Lecture Notes (over-simplified)

Probability Theory I

Fall 2021

University of Helsinki

Preface

There are many classical books on introduction to probability theory. If I write another one for this course, it will be almost surely inferior, as I constantly discover new things that I didn't really understand as a student while reading these classical references. The plan is to give an accessible course with examples that illustrate the general principles and complement the existing references (and usually with the simplest setting), rather than a comprehensive course that is built towards a deep understanding of the theoretical setups. Thus, the measure theory will not be used explicitly in the beginning but rather only alluded to, and we only recall the relevent elements when appealed to.

Below are some references that seem to be popular in the present days, and the course will be mainly based on the two versions.[1]

1. Durrett – Probability: Theory and Examples
2. Feller – An Introduction to Probability Theory and its Applications
3. Jacod, Protter – Probability Essentials
4. Williams – Probability with Martingales
5. Billingsley – Probability and Measure

As a word of warning, and it is a message that has been transmitted from my teacher to my generation: this course will not make you a poker star. Or at least, it is not oriented to make you discover loopholes in casinos.[2]

I strongly recommand you to discuss between peers or use the Ratkomo system, as it is important to train your independence (in the non-probabilistic sense).

Have fun!
Yichao

[1] My personal textbook is this set of lecture notes by Le Gall, in French but freely available: see Chapters 8,9,10 of the file.

[2] But if you learn well and try harder. . . https://www.bbc.com/news/magazine-27519748

Contents

Chapter 1
Probability space and random variables

We recall some basic concepts from measure theory, since a probability space is no more than a measurable space with total mass 1, and the correct way of doing calculations in probability theory uses integration in the sense of Lebesgue. However, bear in mind that the focus of the probability theory are not properties of measurable spaces, but properties of *random variables*, which are "just" measurable functions.

1.1 Probability space

Definition 1 (Probability space). A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ with:

- A set of "outcomes" Ω called **sample space**;
- A σ-field or σ-algebra $\mathcal{F}$ whose elements are called **events**;
- A **probability measure** $\mathbb{P} : \mathcal{F} \to [0, 1]$ that assigns probability to events.

In particular, as a set equipped with a σ-algebra, $(\Omega, \mathcal{F})$ is a *measurable space*. The condition that $\mathcal{F}$ is a σ-algebra, i.e. (non-empty and) stable under taking complement and *countable* unions or intersections, is the assumption for which we can put a (positive) *measure* $\mu : \mathcal{F} \to \mathbb{R}$ which satisfies in particular the *σ-additivity property* for a countable pairwise disjoint sets in Ω. More precisely:

Definition 2 (Sigma-algebra). Let $\mathcal{F}$ be a collection of subsets of some set Ω. We call $\mathcal{F}$ a σ-algebra if

- Trivial elements $\emptyset, \Omega \in \mathcal{F}$;
- Stability by complement: if $A \in \mathcal{F}$ then A^c in $\mathcal{F}$, where the complement is taken with respect to Ω;
- Stability under countable unions and intersections: if $\{A_n\}_{n\geq 1}$ is a countable family of elements in $\mathcal{F}$, then their union $\cup_{n\geq 1} A_n$ and their intersection $\cap_{n\geq 1} A_n$ are also in $\mathcal{F}$.

The couple $(\Omega, \mathcal{F})$ is then called a **measurable space**.

Definition 3 (Measure). Given a measurable space $(\Omega, \mathcal{F})$, a (positive) measure μ assigns to each element A of $\mathcal{F}$ a number $\mu(A) \in [0, \infty]$ (infinity is allowed in general). Furthermore, μ should satisfy the following properties:

- Trivial elements: $\mu(\emptyset) = 0$;
- σ-additivity: for a countable collection of **pairwise disjoint** elements $\{A_n\}_{n \geq 0}$ in $\mathcal{F}$, we have

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n).$$

Remark 1. The terminology sigma- or σ- usually refers to the countability of some family.

Definition 4 (Probability measure). A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a measure with $\mathbb{P}[\Omega] = 1$.

Remark 2. Shen μ is a finite measure, we automatically have $\mu(\emptyset) = 0$. This is because we must have $\mu(\emptyset) + \mu(\Omega) = \mu(\Omega)$ by disjointness of $\emptyset$ with any set, and substracting $\mu(\Omega)$ from both sides yields the result. In particular, this holds for any probability measure μ.

Remark 3. Given a finite (and non-trivial) measure μ, we can renormalize it to a probability measure $\mathbb{P}$ by defining $\mathbb{P}(A)$ as $\frac{\mu(A)}{\mu(\Omega)}$.

Usually we recall some basic properties of a measure now, but we will do a more streamlined version where basic properties from measure theory are recalled when they are needed, and more complicated properties (usually found in an appendix of a textbook) in the weekly reading assignments.

1.2 Random variables

A random variable is nothing but a measurable function.

Definition 5 (Measurable function). Let $f : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ a function between two measurable spaces. We say that f is **measurable** if the *preimage* $f^{-1}(B)$ of any measurable set B in $\mathcal{E}$ is measurable in $\mathcal{F}$, i.e. $B \in \mathcal{E} \implies f^{-1}(B) \in \mathcal{F}$.

Remark 4. The operation f^{-1} above, when applied to a set B, refers to the preimage and gives back a set. It is not the same as the antecedent, which requires bijectivity of f (we almost never use f^{-1} in this sense in the course).

Definition 6 (Random variable). A **random variable** X is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(E, \mathcal{E})$.

Many people write *r.v.* for random variables: I tend to spell them out more often (as well as other terminologies).

Remark 5. An analogy with topology: open sets are the "defining elements" of a topology just as measurable sets are the "defining elements" of a measurable space. A continuous function is an operation that respects the notion of open sets just as a measurable function is an operation that respects the notion of measurable sets.

Example 1 (Coin toss). Consider $\Omega = \{-1, 1\}$ and $\mathcal{F} = \{\emptyset, \{-1\}, \{1\}, \Omega\}$. Then $(\Omega, \mathcal{F})$ is a measurable space. Together with a measure μ such that $\mu(\{-1\}) = \mu(\{1\}) = 1$, the triple $(\Omega, \mathcal{F}, \mu)$ is the so-called probability space of a fair coin toss.

You may recognize that these kinds of objects in mathematics are useful for "transferring" structures between spaces. That is, there is a natural thing to do when you want to measure an event $B \in \mathcal{E}$, knowing that there is no measure directly defined on $(E, \mathcal{E})$ but you dispose of a measurable function to $(E, \mathcal{E})$ from another measurable space with a defined measure.

Definition 7 (Law of a random variable). The **law** of a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ is the probability measure $\mathbb{P}_X$ on $(E, \mathcal{E})$ defined as

$$\mathbb{P}_X[B] = \mathbb{P}[X^{-1}(B)] \tag{1.1}$$

for all $B \in \mathcal{E}$.

So the *law* of a random variable X is nothing else but the *pushforward* of $\mathbb{P}$ by X on $(E, \mathcal{E})$. However, this notion becomes fundamental in applications since it gives us a way of calculating the probability of "observing X in a certain state".

Remark 6. The law $\mathbb{P}_X$ is always a probability measure, since $\mathbb{P}_X[E] = \mathbb{P}[X^{-1}(\Omega)] = \mathbb{P}[\Omega] = 1$.

Remark 7. The identity map $\mathrm{Id} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{F})$ is always a measurable function, i.e. Id is always a random variable from a probability space to itself. The law of Id is then $\mathbb{P}$. For example, we can say that Id from the probability space of a fair coin toss (see above) to itself is a random variable representing a fair coin toss, although this definition is not canonical (i.e. not unique and alternative definitions are possible).

In practise, $\mathbb{P}_X[B]$ is written more as $\mathbb{P}[X \in B] \coloneqq \mathbb{P}[\omega \in \Omega \,;\, X(\omega) \in B]$, which should be interpreted as "the probability of observing X in the state B".

1.3 Real valued random variables

When considering random variables with values in a topological space such as $\mathbb{R}$, it is most natural to make the defining elements of a topological space, i.e. open sets, measurable.

Definition 8 (σ-algebra generated by a family). Let $(\Omega, \mathcal{F})$ be a measurable space and $F = \{A_j\}_{j \in J}$ a family (not necessarily countable) of elements in $\mathcal{F}$. We define the σ-algebra generated by F, denoted as $\sigma(F)$, to be the smallest σ-algebra containing the family F.

The above definition is well-posed since $\mathcal{F}$ is always a σ-algebra containing the family F and σ-algebras are stable under arbitrarily many intersections.

Definition 9 (Borel σ-algebra of $\mathbb{R}$). The Borel σ-algebra of $\mathbb{R}$ is the smallest σ-algebra on $\mathbb{R}$ containing all open sets on $\mathbb{R}$. We usually denote it by $\mathcal{B}(\mathbb{R})$.

Example 2. Let $(\Omega, \mathcal{F})$ be a measurable space and $A \in \mathcal{F}$ a measurable set. Then $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$ containing 4 elements if A is non-trivial.

Example 3 (Generators of $\mathcal{B}(\mathbb{R})$). Recall that $\mathcal{B}(\mathbb{R})$ can be also equivalently generated by any of the following family:

1. Intervals of type $]a, b[\in \mathbb{R}$;
2. Intervals of type $]-\infty, a[, a \in \mathbb{R}$;
3. Intervals of type $]-\infty, a[, a \in \mathbb{Q}$.

We replace the above intervals by closed intervals.

Proof. See measure theory course.

Definition 10 (Real random variable). A **real random variable** is a random variable X from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Usually, the probability space is omitted, $\mathcal{B}(\mathbb{R})$ is implicitly implied, and we write simply $X : \Omega \to \mathbb{R}$. When we really want to emphasize on the probability space, we say that X is $\mathcal{F}$-measurable.

Remark 8. Let μ be a probability measure on $\mathbb{R}$. There is a canonical way of constructing a random variable X on $\mathbb{R}$ having μ as its law: choose X as $\mathrm{Id} : (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The induced measure is obviously μ, but it is also the law of X by definition.

Definition 11 (Absolutely continuous real random variable). We say that $X : \Omega \to \mathbb{R}$ is a **absolutely continuous real random variable** if there exists a Borel function $p : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $\int_{\mathbb{R}} p(x) = 1$, such that

$$\mathbb{P}_X[B] = \int_B p(x)dx$$

for all Borel sets $B \in \mathcal{B}(\mathbb{R})$. Recall that $\mathbb{P}_X[B] = \mathbb{P}[X \in B]$.

Recall that a Borel function is simply a measurable function between two topological spaces equipped with their respective Borel σ-algebras.

Remark 9. The renormalization condition $\int_{\mathbb{R}} p(x) = 1$ comes from $\mathbb{P}_X[\mathbb{R}] = 1$.

Remark 10. The Lebesgue measure λ on $\mathbb{R}$ is the unique measure on $\mathcal{B}(\mathbb{R})$ such that $\lambda(]a, b[) = b - a$ for all $a < b$.

Definition 12 (Probability density function). Let X be an absolutely continuous real random variable. The **probability density function** of X is the Radon-Nikydym derivative $p \coloneqq \frac{d\mathbb{P}_X}{d\lambda}$ of the law of X with respect to the Lebesgue measure on $\mathbb{R}$ in the definition above.

It is often called simply *density* or *p.d.f.*, and sometimes denoted by f in the literature. In particular, we can express the following type of probability:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b p(x)dx,$$

i.e., the probability of X "falling" in the interval $[a, b]$. In particular, one could think of $p(x)dx$ as the probability of X being in the infinitesimal interval $[x, x + dx]$.

Remark 11. The density function is unique modulo a set of Lebesgue measure 0, since modifying it on a Lebesgue null-set does not affect its Lebesgue integrals.

A real random variable X might be continuous, i.e. $\mathbb{P}[X = a] = 0$ for all $a \in \mathbb{R}$, without having a density function, i.e. X might not be absolutely continuous (actually, absolutely continuous refers to absolute continuity of the law of X with respect to the Lebesgue measure). That is, the distribution of X might be "supported" on a set of Lebesgue measure 0, i.e. X is a *singular continuous distribution*. This corresponds to the singular part in the Radon-Nikodym decomposition theorem: a typical example is the so-called Cantor's distribution. We will study more characterizations next week with the notion of *cumulative distribution function*.

Remark 12. In this course, we mainly deal with absolutely continuous random variables, especailly when performing calculations.

1.4 Sigma-algebra of a random variable

Given a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$, we don't need all $\mathcal{F}$ to make X measurable.

Definition 13 (σ-algebra of a random variable). Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ be a random variable. The σ-algebra generated by X, denoted by $\sigma(X)$, is defined as

$$\sigma(X) \coloneqq \{X^{-1}(B), B \in \mathcal{E}\} \subset \mathcal{F}.$$

In practise, most events that we study in this course are given in this form. In some sense, $\sigma(X)$ is when you can infer about the structure of $\mathcal{F}$ by inspecting the random variable X. Sometimes we say that $\sigma(X)$ contains the *information* given by X.

Remark 13. The σ-algebra of X is a σ-algebra: this is because the operation of taking the preimage is distributive with respect to unions and intersections.

Remark 14. Consider a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \{-1, 1\}$ modelizing a fair coin toss. The σ-algebra generated by X is $\sigma(X) = \{\emptyset, X^{-1}(\{-1\}), X^{-1}(\{1\}), \Omega\}$. This is the maximal structure (or partition in some sense) you can deduce on $\mathcal{F}$ by observing the results of a coin toss, that there is a part in $\mathcal{F}$ that leads to -1, and its complement that leads to 1.

We may want to make more several random variables measurable at once.

Definition 14 (σ-algebra of a family of random variables). Given an arbitrary family of random variables $\{X_j\}_{j\in J}$ with value respectively in $(E_j, \mathcal{E}_j)$ for each $j \in J$, the σ-algebra generated by $\{X_j\}_{j\in J}$ is

$$\sigma(\{X_j\}_{j\in J}) \coloneqq \sigma\{X_j^{-1}(B_j), j \in J, B_j \in \mathcal{E}_j\}. \tag{1.2}$$

In other words, it is the smallest σ-algebra that makes each random variable measurable.

Remark 15. The collection of preimages $\{X_j^{-1}(B_j), j \in J, B_j \in \mathcal{E}_j\}$ is in general not enough in the above definition, since it is not in general a σ-algebra. We "complete" it to a σ-algebra by taking the smallest σ-algebra containing this collection of sets.

Up next is a special case of measurability with respect to (the σ-algebra generated by) a random variable.

Proposition 1 (Measurability with respect to a random variable). *Let X be a random variable with value in $(E, \mathcal{E})$, and Y a real random variable. Then Y is $\sigma(X)$-measurable if and only if there exists a measurable function $f : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $Y = f(X)$.*

Proof. If such a measurable function f exists, then Y is $\sigma(X)$-measurable by composition of measurable functions. In the other direction, suppose that Y is $\sigma(X)$-measurable and construct such measurable function f. We will use the fact that any measurable function can be written as the pointwise limit of a sequence of simple functions.

Start by the case of a $\sigma(X)$-measurable simple function $Y = \sum_{j=1}^{n} a_j \mathbf{1}_{A_j}$ with $A_j \in \sigma(X)$. Actually, start by the case that $Y = a_j \mathbf{1}_{A_j}$ with $A_j \in \sigma(X)$ and find the appropriate f. Since $A_j \in \sigma(X)$, we can write A_j as $X^{-1}(B_j)$ for some $B_j \in \mathcal{E}$. Now if $\omega \in A_j$, $Y(\omega) = a_j$ and $X(\omega) \in B_j$; and if $\omega \notin A_j$, $Y(\omega) = 0$. We can then put $f = a_j \mathbf{1}_{B_j}$, and f is $\mathcal{E}$-measurable. I leave you to generalize this to all simple functions Y.

