrt{2\pi}}\int_a^\infty e^{-t^2/2}\mathrm{d}t > \frac{1}{\sqrt{2\pi}}\int_a^{a\sqrt{2}} e^{-t^2/2}\mathrm{d}t > \frac{\sqrt{2}-1}{\sqrt{2\pi}}ae^{-a^2} > e^{-a^2},$$

for a, equivalently, n large enough, which gives

$$\mathbb{P}\left(|X_1| \le a\right)^n \le \left(1 - e^{-a^2}\right)^n = \left(1 - \frac{1}{n}\right)^n \le e^{-1}$$

and

$$\mathbb{E}\max_{i\le n}|X_i| \ge (1-e^{-1})\sqrt{\log n},$$

for n large enough. □

We finish with two remarks left as exercises.

16.20 Remark. The lower bound can be substantially improved to produce an asymptotically exact result

$$\mathbb{E}\max_{i\le n} X_i = (1+o(1))\sqrt{2\log n},$$

for i.i.d. standard Gaussians X_i.

16.21 Remark. If $X_1, \dots, X_n$ are independent Gaussians with mean 0 and variances $\sigma_1^2 \ge \dots \sigma_n^2$, then

$$c \cdot \max_{k\le n}\sigma_k\sqrt{\log(1+k)} \le \mathbb{E}\max_{i\le n} X_i \le C\cdot\max_{k\le n}\sigma_k\sqrt{\log(1+k)}$$

with some positive universal constants c and C.

16.6 A flavour of concentration inequalities

Concentration inequalities concern upper bounds on probabilities $\mathbb{P}\left(|X-a| > \varepsilon\right)$, where a is usually the mean or a median of X (anti-concentration inequalities seek upper bounds on $\sup_a \mathbb{P}\left(|X-a| < \varepsilon\right)$). We have seen a very simple example of a concentration inequality, namely Chebyshev's inequality $\mathbb{P}\left(|X-\mathbb{E}X| > \varepsilon\right) \le \frac{\mathrm{Var}(X)}{\varepsilon^2}$. Another examples are in Exercises 6.19, 6.21, 14.7. To merely give a flavour of concentration inequalities for sums of independent random variables, we shall discuss a basic result for the so-called sub-exponential random variables. We begin with a motivating example.

16.22 Example. Let $X_1, \dots, X_n$ be i.i.d. exponential random variables with parameter 1. Let $X = \frac{1}{n}\sum_{i=1}^n X_i$ which has mean 1. We would like to upper bound $\mathbb{P}\left(X - 1 > \varepsilon\right)$ for $\varepsilon > 0$. Since $\sum_{i=1}^n X_i$ has the distribution Gamma with parameter n, we have an exact expression

$$\mathbb{P}\left(X-1>\varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^n X_i > (1+\varepsilon)n\right) = \int_{(1+\varepsilon)n}^\infty \frac{x^{n-1}}{n!}e^{-x}\mathrm{d}x.$$

However, there is no closed expression for this integral. Chernoff's bound from Lemma 16.11 yields

$$\mathbb{P}\left(X-1>\varepsilon\right) \le \exp\left\{-n\sup_{\lambda>0}(1+\varepsilon)\lambda - \log\mathbb{E}e^{\lambda X_1}\right\} = \exp\left\{-n(\varepsilon - \log(1+\varepsilon)\right\}.$$

From large deviations theory (Cramér's theorem), we know this bound is asymptotically tight because $\frac{1}{n}\log\mathbb{P}\left(X-1>\varepsilon\right)\to -I(1+\varepsilon)$ with the rate function $I(a)=a-1-\log(1+a)$. We point out two features of this bound.

1) As $\varepsilon\to 0$, $\varepsilon-\log(1+\varepsilon)\approx\frac{\varepsilon^2}{2}$ which means, roughly, that X has a Gaussian tail in this regime (which is not surprising because of the central limit theorem).

2) As $\varepsilon\to\infty$, $\varepsilon-\log(1+\varepsilon)\approx\varepsilon$ which means, roughly, that X has an exponential tail in this regime.

We shall present a result which will capture such behaviours in a much greater generality than just for sums of i.i.d. exponentials.

For a random variable X, we define its ψ_1**-norm** as

$$\|X\|_{\psi_1}=\inf\left\{t>0,\ \mathbb{E}e^{|X/t|}\leq 2\right\}.$$

We say that X is **sub-exponential** if $\|X\|_{\psi_1}<\infty$. The following lemma helps decide whether a particular distribution is sub-exponential (we defer its proof to exercises).

16.23 Lemma. *For a random variable X, the following conditions are equivalent.*

(i) X is sub-exponential.

(ii) There are constants $c_1,c_2>0$ such that $\mathbb{P}\left(|X|>t\right)\leq c_1e^{-c_2t}$, for every $t>0$.

(iii) There is a constant $C>0$ such that $\|X\|_p\leq Cp$, for every $p\geq 1$.

(iv) There are constants $\delta,C>0$ such that $\mathbb{E}e^{\lambda|X|}\leq e^{c\lambda}$, for every $\lambda\in(0,\delta)$.

We shall need yet another characterisation of sub-exponentiality saying that for centred random variables, the moment generating function near the origin is bounded by the Gaussian one.

16.24 Lemma. *If X is a random variable with mean 0, then*

$$\mathbb{E}e^{\lambda X}\leq e^{\lambda^2\|X\|_{\psi_1}^2},\qquad |\lambda|<\frac{1}{\|X\|_{\psi_1}}.$$

Proof. If $\|X\|_{\psi_1}=\infty$, there is nothing to do. If $\|X\|_{\psi_1}<\infty$, by homogeneity, we can assume that $\|X\|_{\psi_1}=1$, say for simplicity $\mathbb{E}e^{|X|}=2$ (otherwise, by the definition of infimum, for every $\varepsilon>0$, we find $t_0>1$ with $\mathbb{E}e^{|X/t_0|}<1+\varepsilon$). For $|\lambda|\leq 1$, we have

$$\mathbb{E}e^{\lambda X}=\mathbb{E}\sum_{k=0}^{\infty}\frac{\lambda^kX^k}{k!}=1+\lambda\mathbb{E}X+\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}\mathbb{E}X^k}{k!}$$

(the usage of Fubini's theorem is justified because the integrand is majorised by $e^{|X|}$). Since $\mathbb{E}X=0$ and

$$\left|\sum_{k=2}^{\infty}\frac{\lambda^{k-2}\mathbb{E}X^k}{k!}\right|\leq\sum_{k=2}^{\infty}\frac{\mathbb{E}|X|^k}{k!}=e^{|X|}-1-\mathbb{E}|X|=1-\mathbb{E}|X|\leq 1,$$

for $|\lambda| \le 1$, we get

$$\mathbb{E}e^{\lambda X} \le 1 + \lambda^2 \le e^{\lambda^2}.$$

$\square$

16.25 Theorem (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, each one with mean* 0*. Then for every $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{2\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i\le n}\|X_i\|_{\psi_1}}\right\}\right).$$

Proof. For $\lambda > 0$, by exponential Markov's inequality, independence and Lemma 16.24, we get

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le e^{-\lambda t}\mathbb{E}e^{\lambda\sum_{i=1}^n X_i} = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}e^{\lambda X_i} \le e^{-\lambda t}e^{\lambda^2\sum_{i=1}^n \|X_i\|_{\psi_1}^2},$$

provided that $\lambda < \frac{1}{\|X_i\|_{\psi_1}}$ for every $i \le n$, that is $\lambda < \frac{1}{m}$ with $m = \max_{i\le n}\|X_i\|_{\psi_1}$. It remains to optimise over λ. We let $S = \sum_{i=1}^n \|X_i\|_{\psi_1}^2$. The minimum of the function $-\lambda t + \lambda^2 S$ is attained at $\lambda_0 = \frac{t}{2S}$.

Case 1. $\lambda_0 < \frac{1}{m}$, that is $t < \frac{2S}{m}$. Then we set $\lambda = \lambda_0$ and obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le e^{-t^2/(4S)}.$$

Case 2. $\lambda_0 \ge \frac{1}{m}$, that is $t \ge \frac{2S}{m}$. Then we let $\lambda \to \frac{1}{m}$ and obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le e^{-t/m + S/m^2} \le e^{-t/(2m)},$$

where we use that in this case $\frac{S}{m^2} \le \frac{t}{2m}$.

It remains to observe that these two bounds can be concisely written together as

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{2S}, \frac{t}{m}\right\}\right),$$

which is the assertion. $\square$

It is instructive to see what this gives us in the special case of weighted sums of i.i.d. sub-exponential random variables.

16.26 Corollary. *Let $Y_1, \ldots, Y_n$ be independent random variables, each with mean* 0*, sub-exponential with $\|Y_i\|_{\psi_1} \le K$ for all i for some constant $K > 0$. Then for every $a_1, \ldots, a_n \in \mathbb{R}$ and $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n a_i Y_i > t\right) \le \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{2K^2\sum_{i=1}^n a_i^2}, \frac{t}{K\max_{i\le n}|a_i|}\right\}\right).$$

In particular,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i > t\right) \leq \exp\left(-\frac{n}{2}\min\left\{\frac{t^2}{2K^2}, \frac{t}{K}\right\}\right).$$

The latter can be viewed as a quantitative version of the law of large numbers for sub-exponential random variables. These bounds exhibit the mixture of two behaviours of the Gaussian tail (for small t) and the exponential tail (for large t), as anticipated.

