An introduction to probability theory

Cristina Toninelli

Contents

This is a preliminary version of the notes of the course Introduction to probability for M1 ICFP.
Misprints are very likely to be present: do not hesitate to send me any feedback/corrections (toninelli@ceremade.dauphine.fr).

Chapter 1

A brief introduction and some historical facts

This course is a first introduction to probability theory, we will try to set the basis of the theory with some rigour but skipping most of technical details. Probability is the branch of mathematics that deals with the study of random phenomena, it is one of the most active fields in applied mathematics, and plays a key role in diverse fields including physics , biology, social sciences. The world ”random” does not refer to chaotic (in the sense of ”completely unpredictable”) but rather refers to phenomena that, though not certain, they are to some extent predictable.
Let's start recalling some key dates for the history of probability

- an event often indicated as the beginning of probability is the correspondence between Pascal (1623-1662) and Fermat (1601-1665) concerning games, gambling in particular. There is in particular a famous exchange concerning the so called match problem (*problème de partis*). Two gamblers bet each 10 euros and start playing a series of game each ending with a winner and a looser. The rule is that the player that wins first 3 games will gain the 20 euros. The game is interrupted unfortunately before any of them wins...question: how should one divide fairly the 20 euros based on the number of wins of each player?

- 1657: first book on mathematical modelisation of dice game dates back and is due to Huygens

- end of 17th century: probability theory started to become important outside games, and it was used for an important political/social issue: evaluate mortality and population size. At the beginning of 18th century population theory was developed by Bernoulli, who is considered the ancestor of modern demography

- first half of 18th century: the problem of the estimating errors in astronomy measurement stimulated further the development of probability

- the 18th century saw three key steps in probability theory: law of large numbers (by Bernoulli) and (a first version of) the central limit theorem (by de Moivre), the development of the first studies of statistical inference i.e. the art determining a posteriori probability laws from observation (Bayes and de Laplace)
- 19th century: Gauss error theory, development of the Russian school (Chebyshev, Markov, Lyapunov)
- 20th century: birth of modern probability theory with Kolmogorov and Borel axiomatisation of probability theory in the framework of measure theory. Stochastic process theory flourished and stochastic calculus was born, introduction of differential equations with noise (Levy, Frechet, Doob,Ito,. . .)

Chapter 2

Setting the framework: probability spaces

Probability theory aims at representing in an axiomatic way some concepts which are implicit in common sense and it is constructed in a very empirical way going back and forth from observation to theory. That's why the concepts we will introduce in this chapter will sound familiar to you even if you have never studied probability before!

1 Random experiments and sample spaces

A **random experiment** is an experiment whose result can not exactly be determined in advance, but such that we know in advance the **set of all possible outcomes**. We call Ω the set of possible outcomes, also called the **sample space**

Example 1 (Experiment 1)**.** *Toss once a coin. The set of possible outcomes contains two elements head and tail,* $\Omega = \{head, tail\}$

Example 2 (Experiment 2)**.** *Take n cards each having a different number from 1 to n, mix the deck. The possible outcomes are all the different permutations of the* n *cards.*

$$\Omega := \{\omega = (\omega(1), \dots \omega(n)) : \omega \textit{ is a permutations} of (1, \dots, n))\}$$

2 Events and σ-algebras

Given a random experiment we can define its **events**. Each event is a **sub-set of** Ω, so that the set of events is a subset of the set of all subsets, also called power set, or set of parts *famille de parties*, which we denote by $\mathcal{P}(\Omega)$.

Given two events, A and B, we let

- $A \cup B$ be the event *A or B*
- $A \cap B$ be the event *A and B*
- A^c be the event *not A*

Note that it holds

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

We say that A and B are *disjoint* if the event *A and B* is impossible, namely $A \cap B = \emptyset$.

Example 3 (Example of events for experiment 2)**.**

1. E_1:= *card number 1 is in the second half of the deck*
2. E_2:= *card 2 is not in the deck*
3. E_3:= *card 1 is at the first position*
4. E_4:= *each card is not at its position*

In formulas

1. $E_1 := \{\omega \in \Omega; \omega(1) \geqslant n/2\}$
2. $E_2 := \emptyset$
3. $E_3 := \{\omega \in \Omega : \omega(1) = 1\}$
4. $E_4 := \{\omega \in \Omega : \omega(i) \neq i \ \forall i\} = \cap_{i=1}^n B_i^c$ *where* $B_i := \{\omega \in \Omega : \omega(i) = i\}$

When confusion does not arise we omit the $\omega \in \Omega$, e.g. $B_i := \{\omega(i) = i\}$.

In order to define a probability on the random experiments there are some minimal requirement on the set of events we want to consider (which may or may not include each single outcome of the experiments, i.e. each single element of Ω). This minimal set of requirement sis formalised by saying that the set of events should be a σ**-algebra** (or a σ-field or, in french, a *tribu*), which is defined as follows

Definition 2.1 (σ-algebra)**.** *A set $\mathcal{F}$ of subsets of Ω (i.e. a set of events) is a σ-algebra if*

- $\emptyset, \Omega \in \mathcal{F}$,
- *$\mathcal{F}$ is stable by complement,* $A \in \mathcal{F} \to A^c \in \mathcal{F}$,
- *$\mathcal{F}$ is stable by countable union, i.e.*

$$\textit{If } A_i \in \mathcal{F}\ \forall i \in \mathbb{N} \textit{ it holds } \cup_{i \geqslant 1} A_i \in \mathcal{F}$$

Remark 2.2. *The property of countable union is stronger then what is required for an algebra where only stability under finite union is required (plus the first two conditions).*

Exercise 1. *Prove that the above properties imply also: stability under finite union, stability under countable (and finite) intersection.*

Example 4 (Some examples of σ-algebras)**.**

- *given $E \in \Omega$, then $\{E, E^c, \emptyset, \Omega\}$ is a σ-algebra*
- *the smallest σ algebra is $\{\emptyset, \Omega\}$*
- *$\mathcal{P}(\Omega)$ is a σ-algebra*

3 Probability

Suppose that we repeat N times the same experiment, for example toss N times a coin. We fix an event A (say the event: tail) and we note $N(A)$ the number of times A occurs. Then $N(A)/N$, that we call empirical frequency of the event A, is a random quantity. Our experience says that if the number of tossing becomes large (i.e. for $N \to \infty$) the empirical frequency stabilises around a limit. This limit is our intuitive idea of probability of the event A. Note that if A and B are disjoint it holds $N(A \cup B) = N(A) + N(B)$, thus we will require the probability to stare this additive property. We will also require it to be, as the empirical frequency, a positive number in $[0, 1]$.

Definition 3.1 (Probability and probability space)**.** *Given a σ algebra $\mathcal{F}$ on a set Ω, a probability P is a function $P : \mathcal{F} \to [0, 1]$ s.t.*

- $P(\Omega) = 1$
- *if $\{A_i\}_{i \in \mathbb{N}}$ is a sequence of events that are pairwise disjoint (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$) then*

$$P\left(\cup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i)$$

We call the triple $(\Omega, \mathcal{F}, P)$ a probability space

Remark 3.2. *The notion of probability is intimately related to the (more general) notion of measure. Given a σ algebra $\mathcal{F}$ on a set Ω, a measure on $\mathcal{F}, \Omega$ is a function $\mu : \Omega \to \mathbb{R}^+ \cup \{+\infty\}$ satisfying for all sequence of pairwise disjoint elements of $\mathcal{F}$, $P\left(\cup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i)$. We call $(\Omega, \mathcal{F}, \mu)$ a measure space, so a probability space is a measure space s.t. the measure of the whole space equals one.*

Some easy consequences of the definition of probability.

Proposition 3.3 (Properties of the probability)**.**

- $P(\emptyset) = 0$
- $P(A) = 1 - P(A^c)$
- *if $A_1 \subset A_2$ then $P(A_1) \leqslant P(A_2)$*
- $P(A_1 \cup A_2) + P(A_1 \cup A_2) = P(A_1) + P(A_2)$
- $P(\cup_{i \geqslant 1} A_i) \leqslant \sum_{i \geqslant 1} P(A_i)$

Definition 3.4 (Lim inf and lim sup of sequences of events)**.** *Given a sequence of sets $(A_i)_{n\in\mathbb{N}}$ we define*

$$\liminf_{n\to\infty} A_n := \cup_{n \geqslant 1} \cap_{j \geqslant n} A_j$$

$$\limsup_{n\to\infty} A_n := \cap_{n \geqslant 1} \cup_{j \geqslant n} A_j$$

If $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$ we say that $\lim_{n\to\infty} A_n$ exists and we set it equal to $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$.

Proposition 3.5. *For a non decreasing sequence of events, namely if $A_i \subset A_{i+1}$ for all $i \in \mathbb{N}$*

- *the limit event exists and it equals $\cup_{n \geqslant 1} A_n$*
- $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$

For a non increasing sequence t the limit exists and it holds $\lim_{n\to\infty} A_n = \cap_{n \geqslant 1} A_n$ and again $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$

Proof. For a non decreasing sequence

- Since $\cap_{j \geqslant n} A_j = A_n$ it holds $\liminf_{n\to\infty} A_n = \cup_{n \geqslant 1} A_n$. On the other hand $\cup_{j \geqslant n} A_j = \cup_{j \geqslant 1} A_j$ thus $\limsup_{n\to\infty} A_n = \cup_{j \geqslant 1} A_j$

- Let $B_1 = A_1$ and for $i > 1$ let $B_i = A_i \cap A_{i-1}^c$. Then $\{B_i\}_{i \geqslant 1}$ is a sequence of pair-wise disjoint events. Furthermore it holds $A_n = \cup_{j=1}^n B_j$ which yields $P(A_n) = \sum_{j=1}^n B_j$. On the other hand since $\lim_{n\to\infty} A_n = \cup_{n \geqslant 1} A_n$ this yields $\lim_{n\to\infty} A_n = \cup_{n \geqslant 1} B_n$. Thus we have

$$\lim_{n\to\infty} P(A_n) = \sum_{n \geqslant 1} B_n \text{ and } P(\lim_{n\to\infty} A_n) = \sum_{n \geqslant 1} B_n$$

For a non increasing sequence the proof goes along the same lines. □

3.1 Probability on finite, countable and continuous spaces

If Ω is finite, $\Omega := \{\omega_1, \dots \omega_n\}$, the natural choice of σ-algebra (and actually the only one containing all the results as distinct events), is $\mathcal{P}(\Omega)$. An n-ple $\{p_i\}_{i\in[1,\dots,n]}$ with $p_i \in [0,1]$ for all i and $\sum_{i=1}^n p_i = 1$ defines a probability as follows

$$\forall A \in \mathcal{P}(\Omega) \text{ we let } P(A) = \sum_{i:\omega_i \in A} p_i$$

Example 5. *The choice $\{p_i\}_{i\in[1,\dots,n]}$ with $p_i = 1/n$ for all $i \in [1,\dots,n]$ defines the uniform measure on Ω, $P(\{\omega\}) = \frac{1}{n}$ for all $\omega \in \Omega$.*

If Ω is not finite but countable, namely in bijection with $\mathbb{N}$ so that we can set $\Omega := \{\omega_1, \omega_n \dots\}$, any sequence $(p_i)_{i \geqslant 1}$ of real positive numbers such that $\sum_i p_i = 1$ defines a probability on $\mathcal{F} = \mathcal{P}(\Omega)$.

If $\Omega = \mathbb{R}^d$ there exists a σ-algebra that is called Borel σ-algebra[1] that contains all closed intervals $[a_1, b_1] \times \dots [a_1, b_1]$ for all $a_i < b_i$ (and therefore by stability under complement it also contains all open intervals), and that is the smallest algebra containing all closed intervals. There is a result that we will not prove that says that there exists (and it is unique) a measure λ on $(B(\mathbb{R}^d), \mathbb{R}^d)$ such that the measure of an interval $I := [a_1, b_1] \times \dots [a_1, b_1]$ is $\lambda(I) = \prod_{i=1}^d (b_i - a_i)$. This is the so called Lesbegue measure. Using this result, for any given domain $V \subset R^d$ of Lebesgue measure $\lambda(V)$

[1] The concept of Borel σ algebnra is more general, for all topological spaces. How do we generate a Borel σ algebra for a metric space? take $\mathcal{P}(\Omega)$ and for any $T \subset \mathcal{P}(\Omega)$ let T_σ (resp. T_δ) be all countable unions (intersections) of elements and denoted by $B(\mathbb{R})$. Then the Borel σ algebra can be generated by starting from the collection A of all open subsets and iterating the operation $A \to (A_\delta)_\sigma$ until the first uncountable ordinal

with $\lambda(V) > 0$ and $\lambda(V) < \infty$ we can define a probability P_V that we call uniform probability measure on V s.t. for all $A \in B(\mathbb{R}^d)$ it holds

$$P_V(A) = \frac{\lambda(A \cap V)}{\lambda(V)}.$$

4 Borel-Cantelli lemma

Lemma 4.1 (Borel-Cantelli Lemma)**.** *Consider a probability space $(\Omega, \mathcal{F}, P)$ and let $\{A_n\}_{n\in\mathbb{N}}$ be an infinite sequence of events with summable probabilities, namely s.t.*

$$\sum_{n\in\mathbb{N}} P(A_n) < \infty,$$

then the event that an infinite number of these events occur has probability zero, namely

$$P(\limsup_{n\to\infty} A_n) = 0,$$

where we recall that $\limsup_{n\to\infty} A_n := \cap_{N\geqslant 0} \cup_{n\geqslant N} A_n$.

