An introduction to discrete-time stochastic processes and their applications

Stefano Pagliarani [1]

This version: March 20, 2024

[1]Dipartimento di Matematica, Università di Bologna, Bologna, Italy. **e-mail**: stefano.pagliarani9@unibo.it.

Abbreviations

In these notes the following (standard) abbreviations will be employed:

- r.v.: random variable;
- i.i.d.: independent and identical distributed;
- w.r.t.: with respect to;
- pdf: probability density function;
- pmf: probability mass function;
- cdf: cumulative distribution function;
- mgf: moment generating function;
- cgf: cumulant generating function;
- LLN: law of large numbers;
- CLT: central limit theorem.

Notations

In these notes the following notations will be employed:

- $\mathcal{P}(\Omega)$: power-set of a non-empty set Ω
- $\mathbb{E}_n[\cdot]$: conditional expectation with respect to the σ-algebra $\mathcal{F}_n$, with $(\mathcal{F}_n)$ given filtration
- $m\mathcal{F}$: measurable functions w.r.t. the σ-algebra $\mathcal{F}$
- $m\mathcal{F}^+$: non-negative real-valued measurable functions w.r.t. the σ-algebra $\mathcal{F}$
- $bm\mathcal{F}$: bounded measurable functions w.r.t. the σ-algebra $\mathcal{F}$
- $\mathcal{B} = \mathcal{B}(S)$: Borel σ-algebra on a given subset S of a Euclidian space
- $m\mathcal{B}$: measurable functions w.r.t. a Borel σ-algebra
- $m\mathcal{B}^+$: non-negative real-valued measurable functions w.r.t. a Borel σ-algebra
- $bm\mathcal{B}$: bounded measurable functions w.r.t. a Borel σ-algebra

Contents

Chapter 1

Conditional probability and independence

The term *conditioning* in Probability has a twofold meaning.

In the first place, it may refer to the operation of updating a probability measure, when the uncertainty of a certain event is resolved. The distributions of the random variables and quantities depending on it, such as the expected value, are updated accordingly. In this case, we talk about *conditional probability*, *conditional distribution (or law)* and *conditional expectation* given (or w.r.t.) an event. The conditioning event could be given, for instance, by the outcome of a certain random variable being equal to a certain value. When the a priori probability of an event A is equal to the conditional probability of A given B, the two events are said independent. Independence is a central concept in Probability theory[1] and can be extended to a family of events, as well as to a family of random variables.

Alternatively, one might consider a partition of events that will be observed in the future. The uncertainty related to these events is not yet resolved, and we want to describe how the reference probability measure will change depending on which event of the partition will occur. To each of these events, we can associate the conditional probability given that event. This leads to the definition of a random probability measure, which is called *conditional probability* given (or w.r.t.) a partition. Following the same idea, one can define the *conditional law* and *conditional expectation* of a random variable given (or w.r.t.) a partition. These are all random variables, which will be observed when the uncertainty of the partition will be resolved. As a notable particular case, consider the partition generated by the possible outcomes of a random variables: in this case we talk about conditioning w.r.t. a random variable.

As we shall see, the conditional probability w.r.t. a negligible event cannot be defined in a trivial way. As a consequence, neither the conditional probability with respect to a non-countable partition can be defined with a direct argument as prescribed above. In order to overcome this problem, one needs to introduce the conditional expectation given a σ-algebra, the latter being a less intuitive concept but quite effective in order to define the conditional expectation with respect to negligible events in a meaningful way.

Hereafter, throughout this chapter, $(\Omega, \mathcal{F}, \mathbb{P})$ will denote a generic probability space.

[1] According to many, the concept of independence is what really distinguishes Kolmogorov's Probability theory from measure theory.

1.1 Conditioning with respect to an event

We introduce the concepts of this section with the following

Example 1.1.1. We play the following betting game. Two six-faced balanced dice are rolled: if the sum of the two dice is less or equal than 6 we win 1 Euros, otherwise we loose 1 Euro. We can then represent the physical phenomenon with the probability space give by

$$\Omega := \big\{(i,j) : i,j \in \{1,2,3,4,5,6\}\big\} = \{1,2,3,4,5,6\}^2, \qquad \mathcal{F} = \mathcal{P}(\Omega), \qquad \mathbb{P} = \mathrm{Unif}_\Omega.$$

In order to represent the event of winning the bet, we introduce the random variables

$$X_1(\omega_1,\omega_2) := \omega_1, \qquad X_2(\omega_1,\omega_2) := \omega_2$$

representing, respectively, the outcome of the first and the second die. The event A of winning the bet and the gain random variable Y are thus given by

$$A = \{X_1 + X_2 \leq 6\}, \qquad Y = \mathbf{1}_A - \mathbf{1}_{A^c}.$$

A simple computation yields $\mathbb{P}(A) = \frac{5}{12}$ and $\mathbb{E}[Y] = -\frac{1}{6}$. [2] Assume now, that the two dice are rolled sequentially, as opposed to simultaneously, and that we observe the event $\{X_1 = 5\}$ after the first roll. Now, only a fool would not adjust his view about his winning probability and expected gain in light of the knowledge that the latter even occurred. The physical independence of the two dice rolles, which is implicit in our experiment, would suggest an intuitive way to update the probability of the event A. Given $X_1 = 5$, the one and only outcome compatible with the occurrence of A is $X_2 = 1$. Therefore, being the probability of the outcomes of the second die independent of the outcomes of the first one, we would conclude that the "new" probability of A is given by

$$\mathbb{P}(A|\{X_1 = 5\}) = \mathbb{P}(\{X_2 = 1\}|\{X_1 = 5\}) = \mathbb{P}(\{X_2 = 1\}) = \frac{1}{6}. \tag{1.1.1}$$

Consequently, the "new" expected value of the gain would be

$$\mathbb{E}[Y|\{X_1 = 5\}] = \frac{1}{6} - \frac{5}{6} = -\frac{2}{3}. \tag{1.1.2}$$

Looking at Example 1.1.1, a first question arises: had the two rolls not been independent, how would we have computed the conditional probability of $\{X_2 = 1\}$ given the knowledge of $\{X_1 = 5\}$? This is surely a relevant question, but the reader might soon be compelled with an even more urgent issue. Here we used the independence of the events $\{X_1 = 5\}$ and $\{X_2 = 1\}$ to define the conditional probability of the former w.r.t. the latter. But what is the mathematical definition of independence between two events? Shall not instead independence be defined by the very fact that the conditional probability coincides with the a priori probability? If it is so, how do we define the conditional probability in general?

