Paolo Dai Pra

Stochastic Methods

Lecture Notes

Notations and preliminaries

Given a set Ω and two subsets $A, B \subseteq \Omega$, we use the standard notations for union, intersection and the other operations with sets:

$$\begin{aligned}
A \cup B &:= \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}, \\
A \cap B &:= \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}, \\
A^c &:= \{\omega \in \Omega : \omega \notin A\}, \\
A \setminus B &:= A \cap B^c, \\
A \triangle B &:= (A \setminus B) \cup (B \setminus A) \;=\; (A \cup B) \setminus (A \cap B),
\end{aligned}$$

where the symbol ":=" indicates a definition. Union and intersection can be extended to arbitrary families $\{A_i\}_{i \in I}$ of subsets of Ω:

$$\begin{aligned}
\bigcup_{i \in I} A_i &:= \{\omega \in \Omega : \exists i \in I \text{ such that } \omega \in A_i\}, \\
\bigcap_{i \in I} A_i &:= \{\omega \in \Omega : \forall i \in I \text{ it holds } \omega \in A_i\}.
\end{aligned}$$

We recall the De Morgan's Laws:

$$(A \cup B)^c \;=\; A^c \cap B^c, \qquad (A \cap B)^c \;=\; A^c \cup B^c,$$

and, more generally,

$$\left(\bigcup_{i \in I} A_i\right)^c \;=\; \bigcap_{i \in I} A_i^c, \qquad \left(\bigcap_{i \in I} A_i\right)^c \;=\; \bigcup_{i \in I} A_i^c.$$

We denote by $\mathbb{N} := \{1, 2, 3, \ldots\}$ the set of natural numbers, not including zero; to include zero we adopt the notation $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$. We also use the standard notations for the sets of integers, rationals, real and complex, indicated by $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ e $\mathbb{C}$ respectively, and we set $\mathbb{R}^+ := [0, \infty) = \{x \in \mathbb{R} : x \geq 0\}$, $\mathbb{Q}^+ := \mathbb{Q} \cap [0, \infty)$.

We say that a number $x \in \mathbb{R}$ is *positive* if $x \geq 0$ and *strictly positive* if $x > 0$; similarly, x is *negative* if $x \leq 0$ and *strictly negative* if $x < 0$. The positive and the negative part of a number $x \in \mathbb{R}$ are respectively defined by $x^+ := \max\{x, 0\}$ and $x^- := -\min\{x, 0\} = \max\{-x, 0\}$. Note that $x^+, x^- \geq 0$, $x = x^+ - x^-$, $|x| = x^+ + x^-$.

We use the words "increasing" and "decreasing" in weak sense: a funciton $f : \mathbb{R} \to \mathbb{R}$ is increasing (resp. decreasing) if for all $x > y$ we have $f(x) \geq f(y)$ (resp. $f(x) \leq f(y)$).

The cardinality of a set A, i.e. the number of its elements, will be denoted by $|A|$; thus $|A| < \infty$ indicates that A is a finite set. A set A is countable if there exists a one to one and onto function $f : A \to \mathbb{N}$. Given two sets A, B, the cartesian product $A \times B$ is the set of all pairs (a, b) with $a \in A$ and $b \in B$.

Suppose E is a set, and $A \subseteq E$. The *indicator function* of A, $\mathbf{1}_A : E \to \mathbb{R}$ is defined by

$$\mathbf{1}_A(x) := \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise.} \end{cases}$$

Infinite sums

Given a sequence of real numbers $(x_n)_{n\in\mathbb{N}}$, the *sum of the series* $\sum_{n=1}^{\infty} x_n$ is defined as the limit as $N \to +\infty$ $N \to \infty$ of the partial sums $s_N := \sum_{n=1}^{N} x_n$, assuming such limit exists. When dealing with a family of real numbers $\{x_i\}_{i\in I}$ indexed by an arbitrary set I, to define the *infinite sum* $\sum_{i\in I} x_i$ we proceed as follows.

In the case all summands are positive ($x_i \geq 0$ per ogni $i \in I$) we set

$$\sum_{i\in I} x_i := \sup_{A\subseteq I,\, |A|<\infty} \sum_{j\in A} x_j \in [0, +\infty],$$

where $\sum_{j\in A} x_j$ is just a finite sum. Clearly $\sum_{i\in I} x_i \in [0, +\infty]$ and $\sum_{i\in I} x_i = 0$ if and only if $x_i = 0\ \forall i \in I$. If $\sum_{i\in I} x_i < +\infty$, then we say that the family $\{x_i\}_{i\in I}$ is *summable*: it can be shown that there are at most countably many strictly positive summands ($x_i > 0$) (indeed $|\{i \in I : x_i > \frac{1}{n}\}| < \infty$ for every $n \in \mathbb{N}$).

For a generic family $\{x_i\}_{i\in I}$ (whose summand are not necessarily positive) we say it is *sommabile* if $\sum_{i\in I} |x_i| < \infty$, or, equivalently, if $\sum_{i\in I} x_i^+ < \infty$ e $\sum_{i\in I} x_i^- < \infty$. In this case we set

$$\sum_{i\in I} x_i := \sum_{i\in I} x_i^+ - \sum_{i\in I} x_i^-, \tag{0.1}$$

and we have $\sum_{i\in I} x_i \in (-\infty, +\infty)$. More generally, the infinite sum $\sum_{i\in I} x_i$ can be defined by (0.1) provided at least one of the sums $\sum_{i\in I} x_i^+$, $\sum_{i\in I} x_i^-$ is finite, and in this case $\sum_{i\in I} x_i \in [-\infty, +\infty]$.

We also recall some properties of infinite sums. If $\{x_i\}_{i\in I}$ e $\{y_i\}_{i\in I}$ are summable families, then the family $\{x_i + y_i\}_{i\in I}$ is also summable, and

$$\sum_{i\in I} (x_i + y_i) = \sum_{i\in I} x_i + \sum_{i\in I} y_i.$$

Consider now a family $\{x_{i,j}\}_{(i,j)\in I\times J}$ indexed by the elements of a cartesian products. If it has positive summands ($x_{i,j} \geq 0$ per ogni $i \in I,\ j \in J$), or if it is summable, the following version of Fubini's Theorem holds:

$$\sum_{(i,j)\in I\times J} x_{i,j} \;=\; \sum_{i\in I}\left(\sum_{j\in J} x_{i,j}\right) \;=\; \sum_{j\in J}\left(\sum_{i\in I} x_{i,j}\right). \tag{0.2}$$

Chapter 1
Basic probability theory

In this chapter we briefly review the basic notions of probability needed in the course.

1.1 Probabilities and their basic properties

The aim of probability theory is to make mathematical models for *random experiments*, i.e. those phenomena whose outcome cannot be deterministically predicted. To formulate a probabilistic model the following three ingredients are required:

(i) A set, that we denote by Ω and call the *sample space*, which contains all outcomes of the random experiment.

(ii) A family $\mathscr{A}$ of subsets of Ω, whose elements are called *events*, which satisfies the following properties:

- $\Omega \in \mathscr{A}$;
- if $A \in \mathscr{A}$ then $A^c \in \mathscr{A}$;
- if $(A_n)_{n\in\mathbb{N}}$ is a sequence of events, i.e. $A_n \in \mathscr{A}$ for all n, then

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{A}.$$

In other words the family of events $\mathscr{A}$ contains the sample space, and it is closed for complementation and *countable* unions; this is often expressed by saying that $\mathscr{A}$ is a σ-*algebra*. Using De Morgan's laws it is checked that $\mathscr{A}$ is also closed for countable intersection. Since Ω and $\emptyset$ are events, $\mathscr{A}$ is also close for *finite* unions and intersections. In general $\mathscr{A}$ needs not to be closed for *uncountable* unions or intersections.

(iii) A function

$$\mathrm{P} : \mathscr{A} \to [0,1],$$

called *probability*, satisfying the following properties:

- $\mathrm{P}(\Omega) = 1$;
- For every sequence $(A_n)_{n\in\mathbb{N}}$ of events that are *pairwise disjoint*, i.e. $A_n \cap A_m = \emptyset$ for $n \neq m$, the following identity holds:

$$\mathrm{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathrm{P}(A_n).$$

This property is called σ-*additivity* of the probability.

The triple $(\Omega, \mathscr{A}, \mathrm{P})$ will be called *probability space*.

Remark 1.1. σ-additivity applied to the sequence $A_n = \emptyset$ for all n, immediately implies the obvious property $\mathrm{P}(\emptyset) = 0$. We leave to the reader to check that this last fact allows to show that *finite additivity* also holds, i.e. for a finite family $\{A_1, \ldots, A_n\}$ of pairwise disjoint events,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i).$$

It is good to know, but rather delicate to show, that requiring only finite additivity for a probability is not enough to guarantee that σ-additivity also holds. □

Remark 1.2. For most of the purposes of these notes the precise identification of the σ-algebra of the events $\mathscr{A}$ is not crucial, and one could "assume" that *all* subset of Ω are events, i.e. $\mathscr{A} = \mathscr{P}(\Omega) = \{A : A \subseteq \Omega\}$. Although for Ω finite or countable this can be *really* assumed, there are many relevant example of Ω uncountable and "natural" probabilities which *cannot* be defined on all subsets. For example, if we choose $\Omega = [0,1]$, then we can define P on intervals by setting

$$\mathrm{P}([a,b]) := b - a,$$

for $0 \leq a \leq b \leq 1$. Note that the family of intervals is *not* a σ-algebra. It can be shown that P can be extended to a σ-algebra containing all intervals, but not to all subsets of $\mathbb{R}$. We will see more details later.

□

Example 1.3. The most basic example of probability is the case of *equally likely outcomes*: suppose Ω is a finite set, $\mathscr{A} = \mathscr{P}(\Omega)$, and set

$$\mathrm{P}(A) := \frac{|A|}{|\Omega|}.$$

This is also called the *uniform* probability on Ω. Note that, for each $\omega \in \Omega$, $\mathrm{P}(\{\omega\}) = \frac{1}{|\Omega|}$ does not depend on ω, so all outcomes are equally likely. □

Example 1.4. Let Ω be a generic set, $p : \Omega \to [0, +\infty)$ be such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Then, for every $A \subseteq \Omega$ we can define

$$\mathrm{P}(A) = \sum_{\omega \in A} p(\omega). \tag{1.1}$$

In particular $p(\omega) = \mathrm{P}(\{\omega\})$. It can be shown (details are omitted) that P is a probability on $\mathscr{A} = \mathscr{P}(\Omega)$. We recall that

$$N = \{\omega : p(\omega) > 0\}$$

is at most countable, and clearly $\mathrm{P}(N) = 1$. Thus P is "concentrated" in a *discrete* (i.e. finite or countable) set. The converse is also true. If a probability P on $\mathscr{P}(\Omega)$ is such that $\mathrm{P}(N) = 1$ for a discrete N, then P is of the form (1.1), with

$$p(\omega) = \mathrm{P}(\{\omega\}).$$

□

Next proposition summarizes the basic properties of probability; proofs are simple and left to the reader.

Proposition 1.5. (i) *If $A \subseteq B$ are two events, then*

$$\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A). \tag{1.2}$$

In particular $\mathrm{P}(A) \leq \mathrm{P}(B)$ *and* $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

(ii) *For any two events A, B,*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{1.3}$$

By a nontrivial induction argument (omitted), property (1.3) can be extended to finite unions.

Proposition 1.6. *Consider n events $A_1, A_2, \ldots, A_n$ di Ω. Then*

$$\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} \sum_{\substack{J \subseteq \{1,2,\ldots,n\} \\ \text{such that } |J|=k}} (-1)^{k+1}\, \mathrm{P}\left(\bigcap_{i \in J} A_i\right). \tag{1.4}$$

Propositions 1.5 and 1.6 are concerned with properties of probability related to *finitely many* operations with sets. Some properties concerning *countably many* operations are also useful.

Proposition 1.7. (i) (Lower continuity). *Let $(A_n)_{n\in N}$ be an increasing sequence of events, i.e. $A_n \subseteq A_{n+1}$ for all n. Then*

$$\mathrm{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \lim_{n\to+\infty} \mathrm{P}(A_n). \tag{1.5}$$

(ii) (Upper continuity). *Let $(A_n)_{n\in N}$ be an decreasing sequence of events, i.e. $A_n \supseteq A_{n+1}$ for all n. Then*

$$\mathrm{P}\left(\bigcap_{n\in\mathbb{N}} A_n\right) = \lim_{n\to+\infty} \mathrm{P}(A_n). \tag{1.6}$$

(iii) (Subadditivity). *Let $(A_n)_{n\in N}$ be an arbitrary sequence of events. Then*

$$\mathrm{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \le \sum_{n\in\mathbb{N}} \mathrm{P}(A_n).$$

Proof. First note that lower and upper continuity are equivalent: indeed (1.5) holds fo (A_n) if and only if (1.6) holds for (A_n^c). Thus we only prove upper continuity. Set $B_1 = A_1$, and $B_n := A_n \setminus A_{n-1}$ for $n \ge 2$. It is easily seen that the B_n's are pairwise disjoint, and $A_n = B_1 \cup \cdots \cup B_n$ for each n, which also implies

$$\bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n.$$

Thus, using σ-additivity,

$$\begin{aligned}\mathrm{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) &= \mathrm{P}\left(\bigcup_{n\in\mathbb{N}} B_n\right) = \sum_{n\in\mathbb{N}} \mathrm{P}(B_n)\\ &= \lim_{n\to+\infty}\sum_{k=1}^{n} P(B_k) = \lim_{n\to+\infty} \mathrm{P}(A_n),\end{aligned}$$

where in the last step we have used (finite) additivity to get $\mathrm{P}(A_n) = \sum_{k=1}^n P(B_k)$. Let us now prove subadditivity. By (1.2) and a simple induction argument one shows that, for each n,

$$\mathrm{P}\left(\bigcup_{k=1}^{n} A_k\right) \le \sum_{k=1}^{n} \mathrm{P}(A_k). \tag{1.7}$$

Using lower continuity for the sequence $B_n := \bigcup_{k=1}^n A_k$, the conclusion follows taking the limit as $n \to +\infty$ in (1.7).

□

1.2 Conditional probability and independence

Let A and B be two events, and assume $P(B) > 0$.

Definition 1.8. The *conditional* probability of A *given* B is denoted by $\mathrm{P}(A|B)$ and defined by

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

It is easily verified that, if we fix the event B, the map $A \mapsto \mathrm{P}(A|B)$ is indeed a probability on $\mathscr{A}$. The basic properties of conditional probability are summarized in the following proposition.

Proposition 1.9. (i) *Let $(B_n)_{n=1}^N$ be a finite ($N < +\infty$) or countable family of events, forming a* partition *of Ω, i.e. they are pairwise disjoint and $\bigcup_n B_n = \Omega$. Then for any other event A*

$$\mathrm{P}(A) = \sum_n \mathrm{P}(A|B_n)\,\mathrm{P}(B_n). \tag{1.8}$$

This property is called the Formula of total probability

(ii) *Let A and B be events of nonzero probability. Then*

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

which is called the Bayes' rule *or formula.*

Proof. For the Formula of total probability just observe that, by additivity,

$$\mathrm{P}(A) = \sum_n \mathrm{P}(A \cap B_n)$$

and use the definition of conditional probability.
The Bayes'rule comes from the obvious identity

$$\mathrm{P}(B|A)\,\mathrm{P}(A) = \mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(A \cap B).$$

□

The conditional probability $\mathrm{P}(A|B)$, compared with the *unconditional* probability $\mathrm{P}(A)$, expresses how the *information* concerning the occurrence of the event B modifies our belief that A will occur. In there case the is no modification, i.e. $\mathrm{P}(A|B) = \mathrm{P}(A)$, we say A is independent of B. This seemingly asymmetric notion of independence is actually symmetric, and it is more conveniently given as follows.

Definition 1.10. Two events A and B are said to be *independent* if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B).$$

Note that, whenever $\mathrm{P}(B) > 0$, this is actually equivalent to $\mathrm{P}(A|B) = \mathrm{P}(A)$.

It is useful to extend the notion of independence to general families of events.

Definition 1.11. Let $(A_i)_{i \in I}$ be a family of events, indexed by an arbitrary set I. We say they are independent if for every $J \subseteq I$, with $|J| < +\infty$, we have

$$\mathrm{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathrm{P}(A_i).$$

Note that, for example, for the three events A, B, C to be independent, one needs

$$\begin{aligned}
\mathrm{P}(A \cap B \cap C) &= \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C)\\
\mathrm{P}(A \cap B) &= \mathrm{P}(A)\,\mathrm{P}(B)\\
\mathrm{P}(A \cap C) &= \mathrm{P}(A)\,\mathrm{P}(C)\\
\mathrm{P}(B \cap C) &= \mathrm{P}(B)\,\mathrm{P}(C).
\end{aligned}$$

No one of the above identities are logical consequences of the remaining three.

The notion of independence turns out to be invariant with respect to complementation, as next proposition states.

Proposition 1.12. *Let $(A_i)_{i \in I}$ be a family of independent events. Suppose $(B_i)_{i \in I}$ is obtained from $(A_i)_{i \in I}$ by complementing "some" for the A_i, i.e., for every $i \in I$ either $B_i = A_i$ or $B_i = A_i^c$. Then $(B_i)_{i \in I}$ is a family of independent events.*

Proof. Let $J \subseteq I$ be finite. Assume for the moment there is $j \in J$ such that $B_j = A_j^c$, while $B_i = A_i$ for $i \in J \setminus \{j\}$. Letting $C := \bigcap_{i \in J \setminus \{j\}} A_i$ we have

$$\begin{aligned}
\mathrm{P}\left(\bigcap_{i \in J} B_i\right) &= \mathrm{P}(A_j^c \cap C) = \mathrm{P}(C) - \mathrm{P}(A_j \cap C) =\\
&= \mathrm{P}\left(\bigcap_{i \in J \setminus \{j\}} A_i\right) - \mathrm{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J \setminus \{j\}} \mathrm{P}(A_i) - \prod_{i \in J} \mathrm{P}(A_i)\\
&= (1 - \mathrm{P}(A_j)) \prod_{i \in J \setminus \{j\}} \mathrm{P}(A_i) = \prod_{i \in J} \mathrm{P}(B_i),
\end{aligned}$$

where we have used repeatedly the independence of the $(A_i)_{i\in I}$. When $|\{i\in J : B_i = A_i^c\}| > 1$ we just iterate the above procedure, and obtain in general that $\mathrm{P}\left(\bigcap_{i\in J} B_i\right) = \prod_{i\in J}\mathrm{P}(B_i)$. □

In probability events with probability zero or one are of special interest. In particular, if $\mathrm{P}(A) = 1$ we say that A occurs *almost surely* (in short *a.s.*). We give next an example of how such events may emerge.

Consider a sequence $(A_n)_{n\in\mathbb{N}}$ of events, and a given sample $\omega \in \Omega$. We say that A_n occurs *infinitely often* (i.o) in ω if $\omega \in A_n$ for infinitely many values of n. More formally, define

$$\{A_n \text{ i.o.}\} := \{o \in \Omega : \omega \in A_n \text{ for infinitely many } n\text{'s}\}.$$

It is not hard to realize that

$$\{A_n \text{ i.o.}\} = \bigcap_{n\in\mathbb{N}}\bigcup_{k\geq n} A_k.$$

Proposition 1.13. *(*Borel-Cantelli Lemma*). Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events.*

(i) *If*

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

then

$$\mathrm{P}(A_n \textit{ i.o.}) = 0.$$

(ii) *Assume $(A_n)_{n\in\mathbb{N}}$ is a sequence of* independent *events. If*

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

then

$$\mathrm{P}(A_n \textit{ i.o.}) = 1.$$

Proof. (i) Setting $B_n := \bigcap_{n\in\mathbb{N}}\bigcup_{k\geq n} A_k$, (B_n) is a decreasing sequence of events, so using Proposition 1.7

$$\mathrm{P}(A_n \text{ i.o.}) = \lim_{n\to+\infty}\mathrm{P}(B_n) \leq \lim_{n\to+\infty}\sum_{k\geq n}\mathrm{P}(A_k) = 0,$$

where we have used the fact the a series converges if and only if its tail converges to zero.

(ii) Note that

$$\{A_n \text{ i.o.}\}^c = \bigcup_{n\in\mathbb{N}}\bigcap_{k\geq n} A_k^c.$$

Thus, by subadditivity, it is enough to show that for every $n \in \mathbb{N}$

$$\mathrm{P}\left(\bigcap_{k\geq n} A_k^c\right) = 0. \tag{1.9}$$

By Proposition 1.7 and independence

$$\mathrm{P}(\bigcap_{k\geq n} A_k^c) = \lim_{N\to+\infty} \mathrm{P}\left(\bigcap_{k=n}^{N} A_k^c\right) = \lim_{N\to+\infty} \prod_{k=n}^{N} [1-\mathrm{P}(A_k)].$$

Since the inequality $1-x \leq e^{-x}$ holds for all positive x, we obtain

$$\mathrm{P}(\bigcap_{k\geq n} A_k^c) \leq \lim_{N\to+\infty} \prod_{k=n}^{N} e^{-\mathrm{P}(A_k)} = \exp\left[-\lim_{N\to+\infty} \sum_{k=n}^{N} \mathrm{P}(A_k)\right] = 0,$$

since $\lim_{N\to+\infty}\sum_{k=n}^{N}\mathrm{P}(A_k) = \sum_{k=n}^{+\infty}\mathrm{P}(A_k) = +\infty$ by assumption.

□

Example 1.14. Consider an infinite sequence of independent trials. Denote by A_n the event of success in the n-th trial. These are independent events, by assumption. Suppose first the probability of success is the same, say $p \in (0,1)$ in each trial, i.e. $\mathrm{P}(A_n) \equiv p$. Then $\sum_n \mathrm{P}(A_n) = +\infty$ and $\sum_n \mathrm{P}(A_n^c) = +\infty$, i.e. both win and loss occur infinitely many times.

Assume instead the player *learn* in repeating the game: the probability of success in the n-th trial is $1-a_n$, where (a_n) is a decreasing sequence of numbers in $(0,1)$.

- If $\sum_n a_n = +\infty$, i.e. $\sum_n \mathrm{P}(A_n^c) = +\infty$, then there will be infinitely many loss.
- If, however, $\sum_n a_n < +\infty$, then the player learn “fast enough” so that he/she will keep winning after finitely many loss.

□

1.3 Random variables: generalities

Let $(\Omega, \mathscr{A}, \mathrm{P})$ be a probability space, describing a random experiment. Roughly speaking, for a given set E, an E-valued random variable is a function $X : \Omega \to E$. To any such function it is natural to associate suitable subsets of Ω: given $A \subseteq E$, we denote by $\{X \in A\}$ or $X^{-1}(A)$ the set of outcomes for which X takes values in A, more specifically

$$\{X \in A\} := \{\omega \in \Omega : X(\omega) \in A\}.$$

As we would like these sets to be events, a little care is needed to give a consistent definition of random variable.

Suppose the set E is provided with a σ-algebra of subsets, that we denote by $\mathscr{E}$

Definition 1.15. An E-valued random variable X is a function $X : \Omega \to E$ such that for each $A \in \mathscr{E}$ we have $\{X \in A\} \in \mathscr{A}$.

In mathematics this notion is called *measurability* of the function X. In the case $E = \mathbb{R}$ we canonically choose $\mathscr{E}$ to be the minimal σ-algebra which contains intervals, the so-called *Borel σ-algebra*, that we denote by $\mathscr{B}(\mathbb{R})$. In this case one can show that $X : \Omega \to \mathbb{R}$ is a random variable if and only if $\{X \in I\} \in \mathscr{A}$ for every interval I. In this notes we will never worry about measurability, that will be given as granted.

By Definition 1.15, if X is an E-valued random variable then for every $A \in \mathscr{E}$ we can consider the probability of the event $\{X \in A\}$, that we write simply as $\mathrm{P}(X \in A)$.

Proposition 1.16. *Let X be an E-valued random variable and, for $A \in \mathscr{E}$, set $\mu_X(A) := \mathrm{P}(X \in A)$. Then μ_X is a probability on the σ-algebra $\mathscr{E}$, which is called* law *or* distribution *of the random variable X.*

The proof of Proposition 1.16 is simple and left to the reader. The law μ_X contains the probabilistic properties of the random variable X; random variables will be indeed classified in terms of their law.

A particularly relevant role is played by *real valued* (or simply *real*) random variable: this is the case in which $E = \mathbb{R}$, and $\mathscr{E} = \mathscr{B}(\mathbb{R})$. In this case the law of a real random variable can be characterized in terms of simpler objects. Indeed, a probability μ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is uniquely identified once we know the probability $\mu(I)$ for every interval I. On the other hand,

$$\mu((a,b]) = \mu((-\infty,b] \setminus (-\infty,a]) = \mu((-\infty,b]) - \mu((-\infty,a]),$$

$$\mu([a,b]) = \lim_{n\to+\infty} \mu\left(\left(a - \frac{1}{n}, b\right]\right),$$

(here lower continuity in Proposition 1.6 is used), and similarly

$$\mu((a,b)) = \lim_{n\to+\infty} \mu\left(\left(a, b - \frac{1}{n}\right]\right) \quad \mu([a,b) = \lim_{n\to+\infty} \mu\left(\left[a, b - \frac{1}{n}\right]\right).$$

Thus μ is uniquely determined if we know $\mu((-\infty,x])$ *for each* $x \in \mathbb{R}$. This explains the following definition and result.

Definition 1.17. Let X be a real random variable. The *distribution function* of X is defined by

$$F_X(x) = \mathrm{P}(X \le x) = \mu_X((-\infty,x]).$$

Proposition 1.18. *The distribution function F_X of a real random variable X uniquely characterizes its law. This amounts to say that if X and Y are real random variables with $F_X = F_Y$, then $\mu_X = \mu_Y$ or equivalently*

$$\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$$

for each $A \in \mathscr{B}(\mathbb{R})$.

We collect in next statement some useful properties of the distribution function.

Proposition 1.19. *For the distribution function F_X of a real random variable X the following properties hold:*

(i) *F_X is increasing: $x < y$ implies $F_X(x) \leq F_X(y)$.*

(ii)

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \lim_{x \to +\infty} F_X(x) = 1.$$

(iii) *F_X is right continuous, i.e. for every $x \in \mathbb{R}$*

$$F_X(x^+) := \lim_{y \downarrow x} F_X(y) = F_X(x).$$

(iv) *For every $x \in \mathbb{R}$, defining $F_X(x^-) = \lim_{y \uparrow x} F_X(y)$, we have*

$$F_X(x) - F_X(x^-) = \mathrm{P}(X = x).$$

In other words the points in which F_X is discontinuous are exactly those values that are attained by X with positive probability.

It is very important to notice that when we consider the distribution of a random variable X we often do not care on which sample space Ω this random variable is defined. For instance, in the statement of Proposition 1.18, the random variables X and Y may be defined on *different* sample spaces. There are situations in which it is natural, and relevant, to consider several random variables defined of the *same* sample space. Before introducing some abstract related notion, we give a relevant example.

Example 1.20. We model here a countably infinite sequence of independent trials, each trial having probability of success $p \in (0,1)$. As sample space we take the set Ω of binary sequences $\omega = (\omega_i)_{i=1}^{+\infty} = (\omega_1, \omega_2, \ldots)$, with $\omega_i \in \{0,1\}$: $\omega_i = 1$ (resp. $\omega_i = 0$) indicates that the i-th trial is a success (resp. a failure). Providing Ω with a σ-algebra and a suitable probability P is highly nontrivial, and not so relevant for our purposes. It will be enough to know that P is uniquely identified by the following statement: let $x = (x_1, x_2, \ldots, x_n)$ be a binary sequence of length n, containing k ones and $n - k$ zeroes; then

$$\mathrm{P}(\{\omega : \omega_1 = x_1,\, \omega_2 = x_2,\, \omega_n = x_n\}) = p^k(1-p)^{n-k}.$$

In this probability space several interesting random variables are defined.

(i) For any $n \geq 1$ define $X_n(\omega) := \omega_i$, giving the outcome of the n-th trial.

(ii)
$$Y(\omega) := \min\{n : \omega_n = 1\}$$
which gives the index of the trial in which the first success is obtained.

(iii)
$$S_n(\omega) := \omega_1 + \omega_2 + \cdots + \omega_n = \sum_{k=1}^{n} X_n(\omega)$$
i.e. the number of successes in the first n trials.

(iv) Let a, b be two positive constant; suppose we win a Euros for each success, and lose b Euros for each loss. Then
$$C_n(\omega) = aS_n(\omega) - b(n - S_n(\omega))$$
is the amount won after n trials.

(v) Suppose we play the game in (iv), starting with a sum of s Euros. The random variable
$$R(\omega) := \min\{n : C_n(\omega) \leq -s\}$$
is the time, i.e. the index of the trial, at which we are *ruined*.