The event $\limsup_{n\to\infty} A_n$ is sometimes referred to as $\{A_n$ infinitely often $\}$ or simply $\{A_n$ i.o.$\}$

Proof. Let

$$B_N = \cup_{n\geqslant N} A_n$$

so that

$$\limsup_{n\to\infty} A_n = \cap_{N\geqslant 0} B_N.$$

Notice that $\{B_N\}_{N\in\mathbb{N}}$ is a non increasing sequence of events, therefore

$$P(\limsup_{n\to\infty} A_n) = \lim_{N\to\infty} P(B_N).$$

This, together with

$$P(B_N) \leqslant \sum_{n=N}^{\infty} P(A_n)$$

(which is obtained simply by union bound, i.e. using the property $P(A \cup B) \leqslant P(A) + P(B)$) yields

$$P(\limsup_{n\to\infty} A_n) \leqslant \lim_{N\to\infty} \sum_{n=N}^{\infty} P(A_n) = 0$$

where the last equality follows from the hypothesis $\sum_{n=N}^{\infty} P(A_n) < \infty$. □

A partial converse of the above result is given by the so called second Borel-Cantelli Lemma. We postpone this Lemma to section 3 since we will need the notion of sequence of independent events.

5 Random variables

Often we are more interested in a quantity that depends on the result of the random experiment rather than on the result itself. For example if the experiment is *toss n times a coin* and I tell you that you win 1 euro at each head and you lose 1 euro at each tail, you are probably more interested on your gain (i.e. on the total number of heads) rather than on the result itself (which is an ordered n-ple of symbols of type head and tail). We call such a quantity a random variable. More formally

Definition 5.1 (Discrete random variables)**.** *Given a probability space $(\Omega, \mathcal{F}, P)$ and a discrete (finite or countable) space E, a discrete random variable X with values in E is a measurable function, namely $X : \Omega \to E$ such that*

$$\forall x \in E \text{ it holds } \{X = x\} \in \mathcal{F}.$$

The family of positive real numbers

$$p_X(x) := P(\{\omega \in \Omega : X(\omega) = x\}), \quad x \in E$$

is called probability law of the r.v. X. From the properties of P it follows that $\sum_{x\in E} p_X(x) = 1$[2]*. Note that if we let f be a function on E then $Y = f(X)$ is also a r.v. and it has probability law*

$$p_Y(y) = \sum_{x:f(x)=y} p_x(x) \quad \forall y \in f(E)$$

We have simplified notation by denoting the event $\{\omega \in \Omega : X(\omega) = x\}$ simply by $\{X = x\}$. Note that we use capital letters to denote the r.v. and lowercase letters for their values.

Remark 5.2. *Given X and f, with $X : \Omega \to E$ a discrete r.v. and $f : E \to \mathbb{R}$ a function, also $Y = f(X)$ is a discrete r.v. . The corresponding provability law, p_Y, is easily determined from p_X, indeed it holds for all $y \in f(E)$*

$$p_Y(y) = \sum_{x:f(x)=y} p_X(x).$$

[2] In turn this means that if we consider the probability space E with σ-algebra $\mathcal{P}(E)$, the function $p_X(x)$ defines a probability on it.

Example 6. *Consider a probability space $(\Omega, \mathcal{F}, P)$ and fix an event $A \in \mathcal{F}$ and define the fonction $1_A : \Omega \to \{0,1\}$ by letting $1_A(\omega) = 1$ if $\omega \in A$, $1_A(\omega) = 0$ if $\omega \notin A$. Then $X := 1_A(\omega)$ is a discrete random variable with Bernoulli probability of parameter $p = P(A)$, namely it is a discrete random variable on $\{0,1\}$ and it holds $P_X(1) = p$.*

Definition 5.3 (Real random variables)**.** *Given a probability space $(\Omega, \mathcal{F}, P)$ a real random variable X is a mesurable function on Ω, namely $X : \Omega \to \mathbb{R}$ s.t.*

$$\forall I \text{ interval of } \mathbb{R} \text{ it holds } \{X \in I\} \in \mathcal{F}.$$

The law of X, P_X, is a probability measure on $(\mathbb{R}, B(\mathbb{R}))$ that to any $B \in B(\mathbb{R})$ gives measure

$$P_X(B) = P(\{\omega : X(\omega) \in B\}).$$

If there exists $p : \mathbb{R} \to \mathbb{R}_+$ s.t. holds for all $I \in \mathbb{R}$

$$P(X \in I) = \int_I p(x)dx$$

then we say that the r.v. X has density p (and necessaily p satisfies $\int_{\mathbb{R}} p(x)dx = 1$). It is also useful to define the partition function

$$F(x) := P_X(X \leqslant x)$$

which completely determines the law of the r.v. Note that X has density p iff F is the primitive of p, namely if $F(x) = \int_{-\infty}^{x} p(y)dy, \ \forall x \in \mathbb{R}$.

Remark 5.2 is also true for continuous r.v. provided the function f is sufficiently regular, we will not enter into technicalities here.

Exercise 2. *Use the property of probability that we detailed above to prove that the partition function satisfies*

- *$F : \mathbb{R} \to [0,1]$ is increasing*
- *$\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$*
- *F is continuous on the right*

Do all real random variables have a density? No! Consider for example X a real r.v. with density and $a \in \mathbb{R}$. Consider the r.v. $Y := \max(a, X)$ then its law has a mass at point a and a part with density on (a, ∞).

Proposition 5.4. *Given a r.v. X with density p_X, for each $a, b \in \mathbb{R}$ with $a \neq 0$, the r.v. $Y = aX + b$ also has a density and it holds*

$$p_Y(y) = \frac{1}{|a|} p_X\left(\frac{y-b}{a}\right)$$

Proof. Suppose $a > 0$. For $y \in \mathbb{R}$ oit holds

$$F_Y(y) = P(Y \leqslant y) = P(aX + b \leqslant y) = P\left(X \leqslant \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Thus

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) = \int_{-\infty}^{\frac{y-b}{a}} p_X(x)dx = \int_{-\infty}^{y} p_X\left(\frac{t-b}{a}\right)\frac{1}{a}dt$$

□

5.1 Simulating a random variable of arbitrary law

We will now show that starting from U a uniform r.v. on the interval $[0, 1]$ we can generate a r.v. of arbitrary law.

For example if we want to generate X a Bernoulli random variable of parameter p we can set

$$X := 1_{U>1-p}$$

or

$$X := 1_{U<p}$$

In general if want to generate a r.v. X that has partition function F we let

$$F^{-1}(u) := \min\{x \in \mathbb{R} : F(x) \geqslant u\}$$

(note that F^{-1} is well defined thanks to the fact that F is right continuous) and let

$$X := F^{-1}(U).$$

It is indeed easy to verify that

$$P(X \leqslant x) = P(U \leqslant F(x)) = F(x)$$

(use the fact that $F^{-1}(U) \leqslant x$) iff $U \leqslant F(x)$ and that U is a uniform r.v.) so that the partition function of X is indeed F. Nota that if the r.v. that we want to generate has a density, i.e. if $F(x) := \int_{-\infty}^{x} p(y)dy$ then X is the inverse of F.

Exercise 3. *Fix a real positive number $\lambda > 0$ and set $p(x) := \lambda e^{-\lambda x}$. Calculate F^{-1} to generate a r.v. of density p starting from the uniform r.v. U.*

Exercise 4. *Fix $p = (p_1, \ldots, p_n)$ a probability on a discrete finite set $\{x_1, \ldots, p_n\}$. The partition function is a step function $F(x) = \sum_{x_i \leqslant x} p_i$. Calculate F^{-1} to generate a discrete r.v. with the above probability law starting from the uniform r.v. U. The construction is readily extended to any countable discrete set.*

Chapter 3

Conditioning and independence

1 Conditional probability

The notion of conditional probability arises naturally when we have a partial information on the result of an experiment. Recall from section 3 that the probability of an event A corresponds to the value around which $N(A)/N$ stabilises, where $N(A)$ is the number of times the event A occurs. Suppose now that we know that the event B is verified, the frequency $N(A \cap B)/N(B)$ is an estimation of the probability of A conditioned to B. Hence the following definition is natural:

Definition 1.1 (conditional probability)**.** *Given A and B two events on the same probability space $(\Omega, P, \mathcal{F})$ such that $P(B) > 0$, we call conditional probability of A given B the following quantity*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Remark 1.2. *Note that, for any B s.t. $P(B) > 0$, the function $f : \Omega \to [0, 1]$ with $f(A) = P(A|B)$ is a probability (i.e. satisfies the requirements of Definition 3.1).*

Example 7. *Let T be an integer r.v. that represents a waiting time and set*

$$p(n) := P(T = n), \quad n > 0$$

Fixe a time $t > 0$ and suppose that we know that $T > t$, which is the law of the residual time $X = T - t$ conditioned to this event?

$$p_t(n) := P(T - t = n|T > t) = \frac{P(T - t = n \cap T > t)}{P(T > t)} = \frac{P(T = t + n)}{P(T > t)} = \frac{p(t + n)}{\sum_{m>t} p(m)}$$

Suppose that we add the hypothesis that the law of the residual waiting time is the same as the law of the waiting time, namely suppose that $p_t(n) = p(n)$ then letting $\pi(t) :=$

$\sum_{m>t} p(m)$ *summing the above formula we get*

$$\sum_{n>s} p(n) = \frac{1}{\pi(t)} \sum_{n>s} p(t+s)$$

which implies

$$\pi(s)\pi(t) = \pi(t+s)$$

which in turn implies

$$\pi(s) = a^s, \quad \textit{with } 0 < a < 1$$

where the conditions on a come from the fact that the probability should be positive and should go to zero when $s \to \infty$. This in turn implies that p should have a geometric form, namely

$$p(n) = \pi(n-1) - \pi(n) = (1-a)a^{n-1}.$$

The following result allows to express in a recursive way the probability of the intersection of n events using the conditional probability

Proposition 1.3. *Let $n \in \mathbb{N}$, $n \geqslant 2$ and $A_1, \ldots, A_n$ be events verifying $P(A_1 \cap A_2 \cdots \cap A_{n-1}) > 0$. Then the following holds*

$$P\left(\cap_{i=1}^n A_i\right) = P(A_1) \prod_{i=2}^n P(A_i \mid A_1 \cap \cdots \cap A_{i-1})$$

Proof. Let's start by noticing that for $i \leqslant n$ it holds $P(A_1 \cap \cdots \cap A_{i-1}) \geqslant P(A_1 \cap \cdots \cap A_{n-1}) > 0$ thus the conditional probabilities in the r.h.s. of the formula is well defined. We proceed by induction. For $n = 2$ the result immediately holds by the definition of conditional probability. Then using this result we get

$$P\left(\cap_{i=1}^n A_i\right) = P\left(\left(\cap_{i=1}^{n-1} A_i\right) \cap A_n\right) = P(\cap_{i=1}^{n-1} A_i) P(A_n \mid \cap_{i=1}^{n-1} A_i).$$

The result then follows by induction. □

Proposition 1.4 (Formula of total probability)**.** *Let $\{A_i\}_{i \in I}$ be a partition of Ω, namely $A_i \cap A_j = \emptyset$ for each $i \neq j$ and $\cup_{i \in I} A_i = \Omega$. The for any B it holds*

$$P(B) = \sum_{i \in I} P(B \cap A_i).$$

Furthermore, if $P(A_i) > 0 \; \forall i \in I$ it holds

$$P(B) = \sum_{i \in I} P(B|A_i) P(A_i).$$

In particular, if $0 < P(A) < 1$, it holds

$$P(B) = P(B|A)P(A) + P(B|A^c)(1 - P(A)).$$

Proof. Use $B = B \cap \Omega$ and the property of countable additivity of the probability to obtain the first formula. Then use the definition of conditional probability. □

Proposition 1.5 (Bayes formula)**.** *Given A and B two events s.t. $P(A) > 0$ and $P(B) > 0$, then from the definition of conditional probability it follows immediately that*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Furthermore, if $0 < P(A) < 1$ it holds

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B|A)P(A) + P(B|A^c)(1 - P(A))}$$

and in general if $\{A_i\}_{i \in I}$ is a partition of Ω with $P(A_i) > 0$ for all $i \in I$ it holds

$$P(A_i \mid B) = \frac{P(B \mid A)P(A)}{\sum_{j \in I} P(B|A_j)P(A_j)}$$

Proof. The result immediately follows using the definition of conditional probability and Proposition 1.4 □

Example 8 (An application of Bayes formula)**.** *Suppose that we have a population of individuals and we know that the percentage of people having Covid is $0,04\%$. Suppose we have a test at hand and we know that*

- *if a tested person is ill the test is positive in 99% of cases*
- *if a tested person is sane the test is negative in 98% of cases*

Q. If we test an individual and the test is positive, which is the probability that this individual is ill?

Let's define the two events A=the individual is ill; B=the test is positive. Then the above question translates into: which is the value of $P(A|B)$?

We know that

$$P(A) = \frac{4}{10^4}; \quad P(B|A) = \frac{99}{10^2} \quad ; P(B^c|A^c) = \frac{98}{10^2}$$

We can use Bayes to express

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{4}{10^4}\frac{99}{10^2}\frac{1}{\frac{99}{10^2}\frac{4}{10^4} + \frac{2}{10^2}\frac{10^4-4}{10^4}} \sim 0.02$$

The above conditional probability could seem very small in view of the fact the test has just 2% of false positives. The point is that, in order to have a larger value of reliability of positive test, what should be small is the ratio of the fraction of false positive w.r.t. the fraction of ills. In our case this ration is $0.02/0.04$ which is not $\ll 1$!

2 Independence of events

Given two events , A and B, we say that they are independent if $P(A\cap B) = P(A)P(B)$, which implies, for $P(B > 0)$ that $P(A|B) = P(A)$.

Example 9. *Throw two dices and define the following events*

- A=*the result of the first dice is* 3
- B= *the sum of the results is* 12
- C=*the sum of the results is* 5
- D=*the sum of the results is* 7.