The next approximation argument is more abstract but is almost automatic after using it several times. We know that any $\sigma(X)$-measurable function Y can be approximated by $\sigma(X)$-measurable simple functions Y_n, and let f_n denotes the corresponding $\mathcal{E}$-measurable simple functions as above. It is natural to define $f(x)$ as the limit of $f_n(x)$ for all $x \in E$, but this limit does not necessarily exist, and we put $f(x) = 0$ in this case. Now for any $\omega \in \Omega$, $X(\omega)$ is in the set where the limit of $f_n(x)$ exists, since $f_n(X(\omega)) = Y_n(\omega) \to Y(\omega)$. Now by definition of f, $f(X(\omega)) = \lim f_n(X(\omega)) = Y(\omega)$, and this finishes the proof.

1.5 Classical laws

Discrete laws

1. Uniform distribution: if E is a set with $0 < n < \infty$ elements, then a random variable X with value in $(E, \mathcal{E})$ is uniformly distributed on E if $\mathbb{P}[X = x] = 1/n$ for all $x \in E$.
2. Bernoulli distribution of parameter $p \in [0, 1]$: it is the law of a random variable X with value in $\{0, 1\}$ such that $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$.
3. Binomial distribution $\mathcal{B}(n, p)$ with integer $0 < n < \infty$ and $p \in [0, 1]$: it is the law of a random variable X with value in $\{1, \dots, n\}$ such that $\mathbb{P}[X = k] = C_n^k p^k (1 - k)^{n-k}$.
4. Geometric distribution with parameter $p \in (0, 1)$: it is the law of a random variable X with value in $\mathbb{Z}_{\geq 0}$ such that $\mathbb{P}[X = k] = (1 - p)p^k$. It modelizes e.g. the law of the first appearance of tail in a sequence of independent baised coin tosses.
5. Poisson distribution with parameter $\lambda > 0$: it is the law of a random variable X with value in $\mathbb{Z}_{>0}$ with $\mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$ for all $k \in \mathbb{Z}_{>0}$.

(Absolutely) continuous laws

Recall that the law of a absolutely continuous real random variable X is characterized by its density function $p(x)$.

1. Uniform distribution on (a, b): $p(x) = \frac{1}{b-a} \mathbf{1}_{(a,b)}(x)$.
2. Exponential distribution with parameter $\lambda > 0$: $p(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}_{>0}}(x)$.
3. Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Exponential distributions $X \sim \mathrm{Exp}(\lambda)$ have the special property that $\mathbb{P}[X > a + b] = \mathbb{P}[X > a]\mathbb{P}[X > b]$ for any $a, b > 0$, which later (after you would have learnt about conditional probability) will be interpreted as the "loss of memory" property. Gaussian distributions are also called *normal distributions*, and the case $\mathcal{N}(0, 1)$ is called *standard normal distribution*.

Chapter 2
Calculations with real random variables

We give several concrete examples of calculations: many results are merely translations from a course on Lebesgue integration. However, as Feller puts it:

> ". . . but it should be borne in mind that numerical probabilities are not the principle objects of the theory. Its aim is to discover general laws and construct satisfactory theoretical models."

2.1 Expectation

The expectation of a random variable gives the *average* of the random variable.

Definition 15 (Expectation). Let X be a real random variable. The **expectation** of X, denoted by

$$\mathbb{E}[X] \coloneqq \int_\Omega X(\omega)\mathbb{P}(d\omega),$$

if well-defined in either of the following cases:

- if $X \geq 0$, in which case $\mathbb{E}[X] \in [0, \infty]$;
- if $\mathbb{E}[|X|] = \int |X| d\mathbb{P} < \infty$.

We sometimes also call $\mathbb{E}[X]$ the *mean* of the random variable X.

Remark 16. The expectation is linear, i.e. $\mathbb{E}[aX+bY] = a\mathbb{E}[X]+b\mathbb{E}[Y]$ for all scalar a, b. In fact, all the good properties of an integrable function apply to the expectation since it is just another name for the integral of a measurable function.

We recall briefly some important theorems above convergence of an integral (against a probability measure):

Theorem 1 (Monotone convergence theorem, a.k.a. Beppo-Levi's lemma). *If $X_n \geq 0$ for all n and $\{X_n\}$ increases pointwise and converges to X, then $\mathbb{E}[X_n]$ (increases and) converges to $\mathbb{E}[X]$.*

Theorem 2 (Fatou's lemma). *Consider a sequence of positive real random variables, i.e. $X_n \geq 0$ for all n. Then $\mathbb{E}[\liminf_n X_n] \leq \liminf_n \mathbb{E}[X_n]$.*

The positivity assumption in Fatou's lemma cannot be dropped directly.

Theorem 3 (Dominated convergence theorem). *Consider a sequence of real random variables $\{X_n\}$ dominated by $Z \in L^1$, i.e. $|X_n| \leq Z$, $\mathbb{E}[Z] < \infty$. Then if X_n converges to X, we get $\mathbb{E}[X_n] \to \mathbb{E}[X]$.*

In probability theory, one uses *almost surely* (or *a.s.* in short) instead of *almost everywhere* (or *a.e.*) when the measure in question is a probability measure.

We now revisit the law of a random variable using expectation.

Proposition 2 (A useful representation for the law). *Let X be a real random variable with value in $(E, \mathcal{E})$. Then for all measurable functions $f : (E, \mathcal{E}) \to (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$, we have*

$$\mathbb{E}[f(X)] = \int_E f(x)\mathbb{P}_X(dx).$$

If f is not necessarily positive, the formula remains true under the condition that $\mathbb{E}[|f(X)|] < \infty$.

Proof. By definition, the result is true for $f = \mathbf{1}_B$ with $B \in \mathcal{E}$. By linearity, the result is true for $f = \sum_{k=1}^n b_k \mathbf{1}_{B_k}$, i.e. for all simple functions. Recall that any non-negative measurable function is the pointwise limit of an increasing sequence of non-negative simple functions. Thus, by monotone convergence, the result is true for all non-negative measurable functions. The other case follows similarly by using dominated convergence instead.

Remark 17. It follows from the definition of the Lebesgue integral that $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A]$ for any measurable set $A \in \Omega$, where $\mathbf{1}_A$ is the indicator function.

Remark 18. The converse of this proposition is also true: if the relation above holds for any measurable function f, then the measure $\mathbb{P}_X$ is the law of X. To see this, choose f to be the indicator function $\mathbf{1}_B$ of any measurable set $B \in \mathcal{E}$, and the relation writes $\mathbb{P}[X \in B] = \mathbb{P}_X[B]$, which is the defining relation for the law $\mathbb{P}_X$.

What is important in practise is the following inverse statement: if we can write

$$\mathbb{E}[f(X)] = \int f d\nu,$$

for "sufficiently many" functions f, then we can identify ν as the law of X. We will discuss about the "sufficient many" condition later in this course.

2.2 Moments

Definition 16 (Moment of a random variable). Let X be a real random variable and p a real number. The p-th **moment** of X is the quantity

$$\mathbb{E}[X^p],$$

defined when $X \geq 0$ or $\mathbb{E}[|X|^p] < \infty$.

We define similarly as in the integration course the spaces $L^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p \in [1, \infty]$, with norm $||X||_{L^p} = \mathbb{E}[|X|^p]^{1/p}$ for $p \in [1, \infty)$ (this norm is also written as $|| \cdot ||_p$) and $|| \cdot ||_\infty$ is the (essential) sup-norm, see below.

Remark 19. The 0-th moment of a random variable is always 1 and the first moment of a random variable is just the expectation (provided that it exists).

Remark 20. The expectation of a random variable that is almost surely 0 is 0. Conversely, if the expectation of a positive random variable is 0, then the random variable is almost surely 0.

Remark 21. If the expectation of a positive random variable is finite, then the random variable is almost surely finite.

Remark 22. Knowing all (positive) integer moments of a probability measure does not necessarily characterize its law. Determining some sufficient conditions is known as the *moment problem*.

A lot of inequalities should be recalled here. The most important ones are:

Lemma 1 (Cauchy-Schwarz). *Suppose that some real random variables X, Y are in L^2. Then*

$$\mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2}\mathbb{E}[Y^2]^{1/2}.$$

Remark 23. Taking $X = Y$, we have $\mathbb{E}[|X|]^2 \leq \mathbb{E}[X^2]$.

Remark 24. Sometimes, Cauchy-Schwarz makes sense even when $\mathbb{E}[X^2] = \infty$: we will get a trivial bound of type $x \leq \infty$.

Remark 25. The equality holds in Cauchy-Schwarz if and only if X and Y are almost surely colinear, i.e. there exists some $\alpha \in \mathbb{R}$ such that $\mathbb{P}[Y = \alpha X] = 1$.

Lemma 2 (Hölder). *Suppose that $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in L^q(\Omega, \mathcal{F}, \mathbb{P})$ with $p, q \in [1, \infty]$ and $1/p + 1/q = 1$. Then*

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

Remark 26. Using Hölder, we can recover Cauchy-Schwarz by choosing $p = q = 2$.

Remark 27. Writing $\mathbb{E}[|X|^r] \leq \mathbb{E}[(|X|^r)^{p/r}]$ and apply Hölder with $Y = 1$, we recover that $||X||_{L^r} \leq ||X||_{L^p}$ if $1 \leq r \leq p$: this also shows that $L^p \subset L^r$ if $r \leq p$. This is true even if $p = \infty$ (see below for the definition of the essential-sup norm).

Remark 28. The above remark also shows that the relation $1/p + 1/q \leq 1$ is enough for Hölder to hold in a probability space.

Sometimes q is called the *conjugate exponent*, but it is good to write down the relation everytime you use it.

Lemma 3 (Minkowski). *For $p \geq 1$, we have $||f+g||_{L^p} \leq ||f||_{L^p} + ||g||_{L^p}$. In other words, L^p with $p \geq 1$ is a normed space with norm $|| \cdot ||_{L^p}$.*

Recall that we systematically identify measurable functions which are equal almost everywhere. In particular, the essential sup-norm $|| \cdot ||_\infty$ should be defined as

$$||f||_\infty := \inf\{t \geq 0 \,;\, \mu(|f| > t) = 0\}.$$

Indeed, we can modify f on a set of measure 0 in such a way that the modified function $\tilde{f}$ is bounded by $||f||_\infty$ in absolute value.

Remark 29. When $0 < p < 1$, the useful inequality is $\mathbb{E}[(X+Y)^p] \leq \mathbb{E}[X^p] + \mathbb{E}[Y^p]$ for positive random variables X, Y. In fact, this comes the deterministic sub-additivity inequality that $(x + y)^p \leq x^p + y^p$ for $0 < p < 1$ and $x, y \geq 0$.

Lemma 4 (Jensen). *Suppose that X is a real random variable and $\varphi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ a* convex *function. Then*

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

The positivity assumption on φ is crucial and cannot be dropped.

Remark 30. Once again, by choosing the convex function $\varphi(x) = x^2$, we get $\mathbb{E}[|X|]^2 \leq \mathbb{E}[X^2]$ using a new method.

2.3 Cumulative distribution function

Definition 17 (Cumulative distribution function). Let X be a real random variable. The **cumulative distribution function** of X is defined as

$$F_X(t) := \mathbb{P}[X \leq t] = \mathbb{P}_X(]-\infty, t]), \quad t \in \mathbb{R}.$$

Sometimes it is called just *distribution function*.[1] Also, we will start to gradually stop using the notation $\mathbb{P}_X$ and prefer notations of type $\mathbb{P}[X \leq t]$.

[1] And worse, sometimes it is called *probability distribution function* and abbreviated as *p.d.f.*, which is the same for *probability density function*! Another reason for me to spell everything out in this note.

Proposition 3 (Distribution function determines the law). *The cumulative distribution function of a real random variable X characterizes the law of X.*

Proof. This can be shown by using Dynkin's lemma, cf. reading assignement of last week.

It is easy to see that F is increasing, has limit 0 at $-\infty$ and limit 1 at ∞, and $\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a)$ for $a < b$. The space of cumulative function is essentially characterized by its càdlàg property: see exercise.

Definition 18 (Continuous real random variable). We call a real random variable X **continuous** if its cumulative distribution function F_X is continuous.

In particular, any absolutely continuous real random variable is a continuous real random variable. In some texts, a continuous real random variable refers to an absolutely continuous real random variable (i.e. with probability density function), and by singular distribution they refer to the the somewhat pathological singlar continuous distribution. The terminology might differ, so it might be safe to specify based on the context.

Remark 31. A point of discontinuity where F_X is not continuous corresponds to an *atom* for a probability measure. The law of a continuous random variable has no atoms.

Definition 19 (Singular distribution). We call a real random variable X (continuous) **singular** if its cumulative distribution function F_X is continuous and singular. That is, F_X is continuous, non-constant and the derivative of F_X vanishes almost everywhere.

Remark 32. A typical example of continuous singular random variable is given by the so-called Cantor distribution.

With positive random variables, the following quantity is used more often:

Definition 20 (Tail distribution). The **tail distribution**, or **complementary cumulative distribution function**, is defined as

$$\overline{F}_X(t) := 1 - F_X(t) = \mathbb{P}[X > t].$$

The following inequality is of fundamental importance:

Lemma 5 (Markov's inequality). *Let $X \geq 0$. Then for all $a > 0$,*

$$\mathbb{P}[X > a] \leq \frac{\mathbb{E}[X]}{a}.$$

The following form is often useful: let $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be an increasing function, then for any $a > 0, \mathbb{P}[|X| > a] \leq \frac{\mathbb{E}[\varphi(|X|)]}{\varphi(a)}$. The proof for this follows by considering the auxiliary positive random variable $Y = \varphi(|X|)$. You will learn how to choose the good increasing function φ with experience: it often has to do with the "size" of the tail of the (positive) random variable X.

Proof. Consider $\mathbb{E}[X\mathbf{1}_{X>a}]$. On the one hand, it is smaller than $\mathbb{E}[X]$ by positivity of X. On the other hand, it is larger than $\mathbb{E}[a\mathbf{1}_{X>a}]$ since $X > a$ on the event $\{X > a\}$. It remains to use the linearity of the expectation to write $\mathbb{E}[a\mathbf{1}_{X>a}] = a\mathbb{P}[X > a]$.

Remark 33. One can show that $a\mathbb{P}[X > a] \to 0$ as $a \to \infty$, if $X \in L^1$. Indeed, in the above proof, we see that $\mathbb{E}[a\mathbf{1}_{X>a}] \leq \mathbb{E}[|X|]$ for all a, so that by dominated convergence with the integrable majorant $|X|$, $\mathbb{E}[a\mathbf{1}_{X>a}]$ converge to 0 as a goes to infinity.

Remark 34. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Applying Markov's inequality to the positive random variable $(X - \mathbb{E}[X])^2$ entails the Tchebyshev's inequality

$$\mathbb{P}[|X - \mathbb{E}[X]| > a] \leq \frac{\mathrm{var}(X)}{a^2}, \quad \forall a > 0,$$

where $\mathrm{var}(X)$ is the *variance* of X defined as $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

2.4 Characteristic function

The terminology *characteristic function* in probability is reserved to denote Fourier transforms of probability measure. Functions of type $\mathbf{1}_A$ (or χ_A) will be called *indicator functions*.