□

Consider now n real random variables $X_1, X_2, \ldots, X_n$, defined on the *same* probability space $(\Omega, \mathscr{A}, \mathrm{P})$. We say that $X := (X_1, X_2, \ldots, X_n)$ is a *random vector*: in other words a random vector is a vector whose components are real random variables. Note that, for each $\omega \in \Omega$, $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \in \mathbb{R}^n$, so that X can be viewed as a function from Ω to $\mathbb{R}^n$. In order to properly view it as a *random variable* with values in $\mathbb{R}^n$ we need to provide $\mathbb{R}^n$ with a suitable σ-algebra. This is done as follows. We denote by $\mathscr{B}(\mathbb{R}^n)$ the minimal σ-algebra containing all subsets of $\mathbb{R}^n$ of the form $A_1 \times A_2 \cdots \times A_n$, with $A_1, A_2, \ldots A_n \in \mathscr{B}(\mathbb{R})$. The elements of $\mathscr{B}(\mathbb{R}^n)$ are called *Borel* sets in $\mathbb{R}^n$. With this choice it can be shown that a function $X : (\Omega, \mathscr{A}) \to (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ is a random variable if and only if its components X_i, $i = 1, 2, \ldots, n$, are real random variables.
We can therefore consider the distribution μ_X of a random vector X, which is a probability on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. We sometimes will refer to it as the *joint distribution* of the random variables $X_1, X_2, \ldots, X_n$. If the distribution μ_X of the random vector X is known, then the distribution μ_{X_i} of each component X_i (in this context called *marginal distributions*) is uniquely determined: for instance, if A is a Borel set in $\mathbb{R}$,

$$\begin{aligned}\mu_{X_1}(A) &= \mathrm{P}(X_1 \in A) = \mathrm{P}(X_1 \in A, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}) = \mathrm{P}(X \in A \times \mathbb{R}^{n-1}) \\ &= \mu_X(A \times \mathbb{R}^{n-1}).\end{aligned} \tag{1.10}$$

This is obviously extended to the other components. With a similar argument, from the joint distribution of a random vector one can derive the joint distribution of any *sub-vector*.

In general, the knowledge of the marginal distributions of a random vector does not allow to reconstruct the joint distribution. Indeed, informations on "dependence" among different components cannot be carried by the marginal distributions. An extreme case is that in which there is no dependence at all.

Definition 1.21. Let $X_1, X_2, \dots, X_n$ be real random variables defined on the same probability space $(\Omega, \mathscr{A}, \mathrm{P})$. We say they are independent if for every choice of Borel sets $A_1, A_2, \dots, A_n$,

$$\mathrm{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \mathrm{P}(X_1 \in A_1) \cdot \mathrm{P}(X_2 \in A_2) \cdots \mathrm{P}(X_n \in A_n). \tag{1.11}$$

More generally, the elements of a possibly infinite family of real random variables $(X_i)_{i \in I}$ are said to be independent if the random variables of *any* finite subgroup $(x_i)_{i \in J}$, $|J| < +\infty$, are independent.

Note that if we set $X := (X_1, X_2, \dots, X_n)$, then (1.11) amounts to

$$\mu_X(A_1 \times A_2 \cdots \times A_n) = \prod_{i=1}^{n} \mu_{X_i}(A_i),$$

which only depends on marginal distributions. Since it can be shown that a probability on $\mathbb{R}^n$ in uniquely identified by its values of sets of the form $A_1 \times A_2 \cdots \times A_n$, we conclude that the joint distribution of *independent* random variables is uniquely identified by their marginal distributions.

Remark 1.22. For the sake of simplicity we have defined independence only for real random variables. It is only a notational effort to extent this notion to arbitrary random variables, possibly taking values in different sets, just replacing Borel sets with the elements of the appropriated σ-algebra. Details are left to the reader. □

1.4 Discrete random variables and their expectation

Let X be a random variable taking values on a set E. We say that X is *discrete* if its "possible values" form a set which is finite or countable. More rigorously, if there is $N \subseteq E$ finite or countable (simply *discrete*) such that[1]

$$\mathrm{P}(X \in N) = 1$$

[1] here we are tacitly assuming N to belong to the σ-algebra $\mathscr{E}$ of subsets of E. This will be granted in all cases of interest (e.g. $E = \mathbb{R}^n$) in which "singletons" $\{x\}, x \in E$ are in $\mathscr{E}$, which implies all finite and countable sets are in $\mathscr{E}$.

For discrete random variables, the distribution can be described in terms of a simpler object, the (discrete) *density*.

Definition 1.23. Let X be a discrete, E-valued random variable. Its *density* $p_X : E \to [0,1]$ is defined as follows:

$$p_X(x) := \mathrm{P}(X = x) = \mu_X(\{x\}). \tag{1.12}$$

Note that, if N is a discrete set such that $\mathrm{P}(X \in N) = 1$, then $p_X(x) = 0$ unless $x \in N$. Moreover, for any $A \in \mathscr{E}$,

$$\mathrm{P}(X \in A) = \mathrm{P}(X \in A \cap N) = \sum_{x \in A \cap N} p_X(x) = \sum_{x \in A} p_X(x), \tag{1.13}$$

which shows how to determine the distribution knowing the density. Note that this implies, in particular,

$$\sum_{x \in E} p_X(x) = \mathrm{P}(X \in E) = \mathrm{P}(\Omega) = 1. \tag{1.14}$$

Remark 1.24. It is essential to understand that, although the discrete density in (1.12) could in principle be defined for *any* random variable X, (1.13) strictly depends on the fact X is discrete; more than that: it is equivalent to it. Indeed, in the second equality of (1.13) we have used σ-additivity: the event $\{X \in A \cap N\}$ is the finite or countable union of the disjoint events $\{X = x\}$, $x \in A \cap N$; if X were not discrete, the first equality in (1.13) could not hold, and we could not reduce to finite or countable unions.

Now consider a generic random variable X, and define p_X as in (1.12). Set

$$D := \{x \in E : p_X(x) > 0\}.$$

Note first that D is finite or countable: in fact

$$D = \bigcup_{n \geq 1} \left\{ x \in E : p_X(x) \geq \frac{1}{n} \right\}.$$

Note that the set $D_n := \left\{x \in E : p_X(x) \geq \frac{1}{n}\right\}$ has at most n elements (otherwise we would obtain the absurd statement $\mathrm{P}(X \in A_n) > 1$); so D is the countable union of finite sets, which implies it is at most countable. Moreover, repeating the argument in (1.13)

$$\sum_{x \in E} p_X(x) = \sum_{x \in D} p_X(x) = \mathrm{P}(X \in D) \leq 1.$$

Thus we see that if (1.13), and so (1.14) holds, then $\mathrm{P}(X \in D) = 1$ which means that X is discrete. □

Remark 1.25. In the remaining part of these notes we will often exhibit a "density" p, i.e. a function $p : \mathbb{R} \to [0,1]$ such that $\sum_x p(x) = 1$, and say "let X be a random variable with density X". A skeptical reader may wander whether such random variable exists. A "canonical" way to obtain *one* discrete random variable with density p is the following. Set $(\Omega, \mathscr{A}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, and define a probability P by

$$\mathrm{P}(A) = \sum_{x \in A} p(x);$$

the actual verification that P is a probability is somewhat tedious, and we omit it. In particular, $\mathrm{P}(\{x\}) = p(x)$. Then define the function $X : \Omega \to \mathbb{R}$ by $X(x) = x$. We leave to the reader the simple task of verifying that X is indeed a random variable with density p. □

Consider now a discrete random vector $X = (X_1, X_2, \dots, X_n)$. Its density

$$p_X(x_1, x_2, \dots, x_n),$$

that sometimes we write more extensively as

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n),$$

is called the *joint density* of the real random variables $X_1, X_2, \dots, X_n$, while each density p_{X_i} of the real random variables X_i is called *marginal density*. Specializing to this context (1.10) we obtain (the simple details are omitted) the following statement.

Proposition 1.26. *Let $X = (X_1, X_2, \dots, X_n)$ be a discrete random vector, with joint density $p_X(x_1, x_2, \dots, x_n)$. Then the marginal density p_{X_i} is given by*

$$p_{X_i}(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} p_X(x_1, x_2, \dots, x_n).$$

Independence of discrete random variables can also be characterized in terms of the joint density.

Theorem 1.27. *Let $X = (X_1, X_2, \dots, X_n)$ be a discrete random vector, with joint density p_X. The components $X_1, X_2, \dots, X_n$ are independent if and only if for every $x_1, x_2, \dots, x_n \in \mathbb{R}$*

$$p_X(x_1, x_2, \dots, x_n) = p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n), \tag{1.15}$$

i.e. if and only if the joint density is the product of the marginals.

Proof. Assume $X_1, X_2, \dots, X_n$ are independent, i.e. (1.11) holds. Specializing (1.11) to $A_1 = \{x_1\}, \dots, A_n = \{x_n\}$, we obtain (1.15).

Conversely, assume (1.15) holds. Then, using (1.13), for every choice of Borel sets $A_1, A_2, \ldots, A_n$,

$$\begin{aligned} \mathrm{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) &= \mathrm{P}(X \in A_1 \times A_2 \times \cdots \times A_n) \\ &= \sum_{(x_1,x_2,\ldots,x_n) \in A_1,A_2,\ldots,A_n} p_X(x_1, x_2, \ldots, x_n) \\ &= \sum_{(x_1,x_2,\ldots,x_n) \in A_1,A_2,\ldots,A_n} p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n) \\ &= \sum_{x_1 \in A_1} p_{X_1}(x_1) \cdots \sum_{x_n \in A_n} p_{X_n}(x_n) \\ &= \mathrm{P}(X_1 \in A_1) \cdot \mathrm{P}(X_2 \in A_2) \cdots \mathrm{P}(X_n \in A_n), \end{aligned}$$

where, in the fourth line, we have used the n-dimensional version of (0.2). This completes the proof. □

One of the key notions concerning *real* random variables is that of *expectation*, or *expected value*. We begin by introducing it for discrete random variables.

Definition 1.28. Let X be a real, discrete random variable, with density p_X. Suppose the infinite sum

$$\mathrm{E}(X) := \sum_{x \in \mathbb{R}} x p_X(x) \in [-\infty, +\infty]$$

is well defined. Then $\mathrm{E}(X)$ is called *expectation* (or *expected value*, or *mean value*) of X. In the case $\mathrm{E}(X)$ is finite, which is equivalent to $\sum_x |x| p_X(x) < +\infty$, we say X admits finite expectation, or that it is *integrable*.

Remark 1.29. A constant $a \in \mathbb{R}$ can be identified with a random variable taking the value a with probability one. Clearly, $\mathrm{E}(a) = a$. □

Remark 1.30. Besides constants, the most basic level of random variables are those taking *two values*. Assume we identify these two variables with the reals 0 and 1, consider, for a $\{0,1\}$-valued random variable X, the event

$$A := \{X = 1\} \in \mathscr{A}.$$

Then $X = \mathbf{1}_A$. Since $p_X(1) = \mathrm{P}(A)$, we easily get

$$\mathrm{E}(X) = \mathrm{P}(A) = \mathrm{P}(X = 1).$$

□

It follows from Definition (1.28) that if X is nonnegative, in the sense that $\mathrm{P}(X \geq 0) = 1$, then its expectation is always defined: indeed nonzero summands of the sum $\sum_x x p_X(x)$ are necessarily positive, and a sum whose summands are all nonnegative is always well defined.

To obtain the main properties of expectation, it is very useful to deal with *functions* of a given random variable. Let X be a random variables with values in $(E,\mathscr{E})$, and $f : E \to \mathbb{R}$ be a function such that for every Borel set A,

$$f^{-1}(A) = \{x \in E : f(x) \in A\} \in \mathscr{E}\}. \tag{1.16}$$

Under this last condition it is easily verified that the function $f(X) : \Omega \to \mathbb{R}$ given by $f(X)(\omega) = f(X(\omega))$ is indeed a real random variable.

Proposition 1.31. *Let X be a discrete, E-valued random variable, $f : E \to \mathbb{R}$ be a function such that* (1.16) *holds for every Borel set A. Assume at least one of the following conditions holds:*

$$f(x) \geq 0 \ \textit{ for every } x \in E; \tag{1.17}$$

$$\sum_{x\in E} |f(x)| p_X(x) < +\infty. \tag{1.18}$$

Then the expectation $\mathrm{E}[f(X)]$ *is well defined and*

$$\mathrm{E}[f(X)] = \sum_{x\in E} f(x) p_X(x).$$

Proof. Let $N \subseteq E$ be finite or countable, and such that $\mathrm{P}(X \in N) = 1$, and set $f(N) := \{f(x) : x \in N\}$. Then $f(N)$ is finite or countable, and $\mathrm{P}(f(X) \in f(N)) = 1$. This shows that $f(X)$ is a discrete random variable. Its density is given by

$$p_{f(X)}(u) = \mathrm{P}(f(X) = u) = \mathrm{P}(X \in E_u) = \sum_{x\in E_u} p_X(x),$$

where $E_u := \{x \in E : f(x) = u\}$. Suppose, first, that (1.17) holds. Then, using (0.2) and the fact that $x \in E_u$ if and only if $f(x) = u$,

$$\begin{aligned}
\mathrm{E}[f(X)] &= \sum_{u\geq 0} u p_{f(X)}(u) = \sum_{u\geq 0} u \sum_{x\in E_u} p_X(x) = \sum_{u\geq 0} u \sum_{x\in E} \mathbf{1}_{x\in E_u} p_X(x) \\
&= \sum_{(x,u)\in E\times\mathbb{R}} u\mathbf{1}_{x\in E_u} p_X(x) = \sum_{x\in E}\sum_{u\in\mathbb{R}} u\mathbf{1}_{x\in E_u} p_X(x) \\
&= \sum_{x\in E}\sum_{u\in\mathbb{R}} \mathbf{1}_{x\in E_u} f(x) p_X(x) = \sum_{x\in E} f(x) p_X(x) \sum_{u\in\mathbb{R}} \mathbf{1}_{x\in E_u} = \sum_{x\in E} f(x) p_X(x),
\end{aligned}$$

where the last equality relies of the obvious fact that $\sum_{u\in\mathbb{R}} \mathbf{1}_{x\in E_u} = 1$ for each $x \in E$. This completes the proof for nonnegative f. Otherwise we write $f(x) = f^+(x) - f^-(x)$, use the argument above for $f^{\pm}$ separately, observe that under (1.18), both

$$\sum_{x\in E} f^+(x) p_X(x) = \sum_{u\geq 0} u p_{f(X)}(u)$$

and

$$\sum_{x\in E} f^-(x)p_X(x) = -\sum_{u<0} up_{f(X)}(u)$$

are finite, and

$$\begin{aligned}
\mathrm{E}[f(X)] &= \sum_{u\geq 0} up_{f(X)}(u) = \sum_{u\geq 0} up_{f(X)}(u) + \sum_{u<0} up_{f(X)}(u) \\
&= \sum_{x\in E} f^+(x)p_X(x) - \sum_{x\in E} f^-(x)p_X(x) = \sum_{x\in E} [f^+(x) - f^-(x)]p_X(x) \\
&= \sum_{x\in E_u} f(x)p_X(x).
\end{aligned}$$

□

Using Proposition 1.31 with $f(x) = |x|$, we obtain the following useful characterization of integrability.

Corollary 1.32. *A discrete, real random variable X is integrable if and only if* $\mathrm{E}(|X|) < +\infty$.

A further important consequence of Proposition 1.31 is the following.

Theorem 1.33. *Let X and Y be discrete random variables, defined on the same probability space, which are either both nonnegative of both integrable. Then*

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Proof. The pair (X,Y) is a random vector; denote by $p_{X,Y}$ its joint density. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = x+y$. Thus $X+Y = f(X,Y)$. Using Proposition 1.31, (0.2) and the fact that, under the given conditions all sums below are well defined, we obtain:

$$\begin{aligned}
\mathrm{E}[X+Y] = \mathrm{E}[f(X,Y)] &= \sum_{(x,y)\in\mathbb{R}^2} f(x,y)p_{X,Y}(x,y) = \sum_{(x,y)\in\mathbb{R}^2} (x+y)p_{X,Y}(x,y) \\
&= \sum_{(x,y)\in\mathbb{R}^2} xp_{X,Y}(x,y) + \sum_{(x,y)\in\mathbb{R}^2} yp_{X,Y}(x,y) \\
&= \sum_{x\in\mathbb{R}} x \sum_{y\in\mathbb{R}} p_{X,Y}(x,y) + \sum_{y\in\mathbb{R}} y \sum_{x\in\mathbb{R}} p_{X,Y}(x,y) \\
&= \sum_{x\in\mathbb{R}} xp_X(x) + \sum_{y\in\mathbb{R}} yp_Y(y),
\end{aligned}$$

where we have used the fact that, thanks to Proposition 1.26

$$\sum_{y\in\mathbb{R}} p_{X,Y}(x,y) = p_X(x) \qquad \sum_{x\in\mathbb{R}} p_{X,Y} = p_Y(y).$$

□

Using Theorem 1.33 and basic properties of infinite sums, we summarize below the fundamental properties of expectation. Details are omitted.

Theorem 1.34. *Let X and Y be discrete random variables, and a,b be real numbers. Then the following properties hold:*

(a) *(*Monotonicity*) If $X(\omega) \leq Y(\omega)$ for every $\omega \in \Omega$ and the expectation is well defined both X and Y, then $\mathrm{E}(X) \leq \mathrm{E}(Y)$. Note that since $-|X| \leq X \leq |X|$ we have $-\mathrm{E}(|X|) \leq \mathrm{E}(X) \leq \mathrm{E}(|X|)$, i.e.*

$$|\mathrm{E}(X)| \leq \mathrm{E}(|X|). \tag{1.19}$$

(b) *(*Linearity*) If X and Y are integrable (or if they are nonnegative and $a,b \geq 0$), then*

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

We collect in next result two essentially trivial but useful facts about expectation. We leave the proof to the reader

Proposition 1.35. *(a) If X and Y are discrete, real random variable such that $\mathrm{P}(X = Y) = 1$, then $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ are either both well defined or both ill defined, and in the first case*

$$\mathrm{E}(X) = \mathrm{E}(Y).$$

(b) If X is nonnegative and $\mathrm{E}(X) = 0$, then $\mathrm{P}(X = 0) = 1$.

We conclude this section observing how independence impacts of computation of expectations.

Proposition 1.36. *Let X and Y be real, independent and integrable (resp. nonnegative) discrete random variables. Then their product XY is integrable (resp. nonnegative) and*

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

Proof. Using Proposition 1.31 with $f(x,y) = xy$, Theorem 1.27 and (0.2), we have

$$\mathrm{E}(XY) = \sum_{(x,y)} xy p_{X,Y}(x,y) = \sum_{(x,y)} xy p_X(x) p_Y(y) = \sum_x x p_X(x) \sum_y y p_Y(y) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

□

1.5 Square integrable, discrete random variables

A discrete random variable X such that $E(X^2) < +\infty$, i.e.

$$\sum_{x\in\mathbb{R}} x^2 p_X(x) < +\infty,$$

is said to be *square integrable*. We begin by observing that *a square integrable random variable is necessarily integrable*: indeed, the elementary inequality $|x| \leq 1+x^2$, monotonicity and linearity of expectation give

$$E(|X|) \leq 1 + \mathrm{E}(X^2) < +\infty,$$

so, by Corollary 1.32, X is integrable. The converse is not necessarily true. For example, consider a random variable X, taking values on integers greater or equal to 1, with density

$$p_X(n) = \frac{1}{Zn^3},$$

where $Z := \sum_{n\geq 1} \frac{1}{n^3} < +\infty$. Then

$$E(|X|) = E(X) = \sum_{n\geq 1} n p_X(n) = \sum_{n\geq 1} \frac{1}{Zn^2} < +\infty,$$

but

$$E(X^2) = \sum_{n\geq 1} n^2 p_X(n) = \sum_{n\geq 1} \frac{1}{Zn} = +\infty.$$

This shows in particular that the product of two integrable random variables is not necessarily an integrable random variable (although we have seen this is true under the additional assumption that the random variables are independent). However, the product of two square integrable random variables is integrable.

Proposition 1.37. *Let X and Y discrete, square-integrable random variables, defined on the same probability space. Then their product XY is integrable.*

Proof. Observe that, for every $x, y \in \mathbb{R}$

$$|xy| \leq \frac{1}{2}(x^2 + y^2),$$

as the difference is $\frac{1}{2}(|x| - |y|)^2 \geq 0$. Thus, by monotonicity and linearity of the expectation

$$E(|XY|) \leq \frac{1}{2}[E(X^2) + E(Y^2)] < +\infty.$$

□

Consider now a square-integrable random variable X. The expectation $E\left[(X-a)^2\right]$, for a given constant a, can be seen as an index of how much the distribution of X is "concentrated around a". Note that the map $a \mapsto E\left[(X-a)^2\right]$ takes its maximum at $a = E(X)$:

$$\begin{aligned} E\left[(X-a)^2\right] &= E\left[(X-E(X)+E(X)-a)^2\right] \\ &= E\left[(X-E(X))^2\right] + (E(X)-a)^2 + 2[E(X)-a]E\left[(X-E(X)\right] \\ &= E\left[(X-E(X))^2\right] + (E(X)-a)^2 \geq E\left[(X-E(X))^2\right]. \end{aligned}$$

Thus, the quantity

$$\mathrm{Var}(X) := E\left[(X-E(X))^2\right] = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2,$$

called the *variance* of X, is the basic index of concentration of the distribution of a random variable around its mean. Note that, by Proposition 1.35, $\mathrm{Var}(X) = 0$ if and only if $\mathrm{P}(X = E(X)) = 1$, which is equivalent to the fact that X takes a constant value with probability one.

Suppose X and Y are square integrable, discrete random variables. Linearity of expectation yields

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))].$$

The quantity

$$\mathrm{Cov}(X,Y) := \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] \tag{1.20}$$

is called *covariance* of X and Y, so that we can write

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$

Two random variables such that $\mathrm{Cov}(X,Y) = 0$ are said to be *uncorrelated*.

Proposition 1.38. *If X and Y are discrete, square integrable and independent, then* $\mathrm{Cov}(X,Y) = 0$*, and in particular*

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

The covariance has the following useful properties.

Proposition 1.39. *Let X,Y,Z be discrete, square integrable random variables. Then*

(a)

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X);$$

(b) the covariance is symmetric*, i.e.*

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X);$$

(c) the covariance is linear in each component (or bilinear*), i.e. for each reals a,b*

$$\mathrm{Cov}(aX+bY,Z) = a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z).$$

1.6 Discrete distributions: examples

In this section we collect the examples of discrete random variables that most often appear in applications. Random variables will be classified in terms of their distribution.

1.6.1 Bernoulli and binomial distributions

Consider a sequence of independent trials, each with probability of success $p \in [0,1]$. If we fix $n \geq 1$ and set $X =$ number of successes in the first n trials, we find

$$p_X(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $k = 0,1,\ldots,n$. Any random variable with this distribution is called *Binomial*, and we write

$$X \sim \mathrm{Bin}(n,p).$$

A special case is when $n = 1$, where X takes only the values 0 and 1. In this case X is called *Bernoulli*, and we write

$$X \sim \mathrm{Be}(p).$$

In this last case it is easily seen that

$$\mathrm{E}(X) = p \quad \mathrm{Var}(X) = p(1-p).$$

If, more generally, $X \sim \mathrm{Bin}(n,p)$ is the number of successes in the first n trials, then we can write

$$X = X_1 + X_2 + \cdots + X_n,$$

where $X_i = 1$ if the i-th trial is a success, and $X_i = 0$ otherwise. Since trials are independent, the random variables $X_1, X_2, \ldots, X_n$ are also independent. Thus, by additivity of expectation and Proposition 1.38, we get

$$\mathrm{E}(X) = np \quad \mathrm{Var}(X) = np(1-p).$$

1.6.2 Geometric distributions

In a sequence of independent trials, each with probability of success $p \in (0,1]$, let X denote the index of the *first success*: in other words, the event $\{X = n\}$, $n \geq 1$, means that the first $n-1$ trials are lost, but we win the n-th trial; this event has probability $p(1-p)^{n-1}$. Thus

$$p_X(n) = p(1-p)^{n-1}$$

for $n \geq 1$. Any random variable with this distribution is called *Geometric*, and we write

$$X \sim \mathrm{Geo}(p).$$

To compute mean and variance of X we observe that

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{n\geq 1} np(1-p)^{n-1} = \sum_{n\geq 1} (n-1)p(1-p)^{n-1} + \sum_{n\geq 1} p(1-p)^{n-1} \\
&= \sum_{m\geq 0} mp(1-p)^{m} + 1 = (1-p) \sum_{m\geq 0} mp(1-p)^{m-1} + 1 \\
&= (1-p)\,\mathrm{E}(X) + 1,
\end{aligned}$$

which, solving for $\mathrm{E}(X)$, gives

$$\mathrm{E}(X) = \frac{1}{p}.$$

With a similar trick

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{n\geq 1} n^2 p(1-p)^{n-1} = \sum_{n\geq 1} (n-1+1)^2 p(1-p)^{n-1} \\
&= \sum_{n\geq 1} (n-1)^2 p(1-p)^{n-1} + \sum_{n\geq 1} p(1-p)^{n-1} + 2\sum_{n\geq 1} (n-1)p(1-p)^{n-1} \\
&= (1-p) \sum_{m\geq 1} m^2 p(1-p)^{m-1} + 1 + 2\,\mathrm{E}(X-1) = (1-p)\,\mathrm{E}(X^2) + \frac{2}{p} - 1,
\end{aligned}$$

which yields

$$\mathrm{E}(X^2) = \frac{2}{p^2} - \frac{1}{p}$$

and

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [E(X)]^2 = \frac{1-p}{p^2}.$$

1.6.3 Negative binomial distributions

In a sequence of independent trials, each with probability of success $p \in (0,1]$, let X denote the index of the *r-th success*, with $r \geq 1$. For $r = 1$ we get a geometric random variable. In general, for $n \geq r$, the event $\{X = n\}$ means that:

- the n-th trial is a success;
- in the first $n-1$ trials, exactly $r-1$ successes have been obtained.

Thus

$$p_X(n) = p\binom{n-1}{r-1}p^{r-1}(1-p)^{n-r} = \binom{n-1}{r-1}p^r(1-p)^{n-r}$$

for $n \geq r$. Any random variable with this distribution is called *Negative Binomial*, and we write

$$X \sim \mathrm{NegBin}(r, p).$$

Although we could compute mean and variance from the density, we proceed differently. Note that is X is the index of the *r-th success* in a sequence of independent trials with probability of success $p \in (0,1]$, we can write

$$X = X_1 + X_2 + \ldots + X_r,$$

where X_1 is the number of trials to the first success, X_2 is the number of trials after the first success and to the second success, and so on. It is clear that $X_i \sim \mathrm{Geo}(p)$. It should be intuitive, but it is not so easy to prove (proof is omitted) that all X_i are $\mathrm{Geo}(p)$ and are all independent. This immediately gives

$$\mathrm{E}(X) = \frac{r}{p} \quad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

1.6.4 Poisson distributions

In many applications, e.g counting the number of access to a webpage, or the number of phone calls to a call center, binomial random variables with large n and small p emerge naturally: large number of customers each having a small probability of accessing the service. In these contexts it is convenient to replace the Binomial distribution with a distribution obtained from binomials via a suitable limit. The simplest way to derive such distribution is to assume the parameter p of the binomial scales as n^{-1}, i.e. $p = \frac{\lambda}{n}$ for some $\lambda > 0$. So let $X_n \sim \mathrm{Bin}(n, \lambda/n)$. For any $k \geq 0$

$$\begin{aligned} p_{X_n}(k) &= \frac{n!}{k!(n-k)!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}\frac{\lambda^k}{k!} \end{aligned}$$

as $n \to +\infty$. This motivates to introduce the distribution on $\mathbb{N}_0 := \{0, 1, \ldots\}$ with density $p(n) = e^{-\lambda}\frac{\lambda^n}{n!}$. Any random variable X with this density is called *Poisson* random variable, and we write

$$X \sim \mathrm{Pois}(\lambda).$$

The connection between Poisson random variables and Binomials, or more generally sums of independent Bernoulli random variables, can be made more precise. We in particular state, without proof, the following Theorem, sometimes called *Law of Small Numbers*.

Theorem 1.40. *Let $X_1, X_2, \dots, X_n$ be independent random variables, such that $X_i \sim \mathrm{Be}(p_i)$. Set $S_n := X_1 + X_2 + \cdots + X_n$, and let $W_n \sim \mathrm{Pois}(p_1 + p_2 + \ldots + p_n)$. Then, for every $A \subseteq \mathbb{N}_0$*

$$|\mathrm{P}(S_n \in A) - \mathrm{P}(W_n \in A)| \leq \sum_{k=1}^{n} p_i^2.$$

Thus, the Poisson distribution well approximates the distribution of a sum of independent Bernoulli random variables as soon as $\sum_{k=1}^{n} p_i^2$ is "small". For instance, if $p_i = \frac{\lambda}{n}$ for every i, then $\sum_{k=1}^{n} p_i^2 = \frac{\lambda^2}{n}$, which is actually small for large n.