16.7 Exercises

1. Under the hypothesis of Lemma 16.1, we have that $\lim \frac{1}{n} \log \mathbb{P}(S_n \geq an) = 0$ if and only if $\mathbb{P}(S_n \geq an) = 0$ for all n, if and only if $\mathbb{P}(X_1 \geq a) = 0$.

2. Find the rate functions claimed in Example 16.13.

3. Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables with $\mathbb{E}X_1 = 0$. Suppose that $\mathbb{E}e^{\lambda X_1} = +\infty$ for all $\lambda > 0$. Then for every $a > 0$, we have

$$\frac{1}{n} \log \mathbb{P}(X_1 + \ldots + X_n \geq an) \xrightarrow[n \to \infty]{} 0.$$

This shows that assumption (16.2) is necessary for the exponential convergence in Cramér's theorem.

4. Prove Remark 16.20.

5. Prove Remark 16.21.

6. Let $\psi\colon [0, +\infty) \to [0, +\infty)$ be a convex strictly increasing function with $\psi(0) = 0$. For a random variable X, define

$$\|X\|_\psi = \inf\{t > 0,\ \mathbb{E}\psi(|X|) \leq 1\}.$$

Show that $\|\lambda X\|_\psi = |\lambda| \|X\|_\psi$, $\lambda \in \mathbb{R}$ and for every two random variables X, Y,

$$\|X + Y\|_\psi \leq \|X\|_\psi + \|Y\|_\psi.$$

(This explains the name "ψ_1-norm", where $\psi(x) = \psi_1(x) = e^{|x|} - 1$. Note that the choice $\psi(x) = |x|^p$ gives the familiar L_p norms.)

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 Ω be a set and let $\mathcal{A}$ be an algebra on Ω. Suppose a function $\mathbb{P}\colon \mathcal{A} \to [0, +\infty)$ satisfies*

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

(ii) $\mathbb{P}$ *is finitely additive, that is for every $A_1, \dots, A_n \in \mathcal{A}$ 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 σ-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 Ω 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 Ω which is a σ-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 Ω, 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 $\varnothing \in \mathcal{A}$, we have $\mathbb{P}^*(\varnothing) = 0$. It is also clear that $\mathbb{P}^*$ is monotone, that is if $A \subset B$, then $\mathbb{P}^*(A) \le \mathbb{P}^*(B)$. Finally, we show that $\mathbb{P}^*$ is subadditive, that is for every sets $A_1, A_2, \ldots$, we have

$$\mathbb{P}^*\left(\bigcup_n A_n\right) \le \sum_n \mathbb{P}^*(A_n).$$

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,

$$\mathbb{P}^*\left(\bigcup_n A_n\right) \le \sum_{n,k} \mathbb{P}(B_{n,k}) < \sum_n \mathbb{P}^*(A_n) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we get the desired inequality.

II. We define the following class of subsets of Ω,

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

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

$$\mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^c \cap E) \leq \mathbb{P}^*(E). \tag{A.1}$$

First we show that $\mathcal{M}$ is an algebra and $\mathbb{P}^*$ 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 Ω, we have

$$\begin{aligned}
\mathbb{P}^*(E) &= \mathbb{P}^*(B \cap E) + \mathbb{P}^*(B^c \cap E) \\
&= \mathbb{P}^*(A \cap B \cap E) + \mathbb{P}^*(A^c \cap B \cap E) + \mathbb{P}^*(A \cap B^c \cap E) + \mathbb{P}^*(A^c \cap B^c \cap E) \\
&\geq \mathbb{P}^*(A \cap B \cap E) + \mathbb{P}^*\Big((A^c \cap B \cap E) \cup (A \cap B^c \cap E) \cup (A^c \cap B^c \cap E)\Big) \\
&= \mathbb{P}^*\Big((A \cap B) \cap E\Big) + \mathbb{P}^*\Big((A \cap B)^c \cap E\Big).
\end{aligned}$$

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

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

$$\mathbb{P}^*(A \cup B) = \mathbb{P}^*\Big(A \cap (A \cup B)\Big) + \mathbb{P}^*\Big(A^c \cap (A \cup B)\Big) = \mathbb{P}^*(A) + \mathbb{P}^*(B).$$

By induction, we easily get the desired finite additivity.

Now we argue that $\mathbb{P}^*$ 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

$$\begin{aligned}
\sum_{k=1}^n \mathbb{P}^*(A_k) &= \mathbb{P}^*\left(\bigcup_{k=1}^n A_k\right) \\
&= \mathbb{P}^*\left(A \cap \bigcup_{k=1}^n A_k\right) \\
&\leq \mathbb{P}^*(A)
\end{aligned}$$

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

It remains to show that $\mathcal{M}$ is a σ-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

$$\mathbb{P}^*(K_n \cap E) = \sum_{k=1}^n \mathbb{P}^*(A_k \cap E).$$

The base case $n = 1$ is clear. Further,

$$\begin{aligned}\mathbb{P}^*(K_{n+1} \cap E) &= \mathbb{P}^*(K_n \cap K_{n+1} \cap E) + \mathbb{P}^*(K_n^c \cap K_{n+1} \cap E) \\ &= \mathbb{P}^*(K_n \cap E) + \mathbb{P}^*(A_{n+1} \cap E) \\ &= \sum_{k=1}^n \mathbb{P}^*(A_k \cap E) + \mathbb{P}^*(A_{n+1} \cap E),\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}^*(E) = \mathbb{P}^*(E \cap K_n) + \mathbb{P}^*(E \cap K_n^c) &\geq \sum_{k=1}^n \mathbb{P}^*(A_k \cap E) + \mathbb{P}^*(E \cap K_n^c) \\ &\geq \sum_{k=1}^n \mathbb{P}^*(A_k \cap E) + \mathbb{P}^*(E \cap A^c),\end{aligned}$$

where the last inequality holds because $\mathbb{P}^*$ is monotone (see I.) and $K_n \subset A$. Letting $n \to \infty$ and using subadditivity, we get

$$\mathbb{P}^*(E) \geq \sum_{k=1}^\infty \mathbb{P}^*(A_k \cap E) + \mathbb{P}^*(E \cap A^c) \geq \mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \cap A^c),$$

so A satisfies (A.1).

III. We show 1) $\mathcal{A} \subset \mathcal{M}$ which also gives $\sigma(\mathcal{A}) \subset \mathcal{M}$ because $\mathcal{M}$ is a σ-algebra. Moreover, we show 2) $\mathbb{P}^* = \mathbb{P}$ on $\mathcal{A}$, so $\mathbb{P}^*$ is the desired extension of $\mathbb{P}$ on $\sigma(\mathcal{A})$. The uniqueness follows immediately from Dynkin's theorem on π-λ systems (see Appendix B and Remark 2.11).

To prove 1), take $A \in \mathcal{A}$ and an arbitrary subset E of Ω. Fix $\varepsilon > 0$. By the definition of $\mathbb{P}^*$, there are sets $B_1, B_2, \ldots \in \mathcal{A}$ such that $E \subset \bigcup_n B_n$ and $\sum_{n=1}^\infty \mathbb{P}(B_n) \leq \mathbb{P}^*(E) + \varepsilon$. Since $E \cap A \subset \bigcup (B_n \cap A)$ and $E \cap A^c \subset \bigcup (B_n \cap A^c)$ and $B_n \cap A, B_n \cap A^c \in \mathcal{A}$, by the definition of $\mathbb{P}^*$,

$$\mathbb{P}^*(E \cap A) \leq \sum_n \mathbb{P}(B_n \cap A)$$

and similarly

$$\mathbb{P}^*(E \cap A^c) \leq \sum_n \mathbb{P}(B_n \cap A^c).$$

Adding these up and using the additivity of $\mathbb{P}$ on $\mathcal{A}$, we get

$$\mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \cap A^c) \leq \sum_n (\mathbb{P}(B_n \cap A) + \mathbb{P}(B_n \cap A^c)) = \sum_n \mathbb{P}(B_n) \leq \mathbb{P}^*(E) + \varepsilon,$$

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 $A \subset_n A_n$ 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 $\mathcal{A}$, $C_1 \subset C_2 \subset \ldots$ 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 $\mathcal{A}$ of subsets of a set Ω is a **π-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 Ω is a **λ-system** if $\Omega \in \mathcal{L}$, for every $A, B \in \mathcal{L}$ with $A \subset B$, we have $B \setminus A \in \mathcal{L}$ and for every $A_1, A_2, \ldots \in \mathcal{L}$ such that $A_1 \subset A_2 \subset \ldots$, we have $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

B.1 Remark. If a family is a π-system and a λ-system, then it is a σ-algebra.