Definition 21 (Characteristic function). Let X be a real random variable. The **characteristic function** of X, denoted by $\Phi_X : \mathbb{R} \to \mathbb{C}$, is defined by

$$\Phi_X(\xi) = \mathbb{E}\left[e^{i\xi X}\right] = \int_{\mathbb{R}} e^{i\xi x} \mathbb{P}_X(dx).$$

Remark 35. By triangular inequality, $|\Phi_X|$ is bounded by 1. Later, we will see that Φ_X is uniformly continuous (by dominated convergence).

Remark 36. Suppose that X is an absolutely continuous real random variable with density function $p(x)$. In this case,

$$\Phi_X(\xi) = \int_{\mathbb{R}} e^{i\xi x} p(x) dx.$$

The following result is central, we will prove it later in the course.

Theorem 4 (Characteristic function determines the law). *The characteristic function of any real-valued random variable completely defines its probability distribution.*

To define (the law of) a random variable, it is equivalent to give its characteristic function. Therefore, it is important, at some point, to calculate the characteristic functions of all classical laws. Indeed, there are results that can be resumed as "after some calculations we have found a good-looking characteristic function, so we have identified the law"...

Example 4. Let N be a real random variable distributed as the standard Gaussian $\mathcal{N}(0,1)$. Then $\Phi_N(\xi) = e^{-\xi^2/2}$.

2.5 Some more topics

Here are some topics worth mentioning. It is not planned that we will make explicit use of them in this course, but this is only because we have a limited schedule.

1. L^p-spaces (! – functional analysis).
2. Moment generating function (and/or Laplace transform) (engineering).
3. Cumulant (statistics).

Exercise set I

Exercises marked with ! are important and those with ⋆ are difficult.

Exercise 1 (Absolute value). Let $f : \mathbb{R} \to \mathbb{R}$. Show that $|f|$ is Borel-measurable if and only if f is Borel-measurable. Is this exercise correct?

Exercise 2 (Operations on random variables). Let $X_1, \dots, X_n, \dots$ be a sequence of real random variables, all from the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Show that the followings are also random variables:

$$X_1 + X_2; \quad \sup_n X_n; \quad \liminf_n X_n.$$

Deduce that the set where $\lim_n X_n$ exists is measurable.

Exercise 3 (Variance). Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and define its *variance*

$$\operatorname{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Find a relation between $\mathbb{E}[(X - a)^2]$ and $\operatorname{var}(X)$ for $a \in \mathbb{R}$, and show that

$$\operatorname{var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E}[(X - a)^2].$$

Exercise 4 (! – Pointwise convergence does not imply convergence in mean). Give an example of a sequence of positive real random variables X_n, each with $\mathbb{E}[X] = 1$, where X_n converges pointwise to some limit X_∞ but $\mathbb{E}[X_\infty] = 0$.

Exercise 5 (! – Characteristic functions of a Gaussian variable). Calculate the characteristic function of a Gaussian variable $\mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$.

Exercise 6 (Moments of the semi-circular law). Consider the semi-circular law X with probability density function

$$f(x) = \frac{1}{2\pi}\sqrt{4 - x^2}, \quad x \in [-2, 2].$$

1. Calculate its mean.
2. Calculate its variance.
3. ($\star$) Calculate all positive integer moments of X and find a connection with the so-called Catalan numbers.

Exercise 7 ($\star$ – Càdlàg). The term "càdlàg" (*continue à droite, limite à gauche* in French) means *right continuous with left limit*. More precisely, f is càdlàg means that for all point $t \in \mathbb{R}$, the left limit $f(t-) \coloneqq \lim_{s \uparrow t} f(s)$ exists, and the right limit $f(t+) \coloneqq \lim_{s \downarrow t} f(s)$ (exists and) is equal to $f(t)$.

1. Show that all cumulative distribution functions are càdlàg functions.
2. Show that an increasing, càdlàg function F with limit 0 at $-\infty$ and limit 1 at ∞ is a cumulative distribution function.

Chapter 3
Independence

The notion of independence is fundamental and proper to probability theory. Our focus now shifts drastically from the previous chapter, since the study of independence is probably the principle argument against quotes of type "probability theory is just integration theory". To cite Durrett,

> "Measure theory ends and probability begins with the definition of independence."

3.1 Independence of events

Let us start by the simpliest example of independence.

Definition 22 (Independence of two events). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A pair of events A and B on this space are called independent if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B].$$

In Probability II, we will define the **conditional probability** of the event A knowing B as $\mathbb{P}[A|B] := \mathbb{P}[A \cap B]/\mathbb{P}[B]$, when $\mathbb{P}[B] > 0$. So A and B are independent if and only if $\mathbb{P}[A|B] = \mathbb{P}[A]$, or that "conditioning on B does not change the probability of A".

Remark 37. We cannot condition on an event B of probability 0, but there are certain ways of getting past this in more advanced course.

Remark 38. Notice that $\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$: this somewhat trivial observation is known as Bayes' theorem.

Remark 39. If the event A is independent of itself, then $\mathbb{P}[A] \in \{0, 1\}$. This is an example of a zero-one law that we will study in more detail later.

Remark 40 (Dynkin's lemma). This is an elementary example (arguably the most elementary one) that showcases the use of Dynkin's lemma. Indeed, consider two

events A and B, and suppose we know how to measure $\mathbb{P}[A]$ and $\mathbb{P}[B]$. Now, since $A \cap B$ is in the σ-algebra generated by A and B, can we determine $\mathbb{P}[A \cap B]$? If we can always determine it, then the definition of independence above has no meaning at all, so the answer must be *no*. But once the probability of the intersection $\mathbb{P}[A \cap B]$ is specified, we know how to calculate any probability in $\sigma(A, B)$.

The independence of more than two events is stronger than just pairwise independence. Indeed, the independence property assigns the probability measure on *all* intersections of different numbers of events.

Definition 23 (Independence of finitely many events). We call n events $A_1, \dots, A_n$ independent if for *any* subset $\{j_1, \dots, j_p\} \subset \{1, \dots, n\}$,

$$\mathbb{P}\left[\bigcap_{k=1}^{p} A_{j_k}\right] = \prod_{k=1}^{p} \mathbb{P}\left[A_{j_k}\right].$$

Notice that we require the factorization property on *all subsets* of indices.

Remark 41. Consider two independent fair coin tosses X_1 and X_2 and three events $A_1 = \{X_1 = 1\}$, $A_2 = \{X_2 = 1\}$, $A_3 = \{X_1 = X_2\}$. These events are pairwise independent (check the definition) but not independent as a whole (e.g. knowing A_1 and A_2 happen implies that A_3 happens).

One can rewrite the condition before slightly differently:

Proposition 4 (Independence of finitely many events bis). *The n events $A_1, \dots, A_n$ are independent if and only if for all $B_j \in \sigma(A_j) = \{\emptyset, A_j, (A_j)^c, \Omega\}$,*

$$\mathbb{P}\left[\bigcap_{j=1}^{n} B_j\right] = \prod_{j=1}^{n} \mathbb{P}\left[B_j\right].$$

The proof is a formal manipulation on sets and is omitted. Notice that the product above is indexed from 1 to n, but since some B_j can be Ω, it is still a requirement on all products indexed by subsets $\{j_1, \dots, j_p\}$ of $\{1, \dots, n\}$.

3.2 Independence of sigma-algebras and random variables: finite case

We now prepare for the definition of n independent random variables $X_1, \dots, X_n$. Intuitively, we want to say the knowing the outcome of some of $\{X_1, \dots, X_n\}$ does not give information on other random variables. The "independence of information" is encoded by the σ-algebras.

Definition 24 (Independence of finitely many σ-algebras). Let $\mathcal{B}_1 \ldots, \mathcal{B}_n$ be n sub-σ-algebras of the same σ-algebra $\mathcal{F}$. We say that $\mathcal{B}_1 \ldots, \mathcal{B}_n$ are independent if for all $(B_1, \ldots, B_n) \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_n$, we have

$$\mathbb{P}\left[\bigcap_{j=1}^{n} B_j\right] = \prod_{j=1}^{n} \mathbb{P}\left[B_j\right].$$

Remark 42. Notice that $\mathcal{B}_1 \times \cdots \times \mathcal{B}_n$ is *not* the product σ-algebra $\mathcal{B}_1 \otimes \cdots \otimes \mathcal{B}_n$!

The independence of random variables is then the independence of the σ-algebras that each of them generates. The technical assumption here is that the latter all belong to the same σ-algebra, so the random variables are defined on the same probability space.

Definition 25 (Independence of finitely many random variables). Let $X_1, \ldots, X_n$ be n random variables defined *on the same probability space* with $X_j : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_j, \mathcal{E}_j)$. Then $\{X_1, \ldots, X_n\}$ are independent if and only if $\{\sigma(X_1), \ldots, \sigma(X_n)\}$ are independent. In fact, this is equivalent to saying that, for all $(B_1, \ldots, B_n) \in \mathcal{E}_1 \times \cdots \times \mathcal{E}_n$,

$$\mathbb{P}[X_1 \in B_1, \ldots, X_n \in B_n] = \prod_{j=1}^{n} \mathbb{P}\left[X_j \in B_j\right].$$

The last equivalence follows from recalling that $\sigma(X_j) = \{X_j^{-1}(B), B \in \mathcal{E}_j\}$.

The following factorization property is essential in applications.

Proposition 5 (Joint law of independent random variables). *Let $X_1, \ldots, X_n$ be n random variables defined on the same probability space with different sample spaces, i.e. $X_j : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_j, \mathcal{E}_j)$. Then $(X_1, \ldots, X_n)$ is a random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(E_1 \times \cdots \times E_n, \mathcal{E}_1 \times \cdots \times \mathcal{E}_n)$. The variables $X_1, \ldots, X_n$ are independent if and only if*

$$\mathbb{P}_{(X_1, \ldots, X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}.$$

Furthermore, in this case, we have

$$\mathbb{E}\left[\prod_{j=1}^{n} f_j(X_j)\right] = \prod_{j=1}^{n} \mathbb{E}[f_j(X_j)]$$

for any n-tuple of positive measurable functions $f_j : (E_j, \mathcal{E}_j) \to (\mathbb{R}_{\geq 0}, \mathcal{B}(\mathbb{R}_{\geq 0}))$.

Proof. Recall that a measure μ on the product space $(E_1 \times \cdots \times E_n, \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n)$ is completely determined by the collection $\mu(B_1 \times \cdots \times B_n)$ for all $B_1 \times \cdots \times B_n \in \mathcal{E}_1 \times \cdots \times \mathcal{E}_n$: this is a consequence of Dynkin's lemma (this is actually how the product

measure is defined). Therefore, we only need to show $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$ on elements of type $B_1 \times \cdots \times B_n \in \mathcal{E}_1 \times \cdots \times \mathcal{E}_n$.

By definition,

$$\mathbb{P}_{(X_1,\dots,X_n)}(B_1 \times \cdots \times B_n) = \mathbb{P}[\{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\}],$$

$$\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}(B_1 \times \cdots \times B_n) = \mathbb{P}[X_1 \in B_1] \times \cdots \times \mathbb{P}[X_n \in B_n],$$

but they are equal by the definition of independence above.

We apply Fubini-Tonelli's theorem (i.e. Fubini for positive measurable functions) for the rest. First write

$$\mathbb{E}\left[\prod_{j=1}^{n} f_j(X_j)\right] = \int_{E_1 \times \cdots \times E_n} \prod_{j=1}^{n} f_j(x_j) \mathbb{P}_{(X_1,\dots,X_n)}(dx_1, \dots, dx_n).$$

Since $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$, Fubini's theorem allows us to factorize the integral into

$$\int_{E_1} \cdots \left(\int_{E_{n-1}} \left(\int_{E_n} \prod_{j=1}^{n} f_j(x_j) \mathbb{P}(dx_n) \right) \mathbb{P}(dx_{n-1}) \right) \dots \mathbb{P}(dx_1).$$

Succesively integrating, this is equal to $\prod_{j=1}^{n} \mathbb{E}[f_j(X_j)]$.

The moral here is "independence means multiply". Notice that if the f_j:s are not positive, the factorization still holds if all f_j:s are in L^1 by dominated convergence. In particular, if $X_1, \dots, X_n$ are independent random variables in L^1, then their product is also in L^1 and

$$\mathbb{E}[X_1 \times \cdots \times X_n] = \mathbb{E}[X_1] \times \cdots \times \mathbb{E}[X_n].$$

This is a very strong property! Indeed, it is not true that L^1 is stable under product.

3.3 Independence of real random variables: finite case

We specify the above discussions to the case of real random variables. It is important to note that

Lemma 6. *For all $d \geq 1$, $\mathcal{B}(\mathbb{R}^d) = \mathcal{B}(\mathbb{R})^{\otimes d}$.*

Proof. Admitted (see Section 8.5 of [Williams]).

We will be systematically using $\mathcal{B}(\mathbb{R})^{\otimes d}$ from now on: a generating π-system for this σ-algebra is $\{\prod_{j=1}^{d}(-\infty, a_j]\}_{(a_1,\dots,a_d)\in\mathbb{R}^d}$. A random vector $(X_1, \dots, X_n) \in \mathbb{R}^n$ is formed by n real random variables $X_1, \dots, X_n$: its law $\mathbb{P}_{(X_1,\dots,X_n)}$ defined on

$(\mathbb{R}^n, \mathcal{B}(\mathbb{R})^{\otimes n})$ is specified by the data on the above generating π-system. When $\mathbb{P}_{(X_1,\dots,X_n)} = p(x_1,\dots,x_n)dx_1\dots dx_n$, i.e. absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, we say that it has density $p(x_1,\dots,x_n)$.

Proposition 6 (Joint density and independence). *Let $X_1,\dots,X_n$ be n real random variables.*

1. *Suppose that $X_1,\dots,X_n$ are absolutely continuous, with respective density function p_j. Then the law of the random vector $(X_1,\dots,X_n)$ has density*

$$p(x_1,\dots,x_n) = \prod_j p_j(x_j).$$

2. *Suppose that the law of the random vector $(X_1,\dots,X_n)$ has density that can be written as*

$$p(x_1,\dots,x_n) = \prod_j q_j(x_j),$$

with positive Borel-measurable functions q_j. Then $X_1,\dots,X_n$ are independent, and for each $1 \le j \le n$, the law of X_j has density $p_j = C_j q_j$ with constant $C_j > 0$.

In short, the independence of real random variables with density can be translated into a factorization property.

Proof. The first part is a direct consequence of the factorization propery in the previous proposition. For the second part, we can use Fubini-Tonelli to calculate the density of X_j as

$$p_j(x_j) = \int_{\mathbb{R}^{n-1}} p(x_1,\dots,x_n)dx_1\dots dx_{j-1}dx_{j+1}dx_n$$

where we integrate over all variables but omit dx_j.

By assumption, $p(x_1,\dots,x_n)$ factorizes into products of $q_j(x_j)$, this is

$$\int_{\mathbb{R}^{n-1}} \prod_{j=1}^{n} q_j(x_j)dx_1\dots dx_{j-1}dx_{j+1}dx_n = \left(\prod_{m\neq j} K_m\right) q_j(x_j)$$

with $K_m := \int_{\mathbb{R}} p_m(x_m)dx_m$. As $p(x_1,\dots,x_n)$ is a probability density function, $\prod_{m=1}^n K_m = 1$, and $C_j := \prod_{m\neq j} K_m \neq 0$. This finishes the proof, since now $\mathbb{P}_{(X_1,\dots,X_n)} = \mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}$.

Remark 43. If X_1, X_2 are independent and in L^2, then their covariance $\operatorname{cov}(X_1,X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$ is 0. The converse statement is very wrong!