Definition 1.1.2. Let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The number

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

is called the *probability of A given (or conditioned to) B*.

[2] The game is not fair: it is advantageous for the house, as it always happens for casino games.

Remark 1.1.3 (Bayes formula)**.** Let $A, B \in \mathcal{F}$ with $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{1.1.3}$$

As a simple corollary of the formula above, we have the following famous

Theorem 1.1.4 (Bayes Theorem)**.** *Let I be a countable set, $(A_i)_{i\in I}$ be a partition of non-negligible events of Ω, and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then we have*

$$\mathbb{P}(A_i) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j\in I}\mathbb{P}(B|A_j)\mathbb{P}(A_j)}, \qquad i \in I.$$

Proof. The equality stems from (1.1.3) together with the total probability formula $\mathbb{P}(B) = \sum_{j\in I}\mathbb{P}(B|A_j)\mathbb{P}(A_j)$. □

Definition 1.1.5. Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The function

$$\mathcal{F} \ni A \mapsto \mathbb{P}|_B(A)$$

is called the *conditional probability* $\mathbb{P}$ *given (or w.r.t.)* B.

Remark 1.1.6. The term *probability* in the definition above is justified because the function $\mathbb{P}|_B$ is a probability measure on $(\Omega, \mathcal{F})$.

In light of the previous remark, the following definition is well-posed.

Definition 1.1.7. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. The number

$$\mathbb{E}[X|B] = \mathbb{E}_{\mathbb{P}|_B}[X] := \int_\Omega X d\mathbb{P}|_B$$

is called the *conditional expectation of X given (w.r.t. to)* B.

Example 1.1.8. Consider the betting game of Example 1.1.1. We would like to recover (1.1.1) and (1.1.2), which were obtained heuristically based on the supposed independence of the two rolls, using Definition 1.1.2 and Definition 1.1.7. We have

$$\begin{aligned}\mathbb{P}(A|\{X_1 = 5\}) &= \frac{\mathbb{P}(A \cap \{X_1 = 5\})}{\mathbb{P}(\{X_1 = 5\})} = \frac{\mathbb{P}(\{X_2 = 1\} \cap \{X_1 = 5\})}{\mathbb{P}(\{X_1 = 5\})} \\ &= \frac{\mathbb{P}(\{(5,2)\})}{\mathbb{P}(\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\})} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.\end{aligned}$$

Now, recalling that the gain of the bet is $Y = \mathbf{1}_A - \mathbf{1}_{A^c}$, we have

$$\mathbb{E}[Y|\{X_1 = 5\}] = \mathbb{E}[\mathbf{1}_A|\{X_1 = 5\}] - \mathbb{E}[\mathbf{1}_{A^c}|\{X_1 = 5\}] = \mathbb{P}(A|\{X_1 = 5\}) - \mathbb{P}(A^c|\{X_1 = 5\}) = \frac{1}{6} - \frac{5}{6} = -\frac{2}{3}.$$

Note that, as opposed to what we did in Example 1.1.1, in order to compute the conditional probability and the conditional expectation, we did not have to rely on the independence of the two dice rolls. However, as we shall see in Section 1.2.1, the independence is still present in the above computations as it is "hidden" in the choice of the uniform probability $\mathbb{P}$ on Ω.

The following lemma can be proved following the canonical procedure, which is: (i) prove it first for simple random variables, (ii) then for non-negative measurable r.v.'s, and finally for summable r.v.'s. The details are left for exercise.

Lemma 1.1.9. *For any $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$ we have*

$$\mathbb{E}[X|B] = \frac{1}{\mathbb{P}(B)} \int_B X d\mathbb{P}.$$

Analogously, we can introduce the concept of conditional covariance, and thus of conditional variance, given the event B.

Definition 1.1.10. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a random vector and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. The matrix

$$cov(X|B) := \mathbb{E}\big[(X - \mathbb{E}[X|B])(Y - \mathbb{E}[Y|B])^* \big| B\big]$$

is called the *conditional covariance matrix of X given (or w.r.t. to) B*. Accordingly, if $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a scalar r.v., the number

$$var(X|B) := \mathbb{E}\big[(X - \mathbb{E}[X|B])^2 \big| B\big]$$

is called the *conditional variance of X given (or w.r.t. to) B*.

1.2 Independence

1.2.1 Independence and correlation between events

Consider two events $A, B \in \mathcal{F}$. The intuitive notion of independence is that the knowledge of the occurrence of A does not change the probability of B, and viceversa. To put it in mathematical terms, this corresponds to

$$\mathbb{P}(A|B) = \mathbb{P}(A) \qquad \text{and} \qquad \mathbb{P}(B|A) = \mathbb{P}(B). \tag{1.2.1}$$

However intuitive, this concept of independence gives rise to both technical and conceptual issues when either A or B have null probability.

For instance, if $\mathbb{P}(B) = 0$, then the probability $\mathbb{P}(A|B)$ cannot be defined as in Definition 1.1.2. In other axiomatic approaches to Probability, such as the one based on *coherence*, the conditional probability can be defined regardless the probability of the conditioning event. The representation (1.1.2) for $\mathbb{P}(A|B)$ can then be proved in the particular case of $\mathbb{P}(B) > 0$, and it is referred to as the *Theorem of the conditional expectation*. In this approach, the independence between two arbitrary events A and B is defined exactly by (1.2.1), together with the (pre-)requirement that A and B are logically independent (how can two events be considered independent if the occurrence of one implies the occurrence, or the non-occurrence, of the other?). Indeed, the logical independence is not granted by (1.2.1) if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

The coherent approach to conditional expectation, and to independence, is based on the notion of conditional event $A|B$, which cannot be defined within Kolmogorov's framework. Furthermore, however meaningful from the probabilistic point of view, the utility of this method is doubtful when it comes to actually computing $\mathbb{P}(A|B)$ in the case when $\mathbb{P}(B) = 0$. In Kolmogorov's approach, a more operational notion of independence is preferred.