Then

- *A and B are not independent, indeed $A\cap B = \emptyset$, so $P(A\cap B) = 0$ while $P(A) = 1/6$ and $P(B) = 1/12$.*
- *A and C are not independent, indeed $P(A \cap C) = 1/36$ (A and C occur if the results of the first and second dice are resp. $3, 2$) which is different from $P(C)P(A)$ (indeed $P(C) = P((1, 4) \cup (2, 3) \cup (3, 2) \cup (4, 1)) = 4/36$ and $P(A) = 1/6$.*
- *A and D are independent. Indeed one can readily verify that $P(A) = P(D) = 1/6$ (indeed $P(D) = P((1, 6) \cup (2, 5) \cup (3, 4) \cup (4, 3) \cup (5, 2), \cup(6, 1))$ and $P(A \cap D) = 1/36$.*

Definition 2.1 (Independent events)**.** *Given $(A_i)_{i\in I}$ a family of events on a probability space. We say that they are independent if for any finite subset $J \subseteq I$ with $|J| \geqslant 2$ it holds*

$$P\left(\cap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j) \tag{2.1}$$

Remark 2.2. *Note that it would not be enough to define the pairwise independence to obtain the independence nor to require $P\left(\cap_{i\in I} A_i\right) = \prod_{i\in I} P(A_i)$ to obtain that factorisation holds for any subfamily.*

For example for the experiment in which we throw two dices if we let

- *A_1 first dice yields 1, 2, or 5*
- *A_2 first dice yields 4,5,6*
- *A_3 sum of dices yields 9*

check that $P(A_1) = 1/2$, $P(A_2) = 1/2$, $P(A_3) = 4/36$ *and* $P(A_1 \cap A_2 \cap A_3) = 1/36 = P(A_1)P(A_2)P(A_3)$ *and yet* $P(A_2 \cap A_3) = P((4,5) \cup (5,4) \cup (6,3)) = 3/36 \neq P(A_2)P(A_3) = 1/18$.

The following proposition, that can be proven by induction, allows to better understand the notion of independency of family of events.

Proposition 2.3 (An equivalent definition of independent events)**.** *Given* $(A_i)_{i\in I}$ *a family of events on a probability space. The following facts are equivalent*

(i) the events $A_1, \dots A_n$ *are independent*

(ii) for each choice $B_1, \dots B_n$ *in which each* B_i *is either* A_i *or* A_i^c*, it holds*

$$P(\cap_{i=1}^n B_i) = \prod_{i=1}^n P(B_i) \tag{2.2}$$

Proof. Let's start by proving inductively that (ii) implies (i). For $n = 2$ the result is immediate (set $B_1 = A_1$ and $B_2 = A_2$ and independency follows). Suppose that we know that (ii) implies (i) for n-1 we want to extend the result to n, namely to prove that (2.2) implies (2.1) for all J. If $|J| = n$ the result follows letting $B_i = A_i$ for all i. Suppose $|J| \leqslant n-1$ and $J \subset \{1, \dots n-1\}$ then we have to prove that (2.2) implies the independence of $A_1, \dots A_{n-1}$. Now notice that for all possible choices of $B_i = A_i$ or $B_i = A_i^c$ it holds

$$P(B_1 \cap \cdots \cap B_{n-1}) = P(B_1 \cap \cdots \cap B_{n-1} \cap A_n) + P(B_1 \cap \cdots \cap B_{n-1} \cap A_n^c) =$$

$$= P(B_1) \dots P(B_{n-1})(P(A_n) + P(A_n^c)) = P(B_1) \dots P(B_{n-1})$$

where in the first equality we use that A_n, A_n^c are disjoint and their union is the whole space, and in the second inequality we use (2.2). Therefore we have proven that (2.2) holds also for n-1. Now, by induction, the independence of $A_1, \dots A_{n-1}$ follows.
Let's now prove that (i) implies (ii). In general we want to prove

$$P(A_1^c \cap .. \cap A_k^c \cap A_{k+1} \cdots \cap A_n) = P(A_1^c) ... P(A_k^c) P(A_{k+1}) \dots P(A_n)$$

If $k = 0$ the result follows immediately from (i) (choose $J = I$ in (2.1)). Let's proceed inductively on k, suppose we know that the above factorisation holds for $k-1$ then

$$P(A_1^c \cap .. \cap A_k^c \cap A_{k+1} \cdots \cap A_n) = P(A_1^c \cap .. \cap A_{k-1}^c \cap A_{k+1} \cdots \cap A_n) - P(A_1^c \cap .. \cap A_{k-1}^c \cap A_k \cap A_{k+1} \cdots \cap A_n)$$

using the inductive hypothesis on $k-1$ the desired result follows immediately.

□

An immediate corollary of the above proposition is the following

Corollary 2.4. *Given a family of independents events, if we substitute some of the events with their complementary it still remains a family of independent events.*

When we say that we make n *independent trials* we are referring to n independent events $A_1, \dots A_n$ each with the same probability Given n independent trials, each one with probability p, the probability that at least one event is verified is

$$1 - P(\cap_{i=1}^{n} A_i^c) = 1 - (1-p)^n.$$

Exercise 5 (Borel monkey)**.** *You can now solve the following paradox. Put a monkey in front of a typewriter. Prove that, if the monkey hits uniformly at random a key per second, it will almost surely (i.e. with probability one) compose the Divine Comedy if we give to it infinite time. And if we want that it composes the Divine comedy ($\sim 5 \times 10^5$ characters) with probability at least 0.99, how long should we wait?*

3 Conditional law

We will now generalise the notion of probability conditioned to an event to a probability law of a random variable X conditioned to another random variable Y. This notion will reduce the the one of conditioning on events in the setting $X = \mathbb{1}_A$, $Y = \mathbb{1}_B$.

Definition 3.1 (Conditional law: discrete setting)**.** *Let $X : \Omega \to E$ and $Y : \Omega \to F$ two discrete r.v. on the same probability space $(\Omega, \mathcal{F}, P)$, and let $p_{X,Y}$ be the joint law of the couple X, Y and p_X be the marginal of this law on X. For each $x \in E$ s.t. $p_X(x) > 0$ we let the conditional law of Y given $X = x$ be*

$$p_{Y|X}(y|x) := \frac{p_{X,Y}(x,y)}{p_X(x)} \quad \forall y \in F.$$

Note that $p_{Y|X}(\cdot|x)$ is a probability distribution on F.

Definition 3.2 (Conditional law: continuous setting)**.** *Let (X, Y) be two continuos real r.v. with a density $p_{X,Y}$. We call conditional density of Y given $X = x$ the function*

$$p_{Y|X}(y|x) := \frac{p_{X,Y}(x,y)}{p_X(x)}$$

and conditional law of Y given $X = x$ the probability on $\mathbb{R}$ which has as density $p_{Y|X}(\cdot|x)$

Chapter 4

Random variables: expectation, independence

1 Expectation of a random variable

The value of a random variable depends on the outcome of the random experiment, so it fluctuates. Thus we wish to associate to it its "more likely" value. The first and most important indicator of the most likely value is its expectation.

Definition 1.1 (Expectation)**.**

- *for a discrete r.v.* $X : \omega \to E = \{x_i, i \geqslant 1\}$ *of law* p *we define the expectation as*

$$\mathbf{E}(X) = \sum_{i \geqslant 1} x_i p(x_i)$$

  if $\sum_{i \geqslant 1} |x_i| p(x_i) < \infty$*. In this case we say that* X *is integrable*

- *for a real r.v.* X *of density* p *we define its expectation by*

$$\mathbf{E}(X) = \int_{\mathbb{R}} x p(x) dx$$

  if $\int_{\mathbb{R}} |x| p(x) dx < \infty$*. In this case we say that* X *is integrable*

- *for a (discrete or real) r.v.* X *that is not integrable yet it is positive we can again define its expectation as above. Instead, if the variable is neither positive and nor absolutely convergent, we cannot define its expectation.*

Why do we ask "absolute" convergence? because we do not want the result to depend on how we chose to enumerate the events in the set (easier to understand in the discrete case). A remark: usually in physics jargon expectation is rather called the mean and denoted as $< X >$ instead then with $E(X)$.

1.1 Properties of the expectation

The following proposition easily follows from the definition of expectation.

Proposition 1.2 (Properties of the expectation of discrete r.v.)**.** *Given $X : \Omega \to E$ a discrete r.w. the following holds*

(i) Recall from Remark 5.2 that for any function $f : E \to \mathbb{R}$, $Y = f(X)$ is a discrete r.v.. If $f \geqslant 0$ or if $\sum_y |y| p_Y(y) = \sum_x |f(x)| p_X(x) < \infty$ then it holds

$$\mathbb{E}(f(X)) = \sum_x f(x) p_X(x).$$

(ii) $f \to \mathbb{E}(f(X))$ is linear. In particular, if X is integrable, for all $a, b \in \mathbb{R}$ also $Y = aX + b$ is integrable and

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$$

(iii) Positivity. If X takes only positive vakues then $\mathbf{E}(X) \geqslant 0$ and $\mathbf{E}(X) = 0$ iff $P(X = 0) = 1$.

(iv) Given $f, g : \Omega \to E$ s.t. $\mathbf{E}(f)$ and $\mathbf{E}(g)$ are defined and s.t. $f \leqslant g$ then $E(f) \leqslant E(g)$. In particular, if X is integrable it holds $|\mathbf{E}(\mathbf{X})| \leqslant \mathbf{E}(|X|)$ and if $|X| \leqslant a$ it holds $|E(x)| \leqslant a$ and in X

The analogous results or the continuous case are contained in the following Proposition, which follows from the definition of expectation and the use Fubini theorem to switch the order of integrals (which is guaranteed by absolute convergence).

Proposition 1.3 (Properties of the expectation of real r.v. with density)**.** *Let X be a continuous r.v. with density p*

(i) if X is positive

$$\mathbf{E}(\mathbf{X}) = \int_0^\infty P(X > x) dx$$

(ii) if $f : \mathbb{R} \to \mathbb{R}^+$ is continuous and $f(X)$ also has a density. then

$$\mathbf{E} f(X) = \int_{\mathbb{R}} f(X) p_X(x) dx$$

(iii) if the corresponding expectation are well defined, the same results as in Proposition 1.2(ii),(iii),(iv) apply.

Note that given X, Y two (real or discrete) square integrable r.v., namely s.t. X^2 and Y^2 are integrable, then also the variables XY and $(X+Y)^2$ are integrable. This follows from the above proposition and the inequalities

$$2|XY| \leqslant X^2 + Y^2, \quad (X+Y)^2 \leqslant 2(X^2+Y^2)$$

Exercise 6. *Prove that if X is square integrable, then necessarily X is also integrable. More generally, for any two integers r, s s.t. $0 < s < r$ if $E|X|^r < \infty$ this implies $E|X|^s < \infty$.*

Hint 1.4. *Use $|X|^r = |X|^r(\mathbb{1}_{|X| \geqslant 1} + \mathbb{1}_{|X|<1})$*

Exercise 7. *Prove that $E(X) < \infty$ implies $P(\{X < \infty\}) = 1$.*

Hint 1.5. *Suppose by contradiction that $P(\{X < \infty\}) < 1$ then deduce that there exists $c > 0$ such that for any N it holds $P(\{X > N\}) > c > 0$, and use this to that $E(X)$ cannot be finite.*

Remark 1.6. *The converse is not true: $P(\{X < \infty\}) = 0$ does not imply $E(X) < \infty$. We will see in section* **??** *an example of variable which has infinite mean though it is finite with probability one, the first time a simple symmetric one dimensional r.v. starting from the origin comes back to the origin.*

1.2 Jensen and Markov inequalities

Let us state two key inequalities involving means of random variables.

Proposition 1.7 (Jensen inequality)**.** *Given a probability space $(\Omega, \mathcal{F}, P)$ and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function, for any integrable variable X s.t. also $\phi(X)$ is integrable it holds*

$$\phi(\mathbf{E}X) \leqslant \mathbf{E}(\phi(X)).$$

Proof. From the convexity of ϕ it follows that for any x_0 it holds

$$\phi(x) \geqslant \phi(x_0) + \phi'(x_0)(x - x_0)$$

Therefore

$$\phi(X) \geqslant \phi(E(x)) + \phi'(x_0)(X - E(X)).$$

By taking the expected value of the above expression and using the properties of the expectation we obtain the desired inequality. □

As for Cauchy-Schwartz inequality, you might meet Jensen inequality in other mathematics fields (of course stated in different forms), it is a basic inequality upper bounding the convex function of an integral with the integral of the convex function.