B.2 Theorem (Dynkin)**.** *Let Ω be a set. If a λ-system $\mathcal{L}$ on Ω contains a π-system $\mathcal{A}$ on Ω, then $\mathcal{L}$ contains $\sigma(\mathcal{A})$.*

Proof. Let $\mathcal{L}_0$ be the smallest λ-system containing $\mathcal{A}$. By Remark B.1, it suffices to show that $\mathcal{L}_0$ is a π-system. To this end, we first consider the family

$$\mathcal{C} = \{A \subset \Omega,\ A \cap B \in \mathcal{L}_0 \text{ for every } B \in \mathcal{A}\}.$$

Clearly, $\mathcal{C}$ contains $\mathcal{A}$. Moreover, $\mathcal{C}$ is a λ-system. Indeed,

(i) $\Omega \in \mathcal{C}$ because $\mathcal{A} \subset \mathcal{L}_0$,

(ii) let $U, V \in \mathcal{C}$ with $U \subset V$, then for $B \in \mathcal{A}$,

$$(V \setminus U) \cap B = (U \cap B) \setminus (V \cap B)$$

which is in $\mathcal{L}_0$ because $U \cap B \subset V \cap B$ and $\mathcal{L}_0$ is a λ-system

(iii) let $A_1, A_2, \ldots \in \mathcal{C}$ with $A_1 \subset A_2 \subset \ldots$, then for $B \in \mathcal{A}$, 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 $\mathcal{L}_0$ because $A_i \cap B$ are in $\mathcal{L}_0$ and it is a λ-system.

We thus get that $\mathcal{C}$, as a λ-system containing $\mathcal{A}$, contains the smallest λ-system containing $\mathcal{A}$, that is $\mathcal{L}_0$. This means that $A \cap B \in \mathcal{L}_0$ whenever $A \in \mathcal{L}_0$ and $B \in \mathcal{A}$.

The rest of the proof is a repetition of the same argument. We consider the family

$$\tilde{\mathcal{C}} = \{A \subset \Omega,\ A \cap B \in \mathcal{L}_0 \text{ for every } B \in \mathcal{L}_0\}.$$

By the previous step, we know that $\tilde{\mathcal{C}} \supset \mathcal{A}$. We show that $\tilde{\mathcal{C}}$ is a λ-system, hence, as above, it contains $\mathcal{L}_0$. Therefore, for every $A, B \in \mathcal{L}_0$, $A \cap B \in \mathcal{L}_0$, that is $\mathcal{L}_0$ is a π-system, as required. $\square$

C 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 σ-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,

$$\begin{aligned} A_{\omega_1} &= \{\omega_2 \in \Omega_2,\ (\omega_1, \omega_2) \in A\}, \qquad \omega_1 \in \Omega_1, \\ A^{\omega_2} &= \{\omega_1 \in \Omega_1,\ (\omega_1, \omega_2) \in A\}, \qquad \omega_2 \in \Omega_2. \end{aligned}$$

Similarly, for a function $X \colon \Omega \to \mathbb{R}$, define its section functions

$$\begin{aligned} X_{\omega_1} \colon \Omega_2 \to \mathbb{R}, \qquad & X_{\omega_1}(\omega_2) = X(\omega_1, \omega_2), \qquad \omega_1 \in \Omega_1, \\ X^{\omega_2} \colon \Omega_1 \to \mathbb{R}, \qquad & X^{\omega_2}(\omega_1) = X(\omega_1, \omega_2), \qquad \omega_2 \in \Omega_2. \end{aligned}$$

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*

$$A_{\omega_1} \in \mathcal{F}_2, \qquad A^{\omega_2} \in \mathcal{F}_1.$$

For every $\mathcal{F}$-measurable function $X \colon \Omega \to \mathbb{R}$, every $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, we have

$$X_{\omega_1} \text{ is } \mathcal{F}_2\text{-measurable}, \qquad X^{\omega_2} \text{ is } \mathcal{F}_1\text{-measurable}.$$

If moreover X is nonnegative, we have that

$$\omega_1 \mapsto \int_{\Omega_2} X_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) \quad \text{is } \mathcal{F}_1\text{-measurable}$$

and

$$\omega_2 \mapsto \int_{\Omega_1} X^{\omega_2}(\omega_1) \mathrm{d}\mathbb{P}_1(\omega_1) \quad \text{is } \mathcal{F}_2\text{-measurable}$$

Proof. Let $\mathcal{M}$ be the class of all subsets A of Ω such that for every ω_1, A_{ω_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 π-system generating $\mathcal{F}$. Moreover, it is easy to check that $\mathcal{M}$ is a σ-algebra. Thus $\mathcal{M} \supset \mathcal{F}$. We argue similarly about A^{ω_2}.

To prove the $\mathcal{F}_2$-measurablity of X_{ω_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 $\mathcal{F}_2$ by the previous part because $X^{-1}(B) \in \mathcal{F}$. The $\mathcal{F}_1$-measurability of X^{ω_2}, we proceed in the same way.

Finally, if $X = \mathbf{1}_{B_1 \times B_2}$ for some $B_i \in \mathcal{F}_i$, $i = 1, 2$, we have

$$\int_{\Omega_2} X_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) = \mathbf{1}_{B_1}(\omega_1) \int_{\Omega_2} \mathbf{1}_{B_2}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2),$$

which is clearly $\mathcal{F}_1$-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. $\square$

We define $\mathbb{P} \colon \mathcal{F} \to [0, 1]$ as follows: for $A \in \mathcal{F}$, let $X = \mathbf{1}_A$ and

$$\mathbb{P}(A) = \int_{\Omega_1} \left(\int_{\Omega_2} X_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) \right) \mathrm{d}\mathbb{P}_1(\omega_1).$$

We have the following important result saying that $\mathbb{P}$ is the so-called product measure on Ω.

C.2 Theorem (The uniqueness of product measures)**.** *The set function $\mathbb{P}$ is a unique probability measure on $(\Omega, \mathcal{F})$ such that for every $A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$, we have*

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

Moreover,

$$\mathbb{P}(A) = \int_{\Omega_2} \left(\int_{\Omega_1} X^{\omega_2}(\omega_1) \mathrm{d}\mathbb{P}_1(\omega_1) \right) \mathrm{d}\mathbb{P}_2(\omega_2).$$

Proof. By Lemma C.1, the inner integral in the definition of $\mathbb{P}$ is an $\mathcal{F}_2$-measurable function, thus $\mathbb{P}$ is well defined on $\mathcal{F}$. Clearly, $\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$, so in particular $\mathbb{P}(\Omega) = 1$. If $B_1, B_2, \ldots \in \mathcal{F}$ are disjoint, then so are their sections, that is we have $\mathbf{1}_{\bigcup_n (B_n)_{\omega_1}} = \sum_n \mathbf{1}_{(B_n)_{\omega_1}}$, consequently, by the linearity of integrals, we get that $\mathbb{P}$ is countably-additive. The uniqueness follows from the fact that the product sets $A_1 \times A_2$, $A_i \in \mathcal{F}_i$, form a π-system generating $\mathcal{F}$, combined with Remark 2.11. The formula with the integrals over Ω_1 and Ω_2 swapped follows by considering

$$\tilde{\mathbb{P}}(A) = \int_{\Omega_2} \left(\int_{\Omega_1} X^{\omega_2}(\omega_1) \mathrm{d}\mathbb{P}_1(\omega_1) \right) \mathrm{d}\mathbb{P}_2(\omega_2),$$

checking that $\tilde{\mathbb{P}}$ satisfies the same defining property, $\tilde{\mathbb{P}}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ and using the uniqueness. $\square$

We say that $\mathbb{P}$ is the product of $\mathbb{P}_1$ and $\mathbb{P}_2$, denoted

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

C.3 Theorem (Fubini)**.** *Let $X \colon \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable.*

(i) If $X \geq 0$, then

$$\int_{\Omega_1 \times \Omega_2} X \mathrm{d}\mathbb{P} = \int_{\Omega_1} \left(\int_{\Omega_2} X_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) \right) \mathrm{d}\mathbb{P}_1(\omega_1)$$
$$= \int_{\Omega_2} \left(\int_{\Omega_1} X^{\omega_2}(\omega_1) \mathrm{d}\mathbb{P}_1(\omega_1) \right) \mathrm{d}\mathbb{P}_2(\omega_2).$$

(ii) If

$$\int_{\Omega_1} \left(\int_{\Omega_2} |X_{\omega_1}(\omega_2)| \mathrm{d}\mathbb{P}_2(\omega_2) \right) \mathrm{d}\mathbb{P}_1(\omega_1) < \infty,$$

or

$$\int_{\Omega_2} \left(\int_{\Omega_1} |X^{\omega_2}(\omega_1)| \mathrm{d}\mathbb{P}_1(\omega_1) \right) \mathrm{d}\mathbb{P}_2(\omega_2) < \infty,$$

then

$$\int_{\Omega_1 \times \Omega_2} |X| \mathrm{d}\mathbb{P} < \infty,$$

that is X is $(\Omega, \mathcal{F}, \mathbb{P})$-integrable.

(iii) If X is $(\Omega, \mathcal{F}, \mathbb{P})$-integrable, then

$$\mathbb{P}_1 \left\{ \omega_1 \in \Omega_1, \int_{\Omega_2} |X_{\omega_1}(\omega_2)| \mathrm{d}\mathbb{P}_2(\omega_2) < \infty \right\} = 1,$$
$$\mathbb{P}_2 \left\{ \omega_2 \in \Omega_2, \int_{\Omega_1} |X^{\omega_2}(\omega_1)| \mathrm{d}\mathbb{P}_1(\omega_1) < \infty \right\} = 1$$

and (i) holds.