Definition 1.2.1. Two events $A, B \in \mathcal{F}$ are said independent if

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

Remark 1.2.2. It is trivial to verify that, if $\mathbb{P}(A), \mathbb{P}(B) > 0$, then (1.2.1) holds true, and therefore we recover the probabilistic intuitive notion of independence. However, A and B result independent so long as either $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$. In other words, the definition is meaningless when negligible events are involved. In particular, A and B could be independent according to Definition 1.2.1 and logically dependent at the same time!

Example 1.2.3. Consider again the betting game of Example 1.1.1 and Example 1.1.8. We want to check that the events $\{X_1 = 5\}$ and $\{X_2 = 1\}$ are independent under $\mathbb{P} = Unif_\Omega$. We have

$$\begin{aligned}\mathbb{P}(\{X_1 = 5\} \cap \{X_2 = 1\}) &= \mathbb{P}\big(\{(5,1)\}\big) = \frac{1}{36} = \frac{1}{6}\frac{1}{6} \\ &= \mathbb{P}\big(\{(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\}\big)\,\mathbb{P}\big(\{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)\}\big) \\ &= \mathbb{P}(\{X_1 = 5\})\,\mathbb{P}(\{X_2 = 1\}).\end{aligned}$$

A careful analysis of the computations above reveals that, in order for any pair of events of the type

$$\{X_1 \in H_1\}, \quad \{X_2 \in H_2\}, \qquad H_1, H_2 \subset \{1,2,3,4,5,6\}$$

to be independent, a necessary and sufficient condition is that the probability $\mathbb{P}$ on $\Omega = \{1,2,3,4,5,6\}^2$ is given by

$$\mathbb{P}(i,j) = \mathbb{P}_1(i)\,\mathbb{P}_2(j), \qquad (i,j) \in \Omega,$$

with $\mathbb{P}_1, \mathbb{P}_2$ probabilities on $\{1,2,3,4,5,6\}$.

Definition 1.2.4. Two events $A, B \in \mathcal{F}$ such that $\mathbb{P}(A), \mathbb{P}(B) > 0$ are said positively (negatively) correlated if

$$\mathbb{P}(A|B) > \mathbb{P}(A).$$
$$(<)$$

Remark 1.2.5. Both positive and negative correlation between events are a symmetric properties. Indeed, if $\mathbb{P}(A), \mathbb{P}(B) > 0$, Bayes formula simply yields

$$\mathbb{P}(A|B) > \mathbb{P}(A) \Leftrightarrow \mathbb{P}(B|A) > \mathbb{P}(B).$$

We now employ the definition of independence between two events to introduce the independence between events of one family, and the independence between families of events. Through this section I always denotes a generic non-empty set.

Definition 1.2.6. Let $(A_i)_{i\in I}$ be a family of events. These events are said independent if

$$\mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbb{P}(A_j)$$

for any finite $J \subset I$.

Remark 1.2.7. If the events of a family $(A_i)_{i\in I}$ are independent, then they are pair-wise independent, i.e.

$$\mathbb{P}\big(A_i \cap A_j\big) = \mathbb{P}(A_i)\mathbb{P}(A_j), \qquad i \neq j,$$

but the converse is not true.

Example 1.2.8. Let $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ be the sample space related to the toss of two coins. Assuming the two coins to be balanced and the tosses independent, we set $\mathbb{P} = Unif_\Omega$. Consider now the events

$$A = \{(H,H),(H,T)\} = \{\text{heads for the first coin}\},$$
$$B = \{(H,H),(T,H)\} = \{\text{heads for the second coin}\},$$
$$C = \{(H,T),(H,T)\} = \{\text{heads for only one coin}\}.$$

We have

$$\mathbb{P}(A\cap B) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B), \qquad \mathbb{P}(B\cap C) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C), \qquad \mathbb{P}(A\cap C) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C),$$

and thus the events are pair-wise independent. However,

$$\mathbb{P}(A\cap B\cap C) = 0 \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C),$$

and thus $\{A,B,C\}$ is not a family of independent events.

Definition 1.2.9. Let $(\mathcal{A}_i)_{i\in I}$ be a collection of families of events, i.e. each $\mathcal{A}_i \subset \mathcal{F}$. These families are said independent if

$$\mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J}\mathbb{P}(A_j)$$

for any finite $J \subset I$ and any family $(A_j)_{j\in J}$ with $A_j \in \mathcal{A}_j$.

1.2.2 Independence between random variables

Definition 1.2.10. Let $(X_i)_{i\in I}$ be a family of random variables. These are said independent if $(\sigma(X_i))_{i\in I}$ is a collection of independent σ-algebras.