Next we record a more general consequence of independence of real random variables. This proposition can also serve as mean to shown independence between random variables.

Proposition 7 (Practical criteria for independence). *Let $X_1, \dots, X_n$ be n real random variables. There is equivalence between:*

1. *$X_1, \dots, X_n$ are independent.*
2. *For all $a_1, \dots, a_n \in \mathbb{R}$, $\mathbb{P}[X_1 \leq a_1, \dots, X_n \leq a_n] = \prod_{j=1}^n \mathbb{P}[X_j \leq a_j]$.*
3. *If $f_1, \dots, f_n$ are continuous, compact supported functions from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$, then $\mathbb{E}[\prod_{j=1}^n f_j(X_j)] = \prod_{j=1}^n \mathbb{E}[f_j(X_j)]$.*
4. *The characteristic function of $X = (X_1, \dots, X_n)$ is $\Phi_X(\xi_1, \dots, \xi_n) = \prod_{j=1}^n \Phi_{X_j}(\xi_j)$.*

Proof. We have seen (1) $\implies$ (2) $\implies$ (3) $\implies$ (4). For (4) $\implies$ (1), one uses the injectivity of the Fourier transform on the space of probability measures: a rigorous proof will be provided later in this course.

In practise, the third item above is what one uses most about independent random variables, but the fourth item is also helpful when the characteristic functions are easy to calculate.

Remark 44. The quantity in the second item above, as a generalized cumulative distribution function, characterizes the law of $(X_1, \dots, X_n)$. The quantity in the fourth item also chracterizes the law of $(X_1, \dots, X_n)$ (notice that this is a higher dimensional version of Fourier transform, and only knowning the diagonal values $\Phi_X(\xi, \dots, \xi)$ for all $\xi \in \mathbb{R}$) is not enough. Together, they provide useful checks of independence.

3.4 Sum of two independent real random variables

Sums of independent random variables will be an important subject that occupy the last part of this course. We now study the sum of two independent random variables.

Proposition 8 (Characteristic function and sum of independent random variables). *Let X and Y be two independent real random variables. Then the characteristic function of $X + Y$ is $\Phi_{X+Y}(\xi) = \Phi_X(\xi)\Phi_X(\xi)$.*

Proof. This follows from $\Phi_{X+Y}(\xi) = \mathbb{E}[e^{i\xi(X+Y)}] = \mathbb{E}[e^{i\xi X}]\mathbb{E}[e^{i\xi Y}] = \Phi_X(\xi)\Phi_Y(\xi)$, where the factorization is justified by the independence between X and Y.

Given two probability measures μ and ν on $\mathbb{R}$, recall that we can define their convolution measure $\mu * \nu$ on $\mathbb{R}$ such that for all measurable functions $\varphi : \mathbb{R} \to \mathbb{R}_{\geq 0}$,

$$\int_{\mathbb{R}} \varphi(z)\mu * \nu(dz) = \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(x + y)\mu(dx)\nu(dy).$$

Proposition 9 (Convolution and sum of independent random variables). *Let X and Y be two independent real random variables. Then the law of $X + Y$ is $P_X * P_Y$.*

In particular, if X and Y are continuous with density functions respectively p_X and p_Y, then $X+Y$ is continuous and has density function $p_X * p_Y$ (the last $*$ is stricto sensu *the convolution of functions instead of measures).*

Proof. To calculate the law of $X+Y$, we study $\mathbb{E}[f(X+Y)]$ for all positive measurable functions f. Recall that, by independence, $\mathbb{P}_{(X,Y)}(dxdy) = \mathbb{P}_X(dx)\mathbb{P}_Y(dy)$. We have

$$\mathbb{E}[f(X+Y)] = \int_{\mathbb{R}^2} f(x+y)\mathbb{P}_{(X,Y)}(dxdy) = \int_{\mathbb{R}^2} f(x+y)\mathbb{P}_X(dx)\mathbb{P}_Y(dy),$$

but the last expression is $\int_{\mathbb{R}^2} f(z)\mathbb{P}_X * \mathbb{P}_Y(dz)$ by definition.

In the case where $\mathbb{P}_X(dx) = p_X(x)dx$ and $\mathbb{P}_Y(dy) = p_Y(y)dy$, the last expression above is

$$\int_{\mathbb{R}^2} f(x+y)p_X(x)p_Y(y)dxdy = \int_{\mathbb{R}} f(z)\left(\int_{\mathbb{R}} p_X(x)p_Y(z-x)dx\right)dz$$

where we used the change of variables $z = x+y$. We recognize the convolution of two L^1 functions in the parenthesis, namely $(p_X * p_Y)(z)$.

Example 5. The sum of two independent uniform distributions $\mathcal{U}([0,1])$ uniform has density $p(x) = (1-|x-1|)_+$.

Chapter 4
Sequence of infinitely many random variables

We often want to study an infinite sequence of random variables and say something about its long-term behavior. It would be great interest to be able to predict that something is almost surely going to happen in the future. We give two important results of this type: Borel-Cantelli's lemma(s) and Kolomogorov's zero-one law.

4.1 Independence of an infinite family of random variables

Given n independent random variables $X_1, \dots, X_n$, one can separate them into two collections $Y_1 = (X_1, \dots, X_p)$ and $Y_2 = (X_{p+1}, \dots, X_n)$. It is intuitively clear that Y_1 and Y_2 should be independent (as random vectors), but the justification takes some work. First, we need a technical measure theory lemma.

Lemma 7 (Independence of collections and generated σ-algebras). *We say that a collection of sets (not-necessary σ-algebras) $\mathcal{A}_1, \dots, \mathcal{A}_n$, each containing Ω, is independent if for any $A_j \in \mathcal{A}_j$,*

$$\mathbb{P}\left[\bigcap_{j=1}^{n} A_j\right] = \prod_{j=1}^{n} \mathbb{P}[A_j].$$

If furthermore, each $\mathcal{A}_j$ is a π-system, i.e. stable by finite intersection, then they generate independent σ-algebras $\sigma(A_1), \dots, \sigma(A_n)$.

Proof. As you might suspect, the proof relies on Dynkin's lemma. We follow Theorem 2.1.7 in [Durrett]: it suffices to prove, with the notations above, that $\sigma(A_1), A_2, \dots, A_n$ are independent, and then finish the proof by induction.

For this, notice that $\{A \in \mathcal{F} \ ; \ A \text{ independent of } A_2, \dots, A_n\}$ is a λ-system that contains A_1. Since A_1 is a π-system, Dynkin's lemma shows that $\sigma(A_1)$ is contained in the above λ-system, so that $\sigma(A_1)$ is independent of $A_2, \dots, A_n$.

The following corollary is very practical.

Corollary 1 (Independence by blocks). *Let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be independent σ-algebras. Let $0 = n_0 < n_1 < \cdots < n_p = n$ and divide them into blocks with different indices: $\mathcal{D}_1 = \sigma(\mathcal{B}_1, \ldots, \mathcal{B}_{n_1})$, $\mathcal{D}_2 = \sigma(\mathcal{B}_{n_1+1}, \ldots, \mathcal{B}_{n_2})$, $\ldots$, $\mathcal{D}_p = \sigma(\mathcal{B}_{n_{p-1}+1}, \ldots, \mathcal{B}_n)$. Then the σ-algebras $\mathcal{D}_1, \ldots, \mathcal{D}_p$ are independent.*

In particular, if $X_1, \ldots, X_n$ are independent random variables, the random vectors $Y_1 = (X_1, \ldots, X_p)$ and $Y_2 = (X_{p+1}, \ldots, X_n)$ are independent.

Proof. Exercise (use the previous lemma). See also Theorem 2.1.9 of [Durrett].

Remark 45. Given three independent random variables X_1, X_2, X_3, the random variables $\exp(X_1)$ and $X_2 \cdot X_3$ are independent, since they are measurable functions of X_1 and (X_2, X_3) which are independent.

We generalize the above discussion to define independence of an infinite family of random variables $\{X_j\}_{j \in \mathbb{Z}_{>0}}$. As usual, start with the σ-algebras:

Definition 26 (Independence of an infinite family of σ-algebras). Let $(\mathcal{B}_j)_{j \in J}$ be an (possibly uncountably) infinite family of σ-algebras of $\mathcal{F}$. We say that this family of σ-algebras is independent if and only if every finite collection $(\mathcal{B}_{j_1}, \ldots, \mathcal{B}_{j_p})$ is independent (as a finite family of σ-algebras defined in the previous chapter).

Definition 27 (Independence of an infinite family of random variables). Let $X_1, X_2, \ldots$ be an infinite family of random variables defines on the same probability space. We say that this family of random variables is independent if and only if their σ-algebras are independent (as in the previous definition).

Remark 46. Another way of formulating the previous definition is to say that an infinite family of random variables is independent if and only if every finite collection of this family of random variables is independent.

Example 6 (Lebesgue's construction of independent coin tosses). Consider $X \sim \mathcal{U}[0, 1]$ and write it in the dyadic bases, $X(\omega) = \sum_{j \geq 1} \epsilon_j(\omega) 2^{-j}$. Then $(\epsilon_j)_{j \geq 1}$ is an infinite i.i.d. sequence of random fair coin tosses with value in $\{0, 1\}$. This construction does not use Kolmogorov's extension theorem.

4.2 Infinite sequences of events

To start our discussion, we need to review some materials on infinite sequences. Not only of numbers, but of sets in general.

Definition 28 (Limits of sets). Recall that a σ-algebra X is closed under countable unions and intersections. Let $A_1, \ldots, A_n, \ldots$ be a sequence of sets in X. Then the following sets are also in the σ-algebra X:

1. **Limit supremum**: $\limsup_n A_n = \bigcap_{n=1}^{\infty} \left(\bigcup_{m=n}^{\infty} A_m \right)$.
2. **Limit infimum**: $\liminf_n A_n = \bigcup_{n=1}^{\infty} \left(\bigcap_{m=n}^{\infty} A_m \right)$.

The difference with the more familiar lim sup and lim inf of sequences is that, the inclusion of sets is a partial order, while the comparaison of real numbers is a total order. The above quantities control the fluctuation of the sequence sets in the limits.

Remark 47. It is easy to see that $\liminf_n A_n$ is always smaller than $\limsup_n A_n$ (in the sense of inclusion). If $\liminf_n A_n = \limsup_n A_n$, then we can define $\lim_n A_n$ as this common limit set.

An important remark is that, by continuity of probability measures from below and from above, we have

$$\mathbb{P}\left[\limsup_{n\to\infty} A_n\right] = \lim_{n\to\infty} \downarrow \mathbb{P}\left[\cup_{m=n}^{\infty} A_m\right],$$

and similarly for lim inf (this one is used less often, and so is left as an exercise).

To explain the definition with common words, an element x is in the set $\limsup_n A_n$ if and only if for arbitrarily large n_0, x appears in some set A_m with $m \geq n_0$. A more commonly used definition in probability is:

Proposition 10 ("Infinitely often"). *Let $E_1, \ldots, E_n, \ldots$ be a sequence of events in some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $\limsup_n E_n$ is the set of outcomes that occur infinitely many times with the sequence E_n.*

One can write *i.o.* for infinitely often. You should have already encountered this for deterministic sequences, but this notion is of particular interest for probabilists.

Example 7. As a deterministic example, prime numbers appear infinitely often in the sequence of natural numbers. Actually, there are very nice probabilistic models of prime numbers, starting with the so-called Cramér's model. Google it!

We now extend the "grouping independent random variables by blocks" property to the infinite case.

Proposition 11 (Independence of an infinite family of random variables). *Let $X_1, \ldots, X_n, \ldots$ be independent random variables. Then the random vectors $Y_1 = (X_1, \ldots, X_p)$ and $Y_2 = (X_{p+1}, X_{p+2}, \ldots)$ are independent.*

Proof. We should show that $\mathcal{B}_1 = \sigma(X_1, \ldots, X_p)$ and $\mathcal{B}_2 = \sigma(X_{p+1}, X_{p+2}, \ldots)$ are independent σ-algebras. Notice that $\mathcal{B}_2$ is generated by $\mathcal{D}_j = \sigma(X_{p+1}, \ldots, X_{p+j})$, each of them independent of $\mathcal{B}_1$. Furthermore, $\cup_{j\geq 1} \mathcal{D}_j$ is a π-system that contains Ω. Applying a lemma above shows that $\mathcal{B}_2 = \sigma(\cup_{j\geq 1} \mathcal{D}_j)$ is also independent of $\mathcal{B}_1$.

4.3 Borel-Cantelli lemma(s)

There are two versions of Borel-Cantelli lemma for the lim sup of an infinite sequence of events.

Theorem 5 (First Borel-Cantelli lemma). *Let $(A_j)_{j\geq 1}$ be a sequence of events in $(\Omega, \mathcal{F}, \mathbb{P})$. If $\sum_{j\geq 1} \mathbb{P}[A_j] < \infty$, then $\mathbb{P}[\limsup_j A_j] = 0$, i.e. the probability that infinitely many of A_j occur is 0.*

Proof. It is enough to show that $\lim_{j_0\to\infty} \mathbb{P}[\cup_{j\geq j_0} A_j] = 0$. By union bound, it suffices if $\sum_{j\geq j_0} \mathbb{P}[A_j] \to 0$ as $j_0 \to \infty$. But this is true since the series $\sum_{j\geq 1} \mathbb{P}[A_j]$ converges.

Notice that the following lemma supposes independence while the previous one does not.

Theorem 6 (Second Borel-Cantelli lemma). *Let $(A_j)_{j\geq 1}$ be a sequence of* independent *events in $(\Omega, \mathcal{F}, \mathbb{P})$. If $\sum_{j\geq 1} \mathbb{P}[A_j] = \infty$, then $\mathbb{P}[\limsup_j A_j] = 1$, i.e. the probability that infinitely many of A_j occur is* 1.

Proof. To show that $\mathbb{P}[\limsup_j A_j] = 1$, we should show that $\mathbb{P}[\cup_{j\geq m} A_j] = 1$ for all $m \geq 1$. To estimate the latter, we pass to the complement and show that $\mathbb{P}[\cap_{j\geq m}(A_j)^c] = 0$ for all $m \geq 1$. Using independence of A_j, we have $\mathbb{P}[\cap_{j\geq m}(A_j)^c] = \prod_{j\geq m} \mathbb{P}[(A_j)^c] = \prod_{j\geq m}(1 - \mathbb{P}[A_j])$.

Since $\sum_{j\geq 1} \mathbb{P}[A_j] = \infty$, we also have $\sum_{j\geq m} \mathbb{P}[A_j] = \infty$, and since $1 - \mathbb{P}[A_j] \leq e^{-\mathbb{P}[A_j]}$, we get $\prod_{j\geq m}(1 - \mathbb{P}[A_j]) \leq e^{-\sum_{j\geq m} \mathbb{P}[A_j]} = 0$. This finishes the proof.

Remark 48. The independence assumption in the second Borel-Cantelli lemma cannot be dropped: consider the extreme case where the sequence is completely correlated, i.e. $X_i = X_j$ for all $i \neq j$. One verifies that the sequence of events $A_j \equiv A$ with $0 < \mathbb{P}[A] < 1$ happens infinitely often with probability $\mathbb{P}[A] < 1$, but $\sum_{j\geq 1} \mathbb{P}[A_j] = \infty$.