Proof. (i) By Theorem C.2, the formula holds for $X = \mathbf{1}_A$, $A \in \mathcal{F}$. 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

$$\int_{\Omega} X^+ \mathrm{d}\mathbb{P} < \infty \qquad \text{and} \qquad \int_{\Omega} X^- \mathrm{d}\mathbb{P} < \infty.$$

Thus from (a) applied to X^+,

$$\int_{\Omega} X^+ \mathrm{d}\mathbb{P} = \int_{\Omega_1} \left(\int_{\Omega_2} X^+_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) \right) \mathrm{d}\mathbb{P}_1(\omega_1),$$

which by basic properties of Lebesgue integrals means that

$$\int_{\Omega_2} X^+_{\omega_1}(\omega_2) \mathrm{d}\mathbb{P}_2(\omega_2) < \infty$$

for $\mathbb{P}_1$-a.e. ω_1. Simarly for X^-. Therefore,

$$\mathbb{P}_1 \left\{ \omega_1 \in \Omega_1, \int_{\Omega_2} |X_{\omega_1}(\omega_2)| \mathrm{d}\mathbb{P}_2(\omega_2) < \infty \right\} = 1.$$

In particular, for every ω_1 in this event,

$$\int_{\Omega_2} X_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2) = \int_{\Omega_2} X^+_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2) - \int_{\Omega_2} X^-_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2).$$

For the remaining ω_1, we can set all these integrals to be 0 and then we get

$$\begin{aligned}
\int_{\Omega_1}\left(\int_{\Omega_2} X_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2)\right)\mathrm{d}\mathbb{P}_1(\omega_1) &= \int_{\Omega_1}\left(\int_{\Omega_2} X^+_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2)\right)\mathrm{d}\mathbb{P}_1(\omega_1) \\
&\quad - \int_{\Omega_1}\left(\int_{\Omega_2} X^-_{\omega_1}(\omega_2)\mathrm{d}\mathbb{P}_2(\omega_2)\right)\mathrm{d}\mathbb{P}_1(\omega_1) \\
&= \int_{\Omega_1\times\Omega_2} X^+\mathrm{d}\mathbb{P} - \int_{\Omega_1\times\Omega_2} X^-\mathrm{d}\mathbb{P} \\
&= \int_{\Omega_1\times\Omega_2} X\mathrm{d}\mathbb{P}.
\end{aligned}$$

We proceed in the same way for the swapped order of taking the integrals over Ω_1 and Ω_2. □

Fubini's theorem generalises to σ-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: Infninte products of measures

D.1 Theorem. *Let $\mu_1, \mu_2, \dots$ be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We set*

$$\Omega = \prod_{i=1}^{\infty} \mathbb{R} = \mathbb{R} \times \mathbb{R} \times \dots,$$

$$X_n(\omega_1, \omega_2, \dots) = \omega_n, \qquad (\omega_1, \omega_2, \dots) \in \Omega,$$

and

$$\mathcal{F} = \sigma(X_1, X_2, \dots).$$

There is a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that for every $k \geq 1$ and $A_1, \dots, A_k \in \mathcal{B}(\mathbb{R})$, w have

$$\mathbb{P}\left(A_1 \times \dots \times A_k \times \mathbb{R} \times \dots\right) = \mu_1(A_1) \cdot \ldots \cdot \mu_k(A_k).$$

Moreover, $X_1, X_2, \dots$ are independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mu_{X_i} = \mu_i$.

Proof. For $n \geq 1$, we set

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n).$$

It is a σ-algebra generated by the π-system of the product sets of the form

$$F_n = A_1 \times \dots \times A_n \times \mathbb{R} \times \mathbb{R} \times \dots,$$

where $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$. We consider the algebra

$$\mathcal{A} = \bigcup_{n \geq 1} \mathcal{F}_n$$

along with $\mathbb{P} \colon \mathcal{A} \to [0, 1]$, given by

$$\mathbb{P}\left(F_n\right) = \mu_1(A_1) \cdot \ldots \cdot \mu(A_n)$$

(the product measure). By the construction of the finite product measures (Fubini's theorem), $\mathbb{P}$ is finitely additive on $(\Omega, \mathcal{A})$. Moreover, for each n, $(\Omega, \mathcal{F}_n, \mathbb{P})$ is a probability space and $X_1, \dots, X_n$ are independent. It remains to argue that $\mathbb{P}$ can be extended to a probability measure on $\sigma(\mathcal{A})$ and such an extension will be the desired measure $\mathbb{P}$. Thanks to Carathéodory's theorem, it suffices to verify the condition given in Remark 1.9:

> for every sequence $(H_r)_{r \geq 1}$ of sets in $\mathcal{A}$ with $H_1 \supset H_2 \supset \dots$ such that for some $\varepsilon > 0$, $\mathbb{P}\left(H_r\right) \geq \varepsilon$ for every $r \geq 1$, we have $\bigcap H_r \neq \varnothing$.

We break the argument into several steps.

I. For every r, there is some n_r such that $H_r \in \mathcal{F}_{n_r}$ and then there is an $\mathcal{F}_{n_r}$-measurable bounded function h_r such that

$$\mathbf{1}_{H_r}(\omega) = h_r(\omega_1, \ldots, \omega_{n_r})$$

and

$$\mathbb{E}h_r(X_1, \ldots, X_{n_r}) = \mathbb{P}\left(H_r\right) \geq \varepsilon.$$

II. Define a function $g_r \colon \mathbb{R} \to \mathbb{R}$,

$$g_r(\omega_1) = \mathbb{E}h_r(\omega_1, X_2, \ldots, X_{n_r})$$

(here $\mathbb{E}$ is understood on the probability space $(\Omega, \mathcal{F}_{n_r}, \mathbb{P})$). Since $0 \leq g_r \leq 1$, we have

$$\varepsilon \leq \mathbb{E}h_r = \int g_r \mathrm{d}\mu_1 \leq \mu_1\{g_r \geq \varepsilon/2\} + \frac{\varepsilon}{2}\mu_1\{g_r \leq \varepsilon/2\} \leq \mu_1\{g_r \geq \varepsilon/2\} + \frac{\varepsilon}{2},$$

so

$$\mu_1\{g_r \geq \varepsilon/2\} \geq \frac{\varepsilon}{2}.$$

III. Since $H_r \supset H_{r+1}$, we have $h_r \geq h_{r+1}$, thus

$$g_r(\omega_1) \geq g_{r+1}(\omega_1), \qquad \text{for every } \omega_1 \in \mathbb{R}.$$

This gives that the events $\{g_r \geq \varepsilon/2\}$ decrease, so by the continuity of probability measures and Step II, we get

$$\mu_1\{\forall r \ g_r \geq \varepsilon/2\} \geq \frac{\varepsilon}{2} > 0.$$

Hence, there exists $\omega_1^* \in \mathbb{R}$ such that for every $r \geq 1$,

$$g_r(\omega_1^*) = \mathbb{E}h_r(\omega_1^*, X_2, \ldots, X_{n_r}) \geq \frac{\varepsilon}{2}.$$

IV. Repeating Steps II and III applied to the functions

$$\tilde{g}_r(\omega_2) = \mathbb{E}h_r(\omega_1^*, \omega_2, X_3, \ldots, X_{n_r})$$

yields existence of $\omega_2^* \in \mathbb{R}$ such that for every $r \geq 1$,

$$\mathbb{E}h_r(\omega_1^*, \omega_2^*, X_3, \ldots, X_{n_r}) \geq \frac{\varepsilon}{2^2}.$$

Continuing this procedure (inductively), we obtain an infinite sequence

$$\omega_* = (\omega_1^*, \omega_2^*, \ldots) \in \Omega$$

with the property that for every r,

$$\mathbb{E}h_r(\omega_1^*, \omega_2^*, \omega_3^*, \ldots, \omega_{n_r}^*) = h_r(\omega_1^*, \omega_2^*, \omega_3^*, \ldots, \omega_{n_r}^*) \geq \frac{\varepsilon}{2^{n_r}}.$$

On the other hand,

$$h_r(\omega_1^*, \omega_2^*, \omega_3^*, \ldots, \omega_{n_r}^*) = \mathbf{1}_{H_r}(\omega^*)$$

is either 0 or 1, so it has to be 1, which gives that $\omega_* \in H_r$ and this holds for every r. Therefore, $\bigcap H_r \neq \varnothing$, which shows the desired property allowing to use Carathéodory's theorem and thus finishes the proof. $\square$

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 $\{X \le x\} = \{\omega \in \Omega,\ X(\omega) \le x\} = X^{-1}((-\infty, x])$ 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

$$X = \sum_{k=1}^{n} x_k \, \mathbf{1}_{A_k},$$

for some distinct $x_1, \ldots, x_n \in \mathbb{R}$ (values) and events $A_1, \ldots, A_n$ which form a partition of Ω (we have, $A_k = \{X = x_k\}$).

The **expectation** of the simple random variable X, denoted $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \sum_{k=1}^{n} x_k \mathbb{P}\left(A_k\right).$$

The expectation of a nonnegative random variable X is defined as

$$\mathbb{E}X = \sup\{\mathbb{E}Z,\ Z \text{ is simple and } Z \le X\}.$$

Note that $\mathbb{E}X \ge 0$ because we can always take $Z = 0$. We can have $\mathbb{E}X = +\infty$ (for instance, for a discrete random variable X with $\mathbb{P}\left(X = k\right) = \frac{1}{k(k-1)}$, $k = 2, 3, \ldots$). For an arbitrary random variable X, we write

$$X = X^+ - X^-,$$

where

$$X^+ = \max\{X, 0\} = X \, \mathbf{1}_{\{X \ge 0\}}$$

is the positive part of X and

$$X^- = -\min\{X, 0\} = -X \, \mathbf{1}_{\{X \le 0\}}$$

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.