Mean and variance of $X \sim \mathrm{Pois}(\lambda)$ are easily computed:

$$\mathrm{E}(X) = \sum_{n=0}^{+\infty} n e^{-\lambda} \frac{\lambda^n}{n!} = \lambda e^{-\lambda} \sum_{n=1}^{+\infty} \frac{\lambda^{n-1}}{(n-1)!} = \lambda,$$

where we have used the well known fact that

$$\sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{\lambda}; \tag{1.21}$$

moreover

$$\begin{aligned}\mathrm{E}(X^2) = \mathrm{E}[X(X-1)] + \mathrm{E}(X) &= e^{-\lambda} \sum_{n=0}^{+\infty} n(n-1) \frac{\lambda^n}{n!} + \lambda = e^{-\lambda} \sum_{n=2}^{+\infty} \frac{\lambda^n}{(n-2)!} + \lambda \\ &= e^{-\lambda} \lambda^2 \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} + \lambda = \lambda^2 + \lambda\end{aligned}$$

which gives

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [E(X)]^2 = \lambda.$$

1.6.5 Hypergeometric distributions

Suppose we make n draws from an urn containing N balls, m of which are *red*, the other $N - m$ are *blue*. Denote by X the number of red balls among the n drawn. In the case we draw *with replacement*, i.e. balls are reinserted in the urn immediately after being drawn, then $X \sim \mathrm{Bin}\left(n, \frac{m}{N}\right)$. Suppose, instead, that we draw *without replacement*, in particular $n \leq N$. A simple combinatorial argument shows that

$$p_X(k) = \mathrm{P}(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \tag{1.22}$$

where the above formula holds for every $k \geq 0$ once for binomial coefficients we agree on the convention that $\binom{n}{k} = 0$ whenever $k < 0$ or $k > n$. Any random variable having density given by (1.22) is called *Hypergeometric*, and we write

$$X \sim Hp(n,N,m).$$

To compute mean and variance, it is convenient to derive first the following nice formula:

$$\mathrm{E}\left[\binom{X}{l}\right] = \frac{\binom{m}{l}\binom{n}{l}}{\binom{N}{l}}. \tag{1.23}$$

We prove (1.23) by induction on l. For $l=0$ it is obviously true, as $\binom{h}{0} = 1$ for every $h \geq 0$. Assume then (1.23)holds up to l (for every choice of N,m,n), and let us prove it holds for $l+1$. We first observe that

$$\binom{X}{l+1} = \binom{X}{l}\frac{X-l}{l+1},$$

so that

$$\begin{aligned}
\mathrm{E}\left[\binom{X}{l+1}\right] &= \frac{1}{l+1}\mathrm{E}\left[X\binom{X}{l}\right] - \frac{l}{l+1}\mathrm{E}\left[\binom{X}{l}\right] \\
&= \frac{1}{l+1}\mathrm{E}\left[X\binom{X}{l}\right] - \frac{l}{l+1}\frac{\binom{m}{l}\binom{n}{l}}{\binom{N}{l}},
\end{aligned} \tag{1.24}$$

where the inductive assumption has been used. Thus we have to compute $\mathrm{E}\left[X\binom{X}{l}\right]$. We use the identities

$$k\binom{m}{k} = m\binom{m-1}{k-1}, \quad \binom{k}{l} = \binom{k-1}{l-1} + \binom{k-1}{l} \quad \text{and} \quad \binom{N}{n} = \binom{N-1}{n-1}\frac{N}{n}$$

which gives

$$\begin{aligned}
\mathrm{E}\left[X\binom{X}{l}\right] &= \sum_k k\binom{k}{l}\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = m\sum_k \binom{k}{l}\frac{\binom{m-1}{k-1}\binom{N-m}{n-k}}{\binom{N}{n}} \\
&= m\sum_k\left[\binom{k-1}{l-1} + \binom{k-1}{l}\right]\frac{\binom{m-1}{k-1}\binom{(N-1)-(m-1)}{(n-1)-(k-1)}}{\binom{N-1}{n-1}}\frac{n}{N}.
\end{aligned} \tag{1.25}$$

Making the change of variable $h = k-1$:

$$\begin{aligned}\mathrm{E}\left[X\binom{X}{l}\right] &= m\frac{n}{N}\left[\sum_h \binom{h}{l-1}\frac{\binom{m-1}{h}\binom{(N-1)-(m-1)}{(n-1)-h}}{\binom{N-1}{n-1}} + \sum_h \binom{h}{l}\frac{\binom{m-1}{h}\binom{(N-1)-(m-1)}{(n-1)-h}}{\binom{N-1}{n-1}}\right]\\ &= m\frac{n}{N}\left(\mathrm{E}\left[\binom{Y}{l-1}\right] + \mathrm{E}\left[\binom{Y}{l}\right]\right),\end{aligned} \tag{1.26}$$

where $Y \sim \mathrm{Hyp}(n-1, N-1, m-1)$. By the inductive assumption

$$\mathrm{E}\left[\binom{Y}{l-1}\right] = \frac{\binom{m-1}{l-1}\binom{n-1}{l-1}}{\binom{N-1}{l-1}}$$

and

$$\mathrm{E}\left[\binom{Y}{l}\right] = \frac{\binom{m-1}{l}\binom{n-1}{l}}{\binom{N-1}{l}}$$

which, inserted in (1.26), give

$$\begin{aligned}\mathrm{E}\left[X\binom{X}{l}\right] &= m\frac{n}{N}\left[\frac{\binom{m-1}{l-1}\binom{n-1}{l-1}}{\binom{N-1}{l-1}} + \frac{\binom{m-1}{l}\binom{n-1}{l}}{\binom{N-1}{l}}\right]\\ &= l\frac{\binom{m}{l}\binom{n}{l}}{\binom{N}{l}} + (l+1)\frac{\binom{m}{l+1}\binom{n}{l+1}}{\binom{N}{l+1}},\end{aligned}$$

where we have used repeatedly $k\binom{m}{k} = m\binom{m-1}{k-1}$. Inserting this in (1.24), we easily obtain (1.23).

Note that, since $\binom{X}{1} = X$, (1.23) for $l = 1$ gives

$$\mathrm{E}(X) = n\frac{m}{N}. \tag{1.27}$$

Similarly, since $\binom{X}{2} = \frac{X(X-1}{2}$, (1.23) for $l = 2$ gives

$$\mathrm{E}[X(X-1)] = \frac{m(m-1)n(n-1)}{N(N-1)} = \mathrm{E}(X^2) - \mathrm{E}(X),$$

which gives

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \mathrm{E}[X(X-1)] + \mathrm{E}(X) - [\mathrm{E}(X)]^2\\ &= \frac{nm(N-m)}{N^2}\frac{N-n}{N-1}.\end{aligned} \tag{1.28}$$

Note that if we were sampling with replacement, so $X \sim \mathrm{Bin}\left(n, \frac{m}{N}\right)$, we would get the same mean, but the variance would be $\frac{nm(N-m)}{N^2}$. Thus, the variance of the hypergeometric is smaller by the factor $\frac{N-n}{N-1}$ (which is close to one if $N \gg n$). A way of interpreting al this, and indeed to give a alternative derivation, is to define, for

$i = 1,\dots,N$, the random variable X_i as the indicator function of the event "the i-th ball drawn is red". In the case of sampling with replacement, the X_i are independent Be $\left(\frac{m}{N}\right)$, and thus the variance of X is the the sum of the variances of the X_i's. If we sample without replacement, the X_i's are still Be $\left(\frac{m}{N}\right)$, but are not independent. It is not hard to compute their covariances, and re-obtain (1.28) using Proposition 1.38.

1.7 Expectation of general real random variables

In this section we define expectation for general real random variables. A detailed treatment of this topic is well beyond the purposes of these notes; we will give rigorous definitions but no proofs.

Let $(\Omega, \mathscr{A}, \mathrm{P})$ be a probability space, and $X : \Omega \to \mathbb{R}$ a real random variable. Let $\mathscr{D}^+$ be the family of all nonnegative, discrete random variables defined on $(\Omega, \mathscr{A}, \mathrm{P})$. Suppose first that X is a nonegative random variable. We define the expected value of X as follows:

$$\mathrm{E}(X) := \sup\{\mathrm{E}(Y) : Y \in \mathscr{D}^+, Y \leq X\} \in [0, +\infty]. \tag{1.29}$$

This rather formal definition can be made more constructive as follows. Given a nonegative random variable X, it is not difficult to construct discrete random variables the approximate X from below: for instance, for $n \geq 1$, we can define the discrete random variable X_n as follows:

$$X_n(\omega) = \sum_{k=0}^{+\infty} \frac{k}{2^n} \mathbf{1}_{\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)}(X(\omega)). \tag{1.30}$$

In other words, X_n takes the value $\frac{k}{2^n}$ whenever X takes values between $\frac{k}{2^n}$ and $\frac{k+1}{2^n}$. Note that, for each $\omega \in \Omega$, $X_n(\omega) \leq X_{n+1}(\omega)$, and

$$\lim_{n \to +\infty} X_n(\omega) = X(\omega).$$

By carefully exploiting Definition 1.29, it can be shown that

$$\mathrm{E}(X) = \lim_{n \to +\infty} \mathrm{E}(X_n) = \lim_{n \to +\infty} \sum_{k=0}^{+\infty} \frac{k}{2^n} \mathrm{P}\left(\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\right). \tag{1.31}$$

The convergence result in (1.31) can be generalized to obtain the following fundamental result,: it states that the expectation $\mathrm{E}(X)$ can be computed taking the limit $\lim_{n \to +\infty} \mathrm{E}(X_n)$ along *any* sequence X_n of random variables which approximates X from below.

Theorem 1.41. (Monotone Convergence Theorem). *Let $(X_n)_{n\geq 1}$ be a sequence of nonnegative random variables having the following properties:*

(i) for every $\omega \in \Omega$ and $n \geq 1$, $X_n(\omega) \leq X_{n+1}(\omega)$, i.e. the sequence $X_n(\omega)$ is increasing;

(ii) for every $\omega \in \Omega$,

$$\lim_{n\to+\infty} X_n(\omega) = X(\omega),$$

Then

$$\mathrm{E}(X) = \lim_{n\to+\infty} \mathrm{E}(X_n).$$

If X is a real random variable we can always decompose it as $X = X^+ - X^-$. We say that X *admits expectation* if at least one of the expectations $\mathrm{E}(X^+), \mathrm{E}(X^-)$ is finite, and in this case we set

$$\mathrm{E}(X) = \mathrm{E}(X^+) - \mathrm{E}(X^-).$$

Moreover, if both $\mathrm{E}(X^+) < +\infty$ and $\mathrm{E}(X^-) < +\infty$, then we say X is *integrable*. Many of the results stated in Section 1.4 can be extended to this general context; for ease of reading, we restate them.

Proposition 1.42. *A real random variable X is integrable if and only if $\mathrm{E}(|X|) < +\infty$.*

Theorem 1.43. *Let X and Y be random variables, defined on the same probability space, which are either both nonnegative of both integrable. Then*

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Theorem 1.44. *Let X and Y be random variables, and a, b be real numbers. Then the following properties hold:*

(a) (Monotonicity) *If $X(\omega) \leq Y(\omega)$ for every $\omega \in \Omega$ and the expectation is well defined both X and Y, then $\mathrm{E}(X) \leq \mathrm{E}(Y)$. Moreover*

$$|\mathrm{E}(X)| \leq \mathrm{E}(|X|). \tag{1.32}$$

(b) (Linearity) *If X and Y are integrable (or if they are nonnegative and $a, b \geq 0$), then*

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

Proposition 1.45. *(a) If X and Y are real random variable such that* $\mathrm{P}(X = Y) = 1$*, then* $\mathrm{E}(X)$ *and* $\mathrm{E}(Y)$ *are either both well defined or both ill defined, and in the first case*

$$\mathrm{E}(X) = \mathrm{E}(Y).$$

(b) If X is nonnegative and $\mathrm{E}(X) = 0$*, then* $\mathrm{P}(X = 0) = 1$.

Proposition 1.46. *Let X and Y be real, independent and integrable (resp. nonnegative) random variables. Then their product XY is integrable (resp. nonnegative) and*

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

Having the notion of expected value, we can define square integrable random variables, Variance and Covariance as in Section 1.5. Also the following related results hold in general.

Proposition 1.47. *Let X and Y square-integrable random variables, defined on the same probability space. Then their product XY is integrable.*

Proposition 1.48. *If X and Y are square integrable and independent, then* $\mathrm{Cov}(X,Y) = 0$*, and in particular*

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

Proposition 1.49. *Let X,Y,Z be discrete, square integrable random variables. Then*

(a)

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X);$$

(b) the covariance is symmetric*, i.e.*

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X);$$

(c) the covariance is linear in each component (or bilinear*), i.e. for each reals a,b*

$$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z).$$

Remark 1.50. It is useful to note here that the expectation of a real random variable uniquely depend on its distribution. An even more general statement holds. Let X be a random variable taking values in $(E,\mathscr{E})$, and let μ_X be its law. Consider a function

$f : E \to [0, +\infty)$ such that $\{x : f(x) \in A\} \in \mathscr{E}$ for every Borel set A; thus $f(X)$ is a nonegative random variable. By (1.31),

$$\begin{aligned} \mathrm{E}[f(X)] &= \lim_{n\to+\infty} \sum_{k=0}^{+\infty} \frac{k}{2^n} \mathrm{P}\left(\frac{k}{2^n} \leq f(X) < \frac{k+1}{2^n}\right) \\ &= \lim_{n\to+\infty} \sum_{k=0}^{+\infty} \frac{k}{2^n} \mu_X \left(\left\{x : \frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\right\}\right) \end{aligned}$$

only depends on μ_X. This can be extended to any function $f : E \to \mathbb{R}$ such that $\{x : f(x) \in A\} \in \mathscr{E}$ for every Borel set A, and such that $f(X)$ is integrable. Thus, if two random variables have the same distribution, then all expectations of functions of them are equal. □

1.8 Absolutely continuous distributions: the univariate case

Consider a real random variable X, and denote by F_X its distribution function. By Proposition 1.19 and Remark 1.24, F_X has at most countably many discontinuity points, that are those in the set

$$D := \{x \in E : \mathrm{P}(X = x) > 0\}.$$

Define

$$F_d(x) = \sum_{y \leq x} \mathrm{P}(X = x).$$

Note that F_d is increasing and discontinuous in all points of D, and it "jumps like F_X", i.e.

$$F_d(x) - F_d(x^-) = F_X(x) - F_X(x^-).$$

It follows that the difference

$$F_c(x) := F_X(x) - F_d(x)$$

is continuous, and it is not hard to show that it is still increasing. A deep result in mathematical analysis states that all increasing functions are such that their derivative exists for *almost all* points $x \in \mathbb{R}$, and it is nonnegative[2]. Thus we can define

$$F_{ac}(x) := \int_{-\infty}^{x} F_c'(y)dy.$$

[2] The precise statement is that if $F : \mathbb{R} \to \mathbb{R}$ is increasing, then $F'(x)$ exists except at most for a set of values of x having *Lebesgue measure* zero. The Lebesgue measure can be thought of as the unique way of assigning an "extension" to Borel sets consistent with the requirement that the extension of an interval $[a,b]$ equals $b-a$. See Remark 1.52 for more on this

In the case F_c is sufficiently regular, e.g. it has a continuous derivative, the so-called *Fundamental Theorem of Calculus* states that $F_c = F_{ac}$. This is not, however, always true, and in general only the inequality $F_{ac} \leq F_c$ holds. If we set

$$F_s(x) := F_c(x) - F_{ac}(x),$$

it can be shown that F_s is continuous and increasing, its derivative is zero in almost all x, but it can be not identically zero. Summing all up we have written

$$F_X = F_d + F_{ac} + F_s.$$

The three functions F_d, F_{ac} and F_s are all increasing, but are of very different nature: F_X can only increase by jumps and it is constants between two consecutive jumps, F_{ac} is a “nice” function with the property of being the integral of its derivative, while F_s is quite singular, indeed quite hard to imagine. We typically deal with real random variables such that the *singular part* F_s of their distribution function is identically zero. Moreover, we see that a real random variable is discrete if and only if $F_X = F_d$, i.e. $F_c \equiv 0$.

The decomposition for F_X above suggests the following definition.

Definition 1.51. A real random variable X is called *absolutely continuous* (or we say it has an absolutely continuous distribution) if there exists a nonnegative function f_X, integrable on $(-\infty, +\infty)$, such that for each $x \in \mathbb{R}$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{1.33}$$

The function f_X is called *density* of the random variable X.

Remark 1.52. There are mathematical subtleties related to Definition 1.51, for what concerns the notion of integrability. Although for most purposes Riemann integral (generalized to unbounded intervals) suffices, a solid mathematical foundation needs the more advanced notion of *Lebesgue integral*. We briefly summarize here what it is about.

To begin with, one consider the *Lebesgue measure* m on $\mathscr{B}(\mathbb{R})$, which is the unique assigment

$$m : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$$

which obeys the following properties:

- for every bounded interval $[a,b]$, $m([a,b]) = b - a$;
- σ-additivity holds, i.e. for any sequence $(A_n)_{n\geq 1}$ of *disjoint* Borel sets,

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

Existence and uniqueness of the Lebesgue measure is a deep, hard to prove, result. Now, consider a *Borel-measurable*, nonnegative function f, i.e. a function $f : \mathbb{R} \to [0,+\infty)$ such that $\{x : f(x) \in A\} \in \mathscr{B}(\mathbb{R})$ for every $A \in \mathscr{B}(\mathbb{R})$. It should be remarked that all "common" functions are Borel measurable, in particular continuous functions. Using the same method used in (1.30), we can approximate f by "discrete" functions, i.e. functions taking at most countably many values, and define, in analogy with (1.31),

$$\int f(x)dx := \lim_{n\to+\infty} \frac{k}{2^n} m\left\{x : \frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\right\}. \tag{1.34}$$

It is not hard to show that the above limit exists in $[0,+\infty]$, being the limit of an increasing sequence. Finally, for a general Borel-measurable $f : \mathbb{R} \to \mathbb{R}$, we set

$$\int f(x)dx = \int f^+(x)dx - \int f^-(x)dx, \tag{1.35}$$

whenever at least one of the two summands is finite. This defines the Lebesgue integral. The following version of the Dominated Convergence Theorem stated in Theorem 1.41 for the expectation, holds for the Lebesgue integral: if f_n is an *increasing* sequence of nonnegative, Borel-measurable functions, such that for every $x \in \mathbb{R}$

$$\lim_{n\to+\infty} f_n(x) = f(x),$$

then

$$\int f(x)dx = \lim_{n\to+\infty} \int f_n(x)dx. \tag{1.36}$$

Using the definition above it is easy to show that for any Borel set A

$$\int \mathbf{1}_A(x)dx = m(A).$$

Moreover we define

$$\int_A f(x)dx := \int \mathbf{1}_A(x)f(x)dx.$$

Finally, we mention the fact that Lebesgue integral is consistent with Riemann integral, in the following sense: if f is a Riemann-integrable, nonnegative function, then f is also Lebesgue integrable, and the two integrals coincide. Key properties of the Lebesgue integral are:

- *Monotonicity*: $f \leq g$ implies $\int f(x)dx \leq \int g(x)dx$;
- *Linearity*: $\int(af(x)+bg(x))dx = a\int f(x)dx + b\int g(x)dx$.
- If $f \geq 0$ and $\int f(x)dx = 0$ then $\{x : f(x) > 0\}$ has Lebesgue measure zero.

Among the several reasons to introduce Lebesgue integrals we mention the fact that Lebesgue integrable functions are many more that Riemann integrable ones. For instance, a standard example of non-Riemann integrable function is the indicator function $\mathbf{1}_{\mathbb{Q}\cap[0,1]}$ of the rationals in $[0,1]$. This function is indeed Lebesgue integrable, and its integral is zero (why?)

□

Remark 1.53. For a real, absolutely continuous random variable, the density f_X is *not* uniquely identified, but "almost uniquely". To make this precise, we introduce the following notion: we say that a property of real numbers holds *almost everywhere* if it is true up to a set of real numbers of Lebesgue measure zero. It can be shown that, given a density f_X, a function g is such that

$$\int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} g(t)dt$$

for all $x \in \mathbb{R}$, if and only if $f_X(x) = g(x)$ almost everywhere. □

A careful analysis of the above definitions, that we here omit, shows that if X is a real, absolutely continuous random variable, then for every Borel set B

$$\mathrm{P}(X \in B) = \int_B f_X(x)dx.$$

Consider now a nonnegative, absolutely continuous random variable X, with density f_X. By (1.31),

$$\begin{aligned} \mathrm{E}(X) &= \lim_{n\to+\infty} \sum_{k=0}^{+\infty} \frac{k}{2^n} \mathrm{P}\left(\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\right) \\ &= \lim_{n\to+\infty} \sum_{k=0}^{+\infty} \frac{k}{2^n} \int_{\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)} f_X(x)dx \end{aligned} \tag{1.37}$$

Note that, by (1.36),

$$\sum_{k=0}^{+\infty} \frac{k}{2^n} \int_{\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)} f_X(x)dx = \int h_n(x)dx,$$

where

$$h_n(x) = f_X(x) \sum_{k=0}^{+\infty} \frac{k}{2^n} \mathbf{1}_{\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)}(x)$$

(observe that for every given x the sum above has only one nonzero summand). Moreover, for every x, the sequence $h_n(x)$ is increasing, and it converges to $xf(x)$. Therefore, using (1.37) and again (1.36),

$$\mathrm{E}(X) = \int_{[0,+\infty)} x f_X(x)dx.$$

Adding the fact that, being X nonegative

$$0 = \mathrm{P}(X < 0) = \int_{(-\infty,0)} f_X(x)dx \implies F_X(x) = 0 \text{ almost everywhere on } (-\infty, 0)$$

so that

$$\int_{(-\infty,0)} x f_X(x)dx = 0,$$

we obtain

$$\mathrm{E}(X) = \int x f_X(x)dx. \tag{1.38}$$

Using the fact that $\mathrm{E}(X) = \mathrm{E}(X^+) - \mathrm{E}(X^-)$, one can show that (1.38) holds also for all *integrable* random variables.

A quite similar argument, allows to extend (1.38) as follows.

Proposition 1.54. *Let X be a real, absolutely continuous random variable, $g : \mathbb{R} \to \mathbb{R}$ a Borel measurable function. Then $g(X)$ is a real random variable, and if either $g \geq 0$ or $g(X)$ is an integrable random variable, then*

$$\mathrm{E}[g(X)] = \int g(x) f_X(x)dx.$$

Moreover, $g(X)$ is an integrable random variable if and only if

$$\int |g(x)| f_X(x)dx < +\infty.$$

1.9 Absolutely continuous distributions: examples

.

1.9.1 Uniform distributions

Generating random objects is the key tool in stochastic algorithms. When we generate a random object with a given distribution μ or, more precisely, we *simulate* a random variable with law μ, we say we are *sampling from* μ. In general it is hard to sample from distributions in high dimensional spaces: it requires highly nontrivial combinations of several one-dimensional samples. Basic computer routines generate samples from uniform distributions, that we now describe.

Let $a < b$ be real numbers. We say that a real random variable X is uniform in $[a,b]$ if it is absolutely continuous with density

$$f_X(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x);$$

We write $X \sim \mathrm{U}(a,b)$. This is what makes rigorous the sentence "choose a number between a and b at random". It is easily seen that:

$$\mathrm{E}(X) = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{a+b}{2}, \qquad \mathrm{E}(X^2) = \frac{1}{b-a}\int_a^b x^2\,\mathrm{d}x = \frac{a^2+ab+b^2}{3},$$

so

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{(b-a)^2}{12}.$$

In principle, if we know how to sample from uniform distributions, we can sample from any distribution on $\mathbb{R}$. More precisely, given a probability μ on $\mathbb{R}$, we can find a function $h : \mathbb{R} \to \mathbb{R}$ such that if $Y\mathrm{U}(0,1)$ then $X := h(Y)$ has distribution μ.

Proposition 1.55. *Let μ be any probability on $\mathbb{R}$ and $Y \sim \mathrm{U}(0,1)$. Set $F(x) := \mu((-\infty,x])$. Define*

$$h(y) := \inf\{z \in \mathbb{R} : F(z) \geq y\}$$

(note that $h = F^{-1}$ if F is invertible). Then $h(Y)$ has distribution μ.

Proof. A little work, left to the reader, shows the following equivalence:

$$h(y) \leq x \iff F(x) \geq y.$$

Note also that, for $0 \leq y \leq 1$,

$$F_Y(y) = \int_0^y dx = y.$$

Thus

$$F_X(x) = \mathrm{P}(h(Y) \leq x) = \mathrm{P}(F(x) \geq Y) = F_Y(F(x)) = F(x),$$

Since the distribution function uniquely determine the law, the proof is completed. □

1.9.2 Exponential and Gamma distributions

Given $\lambda > 0$, an absolutely continuous random variable X with density

$$f_X(x) = \lambda e^{-\lambda x}\mathbf{1}_{(0,+\infty)}(x)$$

is called *exponential*, and we write $X \sim \mathrm{Exp}(\lambda)$. Exponential random variables are used to model *waiting times* that are "unpredictable" in the following sense. Suppose T is the time between two consecutive occurrence of a given phenomenon (e.g. arrival of a customer, earthquakes,...), and the last occurrence took place t unit of times before now, i.e. we know $T > t$. We ask for the (conditional) probability of having to way at least h more unit of time, i.e. $\mathrm{P}(T > t+h|T > t)$. If $T \sim \mathrm{Exp}(\lambda)$,

$$\mathrm{P}(T > t) = \lambda \int_t^{+\infty} e^{-\lambda x} dx = e^{-\lambda t},$$

so

$$\mathrm{P}(T > t+h\,|\,T > t) \,=\, \frac{\mathrm{P}(T > t+h)}{\mathrm{P}(T > t)} = e^{-\lambda t} = \mathrm{P}(T > h)$$

In other world, the information $T > t$ is totally useless in computing the distribution of the *remaining waiting time*.

Gamma distributions generalize exponential distributions in the sense that we consider densities proportional to $x^{\alpha-1}e^{-\lambda x}\mathbf{1}_{(0,+\infty)}(x)$, for every $\alpha > 0$. Since a density must integrate to one, we need to compute

$$\int x^{\alpha-1}e^{-\lambda x}\mathbf{1}_{(0,+\infty)}(x)dx = \int_0^{+\infty} x^{\alpha-1}e^{-\lambda x}dx.$$

Observe that, for $\alpha > 0$, this integral is finite. Using the change of variable $y = \lambda x$ we get

$$\int_0^{+\infty} x^{\alpha-1}e^{-\lambda x}dx = \frac{1}{\lambda^\alpha}\int_0^{+\infty} y^{\alpha-1}e^{-y}dy.$$

The integral

$$\Gamma(\alpha) := \int_0^{+\infty} y^{\alpha-1}e^{-y}dy,$$

except for special cases, is not explicitly computable. Its value $\Gamma(\alpha)$ gives the so-called *Euler Gamma function*. Easily, $\Gamma(1) = 1$, and, by an integration by parts, for every $\alpha < 0$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha). \tag{1.39}$$

This implies that for any $n \in \mathbb{N}$,

$$\Gamma(n) = (n-1)!$$

(just prove it by induction).

We can now define Gamma random variables as those having density of the form

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}\mathbf{1}_{(0,+\infty)}(x),$$

and we write $X \sim \mathrm{Gamma}(\alpha,\lambda)$. Note that exponential random variables are special cases, as $\mathrm{Gamma}(1,\lambda) = \mathrm{Exp}(\lambda)$. The change of variable used above can be exploited to show the following property:

$$X \sim \mathrm{Gamma}(\alpha,\lambda) \qquad \Longleftrightarrow \qquad Y := \lambda X \sim \mathrm{Gamma}(\alpha,1) \tag{1.40}$$

(compute F_Y in terms of F_X and change variable). Since $\mathrm{E}(X) = \frac{\mathrm{E}(X)}{\lambda}$ and $\mathrm{Var}(X) = \frac{\mathrm{Var}(Y)}{\lambda^2}$, it is enough to compute mean and variance for $Y \sim \mathrm{Gamma}(\alpha, 1)$: recalling (1.39)

$$\mathrm{E}(Y) = \int_{-\infty}^{+\infty} y\, f_Y(y)\, \mathrm{d}y = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} y^{\alpha}\, e^{-y}\, \mathrm{d}y = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha\,,$$

and, similarly,

$$\mathrm{E}(Y^2) = \int_{-\infty}^{+\infty} y^2\, f_Y(y)\, \mathrm{d}y = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} y^{\alpha+1}\, e^{-y}\, \mathrm{d}y = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \alpha(\alpha+1)\,,$$

giving

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \alpha\,.$$

Finally, for $X := \frac{1}{\lambda} Y \sim \mathrm{Gamma}(\alpha, \lambda)$ we have

$$\mathrm{E}(X) = \frac{\alpha}{\lambda}\,, \qquad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}\,. \tag{1.41}$$

1.9.3 Normal or Gaussian distributions

Distributions called *Normal* or Gaussian are the most used in application, as they model a great variety of phenomena. Any time the outcome of a random phenomenon is the sum of several, nearly independent small contributions, then it can be modeled by a Normal random variable. The theoretical basis of this statement if provided by the Central Limit Theorem, that we will see later.