4.4 Kolmogorov's zero-one law

Suppose that $\mathcal{F}$ is the σ-algebra generated by an infinite sequence of random variables $\{X_j\}_{j\in\mathbb{Z}_{>0}}$. We call $A \in \mathcal{F}$ a **tail event** (or sometimes **asymptotic event**) if it is independent of each finite subset of the random variables $\{X_j\}_{j>0}$. In other words:

Definition 29 (Tail σ-algebra). Consider a sequence of random variables $\{X_j\}_{j\in\mathbb{Z}_{>0}}$ and let $\mathcal{F}_n = \sigma(X_n, X_{n+1}, \dots)$. The *tail σ-algebra*, denoted by $\mathcal{T}$, is the intersection $\mathcal{T} \coloneqq \cap_{n\geq 1}\mathcal{F}_n$.

The idea is that, $\mathcal{F}_n$ contains the "information after time n", and $\mathcal{T}$ is the "information in the remote future". Changing the value of a finite number of X_j does not affect the outcome of a tail event $A \in \mathcal{T}$.

Remark 49. For an infinite sequence of events E_n, the event "E_n happens infinitely often" is a tail event. Let us verify this properly with the definition. Recall that $\limsup_n A_n = \cap_{n=1}^{\infty} \left(\cup_{m=n}^{\infty} A_m\right)$. Now, $\cup_{m=n}^{\infty} A_m \in \mathcal{F}_n$, and $\limsup_n A_n \in \cap_{n=1}^{\infty} \mathcal{F}_n = \mathcal{T}$ is indeed a tail event.

Remark 50. The event "$\{X_j\}_{j \in \mathbb{Z}_{>0}}$ is bounded" is a tail event. Intuitively, this event does not depend on the realizations of finitely many terms of $\{X_j\}_{j \geq 1}$.

Tail events appear in the study of boundness, convergence, recurrence etc., where the answer to the question is independent of finitely many entries. The remarkable property about a tail event is:

Theorem 7 (Kolmogorov's zero-one law). *Let $X_1, X_2, \dots$ be independent random variables and $\mathcal{T}$ its tail σ-algebra. Then any tail event $A \in \mathcal{T}$ has probability either* 0 *or* 1.

Notice that we assume *independence* of the random sequence.

Proof. The (somewhat surprising) idea of that such a tail event $A \in \mathcal{T}$ is *independent of itself*! From this, you get $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$, so that $P[A] \in \{0, 1\}$.

We have already seen that, for all $k \geq 1$, the σ-algebras $\sigma(X_1, \dots, X_k)$ and $\sigma(X_{k+1}, X_{k+2}, \dots)$ are independent. We will generalize the argument to show that $\sigma(X_1, X_2, \dots)$ and $\mathcal{T}$ are independent. Already, $\sigma(X_1, \dots, X_k)$ and $\mathcal{T}$ are independent since $\mathcal{T} \subset \sigma(X_{k+1}, X_{k+2}, \dots)$. But $\cup_{k \geq 1} \sigma(X_1 \dots, X_k)$ is a π-system (and it is equal to $\sigma(X_1, X_2, \dots)$ that contains Ω), so by the same argument as in the beginning of this chapter, $\sigma(X_1, X_2, \dots)$ is independent of $\mathcal{T}$ (strange, isn't it!).

Now if $A \in \mathcal{T}$, as A is also in $\sigma(X_1, X_2, \dots)$, A is independent of itself.

Remark 51. As a σ-algebra, $\mathcal{T}$ is independent of $\mathcal{T}$.

Remark 52. Let $X_1, X_2, \dots$ be independent random variables. Show that if the probability that the sequence $(X_1, X_2, \dots)$ converges is greater than 10^{-100}, then it converges almost surely. Indeed, the event "$\{X_n\}_{n \geq 1}$ converges" is in the tail $\mathcal{T}$, but its probability cannot be 0 by assumption, so its probability must be 1 by Kolmogorov's zero-one law.

Exercise set II

Exercises marked with ! are important and those with ⋆ are difficult.

Exercise 8 (Sum of two independent Gaussian random variables). Consider two independent Gaussian random variables $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Show that $X_1 + X_2$ is distributed as $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Exercise 9 (! – Gaussian vector or not, that is the question). A random vector $(X_1, X_2) \in \mathbb{R}^2$ is called a *Gaussian vector* if every linear combination of $\{X_1, X_2\}$ is a Gaussian random variable.

1. Suppose that (X_1, X_2) is a Gaussian vector with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ (we say that X_1, X_2 are *centered*). Does $\mathbb{E}[X_1 \cdot X_2] = 0$ imply that X_1 and X_2 are independent?
2. Answer the same question above when (X_1, X_2) is not necessarily a Gaussian vector (but each distributed as a centered Gaussian). Hint: you can e.g. "change the sign of a Gaussian with a coin toss".
3. Show that if (X_1, X_2) is a centered Gaussian vector (i.e. Gaussian vector with 0-mean components), its law is completely characterized by the 2×2 matrix $A = (a_{ij})_{i,j \in \{1,2\}}$ with $a_{ij} = \mathbb{E}[X_i \cdot X_j]$. Hint: this is a Hilbert space problem.

Exercise 10 (Minimum of independent exponential random variables). Take two independent random variables $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$ with $\lambda_1, \lambda_2 > 0$. Consider $X = \min(X_1, X_2)$.

1. Show that X is an exponential random variable. Calculate its parameter.
2. Generalize your answer to the minimum of n independent exponential distributions.
3. Does the question work with the maximum?
4. Use the information above to calculate the mean of $\max(X_1, X_2)$.

Exercise 11 (Maximum of Gaussian random variables). Take a sequence of n standard normal random variables $X_1, \ldots, X_n$, each distributed as $X \sim \mathcal{N}(0, 1)$. Denote by $Z = \max(X_1, \ldots, X_n)$. We want to estimate the size of $\mathbb{E}[Z]$.

1. Using Markov inequality to show that $\mathbb{P}[X > a] \leq \exp(-a^2/2)$ for any $a > 0$. Hint: consider $\mathbb{E}[e^{\gamma X}]$.
2. Write $\mathbb{E}[Z] = \frac{1}{\beta}\mathbb{E}[\ln(e^{\beta Z})]$. Deduce that for all $\beta > 0$,

$$\mathbb{E}[Z] \leq \frac{1}{\beta}\mathbb{E}\left[\ln\left(\sum_{j=1}^{n} e^{\beta X_j}\right)\right].$$

3. Apply Jensen's inequality, then optimize on β to conclude that $\mathbb{E}[Z] \leq \sqrt{2\ln n}$.
4. Did we assume the independence of $(X_1, \ldots, X_n)$ somewhere? What can you say if we assume that they are independent ($\star$)?

Exercise 12 (Factorization of planar Gaussian variable). Let U, V be independent random variables with $U \sim \text{Exp}(1)$ and $V \sim \mathcal{U}([0,1])$. Define $(X, Y) = (\sqrt{U}\cos(2\pi V), \sqrt{U}\sin(2\pi V))$.

1. Show that X, Y are independent and identically distributed as $\mathcal{N}(0, 1/2)$. For this, you can try to show that the density function of (X, Y) factorizes.
2. Explain the title of this exercise.

Exercise 13 (Positive random series). Consider a sequence of independent positive random variables $U_1, \ldots, U_n, \ldots$ and investigate the series $\sum_1^\infty U_n$.

1. What are the possible values for $\mathbb{P}\left[\sum_1^\infty U_n < \infty\right]$ using the zero-one law?
2. What can be said about the previous question if $\mathbb{E}\left[\sum_1^\infty U_n\right] < \infty$?
3. Take $U_n = 0$ with probability $1-2^{-n}$ and $U_n = 2^n$ with probability 2^{-n}. Calculate $\mathbb{E}[U_n]$, then $\mathbb{E}\left[\sum_1^\infty U_n\right]$. However, use the Borel-Cantelli lemma to conclude that $\sum_1^\infty U_n < \infty$ almost surely.

[Kahane – Some Random Series of Functions (2nd ed.), p32]

Exercise 14 (Just for fun – a proof by game). Use a fair coin to prove the following identity:

$$\frac{1}{4} + \frac{1}{8} + \frac{2}{16} + \frac{3}{32} + \frac{5}{64} + \cdots = 1. \tag{4.1}$$

where the numerators are given by the Fibonacci sequence.
Hint: consider the following game (for one player) with the diagram

$$\text{Start} \longrightarrow \text{Nothing Happens} \longrightarrow \text{Go Back to Start} \longrightarrow \text{Finish}$$

and the rules:

1. If you hit "Head", advance two steps;
2. If you hit "Tail", advance one step.

To be rigorous, you might want to show that the game stops with probability 1.
[Litchfield – Mathematics Magazine, Vol.67, No.4]

Chapter 5
Convergence in law

The central limit theorem says that, in many cases, renormalized sum of independent random variables "behaves like" a Gaussian random variable. To establish this result, we should define properly what does "behaves like" mean in probability with the definition of convergence in law, and we will study a simple proof with the use of characteristic functions, i.e. Fourier transform of probability measures.

5.1 Convergence of probability measures

In this chapter, we work with probability measures, denoted as μ or $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the latter if it is induced by some real random variable X. This is equivalent to specifying the distribution function $F_X(a) = \mathbb{P}[X \leq a]$ for $a \in \mathbb{R}$: recall also that a distribution function is essentially characterized by its càdlàg property.

Let us already notice that different real random variables can induce the same law, and in the following, we don't specify the probability space of the random variable if only the law of the latter is of interest: this is an abus of language that is allowed only in this chapter.

Definition 30 (Weak convergence of distribution functions). Let $\{F_n\}_{n\geq 1}$ and F be distribution functions on $\mathbb{R}$. We say that F_n converges **weakly** towards F if for all $y \in \mathbb{R}$ continuous point of F, we have $F_n(y) \to F(y)$ as $n \to \infty$.

The above definition is often replaced by the following more practical one in many applications. The easiest proof of these results is to go by the Skorokhod representation theorem and uses the notion of almost sure convergence, that we might come back to in the last chapter, time permitting.

Definition 31 (Weak convergence of probability measures). Let $\mu_1, \mu_2, \ldots$ and μ be probability measures on $\mathbb{R}$. We say that $\{\mu_n\}_{n\geq 1}$ converges **weakly** to μ if and only if, for every bounded continuous function $f \in C_b(\mathbb{R})$, we have $\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx)$ as $n \to \infty$.

Definition 32 (Weak convergence of real random variables). Let $X_1, X_2, \ldots$ and X be real random variables. We say that $\{X_n\}_{n\geq 1}$ converges **weakly** to X if and only if, for every bounded continuous function $f \in C_b(\mathbb{R})$, we have $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$.

Proof. See Theorem 3.2.9 of [Durrett].

We usually say in this case that the sequence of random variables $\{X_n\}_{n\geq 1}$ converges **in law** or **in distribution** towards the random variable X, and denote this convergence by $\xrightarrow{(d)}$.

Remark 53. In the definition of the convergence in law, we do not require that the random variables X_n are defined on the same probability space. This is not the case for the the convergence in probability or the almost sure convergence that we will see in the coming weeks.

Remark 54. In the case of probability measures or random variables, the space of test functions $C_b(\mathbb{R})$ can be replaced by $C_c(\mathbb{R})$, the space of compactly supported continuous functions on $\mathbb{R}$. The reason behind is because when the total mass $\mu(\mathbb{R})$ is fixed, the notion of weak convergence coincides with that of the so-called vague convergence.

We now give some useful conditions for showing weak convergence of probability measures (which we can avoid using in this course).

Theorem 8 (Continuous mapping theorem). *Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function and D_g the set of points $x \in \mathbb{R}$ where g is discontinuous. If X_n converges in law to X and $\mathbb{P}[X \in D_g] = 0$, then $g(X_n)$ converges in law to $g(X)$. Furthermore, if g is bounded, then $\mathbb{E}[g(X_n)]$ converges to $\mathbb{E}[g(X)]$.*

Proof. See Theorem 3.2.10 of [Durrett].

The next theorem gives some equivalent ways of checking the weak convergence.

Theorem 9 (Portmanteau's theorem). *The following statements are equivalent:*

1. *The sequence of random variables $\{X_n\}_{n\geq 1}$ converges in law to X;*
2. *For all open sets G, $\liminf_n \mathbb{P}[X_n \in G] \geq \mathbb{P}[X \in G]$;*
3. *For all closed sets F, $\limsup_n \mathbb{P}[X_n \in F] \leq \mathbb{P}[X \in F]$;*
4. *For all Borel sets A with $\mathbb{P}[X \in \partial A] = 0$, $\lim_n \mathbb{P}[X_n \in A] = \mathbb{P}[X \in A]$, where ∂A is the boundary of A.*

Proof. See Theorem 3.2.11 of [Durrett].

Finally, it is useful to point out that the weak convergence of measures has a topological definition:

Theorem 10 (Lévy's metric). *The function*

$$\rho(F, G) = \inf\{\epsilon \leq 0 \ : \ F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \text{ for all } x\}$$

defines a metric on the space of distribution functions and $\rho(F_n, F) \to 0$ if and only if F_n converges weakly to F.

Proof. Exercise.

5.2 Characteristic function

Recall that the characteristic function of a real random variable X is just the Fourier transform $\Phi_X(\xi) = \mathbb{E}[e^{i\xi X}]$. In other words, the characteristic function of a probability measure μ is just $\Phi_\mu(\xi) = \int_{\mathbb{R}} e^{i\xi x}\mu(dx)$. The study of characteristic functions in probability can very well be a chapter on its own because of its importance.

Proposition 12 (Elementary properties of characteristic functions). *Let X be a real random variable and Φ_X its characteristic function.*

1. *For all $\xi \in \mathbb{R}$, $|\Phi_X(\xi)| \leq \Phi_X(0) = 1$.*
2. *For all $\xi \in \mathbb{R}$, $\overline{\Phi_X(\xi)} = \Phi_X(-\xi) = \Phi_{-X}(t)$.*
3. *The function Φ_X is uniformly continuous.*
4. *Let $a, b \in \mathbb{R}$. Then $\Phi_{aX+b}(\xi) = e^{i\xi b} \cdot \Phi_X(a\xi)$.*

Proof. Exercise.

We shall prove several important results about the characteristic function. In particular, we will show the inversion formula, which justifies its name that the characteristic function completely characterizes the law of a real random variable.

Theorem 11 (The inversion formula). *Let μ be a probability measure on $\mathbb{R}$ and Φ_μ its characteristic function. Then for all interval $(a, b) \subset \mathbb{R}$,*

$$\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-i\xi a} - e^{-i\xi b}}{i\xi}\Phi_\mu(\xi)d\xi.$$

Here $\mu(\{a, b\})$ denotes the point masses of μ at a and b. In particular, if μ is continuous,

$$\mu((a,b)) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-i\xi a} - e^{-i\xi b}}{i\xi}\Phi_\mu(\xi)d\xi.$$

In other words, there is a bijection between probability measures on $\mathbb{R}$ and characteristic functions. Here's a physical interpretation of the formula. Plancherel's

theorem tells us that the Fourier transform is an L^2 isometry, and in particular, for all $f, g \in L^2$, we have $\int_{\mathbb{R}} f\overline{g} = \int_{\mathbb{R}} \hat{f}\overline{\hat{g}}$ by polarization (you can check this by Fubini directly). Apply this to $\mathbf{1}_{(a,b)}$ and μ and we get almost the identity above, except for points a and b at which $\mathbf{1}_{(a,b)}$ is discontinuous and we need to adjust them by the principal values.

To make the above heuristics into an actual mathematical proof, some regularization procedure should be applied. The actual choice of the regularization is a matter of personal taste (the inversion formula is itself a principal value formula).