Proposition 1.8 (Markov inequality)**.** *Given a positive real random variable X, and a positive real number $a > 0$ it hols*

$$P(\{X > a\}) \leqslant \frac{E(X)}{a}$$

Hint 1.9. *Set $X = X\mathbb{1}_{X>a} + X\mathbb{1}_{X \leqslant a}$*

2 Variance and moments

For any integer $k \geqslant 1$ the k-th moment of the real r.v. X is, by definition, $\mathbf{E}X^k$, it this expectation exists. If $k = 1$ it is the expectation, if $k = 2$ we call the centred moment the variance, namely

$$\text{Var}(X) := \mathbf{E}[(X - \mathbf{EX})^{\mathbf{2}}]$$

which, being the mean square displacement from the expectation, represents the dispersion of the law. The quantity which represents the typical fluctuations from the mean is thus

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

By linearity of expectation we get for any real a, b

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

and

$$\text{Var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(\mathbf{X}))^{\mathbf{2}}.$$

Exercise 8. *Calculate the expectation and the variance of X in the following cases*

- *X is a Bernoulli law of parameter p with $p \in [0, 1]$, namely it is a continuous random variable with density $\Omega = \{0, 1\}$ and $P(\{X = 1\}) = p$*
- *X is a binomial $B(n, p)$ with $n \in \mathbb{N}$ $p \in [0, 1]$, namely $\Omega = \mathbb{N}$ and $P(\{X = k\}) = \binom{n}{k} p^k (1 - p)^{n-k}$*
- *X is an exponential variable with parameter $\lambda >$ namely it is a continuous random variable with density $p(x) = \lambda e^{-\lambda x}$ for $x \geqslant 0$, $p(x) = 0$ for $x < 0$*
- *X is a it is a r.v. with uniform law on an interval $[a, b]$ with $a, b \in \mathbb{R}$ $-\infty < a < b < \infty$, namely it is continuous random variable with densty $p(x) = (b - a)^{-1} 1_{[a,b]}(x)$*
- *X is a r.v. with Poisson law of parameter λ, $\lambda > 0$, namely it is a discrete r.v. with $\Omega = \mathbb{N}$ and $p(n) = e^{-\lambda}\lambda^n/n!$*

- *a real variable X of density*

$$p_x(x)\frac{1}{\sigma\sqrt{2\pi}}\exp\{-(x-\mu)^2/(2\sigma^2)\}$$

with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. This law is called normal or Gaussian law of parameters μ and σ^2

Prove also that for $n \to \infty$ nd $p \to 0$ with $np \to \lambda$ the law of the binomial variable $B(n,p)$ goes to the law of the Poissosn variable p_λ, namely for all $n \in \mathbb{N}$ it holds $\binom{n}{k}p^k(1-p)^{n-k} \to B_{n,p}(n) \to e^{-\lambda}\lambda^n/n!$

2.1 Quantifying dispersion: Chebyshev inequality

Theorem 2.1 (Chebychev inequality)**.** *Given X a r.v. s.t. X^2 is integrable, and $a \in \mathbb{R}_+$ it holds*

$$P(|X - \mathbf{E}(X)| \geqslant a) \leqslant \frac{1}{a^2}Var(X).$$

Proof. Let Y be a variable s.t. Y^2 is integrable. Then

$$a^2\mathbb{1}_{|Y| \geqslant a} \leqslant Y^2$$

Thus using $E(f) \leqslant E(g)$ for $f \leqslant g$ we get

$$\mathbf{E}(\mathbb{1}_{|Y| \geqslant a}) = P(|Y| \geqslant a) \leqslant a^2\mathbf{E}(Y^2).$$

Chebychev inequality is obtained by setting $Y = X - \mathbf{E}(X)$. □

2.2 Covariance and correlation

To study simultaneously two or more random variables we need to introduce the notion of covariance and correlation.

Definition 2.2 (Covariance and correlation)**.** *Given X, Y square integrable variables we define their covariance as*

$$\mathrm{Cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right].$$

When $Var(X)$ and $Var(Y)$ are non zero, we define the correlation as

$$Cor(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

From the above definition it follows that

Proposition 2.3.

(i) If $(X_1 \dots X_n)$ are n square integrable r.v. then

$$Var\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} Var(X_j) + 2\sum\sum \mathrm{Cov}(X_\ell, X_m)$$

(ii) $Cor(X, Y) \in [-1, 1]$

Proof. To prove (i) we set $S_n := \sum_{i=1}^{n} X_i$ and notice that

$$S_n - \mathbf{E}(S_n) = \sum_{i=1}^{n} [X_i - \mathbf{E}(X_i)]$$

which yields

$$[S_n - \mathbf{E}(S_n)]^2 = \sum_{i=1}^{n} [X_i - \mathbf{E}(X_i)]^2 + 2\sum_{m=1}^{n}\sum_{\ell<m} [X_\ell - \mathbf{E}(X_\ell)]\,[X_m - \mathbf{E}(X_m)]\,.$$

To prove (ii) we note that the statement is equivalent to

$$|\mathbf{E}\,[Z_1 Z_2]\,| \leqslant \sqrt{\mathbf{E}(Z_1^2)}\sqrt{\mathbf{E}(Z_2^2)}$$

with $Z_1 := X - \mathbf{E}(X)$ and $Z_2 := Y - \mathbf{E}(Y)$. In turn, to prove this inequality it is sufficient to notice that $|\mathbf{E}\,[Z_1 Z_2]\,| \leqslant \mathbf{E}\,[|Z_1 Z_2|]$ and to apply the Cauchy-Schwartz inequality which we state and prove below separately since it is of much more general interest.

□

2.3 Cauchy-Schwartz inequality

Let us state and prove here an inequality which is often used in probability (and actually in many other mathematics fields, where it can take different formulations.

Proposition 2.4 (Cauchy-Schwartz inequality)**.** *Let X, Y be two random variables then*

$$E(|XY|) \leqslant \sqrt{E(X^2)}\sqrt{E(Y^2)}$$

Proof. There are several different ways of proving the inequality, one being the following. For $\lambda \in \mathbb{R}$ define

$$f(\lambda) := \mathbf{E}(|X| + \lambda|Y|)^2 = \lambda^2\mathbf{E}(Y^2) + 2\lambda\mathbf{E}(|XY|) + \mathbf{E}(X^2)$$

Note that $f(\lambda) \geqslant 0$ for all λ. Therefore the discriminant of the above second order equation in λ should be non positive, otherwise we could write $f(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)$ with $\lambda_1, \lambda_2 \in \mathbb{R}$ and find an interval of λ for which $f(\lambda) < 0$. This implies

$$4(\mathbf{E}(|XY|))^2 - 4\mathbf{E}(Y^2)\mathbf{E}(X^2) \leqslant 0$$

which immediately yields the desired result. □

We also mention without proof a useful generalisation of the C-S inequality which is called Holdër inequality.

Proposition 2.5 (Holdër inequality)**.** *For $p, q \in [1, \infty]$ s.t. $1/p + 1/q = 1$ it holds*

$$\mathbf{E}(|XY|) \leqslant (\mathbf{E}(|X^p|)^{1/p})(\mathbf{E}(|X^q|)^{1/q}$$

3 Independent random variables

Definition 3.1 (Independent random variables)**.**

- *Given a probability space $(\Omega, \mathcal{F}, P)$ and n discrete r.v. $X_1, \dots X_n$ with $X_i : \Omega \to E$ on this space, we say that they are independent if*

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \ , \quad \forall (x_1, \dots x_n) \in E^n$$

- *Given n real r.v. $X_1, \dots X_n$ defined on the same probability space $(\Omega, \mathcal{F}, P)$ we say that they are independent if*

$$P(X_i \in I_i, i = 1, \dots n) = \prod_{i=1}^{n} P(X_i \in I_i)$$

  for any choice of the intervals $I_1, \dots I_n$. In particular, if the n-ple$(X_1, \dots X_n)$ has a density, independence corresponds to factorisation of the joint density.

In view of Definition 2.1 of independents of events it could seem strange that we do not require the above property also for any subfamily. However this holds automatically, namely it also holds

$$P(X_{i_1} \in I_{i_1} \cap X_{i_k} \in I_{i_k}) = \prod_{j=1}^{k} P(X_{i_j} \in I_{i_j})$$

since we can get rid of the other variables X_j with $j \notin \{i_1, \dots i_k\}$ by choosing $I_j = \Omega$.

An easy consequence of the above definition is the following.

Proposition 3.2.

- *Given n independent r.v., $X_1, \dots X_n$, and n real functions $f_1, \dots f_n$ such that for all $i \in [1, n]$ $f_i : \mathbb{R} \to \mathbb{R}$ and $f_i(X_i)$ is integrable, it holds*

$$\mathbf{E} \prod_{i=1}^{n} f_i(X_i) = \prod_{i=1}^{n} \mathbf{E} f_i(X_i)$$

- *if X and Y are independent and square integrable it holds*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(y)$$

  which, using proposition 2.3, yields $\mathrm{Cov}(X, Y) = 0$

It is also useful to introduce the weaker notion of sequence of pairwise independent random variables. We say that $X_1, \dots X_n$ are pairwise independent if $\forall i \neq j$ X_i and X_j are independent.

Proposition 3.3. *Given two independent real r.v. X, Y with density p_X, p_Y, so that $p_{X,Y} = p_X p_Y$ their sum has also a density and it holds*

$$p_{X+Y}(z) = \int_{\mathbb{R}} p_X(x) p_Y(z - x) dx$$

Proof. The proof follows immediately by using factorisation of the density of independent random variables and Fubini theorem. □

A natural generalisation of the notion of independent r.v. to an infinite countable set of r.v. is the following.

Definition 3.4 (Sequence of independent random variables)**.** *Given a sequence $\{X_n\}_{n \geqslant 1}$ of random variables it is called a sequence of independent r.v. if for any $n \geqslant 1$ it holds that $X_1, \dots, X_m$ are m independent r.v.*

Lemma 3.5 (Second Borel-Cantelli Lemma)**.** *Consider a probability space $(\Omega, \mathcal{F}, P)$ and let $\{A_n\}_{n \in \mathbb{N}}$ be an infinite sequence of independent events it holds that with non summable probabilities, namely*

$$\sum_{n \in \mathbb{N}} P(A_n) = \infty,$$

then the probability that infinitely many of these event occurs is one, namely

$$P(\limsup_{n \to \infty} A_n) = 1.$$

Proof. The result of the lemma is equivalent to

$$P\left((\cap_{N\geqslant 0}\cup_{n\geqslant N}A_n)^c\right)=0.$$

On the other hand $(\cap_{N\geqslant 0}\cup_{n\geqslant N}A_n)^c=\cup_{N\geqslant 0}\cap_{n\geqslant N}A_n^c$ so that

$$P\left((\cap_{N\geqslant 0}\cup_{n\geqslant N}A_n)^c\right)=\lim_{N\to\infty}P(\cap_{n=N}^{\infty}A_n^c).$$

Using the independence on the events to rewrite the probability of the intersection as a product of the probabilities we get

$$P\left((\cap_{N\geqslant 0}\cup_{n\geqslant N}A_n)^c\right)=\lim_{N\to\infty}\prod_{n=N}^{\infty}P(A_n^c)=\lim_{N\to\infty}\prod_{n=N}^{\infty}(1-P(A_n))=0$$

□

3.1 Coin tossing: an example of i.i.d. variables

Consider the following experiment: throw n times a coin which yields head with probability p and tail with probability $1-p$ at each trial. Thus we set $\Omega=\{0,1\}^n$ and

$$P(\{\omega\})=p^{K(\omega)}(1-p)^{n-K(\omega)},\quad \text{with } K(\omega)=\text{ total number of heads.}$$

We write each $\omega\in\Omega$ as $(\omega_1,\ldots,\omega_n)$ with $\omega_i\in\{0,1\}$ and let $\omega_i=1$ (resp. $\omega_i=0$) correspond to the i-th toss giving head (resp. tail) so that $K(\omega)=\sum_{i=1}^n\omega_i$. Note that K is a random variable taking values in $E:=\{0,1,\ldots,n\}$ and

$$P(\{K=j\})=\sum_{\omega:\sum_{i=1}^n\omega_i=j}P(\{\omega\})=\binom{n}{j}p^j(1-p)^{n-j}$$

Thus K follows the Binomial law with parameters n and p which we denote by $k\sim B(n,p)$.
Note that if we consider the n random variables $X_1,\ldots X_n$ with $X_i(\omega)=\omega_i$, it holds

$$P(X_1=x_1,\ldots,X_n=x_n)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}==\prod_{i=1}^n P(X_i=x_i)$$

thus the variables $X_1,\ldots X_n$ are independent.

We shall now provide an analytic construction of the coin-tossing game with infinite trials. More precisely the goal here is to provide an explicit construction of a probability space $(\Omega,\mathcal{F},P)$ on which we can define $\{X_n\}_{n\in\mathbb{N}}$ that are independent and s.t. each

marginal is Bernoulli(1/2). Consider the probability space $(\Omega, \mathcal{F}, P)$ with Ω the real interval $\mathcal{R} \cap [0,1]$, $\mathcal{F}$ the Borel σ-algebra on this interval and P the uniform measure, that we call λ. For each $\omega \in \Omega$ we can write its unique dyadic expansion [1] which reads

$$\omega = (0, \omega_1, \dots \omega_n \dots) \text{ with } \omega = \sum_{n \geqslant 1} \frac{\omega_n}{2^n}.$$

For each $n \in \mathbb{N}$ let now X_n be the r.v. defined by $X_n(\omega) = \omega_n$ so that $X_n(\omega) \in \{0,1\}$. By definition for each n and for each choice $x_1, \dots x_n \in \{0,1\}^n$ it holds

$$\{X_i = x_i; i \in \{1, \dots, n\}\} = \left\{ \omega \in \left[\sum_{i=1}^{n} \frac{x_i}{2^i}, \sum_{i=1}^{n} \frac{x_i}{2^i} + \frac{1}{2^n} \right] \right\}$$

so that

$$P(\{X_i = x_i; i \in \{1, \dots, n\}) = \frac{1}{2^n}$$

and $\{X_i\}_{i \in \mathcal{N}}$ is a sequence of i.i.d. r.v. of law Ber(1/2).

4 Generating functions

For integer positive r.v., t is convenient to introduce their generating function to simplify calculations on these variables.