We say that a random variable Z is a *standard normal*, and we write $Z \sim \mathrm{N}(0,1)$, if it is absolutely continuous with density

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

The fact that f_Z is a density, i.e. it integrates to one, comes from the identity

$$\int e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}. \tag{1.42}$$

A way of deriving (1.42) is to write

$$\begin{aligned} \left(\int e^{-\frac{x^2}{2}} dx \right)^2 &= \int e^{-\frac{x^2}{2}} dx \int e^{-\frac{y^2}{2}} dy \\ &= \int \int e^{-\frac{x^2+y^2}{2}} dx dy \end{aligned}$$

and pass to polar coordinates (details are omitted). We have:

$$\mathrm{E}(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{x^2}{2}} \, \mathrm{d}x = 0 \, ,$$

since the function$x e^{-x^2/2}$ is integrable and odd. Moreover, integrating by parts,

$$\begin{aligned} \mathrm{Var}(Z) &= \mathrm{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left[-x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \, \mathrm{d}x \right] = 1 \, . \end{aligned}$$

Now, let $\mu \in \mathbb{R}$ and $\sigma > 0$ be two constants, and set $X = \mu + \sigma Z$. It is easily checked that $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Moreover, we can write

$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(\sigma Z + \mu \leq x) = \mathrm{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(Z \leq \frac{x-\mu}{\sigma}\right) .$$

The distribution function F_Z is continuously differentiable, since its derivative is f_Z. Thus also F_X is continuously differentiable. By the Fundamental Theorem of Calculus, a continuously differentiable is the integral of its derivative. So F_X is the distribution function of an absolutely continuous random variable, whose density is

$$f_X(x) = F_X'(x) = \frac{1}{\sigma} F_Z'\left(Z \leq \frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma} f_Z\left(Z \leq \frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

Any random variable with this density is called *Normal*, and we write $X \sim \mathrm{N}(\mu, \sigma^2)$.

The argument used to compute the density of X can be adapted to show the following fact, stating that any affine transformation of a Normal random variable is still Normal.

Proposition 1.56. *Let $X \sim \mathrm{N}(\mu, \sigma^2)$, and $a, b \in \mathbb{R}$ with $a \neq 0$. Then*

$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2) .$$

1.10 Absolutely continuous distributions: the multivariate case

.

We now generalize what seen in Section 1.8 to random vectors. We first briefly summarize the notion of Lebesgue measure and Lebesgue integral in $\mathbb{R}^n$. We say that a subset of $\mathbb{R}^n$ is a *rectangle* if it is of the form

$$[a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_n,b_n].$$

We denote by $\mathscr{B}(\mathbb{R}^n)$ the minimal σ-algebra containing all rectangles, whose elements are called Borel sets, as in dimension one. The Lebesgue measure m_n on $\mathscr{B}(\mathbb{R}^n)$ is uniquely determined by the properties of being σ-additive, and the fact that

$$m_n([a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_n,b_n])=\prod_{i=1}^n(b_i-a_i).$$

The existence and uniqueness of m_n is highly nontrivial, but we accept it. A function $f:\mathbb{R}^n\to\mathbb{R}$ is said *Borel measurable* if for every $A\in\mathscr{B}(\mathbb{R})$ we have that $\{x\in\mathbb{R}^n : f(x)\in A\}\in\mathscr{B}(\mathbb{R}^n)$. The n-dimensional, or multiple Lebesgue integral of f can be defined as in (1.34) and (1.35), simply replacing m by m_n. This integral will be denoted by

$$\int f(x)dx \quad \text{or} \quad \int f(x_1,x_2,\ldots,x_n)dx_1dx_2\cdots dx_n$$

depending whether we want to stress the fact that the variable x is a vector. Finally, as in the one dimensional case, for $B\in\mathscr{B}(\mathbb{R}^n)$ we set

$$\int_B f(x)dx=\int f(x)\mathbf{1}_B(x)dx.$$

One of the most useful properties of multiple Lebesgue integrals is most simply stated in the case $n=2$, and shows that a multiple Lebesgue integral can be reduced to a sequence of one dimensional integrals.

Theorem 1.57. (Fubini) *Consider a Borel-measurable $f:\mathbb{R}^2\to\mathbb{R}$ which is either nonnegative or integrable. Then the following properties hold:*

(i) for almost every $x\in\mathbb{R}$ the integral

$$\int f(x,y)dy$$

is well defined, and the function $x\mapsto\int f(x,y)dy$ is Borel-measurable;

(ii) the integral

$$\int\left[\int f(x,y)dy\right]dx$$

is well defined and coincides with the double integral $\int f(x,y)dxdy$.

We are now ready for the next definition.

Definition 1.58. A n-dimensional random vector X is called absolutely continuous if there exists a integrable, nonnegative $f_X : \mathbb{R}^n \to [0, +\infty)$ such that

$$\mathrm{P}(X \in R) = \int_R f_X(x)dx \tag{1.43}$$

for every n-dimensional rectangle R.

It can be shown that if X is an absolutely continuous random vector then (1.43) actually holds for every Borel set R. In particular, if $n = 2$, consider a two-dimensional random vector (X,Y) and we take R of the form $(-\infty, z] \times \mathbb{R}$, using (1.57) we have

$$\begin{aligned} F_X(z) = \mathrm{P}(X \le z) = \mathrm{P}((X,Y) \in R) &= \int f_{X,Y}(x,y)\mathbf{1}_{(-\infty,z]}(x)dxdy \\ &= \int_{-\infty}^{z} \left[\int f_{X,Y}(x,y)dy \right] dx, \end{aligned}$$

where, in analogy with the discrete case, $f_{X,Y}$ denotes the density of the random vector (X,Y), also called the *joint density* of the random variables X and Y. This shows that X is a real, absolutely continuous random variable with density

$$f_X(x) = \int f_{X,Y}(x,y)dy.$$

More generally, the following "continuous" version of Proposition 1.26 holds.

Proposition 1.59. *Let $X = (X_1, X_2, \ldots, X_n)$ be a absolutely continuous random vector, with joint density $f_X(x_1, x_2, \ldots, x_n)$. Then the marginal density f_{X_i} is given by*

$$f_{X_i}(x_i) = \int f_X(x_1, x_2, \ldots, x_n)dx_1 \cdots dx_{i-1}dx_{i+1} \cdots dx_n.$$

Thus, the components of an absolutely continuous random vector are absolutely continuous random variables. The converse is not necessarily true (examples will be given later). It is true, however, when the components are independent.

Theorem 1.60. *(i) Let $X_1, X_2, \ldots, X_n$ be independent, real, absolutely continuous random variables, and denote by f_{X_i} the density of X_i. Then the vector $X = (X_1, X_2, \ldots, X_n)$ is absolutely continuous with joint density*

$$f_X(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i). \tag{1.44}$$

(ii) Let $X = (X_1, X_2, \dots, X_n)$ be an absolutely continuous random vector. Then its components are independent if and only if (1.44) *holds.*

Note that if X is a n-dimensional random vector and $g : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable, then $g(X)$ is a real random variable. The following extension of Proposition 1.54 can be derived.

Proposition 1.61. *If X is absolutely continuous with density f_X, and if either g is nonnegative or $g(X)$ is integrable, then*

$$\mathrm{E}[g(X)] = \int g(x) f_X(x) dx.$$

Moreover, $g(X)$ is an integrable random variable if and only if

$$\int |g(x)| f_X(x) dx < +\infty.$$

In dealing with Normal random variables we have seen how, given the density of an absolutely continuous real random variable X, to get the density of an affine transform $aX + b$ of X. This can be extended to higher dimension and nonlinear transformations. To state this result, consider a differentiable function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$. Thus, for $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, we have $\varphi(x) = (\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x))$. Thus, for $i, j \in \{1, 2, \dots, n\}$, we can consider the partial derivatives $\frac{\partial \varphi_i}{\partial x_j}$. The $n \times n$ matrix whose entry (i, j) is $\frac{\partial \varphi_i}{\partial x_j}(x)$ is called the *Jacobian matrix* of φ, and we denote it by $D\varphi(x)$.

Theorem 1.62. *Let X be a n-dimensional, absolutely continuous random vector, with density f_X. Let U be an open subset $\mathbb{R}^n$ such that $\mathrm{P}(X \in U) = 1$, and $\varphi : U \to V$ be a* diffeomorphism*, i.e. and invertible function such that both φ and φ^{-1} are continuously differentiable. Then the random vector $Y := \varphi(X)$ is absolutely continuous with density*

$$f_Y(y) = \begin{cases} f_X(\varphi^{-1}(y)) |\det D\varphi^{-1}(y)| & \textit{se } y \in V \\ 0 & \textit{altrimenti} \end{cases}. \tag{1.45}$$

A useful special case is that in which φ is an affine transform, i.e. of the form[3]

$$\varphi(x) = Ax + b$$

where A is a $n \times n$ invertible matrix, and $b \in \mathbb{R}^n$. In this case $Y = AX + b$ has density

[3] note that when the matrix multiplication Ax is considered, x is meant as s *column* vector.

$$f_Y(y) = \frac{1}{|\det(A)|} f_X(A^{-1}(y-b)). \tag{1.46}$$

1.11 Characteristic functions and sums of independent random variables

We first recall the identity for complex numbers

$$e^{ix} = \cos(x) + i\sin(x).$$

Note also that if X is a real random variable, then $|cos(X)| \leq 1$ and $|\sin(X)| \leq 1$. In particular, $\cos(X)$ and $\sin(X)$ are integrable random variables. It therefore makes sense to define

$$\mathrm{E}\left[e^{iX}\right] := \mathrm{E}[\cos(X)] + i\,\mathrm{E}[\sin(X)] \in \mathbb{C}.$$

Definition 1.63. Let X be a real random variable. The function $\varphi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\varphi_X(u) = \mathrm{E}\left[e^{iuX}\right]$$

is called the *characteristic function* of X.

Characteristic functions are useful in many respects, both theoretically and computationally. The first relevant property is that they uniquely identify the law of a real random variable:

$$\mu_X = \mu_Y \iff \varphi_X = \varphi_Y.$$

The "⇒" implication is a consequence of Remark 1.50. The converse is more delicate; we do not give a formal proof, but it is useful to know *why* it is true. Suppose $\varphi_X = \varphi_Y$. To prove that $\mu_X = \mu_Y$ it is enough to show that for all $a < b$, $\mu_X((a,b)) = \mu_Y((a,b))$, since probabilities on $\mathbb{R}$ are uniquely identified by their values on intervals. The above identity can be written as

$$\mathrm{E}\left[\mathbf{1}_{(a,b)}(X)\right] = \mathrm{E}\left[\mathbf{1}_{(a,b)}(Y)\right]. \tag{1.47}$$

Since the function $\mathbf{1}_{(a,b)}$ can be approximated from below by a continuous, piecewise linear function[4], (1.47) is implied by

$$\mathrm{E}[f(X)] = \mathrm{E}[f(Y)] \tag{1.48}$$

for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$. A classical theorem in Analysis states that any continuous function can be approximated by linear combinations of *trigonometric functions*; a way of stating this is to say that there are real numbers

[4] just take a function f which takes value zero outside of (a,b), value one is $[a+\varepsilon, b-\varepsilon]$ with ε small, and interpolates linearly in between.

$u_1, u_2, \dots, u_n$ and complex numbers $a_1, a_2, \dots a_n$ such that

$$f(x) \simeq \sum_{k=1}^{n} a_k e^{iu_k x}, \tag{1.49}$$

so that

$$\mathrm{E}[f(X)] \simeq \sum_{k=1}^{n} a_k \varphi_X(u_k).$$

Since this last expression stays the same if we replace X with Y, so does $\mathrm{E}\left[\mathbf{1}_{(a,b)}(X)\right]$, and so $\mu_X = \mu_Y$. The relevance of this result makes it worth stating it as a theorem.

Theorem 1.64. *For two real random variables X and Y,*

$$\mu_X = \mu_Y \iff \varphi_X = \varphi_Y.$$

The fact that the characteristic function identifies the distribution will be used several times later. Another useful property is the fact that the characteristic function, if known, can be used to compute the *moments* of a real random variable, i.e. the expectations of the form $\mathrm{E}(X^k)$ with $k \geq 1$. Note that the first moment ($k = 1$) is just the expectation of X. If we formally take the derivative of φ_X, interchanging derivative with expectation, we get

$$\frac{d^k}{dx^k}\varphi_X(u) = \varphi_X^{(k)}(u) = \mathrm{E}\left[\frac{d^k}{dx^k}e^{iuX}\right] = \mathrm{E}\left[X^k e^{iuX}\right]. \tag{1.50}$$

In particular

$$\mathrm{E}(X^k) = \varphi_X^{(k)}(0). \tag{1.51}$$

Interchanging derivative and expectation actually requires some assumptions. The precise result we can obtain is the following.

Proposition 1.65. *Let X be a real random variable. Then* $\mathrm{E}(|X|^k) < +\infty$ *if and only if φ_X has k continuous derivatives. In this case* (1.50) *and* (1.51) *hold.*

It is simple to extend the notion of characteristic function to random vectors. If $x, y \in \mathbb{R}^n$,

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i. \tag{1.52}$$

is the usual scalar product on $\mathbb{R}^n$ [5].

[5] x is here intended as a column vector, $\top$ denotes transposition and $x^\top y$ is the usual matrix multiplication.

Definition 1.66. Let X be a n-dimensional random vector. The function $\varphi_X : \mathbb{R}^n \to \mathbb{C}$ given by

$$\varphi_X(u) = \mathrm{E}\left[e^{i\langle u,X\rangle}\right]$$

is called the *characteristic function* of X.

Theorem 1.64 holds in this vector setting too.

We now see how independence of random variables can be established using characteristic functions. Note first that if X is a n-dimensional random vector and Y a m-dimensional random vector (defined in the same probability space) then (X,Y) is $n+m$ dimensional and we write its (joint) characteristic function as

$$\varphi_{X,Y}(u,v) = \mathrm{E}\left[\exp\left(\langle (u,v),(X,Y)\rangle\right)\right] = \mathrm{E}\left[\exp\left(\langle u,X\rangle + \langle v,Y\rangle\right)\right].$$

Theorem 1.67. *Two random vectors X and Y are independent if and only if*

$$\varphi_{X,Y}(u,v) = \varphi_X(u)\varphi_Y(v). \tag{1.53}$$

Proof. Suppose first X and Y are independent. Then also the (complex) random variables $e^{\langle u,X\rangle}$ and $e^{\langle v,Y\rangle}$ are independent[6]. Since Proposition 1.36 is easily extended to complex-valued random variables,

$$\varphi_{X,Y}(u,v) = \mathrm{E}\left[e^{\langle u,X\rangle}e^{\langle v,Y\rangle}\right] = \mathrm{E}\left[e^{\langle u,X\rangle}\right]\mathrm{E}\left[e^{\langle v,Y\rangle}\right] = \varphi_X(u)\varphi_Y(u).$$

Conversely, suppose (1.53) holds, and let X', Y' two independent random vectors such that X has the same distribution of X' and Y has the same distribution of Y'. By what we have just seen, (X,Y) and (X',Y') have the same joint characteristic function, so they have the same joint distribution. Thus, since X' and Y' are independent, so are X and Y. □

When X and Y are real, and we set $u = v$ in (1.53), we obtain the following useful way to characterize the distribution of the sum of two independent random variables.

Proposition 1.68. *Let X and Y be two real, independent random variables. Then*

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u).$$

[6] In general if X and Y are random variables and f,g are functions satisfying (1.16), then

$$\begin{aligned}\mathrm{P}(f(X)\in A, g(X)\in B) &= \mathrm{P}(X\in f^{-1}(A), Y\in g^{-1}(B)) = \mathrm{P}(X\in f^{-1}(A))\mathrm{P}(Y\in g^{-1}(B))\\ &= \mathrm{P}(f(X)\in A)\mathrm{P}(g(X)\in B)\end{aligned}$$

so $f(X)$ and $g(Y)$ are independent.

Remark 1.69. We will make use of Proposition 1.68 to determine the distribution of sums of independent random variables. It is worth mentioning that for discrete or absolutely continuous real random variables, there is an alternative procedure.

To begin with let X and Y be real, discrete random variables. Note that

$$\{X+Y=z\} = \bigcup_x \{X=x, Y=z-x\}$$

where in the above union the events are disjoint and x ranges over the (at most countable) set $\{x : p_X(x) > 0\}$. Thus, by σ-additivity and independence

$$\mathrm{P}(X+Y=z) = \sum_x \mathrm{P}(X=x, Y=z-x) = \sum_x \mathrm{P}(X=x)\,\mathrm{P}(Y=z-x),$$

that can be written as

$$p_{X+Y}(z) = p_X * p_Y(z), \tag{1.54}$$

where

$$p_X * p_Y(z) := \sum_x p_X(x) p_Y(z-x) = \sum_y p_X(z-y) p_Y(y) \tag{1.55}$$

is called the *convolution* of p_X and p_Y. Formulas (1.54) and (1.55) has an analog for absolutely continuous random variables: if X and Y are independent and absolutely continuous, then

$$f_{X+Y}(z) = f_X * f_Y(z), \tag{1.56}$$

where

$$f_X * f_Y(z) := \int f_X(x) f_Y(z-x) dx = \int f_X(z-y) f_Y(y) dy. \tag{1.57}$$

To derive it, apply (1.46) with $b=0$ and

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

and we get the joint density of $(X, X+Y)$:

$$f_{X,X+Y}(x,z) = f_X(x) f_Y(z-x)$$

from which (1.57) follows taking the marginal over the second component. □

We now give several examples where the characteristic function can be computed explicitly, and of use of Proposition 1.68.

1.11.1 Binomial

Let $X \sim \mathrm{Be}(p)$. Clearly

$$\varphi_X(u) = \mathrm{E}\left[e^{iuX}\right] = 1 - p + pe^{it}.$$

More generally, $X \sim \mathrm{Bin}(n,p)$ has the same distribution of the sum $X_1+X_2+\cdots+X_n$ of independent $\mathrm{Be}(p)$. Thus, by Proposition 1.68

$$\varphi_X(u) = \left[1-p+pe^{it}\right]^n.$$

1.11.2 Geometric and negative binomial

Let $X \sim \mathrm{Geo}(p)$. Then, using the fact that the identity

$$\sum_{n=1}^{+\infty} z^n = \frac{z}{1-z}$$

holds for every $z \in \mathbb{C}$ with $|z| < 1$, we have

$$\varphi_X(u) = \mathrm{E}\left[e^{iuX}\right] = \sum_{n=1}^{+\infty} pe^{iun}(1-p)^{n-1} = \frac{p}{1-p}\frac{(1-p)e^{iu}}{1-(1-p)e^{iu}} = \frac{p}{e^{-iu}-1+p}.$$

Having mentioned the fact that $X \sim \mathrm{NegBin}(r,p)$ has the distribution of a sum of r independent $\mathrm{Geo}(p)$, it follows by Proposition 1.68 that

$$\varphi_X(u) = \left[\frac{p}{e^{-iu}-1+p}\right]^r.$$

1.11.3 Poisson

Let $X \sim \mathrm{Pois}(\lambda)$. Then, using the fact that the identity

$$\sum_{n=0}^{+\infty} \frac{z^n}{n!}$$

holds for every $z \in \mathbb{C}$, we have

$$\varphi_X(u) = \mathrm{E}\left[e^{iuX}\right] = \sum_{n=0}^{+\infty} e^{iun}e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda e^{iu}} = \varepsilon^{\lambda(e^{iu}-1)}.$$

An important fact is the following. Suppose $X \sim \mathrm{Pois}(\lambda)$ and $Y \sim \mathrm{Pois}(\mu)$ are independent. Then by Proposition 1.68

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u) = \varepsilon^{(\lambda+\mu)(e^{iu}-1)},$$

which equals the characteristic function of a $\mathrm{Pois}(\lambda+\mu)$. Then, by Theorem 1.64 we get to the following conclusion:

$$X+Y \sim \mathrm{Pois}(\lambda+\mu),$$

thus *the sum of independent Poisson is Poisson*. It is not hard to give an alternative proof of this fact using (1.54).

1.11.4 Exponential and Gamma

One “economic” way of computing the characteristic function of Gamma (and Normal too) random variables consists in making use of the following property of the so-called *analytic* functions. Suppose that, rather than the characteristic function, we compute the *moment generating function*

$$m_X(t) = \mathrm{E}\left[e^{tX}\right]. \tag{1.58}$$

Note that, unlike for the characteristic function, there is no guarantee that $m_X(t) < +\infty$ for $t \neq 0$. Suppose, however, this is the case for all t in an open interval containig 0. It follows that m_X is an analytic function, i.e. it is the sum of its Taylor series. This implies that, preserving this analyticity, m_X extends to a region of the complex plane containing the imaginary axis. Thus the “formal” identity

$$\varphi_X(u) = m_X(iu)$$

is actually correct!

To begin with, let $X \sim \mathrm{Exp}(\lambda)$. Then

$$m_X(t) = \mathrm{E}\left[e^{tX}\right] = \int_0^{+\infty} e^{tx}\lambda e^{-\lambda x}dx = \lambda\int_0^{+\infty} e^{-(\lambda-t)x} = \frac{\lambda}{\lambda-t}$$

for all $t < \lambda$. Thus, by the above argument:

$$\varphi_X(u) = \frac{\lambda}{\lambda-iu}.$$

More generally, for $X \sim \Gamma(\alpha,\lambda)$,

$$m_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^{+\infty} x^{\alpha-1}e^{-(\lambda-t)x}\,\mathrm{d}x.$$

This last integral is finite for $t < \lambda$, and, with the change of variable $x = z/(\lambda-t)$ we get

$$\int_0^{+\infty} x^{\alpha-1}e^{-(\lambda-t)x}\,\mathrm{d}x = \frac{1}{(\lambda-t)^\alpha}\int_0^{+\infty} z^{\alpha-1}e^{-z}\,\mathrm{d}z = \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha}.$$

Summing up, for $t < \lambda$

$$m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha},$$

and

$$\varphi_X(u) = \left(\frac{\lambda}{\lambda - iu}\right)^{\alpha}.$$

Note that, by using (1.51), this provides an alternative way of computing mean and variance of Gamma random variables.

Finally, we observe that if $X \sim \Gamma(\alpha, l)$ and $Y \sim \Gamma(\beta, \lambda)$ are independent, then

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u) = \left(\frac{\lambda}{\lambda - iu}\right)^{\alpha+\beta}$$

which, by Proposition 1.68, implies that $X + Y \sim \Gamma(\alpha + \beta, \lambda)$.

1.11.5 Normal

To begin with, let $Z \sim \mathrm{N}(0,1)$. We have

$$m_Z(t) = \int_{-\infty}^{+\infty} e^{tx - \frac{1}{2}x^2}\,\mathrm{d}x.$$

Using the so-called method of *completion of squares*,

$$tx - \frac{1}{2}x^2 = -\frac{1}{2}(x-t)^2 + \frac{t^2}{2},$$

giving

$$m_Z(t) = e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\,\mathrm{d}x = e^{\frac{t^2}{2}},$$

where we have used the fact that

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\,\mathrm{d}x = 1,$$

since the integrand is the density of a $\mathrm{N}(t,1)$. Letting $t = iu$ we obtain

$$\varphi_Z(u) = e^{-\frac{u^2}{2}}.$$

More generally, $X \sim \mathrm{N}(\mu, \sigma^2)$ has the same distribution of $\sigma Z + \mu$ with $Z \sim \mathrm{N}(0,1)$. Thus

$$\varphi_X(u) = \mathrm{E}\left[e^{iu(\sigma Z + \mu)}\right] = e^{iu\mu}\varphi_Z(\sigma u) = e^{iu\mu}e^{-\frac{\sigma^2 u^2}{2}}. \tag{1.59}$$

Thus, if $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, then

$$\varphi_{X+Y}(u) = \varphi_X(u)\varphi_Y(u) = e^{iu(\mu_1+\mu_2)}e^{-\frac{(\sigma_1^2+\sigma_2^2)u^2}{2}},$$

which, by Proposition 1.68, implies that $X+Y \sim \mathrm{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

1.12 Normal random vectors

Characteristic functions provide a quite effective way of generalizing to more dimensions Normal random variables. Before getting to the point, we need to generalize the argument used in (1.60), where we computed the characteristic function of a random variable obtained as affine transform of another.

Proposition 1.70. *Let X be a m-dimensional random vector, A a $n \times m$ matrix and $b \in \mathbb{R}^n$. Set*

$$Y = AX + b.$$

Then, for all $u \in \mathbb{R}^n$

$$\varphi_Y(u) = e^{i\langle u,b\rangle}\varphi_X(A^\top u).$$

Proof. It is a simple algebraic computation:

$$\varphi_Y(u) = \mathrm{E}\left[e^{i\langle u,AX+b\rangle}\right] = e^{i\langle u,b\rangle}\,\mathrm{E}\left[e^{i\langle u,AX\rangle}\right] = e^{i\langle u,b\rangle}\,\mathrm{E}\left[e^{i\langle A^\top u,X\rangle}\right] = e^{i\langle u,b\rangle}\varphi_X(A^\top u).$$

□

Now, let $Z = (Z_1, Z_2, \ldots, Z_m)$ be a vector whose components are independent $\mathrm{N}(0,1)$. By Theorem 1.67 (actually its obvious n-dimensional generalization), we have

$$\varphi_Z(u) = \prod_{k=1}^{m}\varphi_{Z_k}(u_k) = \prod_{k=1}^{n} e^{-\frac{u_k^2}{2}} = e^{-\frac{1}{2}\langle u,u\rangle} = e^{-\frac{1}{2}\|u\|^2}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$. Let now be A a $n \times m$ matrix, and $\mu \in \mathbb{R}^m$. Then, by Proposition 1.70, setting $X = AZ + \mu$ we have

$$\varphi_X(u) = e^{i\langle u,\mu\rangle}\varphi_Z(A^\top u) = e^{i\langle u,\mu\rangle}e^{-\frac{1}{2}\langle A^\top u, A^\top u\rangle} = e^{i\langle u,\mu\rangle}e^{-\frac{1}{2}\langle \Sigma u,u\rangle},$$

where $\Sigma = AA^\top$. Note that Σ is a $n \times n$ symmetric matrix (i.e. $\Sigma^\top = \Sigma$) and its elements have a clear probabilistic interpretation:

$$\begin{aligned}
\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_i-\mu_1,X_j-\mu_j) &= \mathrm{Cov}\left(\sum_{h=1}^{m} A_{ih}Z_h, \sum_{k=1}^{m} A_{jk}Z_k\right)\\
&= \sum_{h,k=1}^{m} A_{ih}A_{jk}\,\mathrm{Cov}(Z_h,Z_k) \quad \text{(by Proposition 1.49)}\\
&= \sum_{h=1}^{n} A_{ih}A_{jh} \quad \text{(since } \mathrm{Cov}(Z_h,Z_k)=0 \text{ for } h\neq k \text{ and } =1 \text{ for } h=k\text{)}\\
&= \Sigma_{ij}.
\end{aligned}$$

In other words Σ is the so-called *covariance matrix* of the random vector X. Finally, Σ is *positive-semidefinite*, which means that

$$\langle x, \Sigma x\rangle \geq 0 \quad \text{for all } x\in\mathbb{R}^n.$$

Indeed

$$\langle x, \Sigma x\rangle = \langle x, AA^\top x\rangle = \langle A^\top x, A^\top x\rangle = \|A^\top x\|^2 \geq 0.$$

Alternatively, one can see that a covariance matrix in necessarily positive-semidefinite:

$$\begin{aligned}
\langle x, \Sigma x\rangle = \sum_{i,j=1}^{n} \Sigma_{ij}x_ix_j = \sum_{i,j=1}^{n} \Sigma_{ij}x_ix_j &= \sum_{i,j=1}^{n} \mathrm{Cov}(X_i,X_j)x_ix_j\\
= \mathrm{Var}\left(\langle x, X\rangle\right) \geq 0,
\end{aligned}$$

where we have used again Proposition 1.49. All this justifies the following definition.

Definition 1.71. A n dimensional random vector X is called *Normal* or *Gaussian* if its characteristic function φ_X is given by

$$\varphi_X(u) = e^{i\langle u,\mu\rangle}e^{-\frac{1}{2}\langle \Sigma u,u\rangle} \tag{1.60}$$

for some $\mu\in\mathbb{R}^n$ and some symmetric, positive-semidefinite $n\times n$ matrix Σ. We write

$$X\sim \mathrm{N}(\mu,\Sigma).$$

In particular, a vector $Z=(Z_1,Z_2,\ldots,Z_m)$ whose components are $\mathrm{N}(0,1)$ is a Normal random vector $\mathrm{N}(0,\mathbb{I}_n)$, where 0 denotes here the zero vector in $\mathbb{R}^n$ and $\mathbb{I}_n$ is the n-dimensional identity matrix. Moreover, we have seen that any affine transform of $Z\sim\mathrm{N}(0,\mathbb{I}_n)$ is Normal. The converse is true: any Normal random vector has the same distribution of an affine transform of $Z\sim\mathrm{N}(0,\mathbb{I}_n)$. To see this, take $X\sim\mathrm{N}(\mu,\Sigma)$. A result from Linear Algebra guarantees that for every symmetric, positive-semidefinite Σ there is a (not necessarily unique) square matrix A such that $\Sigma = AA^\top$. It follows that X has the same characteristic function, and therefore the same distribution, of $AZ+\mu$. Since composition of two affine transform is an affine transform, we obtain the following useful fact.

Proposition 1.72. *Let X be a Normal random vector. Then any affine transform of X is a Normal random vector. More specifically, let $X \sim \mathrm{N}(\mu,\Sigma)$ be a n dimensional random vector, A a $m \times n$ matrix and $b \in \mathbb{R}^m$. Then*

$$AX + b \sim \mathrm{N}(A\mu + b, A\Sigma A^\top).$$

In particular, if X is a n-dimensional Normal random vector, any component X_i is an affine (actually linear) transform of X, so it is Normal:

$$X \sim \mathrm{N}(\mu,\Sigma) \implies X_i \sim \mathrm{N}(\mu_i, \Sigma_{ii}).$$

Thus the components of a Normal random vector are Normal. But be careful, the converse is not always true: there are *non normal* random vectors whose components are all normal.