E.1 Theorem. *Let X and Y be simple random variables. Then $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$.*

Proof. Let $X = \sum_{k=1}^{m} x_k \mathbf{1}_{A_k}$ and $Y = \sum_{l=1}^{n} y_l \mathbf{1}_{B_l}$ for some reals x_k, y_l and events A_k and B_l are such that the A_k partition Ω and the B_l partition Ω. Then the events $A_k \cap B_l$, $k \leq m$, $l \leq n$ partition Ω and

$$X + Y = \sum_{k \leq m, l \leq n} (x_k + y_l) \mathbf{1}_{A_k \cap B_l}.$$

This is a simple random variable with

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{k \leq m, l \leq n} (x_k + y_l) \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m, l \leq n} x_k \mathbb{P}(A_k \cap B_l) + \sum_{k \leq m, l \leq n} y_l \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k \sum_{l \leq n} \mathbb{P}(A_k \cap B_l) + \sum_{l \leq n} y_l \sum_{k \leq m} \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k \mathbb{P}\left(A_k \cap \bigcup_{l \leq n} B_l \right) + \sum_{l \leq n} y_l \mathbb{P}\left(\bigcup_{k \leq m} A_k \cap B_l \right) \\
&= \sum_{k \leq m} x_k \mathbb{P}(A_k) + \sum_{l \leq n} y_l \mathbb{P}(B_l),
\end{aligned}$$

which is $\mathbb{E}X + \mathbb{E}Y$ and this finishes the proof. □

E.1 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.

E.2 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 $E(aX) = a\mathbb{E}X$,

(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 $\mathbb{E}X \mathbf{1}_A \leq \mathbb{E}X \mathbf{1}_B$.

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 ε 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 $X \geq X\,\mathbf{1}_{\{X \geq 1/n\}} \geq \frac{1}{n}\,\mathbf{1}_{\{X \geq 1/n\}}$, so by (a) we get

$$0 = \mathbb{E}X \geq \mathbb{E}\frac{1}{n}\,\mathbf{1}_{\{X \geq 1/n\}} = \frac{1}{n}\mathbb{P}\left(X \geq 1/n\right),$$

thus $\mathbb{P}\left(X \geq 1/n\right) = 0$, so

$$\mathbb{P}\left(X > 0\right) = \mathbb{P}\left(\bigcap_{n \geq 1}\{X \geq 1/n\}\right) = \lim_{n \to \infty}\mathbb{P}\left(X \geq 1/n\right) = 0.$$

(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

$$Z_n = \sum_{k=1}^{n \cdot 2^n} \frac{k-1}{2^n}\,\mathbf{1}_{\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}} + n\,\mathbf{1}_{\{X \geq n\}}.$$

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 $\varepsilon > 0$,

$$Z \leq (X_n + \varepsilon)\,\mathbf{1}_{\{Z < X_n + \varepsilon\}} + K\,\mathbf{1}_{\{Z \geq X_n + \varepsilon\}}. \tag{E.1}$$

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

$$Z\,\mathbf{1}_A \leq X\,\mathbf{1}_A \qquad \text{and} \qquad Z\,\mathbf{1}_{A^c} \leq Y\,\mathbf{1}_{A^c}.$$

Thus by E.2 (a),

$$\mathbb{E}Z\,\mathbf{1}_A \le \mathbb{E}X\,\mathbf{1}_A \qquad \text{and} \qquad \mathbb{E}Z\,\mathbf{1}_{A^c} \le \mathbb{E}Y\,\mathbf{1}_{A^c}\,.$$

Adding these two inequalities together and using that $\mathbb{E}Z\,\mathbf{1}_A + \mathbb{E}Z\,\mathbf{1}_{A^c} = \mathbb{E}Z$, which follows from linearity of expectation for simple random variables (Theorem E.1), we get

$$\mathbb{E}(X\,\mathbf{1}_A + Y\,\mathbf{1}_{A^c}) - \varepsilon < \mathbb{E}Z \le \mathbb{E}X\,\mathbf{1}_A + \mathbb{E}Y\,\mathbf{1}_{A^c}\,.$$

Sending $\varepsilon \to 0$ finishes the argument. □

Applying the claim to (E.1), we obtain

$$\mathbb{E}Z \le \mathbb{E}X_n + \varepsilon + K\mathbb{P}\left(Z \ge X_n + \varepsilon\right).$$

The events $\{Z \ge X_n + \varepsilon\}$ form a decreasing family (because $X_n \le X_{n+1}$ and their intersection is $\{Z \ge X + \varepsilon\} = \varnothing$ (because $X_n \to X$ and $Z \le X$). Therefore taking $n \to \infty$ in the last inequality gives

$$\mathbb{E}Z \le \lim_n \mathbb{E}X_n + \varepsilon.$$

Taking the supremum over simple random variables $Z \le X$ gives

$$\mathbb{E}X \le \lim_n \mathbb{E}X_n + \varepsilon.$$

Letting $\varepsilon \to 0$, we finish the proof. □

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*

$$\mathbb{E}\liminf_{n\to\infty} X_n \le \liminf_{n\to\infty} \mathbb{E}X_n.$$

Proof. Let $Y_n = \inf_{k\ge n} X_k$. Then this is a nondecreasing sequence which converges to $\liminf_{n\to\infty} X_n$ and $Y_n \le X_n$. Note that

$$\liminf_{n\to\infty} \mathbb{E}X_n \ge \liminf_{n\to\infty} \mathbb{E}Y_n = \lim_{n\to\infty} \mathbb{E}Y_n,$$

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. □

We are ready to prove linearity of expectation for nonnegative random variables.

E.6 Theorem. *Let X and Y be nonnegative random variables. Then*

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

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,

$$\mathbb{E}(X_n + Y_n) = \mathbb{E}X_n + \mathbb{E}Y_n.$$

Letting $n \to \infty$, by the virtue of Lebesgue's monotone convergence theorem, we get in the limit $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$. □

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

(b) $\mathbb{E}(aX) = a\mathbb{E}X$ for every $a \in \mathbb{R}$,

(c) if $X \le Y$, then $\mathbb{E}X \le \mathbb{E}Y$,

(d) $|\mathbb{E}X| \le \mathbb{E}|X|$.

Proof. (a) By the triangle inequality Theorem E.2 (a) and Theorem E.6,

$$\mathbb{E}|X+Y| \le \mathbb{E}(|X| + |Y|) = \mathbb{E}|X| + \mathbb{E}|Y|$$

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

$$\mathbb{E}(X+Y)^+ + \mathbb{E}X^- + \mathbb{E}Y^- = \mathbb{E}(X+Y)^- + \mathbb{E}X^+ + \mathbb{E}Y^+,$$

which rearranged again gives $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$.

(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 $-\mathbb{E}|X| \leq \mathbb{E}X \leq \mathbb{E}|X|$, that is $|\mathbb{E}X| \leq \mathbb{E}|X|$. $\square$

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 $X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)$ and there is an integrable random variable Y such that $|X_n| \leq Y$, then*

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

In particular,

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

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

$$\begin{aligned}\mathbb{E}(2Y) = \mathbb{E}\liminf(2Y - |X_n - X|) &\leq \liminf \mathbb{E}(2Y - |X_n - X|) \\ &= 2\mathbb{E}Y - \limsup \mathbb{E}|X_n - X|.\end{aligned}$$

As a result, $\limsup \mathbb{E}|X_n - X| \leq 0$, so

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

In particular, since by Theorem E.7 (d),

$$|\mathbb{E}(X_n - X)| \leq \mathbb{E}|X_n - X|,$$

we get that the left hand side goes to 0, that is $\mathbb{E}X_n \to \mathbb{E}X$. $\square$

F Appendix: Lindeberg's swapping argument

We begin by a useful observation regarding the definition of weak convergence: for random variables (and random vectors in $\mathbb{R}^n$), the weak convergence is already captured by all compactly supported smooth test functions (instead of all continuous bounded test functions as per the definition).

F.1 Lemma. *Let $\mu, \mu_1, \mu_2, \ldots$ be Borel probability measures on $\mathbb{R}$. The following are equivalent*

(i) for every continuous bounded function $f\colon \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{R}} f \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f \mathrm{d}\mu$$

($\mu_n \to \mu$ weakly),

(ii) for every compactly supported smooth function $f\colon \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{R}} f \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f \mathrm{d}\mu.$$

Proof. Only (ii)⇒(i) requires explanation. Repeating the first part of the proof of Theorem 8.5 with the function $g_{t,\varepsilon}$ replaced with its smooth approximation, we show that (ii) implies (8.1) which we already know is equivalent to (i) (by Theorem 8.5). □

F.2 Remark. Thanks to multidimensional cumulative distribution functions, this characterisation of weak convergence can be extended to Borel probability measures on $\mathbb{R}^n$.

The goal of this section is to present a classical argument of Lindeberg relying on consecutive swapping summands with independent Gaussians, leading to a quantitative version of the central limit theorem. The heart of the argument lies in the following lemma.