Remark 1.2.11. By Definition 1.2.10 and Definition 1.2.9, $(X_i)_{i\in I}$ is a family of independent random variables if and only if, for any finite subset of indices $J = \{j_1,\cdots,j_n\} \subset I$, we have

$$\mathbb{P}\big((X_{j_1}\in H_1)\cap\cdots\cap(X_{j_n}\in H_n)\big) = \mathbb{P}(X_{j_1}\in H_1)\times\cdots\times\mathbb{P}(X_{j_n}\in H_n), \qquad H_1,\cdots,H_n\in\mathcal{B},$$

which is in turn equivalent to

$$\mu_{(X_{j_1},\cdots,X_{j_n})} = \mu_{X_{j_1}}\otimes\cdots\otimes\mu_{X_{j_n}}.$$

Remark 1.2.12. [Discrete case] By Remark 1.2.11, $(X_n)_{n\in\mathbb{N}}$ is a family (sequence) of independent random variables if and only if

$$\mu_{(X_1,\cdots,X_n)} = \mu_{X_1}\otimes\cdots\otimes\mu_{X_n}, \qquad n\in\mathbb{N}.$$

Example 1.2.13. [Canonical construction] Let $I = \{1,\cdots,N\}$ and $(\mu_n)_{n=1,\cdots,N}$ be a family of distributions on $\mathcal{B}(\mathbb{R}^d)$. Set

$$\Omega = \underbrace{\mathbb{R}^d\times\cdots\times\mathbb{R}^d}_{N\text{ times}} = \mathbb{R}^{Nd}, \qquad \mathcal{F} = \underbrace{\mathcal{B}(\mathbb{R}^d)\otimes\cdots\otimes\mathcal{B}(\mathbb{R}^d)}_{N\text{ times}} = \mathcal{B}(\mathbb{R}^{Nd}), \qquad \mathbb{P} = \mu_1\otimes\cdots\otimes\mu_N,$$

and

$$X_n(\omega) = X_n(\omega_1,\cdots,\omega_N) := \omega_n, \qquad \omega\in\Omega, \quad n = 1,\cdots,N.$$

Then, we clearly have

$$\mu_{(X_1,\cdots,X_N)} = \mu_1 \otimes \cdots \otimes \mu_N.$$

In light of Remark 1.2.12, $(X_n)_{n=1,\cdots,N}$ is a sequence of independent random variables such that $X_n \sim \mu_n$ for any n. If $\mu_1 = \cdots = \mu_N$, then the random variables X_n are also identically distributed.

Remark 1.2.14. The canonical construction in Example 1.2.13 can be extended to construct a family $(X_i)_{i\in I}$ of independent random variables with pre-assigned distributions $(\mu_i)_{i\in I}$, for an arbitrary index set I (see Proposition 2.2.20 for $I = \mathbb{N}$). In order to do so, one has to make sense of the, possibly infinite, product space given by

$$\Omega = (\mathbb{R}^d)^I, \qquad \mathcal{F} = \mathcal{B}(\mathbb{R}^d)^{\otimes I} = \mathcal{B}(\mathbb{R}^d)^I, \qquad \mathbb{P} = \mu_i \,^{\otimes i\in I}.$$

Interpreting Ω as the set of all the functions $\omega : I \to \mathbb{R}^d$, $\mathcal{F}$ can be defined as the smallest σ-algebra generated by the finite cylinders

$$C_{i_1,\cdots,i_n}(H) := \big\{\omega \in \Omega : \big(\omega(i_1),\cdots,\omega(i_n)\big) \in H\big\}, \qquad i_1,\cdots,i_n \in I, \quad H \in \mathcal{B}(\mathbb{R}^{n\times d}).$$

The construction of $\mathbb{P}$ is achieved by means of the so-called Kolmogorov Extension Theorem (see Theorem 2.2.18 for particular case of $I = \mathbb{N}$). Basically, $\mathbb{P}$ is the only probability measure on $\mathcal{F}$ such that

$$\mathbb{P}\big(C_{i_1,\cdots,i_n}(H)\big) = \big(\mu_{i_1} \otimes \cdots \otimes \mu_{i_n}\big)(H), \qquad i_1,\cdots,i_n \in I, \quad H \in \mathcal{B}(\mathbb{R}^{n\times d}),$$

thus ensuring that the canonical family $(X_i)_{i\in}$ defined through

$$X_i(\omega) = \omega_i, \qquad \omega \in \Omega, \quad i \in I$$

is a family on independent random variables.

1.3 Conditioning with respect to a countable partition

In the previous section we introduced the concepts of conditional probability and conditional expectations given the occurrence of one specific event. As opposed to one single event, we consider now a countable partition $\mathcal{P}$ of non-negligeble events. These events can be thought as directly related to a partial realization of the random phenomenon, e.g. the realization of a certain random variable, which is observed at some time after the beginning of the experiment. In order to model the way the probabilities will be updated, it does makes sense to define a conditional probability, and thus conditional expectations, for each event in the partition $\mathcal{P}$. This family of conditional probabilities will then define a random probability measure, which will be observed only when the uncertainty regarding the partition $\mathcal{P}$ will be resolved, meaning that the truth value of each of its events will be revealed.

In order to make this concept more tangible, we resume the example of the dice-rolls.

Example 1.3.1. Consider the betting game of Example 1.1.1. Assume now that we have not yet figured out the outcome of the first die. However, we want to describe the how the probability $\mathbb{P}$, and with that the expectation of our gain Y, updates in time, as the outcomes of X_1 and X_2 are revealed to us. We then consider three different times, or stages:

- $t = 0$: non of the dice have been rolled. The partition observable at the this stage is the trivial one, i.e. $\mathcal{P}_0 = \{\Omega\}$.

- $t = 1$: only the first dice has been rolled. At this stage we discover the time value of X_1. The partition observable is

$$\mathcal{P}_1 = \big\{\{X_1 = i\}, i = 1, \cdots, 6\big\},$$

- $t = 2$: the second dice has been rolled too. We now discover the value of X_2. The partition observable at this stage is the finest possible, i.e. $\mathcal{P}_2 = \{(i,j)\}_{i,j=1,\cdots,6}$.