The special form of the characteristic function of a Normal vector has quite useful consequences. Consider a two-dimensional, Normal random vector $(X,Y) \sim \mathrm{N}(\mu,\Sigma)$. The off-diagonal elements $\Sigma_{12} = \Sigma_{21} = \mathrm{Cov}(X,Y)$ are zero if X and Y are independent, by Proposition 1.48. Conversely, assume $\Sigma_{12} = \Sigma_{21} = 0$. If we consider another random vector (X',Y') with independent components, and such that $X' \sim \mathrm{N}(\mu_1, \Sigma_{11})$, $Y' \sim \mathrm{N}(\mu_2, \Sigma_{22})$, by (1.60), the two vectors (X,Y) and (X',Y') have the same characteristic function, and therefore the same joint distribution. Thus, alse the components of (X,Y) are independent. This fact can be stated as follows.

Proposition 1.73. *Two components of a Normal random vector are independent if and only if are uncorrelated, i.e. their covariance is zero.*

Is a Normal random vector absolutely continuous? This is not necessarily true. For instance, take $Z = (Z_1, Z_2) \sim \mathrm{N}(0, \mathbb{I}_2)$, and the matrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

so that $AZ = (Z_1 + Z_2, 0)$. It is not hard to show that AZ, which is a two-dimensional normal random vector, it is not a two-dimensional absolutely continuos random vector, since it takes value on the x-axis that has Lebesgue measure zero in dimension two. Note that AZ is a normal random vector whose covariance matrix $\Sigma = AA^\top$ is *degenerate*, i.e. $\det(\Sigma) = 0$. The same phenomenon occurs whenever $X \sim \mathrm{N}(\mu,\Sigma)$ with $\det(\Sigma) = 0$: in this case X is *not* an absolutely continuous random vector.

If, instead, $X \sim \mathrm{N}(\mu,\Sigma)$ with $\det(\Sigma) \neq 0$, the situation is different. We have seen above that X has the same distribution of $AZ + \mu$, where $Z \sim \mathrm{N}(0, \mathbb{I}_n)$ and A is a square matrix such that $AA^\top = \Sigma$. Since $\det(\Sigma) = \det^2(A)$, it follows that $\det(A) \neq 0$, so A is invertible. Note that by Theorem 1.60, the random vector Z is absolutely continuous, with density

$$f_Z(x_1,x_2,\dots,x_n)=\prod_{k=1}^{n} f_{Z_k}(x_k)=\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}=\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\langle x,x\rangle}=\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\|x\|^2}.$$

The density $f_X = f_{AZ+\mu}$ can therefore be computed via (1.46):

$$\begin{aligned}
f_X(x) &= \frac{1}{|\det(A)|} f_Z(A^{-1}(x-\mu)) = \frac{1}{|\det(A)|}\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\langle A^{-1}x, A^{-1}x\rangle} \\
&= \frac{1}{(2\pi)^{n/2}\,|\det(A)|} e^{-\frac{1}{2}\langle x,(AA^\top)^{-1}x\rangle} \\
&= \frac{1}{(2\pi)^{n/2}\,|\det(A)|} e^{-\frac{1}{2}\langle x,\Sigma^{-1}x\rangle}.
\end{aligned}$$

We have thus obtained the following result.

Proposition 1.74. *Let $X \sim \mathrm{N}(\mu,\Sigma)$. Then X is an absolutely continuous random vector if and only if Σ is invertible (or, equivalently,* $\det\Sigma \neq 0$*). In this case*

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\,|\det(A)|} e^{-\frac{1}{2}\langle x,\Sigma^{-1}x\rangle}.$$

Chapter 2
Limit Theorems and high dimensional computations

Big data are often big in two ways: there are many data and each data element is a large dimensional vector. In this chapter we see how probabilistic asymptotic estimates can be used, to obtain robust informations from data and, sometimes, to reveal counterintuitive features.

2.1 Laws of large numbers and Chernoff bounds

The classical way of modeling sequences of independent measurements is by sequences of random variables which are independent and have the same distribution (in short: *i.i.d.*, independent, identically distributed). So let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. real random variables. Any (real valued) combination of finitely many of them is called a *statistic*. The most basic (and important) statistic is the *sample mean*:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The *Law of Large Numbers* states that the value of the sample mean can be quite accurately predicted if n is large, with a small error probability. Before stating this result, we first mention two useful inequalities.

Proposition 2.1. *(a) (*Markov inequality*). Let X be a nonnegative, integrable random variable. Then for every $\varepsilon > 0$*

$$\mathrm{P}(X \geq \varepsilon) \leq \frac{\mathrm{E}(X)}{\varepsilon}.$$

(b) (Chebischev inequality) *Let X be a square-integrable, real random variable, with mean μ and variance σ^2. Then for every $\varepsilon > 0$*

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. (a) Just take the mean in the inequality

$$\varepsilon \mathbf{1}_{\{X \geq \varepsilon\}} \leq X.$$

(b) Apply Markov Inequality to $(X-\mu)^2$.

□

Chebischev inequality provides a simple proof of a standard version of the Law of Large Numbers.

Theorem 2.2. (Weak law of large numbers). *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d., square integrable random variables, with mean μ and variance σ^2. For every $\varepsilon > 0$*

$$\lim_{n\to+\infty} P(|\overline{X}_n - \mu| \geq \varepsilon) = 0.$$

Proof. Since, for every square-integrable random variable X and $a \in \mathbb{R}$, we have $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$,

$$\operatorname{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}(X_i) = \frac{\sigma^2}{n}, \tag{2.1}$$

where we have also used Proposition 1.48. Thus, by Chebischev inequality:

$$P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}, \tag{2.2}$$

which clearly tends to zero as $n \to +\infty$. □

Inequality (2.2) establishes the "degree of concentration" of the sample mean around the mean μ, in terms of the sample size n. In particular, if a given degree of concentration is required, it allows to find the sample size n that guarantee it. Note that the upper bound in (2.2) uses very little details of the distribution of the random variables X_n, indeed only the variance. This estimate could be considerably improved if more assumptions are made. We now illustrate a class of estimates, called *Chernoff estimates*, that improve (2.2) in terms of rate of convergence to zero as $n \to +\infty$. These estimates make use of the moment generating function we have defined in (1.58):

$$m_X(t) = E\left[e^{tX}\right].$$

Consider then a sequence $(X_n)_{n\geq 1}$ of real, i.i.d. random variables. Since the X_n's have the same distribution, they have the same moment generating function

$$m(t) := m_{X_n}(t).$$

We assume $m(t) < +\infty$ for all t in an open interval containing zero; it can be shown that this assumption is stronger than square integrability. It is not hard to prove versions for the moment generating function of Proposition 1.65 and Proposition 1.68; in particular

$$m'(0) = \mu := \mathrm{E}(X_n) \tag{2.3}$$

and

$$m_{X_1+X_2+\cdots+X_n}(t) = m_{X_1}(t) \cdot m_{X_2}(t) \cdots m_{X_n}(t) = m^n(t). \tag{2.4}$$

Now, for every $t > 0$

$$\begin{aligned} \mathrm{P}\left(\overline{X}_n \geq \mu + \varepsilon\right) &= \mathrm{P}\left(X_1 + X_2 + \cdots + X_n \geq n(\mu + \varepsilon)\right) \\ &= \mathrm{P}\left(e^{t(X_1+X_2+\cdots+X_n)} \geq e^{tn(\mu+\varepsilon)}\right) \\ &\leq \frac{\mathrm{E}\left[e^{t(X_1+X_2+\cdots+X_n)}\right]}{e^{tn(\mu+\varepsilon)}}, \end{aligned}$$

where in the last step we have used Markov inequality for the random variable $e^{t(X_1+X_2+\cdots+X_n)}$. This can be rewritten as

$$\begin{aligned} \mathrm{P}\left(\overline{X}_n \geq \mu + \varepsilon\right) \leq \frac{m_{X_1+X_2+\cdots+X_n}(t)}{e^{tn(\mu+\varepsilon)}} &= \frac{m^n(t)}{e^{tn(\mu+\varepsilon)}} \\ = \exp\left[n \log m(t) - nt(\mu+\varepsilon)\right] &= \exp\left[-n\left(t(\mu+\varepsilon) - \log m(t)\right)\right]. \end{aligned}$$

Define

$$g(t) := t(\mu + \varepsilon) - \log m(t). \tag{2.5}$$

If we can find t^* such that $g(t^*) > 0$, we have

$$\mathrm{P}\left(\overline{X}_n \geq \mu + \varepsilon\right) \leq e^{-g(t^*)n}. \tag{2.6}$$

This is an exponential decay for the "upper tail" of $\overline{X}_n - \mu$, much better than the estimate in (2.2), which only guarantee a decay of order $\frac{1}{n}$. It is easy to show that such t^* exists. Indeed note that, since $m(0) = 1$, we have $g(0) = 0$. Moreover

$$g'(0) = \mu + \varepsilon - \frac{m'(0)}{m(0)} = \varepsilon > 0,$$

which implies that $\gamma(t)$ is strictly positive for some $t > 0$. Note that the best possible estimate would be obtained by maximizing $g(t)$ over t; this, however, may be a hard task. In order to obtain explicit bounds we must compute $m(t)$. Before considering specific examples, we remark that estimates for the *lower* tails can be obtained similarly: for $t > 0$

$$\begin{aligned}
\mathrm{P}\left(\overline{X}_n \le \mu-\varepsilon\right) &= \mathrm{P}\left(X_1+X_2+\cdots+X_n \le n(\mu-\varepsilon)\right)\\
&= \mathrm{P}\left(e^{-t(X_1+X_2+\cdots+X_n)} \ge e^{-tn(\mu-\varepsilon)}\right)\\
&\le \frac{\mathrm{E}\left[e^{-t(X_1+X_2+\cdots+X_n)}\right]}{e^{-tn(\mu-\varepsilon)}}\\
&= \frac{m^n(-t)}{e^{-tn(\mu-\varepsilon)}}\\
&= \exp\left[-n\left(-t(\mu-\varepsilon)-\log m(-t)\right)\right].
\end{aligned}$$

One shows as before that

$$h(t) := -t(\mu-\varepsilon)-\log m(-t) > 0 \tag{2.7}$$

for some $t>0$, and we obtain exponential decay for the lower tails too:

$$\mathrm{P}\left(\overline{X}_n \le \mu-\varepsilon\right) \le e^{-h(t^*)n}. \tag{2.8}$$

2.1.1 Bernoulli

Let $X_n \sim \mathrm{Be}(p)$. Clearly

$$m(t) := m_{X_n}(t) = 1-p+pe^t.$$

Moreover $\mu = p$. Taking, in (2.5), $\varepsilon = \delta p$ we have

$$g(t) := tp(p+\delta)-\log\left(1-p+pe^t\right).$$

Since for all $x\in\mathbb{R}$ we have that $1+x\le e^x$, we can bound

$$m(t) = 1-p+pe^t \le e^{p(e^t-1)},$$

we obtain

$$g(t) \ge tp(1+\delta)-p(e^t-1) =: \tilde{g}(t).$$

By elementary calculus one shows that $\tilde{g}(t)$ is maximized at $t^* := \log(1+\delta)$, so that

$$g(t^*) \ge p(1+\delta)\log(1+\delta)-p\delta.$$

To get a simpler lower bound observe that (exercise) for every $x\ge 0$

$$\log(1+x) \ge \frac{2x}{2+x}$$

which, multiplying by $1+x$ gives

$$(1+x)\log(1+x) - x \geq \frac{x^2}{2+x},$$

yielding

$$g(t^*) \geq p\frac{\delta^2}{2+\delta}.$$

Thus, by (2.6),

$$\mathrm{P}\left(\overline{X}_n \geq p(1+\delta)\right) \leq e^{-pn\frac{\delta^2}{2+\delta}}.$$

Bounds for the lower tails are obtained similarly: for $0 < \delta < 1$, set $\varepsilon := p\delta$; by (2.7)

$$\begin{aligned} h(t) &:= -tp(1-\delta) - \log\left(1-p+pe^{-t}\right) \\ &\geq -tp(1-\delta) - \log e^{-p(1-e^{-t})} = -tp(1-\delta) + p(1-e^{-t}) =: \tilde{h}(t), \end{aligned}$$

which is maximized at $t^* := \log\frac{1}{1-\delta}$. Using the inequality $\log(1-\delta) \geq -\delta + \frac{\delta^2}{2}$, we end up with

$$\mathrm{P}\left(\overline{X}_n \leq p(1-\delta)\right) \leq e^{-pn\frac{\delta^2}{2}}.$$

We now summarized our achievements.

Proposition 2.3. *Let $X_1, X_2, \ldots, X_n \sim \mathrm{Be}(p)$ independent. Then*

Upper Tail: for every $\delta > 0$

$$\mathrm{P}\left(\overline{X}_n \geq p(1+\delta)\right) \leq e^{-pn\frac{\delta^2}{2+\delta}};$$

Lower Tail: for every $0 < \delta < 1$

$$\mathrm{P}\left(\overline{X}_n \leq p(1-\delta)\right) \leq e^{-pn\frac{\delta^2}{2}}.$$

It is easy to provide a symmetric version of Proposition 2.3, that will be used later. Suppose, to begin with, $p \leq \frac{1}{2}$, and set $\varepsilon := p\delta$. Note that we may assume $p\delta < 1$ for, otherwise, both probabilities in Proposition 2.3 are zero. The upper tail estimate can be rewritten as

$$\mathrm{P}\left(\overline{X}_n \geq p+\varepsilon\right) \leq e^{-n\frac{a^2}{2p+p\delta}} \leq e^{-n\frac{a^2}{2}}.$$

Similarly, the lower estimate becomes

$$\mathrm{P}\left(\overline{X}_n \leq p-\varepsilon\right) \leq e^{-n\frac{\varepsilon^2}{2p}} \leq e^{-n\frac{a^2}{2}},$$

The assumption $p \leq \frac{1}{2}$ can be avoided by observing that, otherwise, we have the inequalities for $Y_n := 1 - X_n$, that end up to the same bounds for X_n. Summing all up we get

$$\mathrm{P}\Big|(\overline{X}_n - p| \geq \varepsilon\Big) \leq 2e^{-n\frac{a^2}{2}}. \tag{2.9}$$

2.1.2 Exponential

Chernoff bounds for the exponential distribution are even simpler. Let $X_n \sim \mathrm{Exp}(\lambda)$.

$$m(t) = \lambda \int_0^{+\infty} e^{tx} e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$. Thus

$$g(t) = t\left(\frac{1}{\lambda} + \varepsilon\right) - \log \frac{\lambda}{\lambda - t},$$

which attains its maximum at $t^* := \frac{\varepsilon\lambda^2}{1+\varepsilon\lambda}$, for which

$$g(t^*) = \lambda\varepsilon - \log(1 + \lambda\varepsilon).$$

To get a friendlier bound, we use the inequality (exercise)

$$\log(1+x) \leq x - \frac{x^2}{2(1+x)}, \tag{2.10}$$

for $x \geq 0$, which gives

$$g(t^*) \geq \frac{\lambda^2\varepsilon^2}{2(1+\lambda\varepsilon)}.$$

For the lower tail, we have, for $0 < \varepsilon < \frac{1}{\lambda}$

$$h(t) = -t\left(\frac{1}{\lambda} - \varepsilon\right) - \log \frac{\lambda}{\lambda + t},$$

which attains its maximum at $t^* := \frac{\varepsilon\lambda^2}{1-\varepsilon\lambda}$, for which

$$h(t^*) = -\lambda\varepsilon - \log(1 - \lambda\varepsilon).$$

Now we use the inequality (exercise)

$$\log(1-x) \leq -x - \frac{x^2}{2}, \tag{2.11}$$

to obtain

$$h(t^*) \geq \frac{\varepsilon^2\lambda^2}{2}.$$

Setting $\varepsilon = \frac{\delta}{\lambda}$, we have obtained:

Proposition 2.4. *Let* $X_1, X_2, \dots, X_n \sim \mathrm{Exp}(\lambda)$ *independent. Then*
Upper Tail: for every $\delta > 0$

$$\mathrm{P}\left(\overline{X}_n \geq \frac{(1+\delta)}{\lambda}\right) \leq e^{-n\frac{\delta^2}{2(1+\delta)}};$$

Lower Tail: for every $0 < \delta < 1$

$$\mathrm{P}\left(\overline{X}_n \leq \frac{(1-\delta)}{\lambda}\right) \leq e^{-n\frac{\delta^2}{2}}.$$

Remark 2.5. Note that the above bounds do not depend on λ. This is due to the fact that Exponential random variables are *scale-invariant*:

$$X \sim \mathrm{Exp}(\lambda) \iff \frac{X}{\lambda} \sim \mathrm{Exp}(1).$$

□

2.1.3 Square of Gaussians

For later use we consider the following further example. Let $Z_1, Z_2, \dots, Z_n \sim \mathrm{N}(0,1)$ independent, and set $X_i := Z_i^2$. We first compute the moment generating function

$$m(t) = \mathrm{E}\left[e^{tX_i}\right] = \frac{1}{2\pi}\int e^{tx^2} e^{-\frac{x^2}{2}}\, dx = \frac{1}{2\pi}\int e^{-\frac{x^2}{2}(1-2t)} dx.$$

This last integral is finite if $t < \frac{1}{2}$. Recalling that

$$\int e^{-\frac{x^2}{2\sigma^2}}\, dx = \sigma\sqrt{2\pi}$$

and using it for $\sigma^2 = \frac{1}{1-2t}$, we obtain

$$m(t) = \sqrt{\frac{1}{1-2t}}.$$

Thus, noting that $\mu = \mathrm{E}(X_i) = 1$,

$$g(t) = t(1+\varepsilon) - \frac{1}{2}\log\left(\frac{1}{1-2t}\right),$$

which is maximized at $t^* = \frac{\varepsilon}{2(1+\varepsilon)}$, giving

$$g(t^*) = \frac{\varepsilon}{2} - \frac{1}{2}\log(1+\varepsilon) \geq \frac{1}{2}\left[\varepsilon - \left(\varepsilon - \frac{\varepsilon^2}{2(1+\varepsilon)}\right)\right] = \frac{\varepsilon^2}{4(1+\varepsilon)},$$

where we have used (2.10). For the lower tail, choosing $0 < \varepsilon < 1$,

$$h(t) = -t(1-\varepsilon) - \frac{1}{2}\log\frac{1}{1+2t},$$

which attains the maximum at $t^* = \frac{\varepsilon}{2(1-\varepsilon)}$, for which

$$h(t^*) = -\frac{\varepsilon}{2} - \frac{1}{2}\log(1-\varepsilon) \geq \frac{1}{2}\left[-\varepsilon - \left(-\varepsilon - \frac{\varepsilon^2}{2}\right)\right] = \frac{\varepsilon^2}{4},$$

wher we have used (2.11). Summing up:

Proposition 2.6. *Let $X_1, X_2, \dots, X_n$ be i.i.d., each being the square of a standard normal. Then*

Upper Tail: for every $\varepsilon > 0$

$$\mathrm{P}\left(\overline{X}_n \geq 1+\varepsilon\right) \leq e^{-n\frac{\varepsilon^2}{4(1+\varepsilon)}};$$

Lower Tail: for every $0 < \varepsilon < 1$

$$\mathrm{P}\left(\overline{X}_n \leq 1-\varepsilon\right) \leq e^{-n\frac{\varepsilon^2}{4}}.$$

These two estimates imply that, for $0 < \varepsilon < 1$,

$$\mathrm{P}\left(|\overline{X}_n - 1| \geq \varepsilon\right) \leq 2e^{-n\frac{\varepsilon^2}{8}}.$$

2.2 Central Limit Theorem

Computations with random variables may greatly benefit from approximating their distribution with a "simpler" one. To have a measure of the goodness of the approximation, it is essential to have a notion of convergence for distributions of random variables. Here we only deal with the case of real random variables.

Definition 2.7. Let $(X_n)_{n\geq 1}$ be a sequence of real random variables. We say that X_n *converges* to a random variable X *in distribution* if for each function $f : \mathbb{R} \to \mathbb{R}$ continuous and bounded, we have

$$\lim_{n\to+\infty} \mathrm{E}[f(X_n)] = \mathrm{E}[f(X)].$$

This definition of convergence in distribution can be characterized in terms that are often more useful in applications. The proof is omitted.

Proposition 2.8. *A sequence $(X_n)_{n\geq 1}$ of real random variables converges in distribution to X if and only if*

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$$

for all x in which F_X is continuous.

Recalling that, knowing F_X, we can compute $\mathrm{P}(X \in I)$ for any interval I, Proposition 2.8 says that if $X_n \to X$ in distribution, then

$$\lim_{n\to+\infty} \mathrm{P}(X_n \in I) = \mathrm{P}(X \in I)$$

for any interval I whose endpoints are points in which F_X is continuous.

Convergence in distribution can be established in various ways, but the most effective involves characteristic functions.

Theorem 2.9. *A sequence $(X_n)_{n\geq 1}$ of real random variables converges in distribution to X if and only if*

$$\lim_{n\to+\infty} \varphi_{X_n}(u) = \varphi_X(u) \tag{2.12}$$

for all $u \in \mathbb{R}$.

Sketch of the proof. Suppose that $X_n \to X$ in distribution. Since cosine and sine are bounded continuous functions, it follows form definition 2.7 that

$$\begin{aligned}\lim_{n\to+\infty} \varphi_{X_n}(u) &= \lim_{n\to+\infty} \mathrm{E}[\cos(uX_n)] + i \lim_{n\to+\infty} \mathrm{E}[\sin(uX_n)] \\ &= \mathrm{E}[\cos(uX)] + i\,\mathrm{E}[\sin(uX)] = \varphi_X(u).\end{aligned}$$

Conversely, suppose (2.12) holds, and let f be a bounded continuous function. Approximating f by trigonometric functions as in (1.49),

$$f(x) \simeq \sum_{k=1}^{m} a_k e^{iu_k x}$$

we have that

$$\mathrm{E}[f(X_n)] \simeq \sum_{k=1}^{m} a_k \varphi_{X_n}(u_k) \overset{n\to\infty}{\longrightarrow} \sum_{k=1}^{m} a_k \varphi_X(u_k) \simeq \mathrm{E}[f(X)].$$

The actual proof consists in making rigorous these approximations. □

Theorem 2.9 allows an elegant proof of the most fundamental theorem in Probability.

Theorem 2.10. (Central Limit Theorem) *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d., square integrable random variables, with mean μ and variance $\sigma^2 > 0$. Let*

$$Z_n := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}.$$

Then Z_n converges in distribution to $Z \sim \mathrm{N}(0,1)$.

Proof. Note that

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i,$$

where

$$Y_i := \frac{X_i - \mu}{\sigma}$$

are independent and identically distrubuted random variables. By Proposiiton 1.68

$$\varphi_{Z_n}(u) = \prod_{i=1}^n \varphi_{Y_i}(u/\sqrt{n}) = \left(\varphi_{Y_1}(u/\sqrt{n})\right)^n.$$

We now take the Taylor expansion

$$\varphi_{Y_1}(x) = 1 + \varphi_{Y_1}'(0)x + \frac{1}{2}\varphi_{Y_1}''(0)x^2 + o(x^2)$$

with $x = \frac{u}{\sqrt{n}}$, and use (1.51) and the fact that $\mathrm{E}(Y_1) = 0$, $\mathrm{E}(Y_1^2) = 1$; we get

$$\varphi_{Z_n}(u) = \left(1 - \frac{u^2}{2n} + o(1/n)\right)^n \overset{n\to\infty}{\longrightarrow} e^{-\frac{u^2}{2}} = \varphi_Z(u),$$

and the conclusion follows from Theorem 2.9.

□

The Central Limit Theorem states that the sum of i.i.d. random variables, when suitably rescaled, have an approximate normal distribution (which is *exactly* true if the summand are normal). Conditions for the validity of the Central Limit Theorem can be greatly relaxed, allowing e.g. the X_i to have different distributions. This explain the ubiquity of normal distribution in the real world: whenever randomness comes from several small and independent contributions, its distribution is nearly normal!

2.3 “Strange” behaviors of high dimensional data

So far in this Chapter we have investigated the behavior of a sample of independent data as the size of the sample goes to infinity. We now assume the data belong to a high dimensional space, and study their properties as the dimension of this space goes to infinity. In particular, we investigate here the behavior of high dimensional normal random vectors. Since all normals can be obtained as linear transform of a normal with identity covariance matrix, we state most of the resulta for this particular case. Many results can be adapted to more general cases.

Consider then $X \sim \mathrm{N}(0, \mathbb{I}_d)$, where d is the dimension of the vector. One first property, that easily follows from Proposition 1.74, is the so-called *spherical symmetry* of the distribution of X: if A is a $d \times d$ rotation matrix, i.e. $AA^\top = \mathbb{I}_d$, then AX has the same distribution as X.

2.3.1 *Gaussian annuli*

Our first result states that the norm of X is concentrated around its mean, if the dimension of X is large. We recall that

$$\|X\|^2 = \langle X, X \rangle = \sum_{i=1}^{d} X_i^2$$

and, since $X_i \sim \mathrm{N}(0,1)$, it follows that $\mathrm{E}\left[\|X\|^2\right] = d$.

Theorem 2.11. *(*Gaussian Annulus Theorem*) For any constant $c > 0$*

$$\mathrm{P}\left[\left|\|X\|^2 - d\right| \geq c\sqrt{d}\right] \leq 2e^{-c^2/8}.$$

Proof. Just apply Proposition 2.6 with $n = d$ and $\varepsilon = \frac{c}{\sqrt{d}}$. □

Thus, for large d, with high probability the deviations of $\|X\|^2$ form d are at most of order $\sqrt{d} \ll d$, i.e. X belongs to a thin (comparatively to its radius) annulus around the sphere of radius $\sqrt{d}$. We express this by writing

$$\|X\|^2 = d + O(\sqrt{d}).$$

2.3.2 *Near orthogonality*

If x and y are two d-dimensional vectors, we say they are orthogonal if $\langle x, y \rangle = 0$. Since

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle,$$

it follows that orthogonality is equivalent to the statement

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2.$$

Given x and y we can consider the corresponding unit vectors (i.e. vectors of norm 1) $\frac{x}{\|x\|}$ and $\frac{y}{\|y\|}$. Since

$$\langle x,y\rangle = \|x\|\,\|y\|\langle \frac{x}{\|x\|}, \frac{y}{\|y\|}\rangle,$$

we say the x and y are *nearly orthogonal* if $\langle x,y\rangle$ is much smaller than $\|x\|\,\|y\|$.

Theorem 2.12. *Let $y \in \mathbb{R}^d$, and $X \sim \mathrm{N}(0,\mathbb{I}_d)$. Then for all $c > 0$*

$$\mathrm{P}\left(|\langle X,y\rangle| > c\|y\|\right) \leq \frac{2}{\sqrt{2\pi}} \frac{e^{-c^2}}{c}. \tag{2.13}$$

Moreover X and y are nearly orthogonal with high probability (i.e. for a probability close to one) for d large.

Proof. Since the distribution of X is left invariant by rotations, we can rotate y to make it parallel to the first reference axis. So there is no loss of generality to assume $y_1 = \|y\|$, and $y_i = 0$ for $i > 1$. Thus $\langle X,y\rangle = X_1\|y\|$. Therefore

$$\mathrm{P}\left(|\langle X,y\rangle| > c\|y\|\right) = \mathrm{P}(|X_1| > c) = 2\,\mathrm{P}(X_1 > c) = \frac{2}{\sqrt{2\pi}} \int_c^{+\infty} e^{-\frac{x^2}{2}}\,dx.$$

Thus, setting

$$\alpha(c) := \int_c^{+\infty} e^{-\frac{x^2}{2}}\,dx$$

and

$$\beta(c) := \frac{e^{-c^2}}{c},$$

all we need to prove is that $\alpha(c) \leq \beta(c)$ for all $c > 0$. Since

$$\lim_{c\to+\infty} \alpha(c) = \lim_{c\to+\infty} \beta(c) = 0,$$

it is enough to show that

$$\alpha'(c) \geq \beta'(c)$$

for all $c > 0$, which is checked very easily (left to the reader). This shows that, with high probability, $|\langle X,y\rangle|$ is bounded by $c\|y\|$ for a sufficiently large c that does not depend on d or on the choice of y. Since, on the other hand, we have seen that $\|X\|$ is $\sqrt{d} + O(d^{1/4})$, it follows that $|\langle X,y\rangle|$ is much smaller than $\|X\|\,\|y\|$, i.e. X and y are nearly orthogonal. □

Remark 2.13. Does (2.13) holds if we replace y by a random vector Y? In general it does not: take, as an example, $Y = X$. It is intuitive, but not so easy to prove (we omit any further details), that it holds if Y and X are *independent*, □

2.3.3 Separating Gaussians

It often happens in practice that collected data are sampled from two or more different distributions; this occurs, for instance, when collecting data from populations in which some *discrete* unknown factor affects the distribution of data. In these case one looks for algorithms for separating data coming from different distributions. We consider here data coming form two different d-dimensional normal distributions, with the same covariance matrix but different means; after a suitable linear transformation, we reduce to the case in which the covariance matrix is $\mathbb{I}_d$.