Definition 4.1 (Generating function)**.** *Given $X : \Omega \to \mathbb{N}$ a positive r.v. with integer values, we let*

$$G_X(z) = \mathbf{E}(z^X) = \sum_{x \in \mathbb{N}} z^x p_X(x), \quad \forall z \in \mathbb{C}, \ |z| \leqslant 1$$

Exercise 9. *Calculate the generating function of the following r.v.*

- *X distributed as a geometric function of parameter a with $a \in \mathbb{R} \cap (0,1)$, namely a function s.t.*
$$P(X = i) = (1-a)^{i-1} a, \ \forall i \in \mathbb{N}^+$$
- *X distributed as a binomial $B(n,p)$*

For a positive r.v. with integer values that can also take infinite value, namely for $X : \Omega \to \mathbb{N} \cup \{\infty\}$ we can define the generating function in the same way and we get

$$P(X = \infty) = 1 - \sum_{k \geqslant 0} P(X = k) = 1 - \sum_{z \to 1} G_X(z).$$

From the above definition the following properties of the generating function follow:

[1] We exclude expansion ending with an infinity of 1's which makes the dyadic expansion unique

Proposition 4.2 (Properties of the generating functions)**.**

- *G_X is a series of radius of convergence $\geqslant 1$*
- *G_X is a continuous function on $|z| \leqslant 1$, infinitely differentiable*
- *from the knowledge of G_X we can completely determine the law of X, indeed*

$$\left[\frac{d^n}{dz^n} G_X(z)\right]_{z=0} = n! p_X(n)$$

- *In particular it holds*

$$\lim_{z\to 1} \frac{d}{dz} G_X(z) = \sum_{n\in\mathbb{N}} n p_X(n) = \mathbf{E}(X)$$

- *more generally for $k \geqslant 1$*

$$\lim_{z\to 1} \frac{d^k}{dz^k} G_X(z) = \mathbf{E}\left[X(X-1)\dots(X-k+1)\right]$$

which yields in particular

$$\mathbf{E}\left[X^2\right] = \mathbf{E}\left[X(X-1)\right] + \mathbf{E}\left[X\right] = \lim_{z\to 1}\left[\frac{d^2}{dz^2} G_X(z) + \frac{d}{dz} G_X(z)\right]$$

- *given n independent integer positive r.v. $X_1, \dots X_n$, the generating function of the sum equals the product of the generating functions, namely letting $S_n = X_1 + \dots X_n$ it holds*

$$G_{S_n}(z) = \prod_{i=1}^{n} G_{X_i}(z)$$

The proof of the proposition is left as an exercise

Hint 4.3. *For the last point use $G_{S_n}(z) = \mathbf{E}(z^{S_n}) = \mathbf{E}\prod_{i=1}^n z_i^X = \prod_{i=1}^n \mathbf{E} z_i^X$.*

5 Conditional expectation

Recall the definitions of conditional laws given in Section 3 , we are now ready to introduce the notion of conditional expectation.

Definition 5.1 (Conditional expectation)**.** *Given $X : \Omega \to E$ and $Y : \Omega \to F$ two discrete r.v. we let the conditional expectation of Y given X be the random variable $\phi : E \to \mathbb{R}$ which associates to $x \in E$ the value $E(Y|X = x)$ defined as*

$$E(Y|X = x) := \sum y p_{Y|X}(y|x).$$

Given (X, Y) a couple of real random variables with a density, the conditional expectation of Y given X is a real random variable $\phi : \mathbb{R} \to \mathbb{R}$ which associates to $x \in \mathbb{R}$ s.t. $p_X(x) \neq 0$ the value

$$E(Y|X = x) := \int_{\mathbb{R}} y p_{Y|X}(y|x) dy.$$

From the above definition the following properties follow

Proposition 5.2 (Properties of the conditional expectation)**.**

(i) $E(E(Y|X)) = E(Y)$

(ii) for all real a, b it holds $E(aY + bZ|X) = aE(Y|X) + bE(Z|X)$

(iii) $E(Y|X) \geqslant 0$ *if* $Y \geqslant 0$

(iv) $E(1|X) = 1$

(v) $E(Y g(X)|X) = g(X)E(Y|X)$

Proof. Let's prove (i) in the discrete case, the continuous case is analogous (with sums replaced by integrals).

$$E(E(Y|X)) = \sum_x p_X(x) E(Y|X = x) = \sum_{x,y} y p_Y(y|x) p_X(x) = \sum_{x,y} y p_{X,Y}(x, y) = \sum_y p_y(y) = E(Y)$$

Properties (ii)-(iv) are an easy consequence of the analogous properties for the expectation. Property (v) follows from the fact that when we evaluate the conditional expectation given X, this variable should be considered as a constant. □

Note that if (X, Y) are two independent r.v. (see Definition 3.1) then $p_{Y|X}(\cdot|x) = p_Y(\cdot)$ and $E(Y|X)$ is a constant function equal to $E(Y)$. Therefore we can reobtain more directly the first result of Proposition 3.2:

$$E(f(X)g(Y)) = E(E(f(X)g(Y)|X)) = E(f(X)E(g(Y)|X)) =$$

$$= E(f(X)E(g(Y)) = E(f(X))E(g(Y))$$

The first point of Proposition 5.2 often turns out to be very useful to evaluate the expectation of a random variable. Indeed, it is sometimes easier to proceed by conditioning on another auxiliary variable and then taking the expectation.

Example 10. *The number of people passing in front of Eiffel tower every day follows a poisson law of parameter $\lambda > 0$. Each person decides (independently from the others) to stop and visit the tower with probability p. Calculate the mean number of persons that visit the Eiffel tower today in terms of λ and p.*
We note V the r.v. representing the number of people visiting the tower today and N the number of peoples passing in front. We know that

$$p_{V|N}(v|n) = \binom{n}{v} p^v (1-p)^{n-v}$$

and that

$$p_N(n) = \frac{e^{-\lambda}\lambda^n}{n!}$$

So we get

$$E(V) = E(E(V|N)) = \sum_n \frac{e^{-\lambda}\lambda^n}{n!} \sum_v v \binom{n}{v} p^v (1-p)^{n-v} = \sum_n \frac{e^{-\lambda}\lambda^n}{n!} pn = \lambda p$$

Example 11. *The number of clients arriving every day to a barber shop is a random variable N of mean c and variance v_c. The duration of the haircut required by client i is a random variable X_i. We suppose $\{X_i\}_{i\in\mathbb{N}}$ are i.i.d. and square integrable of mean t and variance v_t. Determine the mean and the variance of the total time T the barber spends cutting hairs per day (in terms of c, t, v_c, v_t).*

Since $T = \sum_{n\in\mathbb{N}} \mathbb{1}_{N=n} \sum_{i=1}^n X_i$ it holds

$$E(T) = E(E(T|N)) = \sum_{n\in\mathbb{N}} p(n)nt = ct$$

$$E(T^2) = E(E(T^2|N)) = \sum_{n\in\mathbb{N}} p(n)E(\sum_{i,j=1}^n X_iX_j) = \sum_{n\in\mathbb{N}} p(n)\left[nE(X_i^2) + n(n-1)t^2\right] = cE(X_i^2)+E(N^2)t^2-ct^2$$

Hence

$$Var(T) = cE(X_i^2) + E(N^2)t^2 - ct^2 - c^2t^2 = cVar(X_i) + t^2Var(N) = cv_t + t^2v_c$$

6 Random sums of random variables

In both examples of previous section we were handling with a sum of N independent realisation of a random variables, where N itself was a random variable. In both cases we discovered that the mean of the sum is given by the product of the mean of N and of the mean of a single realisation of the random variable. The following result proves that this is a general fact and it gives a full characterisation of the distribution of the random sum.

Proposition 6.1. *Given a sequence $\{X_n\}_{n\in\mathbb{B}}$ i.i.d. integer positive random variables and ν an integer positive r.v. independent from this sequence, consider the variable S defined by setting $S = 0$ if $\nu = 0$ and otherwise*

$$S := \sum_{i=1}^{\nu} X_i.$$

Then the generating function of S satisfies

$$E(z^S) = G(\phi(z)),$$

where

$$G(z) = E(z^\nu) \quad \textit{and} \quad \phi(z) = E(z^{X_i}).$$

Proof. Let p be the distribution of ν, then

$$E(z^S) = E(E(z^S|\nu)) = \sum_{n=1}^{\infty} p(n) \prod_{i=1}^{n} E(z^{X_i}) = \sum_{n=1}^{\infty} p(n)\phi(z)^n = G(\phi(z))$$

□

Exercise 10. *Suppose that the hypothesis of proposition 6.1 hold with $\nu \sim$ Poisson of parameter λ and $X_i \sim$ Bernoulli of parameter p. Prove that S follows a Poisson law of parameter λp.*
Suppose that the hypothesis of proposition 6.1 hold with $\nu \sim$ a geometric distribution of mean a (namely $P(X = k) = (1 - \frac{1}{a})^{k-1}\frac{1}{a}$ for $k \in [1, 2, \dots]$) and X_i a geometric distribution of mean b. Prove that S follows a geometric distribution of mean ab.

Chapter 5

Infinite sequences of random variables

1 Convergences

We shall introduce here different notions if convergence.

Definition 1.1. *Given $\{X_n\}_{n \geqslant 0}$ a sequence of real random variables and X a random variable we say that*

- *(X_n) converges almost surely (a.s) to X if $P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$. In this case we write $X_n \xrightarrow[n\to\infty]{a.s.} X$*
- *(X_n) converges in L^p to X with $p \geqslant 1$, $p \in \mathbb{N}$, if $E(|X - X_n|^p) \to 0$. In this case we write $X_n \xrightarrow{L^p} X$*
- *(X_n) converges in probability to X if $\forall \epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$. In this case we write $X_n \xrightarrow{P} X$*
- *(X_n) converges in law to X if for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$ it holds $E(f(X_n)) \to E(f(X))$. In this case we write $\mu_{X_n} \to \mu_X$.*

2 Links among the different notions of convergence

Proposition 2.1.

(i) Almost sure convergence implies convergence in probability

(ii) Convergence in L^r for any $r \geqslant 1$ implies convergence in probability

(iii) Convergence in probability implies convergence in law.

Proof. (i) Fix $\epsilon > 0$, the sequence $(f_n)_{n \geqslant 0}$ with $f_n = \mathbb{1}_{|X_n - X| > \epsilon}$ converges a.s. towards 0. Note also that $f_n < 1$ for all n, thus applying the dominated convergence theorem 0.6 we get $\lim_{n\to\infty} E(f_n) = \lim_{n\to\infty} P(|X_n - X| > \epsilon) \to 0$

(ii) Fix $\epsilon > 0$. Then proceed as in the proof of Markov inequality to get

$$E(|X_n - X|^p) = \int_\Omega |X_n - X|^p dP =$$

$$= \int_{|X_n - X| > \epsilon} |X_n - X|^p dP + \int_{|X_n - X| \leqslant \epsilon} |X_n - X|^p \geqslant \epsilon^p P(|X_n - X| > \epsilon$$

Hence, since by convergence in L^p it holds $\lim_{n\to\infty} E(|X_n - X|^p) \to 0$, convergence in probability follows.

(iii) Given f a real continuous function with compact support. Then f is absolutely contunuous, i.e. for all $\epsilon > 0$ there exists $\alpha > 0$ s.t. $|x - y| < \alpha$ impolies $|f(x) - f(y)| < \epsilon$. Then note that

$$|E(f(X_n) - f(X)| \leqslant E(|f(X_n) - f(X)|) = E(|f(X_n) - f(X)|(\mathbb{1}_{|X_n - X| \geqslant \alpha} + \mathbb{1}_{|X_n - X| < \alpha})$$

$$\leqslant 2 \sup f P(\{|X_n - X| \geqslant \alpha\})$$

Furthermore, by the hypothesis of convergence ion probability, for n sufficiently large it holds $P(\{|X_n - X| \geqslant \alpha\}) < \epsilon$ hence we get for n sufficiently large and for any fixed ϵ $|E(f(X_n) - f(X)| \leqslant 3\epsilon$ which implies $E(f(X_n)) \to E(f(X)$ and hence convergence in law.

□

Remark 2.2. *The implications of proposition 2.1 among the different types of convergence are the only ones that hold in general. Indeed for all other types of applications it is possible to construct counterexamples*

The dominated convergence theorem 0.6 applied to the case in which the measure μ is a probability measure yields a link between almost sure convergence and L^1 convergence under additional hypothesis.

Proposition 2.3 (Dominated convergence)**.** *Let (X_n) convergence a.s. towards X, and suppose there exists an integrable random variable Y (i.e. s.t. $E(|Y| < \infty)$ s.t. $|X_n| \leqslant |Y|$ for all n. Then convergence in L^1 holds, namely $X_n \xrightarrow{L^1} X$.*

Proposition 2.4. *If $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$ with $* =$ a.s. or P or L^p, then*

$$X_n + Y_n \xrightarrow{*} X + Y$$

Exercise 11. *Prove the above proposition*

Proposition 2.5. *If $\{X_n\}_{n \geqslant 0}$ converges in probability to X, there exists a sub-sequence (X_{n_k}) that converges to X a.s.*

Proof. The proof is an application of Borel-Cantelli lemma (see Lemma 4.1). For all $k \geqslant 1$ we can find n_k s.t.

$$P(|X_{n_k} - X| > \frac{1}{k}) \leqslant \frac{1}{k^2},$$

which implies

$$\sum P(|X_{n_k} - X| > \frac{1}{k}) < \infty$$

which, using Borel-Cantelli, implies that the probability that A_k i.o. with $A_k := |X_{n_k} - X| > \frac{1}{k}$ has probability zero. Hence there exists N s.t. the complementary events A_k^c occur with probability one for $k > N$. This implies $X_{n_k} \to X$ a.s. □

Chapter 6

Law of large numbers (LLN)

The law of large numbers is a result that formalises the experimental observation that if we repeat n times the same experiment, the empirical mean of a random variable that depends on the outcome of the experiment concentrates around the mean .