Our reference probability at $t = 0$ is then $\mathbb{P}_0 = \mathbb{P}$, i.e. the a priori probability. As we already saw in Example 1.1.1 and Example 1.1.8, under this probability we have

$$\mathbb{E}_{\mathbb{P}_0}[Y] = \mathbb{E}[Y] = -\frac{2}{3}.$$

At time $t = 2$, the information available to us will be total. Precisely, we will be able to observe the realized pair $\{(i,j)\} \in \Omega$. Therefore, our conditional probability at his stage will collapse onto such pair. The family $\mathbb{P}_2$ of possible conditional probabilities (one for each possible pair) forms then a random probability measure, observable only at time $t = 2$, which concentrates the whole mass onto the realized pair. Precisely,

$$\mathbb{P}_2(i,j) = \delta_{(i,j)}, \qquad (i,j) \in \Omega.$$

Consequently, also the expectation of Y a $t = 2$, meant as the expectation of Y w.r.t. $\mathbb{P}_2$, is a random variable, which coincides with Y itself, i.e.:

$$\mathbb{E}_{\mathbb{P}_2(i,j)}[Y] = Y(i,j), \qquad (i,j) \in \Omega.$$

We now go on to consider the intermediate stage $t = 1$. In this case the information available to us is only partial. If the event $\{X_1 = i\}$ is realized, the initial probability $\mathbb{P}$ will be updated by $\mathbb{P}|_{\{X_1=i\}}$. Since the outcome of the first roll is random, the family $\mathbb{P}_1$ of such conditional probabilities forms a random probability measure, observable at $t = 1$, as it only depends on the outcome of the first roll. Precisely:

$$\mathbb{P}_1(i,j) = \mathbb{P}|_{\{X_1=i\}}, \qquad (i,j) \in \Omega.$$

Consequently, the expectation of Y a $t = 1$, meant as the expectation of Y w.r.t. $\mathbb{P}_1$, is a random variable observable at $t = 1$, i.e.:

$$\mathbb{E}_{\mathbb{P}_1(i,j)}[Y] = \mathbb{E}[Y|\{X_1 = i\}], \qquad (i,j) \in \Omega.$$

Note that, by definition, the random measure $\mathbb{P}_1$ and the random variable $\mathbb{E}_{\mathbb{P}_1}[Y]$ are constant on the events of the partition $\mathcal{P}_1$. In particular, they are measurable[3] with respect to the σ-algebra generated by this partition. Note also that $\mathbb{P}_0$ ($\mathbb{E}_{\mathbb{P}_0}[Y]$) and $\mathbb{P}_2$ ($\mathbb{E}_{\mathbb{P}_2}[Y]$) are constant on the trivial partitions $\mathcal{P}_0 = \{\Omega\}$ and $\mathcal{P}_2 = \big\{\{\omega\}, \omega \in \Omega\big\}$, respectively.

Hereafter, throughout this section, we denote by I a countable set and by

$$\mathcal{P} := (A_i)_{i \in I}$$

a countable partition of non-negligible events of Ω. Precisely, this means that

$$\Omega = \biguplus_{i \in I} A_i \qquad \text{and} \qquad \mathbb{P}(A_i) > 0, \quad i \in I. \tag{1.3.1}$$

[3]Here we mean that $\mathbb{P}_1(A) \in m\sigma(\mathcal{P})$ for any $A \in \mathcal{F}$.

Definition 1.3.2. For any $F \in \mathcal{F}$, the function

$$\Omega \ni \omega \mapsto \sum_{i \in I} \mathbf{1}_{A_i}(\omega)\, \mathbb{P}(F|A_i) =: \mathbb{P}(F|\mathcal{P})(\omega) = \mathbb{P}|_{\mathcal{P}}(\omega)(F) \tag{1.3.2}$$

is called the *conditional probability of F given (w.r.t. to) $\mathcal{P}$*. In light of Remark 1.1.6, the function

$$\mathcal{F} \ni F \mapsto \mathbb{P}|_{\mathcal{P}}(\omega)(F)$$

is a probability measure for any $\omega \in \Omega$. Thus, the function

$$\Omega \ni \omega \mapsto \mathbb{P}|_{\mathcal{P}}(\omega)$$

is a random probability measure and is called the *conditional probability $\mathbb{P}$ given (or w.r.t.) $\mathcal{P}$*.

Remark 1.3.3. Note that the definition above is well posed because of (1.3.1). On the one hand, the fact that all the events A_i are non negligible implies that all the conditional probabilities $\mathbb{P}(F|A_i)$ are well defined. The fact that $\mathcal{P}$ is a partition implies that $\mathbb{P}(F|\mathcal{P})(\omega)$ is necessary equal to $\mathbb{P}(F|A_i)$ for one and only one $i \in I$.

Example 1.3.4. Looking at Example 1.3.1 above, we can now see that the random probabilities $\mathbb{P}_0$, $\mathbb{P}_1$, $\mathbb{P}_2$ coincide with the conditional probabilities $\mathbb{P}|_{\mathcal{P}_0}$, $\mathbb{P}|_{\mathcal{P}_1}$, $\mathbb{P}|_{\mathcal{P}_2}$, respectively.

Definition 1.3.5. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. The function

$$\Omega \ni \omega \mapsto \mathbb{E}[X|\mathcal{P}](\omega) := \mathbb{E}_{\mathbb{P}|_{\mathcal{P}}(\omega)}[X]$$

is called the *conditional expectation of X given (w.r.t. to) $\mathcal{P}$*.

Remark 1.3.6. By (1.3.2) we can write

$$\mathbb{P}|_{\mathcal{P}}(\omega) = \sum_{i \in I} \mathbf{1}_{A_i}(\omega)\, \mathbb{P}|_{A_i}, \qquad \omega \in \Omega,$$

and thus we have

$$\mathbb{E}[X|\mathcal{P}] = \sum_{i \in I} \mathbf{1}_{A_i}\, \mathbb{E}[X|A_i].$$

Remark 1.3.7. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then it is obvious that $\mathbb{E}[X|\mathcal{P}]$ is a random variable. In particular, the following two properties hold true:

(i) $\mathbb{E}[X|\mathcal{P}] \in m\sigma(\mathcal{P})$;