So we consider two random variables X and Y, that we know have either distribution $\mathrm{N}(\mu, \mathbb{I}_d)$ or distribution $\mathrm{N}(\nu, I_d)$. We show here that for d large one can detect whether X and Y have or not the same distribution, provided μ and ν are sufficiently separated. Note that if ν is obtained from μ by adding a constant to each component, then $\|\mu-\nu\|^2$ is of order d. Here we require much less: that $\|\mu-\nu\|^2$ is much larger that $\sqrt{d}$.

Suppose first that X and Y come from the same distribution. Then $X - Y = X' - Y'$ with $X'_i = X_1 - \mathrm{E}(X_i)$ (same for Y'), so that X' and Y' are independent $\mathrm{N}(0, \mathbb{I}_d)$. By Theorem 2.12 and Remark 2.13, with high probability

$$|\langle X', Y'\rangle| \leq c\|Y'\|.$$

Moreover, by Theorem 2.11, $\|Y'\|$ is close to $\sqrt{d}$ up to deviations of order $d^{1/4}$. Thus, using again Theorem 2.11,

$$\|X-Y\|^2 = \|X'\|^2 + \|Y'\|^2 - 2\langle X', Y'\rangle = 2d + O(\sqrt{d}).$$

Suppose, instead, that $X \sim \mathrm{N}(\mu, \mathbb{I}_d)$ and $Y \sim \mathrm{N}(\nu, I_d)$. Then

$$X - Y = X' - Y' + (\mu - \nu).$$

Again by Theorem 2.12 and Remark 2.13, with high probability the three vectors X', Y' and $\mu - \nu$ are nearly orthogonal, so by Theorem 2.11

$$\|X-Y\|^2 = \|X'\|^2 + \|Y'\|^2 + \|\mu-\nu\|^2 + O(\sqrt{d}) = 2d + \|\mu-\nu\|^2 + O(\sqrt{d}).$$

If $\|\mu-\nu\|^2 \gg \sqrt{d}$ then this distance is much larger than in the case where the variables have the same distribution. This fact may be used to design statistical test for detecting whether X and Y have or not the same distribution, as well as, in the case of several data, to cluster them in groups sampled from the same distribution.

2.4 An application to Machine Learning

Classification is one of the core problems in Machine Learning. Assume we are given a set S, that we call *instance space*. For instance, the elements of S are email messages, or a set of biometrical indexes of an individual; typically, one choose $S = \{0,1\}^d$ or $S = \mathbb{R}^d$. We receive, or collect, *data* from S; in mathematical terms, we represent these data as a sequence $(X_j)_{j\geq 1}$ of i.i.d, S-valued random variables, with common distribution μ. We also consider a subset $C \subseteq S$; in the examples above, C could be the set of "SPAM" email, or the set of individuals affected by a given pathology. Suppose we have observed n data $(X_1, X_2, \dots, X_n)$, that is called *training sequence*; these date are "fully understood", i.e. for each $i = 1, 2, \dots, n$ we know whether $X_i \in C$ or $X_i \in C^c$. The aim is to determine, on the basis of the training data, a "simple" rule that will be use to decide whether a future datum, say X_{n+1}, is in C or not.

In order to formalize this notion, let $\mathscr{H}$ be a family of subsets of S, i.e. $\mathscr{H} \subseteq \mathscr{P}(S)$. An element H of $\mathscr{H}$ is called *classification rule*. In principle one could consider $\mathscr{H} = \mathscr{P}(S)$, but in practice a much smaller set of possible classification rules is desirable. For $H \in \mathscr{H}$ we define the *true error* of H as

$$err(H) := \mathrm{P}(X_{n+1} \in H \Delta C) = \mu(H \Delta C),$$

where $H \Delta C = (H \setminus C) \cup (C \setminus H)$. In other words $err(H)$ is the probability of having a wrong classification of X_{n+1} if we classify using H in place of C. The *training error* of H is the fraction of the training data that get wrong classification using H in place of C:

$$err_n(H) = \frac{|\{i = 1, 2, , \dots, n \text{ such that } X_i \in H \Delta C\}}{n}.$$

A very natural way of choosing the best classification rule is to choose the $H \in \mathscr{H}$ that minimizes the training error. It may happen, however, that minimizing the training error does not correspond to mimimizing the true error and, although the training error is low, the true error may be large. This phenomenon is called *overfitting*.

We now show that, provided the size of the training sequence is sufficiently large, the training and the true error are acually close, with high probability, *for all* classification rule, so overfitting is not an issue. This result assume $\mathscr{H}$ to be a finite set. With a little more effort various results with $\mathscr{H}$ infinite can be also obtained.

Theorem 2.14. *Let $\varepsilon, \delta > 0$, and assume the size of the training sequence n is such that*

$$n \geq \frac{2}{\varepsilon^2}\left(\log|\mathscr{H}| + \log\left(\frac{2}{\delta}\right)\right). \tag{2.14}$$

Then

$$\mathrm{P}(|err(H) - err_n(H)| \geq \varepsilon \textit{ for some } H \in \mathscr{H}) \leq \delta.$$

Proof. Define that random variables:

$$Y_i^H := \mathbf{1}_{\{X_i \in H \Delta C\}}.$$

Note that the Y_i^H are i.i.d. Bernoulli random variables, with $E(Y_i^H) = err(H)$ and $\overline{Y}_n^H = err_n(H)$. So, by the Chernoff bound (2.9)

$$\mathrm{P}(|err(H) - err_n(H)| \geq \varepsilon) \leq 2e^{-n\frac{\varepsilon^2}{2}}.$$

Therefore

$$\mathrm{P}(|err(H) - err_n(H)| \geq \varepsilon \text{ for some } H \in \mathscr{H}) \leq |\mathscr{H}| 2e^{-n\frac{\varepsilon^2}{2}}.$$

Taking n as in (2.14), this last expression is less or equal to δ.

□

Chapter 3
Markov chains and Markov Chain Monte Carlo

Markov Chains are sequences of random variables with a mutual dependence of a special form that generalizes the i.i.d. sequences we have seen so far. Markov Chains are used to model real dynamics, with applications ranging from Physics to Economics, from Biology to Sociology. They are also used as computational tools to solve "hard problems", which may or may not come from probability.

3.1 Markov property

In this Chapter we denote by S a finite set, that we call the *state space*. Whenever we have a sequence $x = (x_n)_{n \geq 0}$ with elements in S, and $0 \leq m < n$, we set

$$x_m^n := (x_m, x_{m+1}, \dots, x_{n-1}, x_n).$$

Definition 3.1. A *Markov Chain* with values in S is a sequence $(X_n)_{n \geq 0}$ of S-valued random variables which have the following property: for each $n \geq 0$ and each sequence $x = (x_n)_{n \geq 0}$ with elements in S

$$\mathrm{P}(X_{n+1} = x_{n+1} | X_0^n = x_0^n) = \mathrm{P}(X_{n+1} = x_{n+1} | X_n = x_n). \tag{3.1}$$

The property (3.1) is called *Markov Property*. It states that the random sequence have "short memory": if the trajectory of the chain is known up to time n, and we want to compute the conditional probability of reaching a given state at time $n+1$, only the state at time n matters.

The conditional probabilities $\mathrm{P}(X_{n+1} = x | X_n = y)$ are called *one step transition probabilities*, or simply transition probabilities. We observe that with the transition probabilities we can recover the joint distribution of $(X_0, X_1, \dots, X_n) = X_0^n$ for all n, provided the *initial distribution*, i.e. the distribution of X_0, is known. To see this observe that

$$\begin{aligned}\mathrm{P}(X_0^n = x_0^n) &= \mathrm{P}(X_n = x_n | X_0^{n-1} = x_0^{n-1})\,\mathrm{P}(X_0^{n-1} = x_0^{n-1}) \\ &= \mathrm{P}(X_n = x_n | X_{n-1} = x_{n-1})\,\mathrm{P}(X_0^{n-1} = x_0^{n-1}),\end{aligned}$$

where the Markov property has been used in the last step. Iterating the same procedure one gets

$$\mathrm{P}(X_0^n = x_0^n) = \mathrm{P}(X_n = x_n | X_{n-1} = x_{n-1})\,\mathrm{P}(X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}) \cdots \mathrm{P}(X_1 = x_1 | X_0 = x_0)\,\mathrm{P}(X_0 = x_0), \tag{3.2}$$

which only depends on the one step transition probabilities and the initial distribution.

A Markov chain is said to be *homogeneous* if the one step transition probability $\mathrm{P}(X_{n+1} = x | X_n = y)$ depends on x and y buy *not* on n: in other words, the updating rules of the chain are constant in time. In this case we set

$$P_{yx} := \mathrm{P}(X_{n+1} = x | X_n = y).$$

These are the entries of a $|S| \times |S|$ matrix P, called the *transition matrix* of the chain. Thus, by (3.2), we can write

$$\mathrm{P}(X_0^n = x_0^n) = \mathrm{P}(X_0 = x_0) \prod_{k=0}^{n-1} P_{x_{k+1},x_k}. \tag{3.3}$$

Note that the matrix P has nonnegative entries, and

$$\sum_{x \in S} P_{yx} = \sum_{x \in S} \mathrm{P}(X_{n+1} = x | X_n = y) = 1,$$

i.e. the sum of the entries in each row is one. A matrix with these properties is called a *stochastic matrix*. It can be shown that for any stochastic matrix there exist a Markov Chain having it as transition matrix.

Consider now a Homogeneous Markov Chain with transition probability P, and the *two steps* transition probabilities $\mathrm{P}(X_{n+2} = x | X_n = y)$. Note that

$$\begin{aligned}\mathrm{P}(X_{n+2} = x | X_n = y) &= \sum_{z \in S} \mathrm{P}(X_{n+2} = x, X_{n+1} = z | X_n = y) \\ &= \sum_{z \in S} \mathrm{P}(X_{n+2} = x | X_{n+1} = z, X_n = y)\,\mathrm{P}(X_{n+1} = z | X_n = y) \\ &= \sum_{z \in S} \mathrm{P}(X_{n+2} = x | X_{n+1} = z)\,\mathrm{P}(X_{n+1} = z | X_n = y) \\ &= \sum_{z \in S} P_{yz} P_{zx},\end{aligned}$$

where in the second line we have used that fact that, given three events A, B, C,

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

easy to verify, and in the third line we have used the Markov Property. Note that $\sum_{z \in S} P_{yz} P_{zx}$ is the entry yx of the matrix $P \cdot P = P^2$. Thus P^2 is the two steps transition matrix. Iterating the same argument, we can show that the *k step* transition probabilities are

$$\mathrm{P}(X_{n+k} = x | X_n = y) = P^k_{yx},$$

where P^k is the k-fold product of P. In particular, for every $n \geq 1$

$$P^n_{yx} = \mathrm{P}(X_n = x | X_0 = y).$$

Note that, applying the Formula of total probability (1.8),

$$\mathrm{P}(X_n = x) = \sum_{y \in S} \mathrm{P}(X_n = x | X_0 = y)\, \mathrm{P}(X_0 = y). \tag{3.4}$$

It follows that, setting $\nu^{(n)}_y := \mathrm{P}(X_n = y)$ and interpreting each $\nu^{(n)} = (\nu^{(n)}_y)_{y \in S}$ as a *row vector*, the identity

$$\nu^{(n)} = \nu^{(0)} P^n \tag{3.5}$$

holds for all $n \geq 0$. In other words, the matrix multiplication of the initial distribution with the n steps transition matrix returns the distribution of the chain at time n.

3.2 Stationary distributions

Here and in what follows we continue to identify distributions on S with row vectors with nonnegative entries summing up to one.

Definition 3.2. A distribution π on S is said *stationary* for a homogeneous Markov Chain with transition matrix P if

$$\pi P = \pi,$$

which implies that if $X_0 \sim \pi$ then $X_n \sim \pi$ for all $n \geq 0$.

From an algebraic viewpoint, a stationary distribution can be seen as a left-eigenvector of the matrix P corresponding to the eigenvalue 1. It is easy to show that 1 is indeed an eigenvalue of P. To see this, since P and $P^\top$ have the same eigenvalues, it is enough to show that the equation $Pv = v$ admits a nonzero solution; but, setting $v_x = 1$ for all $x \in S$, $Pv = v$ follows from the fact that the sum of the entries of any row of P is one.

Note that the existence of a nonzero solution for $\pi P = \pi$ does not guarantee that a stationary distribution exists, as eigenvectors may have negative entries. It turns out, however, that any Markov Chain has at least one stationary distribution

Proposition 3.3. *For any stochastic matrix P there exists at least a stationary distribution for a homogeneous Markov Chain with transition matrix P.*

Proof. Given any distribution ν, set

$$\mu_n := \frac{1}{n}\sum_{j=1}^{n} \nu P^j.$$

Note that μ_n is the arithmetic mean of the distribution of the chain at time j, assuming the initial distribution is ν. It follows that μ_n is itself a distribution, i.e. it belongs to the set

$$\mathscr{C}_S : \left\{ v \in \mathbb{R}^S : v_x \geq 0 \; \forall \, x \in S, \, \sum_{x \in S} v_x = 1 \right\}.$$

$\mathscr{C}_S$ is a closed and bounded subset of a finite-dimensional space, so it is compact. This implies that the sequence $(\mu_n)_{n\geq 1}$ has a convergent subsequence μ_{n_k}, i.e. for all $x \in S$

$$\lim_{k\to+\infty} (\mu_{n_k})_x =: \pi_x,$$

for some distribution π. We show that π is stationary:

$$\begin{aligned} \pi P - \pi &= \lim_{k\to+\infty} \frac{1}{n_k}\sum_{j=1}^{n_k} \nu P^{j+1} - \lim_{k\to+\infty} \frac{1}{n_k}\sum_{j=1}^{n_k} \nu P^j \\ &= \lim_{k\to+\infty} \frac{1}{n_k}\left[\nu P^{n_k+1} - \nu P\right] = 0 \end{aligned}$$

where in the last step we have used the fact that νP^{n_k+1} and νP have entries between 0 and 1, so they go to zero when divided by n_k. □

Looking for stationary distributions is in general a hard task. Note that finding stationary distributions amounts to solve a system of linear equations, whose dimension is the number of elements of the state space S. When $|S|$ is very large this can be computationally problematic; it is often difficult even to verify that a given candidate distribution is stationary. The following sufficient condition provides a useful criterion.

Definition 3.4. A distribution π is said to be *reversible* for a Markov chain with transition probability P if for $x, y \in S$ the following condition, called *detailed balance* condition, holds:

$$\pi_y P_{yx} = \pi_x P_{xy}. \tag{3.6}$$

Proposition 3.5. *A reversible distribution is stationary.*

Proof. We have

$$(\pi P)_x = \sum_{y \in S} \pi_y P_{yx} = \sum_{y \in S} \pi_x P_{xy} = \pi_x \sum_{y \in S} P_{xy} = \pi_x,$$

where in the second equality we have used detailed balance, and in the last the fact the entries of a row of P sum up to one. □

It is easy to show that not all stationary distribution for a transition matrix P are necessarily reversible. Consider, for instance, the transition matrix on $S = \{1,2,3\}$

$$P = \begin{pmatrix} 2/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 0 & 2/3 & 1/3 \end{pmatrix}.$$

Note that $\pi = (1/3, 1/3, 1/3)$ is stationary, but not reversible, as

$$\pi_1 P_{12} = 0 \neq \pi_2 P_{21} = 1/9.$$

π is actually the *unique* stationary distribution for this chain, as next results will show.

3.3 Irreducibility and uniqueness of the stationary distribution

Definition 3.6. A *(directed) graph* is a pair (S,E) where S is a finite or countable set and E is a subset of $S \times S$.

The elements of S are said *nodes* of the graph, while that elements of E are called *links* or *edges*: if $(x,y) \in E$ we say there is a link from x to y. We say that a graph is *connected* if for any choice of two nodes $x \neq y$, they are joined by a sequence of links, i.e. there exist links of the form

$$(x,x_1),\ (x_1,x_2),\ \dots,(x_{n-2},x_{n-1}),\ (x_{n-1},y).$$

Connectedness of a graph has the following useful characterization, whose proof is left as exercise.

Proposition 3.7. *A graph (S,E) is connected if and only if for every nonempty and proper subset A of S there is a link from an element of A and an element of A^c.*

To a transition matrix P on the state space S we can associate a graph by defining the set of links E as follows:

$$(x,y) \in E \iff P_{xy} > 0.$$

In other words, we say there is a link from x to y if the chain can go from x to y in one step with strictly positive probability.

Definition 3.8. A transition probability P is said to be *irreducible* if its associated graph is connected.

It is not hard to see (exercise) that P is irreducible if and only if for every $x \neq y$ there exist $n \geq 1$ such that $P^n_{xy} > 0$. Thus, P is irreducible if for every $x \neq y$ there exist $n \geq 1$ such that the chain can go from x to y in n steps with strictly positive probability.

Theorem 3.9. *Let P be a irreducible transition probability. Then:*

(a) there exists a unique stationary distribution π;

(b) the stationary distribution gives strictly positive probability to all states, i.e. $\pi_x > 0$ for all $x \in S$;

(c) for any initial distribution ν

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} \nu P^j = \pi.$$

Proof. (a) We know by Proposition 3.3 that one stationary distribution π exists. The column vector $v := \pi^\top$ is a solution of the linear, homogeneous linear system

$$(P - \mathbb{I})^\top v = 0, \tag{3.7}$$

where $\mathbb{I}$ is the identity matrix. We know from linear algebra that the set of solutions of (3.7) form a vector space whose dimension is $d = |S| - \operatorname{rank}((P - \mathbb{I})^\top)$. If we show that $d = 1$ then we are done: in this case all solutions are of the form kv for $k \in \mathbb{R}$ and, among them, v is the only one which is a distribution. The rank of a matrix equals the rank of its transpose: so it is enough to show that the equation

$$(P - \mathbb{I})w = 0 \tag{3.8}$$

have solutions forming a vector space of dimension one. Note that, since P is a stochastic matrix, one solution of (3.8) is the vector $\mathbf{1}$ whose components are all equal to one. If $d > 1$ there must be another solution w of (3.8) which is not of the form $k\mathbf{1}$; thus amounts to say that the components of w are not all equal. So let α be the maximum of the components of w:

$$\alpha := \max\{w_x : x \in S\},$$

and let $A := \{x \in S : w_x = \alpha\}$. Note that $A \neq \emptyset$ and $A^c \neq \emptyset$. By Proposition 3.7 there exists $x \in A$ and $z \in A^c$ with $P_{xz} > 0$. We have

$$\alpha = w_x = (Pw)_x = \sum_{y \in S} P_{xy} w_y < \sum_{y \in S} P_{xy} \alpha = \alpha,$$

where the strict inequality comes form the fact that $P_{xz} w_z < P_{xz} \alpha$. So we have obtained $\alpha < \alpha$; the contradiction comes from having assumed $d > 1$.

(b) Set $A := \{x \in S : \pi_x = 0\}$; clearly $A \neq S$. Suppose $A \neq \emptyset$. Then there exist $x \in A$ and $z \in A^c$ with $P_{zx} > 0$. Since π is stationary

$$0 = \pi_x = \sum_{y \in S} \pi_y P_{yx} \geq \pi_z P_{zx} > 0,$$

which is a contradiction. So $A = \emptyset$.

(c) Set

$$\mu_n := \frac{1}{n} \sum_{j=1}^{n} \nu P^j. \tag{3.9}$$

In the proof of Proposition 3.3 we have seen that any convergent subsequence of (μ_n) converges to a stationary distribution. Since there is a unique stationary distribution, this implies that the sequence (μ_n) has a unique limit point, so it is convergent.

□

It is easy to exhibit a random variable with distribution μ_n given in (3.9). Let $(X_k)_{k \geq 1}$ be a Markov Chain with transition matrix P and initial distribution ν. Given $n \geq 1$ we select at random a time T_n between 1 and n, independently of the evolution of the chain. This means that T_n is a random variable independent of $X_0, X_1, \ldots, X_n$, and $\mathrm{P}(T_n = k) = \frac{1}{n}$ for all $k = 1, 2, \ldots, n$. Then X_{T_n} has distribution μ_n. Indeed

$$\begin{aligned}
\mathrm{P}(X_{T_n} = x) &= \sum_{k=1}^{n} \mathrm{P}(X_{T_n} = x, T_n = k) \\
&= \sum_{k=1}^{n} \mathrm{P}(X_k = x, T_n = k) \\
&= \sum_{k=1}^{n} \mathrm{P}(X_k = x)\, \mathrm{P}(T_n = k) \\
&= \frac{1}{n} \sum_{k=1}^{n} \mathrm{P}(X_k = x) = \frac{1}{n} \sum_{k=1}^{n} (\nu P^k)_x = (\mu_n)_x.
\end{aligned}$$

Thus, if we have a good algorithm for generating trajectories of a Markov Chain, we can *approximately sample* form the stationary distribution, i.e. generate a random variable whose distribution is close to the stationary distribution. This is the basis of Markov Chain Monte Carlo.

It is natural to conjecture, for a irreducible Markov Chain, that $\mu P^n \to \pi$ as $n \to +\infty$. This is not true in general. For instance, consider the transition matrix on the state space $\{1, 2\}$

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{3.10}$$

The unique stationary distribution is $\pi = (1/2, 1/2)$. If we take ν such that $\mu_1 = p \neq 1/2$, then $(\nu P^n)_1 = p$ for n even and $= 1 - p$ for n odd, so νP^n does not have limit as $n \to +\infty$. In order to rule out similar periodic-like behaviors some assumptions are needed.

Definition 3.10. Let P be a transition matrix and $x \in S$. The *period* of x is defined as the largest common divisor of the integers n such that $P^n_{xx} > 0$.

Proposition 3.11. *In an irreducible Markov Chain all states have the same period.*

We omit the rather simple proof of this result. Note that, in the example (3.10), states have period 2.

Definition 3.12. A Markov chain is said to be *aperiodic* if all states have period 1.

Theorem 3.13. *Let P be the transition matrix of a irreducible, aperiodic Markov Chains, and denote by π its unique stationary distribution. Then for every initial distribution ν*

$$\lim_{n \to +\infty} \nu P^n = \pi.$$

The proof of this result is not given here. We note however that given a irreducible Markov Chain with stationary distribution π which is *not* aperiodic, it is simple to modify its transition matrix P to make it aperiodic. For $0 < \varepsilon < 1$ set

$$P_\varepsilon := \varepsilon \mathbb{I} + (1 - \varepsilon) P.$$

It is easy to see that P_ε irreducible and it has stationary distribution π. Moreover it is aperiodic, as $P_{xx} \geq \varepsilon > 0$ for all $x \in S$. It is called the *lazy* version of P, where that name is motivated by the fact that with $\varepsilon > 0$ we increase the probability that the chain does not move in one time step.

3.4 Markov Chain Monte Carlo

In this section we address the following problem: given a (finite) set S and a distribution π on S, generate a sequence $(X_n)_{n \geq 1}$ of random variables such that the following law of large numbers holds for every function $f : S \to \mathbb{R}$:

$$\lim_{n\to+\infty} \mathrm{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \pi[f]\right| \geq \varepsilon\right] = 0 \tag{3.11}$$

for all $\varepsilon > 0$, where

$$\pi[f] := \sum_{x\in S} f(x)\pi(x).$$

Note that (3.11) would follow from Theorem 2.2 if the X_n were independent and with distribution π. However, generating independent samples for a given distribution could be very costly. Alternatively to independent samples, one could generate dependent samples forming a Markov Chain. Next Theorem shows that irreducible Markov Chains indeed do the job; later, in illustrating examples, we will see that Markov Chains could be computationally much more economic than independent samples.

Theorem 3.14. *Let $(X_n)_{n\geq 0}$ be an irreducible Markov Chain with invariant distribution π. Then for every $f : S \to \mathbb{R}$ (3.11) holds.*

Idea of the proof. Denote by ν the distribution of X_0 and set

$$\mu_n := \frac{1}{n}\sum_{j=1}^{n} \nu P^j.$$

Note that

$$\mu_n[f] = \mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right].$$

By Theorem 3.13, μ_n converges to π, so

$$\pi[f] = \lim_{n\to+\infty} \mu_n[f].$$

Therefore, for a given $\varepsilon > 0$, if n is large enough, $|\mu_n[f] - \pi[f]| < \frac{\varepsilon}{2}$, so

$$\mathrm{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \pi[f]\right| \geq \varepsilon\right] \leq \mathrm{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mu_n[f]\right| \geq \frac{\varepsilon}{2}\right] \leq \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right)}{\varepsilon^2/4}.$$

Thus the conclusion follows if we prove that

$$\lim_{n\to+\infty} \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right) = 0.$$

The proof of this last fact requires a refinement of Theorem 3.13, and it is omitted here. □

We have observed before that a random variable with distribution π_n is given by X_{T_n} where T_n is uniformly distributed in $\{1,2,\ldots,n\}$ and independent on the Markov

Chain. In particular, by generating a Markov Chain with stationary distribution π we can generate a sample from a distribution close to π.

We now illustrate some prototypical problems in which generating random variables with the above properties is relevant.

3.4.1 Examples

3.4.1.1 Bayesian Statistics: computing a-posteriori probabilities

Consider a continuous *statistical model* i.e. a family of densities $f(x;\theta)$ of real, absolutely continuous random variables, depending on a parameter $\theta \in S$ where S, a finite set, could be the result of a discretization of a continuous parameter. In the Bayesian context one assumes a *a priori* distribution $\pi(\theta)$ for the parameter θ. Given a dataset $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, the *a posteriori* distribution of θ is given by

$$\pi_{\mathbf{x}}(\theta) = \frac{f(\mathbf{x};\theta)\pi(\theta)}{\sum_{\gamma\in S} f(\mathbf{x};\gamma)\pi(\gamma)},$$

where

$$f(\mathbf{x};\theta) := \prod_{k=1}^{n} f(x_k;\theta).$$

One is interested in computing averages of functions $g : S \to \mathbb{R}$ with respect to the a posteriori distribution, i.e.

$$\sum_{\theta\in S} g(\theta)\pi_{\mathbf{x}}(\theta).$$

There are cases in which S is too large for this sum to be computed directly; even the normalization factor $\sum_{\gamma\in S} f(\mathbf{x};\gamma)\pi(\gamma)$ in the definition of $\pi_{\mathbf{x}}$ could be uncomputable. If, with a modest computational effort, we can simulate random variables $X_1, X_2, \ldots, X_n$ for which (3.11) holds with $\pi = \pi_{\mathbf{x}}$, then the sample average

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

provides, with high probability if n is sufficiently large, a good approximation for $\sum_{\theta\in S} g(\theta)\pi_{\mathbf{x}}(\theta)$.

3.4.1.2 Establishing "mean properties" of complex random models: the Ising model on a graph

Probabilistic models are quite popular and useful in many application domains, such as engineering, social sciences, biology, physics.... Complex problems often require complex mathematical models, that may be hard to treat analytically. We mention here one example: the *Ising model* on a graph. Originally introduced as a stylized model for a crystal, this model and its variations are used in a variety of fields.

Let $G = (V,E)$ be a graph, according to Definition 3.6. We say that G is *undirected* if E is symmetric, i.e. $(i, j) \in E \Rightarrow (j,i) \in E$. If $(i, j) \in E$ and $i \neq j$ we say that i and j are *nearest neighbors*, or simply neighbors, and we write $i \sim j$.

Consider now an undirected, connected, finite graph, i.e. V is a finite set, and set $S = \{-1,1\}^V$: the elements of S are binary sequences $x = (x_i)_{i\in V}$, $x_i \in \{-1,1\}$ indexed by vertices: we say that x assigns to each vertex i a *spin* x_i. Elements of S are called *configurations*. We define the function $H : S \to \mathbb{R}$ as follows

$$H(x) = -\sum_{i,j:i\sim j} x_i x_j.$$

$H(x)$ is called the *energy* of the configuration x. Note that a pair of neighbors that "agree", i.e. they have the same spin, gives a contribution -1 to the energy, while a pair of neighbors that disagree gives a contribution $+1$. Since the graph is connected, there are only two configurations of minimal energy, those with constant spin. For a parameter $\beta \in \mathbb{R}$ define the distribution on S given by

$$\pi_\beta(x) = \frac{e^{-\beta H(x)}}{Z_\beta},$$

where $Z_\beta = \sum_{y\in S} e^{-\beta H(y)}$ is the *normalization factor*, needed for π_β to be a probability distribution. Note that for $\beta = 0$, π_β is the uniform distribution on S. For $\beta > 0$, π_β "favors" configurations with low energy, i.e. those where most of the neighboring spins are aligned. Exactly the opposite occurs for $\beta < 0$. This model is called the *ferromagnetic* (resp. *antiferromagnetic*) *Ising model* for $\beta > 0$ (resp. $\beta < 0$). Note that $|S| = 2^{|V|}$, so that for even moderately large V, S has a huge number of elements; in particular, no direct numerical evaluation of Z_β is possible. Thus generating S-valued random variables with distribution π_β is particularly challenging.