1 Weak LLN

Theorem 1.1 (Weak LLN)**.** *Consider a sequence $\{X_i\}_{i\in\mathbb{N}M}$ of independent square integrable r.v. with the same law and define their empirical mean $\overline{X_n}$ as*

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then $\overline{X_n}$ convergences in probability to $\mathbf{E}(X_1)$, namely

$$P(|\overline{X_n} - \mathbf{E}(X_1)| \geqslant \epsilon) = 0, \ \ \forall \epsilon > 0$$

Proof. By linearity of expectation it holds

$$\mathbf{E}(\overline{X_n}) = 1/n \sum_{i=1}^{n} \mathbf{E}(X_i) = \mathbf{E}(X_1).$$

By using $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$ and independence we get

$$\mathrm{Var}(\overline{X_n}) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}$$

with $\sigma^2 = \mathrm{Var}(X_i)$. Thus applying Chebyshev inequality (see Theorem 2.1) to $\overline{X_n}$ we get

$$P(|\overline{X_n} - \mathbf{E}(X_1)| \geqslant \epsilon) \leqslant \frac{1}{\epsilon^2}\frac{\sigma^2}{n}$$

which concludes the proof. □

2 Strong LLN

A stronger result holds

Theorem 2.1 (Strong LLN)**.** *Consider a sequence $\{X_i\}_{i\in\mathbb{N}}$ of independent integrable r.v. with the same law. Their empirical mean $\overline{X_n}$ converges with probability 1 to $\mathbf{E}X_1$, namely*

$$P(\lim_{n\to\infty} \overline{X_n} = \mathbf{E}(X_1)) = 1,$$

Remark 2.2. *An even stronger LLN holds: independence can be substituted by pair-wise independence (strong law of Etemadi)*

Proof. For simplicity we assume $\mathbf{E}(X_1) = 0$ (otherwise we can recenter the variable, make the proof for $X_i - \mathbf{E}(X_1)$ and then deduce the desired result). We will also make th (non innocent but simplifying!) hypothesis that $\mathbf{E}(X_1^4) < \infty$ (if it is not the case the proof is more involved). Then

$$\mathbf{E}(\overline{X_n}^4) = \frac{1}{n^4}\sum_{ijkl}\mathbf{E}(X_iX_jX_kX_l) = \frac{1}{n^4}\left[\sum_{i\neq j}\mathbf{E}(X_i^2X_j^2) + \sum_i \mathbf{E}(X_i^4)\right] =$$

$$= \frac{1}{n^4}\left[n(n-1)(\mathbf{E}(X_1^2))^2 + n\mathbf{E}(X_1^4)\right] = O\left(\frac{1}{n^2}\right)$$

were we used the fact independence and the fact that X_i have zero mean to get $\mathbf{E}(X_iX_j^3) = \mathbf{E}(X_i)\mathbf{E}(X_j^3) = 0$. Analogously, for $j \neq j \neq k$ it holds $\mathbf{E}(X_iX_jX_k^2) = 0$, $j \neq j \neq k \neq l$ $\mathbf{E}(X_iX_jX_kX_l) = 0$.

Note that this result implies that

$$\mathbf{E}(\sum_n \overline{X_n}^4) = \sum_{n\geqslant 1}\mathbf{E}(\overline{X_n})^4 < \infty$$

which in turn implies that the r.v. $\sum_n \overline{X_n}^4$ is finite with probability 1 and therefore the variable $\overline{X_n}$ converges to 0 with probability 1.

□

Exercise 12. *Prove that if we repeat independently n times the same random experiment and we consider the frequency of an event A, then as the number of times we repeat the experiment is sent to infinity, the frequency converges with probability 1 to the probability of tghe event.*

3 An application of LLN: Monte Carlo method to evaluate integrals

To be filled

4 An application of LLN: equipartition of probabilities

Let X be a random variable with density p_X and living on a finite space, $X : \Omega \to F$ with $|F| < \infty$. We define the entropy of X as

$$H_X := -\sum_{x \in F} p_X(x) \log p_X(x) = \mathbf{E}(-\log p_X)$$

where we set $xlog_x = 0$ for $x = 0$ and the equality follows from the property of expectation (see Proposition 1.3). Note that the entropy is the avarage of the level of information brought by an event: if an event is typical i.e. has high probability (resp. atypical) its occurrence brings little (rep. high) information. Indeed, $x \log x$ is an decreasing function of x for $x \in [0, 1]$. By using the LLN we will prove that the entropy "measures the disorder" of the law of X, in a way that will be quantified in Theorem 4.2.

Proposition 4.1. *The following holds*

- $0 \leqslant H_x \leqslant \log |F|$
- $H_X = 0$ *iff* X *is constant a.s.*
- $H_X = \log |F|$ *iff* X *is uniformly distributed on* F.

Proof. To be filled □

Theorem 4.2 (Asymptotic equipartition)**.** *Given a probability space* $(\Omega, \mathcal{F}, P)$*, a finite space* F *and a r.v.* $X : \Omega \to F$*, let* $(X_1, \dots, X_n)$ *be* n *i.i.d. r.v. with the same law as* X*. Then, for any* $\epsilon > 0$ *the following holds:*

(i) $\exists A_n \subset F^n$ *s.t. for any* $n \in \mathbb{N}$ *it holds* $|A_n| \leqslant \exp(n(H_X + \epsilon))$ *and s.t.*

$$\lim_{n \to \infty} P((X_1, \dots, X_n) \in A_n) = 1$$

(ii) for any $B_n \subset F^n$ *s.t.* $|B_n| \leqslant \exp(n(H_X - \epsilon))$ *it holds*

$$\lim_{n \to \infty} P((X_1, \dots, X_n) \in B_n) = 0$$

In other words, the set of typical sequences has roughly e^{nH_X} elements and a typical sequence has probability $p(x_1, \dots x_n) = e^{-nH_X}$. This result is at the base of *data compression* techniques in information theory. Indeed, if we want to transmit a signal corresponding to the sequence $X_1, \dots X_n$, thanks to Theorem 4.2 we can compress data and represent all typical sequences with $n \log_2 eH_X$ bits. Then we can assign to the remaining (atypical) sequences longer code words.

Proof. (i)Let

$$A_n = A_n^\epsilon := \left\{ x \in F^n : -\frac{1}{n} \sum_{i=1}^{n} \log p_X(x_i) \in [H_X - \epsilon, H_X + \epsilon] \right\}.$$

We shall prove that A_n satisfies both requirements...To be filled
(ii) Fix $B_n \subset F^n$ and let $\tilde{A}_n = A_n^{\epsilon/2}$. It holds

$$P(B_n) = P(B_n \cap \tilde{A}_n) + P(B_n \cap \tilde{A}_n^c) \tag{4.1}$$

Note that

$$P(B_n \cap \tilde{A}_n^c) \leqslant P(\tilde{A}_n^c)$$

which goes to zero as $n \to \infty$ thanks to the proof of point (i). We will now prove that also the first terms on the r.h.s. of 4.1 goes to zero as $n \to \infty$. Indeed it holds

$$P(B_n \cap \tilde{A}_n) = \sum_{(x_1,\dots x_n) \in B_n \cap \tilde{A}_n} p_X(x_1) \dots p_X(x_n) \leqslant \sum_{(x_1,\dots x_n) \in B_n \cap \tilde{A}_n} \exp(-n(H_X - \epsilon/2)) =$$

$$= |B_n \cap \tilde{A}_n| \exp(-n(H_X - \epsilon/2)) \leqslant \exp(n(H_X - \epsilon)) \exp(-n(H_X - \epsilon/2)) = \exp(-n/2\epsilon)$$

where for the last inequality we used $|B_n| \leqslant \exp(n(H_X - \epsilon))$ and $|B_n \cap \tilde{A}_n| \leqslant |B_n|$. □

Chapter 7

Central limit theorem

1 CLT: statement and sketch of the proof

Theorem 1.1. *Given $\{X_n\}_{n \geqslant 1}$ a sequence of independent and identically distributed (i.i.d.) r.v. that have mean μ, are square integrable, and have non zero variance σ^2, for each $I \subset \mathbb{R}$ it holds*

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \in I\right) \to \int_I g(x)dx$$

with

$$g(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}$$

and $S_n = \sum_{i=1}^n X_i$

Analogously, we could say that in the limit $n \to \infty$ the partition function of the variable $Z_n := \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$ with $\bar{X}_n = S_n/n$ converges to the partition function of standard gaussian variable.

Proof. To be filled □

The assumptions of the CLT can be weakened. In particular, the hypothesis of equidistribution does not play an essential role.
If the $\{X_i\}_{i \geqslant 1}$ are independent, of finite second moment and their tails satisfy the following condition (known as Lindeberg condition)

$$\lim_{n\to\infty} \frac{1}{v_n^2} \sum_{i=1}^n E((X_i - \mu_i)^2 \mathbb{1}_{|X_i - \mu_j| > \epsilon v_n}) = 0 \tag{1.1}$$

with $\mu_i = E(X_i)$, $v_n^2 := \sum_1^n Var(X_i)$, CLT continues to holds. Note that the above condition requires that the S_n is "very random", its variance is required to diverge as

$n \to \infty$.

2 An application: Gaussian law of errors

This "law" states that the measurements of a quantity which is subject only to accidental errors are distributed normally around the mean of the observations.

This can be understood by considering that the result of a measurement of a quantity can be represented via a random variable $M = q + E$ where q is a constant (the true value of the quantity) and E is a random variable corresponding to the sum of all the sources of errors in the measurement. It is natural to expect that $E = \sum_{i=1}^{n} e_i$ with n large (there are many independent causes of error) and e_i independent. Furthermore, if we have removed systematic errors, each source of error is a centred variable (it can take both positive and negative value with the same probability, otherwise it would be a systematic error) and it is natural to assume that all the sources of error is predominant so that the Lindeberg condition 1.1 holds. If the above assumptions hold, by CLT, we get $E/\sum_{i=1}^{n} \sigma(e_i)^2 \sim \mathcal{N}(0,1)$, namely the total error properly renormalised follows a standard Gaussian law.

Proposition 2.1. *Given a r.v. X with density p_X, for each $a, b \in \mathbb{R}$ with $a \neq 0$, the r.v. $Y = aX + b$ also has a density and it holds*

$$p_Y(y) = \frac{1}{|a|} p_X(\frac{y-b}{a})$$

Proof. Suppose $a > 0$. For $y \in \mathbb{R}$ oit holds

$$F_Y(y) = P(Y \leqslant y) = P(aX + b \leqslant y) = P\left(X \leqslant \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Thus

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) = \int_{-\infty}^{\frac{y-b}{a}} p_X(x)dx = \int_{-\infty}^{y} p_X\left(\frac{t-b}{a}\right)\frac{1}{a}dt$$

□

By using proposition 2.1 it follows that the measurement M are distributed according to a gaussian law of mean q and variance $\sigma^2 := \sum_{i=1}^{n} \sigma(e_i)^2$, namely the probability density of M is

$\sim \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-q)^2/2\sigma^2}$

3 Another application: confidence intervals

TO BE FILLED

Chapter 8

A rapid overview of two notable probabilistic models

1 Random walks: transience and recurrence

1.1 Random walks on $\mathbb{Z}$

Consider a sequence $X_1, X_2, \dots$ of independent discrete r.v. with $P(X_i = 1) = p$, $P(X_i = -1) = 1 - p$. The simple random walk starting from the ori gin is the sequence of variables $\{S_n\}_{n\in\mathbb{N}}$ defined as the partial sums, namely

$$S_n = \sum_{i=1}^{n} X_i, \ \forall n \geqslant 1, \ \ S_0 = 0$$

An interesting quantity is the first time τ at which the random walks comes back to the origin, defined as

$$\tau = \min\{n \geqslant 1 : S_n = 0\}, \ \ \tau \in \mathbb{N} \cup \{\infty\},$$

whose law τ can be completely determined. We say that a random walk is *recurrent* if $P(\tau < \infty) = 1$, *transient* otherwise. Recall the definition of generating function

$$G_X(z) = \mathbf{E}(z^X) = \sum_{x\in\mathbb{N}} z^x p_X(x), \ \ \forall z \in \mathbb{C}, \ |z| \leqslant 1$$

The following holds

Proposition 1.1. *The generating function of τ satisfies*

$$G_\tau(z) = 1 - \sqrt{1 - 4p(1-p)z^2}.$$

This implies that

- $P(\tau < \infty) = \lim_{z\to 1} G_\tau(z) = 1 - \sqrt{1 - 4p + 4p^2} = 1 - |2p - 1|$.
  In particular, the random walk is recurrent iff $p = 1/2$

- $\mathbf{E}(\tau) = \lim_{z\to 1} G'_\tau(z) = (1-4p(1-p))^{-1/2}$.
  Note that in the symmetric case, i.e. for $p = 1-p = 1/2$, it holds $\mathbf{E}(\tau) = \infty$

Proof. Let

$$Q(z) := \sum_{n=0}^{\infty} a_n z^n \quad \text{with} \quad a_n = P(S_n = 0).$$

Then notice that

$$a_n = \sum_{k=1}^{n} P(S_n = 0|\tau = k)P(\tau = k) = \sum_{k=1}^{n} a_{n-k}P(\tau = k)$$

where we used the fact that X_i being independent it holds $P(S_n = 0|\tau = k) = P(S_{n-k} = 0)$. Then multiplying by z^n and summing on n the previous equality we get

$$Q(z) = 1 + \sum_{n \geqslant 1}\sum_{k=1}^{n} a_{n-k}z^{n-k}P(\tau = k)z^k = 1 + Q(z)G_\tau(z),$$

thus

$$G_\tau(z) = 1 - \frac{1}{Q(z)}.$$

On the other hand using

$$P(S_n = 0) = \binom{n}{n/2} p^{n/2}(1-p)^{n/2} \ \ \forall n \text{ even}$$

and $P(S_n = 0) = 0$ for all n odd we get

$$Q(z) = \sum_{k \geqslant 0} \binom{2k}{k} (p(1-p)z^2)^k.$$

The above series can be summed yielding

$$Q(z) = (1-4p(1-p)z^2)^{-1/2}$$

which together with the above formula connecting Q to G_τ yields the desired result.

□

2 Branching process

We will discuss in this section the simplest branching model, the so called Galton Watson model. This processes describe the evolution of the number of individuals of a population along its subsequent generations. An historical note: the model was first introduced in the end of the 19th century, in order to study the probability of extinction of names of aristocratic families, hence the fact that it is a single individual (the male) that give birth to individual (i.e. transmits the name to its male children). The model has been later applied to several other fields including biology (evolution of bacterial colonies), epidemiology (spread of infections), chemistry (dynamics of chain reactions)..