F.3 Lemma. *Let $X_1, \ldots, X_n$ be independent random variables, each with mean 0 and variance 1. Let Z be a standard Gaussian random variable. For every smooth function $f\colon \mathbb{R} \to \mathbb{R}$ with bounded derivatives up to order 3, we have*

$$\left|\mathbb{E} f\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E} f(Z)\right| \leq C \frac{\|f'''\|_\infty \sum_{k=1}^n \mathbb{E}|X_k|^3}{n^{3/2}},$$

where C is a universal positive constant. One can take $C = \frac{1+\sqrt{8/\pi}}{6}$.

Proof. Let $Z_1, \ldots, Z_n$ be i.i.d. copies of Z. Since Z has the same distribution as $\frac{Z_1+\ldots+Z_n}{\sqrt{n}}$, we have

$$\begin{aligned}
\mathbb{E} f\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E} f(Z) &= \mathbb{E} f\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E} f\left(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}}\right) \\
&= -\sum_{k=0}^{n-1} \big[\mathbb{E} f(S_k) - \mathbb{E} f(S_{k+1})\big],
\end{aligned}$$

where the telescoping sum involves

$$S_k = \frac{X_1 + \cdots + X_k + Z_{k+1} + \cdots + Z_n}{\sqrt{n}}, \qquad 0 \le k \le n,$$

with the convention $S_0 = \frac{Z_1+\cdots+Z_n}{\sqrt{n}}$ and $S_n = \frac{X_1+\cdots+X_n}{\sqrt{n}}$. It thus suffices to show that for every $0 \le k \le n-1$,

$$|\mathbb{E}f(S_k) - \mathbb{E}f(S_{k+1})| \le C\frac{\mathbb{E}|X_k|^3\|f'''\|_\infty}{n^{3/2}}.$$

Letting $V = \frac{X_1+\cdots+X_k+Z_{k+2}+\cdots+Z_n}{\sqrt{n}}$, we have

$$S_k = V + \frac{Z_{k+1}}{\sqrt{n}}, \qquad S_{k+1} = V + \frac{X_{k+1}}{\sqrt{n}}$$

and by Taylor expansion for f with Lagrange's remainder,

$$\begin{aligned}
f(S_k) &= f(V) + f'(V)\frac{Z_{k+1}}{\sqrt{n}} + \frac{f''(V)}{2}\frac{Z_{k+1}^2}{n} + \frac{f'''(\theta)}{6}\frac{Z_{k+1}^3}{n^{3/2}},\\
f(S_{k+1}) &= f(V) + f'(V)\frac{X_{k+1}}{\sqrt{n}} + \frac{f''(V)}{2}\frac{X_{k+1}^2}{n} + \frac{f'''(\theta')}{6}\frac{X_{k+1}^3}{n^{3/2}},
\end{aligned}$$

where θ and θ' denote mean points. Crucially, V is independent of Z_{k+1} as well as of X_{k+1}, so after taking the expectation and subtracting and using that Z_{k+1} and X_{k+1} have matching moments up to order 2, we obtain

$$|\mathbb{E}f(S_k) - \mathbb{E}f(S_{k+1})| \le \frac{\|f'''\|_\infty}{6n^{3/2}}(\mathbb{E}|Z_{k+1}|^3 + \mathbb{E}|X_{k+1}|^3).$$

By Hölder's inequality, $\mathbb{E}|X_{k+1}|^3 \ge (\mathbb{E}|X_{k+1}|^2)^{3/2} = 1 = \sqrt{\frac{\pi}{8}}\mathbb{E}|Z_{k+1}|^3$, which concludes the argument with $C = \frac{1+\sqrt{\frac{8}{\pi}}}{6} < 0.44$. □

Alternative proof of the vanilla central limit theorem – Theorem 10.5.

Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable with mean 0 and variance 1. Let Z be a standard Gaussian random variable. If $\mathbb{E}|X_1|^3 < \infty$, by Lemma F.3,

$$\left|\mathbb{E}f\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) - \mathbb{E}f(Z)\right| \le C\frac{\mathbb{E}|X_1|^3\|f'''\|_\infty}{n^{1/2}},$$

for every smooth compactly supported function f. Since the right hand side converges to 0 as n goes to ∞, Lemma F.1 finishes the proof in the case of finite third moment.

If X_1 does not have a finite third moment, we use a truncation argument. Fix $\varepsilon > 0$. For each k, let

$$\begin{aligned}
Y_k &= X_k\,\mathbf{1}_{\{|X_k|\le\varepsilon\sqrt{n}\}} - \mu_n,\\
Y_k' &= X_k\,\mathbf{1}_{\{|X_k|>\varepsilon\sqrt{n}\}} + \mu_n,
\end{aligned}$$

with

$$\mu_n = \mathbb{E} X_k \, \mathbf{1}_{\{|X_k| \le \varepsilon\sqrt{n}\}},$$

so that

$$X_k = Y_k + Y_k'$$

and

$$\mathbb{E} Y_k = \mathbb{E} Y_k' = 0.$$

Let

$$\sigma_n = \mathrm{Var}(Y_k) = \mathbb{E} X_k^2 \, \mathbf{1}_{\{|X_k| \le \varepsilon\sqrt{n}\}} - \mu_n^2.$$

By Lebesgue's dominated convergence theorem,

$$\mu_n \xrightarrow[n\to\infty]{} 0,$$
$$\sigma_n^2 \xrightarrow[n\to\infty]{} 1.$$

Let $f \colon \mathbb{R} \to \mathbb{R}$ be a smooth function with compact support. Fix $\delta > 0$. From Lemma F.3 applied to the Y_k,

$$|\mathbb{E} f\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right) - \mathbb{E} f(\sigma_n Z)| \le C \frac{\mathbb{E}|Y_1|^3 \|f'''\|_\infty}{n^{1/2}\sigma_n^3}.$$

Using $(a+b)^3 \le 4(a^3 + b^3)$, we obtain

$$\mathbb{E}|Y_1|^3 \le 4(\mathbb{E}|X_1|^3 + |\mu_n|^3) \le 4(\varepsilon\sqrt{n}\mathbb{E}|X_1|^2 + |\mu_n|^3) = 4(\varepsilon\sqrt{n} + |\mu_n|^3),$$

so that for n large enough,

$$\left|\mathbb{E} f\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right) - \mathbb{E} f(\sigma_n Z)\right| < \delta.$$

Since f is continuous and bounded, $\mathbb{E} f(\sigma_n Z) \to \mathbb{E} f(Z)$, so

$$|\mathbb{E} f(\sigma_n Z) - \mathbb{E} f(Z)| < \delta,$$

also for all n large enough. It remains to deal with the Y_k'. They all have mean 0 and

$$\mathbb{E}|Y_k'|^2 = \mathbb{E}|Y_1'|^2 = \mathrm{Var}(Y_1') = \mathbb{E}|X_1|^2 \, \mathbf{1}_{\{|X_1| > \varepsilon\sqrt{n}\}} - \mu_n^2,$$

with the right hand side converging to 0 as $n \to \infty$, by Lebesgue's dominated convergence theorem. Thus for n large enough,

$$\mathbb{E}\left|\frac{Y_1' + \cdots + Y_n'}{\sqrt{n}}\right| \le \left(\mathbb{E}\left|\frac{Y_1' + \cdots + Y_n'}{\sqrt{n}}\right|^2\right)^{1/2} = (\mathbb{E}|Y_1'|^2)^{1/2} < \delta.$$

Since f is smooth and compactly supported, it is Lipschitz, say with constant L, and because $X_k - Y_k = Y_k'$, we finally get

$$\begin{aligned}\left|\mathbb{E}f\left(\frac{X_1+\dots+X_n}{\sqrt{n}}\right)-\mathbb{E}f(Z)\right| &\leq \left|\mathbb{E}f\left(\frac{X_1+\dots+X_n}{\sqrt{n}}\right)-\mathbb{E}f\left(\frac{Y_1+\dots+Y_n}{\sqrt{n}}\right)\right| \\ &\quad+\left|\mathbb{E}f\left(\frac{Y_1+\dots+Y_n}{\sqrt{n}}\right)-\mathbb{E}f(\sigma_n Z)\right| \\ &\quad+\left|\mathbb{E}f(\sigma_n Z)-\mathbb{E}f(Z)\right| \\ &\leq L\mathbb{E}\left|\frac{Y_1'+\dots+Y_n'}{\sqrt{n}}\right|+2\delta<(L+2)\delta,\end{aligned}$$

for all n large enough. This shows that

$$\mathbb{E}f\left(\frac{X_1+\dots+X_n}{\sqrt{n}}\right)\xrightarrow[n\to\infty]{}\mathbb{E}f(Z),$$

Lemma F.1 thus gives

$$\frac{X_1+\dots+X_n}{\sqrt{n}}\xrightarrow[n\to\infty]{d}Z,$$

as desired. $\square$

Lemma F.3 is powerful enough to also give quantitative bounds in the central limit theorem (weaker than the optimal one provided by the Berry-Esseen theorem, but with a much simpler proof). We need an elementary fact providing approximations of indicator functions by smooth functions.