Proof. The idea of the proof is to examine the above heuristic about the Fourier transform of the indicator function $\mathbf{1}_{(a,b)}$ via Fubini's theorem. Consider

$$\begin{aligned} I_T &= \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-i\xi a} - e^{-i\xi b}}{i\xi}\Phi_\mu(\xi)d\xi \\ &= \frac{1}{2\pi}\int_{-T}^{T}\int_{\mathbb{R}} \frac{e^{-i\xi a} - e^{-i\xi b}}{i\xi} e^{i\xi x}\mu(dx)d\xi \\ &= \frac{1}{2\pi}\int_{-T}^{T}\int_{\mathbb{R}}\int_{a}^{b} e^{-i\xi y}e^{i\xi x}dy\mu(dx)d\xi \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_a^b\int_{-T}^{T} e^{-i\xi(y-x)}d\xi dy\right)\mu(dx), \end{aligned}$$

where the use of Fubini for fixed T, a, b is allowed since the modulus of the integrand in the last expression is always bounded by 1.

It follows that, to prove the inversion formula, we only need to study the integral

$$J_T(x) = \int_a^b\int_{-T}^{T} e^{-i\xi(y-x)}d\xi dy = \int_{a-x}^{b-x}\int_{-T}^{T} e^{-i\xi y}d\xi dy$$

and show that $\lim_{T\to\infty} J_T(x) = 2\pi\mathbf{1}_{(a,b)}(x) + \pi\mathbf{1}_{\{a,b\}}(x)$.

For this, we need to use the symmetry of the interval $[-T, T]$. Integrating the variable ξ over $[-T, T]$, we have

$$J_T(x) = \int_{a-x}^{b-x}\frac{e^{-iTy} - e^{iTy}}{-iy}dy = \int_{a-x}^{b-x}\frac{2\sin(Ty)}{y}dy = 2\int_{T(a-x)}^{T(b-x)}\frac{\sin(y)}{y}dy.$$

where we used a change of variables $Ty \mapsto y$ in the last step. Now use the value of the improper integral

$$\lim_{T\to\infty}\int_0^T \frac{\sin(y)}{y}dy = \frac{\pi}{2}$$

to conclude.

Remark 55. That the limit in the inversion formula (a sort of principal value at infinity) exists is a non-trivial fact. Compare this with the Fourier series of a periodic function (of bounded variation) at a point of discontinuity.

Remark 56 (A partial converse result). One can check using the inversion formula that $\Phi_X = \Phi_{-X}$ if and only if X and $-X$ are equal in law, i.e. when X is symmetric.

Remark 57. One can prove the integral formula for the sinc function

$$\int_0^\infty \frac{\sin(x)}{x} dx = \frac{\pi}{2}$$

using complex analysis: see Exercise E.16.1 of [Williams] for a hint. There is at least another approach by considering

$$G(t) = \int_0^\infty \frac{\sin(x)}{x} e^{-tx} dx$$

and show that $G'(t) = -\frac{1}{1+t^2}$.

Corollary 2 (Characterization of the law). *Let μ, ν be two probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $\Phi_\mu = \Phi_\nu$, then $\mu = \nu$.*

Proof. Let C be the set of points where both μ and ν are continuous: its complement $D = \mathbb{R} \setminus C$ is at most countable. For all $a, b \in C$, the previous theorme shows that $\mu((a,b)) = \nu((a,b))$. It remains to see that intervals of type (a,b) with $a, b \in C$ is a π-system that generates $\mathcal{B}(\mathbb{R})$.

By studying the characteristic function, one can also gain a lot of information about the probability distribution without explicitly identifying it. We give a list of some basic examples: unless otherwise specified, we denote below by μ a probability measure on $\mathbb{R}$ and Φ_μ its characteristic function.

Proposition 13 (Atoms from the characteristic function). *For all $a \in \mathbb{R}$, we have*

$$\mu(\{a\}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-i\xi a} \Phi_\mu(\xi) d\xi.$$

Heuristically, this is what we get when we take $|b-a| \to 0$ in the inversion formula. One can imitate the proof of the inversion formula to prove this.

Proposition 14 (Absolute continuity from the characteristic function). *Suppose that Φ_μ is integrable, i.e. $\int_{\mathbb{R}} |\Phi_\mu(\xi)| d\xi < \infty$, then μ has (bounded continuous) density*

$$p(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\xi x} \Phi_\mu(\xi) d\xi.$$

Heuristically, we take $|b-a| \to 0$ and the derivative in the inversion formula to guess the density. One can prove this with the dominated convergence version of Fubini's theorem.

The next proposition tells us that the moments of a random variable can be recovered by the local behavior of its characteristic function near the origin.

Proposition 15 (Moments from the characteristic function). *Let X be a real random variable in L^n, i.e. $\mathbb{E}[|X|^n] < \infty$ for some integer $n \geq 1$. Then*

$$\mathbb{E}[X^n] = (-i)^n \frac{d^n}{(d\xi)^n}\Big|_{\xi=0} \Phi_X(\xi).$$

We use differentiation under the integral sign to calculate the $n-th$ derivative of Φ_X at all points $\xi \in \mathbb{R}$, then specialize at the point $\xi = 0$.

We now record a partial converse of the calculation of moments that will be useful in the proof of the central limit theorem later.

Proposition 16 (Second derivative and second moment). *Let X be a real random variable and Φ_X its characteristic function. If $\mathbb{E}[|X|^2] < \infty$, then for ξ close to 0,*

$$\Phi_X(\xi) = 1 + i\xi\mathbb{E}[X] - \frac{\xi^2}{2}\mathbb{E}[X^2] + o(\xi^2).$$

Conversely, if $\limsup_{\xi \downarrow 0} \frac{\Phi_X(\xi)+\Phi_X(-\xi)-2\Phi_X(0)}{\xi^2} > -\infty$, then $\mathbb{E}[|X|^2] < \infty$.

Proof. The first part is an application of the previous proposition: one checks that by Taylor expansion that the error term is bounded by $t^2\mathbb{E}[|tX^2|] = o(t^2)$. Indeed, as an elementary exercise (see Lemma 3.3.19 of [Durrett]), one checks that

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right).$$

In the special case of $n = 2$, the second term is obtained as

$$e^{ix} - \left(1 + ix + \frac{(ix)^2}{2}\right) = \frac{i^3}{2}\int_0^x (x-s)^2 e^{is} ds = i^2 \int_0^x (x-s)(e^{is} - 1) ds,$$

and since $|e^{is} - 1| \leq 2$, integrating yields an error term of order $|x|^2$.

The second part relies on the positivity of $-\frac{e^{i\xi x}+e^{-i\xi x}-2}{\xi^2} = 2\frac{1-\cos(\xi x)}{\xi^2} \geq 0$, and this quantity converges to x^2 as ξ goes to 0. By Fatou's lemma, $\mathbb{E}[X^2]$ is

$$\begin{aligned} \int_{\mathbb{R}^2} x^2 \mathbb{P}_X(dx) &\leq \liminf_{\xi \downarrow 0} \int_{\mathbb{R}^2} -\frac{e^{i\xi x} + e^{-i\xi x} - 2}{\xi^2} \mathbb{P}_X(dx) \\ &= -\limsup_{\xi \downarrow 0} \frac{\Phi_X(\xi) + \Phi_X(-\xi) - 2\Phi_X(0)}{\xi^2}, \end{aligned}$$

this shows the boundedness in L^2.

Remark 58. The similar statement for $\Phi_X'(0)$ and L^1 is wrong. Counter example: integer-valued random variable X with $\mathbb{P}[X = k] = \mathbb{P}[X = -k] = \frac{1}{Z}\frac{1}{k^2 \ln(k)}$ for all $k \geq 3$ and $\mathbb{P}[|X| \leq 2] = 0$ (and Z is a renormalization constant so that we have a probability measure).

5.3 Lévy's continuity theorem

Let us first make an important observation, direct consequence of the above definitions.

Proposition 17 (Convergence in law implies convergence of characteristic functions). *Suppose that the real random variables $X_1, X_2, \dots$ converges in law to a real random variable X. Then Φ_{X_n} converges to Φ_X.*

Proof. For all $\xi \in \mathbb{R}$, $x \mapsto e^{i\xi x}$ is a continuous bounded function on $\mathbb{R}$. The result follows from the definition of weak convergence for random variables.

The remarkable fact about characteristic functions and the convergence in law is that the converse statement of the above holds.

Theorem 12 (Lévy's continuity theorem: simple version). *Consider real random variables $X_1, X_2, \dots$ and X and their characteristic functions. Then X_n converges in law to X if and only if the sequence of characteristic functions Φ_{X_n} converges pointwise to Φ_X.*

Remark 59. We refer to Theorem 3.3.17 of [Durrett] for a detailed version of Lévy's continuity theorem. We avoid this version since it makes use of the notion of the tightness of measures, which takes time to prepare and is not very useful for this course (but this is a central notion in the study of convergence of measures).

We now give a (somewhat abstract) proof for the above simple version of Lévy's continuity theorem, just for completeness but you can skip it.

Proof. We have to use some preliminaries:

- When dealing with weak convergence of probability measures, we can replace the space of test functions $f \in C_b(\mathbb{R})$ by the space of continuous and compactly supported $f \in C_c(\mathbb{R})$ (see above);
- Fourier transforms is an automorphism on the space of Schwartz functions $\mathcal{S}$, and the inverse of the Fourier transform in this case is given by the classical inversion formula;
- The Schwartz space $\mathcal{S}$ is dense in $C_c(\mathbb{R})$: this can be shown by the classical Stone-Weierstrass approximation theorem.

Now take a Schwartz function $\varphi \in \mathcal{S}$ and consider its Fourier transform $\hat{\varphi}$. Fourier inversion formula works in the Schwartz space, so $\varphi(x) = \int_{\mathbb{R}} e^{i\xi x}\hat{\varphi}(\xi)d\xi$. Apply to $x = X_n$, we get

$$\varphi(X_n) = \int_{\mathbb{R}} e^{i\xi X_n}\hat{\varphi}(\xi)d\xi.$$

Since φ is Schwartz, we can apply Fubini and take the expectation inside the integral:

$$\mathbb{E}[\varphi(X_n)] = \mathbb{E}\left[\int_{\mathbb{R}} e^{i\xi X_n}\hat{\varphi}(\xi)d\xi\right] = \int_{\mathbb{R}} \mathbb{E}\left[e^{i\xi X_n}\right]\hat{\varphi}(\xi)d\xi = \int_{\mathbb{R}} \Phi_{X_n}(\xi)\hat{\varphi}(\xi)d\xi.$$

Similar formula holds for $\mathbb{E}[\varphi(X)]$. Since $\hat{\varphi}$ is also Schwartz, dominated convergence tells us that if $\Phi_{X_n} \to \Phi_X$ then

$$\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$$

as long as φ is Schwartz. To finish the proof, approximate any function $f \in C_c(\mathbb{R})$ by $\varphi \in \mathcal{S}$ and conclude using dominated convergence.

5.4 Central limit theorem

We have already seen that Fourier transform maps convolution to multiplication, or in the language of probability, if X and Y are independent real random variables, then $\Phi_{X+Y} = \Phi_X \cdot \Phi_Y$. To prepare for the central limit theorem, let us record a direct corollary.

Corollary 3 (Sum of independent random variables). *Let $X_1, X_2, \dots$ be independent real random variables. Then $\Phi_{X_1+\cdots+X_n} = \Phi_{X_1} \times \cdots \times \Phi_{X_n}$.*

We have enough prerequisites now to prove the central limit theorem. Recall that the characteristic function of a standard Gaussian variable $\mathcal{N}$ is $\Phi_{\mathcal{N}}(\xi) = e^{-\xi^2/2}$.

Theorem 13 (Central limit theorem). *Let $\{X_n\}_{n\geq 1}$ be a sequence of independent identically distributed random variables. Suppose furthermore that they are centered, i.e. $\mathbb{E}[X_1] = 0$ and in L^2, with $\sigma^2 = var[X_1] = \mathbb{E}[(X_1)^2]$. Then*

$$\frac{1}{\sqrt{n}}\,(X_1 + \cdots + X_n) \overset{(d)}{\to} \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ is the centered Gaussian distribution with variance σ^2.

Proof. Denote by $Z_n = \frac{1}{\sqrt{n}}\,(X_1 + \cdots + X_n)$. By the previous corollary, the characteristic function of Z_n is

$$\Phi_{Z_n}(\xi) = \left(\Phi_{X_1}(\xi/\sqrt{n})\right)^n.$$

Since X_1 has finite second moment, its characteristic function writes

$$\Phi_{X_1}(\xi) = 1 - \frac{\sigma^2}{2}\xi^2 + o(\xi^2)$$

for $\xi \to 0$. Therefore, for any fixed $\xi \in \mathbb{R}$, as $n \to \infty$,

$$\Phi_{Z_n}(\xi) = \left(1 - \frac{\sigma^2}{2n}\xi^2 + o(1/n)\right)^n \to \exp\left(-\frac{\sigma^2\xi^2}{2}\right).$$

The last one is the characteristic function of $\mathcal{N}(0, \sigma^2)$, and we conclude by Lévy's continuity theorem.

Remark 60. In most textbooks, the condition that X_n are centered is not included in the theorem. This is only because I have decided to put the central limit theorem before the easier-to-prove laws of large numbers.

5.5 Applications and various extensions

We have limited time and we only covered the basics that lead to the central limit theorem. Below is a list of topics for further studies (with my personal taste).

- Stirling's formula from central limit theorem.
- Concentration of measures.
- Helly's selection theorem and tightness of a sequence of measures.
- Bochner's theorem, Khinchine's theorem and Pólya's criteria.
- The moment problem.
- Different versions and variants of the central limit theorem.
- Infinitely divisible laws and infinitely divisible characteristic functions.

Chapter 6
Convergence in probability

The convergence in probability expresses the idea that a sequence of random variables $\{X_n\}_{n\geq 1}$ goes "very close" to some other random variable X with "high probability", without necessarily having a pointwise convergence.

6.1 Convergence in probability

Let us start with a usual definition for the convergence in probability. Notice that we require all random variables to be defined on the same probability space.

Definition 33 (Convergence in probability). Consider real random variables $\{X_n\}_{n\geq 1}$ and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say $\{X_n\}_{n\geq 1}$ **converges in probability** to X if for all $\epsilon > 0$,

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

We usually denote this by $X_n \overset{(\mathrm{P})}{\to} X$.

Remark 61. In some sense, X_n and X can still be very different on a small portion E_n of Ω, and we only require that the size of E_n goes to 0 while the sequence $\{E_n\}_{n\geq 1}$ can fluctuate and move around in Ω.

We will always assume that, when speaking of convergence in probability, that all the random variables are defined on the same probability space. In this case, one should care about the uniqueness of the limit in probability.

Proposition 18 (Uniqueness of the limit in probability). *Suppose that $\{X_n\}_{n\geq 1}$ converges in probability. Then the limit is unique $\mathbb{P}$-almost everywhere.*

Proof. Suppose we have two limits X and Y. Then the set on which they differ at least by ϵ has $\mathbb{P}$-measure at most

$$\mathbb{P}[|X-Y|>\epsilon] \leq \mathbb{P}[|X-X_n|>\epsilon/2] + \mathbb{P}[|Y-X_n|>\epsilon/2]$$

for all n. In other words, if $|X-Y|>\epsilon$ then either $|X-X_n|>\epsilon/2$ or $|Y-X_n|>\epsilon/2$, and we apply the union bound. Now the right hand side above converges to 0 as ϵ goes to 0, so $\mathbb{P}[X \neq Y] = 0$. That is, the limit in probability is $\mathbb{P}$-almost everywhere unique.