We call Z_n the number of individuals in generation n and we suppose that

- in the first generation there is only one individual
- each individual gives independently birth to a certain number of children in the next generation
- the number of children of each individual has identical distribution

Namely we let

$$Z_0 = 1, \;\; Z_{n+1} = \sum_1^{Z_n} \xi_i \;\; \forall n \geqslant 0$$

with ξ_i i.i.d. and

$$\phi(z) = E(z^{\xi}).$$

By proposition 6.1 we have that, letting $G_n := G_{Z_n}$, it holds

$$G_{n+1} = G_n(\phi(z)) \quad \forall n \geqslant 0.$$

Noticing that $G_0(z) = z$ we get

$$G_n(z) = \phi_n(z) \;\; \forall n > 0 \quad \text{where} \;\; \phi(z) = \phi \cdot \phi ... \cdot \phi \text{ n times} .$$

We are interested in the probability that the population gets extincted, namely on the probability of the event

$$E = \{\exists n \in \mathbb{N} : Z_n = 0\} = \cup_n A_n \;\; \text{where} \;\; A_n := \{Z_n = 0\}.$$

Since $\{A_n\}_{n \geqslant 1}$ is a sequence increasing events (due to the fact that $Z_n = 0$ implies $Z_{n+1} = 0$), by proposition 3.5, it holds

$$E = \lim_{n \to \infty} A_n, \;\; P(E) = \lim_{n \to \infty} \phi_n(0)$$

where we used the fact that, since Z_n is an integer positive r.v., it holds $P(Z_n = 0) = \phi_n(0)$.

If we exclude the case $\phi(z) = z$, which corresponds to the (uninteresting) situation in which with probability one each individual gives birth to one child, the following holds:

Theorem 2.1.

(i) If $E(\xi) \leqslant 1$ then $P(E) = 1$, namely the population gets extinguished a.s.

(ii) If $E(\xi) > 1$ it holds $P(E) = \sigma$ with σ the unique solution of the equation $\phi(\sigma) = \sigma$ in $[0, 1)$.

Proof. Consider the function $\phi(z)$ in the interval $[0, 1]$. Notice that

$$\phi'(z) = E(\xi z^{\xi-1}) \quad \text{and} \quad \phi''(z) = E(\xi(\xi-1)z^{\xi-2})$$

which are both positive since they are the expectation of real discrete positive r.v. Furthermore it holds $\phi(1) = 1$. Thus, thanks to the fact that we excluded the case $\phi(z) = z$, only two situations may happen (see Fig.8.1)

- if $\phi'(1) \leqslant 1$ it holds $z \leqslant \phi(z)$ and tehrefore $\phi_n(z) \leqslant \phi_{n+1}(z)$, and

$$P(E) = \lim_{n\to\infty} \phi_n(0) = 1$$

- if $\phi'(1) > 1$ it holds $z \leqslant \phi(z) \leqslant \sigma$ for $z \leqslant \sigma$ and $z \geqslant \phi(z) \geqslant \sigma$ if $z \geqslant \sigma$. Furthermore it holds

$$P(E) = \lim_{n\to\infty} \phi_n(0) = \sigma$$

□

Remark 2.2. *Notice that (ii) is coherent with the fact that if the distribution of the number of children of an individual is such that the probability to have zero child is zero, the extinction event should have probability zero. Indeed in this case it holds $\phi(0) = 0$ and $\sigma = 0$*

The regime $E(\xi) < 1$ is called subrcritical, the regime $E(\xi) > 1$ supercritical and the case $E(\xi) = 1$ critical. The following proposition gives additional informations on the speed of extinction for the subcritical and critical case and on the explosion of the population size in the supercritical case when we condition on the non extinction event.

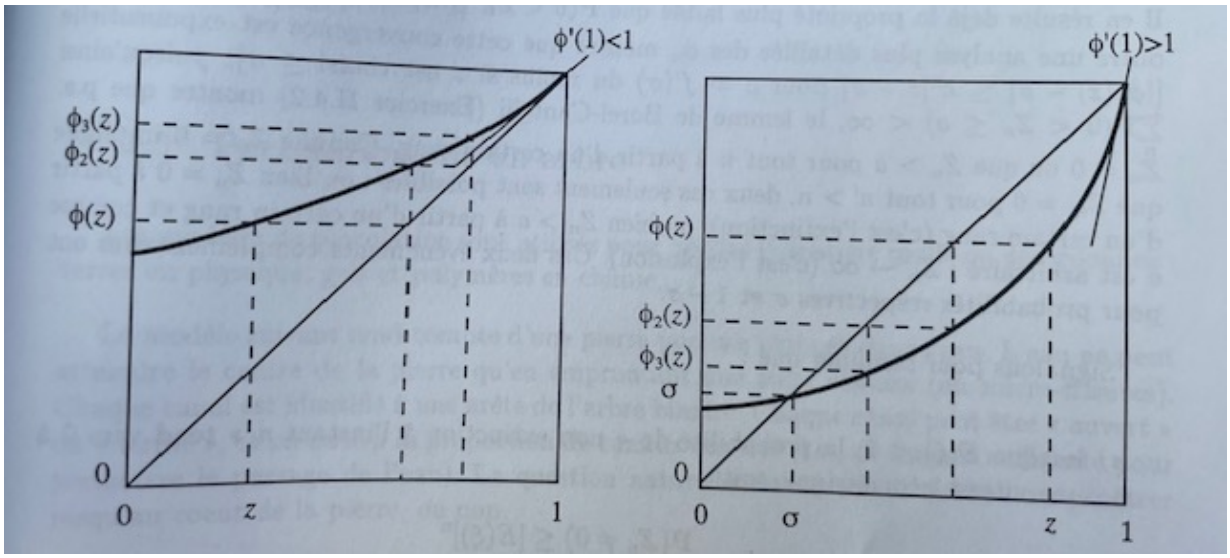

Figure 8.1: The function $\phi(z)$ in the case $E(\xi) \leqslant 1$ (left figure) and $E(\xi) > 1$. In both figures for a chosen z we also indicate the iterated values $\phi_2(z) = \phi(\phi(z))$, $\phi_3(z) = \phi(\phi(\phi(z)))$ which show the convergence to 1 in the case $E(\xi) \leqslant 1$ and to the point σ s.t. $\phi(\sigma) = \sigma$ when $E(\xi) > 1$

Proposition 2.3.

(i) For $E(\xi) < 1$ it holds

$$P(Z_n \neq 0) \leqslant E(\xi)^n$$

thus extinction occurs exponentially fast

(ii) For $E(\xi) = 1$, if $E(\xi^2) < \infty$ it holds for n large

$$P(Z_n \neq 0) \sim \frac{2}{E(\xi^2 - \xi)} \frac{1}{n}$$

(iii) For $E(\xi) > 1$

a) only two events may occur: either extinction or explosion (namely divergence of Z_n)

b) if $E(\xi^2) < \infty$ and we condition on non extinction it holds

$$\frac{Z_n}{E(\xi)^n} \to W \text{ a.s.}$$

with W a positive finite r.v.

We provide a proof only of point (i) and a proof of (iii)(a), the other cases need a longer proof and more refined tools.

Proof. In order to prove (i) we notice that

$$P(Z_n \neq 0) \leqslant E(Z_n) = E(E(Z_n|Z_{n-1})) = E(Z_{n-1}E(\xi)) = E(\xi)^n.$$

In order to prove (iii)-a) we notice that for any $a > 0$ integer and any $z \in (0,1)$ it holds

$$P(0 < Z_n \leqslant a)z^a \leqslant \sum_{k \geqslant 1} P(Z_n = k)z^k$$

where we used the fact that $z^x > z^a$ for $x \leqslant a$ and $z \in (0,1)$. Then we use

$$\sum_{k \geqslant 1} P(Z_n = k)z^k = \phi_n(z) - P(Z_n = 0) = \phi_n(z) - \phi_n(0).$$

This, together with the previous inequality, yields

$$P(A_{n,a}) \leqslant z^{-a}\left[\phi_n(z) - \phi_n(0)\right] \text{ with } A_{n,a} := \{0 < Z_n \leqslant a\}$$

By analysing the recursive equation for $\phi_n(z)$ it is then possible to show that $\phi_n(z) - \phi_n(0)$ converges to zero exponentially fast. Then, applying Borel-Cantelli lemma 4.1 we get that the event $\{A_{n,a}$ occurs i.o.$\}$ has probability zero. In other words, there exists N s.t. for all $n \geqslant N$, the event $A^c_{n,a}$ always occur. Noticing that $A^c_{n,a} = \{Z_n = 0\} \cup \{Z_n \geqslant a\}$ and that $Z_n = 0$ implies $Z_{n+1} = 0$, we deduce that either extinction occurs or $Z_n > a$ for all $n \geqslant N$. The latter case implies explosion thanks to the arbitrariness of a.

□

Exercise 13. *Fix $p \in (0,1)$ and let the number of children of an individual be distributed as $P(\xi = k) = (1-p)^k p$, $k \in \mathbb{N}$. Prove that the extinction probability equals 1 if $p \geqslant 1/2$ and it equals $p(1-p)^{-1}$ if $p < 1/2$.*

Chapter 9

Large deviations

The aim of the large deviation theory is to quantify the occurrence of macroscopic fluctuations in random systems. We will treat in detail only the easiest example, the large deviations for sums of i.i.d. random variables, and give some hints on more general results.

1 Large deviations for sums of i.i.d. random variables

Theorem 1.1 (Cramer's theorem). *Let $(X_n)_{n \geqslant 1}$ be a sequence of i.i.d. real random variables s.t.*

$$\phi(t) = E(e^{tX_1}) < \infty \ \ \forall t \in \mathbb{R}.$$

Let $S_n := \sum_{i=1}^n X_i$ and let I be the Fenchel-Legendre transformation of $\log \phi(t)$, namely

$$I(x) := \sup_{t \in \mathbb{R}} \{tx - \log \phi(t)\}. \tag{1.1}$$

Then

- *for any closed interval [1] $C \subset R$ it holds*

$$\limsup_{n \to \infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \in C\right) \leqslant -\inf_{x \in C} I(x) \tag{1.2}$$

- *for any open interval $A \subset R$ it holds*

$$\liminf_{n \to \infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \in A\right) \geqslant -\inf_{x \in A} I(x) \tag{1.3}$$

[1] namaly $C = [a, b]$ including the cases $a = -\infty$ or $b = \infty$

Corollary 1.2 (Bound on the tail on S_n)**.** *Let $a > E(X_1)$ then from Cramer's theorem we get*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(S_n \in [a,\infty)) \leqslant -\min_{x \geqslant a} I(x)$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log P(S_n \in [a,\infty)) \geqslant \liminf_{n\to\infty} \frac{1}{n} \log P(S_n \in (a,\infty)) \geqslant -\inf_{x>a} I(x).$$

The two above inequalities, together with the fact that $I(x)$ is non decreasing in the interval $[E(X_1),\infty)$ (see proposition 1.3), yield

$$P(S_n \geqslant na) \simeq e^{-nI(a)} \tag{1.4}$$

where the symbol $\simeq$ stands here for logarithmic equivalence [2].

Before proving Cramer's theorem let us discuss some properties of the function $I(x)$ defined in (1.1), which us called rate function.

Proposition 1.3. *The rate function I satisfies the following properties*

(i) I is a convex non negative function and $I(E(X_1)) = 0$

(ii) I is increasing on $[E(X_1),\infty)$ and decreasing on $(-\infty, E(X_1)]$

(iii) for $x \geqslant E(X_1)$ it holds

$$I(x) = \sup_{t \geqslant 0} (tx - \log \phi(t))$$

(iv) for $x \leqslant E(X_1)$ it holds

$$I(x) = \sup_{t \leqslant 0} (tx - \log \phi(t))$$

(v) given a function f we ley $\mathcal{D}_f := \{x : f(x) < \infty\}$ and let $\mathcal{D}_f^0$ be the interior of $\mathcal{D}_f$. Then I is strictly convex and infinitely differentiable in $\mathcal{D}_I^0$ and, for any $\bar{x} \in \mathcal{D}_I^0$ there is a unique $\bar{t} \in \mathcal{D}_{\log\phi(t)}^0$ s.t. $\bar{x}$ and $\bar{t}$ are in duality, namely s.t.

$$\bar{x} = \frac{d \log \phi(t)}{dt}|_{t=\bar{t}}$$

[2] Given two sequences a_n, b_n we say that they are logarithmically equivalent and denote this as $a_n \simeq b_n$ if

$$\lim_{n\to\infty} \frac{1}{n} (\log a_n - \log b_n) = 0.$$

and

$$\bar{t} = I'(\bar{x}).$$

Furthermore it holds

$$I(\bar{x}) = \bar{t}\bar{x} - \log \phi(\bar{t}).$$

Proof. (i)Convexity is established as follows. Let $\lambda \in (0,1)$ then

$$I(\lambda x + (1-\lambda)y) = \sup_{t\in\mathbb{R}}(t(\lambda x + (1-\lambda)y) - \log\phi(t)) =$$

$$\sup_{t\in\mathbb{R}}(\lambda(tx - \log\phi(t)) + (1-\lambda)(t(1-y) - \log\phi(t)) \leqslant \lambda I(x) + (1-\lambda)I(y).$$

Non negativity follows from the fact that $\phi(t) = 0$ thus $I(x) \geqslant 0x - \log\phi(0) = 0$.
The fact that $I(E(X)) = 0$ can be proved by Jensen inequality

$$\phi(t) = E(e^{tX_1}) \geqslant e^{tE(X_1)}$$

so that for all $t \in \mathbb{R}$ it holds

$$tE(X_1) - \log\phi(t) \leqslant tE(X_1) - tE(X_1) = 0.$$

(ii) This result follows by convexity and the fact that $I(E(X_1)) = 0$
(iii) For $x > E(X_1)$ and $t < 0$ it holds

$$tx - \log\phi(t) \leqslant t(x - E(X_1)) < 0$$

where in the first inequality we used $\phi(t) \geqslant e^{tE(X_1)}$. Thus necessarily since $I(x) \geqslant 0$ it should be

$$I(x) = \sup_{t\geqslant 0}(tx - \log\phi(t)) \quad \text{if } x > E(x_1)$$

(iv) Analogously for $x < E(X_1)$ and $t > 0$ we have

$$tx - \log\phi(t) \leqslant t(x - E(X_1)) < 0.$$

This implies

$$I(x) = \sup_{t\leqslant 0}(tx - \log\psi(t)) \quad \text{if } x < E(x_1).$$

(v) Follows from the properties of the Legendre transform □

We are now ready to prove Cramer's Theorem. We will divide the proof into two parts: proof of the upper bound (1.2) and proof of the lower bound (1.3)