F.4 Lemma. *For every real number t and positive ε, there is a smooth nonnegative function f equal to 1 on $(-\infty,t]$, equal to 0 on $[t+\varepsilon,\infty)$ with $f\leq 1$ and $|f'''|\leq 200\varepsilon^{-3}$ everywhere.*

Proof. Consider

$$h(x)=\begin{cases}1, & x\in(-\infty,0],\\ \exp\left(-\frac{1}{1-x^2}\right), & x\in(0,1),\\ 0, & x\in[1,+\infty).\end{cases}$$

This function has the desired properties when $t=0$ and $\varepsilon=1$. Moreover, it can be checked that $\|h'''\|_\infty<200$. In general, we take the function $f(x)=h(\frac{x+t}{\varepsilon})$. $\square$

F.5 Theorem. *Let $X_1,\dots,X_n$ be i.i.d. random variables with mean 0, variance 1 and finite third moment. Let Z be a standard Gaussian random variable. We have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{X_1+\dots+X_n}{\sqrt{n}}\leq t\right)-\mathbb{P}\left(Z\leq t\right)\right|\leq 3\frac{(\mathbb{E}|X|^3)^{1/4}}{n^{1/8}}.$$

Proof. Denote $Z_n=\frac{X_1+\dots+X_n}{\sqrt{n}}$. Fix $t\in\mathbb{R}$ and $\varepsilon>0$. Let f be the function provided by Lemma F.4. In particular, $\mathbf{1}_{(-\infty,t]}(x)\leq f(x)$, for every $x\in\mathbb{R}$, thus

$$\mathbb{P}\left(Z_n\leq t\right)=\mathbb{E}\,\mathbf{1}_{(-\infty,t]}(Z_n)\leq\mathbb{E}f(Z_n)\leq\mathbb{E}f(Z)+\frac{200(1+\sqrt{8/\pi})}{6}\frac{\mathbb{E}|X_1|^3}{\varepsilon^3\sqrt{n}},$$

where the last inequality follows from Lemma F.3. Since $f(x) \leq \mathbf{1}_{(-\infty,t+\varepsilon]}(x)$ and the density of Z is bounded by $\frac{1}{\sqrt{2\pi}}$, we also have

$$\mathbb{E}f(Z) \leq \mathbb{E}\,\mathbf{1}_{(-\infty,t+\varepsilon]}(Z) \leq \mathbb{P}\left(Z \leq t\right) + \frac{\varepsilon}{\sqrt{2\pi}}.$$

As a result,

$$\mathbb{P}\left(Z_n \leq t\right) - \mathbb{P}\left(Z \leq t\right) \leq \frac{1}{\sqrt{2\pi}}\varepsilon + \frac{200(1+\sqrt{8/\pi})}{6}\frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\varepsilon^{-3}$$

and optimising over ε yields

$$\mathbb{P}\left(Z_n \leq t\right) - \mathbb{P}\left(Z \leq t\right) \leq C(\mathbb{E}|X_1|^3)^{1/4}n^{-1/8}$$

with $C = \left(\frac{(3\sqrt{2\pi})^{1/4}}{\sqrt{2\pi}} + (3\sqrt{2\pi})^{-3/4}\right)\left(\frac{200(1+\sqrt{8/\pi}}{6}\right)^{1/4} = 2.68...$ Similar arguments lead to an identical lower bound (we approximate the indicator $\mathbf{1}_{(-\infty,t-\varepsilon]}$ using Lemma F.4 with $t-\varepsilon$ and ε). This finishes the proof. $\square$

G Appendix: The moment method

H Appendix: Feller's converse to the central limit theorem

H.1 Theorem. *Let $\{X_{n,}\}_{n\geq 1, 1\leq k\leq n}$ be a triangular array of random variables with $\mathbb{E}X_{n,k}^2 < \infty$ for each n and k, such that for every $n \geq 1$, the variables $X_{n,1}, \ldots, X_{n,n}$ are independent. Suppose that $\mathbb{E}X_{n,k} = 0$ for each n and k and for each n,*

$$\sum_{k=1}^{n} \mathbb{E}X_{n,k}^2 = 1.$$

For $\varepsilon > 0$, set

$$L_n(\varepsilon) = \sum_{k=1}^{n} \mathbb{E}\bar{X}_{n,k}^2 \,\boldsymbol{1}_{\{|X_{n,k}|>\varepsilon\}} .$$

Let $Z_n = \sum_{k=1}^{n} X_{n,k}$. Suppose that

$$\max_{1\leq k\leq n} \mathbb{E}X_{n,k}^2 \xrightarrow[n\to\infty]{} 0. \tag{H.1}$$

If the sequence $(Z_n)_n$ converges in distribution to a standard Gaussian random variable, then

$$\text{for every } \varepsilon > 0, \quad L_n(\varepsilon) \xrightarrow[n\to\infty]{} 0. \tag{H.2}$$

Proof. Let $\phi_{n,k}(t) = \mathbb{E}e^{itX_{n,k}}$ be the characteristic function of $X_{n,k}$. By Lemma 10.4,

$$|1 - \phi_{n,k}(t)| = |\mathbb{E}(1 + itX_{n,k} - e^{itX_{n,k}})| \leq \frac{t^2}{2}\mathbb{E}X_{n,k}^2, \tag{H.3}$$

which thanks to (H.1), for every fixed t, converges to 0 (uniformly in k) as $n \to \infty$. We fix t and define $R_n(t)$ by

$$R_n(t) = \frac{t^2}{2} - \sum_{k=1}^{n} \mathrm{Re}(1 - \phi_{n,k}(t)).$$

Claim. For every t, $|R_n(t)| \to 0$ as $n \to \infty$.

Proof. Fix t. Using (H.3),

$$|1 - \phi_{n,k}(t)| \leq \frac{t^2}{2} \max_{k\leq n} \mathbb{E}X_{n,k}^2 \to 0$$

as $n \to \infty$, so, we have that for sufficiently large n and all $k \leq n$, $|1 - \phi_{n,k}(t)| < \frac{1}{2}$ and $|\mathrm{Arg}(\phi_{n,k}(t))| < \frac{\pi}{4}$, say. Using the principal value of the complex log function, we can thus write

$$\sum_{k=1}^{n} \log \phi_{n,k}(t) = \log \prod_{k=1}^{n} \phi_{n,k}(t)$$

and by the assumption of the theorem, the right hand side converges to $-t^2/2$. Therefore,

$$\begin{aligned}|R_n(t)| &\leq \left|\frac{t^2}{2} - \sum_{k=1}^{n}(1-\phi_{n,k}(t))\right| \\ &\leq \left|\frac{t^2}{2} + \log\prod_{k=1}^{n}\phi_{n,k}(t)\right| + \left|\sum_{k=1}^{n}\Big[-\log\phi_{n,k}(t) - (1-\phi_{n,k}(t))\Big]\right|.\end{aligned}$$

The first term goes to 0. Setting $z = 1-\phi_{n,k}(t)$, we have $|z| \leq \frac{1}{2}$ and using the Taylor series $-\log(1-z) = \sum_{k=1}^{\infty}\frac{z^k}{k}$ for the principal value of log, we obtain

$$|-\log\phi_{n,k}(t) - (1-\phi_{n,k}(t))| = |-\log(1-z) - z| \leq \sum_{k=2}^{\infty}\frac{|z|^k}{k} \leq |z|^2\sum_{k=2}^{\infty}\frac{(1/2)^{k-2}}{2} = |z|^2.$$

Thus for the second term we get

$$\begin{aligned}\left|\sum_{k=1}^{n}\Big[-\log\phi_{n,k}(t) - (1-\phi_{n,k}(t))\Big]\right| &\leq \sum_{k=1}^{n}|1-\phi_{n,k}(t)|^2 \\ &\leq \max_{k\leq n}|1-\phi_{n,k}(t)|\sum_{k=1}^{n}|1-\phi_{n,k}(t)|\end{aligned}$$

which, by virtue of (H.3), is upper bounded by

$$\max_{k\leq n}|1-\phi_{n,k}(t)|\frac{t^2}{2}\sum_{k=1}^{n}\mathbb{E}X_{n,k}^2 = \frac{t^2}{2}\max_{k\leq n}|1-\phi_{n,k}(t)| \leq \frac{t^4}{4}\max_{k\leq n}\mathbb{E}X_{n,k}^2.$$

The right hand side goes to 0 which finishes the proof of the claim. □

Having the claim, we finish the proof as follows. We fix $\varepsilon > 0$ and write

$$\begin{aligned}R_n(t) &= \frac{t^2}{2} - \sum_{k=1}^{n}\mathbb{E}[1-\cos(tX_{n,k})] \\ &= \frac{t^2}{2} - \sum_{k=1}^{n}\mathbb{E}[1-\cos(tX_{n,k})]\,\mathbf{1}_{\{|X_{n,k}|\leq\varepsilon\}} - \sum_{k=1}^{n}\mathbb{E}[1-\cos(tX_{n,k})]\,\mathbf{1}_{\{|X_{n,k}|>\varepsilon\}}.\end{aligned}$$

If $|x| > \varepsilon$, we have $1-\cos(tx) \leq 2 \leq 2\frac{x^2}{\varepsilon^2}$. This is how we bound the second sum. The first sum will be bounded using $1-\cos(tx) \leq \frac{t^2x^2}{2}$. These yield

$$R_n(t) \geq \frac{t^2}{2} - \frac{t^2}{2}\sum_{k=1}^{n}\mathbb{E}X_{n,k}^2\,\mathbf{1}_{\{|X_{n,k}|\leq\varepsilon\}} - \frac{2}{\varepsilon^2}\sum_{k=1}^{n}\mathbb{E}X_{n,k}^2 = \frac{t^2}{2}L_n(\varepsilon) - \frac{2}{\varepsilon^2},$$

equivalently,

$$L_n(\varepsilon) \leq \frac{4}{t^2\varepsilon^2} + \frac{2R_n(t)}{t^2}.$$

This shows that $L_n(\varepsilon) \to 0$ as $n \to \infty$ (given $\delta > 0$, we fix t such that $\frac{4}{t^2\varepsilon^2} < \delta/2$ and for this t, for all n large enough, we have $\frac{2R_n(t)}{t^2} < \delta/2$, by the claim). □

I Appendix: Uniform integrability

We recall the definition: a family of random variables $\{X_t\}_{t\in T}$ is **uniformly integrable** if for every $\varepsilon > 0$, there is $K > 0$ such that for all $t \in T$, we have $\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} \le \varepsilon$.