(ii) for any $A \in \sigma(\mathcal{P})$ we have

$$\mathbb{E}[X|A] = \mathbb{E}\big[E[X|\mathcal{P}]\big|A\big].$$

While (i) is immediate, in order to see (ii) we need to observe that any event $A \in \sigma(\mathcal{P})$ can be written as

$$A = \biguplus_{j \in J} A_j \qquad \text{for a} \quad J \subset I.$$

Thus

$$\begin{aligned}\mathbb{E}\big[E[X|\mathcal{P}]\big|A\big] &= \frac{1}{\mathbb{P}(A)} \int_A E[X|\mathcal{P}]d\mathbb{P} = \frac{1}{\mathbb{P}(A)} \int_\Omega \Big(\sum_{j \in J} \mathbf{1}_{A_j}\, \mathbb{E}[X|A_j] \Big) d\mathbb{P} = \frac{1}{\mathbb{P}(A)} \sum_{j \in J} \mathbb{P}(A_j)\, \mathbb{E}[X|A_j] \\ &= \frac{1}{\mathbb{P}(A)} \sum_{j \in J} \int_{A_j} X d\mathbb{P} = \frac{1}{\mathbb{P}(A)} \int_A X d\mathbb{P} = \mathbb{E}[X|A].\end{aligned}$$

Remark 1.3.8. For any $F \in \mathcal{F}$ we have $\mathbb{P}(F|\mathcal{P}) = \mathbb{E}[\mathbf{1}_F|\mathcal{P}]$. Therefore, in light of Remark 1.3.7 we also have

(i) $\mathbb{P}(F|\mathcal{P}) \in m\sigma(\mathcal{P})$;

(ii) for any $A \in \sigma(\mathcal{P})$ we have

$$\mathbb{P}(F|A) = \mathbb{E}\big[\mathbb{P}(F|\mathcal{P})\big|A\big].$$

We now consider a particular case of countable partition, which is the one that can be observed by observing a random variable. If $Y \in m\mathcal{F}$ then the family

$$p(Y) := \big\{\{Y = j\}, j \in Y(\Omega)\big\}$$

is a partition. We call it the *partition induced by the random variable* Y. Obviously, if $Y(\Omega)$ is countable then $p(Y)$ is also countable, and so the next definition is well posed.

Definition 1.3.9. Let $Y \in m\mathcal{F}$ with countable values and such that $p(Y)$ does not contain negligible events. For any $F \in \mathcal{F}$ we call the random variable

$$\mathbb{P}(F|Y) := \mathbb{P}(F|p(Y))$$

the *conditional probability of* F *given (w.r.t. to)* Y. We also call the random probability

$$\mathbb{P}|_Y := \mathbb{P}|_{p(Y)}$$

the *conditional probability* $\mathbb{P}$ *given (or w.r.t.)* Y. Finally, for a given $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, the random variable

$$\mathbb{E}[X|Y] := \mathbb{E}[X|p(Y)]$$

is called *conditional expectation of* X *given (w.r.t. to)* Y.

Since the events in $p(Y)$ are the elementary events that are observable by observing the random variable Y, the conditional probability $\mathbb{P}(F|Y)$ describes how the probability $\mathbb{P}$ is updated when the outcome of Y becomes known.

Remark 1.3.10. If follows from the definitions of $\mathbb{P}(F|p(Y))$ and of $\mathbb{E}[X|p(Y)]$, and from Remark 1.3.6, that

$$\mathbb{P}(F|Y) = \mathbb{P}(F|Y = y)|_{y=Y}, \qquad \mathbb{E}[X|Y = y]|_{y=Y}.$$

Example 1.3.11. Looking at Example 1.3.1 above, we can now see that the random probabilities $\mathbb{P}_1$, $\mathbb{P}_2$ coincide with the conditional probabilities $\mathbb{P}|_{X_1}$, $\mathbb{P}|_{(X_1,X_2)}$, respectively. Accordingly, $\mathbb{E}_{\mathbb{P}_1}[Y], \mathbb{E}_{\mathbb{P}_2}[Y]$ coincide, respectively, with $\mathbb{E}_{X_1}[Y], \mathbb{E}_{(X_1,X_2)}[Y]$.

1.4 Conditional expectation with respect to a σ-algebra

However intuitive, the definitions of conditional probability and conditional expectation, given an event or a partition of events, as defined in Sections 1.1 and 1.3 start creaking when the conditioning events are negligible. This cannot be avoided, for instance, when conditioning with respect to an absolutely continuous random variable, like in the following

Example 1.4.1. Consider the betting game in Example 1.1.1 with only one change: instead of two six-faced balanced dice, we roll two $[0,6]$-*continuous* dice. Namely, the probability space is modified as follows:

$$\Omega := [0,6]^2, \qquad \mathcal{F} = \mathcal{B}(\Omega), \qquad \mathbb{P} = \mathrm{Unif}_\Omega.$$

This time, a simple computation yields $\mathbb{P}(A) = \frac{1}{2}$ and thus $\mathbb{E}[Y] = 0$. Again, assume now that the two dice are rolled sequentially and that we observe the event $\{X_1 = 5\}$ after the first roll. Given $X_1 = 5$, the only outcomes compatible with the occurrence of A are given by the event $\{X_2 \leq 1\}$. Therefore, being the two rolls independent, it is intuitive to expect

$$\text{“} \quad \mathbb{P}(A|\{X_1 = 5\}) = \mathbb{P}(\{X_2 \leq 1\}|\{X_1 = 5\}) = \mathbb{P}(\{X_2 \leq 1\}) = \frac{1}{6} \quad \text{”}, \tag{1.4.1}$$

and thus

$$\mathbb{E}[Y|\{X_1 = 5\}] = \frac{1}{6} - \frac{5}{6} = -\frac{2}{3}.$$

However, the quantities and the equalities in (1.4.1) do not make sense in terms of the usual definitions of conditional probability and independence of events. In the first place, the conditional probability $\mathbb{P}(A|\{X_1 = 5\})$ is not well defined because $\mathbb{P}(X_1 = 5) = 0$. Therefore, the first and second equalities in (1.4.1) cannot be justified either. Furthermore, what happens if $\mathbb{P}$ is no longer uniform and is such that X_1 and X_2 are not independent? In this case, even the intuition that lead us through (1.4.1) vanishes. However, it is intuitive that the winning probability has to be somehow updated according to the outcome of the first die.