Proof of inequality (1.2). Let the open set be $C = [a, b]$. If $E(X) \in C$ and $E(X) \neq a$, $E(X) \neq b$, the result immediately follows from the law of large numbers (and the fact that $I(E(X)) = 0$). Otherwise either (i) $a \geqslant E(X)$ or (ii) $b \leqslant E(X)$ and we treat separately the two cases.
(i)

$$P\left(\frac{S_n}{n} \in C\right) \leqslant P\left(\frac{S_n}{n} \geqslant a\right)$$

In turn for $t > 0$ and using Markov inequality it holds

$$P\left(\frac{S_n}{n} \geqslant a\right) = P(e^{tS_n} \geqslant e^{tna}) \leqslant e^{-tna} E(e^{tS_n}) = e^{-tna}\phi(t)^n.$$

Putting the two inequalities above together we get

$$\frac{1}{n}\log P(\frac{S_n}{n} \in C) \leqslant -\sup_{t>0}(ta - \log\phi(t)) = -I(a) = -\inf_{x\in C} I(x)$$

where for the second last equality we used Proposition 1.3 (iii) and for the last equality we used Proposition 1.3 (ii).
(ii) The proof follows along the same way:

$$P\left(\frac{S_n}{n} \in C\right) \leqslant P\left(\frac{S_n}{n} \leqslant b\right)$$

In turn for $t < 0$ and using Markov inequality it holds

$$P\left(\frac{S_n}{n} \leqslant b\right) = P(e^{tS_n} \geqslant e^{tnb}) \leqslant e^{-tnb} E(e^{tS_n)} = e^{-tnb}\phi(t)^n.$$

Putting the two inequalities above together we get

$$\frac{1}{n}\log P(\frac{S_n}{n} \in C) \leqslant -\sup_{t<0}(tb - \log\phi(t)) = -I(b) = -\inf_{x\in C} I(x)$$

where for the second last equality we used Proposition 1.3 (iv) and for the last equality we used Proposition 1.3 (ii).

□

Proof of inequality (1.3). Assume $\inf_{x\in A} I(x) < \infty$, otherwise the statement is trivial. Since A is open this implies $A \cap \mathcal{D}_I^0 \neq 0$, namely there exists $\bar{x} \in A \cap \mathcal{D}_I^0$. Let

$$B_{n,\epsilon} := \left((x_1, \ldots x_n) \in \mathbb{R}^n : \frac{S_n}{n} \in A; \left|\frac{S_n}{n} - \bar{x}\right| < \epsilon\}\right)$$

Since A is open, we can fix ϵ small enough so that $B_{n,\epsilon}$ is not empty and open. Let now $\bar{\lambda}$ be the dual of $\bar{x}$ (see Proposition 1.3) (v), it holds

$$P\left(S_n \in A\right) \geqslant \int_{B_{n,\epsilon}} \prod_{i=1}^{n} \mu(dx_i) = \int_{B_{n,\epsilon}} e^{-\bar{\lambda} S_n + n \log \phi(\bar{\lambda})} \prod_{i=1}^{n} \mu_{\bar{\lambda}}(dx_i) \tag{1.5}$$

where $\mu(dx)$ is the common probability measure of the i.i.d. random variables and for any λ we let

$$\mu_\lambda(dx) := e^{\lambda x - \log \phi(\lambda)} \mu(dx).$$

Assume $\bar{\lambda} > 0$ (the proof in the other case is analogous), then the r.h.s. of (1.5) is lower bounded by

$$e^{-\bar{\lambda} n(\bar{x}+\epsilon) + n \log \phi(\bar{\lambda})} \int_{B_{n,\epsilon}} \prod_{i=1}^{n} \mu_{\bar{\lambda}}(dx_i) = e^{-nI(\bar{x}+\epsilon)} \int_{B_{n,\epsilon}} \prod_{i=1}^{n} \mu_{\bar{\lambda}}(dx_i).$$

This, together with (1.5) yields

$$\frac{1}{n} \log P\left(S_n \in A\right) \geqslant -I(\bar{x}+\epsilon) + \frac{1}{n} \log \int_{B_{n,\epsilon}} \prod_{i=1}^{n} \mu_{\bar{\lambda}}(dx_i). \tag{1.6}$$

Now notice that $\mu_\lambda(dx)$ is a probability measure and

$$\int_{\mathbb{R}} x \mu_\lambda(dx) = \frac{dE(e^{\lambda x})}{d\lambda}.$$

Therefore, when $\lambda = \bar{\lambda}$, the above mean equals $\bar{x}$ thanks to Proposition 1.3(v). Then, the weak law of large number implies

$$\int_{B_{n,\epsilon}} \prod_{i=1}^{n} \mu_{\bar{\lambda}}(dx_i) \to 1$$

which, together with (1.6) implies

$$\frac{1}{n} \log P\left(S_n \in A\right) \geqslant -I(\bar{x}+\epsilon). \tag{1.7}$$

Since the l.h.s. of this equation does not depend on ϵ and I is continuous we can send $\epsilon \to 0$ and we get

$$\frac{1}{n} \log P\left(S_n \in A\right) \geqslant -I(\bar{x}). \tag{1.8}$$

Repeating the argument for any $\bar{x} \in A \cap \mathcal{D}_I^0$ inequality (1.3) is proved. □

Exercise 14. *Let $(X_n)_n \geqslant 1$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$. Prove that for all $a \in (0, 1]$ it holds*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{S_n}{n} \geqslant a\right) = \log 2 + \frac{1-a}{2} \log \frac{1-a}{2} + \frac{1+a}{2} \log \frac{1+a}{2}.$$

Exercise 15. *Let X be a gaussian variable of variance σ^2 and mean m, prove that the rate deviation function of a sum of i.i.d. random variables distributed as X is*

$$I(x) = \frac{(x-m)^2}{2\sigma^2}$$

2 Some generalities

Given a family $\{P_n\}_{n \geqslant 1}$ of probability distributions on Ω we say that the family satisfies a large deviation principle with good rate function $I(\cdot)$ it there exists $I : \Omega \to [0, \infty]$ s.t.

- I is lower semicontinuous, namely for all $x \geqslant 0$ the set $\{x \in \Omega : 0 \leqslant I(x) \leqslant c\}$ is closed
- For each $\ell < \infty$ the set $\{x : I(x) \leqslant \ell\}$ is compact in Ω
- For each closed set $C \subset \Omega$ it holds

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leqslant -\inf_{x\in C} I(x) \tag{2.1}$$

- for each open set $A \subset \Omega$ it holds

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n(A) \geqslant -\inf_{x\in A} I(x) \tag{2.2}$$

It is not difficult to verify that in the case P_n are the distributions of partial sums of i.i.d. variables (namely in the setting of the previous section), (1.1) is a good rate function.

Proposition 2.1. *If a sequence $(P_n)_{n \geqslant 1}$ satisfies a LDP, then the associated rate function is unique.*

Lemma 2.2 (Varadhan's lemma)**.** *Let $(P_n)_{n \geqslant 1}$ be a sequence of probability measures on Ω satisfying an LDP with rate function I. Let $F : \Omega \to \mathbb{R}$ a continuous function that is bounded from above. Then it holds*

$$\lim_{n\to\infty} \frac{1}{n} \log \int_\Omega e^{nF(x)} P_n(dx) = \sup_{x\in\Omega} \{F(x) - I(x)\}$$

Remark 2.3. *Large deviation theory has a natural application in many problems in statistical physics. The easiest example is the case of the Curie Weiss model for which the large deviation function can be explicitly determined (see e.g.* [?Velenik]*).*

Appendix A

Recalling some integration results

Definition 0.1 (Measurable functions)**.** *Given two measurable spaces* $(\Omega, \mathcal{F})$ *and* $(\Omega', \mathcal{F}')$ *and a function* $f : \Omega \to \Omega'$*, we say that* f *is measurable if* $f^{-1}(\mathcal{F}') \in \Omega$*, namely if*

$$\forall B \in \mathcal{F}', \ \{x \in \Omega, f(x) \in B\} \in \mathcal{F}.$$

For example, continuos functions from $\mathbb{R}^k \to \mathbb{R}^m$, are measurable if we endow these spaces with their Borel σ-algebras.

Proposition 0.2. *Any positive measurable function is the point-wise limit of an increasing sequence of positive step-functions*

On a measurable space $(\Omega, \mathcal{F})$ one can define a measure $\mu : \Omega \to \mathbb{R}^+ \cup \{\infty\}$ by requiring that for any sequence of pair-wise disjoint events it holds

$$\mu(\cup_n A_n) = \sum_n \mu(A_n).$$

Then the integral of a function w.r.t. the measure is defined as follows

Definition 0.3 (Integral w.r.t. a measure)**.** *Given a measure space* $(\Omega, \mathcal{F}, \mu)$ *and a measurable function* $f : \Omega \to F$ *we defines the integral w.rt.* μ *of* f *as follows*

- *if* f *is a positive step functions, i.e. of the form* $f = \sum_{k=0}^n \alpha_k \mathbb{1}_{A_k}$ *we set*

$$\int_\Omega f(x) d\mu(x) := sum_{k=0}^n \alpha_k \mu(A_k)$$

- *if* f *is positive, using Proposition 0.2 we set*

$$\int_\Omega f(x) d\mu(x) := \lim_{k \to \infty} \int_\Omega f_k(x) d\mu(x)$$

  which can be proven not to depend on the chosen sequence

- *for a generic l measurable f we set $f = f^+ - f^-$ with $a^+ := \max(a, 0)$ and $a^- := \min(-a, 0)$. Then if $\int_\Omega f^+(x)d\mu(x) < \infty$ and $\int_\Omega f^-(x)d\mu(x) < \infty$ we say that f is integrable and set*

$$\int_\Omega f(x)d\mu(x) = \int_\Omega f^+(x)d\mu(x) - \int_\Omega f^-(x)d\mu(x).$$

Note that, since $|f| = f^+f^-$, if f is integrable also $|f|$ is integrable.

Let us state some fundamental results of integration theory.

Theorem 0.4 (Monotone convergence)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions s.t $0 \leqslant f_n \leqslant f_{n+1}$ for any n. Let f be defined as the pointwise limit of f_n namely*

$$f(x) = \lim_{n\to\infty} f_n(x) \ \ \forall x \in \Omega.$$

Then

$$\int_\Omega f(x)d\mu(x) = \lim_{n\to\infty} \int_\Omega f_n(x)d\mu(x).$$

Lemma 0.5 (Fatou lemma)**.** *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions s.t $0 \leqslant f_n$ for any n. Let f be defined setting $f(x) = \liminf_n f_n(x)$ for all x. Then*

$$\int_\Omega f(x)d\mu(x) \leqslant \liminf_n \int_\Omega f_n(x)d\mu(x)$$

Theorem 0.6 (Dominated convergence)**.** *Let g be an integrable function on $(\Omega, \mathcal{F}, \mu)$, and $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions s.t $|f_n| \leqslant g$ for any n. If $\lim_{n\to\infty} f_n(x) = f(x)$ almost everywhere (i.e. except on a set of measure zero) then*

$$\int_\Omega f(x)d\mu(x) = \lim_{n\to\infty} \int_\Omega f_n(x)d\mu(x).$$

Theorem 0.7 (Fubini-Tonelli and Fubini theorems)**.** *Given two measure spaces $(\Omega, \mathcal{F}, \mu)$ and $(\Omega', \mathcal{F}', \mu')$ s.t. μ and ν are σ-finite [1]) and $f : \Omega \times \Omega \to \mathbb{R}^+$ that is $\Omega \times \Omega'$ measurable, then*

$$\int_{\Omega\times\Omega'} f(x, y)d(\mu \otimes \mu')(x, y) = \int_\Omega \left(\int_{\Omega'} d\mu'(y)\right) d\mu(x) = \int_{\Omega'} \left(\int_\Omega d\mu(x)\right) d\mu'(y)$$

Furthermore, f is integrable on $\Omega \times \Omega'$ iff the two above quantities are finite.
If there exists $G : \Omega' \to \mathbb{R}$ integrable on $(\Omega', \mathcal{F}'\mu')$ s.t. μ'-almost surely

$$\int_\Omega |f(x, y)|d\mu(x) \leqslant G(y)$$

then f is integrable on $\Omega \times \Omega'$.

[1]

Let us recall also the notion of absolute continuity of measures

Definition 0.8. *Let $\Omega, \mathcal{F}$ be a measurable space and μ, ν be two measures on this space. We say that μ is absolutely continuous w.r.t. ν (in formulas $\mu \ll \nu$) if for all $A \in \mathcal{F}$ s.t. $\nu(A) = 0$ on a $\mu(A) = 0$. We say that μ and ν are equivalent (in formulas $\mu \equiv \nu$) if $\mu \ll \nu$ and $\nu \ll \mu$.*

The following result guarantees that absolute continuity implies the existence of a density.

Theorem 0.9 (Radon-Nikodym derivative)**.** *Let $\Omega, \mathcal{F}$ be a measurable space and μ, ν be two measures on this space s.t. $\mu(E) < \infty$, ν is σ-finite and $\mu \ll \nu$. Then there exists a unique $h \in L^1(\nu)$ s.t. for all $A \in \mathcal{F}$ it holds $\mu(A) = \int_A h d\nu$. We call h the Radon-Nikodym derivative of μ w.r.t. ν. In formulas*

$$d\mu = h d\nu \quad \text{or} \quad h = \frac{d\mu}{d\nu}.$$