The Ising model and its modifications are widely used to model the equilibrium behavior of many complex networks. Beyond the traditional applications to Physics, it is used to model opinions in networks of individuals: the opinions (spins) of an individuals (vertices) on a given subject favors, for the ferromagnetic model, configurations in which neighbors tend to have the same opinion. Suppose, for instance, we want to understand how correlated are opinions of different individuals. This means that, if $X = (X_i)_{i\in V}$ is a random variable with distribution π_β, we want to compute $\mathrm{Cov}(X_i,X_j)$ for given $i \neq j$. Note that if, for $x \in S$, we denote by $-x$ the configuration in which all spins are changed, then $H(-x) = H(x)$, which implies $\pi_\beta(-x) = \pi_\beta(x)$. Thus

$$\begin{aligned}\mathrm{E}(X_i) &= \sum_{x\in S} x_i \pi_\beta(x) = \sum_{x\in S:x_i=1} \pi_\beta(x) - \sum_{x\in S:x_i=-1} \pi_\beta(x) \\ &= \sum_{x\in S:x_i=1} [\pi_\beta(x) - \pi_\beta(-x)] = 0.\end{aligned}$$

This implies that

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) = \sum_{x \in S} x_i x_i \pi_\beta(x).$$

If we are able to generate a sequence (X_n) of S-valued random variables such that (3.11) holds for $\pi = \pi_\beta$, then

$$\sum_{x \in S} x_i x_i \pi_\beta(x) \simeq \frac{1}{n} \sum_{k=1}^{n} X_{k,i} X_{k,j}$$

for n large where $X_{k,i} = (X_k)_i$ is the i-th component of X_k.

3.4.1.3 "Hard" optimization problems

Let S be a finite set, and $F : S \to \mathbb{R}$. Suppose we want to find the minimum value $\min(F)$ of F, and (at least) a value of x such that $F(x) = \min(F)$. Define the distribution on X given by

$$\pi_\beta(x) := \frac{e^{-\beta F(x)}}{Z_\beta},$$

where $\beta > 0$ and $Z_\beta = \sum_{x \in S} e^{-\beta F(x)}$. If β is large, this distribution "favors" elements of S where F takes values close to the minimum $m := \min(F)$. This can be made precise as follows. For $\varepsilon > 0$ let $A_\varepsilon := \{x \in S : F(x) < m + \varepsilon\}$, let x^* be such that $F(x^*) = m$. Then

$$Z_\beta = \sum_{x \in S} e^{-\beta F(x)} \geq \sum_{x \in A_\varepsilon^c} e^{-\beta F(x)} + e^{-\beta m}$$

Thus

$$\pi_\beta(A_\varepsilon^c) = \sum_{x \in A_\varepsilon^c} \pi_\beta(x) \leq \frac{\sum_{x \in A_\varepsilon^c} e^{-\beta F(x)}}{\sum_{x \in A_\varepsilon^c} e^{-\beta F(x)} + e^{-\beta m}}.$$

Since

$$\sum_{x \in A_\varepsilon^c} e^{-\beta F(x)} \leq \sum_{x \in A_\varepsilon^c} e^{-\beta(m+\varepsilon)} = |A_\varepsilon^c| e^{-\beta(m+\varepsilon)} = |S| e^{-\beta(m+\varepsilon)}$$

and the function $u \mapsto \frac{u}{u + e^{-\beta m}}$ is increasing, we have

$$\pi_\beta(A_\varepsilon^c) \leq \frac{|S| e^{-\beta(m+\varepsilon)}}{|S| e^{-\beta(m+\varepsilon)} + e^{-\beta m}} \leq |S| e^{-\beta \varepsilon},$$

then $\pi_\beta(A_\varepsilon^c)$ can be made arbitrarily small by taking β large. Thus, if we can generate a random variable X with distribution close to π_β with β large, then with high probability $F(X) \leq m + \varepsilon$, so X is "nearly" a minimizer for F. This provides an *approximate solution* to the minimization problem.

3.4.1.4 Counting problems

Counting the elements of a set satisfying a specific property is a central problem in combinatorics and theoretical computer science. Problems of this sort can be quite hard computationally. Stochastic algorithms are often very useful in this context; we illustrate here a specific example. Let $G = (V,E)$ be a undirected graph, and C a finite set of *colors*. The elements of C^V are called *colorings* of the graph. We say that a coloring is *admissible* if there are no two neighbors with the same color. We indicate with $A(G,C)$ the set of admissible colorings:

$$A(G,C) := \{x \in C^V : i \sim j \ \Rightarrow \ x_i \neq x_j\}.$$

It is not hard to find a sufficient condition for $A(G,C)$ to be non empty. For $i \in V$, let $d(i)$ denote the number of neighbors of i, and set $\Delta = \Delta(G) := \max\{d(i) : i \in V\}$. The number $d(i)$ is called the *degree* of the vertex i, and Δ the *maximal degree* of the graph. If $|C| \geq \Delta + 1$ the following algorithm, called *greedy algorithm*, produces an admissible coloring.

- Order arbitrarily V and C. Then color sequentially the vertices by assigned to each of then the first available color, i.e. the first color that does not violate admissibility.

For a given graph G, one may need less that $\Delta + 1$ colors to obtain an admissible coloring. It is known that an admissible coloring with the minimum number of colors can be obtained by a greedy algorithm if one uses a suitable ordering of V. Finding this ordering is a very hard problem for large graphs (NP hard).

The problem we consider is the following counting problems: given G and C such that $A(G,C) \neq \emptyset$, how many elements $A(G,C)$ has?

Suppose we can exhibit $y \in A(G,C)$, and fix an ordering $V = \{v_1, v_2, \ldots v_n\}$ for the $n = |V|$ elements of V. Define, for $k = 0, 1, \ldots, n$

$$A_k := \{x \in A(G,C) : x_{v_j} = y_{v_j} \text{ for } j = k+1, k+2, \ldots, n\}.$$

Note that $A_0 = \{y\}$ and $A_n = A(G,C)$. Denote by Π_k the uniform measure on A_k and suppose we can generate random variables for which (3.11) holds for $\pi = \pi_k$. Since $A_{k-1} \subseteq A_k$, this can be used to estimate

$$\pi_k[\mathbf{1}_{A_{k-1}}] = \pi_k(A_{k-1}) = \frac{|A_{k-1}|}{|A_k|}.$$

Observing that

$$|A(G,C)| = |A_n| = \prod_{k=1}^{n} \frac{|A_k|}{|A_{k-1}|} = \prod_{k=1}^{n} \frac{1}{\pi_k(A_{k-1})},$$

from the estimates for $\pi_k(A_{k-1})$ we can obtain estimates for $|A(G,C)|$.

3.4.2 The Metropolis algorithm

Given a distribution π on a finite set S, with $\pi_x > 0$ for all $x \in S$, our aim is to construct an irreducible Markov Chain $(X_n)_{n\geq 0}$ with stationary distribution π.

We begin by assigning a *reference* transition matrix Ψ, which has the property of being irreducible and symmetric: $\Psi_{yx} = \Psi_{xy}$ for all $x, y \in S$. Note that symmetry is equivalent to the fact that the uniform distribution on S is reversible for Ψ. In practice, this matrix Ψ has the role of assigning the transition that occur with positive probability; in general one chooses Ψ to be a *sparse* matrix, i.e. a matrix with many zero entries, being careful, however, to fulfill the irreducibility property.

The idea now is to "perturb" Ψ to obtain a transition matrix for which π is reversible.To this end consider a matrix A whose entries are in $(0,1]$, i.e. $A_{xy} \in (0,1]$ for all $x, y \in S$. Define

$$P_{xy} = \begin{cases} \Psi_{xy} A_{xy} & \text{for } x \neq y \\ 1 - \sum_{z \neq x} \Psi_{xz} A_{xz} & \text{for } x = y \end{cases}$$

Note that P is a stochastic matrix. Moreover, π is reversible for P if and only if for every $x \neq y$

$$\pi_x \Psi_{xy} A_{xy} = \pi_y \Psi_{yx} A_{yx}.$$

Since $\Psi_{xy} = \Psi_{yx}$ it is natural to choose A such that $\pi_x A_{xy} = \pi_y A_{yx}$ for all $x \neq y$. The are many choices for A, in general. We select the following one:

$$A_{xy} := \min\left(1, \frac{\pi_y}{\pi_x}\right) = 1 \wedge \frac{\pi_y}{\pi_x}.$$

Indeed

$$\pi_x A_{xy} = \pi_x \left[1 \wedge \frac{\pi_y}{\pi_x}\right] = \pi_x \wedge \pi_y$$

is symmetric in x, y.

Example 3.15. Let $G = (V, E)$ be a connected, undirected, *regular* graph, i.e. a graph in which all vertices have the same degree d. Let $F : V \to \mathbb{R}$ be a given function and, as in Section 3.4.1.3, define

$$\pi_\beta(x) := \frac{e^{-\beta F(x)}}{Z_\beta},$$

As reference transition matrix Ψ we choose that of the so-called *simple random walk* on G:

$$\Psi_{xy} = \begin{cases} \frac{1}{d} & \text{if } x \sim y \\ 0 & \text{otherwise.} \end{cases}$$

In this simple random walk, the dynamics consists in choosing at random one of the neighbors, and move there. It easily seen that

$$A_{xy} := 1 \wedge \frac{\pi_y}{\pi_x} = \exp\left[-\beta(F(y)-F(x))^+\right]$$

where, we recall, for a real number $a^+ := \max(a,0)$. Summing up, the transition matrix of the *Metropolis Markov chain* with stationary distribution π_β is given by:

$$P_{xy} = \begin{cases} \frac{1}{d}\exp[-\beta(F(y)-F(x))^+] & \text{for } x \neq y \\ 0 & \text{for } y \neq x \text{ and } y \not\sim x \\ 1 - \sum_{z \neq x} P_{xy} & \text{for } x = y \end{cases}$$

On the basis of this transition matrix, it is not hard to write an algorithm for the simulation of this chain, that can be then easily translated into a specific programming language. For a given $x \in V$, let $x^{(1)}, x^{(2)}, \dots, x^{(d)}$ be an ordering of its neighbors. As initial distribution ν for the chain it is simplest to assume it is concentrated in a specific vertex $z \in V$. The following algorithm produces the first N states $X_1, X_2, \dots, X_N$ of the chain. The input is the initial state z and the number of steps N. In this code we only generate random numbers with distribution $U(0,1)$. Denoting by $\lfloor u \rfloor$ the integer part of a real number u, observe that if $U \sim U(0,1)$ then $\lfloor dU \rfloor + 1$ takes each of the values $1, 2, \dots, d$ with probability $\frac{1}{d}$

```
Step 1. Define variables $x, y \in V$, $X = (X_1, X_2, \dots, X_N) \in V^N$, $k$ integer.
Step 2. Set $x = z$, $k = 1$.
Step 3. Generate $U \sim U(0,1)$ and set $y = x^{(\lfloor dU \rfloor + 1)}$.
Step 4. If $F(y) \leq F(x)$
          then set $x = y$
          else generate $U \sim U(0,1)$; if $U \leq \exp[-\beta(F(y)-F(x))]$ then set $x = y$
                                                                        else do nothing.
Step 5 Set $X_k = x$ and $k = k + 1$. If $k \leq N$ then return to step 3, else stop.
```

Note that running this Markov chain only involves computing differences of values of F between nearest neighbors. In particular, the normalization factor Z_β is never computed.

□

Remark 3.16. A relevant case to which the above example applies is that of $V = \{0,1\}^n$ being the space of binary sequences of length n, with the following graph structure: two sequences are neighbors if and only if they differ in a single component only. In this case $d = n - 1$ □

Remark 3.17. As illustrated in Section 3.4.1.3, the algorithm above can be used to find approximate minima of F, choosing β large. Note that this Metropolis Markov Chain is reminiscent of a classical *gradient algorithm* to search minima of F. In a classical gradient algorithm one would choose at random a neighbor y of the current state x, and move to y if $F(y) \leq F(x)$, otherwise one stays at x. Note that this algorithm gets trapped whenever x is a *local minimum* for F, even though it is not a global minimum. To avoid this phenomenon, the Metropolis algorithm assigns a positive probability $\exp[-\beta(F(y)-F(x))]$ (small for β large) of moving to y even though $F(y) > F(x)$. Thus, formally, the classical gradient algorithm corresponds to $\beta = +\infty$. □

Remark 3.18. When the goal is to find minima of F, it could be convenient to let the parameter β depend on the time n. This produces a *time inhomogeneous* Markov Chain, whose transition probabilities depend on time. To have a high probability of getting a minimum one lets $\beta = \beta_n \to +\infty$ as $n \to +\infty$. If, on one hand, one would be tempted to let β_n grow very fast in n, this could lead to the dynamics getting trapped in local minima. For this reason, a "moderate growth", e.g. $\beta_n = \log n$ is often more reliable. This procedure is called *simulated annealing*. Annealing is a process in metallurgy where a metal is first heated and then slowly cooled to avoid impurities in the reticular structure. In the algorithm above $\frac{1}{\beta}$ could be interpreted as a "temperature". □

3.4.3 The Gibbs sampler

In the Ising model and in the coloring problem we have seen examples where the state space is a set $S \subseteq C^V$, where C and V are given finite sets. For distributions on sets of this form, the *Gibbs sampler* provides a simple and effective way of constructing Markov Chains on S with a given stationary distribution.

Given $x \in S$ and $i \in V$, we set

$$\Omega(x,i) := \{y \in S : y_j = x_j \text{ for all } j \in V,\, j \neq i\}$$

be the set of elements of S that coincide with x except at most for the value at i. Clearly $y \in \Omega(x,i)$ if and only if $x \in \Omega(y,i)$. We can thus provide S with a structure of undirected graph, defining the set of edges E as follows: $(x,y) \in E$ if and only if $y \in \Omega(x,i)$ for some $i \in V$. We assume in what follows that (S,E) is a connected graph; this is obviously true if $S = C^V$, but may fail otherwise.

Suppose π is a distribution on S such that $\pi_x > 0$ for every $x \in S$. Let X be a S-valued random variable with distribution π. Given $x \in S$, $i \in V$ and $y \in \Omega(x,i)$, consider the conditional probability

$$\mathrm{P}(X = y | X \in \Omega(x,i)) = \frac{\pi_y}{\sum_{z \in \Omega(x,i)} \pi_z}. \tag{3.12}$$

The dynamics of the Gibbs sampler are described in the following two steps, in which x denotes the current state of the Markov Chain: (a) select at random, with uniform probability, a point $i \in V$; (b) select randomly an element $y \in \Omega(x,i)$ with the probability given in (3.12). This corresponds to the transition matrix:

$$P_{xy} := \begin{cases} \frac{1}{|V|} \frac{\pi_y}{\sum_{z \in \Omega(x,i)} \pi_z} & \text{for } y \in \Omega(x,i),\, y \neq x \\ 0 & \text{if there is no } i \in V \text{ with } y \in \Omega(x,i) \\ 1 - \sum_{z \neq x} P_{xz} & \text{if } y = x. \end{cases} \tag{3.13}$$

Proposition 3.19. *The distribution π is reversible for the transition matrix P given in* (3.13).

Proof. Just note that

$$\pi_x P_{xy} = \begin{cases} \frac{1}{|V|} \frac{\pi_x \pi_y}{\sum_{z \in \Omega(x,i)} \pi_z} & \text{for } y \in \Omega(x,i) \\ 0 & \text{otherwise} \end{cases}$$

is symmetric in x, y since $\Omega(x,i) = \Omega(y,i)$. □

We observe that the connectedness of the graph (S,E) guarantees that the Markov Chain with transition matrix P is irreducible.

Example 3.20. In the context of the model illustrated in Section 3.4.1.4, let π the the uniform distribution on the set $A(G,C) \subseteq C^V$ of the admissible colorings of a graph $G = (V,E)$. Denoting by Δ the maximal degree of the graph G, we assume $|C| \geq \Delta + 1$. It can be shown that under this condition, if we provide $A(G,C)$ with the graph structure illustrated above, the we obtain a connected graph. Note that since π is uniform, the conditional probability in (3.12) is simply given by $\frac{1}{|\Omega(x,i)|}$: in other words the color to replace x_i is simply chosen with uniform probability among the colors that do not violate admissibility. More explicitly, for $x \in S$, $i \in V$ and $c \in C$, let $x^{i,c}$ be the element of C^V obtained from x by replacing x_i with c. Moreover, set

$$C(x,i) := \{c \in C : x^{i,c} \in S\}.$$

The dynamics of the Gibbs sampler can be described as follows, assuming the initial distribution is concentrated on a given $z \in S$:
Step 1 Set $x = z$.
Step 2 Select at random, with uniform probability, $i \in V$.
Step 3 Select at random $c \in C(x,i)$.
Step 4 Set $x = x^{i,c}$ and return to Step 2.

□

Example 3.21. We consider now the Ising Model introduced in Section 3.4.1.2. Recall that $G = (V,E)$ is a finite graph, $S = \{-1,1\}^V$, and

$$\pi_\beta(x) = \frac{e^{-\beta H(x)}}{Z_\beta},$$

where

$$H(x) = -\sum_{i,j:i \sim j} x_i x_j.$$

Recall that $i \sim j$ means that $(i,j) \in E$. For $x \in S$ and $i \in V$, we denote by x^i the element of S obtained from x by replacing x_i with $-x_i$. Note that $\Omega(x,i) = \{x, x^i\}$.

Except for the diagonal elements, the only nonzero elements of the transition matrix of the Gibbs samples are P_{x,x^i} for $i \in V$, which are given by

$$P_{x,x^i} = \frac{1}{|V|} \frac{e^{-\beta H(x^i)}}{e^{-\beta H(x^i)} + e^{-\beta H(x)}} = \frac{1}{|V|} \frac{1}{1 + e^{-\beta [H(x^i) - H(x)]}} = \frac{1}{|V|} \frac{1}{1 + e^{\beta x_i S_i}},$$

where

$$S_i := \sum_{j:j \sim i} x_j.$$

We can therefore write an explicit algorithm for the simulation of the first N steps of this Markov Chain. For definiteness, we set $V := \{1, 2, \ldots, m\}$ with edges between consecutive points and *periodic boundary condition*, i.e. 1 and m are neighbors. In what follows integers are meant modulo m, in particular. $0 = m$ and $m + 1 = 1$. The initial condition is set equal to $\mathbf{1} = (1, 1, \ldots, 1)$.

```
Step 1. Define variables $x, y \in S = \{0,1\}^m$, $X = (X_1, X_2, \ldots, X_N) \in S^N$, $i, k$ integers.
Step 2. Set $x = \mathbf{1}$ and $k = 1$.
Step 3. Generate $U \sim U(0,1)$ independent. Set $i := \lfloor mU \rfloor + 1$ and $y = x^i$.
Step 4. Generate $V \sim U(0,1)$.
If $V \leq \frac{1}{1+e^{\beta x_i (x_{i-1}+x_{i+1})}}$
        then set $x = y$
        else do nothing.
Step 5 Set $X_k = x$ and $k = k + 1$. If $k \leq N$ then return to step 3, else stop.  □
```

Chapter 4
Random Networks

Complex networks are so pervasive in Science that the data we have to deal with are often related to a network, and the aim of a Data Scientist often consists in identifying its properties. The most popular complex networks are the *social networks* (Facebook, Instagram, Twitter...): individuals are connected by "virtual" links which carry a huge variety of informations. These networks are quite relevant in the formation of opinions on many different matters, and they impact, for instance, the political debate and strategies all over the world. Other human networks may have links of more "physical" nature, and play a key role, for example, in understanding, controlling or preventing the spread of epidemics. Electric and telephone networks provide examples of networks related to technology and engineering applications. Complex networks also emerge in nature, for instance in fluvial geomorphology.

All these networks have in common the feature of being *big*, so big that a detailed description is impossible or scarcely useful. *Statistical* properties of these networks are more interesting: this has motivated and stimulated the development of probabilistic models for networks. In this Chapter we first describe the models with which the study of random graph has started. We then study some class of *evolving* random graph, whose properties reflects more accurately some features of real complex networks.

4.1 Erdós-Renyi random graphs

We consider a graph with n vertices, identified with the elements of the set $V := \{1,2,\dots,n\}$. The edges of the graph are built with the following probabilistic rule. For each *unordered* pair of distinct vertices $\{i,j\}$ we introduce a random variable $\xi_{ij} \sim \mathrm{Be}(p)$, where $p \in (0,1)$ is a given parameter, and the ξ_{ij}'s are all independent. We then define the set of edges as follows

$$E := \{(i,j) : \xi_{ij} = 1\}. \tag{4.1}$$

Note that the resulting graph is undirected, and each two vertices i and j are linked by an edge with probability p, independently from all other edges. We denote by $G(n,p)$ the resulting random graph; these networks are called *Erdós-Renyi* random graphs.

Observe that the degree of a vertex i is given by

$$d(i) = \sum_{j:j\neq i} \xi_{ij},$$

so it has distribution $\mathrm{Bin}(n-1,p)$. For n large and p small this is well approximated by the Poisson distribution of parameter np (see Theorem 1.40).

We are interested in studying asymptotic properties of Erdós-Renyi graphs, namely properties that hold when n is large. It is moreover convenient to let $p = p_n$ to depend on n; for instance, if p_n scales in n as $\frac{1}{n}$, then the average number of neighbors of each vertex is of order one. It turns out, as we will see in some examples, that the validity of many properties in $G(n,p_n)$ is very sensitive to small changes of p_n. The following definition expresses this sensibility in formal terms.

Definition 4.1. Let $(a_n)_{n\geq 1}$ be a given sequence in $(0,1)$. We say that property $\mathscr{P}$ has a *sharp phase transition* with respect to the *converging* sequence (a_n) if

$$\lim_{n\to+\infty} \mathrm{P}(\text{property } \mathscr{P} \text{ holds for } G(n,p_n)) = \begin{cases} 0 \text{ if } p_n \leq c\cdot a_n \text{ for some } c<1 \\ 1 \text{ if } p_n \geq c\cdot a_n \text{ for some } c>1 \end{cases}$$

or

$$\lim_{n\to+\infty} \mathrm{P}(\text{property } \mathscr{P} \text{ holds for } G(n,p_n)) = \begin{cases} 1 \text{ if } p_n \leq c\cdot a_n \text{ for some } c<1 \\ 0 \text{ if } p_n \geq c\cdot a_n \text{ for some } c>1 \end{cases}$$

In other words, when a sharp phase transition occurs, the probability that property $\mathscr{P}$ holds in $G(n,p_n)$ with n large changes from nearly zero to nearly one by letting p_n crossing the threshold a_n.

In most cases, a property $\mathscr{P}$ for $G(n,p_n)$ is expressed by the condition $\{X_n > 0\}$ where X_n is a random variable taking nonnegative integer values. Proving a sharp phase transition consists in finding conditions for which $\mathrm{P}(X_n>0)\to 0$ as $n\to+\infty$ or rather $\mathrm{P}(X_n>0)\to 1$ as $n\to+\infty$. The basic tool for proving that $\mathrm{P}(X_n>0)\to 0$ is the so-called *first moment method*, which is illustrated in next Proposition.

Proposition 4.2. *Suppose*

$$\lim_{n\to+\infty} \mathrm{E}(X_n) = 0.$$

Then

$$\lim_{n\to+\infty} \mathrm{P}(X_n>0) = 0.$$

Proof. This follows easily form the fact that

$$P(X_n > 0) = \sum_{k=1}^{+\infty} P(X_n = k) \leq \sum_{k=1}^{+\infty} k P(X_n = k) = E(X_n).$$

□

In order to get conditions guaranteeing that $P(X_n > 0) \to 1$ we use a slightly more complicated method involving the computation of the second moment of X_n, and for this reason called the *second moment method*.

Proposition 4.3. *Suppose*

$$\lim_{n\to+\infty} \frac{E(X_n^2)}{E^2(X_n)} = 1.$$

Then

$$\lim_{n\to+\infty} P(X_n > 0) = 1.$$

Proof. Note first that

$$\{X_n > 0\}^c = \{X_n = 0\} \subseteq \{|X_n - E(X_n)| \geq E(X_n)\}.$$

Thus, by Chebischev inequality:

$$1 - P(X_n > 0) = P(X_n = 0) \leq P(|X_n - E(X_n)| \geq E(X_n)) \leq \frac{\mathrm{Var}(X_n)}{E^2(X_n)} = \frac{E(X_n^2)}{E^2(X_n)} - 1,$$

which, by assumption, converges to zero. □

We next provide three examples of properties having a sharp phase transition. For the two of them the first and the second moment methods will suffice to establish phase transition, while more sophisticated tools will be needed for the third.

4.1.1 Existence of isolated vertices

A vertex is isolated if it possesses no neighbors. If $p_n \equiv 0$ then in $G(n, p_n)$ all vertices are isolated. How much p_n has to be increased to rule out the existence of isolated vertices? To answer this question we define

$$Y_i := \begin{cases} 1 \text{ if vertex } i \text{ is isolated} \\ 0 \text{ otherwise.} \end{cases}$$

So $X_n := \sum_{i=1}^n Y_i$ is the number of isolated vertices. So the property "there are isolated vertex" corresponds to the event $\{X_n > 0\}$. In what follows we consider the random variables ξ_{ij} that we have used in (4.1) to define $G(n,p)$ (we are for the moment omitting the index n in p_n. Note that

$$Y_i = 1 \quad \Longleftrightarrow \quad \sum_{j \neq i} \xi_{ij} = 0.$$

But $\sum_{j \neq i} \xi_{ij} \sim \mathrm{Bin}(n-1,p)$, so

$$\mathrm{P}(Y_i = 1) = \mathrm{E}(Y_i) = (1-p)^{n-1}.$$

Therefore

$$\mathrm{E}(X_n) = n(1-p_n)^{n-1} = \exp\left[(n-1)\log(1-p_n) + \log(n)\right].$$

Note that this expectation decreases if we increase p_n. If p_n is bounded from below by a strictly positive constant, then $(n-1)\log(1-p_n) + \log(n) \to -\infty$ as $n \to +\infty$, so $\mathrm{E}(X_n) \to 0$. Suppose, now, that $p_n \to 0$ as $n \to +\infty$, so that $\log(1-p_n) \simeq p_n$, giving

$$\mathrm{E}(X_n) \simeq \exp\left[-np_n + \log(n)\right]. \tag{4.2}$$

This shows that $\mathrm{E}(X_n) \to 0$ if $p_n \geq c\frac{\log(n)}{n}$ with $c > 1$; thus, by the first moment method, the probability of having at least one isolated vertex goes to zero.

In order to establish sharp phase transition, we need to show that whenever $p_n \leq c\frac{\log(n)}{n}$ with $c < 1$, then $\mathrm{P}(X_n > 0) \to 1$ as $n \to +\infty$. To see this we use the second moment method. The key step consists in computing $\mathrm{E}(X_n^2)$. Note that

$$X_n^2 = \sum_{i,j=1}^{n} Y_i Y_j = D + N$$

where

$$D := \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i = X_n$$

is the contribution of the "diagonal" terms, while

$$N := \sum_{i \neq j} Y_i Y_j$$

is the contribution of the "non-diagonal" terms. Note that if $p_n \leq c\frac{\log(n)}{n}$ with $c < 1$, then

$$\mathrm{E}(X_n) \simeq \exp\left[-np_n + \log(n)\right] \to +\infty$$

as $n \to +\infty$, so

$$\frac{\mathrm{E}(D)}{\mathrm{E}^2(X_n)} = \frac{\mathrm{E}(X_n)}{\mathrm{E}^2(X_n)} \to 0. \tag{4.3}$$

Next step consists in computing

$$\mathrm{E}(N) = \sum_{i \neq j} \mathrm{E}(Y_i Y_j).$$

Note that $Y_i Y_j$ is a Bernoulli random variable, and it takes value 1 if and only if both i and j are isolated. This happens if and only if

$$\sum_{k: k \neq i,j} [\xi_{ik} + \xi_{jk}] + \xi_{ij} = 0.$$

Since all variables in this last sum are independent $\mathrm{Be}(p)$, this has probability $(1 - p_n)^{2n-3}$. Thus, always under the condition $p_n \leq c \frac{\log(n)}{n}$ with $c < 1$,

$$\begin{aligned} \mathrm{E}(N) &= \sum_{i \neq j} \mathrm{E}(Y_i Y_j) = n(n-1)(1-p_n)^{2n-3} \\ &= \exp\left[(2n-3)\log(1-p_n) + \log n + \log(n-1)\right] \simeq \exp\left[-2np_n + 2\log(n)\right]. \end{aligned}$$

Comparing this with (4.2), we have

$$\frac{\mathrm{E}(N)}{\mathrm{E}^2(X_n)} \to 1.$$

In conclusion

$$\frac{\mathrm{E}(X_n^2)}{\mathrm{E}^2(X_n)} = \frac{\mathrm{E}(D)}{\mathrm{E}^2(X_n)} + \frac{\mathrm{E}(N)}{\mathrm{E}^2(X_n)} \to 1.$$

By the second order method this implies that $\mathrm{P}(X_n > 0) \to 1$.

Summing up, we have proved the following result.

Proposition 4.4. *The property $\mathscr{P}$ = "there are isolated vertices" has a sharp phase transition w.r.t. the sequence* $\frac{\log(n)}{n}$.

Rather surprisingly, it can be shown that if $p_n \geq c \frac{\log(n)}{n}$ with $c > 1$, not only, with high probability, there are no isolated vertices, but much more that this holds true: $G(n, p_n)$ is connected with a probability that goes to one as $n \to +\infty$. This would require slightly more accurate arguments, that we omit here.