The previous example shows the need for a rigorous definition of conditional probability, and expectation, with respect to a given negligible event, or with respect to a partition containing negligible events. The mathematical object that allows to achieve this goal is the *conditional expectation with respect to a σ-algebra*.

Definition 1.4.2. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. A random variable Z that satisfies:

(i) $Z \in m\mathcal{G}$,

(ii) for any $A \in \mathcal{G}$ such that $\mathbb{P}(A) > 0$,

$$\mathbb{E}[X|A] = \mathbb{E}[Z|A] \tag{1.4.2}$$

is called a *version of the conditional expectation of X given (or w.r.t.) $\mathcal{G}$*. We also denote by $\mathbb{E}[X|\mathcal{G}]$ the set of such random variables.

Remark 1.4.3. Condition (ii) in Definition 1.4.2 is equivalent to

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[Z\mathbf{1}_A], \qquad A \in \mathcal{G}. \tag{1.4.3}$$

In particular, while this condition coincides with (1.4.2) if $\mathbb{P}(A) > 0$, it is trivially satisfied for any r.v. Z if $\mathbb{P}(A) = 0$.

Remark 1.4.4. In general $\mathbb{E}[X|\mathcal{G}]$ contains more than one random variable. Indeed, if $Z \in \mathbb{E}[X|\mathcal{G}]$ and $Z' \in m\mathcal{G}$ is such that $Z = Z'$ almost surely, then $Z' \in \mathbb{E}[X|\mathcal{G}]$ too. Note that the measurability condition (i) for Z' is not implied simply by $Z \overset{\text{a.s.}}{=} Z'$ unless $\mathcal{G}$ contains all the negligible events of $\mathcal{F}$.

The next result gives us the (almost surely) uniqueness of the conditional expectation with respect to a σ-algebra.

Proposition 1.4.5. *Given $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra, $\mathbb{E}[X|\mathcal{G}]$ is an equivalence class on $m\mathcal{G}$ w.r.t. the equivalence relation given by "$\overset{a.s.}{=}$". In particular, we have*

$$Z, Z' \in \mathbb{E}[X|\mathcal{G}] \Longrightarrow Z = Z' \ \mathbb{P}\text{-a.s. .} \tag{1.4.4}$$

Proof. In light of Remark 1.4.4, we only have to prove (1.4.4). It is not restrictive to only consider X with real values. We prove that, if $X, X' \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $X \leq X'$ almost surely, then

$$Z \leq Z' \text{ a.s.}, \qquad Z \in \mathbb{E}[X|\mathcal{G}],\ Z' \in \mathbb{E}[X'|\mathcal{G}].$$

Assume that $\mathbb{P}(Z > Z') > 0$. This implies

$$0 < \mathbb{E}[(Z - Z')\mathbf{1}_{Z>Z'}] = \mathbb{E}[Z\mathbf{1}_{Z>Z'}] - \mathbb{E}[Z'\mathbf{1}_{Z>Z'}]$$

(by (1.4.3))

$$= \mathbb{E}[X\mathbf{1}_{Z>Z'}] - \mathbb{E}[X'\mathbf{1}_{X>Z'}] = \mathbb{E}[(X - X')\mathbf{1}_{Z>Z'}] \leq 0.$$

□

Remark 1.4.6. The proof of Proposition 1.4.5 actually shows that

$$X \overset{\text{a.s.}}{=} X' \Rightarrow \mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X'|\mathcal{G}].$$

Therefore, the operation of conditioning w.r.t. the σ-algebra $\mathcal{G}$ can be seen as a map from the equivalence classes of $m\mathcal{F}$ to those of $m\mathcal{G}$, i.e.

$$[X] \mapsto \mathbb{E}[X|\mathcal{G}],$$

where $[X]$ is the set of $X' \in m\mathcal{F}$ such that $X = X'$ almost surely.

The next remark connects the notion of conditional expectation given a countable partition, introduced in Section 1.3, with the conditional expectation given a σ-algebra.

Remark 1.4.7. If $\mathcal{G} = \sigma(\mathcal{P})$ with $\mathcal{P}$ countable partition of non-negligible events, then Remark 1.3.7 shows that $\mathbb{E}[X|\mathcal{P}]$ is a version of $\mathbb{E}[X|\mathcal{G}]$. Furthermore, any other version would have to be constantly equal to $\mathbb{E}[X|A_i]$ on all the events $A_i \in \mathcal{P}$, due to the measurability condition (i) in Definition 1.4.2. Therefore $\mathbb{E}[X|\mathcal{P}]$ is the only version of $\mathbb{E}[X|\mathcal{G}]$.

The next result gives us instead the existence of a version of the conditional expectation with respect to an arbitrary σ-algebra.

Theorem 1.4.8. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. There exists a random variable $Z \in \mathbb{E}[X|\mathcal{G}]$.*

Proof. It is not restrictive to consider X with non-negative real values. It can be checked that the function

$$\mathcal{G} \ni G \mapsto \int_G X d\mathbb{P} =: \mathbb{Q}(G)$$

is a finite measure on $(\Omega, \mathcal{G})$. This is due to the assumption $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. In particular, it is obvious that $\mathbb{Q} <<_{|\mathcal{G}} \mathbb{P}$, and thus Theorem A.0.1 (Radon-Nikodym) yields that $\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{G}}$ is not empty. In particular, any $Z \in \frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{G}}$ belongs to $m\mathcal{G}$ and satisfies (1.4.3), therefore it belongs to $Z \in \mathbb{E}[X|\mathcal{G}]$. □

Remark 1.4.9. Note that the proof of Theorem 1.4.8 also yields the (almost surely) uniqueness of the conditional expectation of X with respect to $\mathcal{G}$. However, the proof of Proposition 1.4.5 is more direct and does not require the summability of X.