Remark 62. It is customary in probability theory to identify objects that are equal almost everywhere. That is, for random variables, we also work under the equivalence class induced by the relation $X \sim Y$ if and only if $\mathbb{P}$-almost everywhere, we have $X = Y$.

Remark 63. We cannot ask the same question for the convergence in law, since random variables that are different almost everywhere may induce the same law.

The convergence in probability can be realized as a complete metric space on the space of real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.

Proposition 19 (Complete metric space structure for the convergence in probability). *The convergence in probabability can be characterized by the complete metric $d(X,Y) = \mathbb{E}[\min(|X-Y|,1)]$. Otherwise said, the space of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is a Banach space under the distance d, and convergence in probability is equivalent to convergence in the distance d.*

Proof. It is routine to verify that d is a distance. The fact that d characterizes the convergence in probability follows from that for all $\epsilon < 1$,

$$d(X,Y) \leq \mathbb{E}[\min(|X-Y|,1)(\mathbf{1}_{|X-Y|\leq\epsilon} + \mathbf{1}_{|X-Y|>\epsilon})] \leq \epsilon + \mathbb{P}[|X-Y|>\epsilon],$$
$$\mathbb{P}[|X-Y|>\epsilon] \leq \epsilon^{-1}\mathbb{E}[\min(|X-Y|,1)] = \epsilon^{-1} d(X,Y).$$

The proof of completeness is postponed to next week when the notion of almost sure convergence would be defined.

We also mention the continuous mapping theorem.

Proposition 20 (Continuous mapping theorem). *Suppose that a sequence of real random variables $\{X_n\}_{n\geq 0}$ converges in probability towards X. Then if g is a continuous real function, the sequence $\{g(X_n)\}_{n\geq 0}$ converges in probability towards $g(X)$.*

Proof. Exercise.

Observe that the convergence in probability is stronger than the convergence in law:

Proposition 21 (Convergence in probability implies convergence in law). *Suppose that a sequence of real random variables $\{X_n\}_{n\geq 0}$ converges in probability to a random variable X. Than $\{X_n\}_{n\geq 0}$ converges in law to X.*

Proof. Let $f \in C_c(\mathbb{R})$ be a continuous, compactly supported function, and show that $\mathbb{E}[f(X_n)]$ converges to $\mathbb{E}[f(X)]$. For all $\epsilon > 0$,

$$\mathbb{E}[|f(X_n)-f(X)|] \leq \mathbb{E}[|f(X_n)-f(X)|\mathbf{1}_{|X_n-X|>\epsilon}]+\mathbb{E}[|f(X_n)-f(X)|\mathbf{1}_{|X_n-X|\leq\epsilon}].$$

Since $\mathbb{P}[|X_n - X| > \epsilon] \to 0$ and $|f(X_n) - f(X)| \leq 2||f||_\infty$, the first term goes to 0 as n goes to 0. The second term also vanishes in the $n \to \infty$ limit by the uniform continuity of f. This shows the convergence in law since $|\mathbb{E}[f(X_n) - f(X)]| \leq \mathbb{E}[|f(X_n) - f(X)|] \to 0$ as $n \to \infty$.

A partial converse is true when the limit X is (almost surely) a constant.

Proposition 22 (Convergence in law to a constant implies convergence in probability). *If a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in law to a constant c, than it converges in probability to c.*

Proof. Consider the continuous bounded function $f(x) = \min(|x - c|, 1)$. By convergence in law, $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(c)]$. But this is $\mathbb{E}[\min(|X_n - c|, 1)] \to 0$, which is an equivalent definition for the convergence in probability.

Remark 64. Actually, if X_n and X are independent for every n, the only possible case where the convergence in probability can happen is when X is a constant.

A further result in this direction is known as Slutsky's theorem. The proof of Slutsky's theorem is omitted in this note.

Theorem 14 (Slutsky's theorem). *Suppose that $\{X_n\}_{n\geq 1}$, $\{Y_n\}_{n\geq 1}$ are sequences of real random variables, $\{X_n\}_{n\geq 1}$ converging in law to X and $\{Y_n\}_{n\geq 1}$ converging in law to a constant $c \in \mathbb{R}$.*

1. *The sequence $\{Y_n\}_{n\geq 1}$ converges in fact in probability to c.*
2. *The sequence of vectors $\{(X_n, Y_n)\}_{n\geq 1}$ converges in law to (X, c).*

6.2 Convergence in L^p

There is a practical sufficient condition for establishing convergence in probability.

Definition 34 (Convergence in L^p). Let $1 \leq p < \infty$. A sequence of real random variables $\{X_n\}_{n\geq 0}$ converges in L^p towards some real random variable X if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

The finiteness of the p-th moment yields some information on the tail of the difference $|X_n - X|$ by Markov's inequality.

Proposition 23 (Convergence in L^p implies convergence in probability). *Let $1 \leq p < \infty$ and suppose that $\{X_n\}_{n\geq 0}$ converges in L^p towards some real random variable X. Then $\{X_n\}_{n\geq 0}$ converges in probability to X.*

Proof. For fixed $\epsilon > 0$, we have, as $n \to \infty$,

$$\mathbb{P}[|X_n - X| > \epsilon] \leq \frac{\mathbb{E}[|X_n - X|^p]}{\epsilon^p} \to 0.$$

where we used Markov's inequality and then the convergence in L^p. This shows the convergence in probability.

Notice that the above proof actually works for $0 < p < \infty$. Notice also that convergence in L^p implies convergence of the L^p-norms of the random variables.

Proposition 24 (Convergence of L^p-norms). *et $1 \leq p < \infty$ and suppose that $\{X_n\}_{n\geq 0}$ converges in L^p towards some real random variable X. If furthermore, all random variables are in L^p, then $||X_n||_{L^p}$ converges to $||X||_{L^p}$ as well.*

Proof. This is a consequence of the Minkowski's inequality.

Remark 65. Sometimes, convergence in L^1 is called *convergence in mean*.

6.3 Weak law of large numbers

The law of large numbers deals with the average of i.i.d. random variables. As a warm-up, we show a simple weak law of large numbers.

Theorem 15 (L^2 weak law of large numbers). *Let $\{X_j\}_{j\geq 1}$ be a sequence of independent identically distributed random variables on the same probability space. Suppose that they are in L^1, i.e. the average $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ exists. Suppose furthermore that $\mathbb{E}[(X_1 - \mu)^2] = \sigma^2 < \infty$. Then the average $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ converges in probability to the constant μ.*

Proof. First, we can suppose that $\mu = 0$ by replacing X_j by $X_j - \mu$: this will simplify the following proof and we shall show that S_n converges in probability to the constant 0. It is also useful to observe that, by independence, $\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)$.

It follows that $\mathbb{E}\left[(S_n/n)^2\right] = \sigma^2/n$, so that $\{S_n/n\}_{n\geq 1}$ converges in L^2 towards 0. This implies that $\{S_n/n\}_{n\geq 1}$ converges in probability to 0.

Remark 66. Notice that we actually only used the non-decorrelation, i.e. $\mathrm{cov}(X_j, X_k) = 0$, rather then the independence property.

Remark 67. You can also prove this result by using the central limit theorem and use (the simple) Slutsky to upgrade the convergence in law to a convergence in probability.

We now prepare ourselves to get rid of the L^2 assumption above. I personally call this the "L^1-L^2 trick", as it is a general strategy that finds itself useful in many applications. The guiding idea is that a L^2 estimate is worse than a L^1 estimate on the tail of a random variable, so we should treat the small values with a L^2 method and the large values with a L^1 method.

Theorem 16 (Weak law of large numbers). *Let $\{X_j\}_{j\geq 1}$ be a sequence of independent identically distributed random variables on the same probability space. Suppose that they are in L^1, i.e. the average $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ exists. Then the average $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ converges in probability to the constant μ.*

Proof. Suppose that $\mu = 0$. For $L > 0$, write $X_j = X_j \mathbf{1}_{\{|X_j|>L\}} + X_j \mathbf{1}_{\{|X_j|\leq L\}}$ and

$$\frac{S_n}{n} = \frac{\sum_{j=1}^{n} X_j \mathbf{1}_{\{|X_j|>L\}}}{n} + \frac{\sum_{j=1}^{n} X_j \mathbf{1}_{\{|X_j|\leq L\}}}{n} =: \frac{S_n^{>}}{n} + \frac{S_n^{\leq}}{n}.$$

By dominated convergence, $\mathbb{E}[|X_j|\mathbf{1}_{\{|X_j|>L\}}] \to 0$ as $L \to \infty$ (see the remark after Markov's inequality). Therefore, linearity of expectation yields that as $L \to \infty$,

$$\mathbb{E}\left[\left|\frac{S_n^{>}}{n}\right|\right] = \frac{\sum_{j=1}^{n} \mathbb{E}[|X_j|\mathbf{1}_{\{|X_j|>L\}}]}{n} \to 0.$$

Therefore, given $\epsilon > 0$ and $\delta > 0$, Markov's inequality yields that for $L = L(\epsilon)$ large enough,

$$\mathbb{P}\left[\left|\frac{S_n^{>}}{n}\right| > \frac{\epsilon}{2}\right] \leq \frac{2}{\epsilon}\mathbb{E}[|X_j|\mathbf{1}_{\{|X_j|>L\}}] \leq \frac{\delta}{2}.$$

To show the weak law of large numbers, it remains to see (via the union bound) that, for large enough n,

$$\mathbb{P}\left[\left|\frac{S_n^{\leq}}{n}\right| > \frac{\epsilon}{2}\right] \leq \frac{\delta}{2},$$

but this is true since for fixed $L > 0$, the truncated random variable $X_j \mathbf{1}_{\{|X_j|\leq L\}}$ is in L^2, and the above equation follows from the L^2 weak law of large numbers.

Chapter 7
Almost sure convergence

We finish this course with the strongest form of convergence of random variables: the almost sure convergence. With the large of large numbers, we can finally justify mathematically our intuition that the probabability of an event corresponds to its frequency of apparition when the same experiment is repeated a large number of times.

7.1 Almost sure convergence

The almost sure convergence is the pointwise convergence in probability.

Definition 35 (Almost sure convergence). Consider real random variables $\{X_n\}_{n\geq 1}$ and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $\{X_n\}_{n\geq 1}$ converges **almost surely** to X if for $\mathbb{P}$-almost all $\omega \in \Omega$, $X_n(\omega) \to X(\omega)$ as $n \to \infty$. Otherwise said,

$$\mathbb{P}[\omega \in \Omega \,;\, X(\omega) = \lim_{n\to\infty} X_n(\omega)] = 1.$$

We usually denote this by $X_n \overset{\text{a.s.}}{\to} X$, where **a.s.** stands for almost sure(ly).

The almost sure convergence is the strongest mode of convergence.

Proposition 25 (Almost sure convergence implies convergence in probability). *Consider real random variables $X_1, \ldots, X_n$ and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that $\{X_n\}_{n\geq 1}$ converges almost surely to X. Then $\{X_n\}_{n\geq 1}$ converges in probability to X.*

Proof. Suppose that $\{X_n\}_{n\geq 1}$ converges almost surely to X and show that as $n \to \infty$, we have $\mathbb{E}[\min(|X_n - X|, 1)] \to 0$ (recall the complete metric space structure for the convergence in probability). But this follows immediately from dominated convergence.

We have the partial converse of the above:

Proposition 26 (Convergence in probability implies subsequential almost sure convergence). *Consider real random variables $X_1, \dots, X_n$ and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that $\{X_n\}_{n\geq 1}$ converges in probability to X. Then $\{X_n\}_{n\geq 1}$ has a subsequence that converges almost surely to X.*

Proof. This is a general fact of Banach space. The proof is actually contained in the proof of the completeness of the metric of convergence in probability: let us prove the latter now.

To see the completeness of the metric $d(X,Y) = \mathbb{E}[\min(|X-Y|, 1)]$, we should prove that all Cauchy sequences $\{X_n\}_{n\geq 1}$ under d converge to some real random variable X in the limit. As it is customary, choose a subsequence $\{Y_n\}_{n\geq 1}$ of the Cauchy sequence $\{X_n\}_{n\geq 1}$ such that $d(Y_n, Y_{n+1}) \leq 2^{-n}$. It follows that

$$\sum_{n\geq 1} d(Y_n, Y_{n+1}) = \sum_{n\geq 1} \mathbb{E}[\min(|Y_n - Y_{n+1}|, 1)] = \mathbb{E}[\sum_{n\geq 1} \min(|Y_n - Y_{n+1}|, 1)] < \infty,$$

so $\sum_{n\geq 1} \min(|Y_n - Y_{n+1}|, 1) < \infty$ almost surely.

Actually, this implies that $\sum_{n\geq 1} |Y_n - Y_{n+1}| < \infty$ almost surely, since there can be only finitely many terms for which $\min(|Y_n - Y_{n+1}|, 1) = 1$ when the sum converges. Then $Y_1 + \sum_{n\geq 1}(Y_{n+1} - Y_n)$ is almost surely absolutely convergent, therefore convergent, and by denoting its limit as X, we have $Y_n \to X$ almost surely.

Now by dominated convergence, $\lim_{n\to\infty} \mathbb{E}[\min(|Y_n - X|, 1)] = 0$, so Y_n converges in probability to X. It is not hard to see that by construction, X_n converges in probability to X as well.

The relation between almost sure convergence and convergence in law is more abstract.

Proposition 27 (Almost sure convergence implies convergence in law). *Consider real random variables $\{X_n\}_{n\geq 1}$ and X defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and suppose that $\{X_n\}_{n\geq 1}$ converges almost surely to X. Then $\{X_n\}_{n\geq 1}$ converges in law to X.*

Proof. We have seen already that almost sure convergence implies convergence in probability, which in turn implies convergence in law.

The partial converse statement is given by Skorokhod's representation theorem. The latter says that convergence in law can be realized by almost sure convergence, but in possibly a different probability space (recall that the convergence in law do not require that all random variables are defined on the same probability space, it is a definition about the laws of the random variables, i.e. their induced probability measures).

Theorem 17 (Skorokhod's representation theorem). *Suppose $\{Y_n\}_{n\geq 1}$ is a sequence of real random variables converging to some real random variable Y; these random variables are not necessarily defined on the same probability space. Then*

there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *with real random variables* $\{X_n\}_{n\geq 1}$ *and* X *on this probability space such that:*

1. *For each* $n \geq 1$, X_n *and* Y_n *are equal in law;* X *and* Y *are also equal in law.*
2. *The sequence* $\{X_n\}_{n\geq 1}$ *converges almost surely to* X.

Proof. See Theorem 3.2.8 of [Durrett].

Remark 68. The Skorokhod representation theorem is a practical device to simplify many proofs, including some general facts about the convergence in law.

7.2 Random variable in the tail σ-algebra

We prepare ourselves for the law of large numbers. Recall that Kolmogorov's zero-one law says that if $\{X_n\}_{n\geq 1}$ are independent random variables and $\mathcal{T}$ the associated tail σ-algebra, then $\mathcal{T}$ is independent of itself. One consequence is the following:

Proposition 28 (Random variable in the tail σ-algebra). *Suppose that* $\{X_n\}_{n\geq 1}$ *are independent random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* $\mathcal{T}$ *the associated tail* σ*-algebra. Then if some random variable* X *is* $\mathcal{T}$*-measurable,* X *is almost surely constant in* $[-\infty, \infty]$.