We start with two easy criteria guaranteeing uniform integrability.

I.1 Lemma. *Let $\{X_t\}_{t\in T}$ be a family of random variables such that there is a nonnegative random variable Y with $\mathbb{E}Y < \infty$ and $|X_t| \le Y$ for every $t \in T$. Then the family $\{X_t\}_{t\in T}$ is uniformly integrable.*

I.2 Remark. In particular, if $\{X_t\}_{t\in T}$ is a finite family of integrable random variables, then it is uniformly integrable (we simply take $Y = \sum_{t\in T} |X_t|$).

Proof of Lemma I.1. We have, $\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} \le \mathbb{E}Y\,\mathbf{1}_{\{Y>K\}}$ and the right hand side goes to 0 as $K \to \infty$ (by Lebesgue's dominated convergence theorem). $\square$

I.3 Lemma. *Let $\{X_t\}_{t\in T}$ be a family of random variables bounded in L_p for some $p > 1$ (that is, $M = \sup_{t\in T}\mathbb{E}|X_t|^p < \infty$). Then the family $\{X_t\}_{t\in T}$ is uniformly integrable.*

Proof. Write $p = 1 + \delta$ with $\delta > 0$. We have,

$$\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} = \mathbb{E}|X_t|^{1+\delta}|X_t|^{-\delta}\,\mathbf{1}_{\{|X_t|>K\}} \le K^{-\delta}\mathbb{E}|X_t|^{1+\delta} \le K^{-\delta}M,$$

where $M = \sup_{t\in T}\mathbb{E}|X_t|^p$, so that the right hand side goes to 0 as $K \to \infty$. $\square$

There is an equivalent definition of uniform integrability which perhaps explains the name better and is often useful.

I.4 Theorem. *Let $\{X_t\}_{t\in T}$ be a family of random variables. It is uniformly integrable if and only if the following two conidtions hold*

(i) it is bounded in L_1, that is $\sup_{t\in T}\mathbb{E}|X_t| < \infty$,

(ii) for every $\varepsilon > 0$, there is $\delta > 0$ such that for every event A with $\mathbb{P}(A) < \delta$ and every $t \in T$, we have $\mathbb{E}|X_t|\,\mathbf{1}_A < \varepsilon$.

Proof. "$\Rightarrow$": Fix $\varepsilon > 0$ and choose $K > 0$ such that $\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} \le \varepsilon$ for every $t \in T$. For an event A, we have

$$\mathbb{E}|X_t|\,\mathbf{1}_A = \mathbb{E}|X_t|\,\mathbf{1}_A\,\mathbf{1}_{\{|X_t|\le K\}} + \mathbb{E}|X_t|\,\mathbf{1}_A\,\mathbf{1}_{\{|X_t|>K\}} \le K\mathbb{P}(A) + \varepsilon < 2\varepsilon$$

if $\mathbb{P}(A) < \delta = \frac{\varepsilon}{K}$. This shows (ii). Taking $A = \Omega$, the first inequality shows (i).

"$\Leftarrow$": Fix $\varepsilon > 0$ and take δ provided by (ii). We would like to choose K such that $\mathbb{E}|X_t|\,\mathbf{1}_{\{|X_t|>K\}} < \varepsilon$ for every $t \in T$. Let $A = \{|X_t| > K\}$. Let $M = \sup_{t\in T}\mathbb{E}|X_t|$ which is finite by (i). By Chebyshev's inequality,

$$\mathbb{P}(A) \le \frac{1}{K}\mathbb{E}|X_t| \le \frac{M}{K} < \delta$$

provided that $K > \frac{M}{\delta}$. Thus, for this choice of K, what we want follows from (ii) with $A = \{|X_t| > K\}$. $\square$

I.5 Remark. Let X be an integrable random variable. Combining Remark I.2 (applied to the one-element family $\{X\}$) with (ii) of Theorem I.4, we get a "uniform continuity" property of expectation:

$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall A : \mathbb{P}(A) < \infty \qquad \mathbb{E}|X|\,\mathbf{1}_A < \varepsilon. \tag{I.1}$$

The usefulness of uniform integrability lies in the fact that it captures L_p convergence.

I.6 Theorem. *Let $p > 0$. For random variables $X, X_1, X_2, \dots$ in L_p, we have that $X_n \to X$ in L_p if and only if the following two conditions hold*

(i) $X_n \to X$ in probability,

(ii) $\{|X_n|^p\}_{n\geq 1}$ is a uniformly integrable family.

Proof. "$\Rightarrow$": Clearly (i) holds (see Theorem 6.14). To see (ii), we shall use Theorem I.4. Fix $\varepsilon > 0$. Let A be an event. First we use $|a+b|^p \leq C(|a|^p + |b|^p)$, $a, b \in \mathbb{R}$ ($C = \max\{2^{p-1}, 1\}$ is good), to bound

$$\mathbb{E}|X_n|^p\,\mathbf{1}_A \leq C(\mathbb{E}|X|^p\,\mathbf{1}_A + \mathbb{E}|X_n - X|^p\,\mathbf{1}_A).$$

For large n, say $n > N$, we have $\mathbb{E}|X_n - X|^p\,\mathbf{1}_A \leq \mathbb{E}|X_n - X|^p < \varepsilon$. For $n \leq N$, we can choose δ from (ii) of Theorem I.4 applied to the finite family $\{|X|, |X_n - X|, n \leq N\}$ to bound each term by ε. This finishes the argument.

"$\Leftarrow$": Fix $\varepsilon > 0$. Let δ be chosen from (ii) of Theorem I.4 for the uniformly integrable family $\{|X_n - X|^p\}_{n\geq 1}$ (it is uniformly integrable because $|X_n - X|^p \leq C(|X_n|^p + |X|^p)$). We have,

$$\begin{aligned}\mathbb{E}|X_n - X|^p &= \mathbb{E}|X_n - X|^p\,\mathbf{1}_{\{|X_n - X| \leq \varepsilon\}} + \mathbb{E}|X_n - X|^p\,\mathbf{1}_{\{|X_n - X| > \varepsilon\}} \\ &\leq \varepsilon^p + \mathbb{E}|X_n - X|^p\,\mathbf{1}_{\{|X_n - X| > \varepsilon\}}.\end{aligned}$$

Let $A = \{|X_n - X| > \varepsilon\}$. For n large enough, by (i), $\mathbb{P}(A) < \delta$, thus the second term for all such n is bounded by ε. This shows that $X_n \to X$ in L_p. $\square$

I.7 Example. Consider the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{Leb})$ and the random variables $X_n = n\,\mathbf{1}_{[0,1/n]}$, $n \geq 1$. Then, $\mathbb{E}|X_n| = 1$, so $(|X_n|)$ is bounded in L_1, but the family is not uniformly integrable: for every $K > 0$, we have $\mathbb{E}|X_n|\,\mathbf{1}_{\{|X_n| > K\}} = \mathbb{E}|X_n| = 1$ for all $n \geq K$. In this example, $X_n \to 0$ a.s. and in probability, but not in L_1 (see Example 6.15).

References

[1] De Acosta, A. A new proof of the Hartman–Wintner law of the iterated logarithm. *Ann. Probab.* 11 (1983), no. 2, 270–276.

[2] Bolthausen, E., An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete* 66 (1984), no. 3, 379–386.

[3] Durrett, R., Probability: theory and examples. Fourth edition. Cambridge Series in Statistical and Probabilistic Mathematics, 31. *Cambridge University Press, Cambridge*, 2010.

[4] Feller, W., An introduction to probability theory and its applications. Vol. I and II *John Wiley & Sons, Inc.*, 1968.

[5] Grimmett, G., Welsh, D., Probability – an introduction. Second edition. *Oxford University Press*, Oxford, 2014.

[6] Jakubowski, R., Sztencel, R., Wstep do teorii prawdopodobieństwa (in Polish), *Script, Warszawa* 2001.

[7] Rosenthal, J., A first look at rigorous probability theory. Second edition. *World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ*, 2006.

[8] Tyurin, I. S., Improvement of the remainder in the Lyapunov theorem. (Russian) *Teor. Veroyatn. Primen.* 56 (2011), no. 4, 808–811; translation in *Theory Probab. Appl.* 56 (2012), no. 4, 693–696.