4.1.2 Diameter 2

We recall that a *path* in a graph is a finite sequence of adjacent edges, and its length is the number of edges it contains. The *graph distance* between two vertices is the length of the (not necessarily unique) shortest path joining the two vertices: if there is no such path the distance is infinite. The *diameter* of a graph is the maximum distance between pairs of vertices (thus it is $+\infty$ is the graph is not connected). One of the evidences in real complex networks is that they have a "small" diameter, compared to the number of vertices. Note that if $p = 1$ then $G(n, p)$ is the *complete graph*: all pairs of vertices are linked by an edge, so the diameter of the graph is one. For every p_n bounded above by a constant < 1, with high probability there are two vertices that are not linked by an edge, so the diameter is at least 2. We now look for conditions under which the diameter of $G(n, p_n)$ is not larger that 2.

Given two distinct vertices i, j, we set

$$Y_{ij} := \begin{cases} 1 \text{ if the distance between } i \text{ and } j \text{ is at least } 3 \\ 0 \text{ otherwise,} \end{cases}$$

and define

$$X_n := \sum_{\{i,j\}:i\neq j} Y_{ij},$$

where this last sum ranges over all the $\binom{n}{2} = \frac{n(n-1)}{2}$ unordered pairs of distinct vertices. Observe that $X_n > 0$ if and only if the diameter of the graph is at least 3. Note also that $Y_{ij} = 1$ if and only if the following conditions hold:

- there is no edge between i and j, i.e. $\xi_{ij} = 0$. This has probability $1 - p_n$;
- there is no vertex k linked to both i and j, i.e. for all $k \neq i, j$

$$\xi_{ik} + \xi_{jk} < 2.$$

  This has, for a given k, probability $1 - p_n^2$.

Since the conditions above involve independent random variables, we have

$$\mathrm{P}(Y_{ij} = 1) = (1 - p_n)(1 - p_n^2)^{n-2}.$$

Thus

$$\begin{aligned} \mathrm{E}(X_n) &= \frac{n(n-1)}{2}(1-p_n)(1-p_n^2)^{n-2} \\ &= \exp\left[(n-2)\log(1-p_n^2) + \log(1-p_n) + \log(n) + \log(n-1) - \log(2)\right]. \end{aligned}$$

As in the previous example, this expectation is decreasing in p_n. If we assume $p_n \to 0$, then

$$\mathrm{E}(X_n) \simeq \exp\left[-np_n^2 + 2\log(n) - \log(2)\right], \tag{4.4}$$

from which it is easily seen that $\mathrm{E}(X_n) \to 0$ as soon as $p_n \geq c\sqrt{\frac{2\log(n)}{n}}$ with $c > 1$, and therefore, under this condition, by the first moment method $G(n, p_n)$ has with high probability diameter at most 2.

We now aim at proving a sharp phase transition. So assume $p_n \leq c\sqrt{\frac{2\log(n)}{n}}$ with $c < 1$, and we try to use the second order method to show that $\mathrm{P}(X_n > 0) \to 1$. Inspired by what done for isolated vertices, we write

$$X_n^2 = \sum_{\{i,j\},\{k,l\}} Y_{ij}Y_{kl} = D + N_3 + N_4,$$

with

$$D = \sum_{\{i,j\}: i \neq j} Y_{ij}^2 = \sum_{\{i,j\}: i \neq j} Y_{ij} = X_n,$$

$$N_3 = \sum_{\{i,j\},\{k,l\}: |\{i,j,k,l\}|=3} Y_{ij} Y_{kl} = \sum_{i,j,k \text{ distinct}} Y_{ij} Y_{ik},$$

$$N_4 = \sum_{i,j,k,l \text{ distinct}} Y_{ij} Y_{kl}.$$

Since $p_n \leq c\sqrt{\frac{2\log(n)}{n}}$ with $c < 1$, then by (4.4), $\mathrm{E}(X_n) \to +\infty$, so

$$\frac{D}{\mathrm{E}^2(X_n)} = \frac{\mathrm{E}(X_n)}{\mathrm{E}^2(X_n)} \to 0. \tag{4.5}$$

We now consider N_3. Give the three distinct vertices i, j, k, we have that $Y_{ij} Y_{ik} = 1$ if and only if the following independent events occur simultaneously:

- there is no edge between i, j and between i, k. This has probability $(1 - p_n)^2$.
- for each vertex $m \neq i, j, k$, one of the following must hold: either $\xi_{im} = 0$ or $\xi_{im} = 1$ and $\xi_{jm} + \xi_{km} = 0$. The total probability that this occurs in $1 - p_n + p_n(1 - p_n)^2$.

Summing up,

$$\mathrm{P}(Y_{ij} Y_{ik} = 1) = (1 - p_n)^2 \left[1 - p_n + p_n(1 - p_n)^2\right]^{n-3},$$

so

$$\begin{aligned}
\mathrm{E}(N_3) &= 2\binom{n}{3}(1 - p_n)^2 \left[1 - p_n + p_n(1 - p_n)^2\right]^{n-3} \\
&= \frac{n(n-1)(n-2)}{3} \left[1 - p_n + p_n(1 - p_n)^2\right]^{n-3} \\
&= \exp\left[(n-3)\log(1 - 2p_n^2 + p_n^3) + \log(n) + \log(n-1) + \log(n-2) - \log(3)\right] \\
&\simeq \exp\left[-2np_n + 3\log(n) - \log(3)\right].
\end{aligned}$$

Since, by (4.4),

$$\mathrm{E}^2(X_n) \simeq \exp\left[-2np_n^2 + 4\log(n) - 2\log(2)\right],$$

it follows that

$$\frac{\mathrm{E}(N_3)}{\mathrm{E}^2(X_n)} \to 0. \tag{4.6}$$

We finally compute $\mathrm{E}(N_4)$. We are not actually going to compute it exactly, as only an upper bound will be needed. Given four distinct vertices i, j, k, l, if $Y_{ij} Y_{kl} = 1$ then the following statement must hold:

- for any $m \neq i, j, k, l$

$$\xi_{im} + \xi_{jm} < 2 \text{ and } \xi_{km} + \xi_{lm} < 2,$$

which occurs with probability $(1 - p_n^2)^2$.

Thus

$$\mathrm{E}(Y_{ij}Y_{kl}) \leq (1 - p_n^2)^{2(n-4)}.$$

The number of summands in N_4 is $\binom{n}{2}$ (number of ways of choosing the unordered pair $\{i, j\}$ times $\binom{n-2}{2}$ (number of ways of choosing the second unordered pair $\{j, k\}$. Thus

$$\begin{aligned}
\mathrm{E}(N_4) &\leq \frac{n(n-1)(n-2)(n-3)}{4}(1 - p_n^2)^{2(n-4)} \\
&= \exp\left[2(n-4)\log(1 - p_n^2) + \log(n) + \log(n-1) + \log(n-2) + \log(n-3) - \log(4)\right] \\
&\simeq \exp\left[-2np_n^2 + 4\log(n) - 2\log(2)\right] \\
&\simeq \mathrm{E}^2(X_n)
\end{aligned}$$

where the last step follows from (4.4). Putting all estimates above together we obtain

$$\frac{E(X_n^2)}{\mathrm{E}^2(X_n)} = \frac{\mathrm{E}(D) + \mathrm{E}(N_3) + \mathrm{E}(N_4)}{\mathrm{E}^2(X_n)} \leq 1 + o(1). \tag{4.7}$$

Since, being $E(X_n^2) \geq \mathrm{E}^2(X_n)$, the above ratio is always greater or equal to one, it follows that $\frac{E(X_n^2)}{\mathrm{E}^2(X_n)} \to 1$. The second order method therefore implies that, for $p_n \leq c\sqrt{\frac{2\log(n)}{n}}$ with $c < 1$, $\mathrm{P}(X_n > 0) \to 1$, so with high probability the diameter is at least three. Summing up, we have shown the following result.

Proposition 4.5. *The property $\mathscr{P}$ = "$G(n, p_n)$ has diameter at most two" has a sharp phase transition w.r.t. the sequence* $\sqrt{\frac{2\log(n)}{n}}$.

4.1.3 Giant component

We have seen that if $p_n \leq c\frac{\log(n)}{n}$ with $c < 1$, then there are isolated vertices in $G(n, p_n)$. In particular, it is not connected. In this section we consider the case in which $p_n = \frac{c}{n}$ for some $c > 0$: since $\frac{1}{n} \ll \frac{\log(n)}{n}$, it follows that $G(n, c/n)$ is not connected.

Given a vertex i, the *connected component* of i, denoted by $\mathrm{cc}(i)$, is the set of all vertices that can be reached from i following a path, including i itself. In order to explore the connected component of i we introduce the following *search* algorithm. In this algorithm, to each vertex a label is assigned; the possible labels are: *undiscovered*, *discovered*, *explored*. Initially, vertex i is discovered, while all other vertices are undiscovered. In the first step we consider vertices that are neighbors of i; we label them "discovered", while i gets the label "explored". The generic step of the algorithm is as follows: take the smallest discovered vertex j (smallest for the usual order in $\{1,2,\ldots,n\}$). Then label j "explored", and all its undiscovered neighbors (if any) are labelled "discovered". The algorithm stops when there are no more discovered vertices; moreover $|\mathrm{cc}(i)| = k$ if and only if the algorithm stops after k steps. Let also z_k denote the number of vertices that get discovered within the first k steps, not counting the vertex i. Since after k steps, exactly k vertices get the label "explored", the search algorithm stops as soon as $z_k = k-1$.

A different, but essentially equivalent way of exploring the connected component of the vertex i, consists in introducing a "parallel" version of the search algorithm, that we call *branching algorithm*. We still start from the vertex i, and set $w_1 = 1$. We call $\{i\}$ the *first generation*. All neighbors of i form the *second generation*, and we denote by w_2 its size. In general, given the k-th generation, the $k+1$-st generation is comprised by the vertices that do not belong to any of the first k generations, but are neighbors to a vertex of the k-th generation.

We first consider the case with $c < 1$. We show that in this case all connected components are "small", i.e. at most of size $O(\log(n))$.

Proposition 4.6. *Suppose $c < 1$. Then there is a constant $a > 0$ such that the probability that $G(n,c/n)$ has a connected component of size greater than $a\log(n)$ goes to zero as $n \to +\infty$.*

Proof. The key argument is to compute the distribution of z_k, the number of vertices discovered within the first k steps in the search algorithm starting from vertex i. Fix a vertex $j \neq i$. The probability that it is discovered in a given step of the algorithm in $\frac{c}{n}$, independently in different steps and for different vertices. Thus the probability that vertex j remains undiscovered after k steps is $\left(1-\frac{c}{n}\right)^k$, so that

$$z_k \sim \mathrm{Bin}\left(n-1, 1-\left(1-\frac{c}{n}\right)^k\right).$$

In what follows we will use the inequalities, that holds for $x \in (0,1)$:

$$1 - kx \leq (1-x)^k \leq 1 - kx + \frac{k^2x^2}{2}, \tag{4.8}$$

so

$$1-\left(1-\frac{c}{n}\right)^k \leq \frac{kc}{n}.$$

This implies

$$\mathrm{E}(z_k) \leq ck.$$

To get a lower bound for $\mathrm{E}(z_k)$ we use the upper inequality in (4.8):

$$\begin{aligned}\mathrm{E}(z_k) &= (n-1)\left[1-\left(1-\frac{c}{n}\right)^k\right] \geq (n-1)\left[\frac{kc}{n}-\frac{k^2c^2}{2n^2}\right] \\ &\geq (n-1)\left[\frac{kc}{n}-\frac{kc}{2n}\right] \geq \frac{kc}{3},\end{aligned}$$

where we have used the facts that $\frac{kc}{n} < 1$ (as necessarily $k \leq n$) and that, for $n \geq 3$, $\frac{n-1}{n} \geq \frac{2}{3}$. Therefore

$$z_k = k-1 \;\Rightarrow\; z_k \geq \mathrm{E}(z_k) + (1-c)k - 1 \;\Rightarrow\; z_k \geq \mathrm{E}(z_k) + \frac{1-c}{c}\mathrm{E}(z_k) - 1 \geq \mathrm{E}(z_k) + \frac{1-c}{2c}\mathrm{E}(z_k)$$

for k large, as $\mathrm{E}(z_k) \geq \frac{kc}{3} \gg 1$. By the upper tail Chernoff bound in Proposition 2.3 with $\delta = \frac{1-c}{c}$ we obtain, for some constant $b > 0$,

$$\mathrm{P}(z_k = k) \leq \mathrm{P}(z_k \geq \mathrm{E}(z_k) + \delta\, \mathrm{E}(z_k)) \leq e^{-b\mathrm{E}(z_k)}. \tag{4.9}$$

Thus, by (4.10) and the above lower bound for $\mathrm{E}(z_k)$, for a possibly different constant $b > 0$

$$\mathrm{P}(z_k = k) \leq e^{-bk}.$$

Since, as already observed, $z_k = k-1$ if and only if $\mathrm{cc}(i) = k$,

$$\mathrm{P}(|\mathrm{cc}(i)| > a\log(n)) = \sum_{k > a\log(n)} \mathrm{P}(z_k = k) \leq \sum_{k > a\log(n)} e^{-bk} \leq Ce^{-ba\log(n)} = \frac{C}{n^{ab}}.$$

Thus, if we choose $a = \frac{2}{b}$ we have

$$\mathrm{P}(|\mathrm{cc}(i)| > a\log(n)) \leq \frac{C}{n^2}.$$

Finally

$$\mathrm{P}(\exists i:\, |\mathrm{cc}(i)| > a\log(n)) \leq \sum_{i=1}^{n} \mathrm{P}(|\mathrm{cc}(i)| > a\log(n)) \leq n\frac{C}{n^2} \to 0$$

as $n \to +\infty$, and the proof is completed. □

We now consider the case $c > 1$. Next result shows that there must exist a connected component of order larger than $O(\log(n))$.

Proposition 4.7. *Suppose $c > 1$ and let $C > 0$ an arbitrary constant. Then there is $\alpha < 1$ such that for every vertex i and n large enough*

$$\mathrm{P}(|\mathrm{cc}(i)| \leq C\log(n)) \leq \alpha.$$

Sketch of the proof. For this proof, that is a bit complicated to put in totally rigorous terms, we use the branching algorithm illustrated above.

Let $m := n - C\log(n)$. If $|\mathrm{cc}(i)| \le C\log(n)$ then any vertex of the k-th generation has a number of neighbors of the $k+1$-st generation which is bounded below by a binomial $\mathrm{Bin}(m, c/n)$. Thus we can bound from below the sizes w_k by the sizes v_k of the generations of the branching process defined as follows: $v_1 = 1$; each individual of the k-th generation gives birth to a random number $\mathrm{Bin}(m, c/n)$ of individuals of the $k+1$-st generation, independently of all other individuals. The fact that the branching algorithm stops ($w_k = 0$ as soon as $\mathrm{cc}(i)$ is fully explored) implies that also $v_k = 0$ for some k. Note that if $v_k = 0$ then also $v_{k+1} = 0$, so $q_k := \mathrm{P}(v_k = 0)$ is increasing in k. Summing up, we have

$$\mathrm{P}(|\mathrm{cc}(i)| \le C\log(n)) \le q := \lim_{k\to+\infty} q_k.$$

We are therefore left to show that $q < 1$. Note that

$$q_{k+1} = \mathrm{P}(v_{k+1} = 0) = \sum_{i=0}^{m} \mathrm{P}(v_{k+1} = 0 | v_2 = i)\,\mathrm{P}(v_2 = i).$$

By assumption, $v_2 \sim \mathrm{Bin}(m, c/n)$. Moreover, each one of the v_2 individuals of the second generation is the starting point of branching process, independent of the branching processes starting from other individuals. Thus $\mathrm{P}(v_{k+1} = 0 | v_2 = i)$ coincides with the probability that i independent branching processes die out within k steps, i.e. q_k^i. Therefore

$$q_{k+1} = \sum_{i=0}^{m} \mathrm{P}(v_2 = i) q_k^i, \tag{4.10}$$

with initial condition $q_1 = 0$ (and with the convention $0^0 = 1$). Consider the polynomial

$$P(x) = \sum_{i=0}^{m} \mathrm{P}(v_2 = i) x^i.$$

Note that $P(x)$ is increasing and convex in $[0,1]$ (as $P'(x)$ and $P''(x)$ are positive), $P(0) = \mathrm{P}(v_2 = 0) > 0$,

$$P(1) = \sum_{i=0}^{m} \mathrm{P}(v_2 = i) = 1.$$

Moreover

$$P'(1) = \sum_{i=0}^{m} i\,\mathrm{P}(v_2 = i) = \mathrm{E}(v_2) = \frac{mc}{n}.$$

Recalling that $m := n - C\log(n)$ and $c > 1$, for n sufficiently large $\frac{mc}{n} > 1$, so $P'(1) > 1$. As a consequence, there must be a unique $\alpha \in (0,1)$ such that $P(\alpha) = \alpha$. Since, by (4.10), $q_{k+1} = P(q_k)$, by monotonicity of P, we obtain $q_k \le \alpha$ for all k, and this completes the proof. □

We now show that a connected component of order larger than $O(\log(n))$ is necessarily *giant*, i.e. of order $O(n)$.

Proposition 4.8. *Suppose $c > 1$. There are constants $a, b > 0$ such that*

$$\lim_{n\to+\infty} \mathrm{P}(\exists\, i : a\log(n) \le \mathrm{cc}(i) \le bn) = 0.$$

Proof. Recall that $\mathrm{cc}(i) = k$ if and only if $z_k = k-1$, where $z_k \sim \mathrm{Bin}\left(n-1, 1-\left(1-\frac{c}{n}\right)^k\right)$. By (4.8),

$$\mathrm{E}(z_k) = (n-1)\left[1-\left(1-\frac{c}{n}\right)^k\right] \ge (n-1)\left[\frac{kc}{n} - \frac{k^2c^2}{2n^2}\right].$$

Therefore

$$\mathrm{E}(z_k) - (k-1) \ge \frac{c(n-1)-n}{n}k - (n-1)\frac{k^2c^2}{2n^2}.$$

If $k \le bn$ with b sufficiently small, we have that $\mathrm{E}(z_k) - k \ge \delta k$ for a sufficiently small δ. Using a Chernoff bound as in Proposition 4.6,

$$\mathrm{P}(z_k = k) \le \mathrm{P}(z_k \le \mathrm{E}(z_k) - \delta k) \le e^{-\gamma k}$$

for some $\gamma > 0$. Therefore

$$\mathrm{P}(a\log(n) \le \mathrm{cc}(i) \le bn) \le \sum_{k > a\log(n)} e^{-\gamma k} \le Ce^{-\gamma a \log(n)} \le \frac{C}{n^2}$$

if a is sufficiently large, for some constant $C > 0$. Finally

$$\mathrm{P}(\exists\, i : a\log(n) \le \mathrm{cc}(i) \le bn) \le n\mathrm{P}(a\log(n) \le \mathrm{cc}(i) \le bn) \le \frac{C}{n} \to 0$$

and this completes the proof. □

It can also be shown that, for $c > 1$, there is a *unique* giant component of order $O(n)$.

4.2 Random graphs evolving in time

Real complex networks often evolve in time: new vertices enter the network, and new edges are added.

4.2.1 Networks growing without preferential attachment

The simplest model for the network's evolution is the following. Start at time $t = 1$ with 1 vertex and no edge. At each time $t > 1$ we first insert a new vertex, then pick at random, with uniform probability, two vertices and insert an edge linking them, unless such edge already exists. Denote by $N_k(t)$ the number of vertices that at time t have degree k. In particular, N_0 is the number of isolated vertices. At each time t, $N_0(t)$ increases by one at the addition of the new vertex, and decreases by $D_0(t+1)$ where

$$\begin{aligned} \mathrm{P}(D_0(t+1)=2|N_0(t)=m) &= \frac{(m+1)m}{(t+1)t} \\ \mathrm{P}(D_0(t+1)=1|N_0(t)=m) &= \frac{2(m+1)(t-m)}{(t+1)t}. \end{aligned}$$

Therefore, using the fact that

$$\mathrm{P}(D_0(t+1)=i) = \sum_m \mathrm{P}(D_0(t+1)=2|N_0(t)=m)\,\mathrm{P}(N_0(t)=m),$$

we have

$$\begin{aligned} \mathrm{E}(D_0(t+1)) &= \mathrm{P}(D_0(t+1)=1) + 2\,\mathrm{P}(D_0(t+1)=2) \\ &= \sum_m \frac{2(m+1)}{t+1}\,\mathrm{P}(N_0(t)=m) \\ &= \frac{2(\mathrm{E}(N_0(t))+1)}{t+1}. \end{aligned}$$

S $N_0(t+1) = N_0(t) + 1 - D_0(t+1)$, letting $d_0(t) = (\mathrm{E}(N_0(t))$ we have

$$d_0(t+1) = d_0(t) + 1 - \frac{2(d_0(t)+1)}{t+1} \simeq d_0(t) + 1 - \frac{2d_0(t)}{t} \tag{4.11}$$

where this last approximation holds for t large. A similar argument can be used for $N_k(t)$ with $k>0$: it increases by $S_k(t+1) = 0,1,2$, the number of vertices of degree $k-1$ selected for the new edge (unless the chosen pair is already connected), and it decreases by $D_k(t+1) = 0,1,2$, the number of vertices of degree k selected for the new edge. Since the number of edges grows like t while the number of pairs of vertices grows like t^2, it is very unlikely, for large t, to choose a pair which is already linked by an edge. Thus we may ignore this possibility; repeating the argument above for the computation of $\mathrm{E}(D_0)$ and we obtain

$$\mathrm{E}(S_k(t+1)) \simeq \frac{2\,\mathrm{E}(N_{k-1}(t))}{t}$$

and

$$\mathrm{E}(D_k(t+1)) \simeq \frac{2\,\mathrm{E}(N_k(t))}{t}$$

giving, for $d_k(t) = (\mathrm{E}(N_k(t))$

$$d_k(t+1) \simeq d_k(t) + \frac{2d_{k-1}(t)}{t} - \frac{2d_k(t)}{t}. \tag{4.12}$$

We now solve the approximate equations (4.11) and (4.12). It is reasonable to conjecture a priori that $d_k(t)$ grows linearly in t, i.e. it is of the form $d_k(t) = \rho_k t$. Inserting this in (4.11) we get

$$(t+1)\rho_0 \simeq \delta_0 t + 1 - 2\rho_0 \quad \Rightarrow \quad \rho_0 \simeq \frac{1}{3},$$

and in (4.12)

$$(t+1)\rho_k \simeq t\rho_k + 2\rho_{k-1} - 2\rho_k \quad \Longleftrightarrow \quad \rho_k \simeq \frac{2}{3}\rho_{k-1},$$

giving, in conclusion,

$$\rho_k \simeq \frac{1}{3}\left(\frac{2}{3}\right)^k. \tag{4.13}$$

Thus, the average fraction of vertices of degree k decays exponentially in k. Note that this is very different from the case of $G(n, p_n)$, where the

$$\rho_k = \frac{\mathrm{E}(N_k)}{n} = \binom{n}{k} p_n^k (1-p_n)^{n-k}. \tag{4.14}$$

The expressions (4.13) and (4.14) can be compared for $p_n = \frac{c}{n}$, so that both graphs have a number of edges of the order of the number of sites. In this case, for $G(n, c/n)$

$$\rho_k \simeq e^{-c}\frac{c^k}{k!},$$

which decays faster than exponentially in k. Thus the growing graphs we have just introduced have a degree distribution with a slower decay than that of $G(n, c/n)$. The data for many real graphs show that they have a degree distribution with an even slower decay, namely ρ_k decays as a power of $\frac{1}{k}$. A model with this property is illustrated below.

4.2.2 Networks growing with preferential attachment

We consider here a network evolving as follows. For simplicity, we start at time $t = 2$ with two vertices (labelled by 1 and 2) linked by an edge. At any time a new vertex enters the network, so at time t there are t vertices, and one edge is added, linking the new vertex to one of the others. The vertex $t+1$, that enters at time $t+1$, is linked with vertex i, with $i \leq t$, with probability

$$q_i(t) := \frac{d_i(t) + \delta}{2(t-1) + t\delta},$$

where $\delta > -1$ and $d_i(t)$ is the degree of vertex i at time t. The choice $\delta = 0$ correspond to the "pure" preferential attachment rule: vertex i is chosen with a probability proportional to its degree. Letting $\delta \to +\infty$ all vertices are chosen with the same probability. In other words, for $i \leq t$:

$$\mathrm{P}(d_i(t+1) = k | d_i(t) = k-1) = \frac{k-1+\delta}{2(t-1)+t\delta}$$

and

$$\mathrm{P}(d_i(t+1) = k | d_i(t) = k) = 1 - \frac{k+\delta}{2(t-1)+t\delta}.$$

Denoting by $D_i(t) := \mathrm{E}[d_i(t)]$, we have

$$\begin{aligned}
D_i(t+1) &= \sum_k k\mathrm{P}(d_i(t+1) = k) \\
&= \sum_k k\,[\mathrm{P}(d_i(t+1) = k|d_i(t) = k-1)\,\mathrm{P}(d_i(t) = k-1) + \mathrm{P}(d_i(t+1) = k|d_i(t) = k)\,\mathrm{P}(d_i(t) = k)] \\
&= \sum_k k\left[\frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) + \left(1 - \frac{k+\delta}{2(t-1)+t\delta}\right)\mathrm{P}(d_i(t) = k)\right] \\
&= \sum_k \left[(k-1+1)\frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) + k\mathrm{P}(d_i(t) = k) - k\frac{k+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k)\right] \\
&= \sum_k (k-1)\frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) + \sum_k \frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) \\
&\quad + \sum_k k\mathrm{P}(d_i(t) = k) - \sum_k k\frac{k+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k) \\
&= \frac{D_i(t)-1+\delta}{2(t-1)+t\delta} + D_i(t),
\end{aligned}$$

where we have used the facts that

$$\sum_k (k-1)\frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) = \sum_k k\frac{k+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k)$$

and

$$\sum_k \frac{k-1+\delta}{2(t-1)+t\delta}\mathrm{P}(d_i(t) = k-1) = \mathrm{E}\left[\frac{d_i(t)-1+\delta}{2(t-1)+t\delta}\right] = \frac{D_i(t)-1+\delta}{2(t-1)+t\delta}.$$

Thus:

$$D_i(t+1) - D_i(t) = \frac{D_i(t)-1+\delta}{2(t-1)+t\delta}, \tag{4.15}$$

which is a difference equation, to be solved with initial condition at time $d-i$: $D_i(i) = 1$. Although this equation could be dealt with, we rather consider the associated differential equation, which is easier and behaves similarly:

$$\frac{d}{dt}D_i(t) = \frac{D_i(t) - 1 + \delta}{2(t-1) + t\delta}$$
$$D_i(i) = 1$$

whose unique solution is given by

$$D_i(t) = 1 + \delta\left[\left(\frac{t(2+\delta)-2}{i(2+\delta)-2}\right)^{\frac{1}{1+\delta}} - 1\right] \simeq 1 + \delta\left[\left(\frac{t}{i}\right)^{\frac{1}{1+\delta}} - 1\right]$$

where this last approximation is for i and t large. Note that, using this approximation,

$$D_i(t) \geq k \;\text{ whenever }\; \frac{i}{t} \leq \left(1 + \frac{k-1}{\delta}\right)^{-(1+\delta)}.$$

Thus, at a large time t, the fraction of vertices whose *average degree* is larger than k is approximately $\left(1 + \frac{k-1}{\delta}\right)^{-(1+\delta)}$. A (harder!) analysis of the variance of $d_i(t)$ allows to show that the previous statement holds without averaging: if $\sigma_k(t)$ denotes the (random) fraction of vertices with degree $\geq k$ at time t, then

$$\lim_{t\to+\infty} \sigma_k(t) = \left(1 + \frac{k-1}{\delta}\right)^{-(1+\delta)}.$$

It follows that the fraction $\rho_k(t)$ of vertices having degree k is

$$\rho_k(t) = \sigma_k(t) - \sigma_{k+1}(t) \to \left(1 + \frac{k-1}{\delta}\right)^{-(1+\delta)} - \left(1 + \frac{k}{\delta}\right)^{-(1+\delta)} \simeq \frac{\text{const.}}{k^{(2+d)}}$$

for k large. So the degree distribution has a power low decay!

Note that the behavior is quite different if we let $\delta \to +\infty$ in (4.15), i.e. we remove the preferential attachment. Indeed, in the continuous approximation we would get

$$\frac{d}{dt}D_i(t) = \frac{1}{t}$$
$$D_i(i) = 1$$

whose unique solution is given by

$$D_i(t) = 1 + \log\left(\frac{t}{i}\right),$$

giving

$$D_i(t) \geq k \;\text{ whenever }\; \frac{i}{t} \leq e^{-(k+1)}.$$

The same argument as above gives

$$\lim_{t\to+\infty}\sigma_k(t) = e^{-(k+1)},$$

so that

$$\rho_k(t) = \sigma_k(t) - \sigma_{k+1}(t) \to e^{-(k+1)} - e^{-(k+2)} = \text{const}.e^{-k},$$

so the degree distribution has an exponential decay.