Proof. If X is $\mathcal{T}$-measurable, then $A = \{X \leq a\} \in \mathcal{T}$ is a tail event, so $\mathbb{P}[X \leq a] \in \{0, 1\}$. This implies that the cumulative function F_X can only take two values: 0 or 1. But $F_X(a)$ is increasing in a, so if $t = \inf\{a \in [-\infty, \infty] \ ; \ F_X(a) = 1\}$ the moment where $F_X(a)$ jumps from 0 to 1, then $X = t$ almost surely.

Corollary 4. *Suppose that* $\{X_j\}_{j\geq 1}$ *are independent random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$ *and* $S_n = X_1 + \cdots + X_n$ *for* $n \geq 1$. *Then* $\limsup_{n\to\infty} \frac{S_n}{n}$ *is almost surely constant.*

Proof. This is because $\limsup_{n\to\infty} \frac{S_n}{n}$ is a tail random variable.

Therefore, if we know that the limit $\lim_{n\to\infty} \frac{S_n}{n}$ exists, then it is almost surely a constant in $[-\infty, \infty]$. The strong law of large numbers establishes the convergence and identifies its value in the case of i.i.d. random variables.

We cannot resist mentioning another application on simple random walks.

Proposition 29 (Recurrence of $1d$ simple random walk). *Suppose that* $\{X_n\}_{n\geq 1}$ *are independent random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$, *each distributed as a fair coin toss in* $\{-1, 1\}$. *Let* $S_n = X_1 + \cdots + X_n$ *for* $n \geq 1$. *Then almost surely,*

$$\limsup_{n\to\infty} S_n = \infty \quad \textit{and} \quad \liminf_{n\to\infty} S_n = -\infty.$$

In particular, $S_n = 0$ *for infinitely many* $n \geq 1$, *i.e.* S_n *is recurrent (it returns infinitely many times to its initial state).*

Proof. We have already seens that $\mathbb{P}[\limsup_{n\to\infty} S_n = \infty] \in \{0,1\}$. A similar argument shows that $\mathbb{P}[\liminf_{n\to\infty} S_n = -\infty] \in \{0,1\}$. But by symmetry, these two probabilities are equal. The result follows if we can show that $\{S_n\}_{n\geq 1}$ is almost surely not bounded, since then by union bound,

$$\mathbb{P}[\limsup_{n\to\infty} S_n = \infty] + \mathbb{P}[\liminf_{n\to\infty} S_n = -\infty] \geq \mathbb{P}[\{S_n\}_{n\geq 1} \text{ is unbounded}] = 1,$$

so the only possibility is that both of these probabilities are 1.

To show that $\{S_n\}_{n\geq 1}$ is almost surely unbounded, we just need a crude estimate that $\{S_n\}_{n\geq 1}$ is almost surely not included in any interval $[-p,p]$ for any $p \geq 0$: the unboundedness follows then from the union bound

$$\begin{aligned}\mathbb{P}[\{S_n\}_{n\geq 1} \text{ is bounded}] &\leq \mathbb{P}[\exists p \geq 1, \forall n \geq 1, S_n \in [-p,p]] \\ &\leq \sum_{p\geq 1} \mathbb{P}[\forall n \geq 1, S_n \in [-p,p]].\end{aligned}$$

To show that $\{S_n\}_{n\geq 1}$ is almost surely not bounded in $[-p,p]$, it suffices to show that almost surely, $\{S_n\}_{n\geq 1}$ has a consecutive sequence of 1:s of length $2p+2$. This is an easy application of the (second) Borel-Cantelli's lemma and finishes the proof.

7.3 Law of large numbers

The strong law of large numbers upgrades the mode of convergence of the weak law of large numbers: we are interested in results about almost sure convergence in lieu of convergence in probability. Let us start with a simple version.

Theorem 18 (L^4 strong law of large numbers). *Let $\{X_j\}_{j\geq 1}$ be a sequence of independent identically distributed random variables on the same probability space. Suppose that they are in L^1, i.e. the average $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ exists. Suppose furthermore that $\mathbb{E}[X_1^4] < \infty$. Then the average $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ converges almost surely to the constant μ.*

Proof. Again we suppose $\mu = 0$. The idea is to show that $\mathbb{E}[(S_n/n)^4]$ decays fast enough. To see this, develop the factor $(X_1+\cdots+X_n)^4$ and observe that

$$\mathbb{E}[(X_1+\cdots+X_n)^4] \leq n\mathbb{E}[X_1^4] + 3n(n-1)\left(\mathbb{E}[X_1^2]\right)^2 \leq Cn^2.$$

It follows that $\mathbb{E}[(S_n/n)^4] \leq Cn^{-2}$, which is summable. By Fubini-Tonelli, we have

$$\mathbb{E}\left[\sum_{n\geq 1}(S_n/n)^4\right] < \infty.$$

This implies that $\sum_{n\geq 1}(S_n/n)^4 < \infty$ almost surely, which in turn implies that S_n/n converges to 0 almost surely.

As an importance (and largely popularized) consequence:

Corollary 5 (Frequency of apparition). *If $\{A_n\}_{n\geq 0}$ is a sequence of independent events in $(\Omega, \mathcal{F}, \mathbb{P})$ with same probability, then the following convergence holds almost surely:*

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{A_j} \to \mathbb{P}[A_1].$$

Proof. The indicator function $\mathbf{1}_{A_1}$ is bounded in L^4: therefore the L^4 strong law of large numbers applies.

We can loosen the L^4 restriction above. There are many proofs of this ultraclassical result, we select one that is adapted to our knowledge about probability theory.

Theorem 19 (Strong law of large numbers). *Let $\{X_j\}_{j\geq 1}$ be a sequence of independent identically distributed random variables on the same probability space. Suppose that they are in L^1, i.e. the average $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ exists. Then the average $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ converges almost surely to the constant μ.*

Proof. See Theorem 2.4.1 (the proof finishes at Theorem 2.4.5) of [Durrett]. The proof that I prefer uses the theory of martingales (which you will learn in Probability Theory II).

Remark 69. If X_1 is positive and $\mathbb{E}[X_1] = \infty$, applying the theorem to $\min(X_1, k)$ for positive integer k shows that S_n converges almost surely to ∞.

7.4 Some classical applications

Some other topics related to this chapter:

- Law of the iterated logarithms;
- Cramér's theorem on large deviations;
- Kolmogorov's three series theorem;
- Lévy's theorem on the equivalence of different convergences in the case of the sum of i.i.d. random variables.

Exercise set III

Exercises marked with ! are important and those with ⋆ are difficult.

Exercise 15 (! – Convergence in law of Gaussian variables). Suppose that a sequence of real Gaussian variables $X_1, X_2, \dots$ converges in law to some random variable X. Show that X must be a Gaussian variable as well.

Exercise 16 (! – Sublinearity of simple random walk). We call ϵ a Rademacher random variable if it takes value in $\{-1, 1\}$ with equal probability. Consider an i.i.d. sequence $\{\epsilon_n\}_{n\geq 1}$ of Rademacher random variables and show that, for all $t > 0$,

$$\mathbb{P}\left[\sum_{j=1}^{n} \epsilon_j \geq t\right] \leq \exp\left(-\frac{t^2}{2n}\right).$$

Deduce that a simple random walk $S_n = \sum_{j=1}^{n} \epsilon_j$ is sub-linear, i.e. for all $a > 0$, almost surely there exists some $n_0 \geq 1$ such that $S_n \leq an$ for all $n \geq n_0$.

Exercise 17 (Minimum of independent uniform distributions). Let $U_1, \dots, U_n, \dots$ be a sequence of i.i.d. uniform distributions on $[0, 1]$. Show that

$$\lim_{n\to\infty} n \min\{U_1, \dots, U_n\}$$

converges in law to $X \sim \mathrm{Exp}(1)$. Could you have guessed this result using a characteristic property of the exponential distribution?

Exercise 18 (Convergence in probability does not imply almost sure convergence). Consider $X_j = \mathcal{B}(0, 1/j)$ a sequence of independent biaised coin tosses in

$\{0, 1\}$ for $j \geq 1$, each X_j taking the value 1 with probability $1/j$. Show that X_j converges in probability to 0, but X_j converges almost surely to 1.

Exercise 19 (Convergence in probability and topology). Show that a sequence of real random variables $X_1, X_2, \ldots$ converges in probability towards a random variable X if and only if

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n - X|}{1 + |X_n - X|}\right] = 0.$$

Show that the space of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with the distance

$$d(X, Y) = \mathbb{E}\left[\frac{|X - Y|}{1 + |X - Y|}\right]$$

is a Banach space.

Exercise 20 (! – Slutsky's theorem). Suppose that X_n, Y_n are sequences of real random variables, X_n converging in law to X and Y_n converging in law to a constant $c \in \mathbb{R}$.

1. Show that Y_n converges in fact in probability to c.
2. Show that the vector (X_n, Y_n) converges in law to (X, c).
3. Give an example where if we suppose that Y_n converges in law to a general random variable Y, the result above about the joint convergence is wrong.

Exercise 21 (Almost complete convergence). Given random variables $\{X_n\}_{n\geq 1}$ and X on the same probability space, we say that $\{X_n\}_{n\geq 1}$ converges almost completely (n.b.: never used this terminology myself) to X if for all $\epsilon > 0$,

$$\sum_{n\geq 1} \mathbb{P}[|X_n - X| > \epsilon] < \infty.$$

Show that $\{X_n\}_{n\geq 1}$ converges almost surely to X. Deduce that every sequence of random variables converging in probability has a *subsequence* converging almost surely.

Exercise 22 (Centered random walks). Let $\{X_n\}_{n\geq 1}$ be a sequence of i.i.d. real random variables defined on the same probability space and $S_n = X_1 + \cdots + X_n$. Suppose that $\mathbb{E}[X_1] = 0$ and $0 < \mathbb{E}[(X_1)^2] < \infty$. An example is given by taking $\{X_n\}_{n\geq 1}$ to be an independent sequence of fair toin cosses.

1. Let $\{X_{j_k}\}_{k\geq 1}$ denote a subsequence of X_n. Use Kolmogorov's zero-one law to show that

$$\mathbb{P}\left[\limsup_{k\to\infty} \frac{S_{j_k}}{\sqrt{j_k}} = \infty\right] = 1; \quad \mathbb{P}\left[\liminf_{k\to\infty} \frac{S_{j_k}}{\sqrt{j_k}} = -\infty\right] = 1.$$

Hint: use the central limit theorem to give a lower bound for each probability above.

2. Deduce that the sequence $\{X_n\}_{n\geq 1}$ does not converge in probability.

Exercise 23 ($\star$ – Longest head-runs). Consider n independent fair coin tosses $\{X_j\}_{1\leq j\leq n}$ with value in $\{-1, 1\}$. Consider L_n, the length of the longest consecutive appearances of 1:s. More precisely, let $l_k = \max\{m\ ;\ X_{k-m+1} = \cdots = X_k = 1\}$ be the longest run of 1:s at time k, and $L_n = \max_{1\leq k\leq n} l_k$ the global longest run of 1:s.

We will prove that $\frac{L_n}{\ln_2 n} \to 1$ almost surely as $n \to \infty$.

1. Show that for all $\epsilon > 0$, $\mathbb{P}[l_k \geq (1+\epsilon)\ln_2 k] \leq n^{-(1+\epsilon)}$. Use the summability of $n^{-(1+\epsilon)}$ and Borel-Cantelli's lemma to conclude that $\limsup_n \frac{L_n}{\ln_2 n} \leq 1$.
2. For the other direction, break n tossings into disjoint blocks of length approximately $(1-\epsilon)\ln_2 n$ each. For one block of this length, calculate approximately the probability to have only 1:s inside the box. Estimate the number of blocks and then the probability that, at time n, non of these boxes has only 1:s inside. Relate the last calculation with $\mathbb{P}[L_n \leq (1-\epsilon)\ln_2 n]$ and use Borel-Cantelli to conclude (and write a proper version without the "approximately").

[Schilling – The Longest Run of Heads; Durrett – Probability: Theory and Examples (v5), Example 2.3.12; Williams – Probability with martingales, Exercise E.4.4]

Key results of this course

We recollect some important elements of this course.

Part I: Foundations

- Random variable: definition and associated σ-algebra.
- Expectation: different results from integration theory, including convergence results and inequalities.
- Probability inequalities: Markov inequality and its variants.
- Law of random variable: cumulative distribution function, density function, characteristic function and calculations.
- Tricks and tips: be aware of expressions of type $\mathbb{E}[f(X)]$, $\mathbb{P}[X > a]$, and careful about interchanging limits.

Part II: Independence

- Independence: factorization property.
- Product space: Fubini theorem, joint law of (independent) random variables.
- Calculations: detecting independence, sum of independent random variables, convolution and use of characteristic function.
- Sequence of random variables: Borel-Cantelli's lemmas and Kolmogorov zero-one.
- Tricks and tips: be aware of tail events, and check conditions before using Fubini and/or independence property.

Part III: Convergences

- Convergences: definition and relations between different modes of convergence.
- Characteristic function: on the law of a random variable and the convergence in law of a sequence of random variables.
- Applications: at least be familiar with one classical application for each mode of convergence.
- Tricks and tips: remember to try characteristic for convergence in law, moments for convergence in probability and Borel-Cantelli for almost sure convergence.

Mock exam: Probability Theory I

Rules:

- Grading: each exercise has two (2) questions. Answering one (1) of them correctly gains you eight (8) point, and answering two (2) of them correctly gains you twelve (12) points. The maximum point is thirty (30).
- Items: lecture notes and textbooks are allowed. Calculators, programs, phones and webpages (e.g. Wikipedia or forum) are not allowed. Do not share the exam nor your solutions. Water and snack are recommended.

Exercise 24 (Uniform distribution). Let $X \sim \mathcal{U}([0,1])$, that is, X has density function $p(x) = \mathbf{1}_{\{0 \le x \le 1\}}$.

1. Calculate the mean and the variance (alternatively, the second moment) of X.
2. Calculate the law of $Y_n = X^n$ for $n \ge 1$.

Exercise 25 (Paley-Zygmund inequality). Let $Z \ge 0$ be a random variable with finite variance. Let $0 \le \theta \le 1$.

1. Show that $\mathbb{E}[Z\mathbf{1}_{Z > \theta\mathbb{E}[Z]}] \le \mathbb{E}[Z^2]^{1/2}\mathbb{P}[Z > \theta\mathbb{E}[Z]]^{1/2}$.
2. Show that $\mathbb{P}[Z > \theta\mathbb{E}[Z]] \ge (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$.

The last inequality is known as the Paley-Zygmund inequality. While Markov's inequality gives an upper bound on the tail of Z, Paley-Zygmund inequality gives a lower bound and is often used as a partial converse of Markov's inequality.

Exercise 26 (Maximum of independent standard Gaussian variables). Let $\{X_n\}_{n\geq 1}$ be independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, each one distributed as a standard normal Gaussian $\mathcal{N}(0,1)$. You will need the following fact: if $X \sim \mathcal{N}(0,1)$, then we have the following Gaussian tail bound: for all $t > 0$,[1]

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2} \leq \mathbb{P}[X > t] \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}.$$

1. Calculate $b_n = \mathbb{P}[X_n > (2\log n)^{1/2}]$ and show that $\sum_{n\geq 1} b_n = \infty$. Deduce that, almost surely,
$$\limsup_{n\to\infty} \frac{X_n}{(2\log n)^{1/2}} \geq 1.$$
2. Show that $\limsup_{n\to\infty} \frac{X_n}{(2\log n)^{1/2}} = 1$.

[1] In practise (and in this exercise), it is enough to retain that $\mathbb{P}[X > t]$ behaves like $\frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}$ at a first approximation.