Remark 1.4.10. It is possible to prove that $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ implies $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ for any $Z \in \mathbb{E}[X|\mathcal{G}]$. Precisely, Z can be seen as an orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$ with respect to the scalar product on $L^2(\Omega, \mathcal{F}, \mathbb{P})$ given by

$$\langle W, Y \rangle_{L^2} := \mathbb{E}[WY], \qquad W, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P}).$$

We refer to [2, Proposition A.112] for a detailed proof.

We now want to prove some important properties of the conditional expectation given a σ-algebra. It is important to remind that, in general, the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ is an equivalence class that contains infinite versions. We thus need to introduce some algebraic operations between equivalence classes.

Definition 1.4.11. Let $X, Y \in m\mathcal{G}$. Then we write:

(i) $[X] \leq [Y]$ if $Z \leq Z'$ almost surely for any $Z \in [X]$ and $Z' \in [Y]$;

(ii) $[X] + [Y]$ to denote the sum between $[X]$ and $[Y]$, namely the equivalence class of the random variables in $m\mathcal{G}$ that can be written as $Z + Z'$ with $Z \in [X]$ and $Z' \in [Y]$;

(iii) $[X]\,[Y]$ to denote the product between $[X]$ and $[Y]$, namely the equivalence class of the random variables in $m\mathcal{G}$ that can be written as $Z\,Z'$ with $Z \in [X]$ and $Z' \in [Y]$;

In order to improve readability, it is useful (and wise) to slightly abuse the notation and put $\mathbb{E}[X|\mathcal{G}]$ at the same level of a random variable. For instance, it is clear by definition that X is a (not the only) version of $\mathbb{E}[X|\sigma(X)]$. However, we will write

$$X = \mathbb{E}[X|\sigma(X)] \qquad \text{in place of} \qquad X \in \mathbb{E}[X|\sigma(X)].$$

In general, we will employ throughout this notes the following

Notation 1.4.12. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra, and let $Z \in m\mathcal{G}$. Then we write:

(a) $Z = \mathbb{E}[X|\mathcal{G}]$ in place of $Z \in \mathbb{E}[X|\mathcal{G}]$;

(b) $Z \leq \mathbb{E}[X|\mathcal{G}]$ ($Z \geq \mathbb{E}[X|\mathcal{G}]$) in place of $[Z] \leq \mathbb{E}[X|\mathcal{G}]$ ($[Z] \geq \mathbb{E}[X|\mathcal{G}]$);

(c) $Z + \mathbb{E}[X|\mathcal{G}]$ in place of $[Z] + \mathbb{E}[X|\mathcal{G}]$;

(d) $Z\,\mathbb{E}[X|\mathcal{G}]$ in place of $[Z]\,\mathbb{E}[X|\mathcal{G}]$.

In the next proposition we list a first set of properties of the conditional expectation. These should be understood in terms of the relations and operations introduced in Definition 1.4.11 and according to the notation adopted above. Equivalently, the reader may regard the following properties, as well as the similar ones that we will encounter in the sequel, as relations that hold, almost surely, for any representative of the equivalence classes.

Proposition 1.4.13. *Let $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then the following properties hold true:*

1. *(linearity) for any $\alpha, \beta \in \mathbb{R}$ we have*

$$\mathbb{E}[\alpha X + \beta Y|\mathcal{G}] = \alpha\,\mathbb{E}[X|\mathcal{G}] + \beta\,\mathbb{E}[Y|\mathcal{G}];$$

2. *(monotoniciy) if $X \leq Y$ a.s. then*

$$\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}];$$

3. *if $X \in m\mathcal{G}$ then*

$$\mathbb{E}[X|\mathcal{G}] = X; \tag{1.4.5}$$

4. *if $\sigma(X)$ and $\mathcal{G}$ are independent, then*

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X];$$

5. *(tower property) if $\mathcal{H} \subset \mathcal{G}$ is another σ-algebra, then*

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big] = \mathbb{E}[X|\mathcal{H}]. \tag{1.4.6}$$

   In particular, taking $\mathcal{H} = \{\Omega, \emptyset\}$ we have the so-called total probability formula*, i.e.*

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big] = \mathbb{E}[X]; \tag{1.4.7}$$

6. *if $Y \in m\mathcal{G}$ is such that $YX, Y\mathbb{E}[X|\mathcal{G}] \in L^1(\Omega, \mathbb{P})$, then*

$$\mathbb{E}[YX|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]. \tag{1.4.8}$$

   In particular, (1.4.8) *holds true if either:*

   (a) $Y \in bm\mathcal{G}$,

   (b) $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Before proving Proposition 1.4.13, we state the following

Lemma 1.4.14. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. A random variable $Z = \mathbb{E}[X|\mathcal{G}]$ if and only if:*

(i) $Z \in m\mathcal{G}$;

(ii)′ we have

$$\mathbb{E}[XW] = \mathbb{E}[ZW]$$

for any $W \in \mathcal{G}$ such that $XW, \mathbb{E}[X|\mathcal{G}]W \in L^1(\Omega, \mathbb{P})$.

Proof. The non-trivial part of the statement is that property (ii)′ above holds true for $Z = \mathbb{E}[X|\mathcal{G}]$. It is not restrictive to assume $W \geq 0$. There exists a sequence of simple random variables $(W_n)_{n\in\mathbb{N}}$ such that $W_n \nearrow W$. Therefore, by Lebesgue dominated convergence theorem we obtain

$$\begin{aligned}\mathbb{E}[XW] &= \lim_{n\to+\infty} \mathbb{E}[XW_n] = \quad \text{(by linearity)}\\ &= \lim_{n\to+\infty} \mathbb{E}[ZW_n] = \mathbb{E}[ZW],\end{aligned}$$

where we used once more Lebesgue dominated convergence theorem in the last equality. □

Proof of Proposition 1.4.13. 1. It stems from the linearity of the conditional expectation.

2. It was already proved in the proof of Proposition 1.4.5.

3. It is straightforward, by definition.

4. Set $Z := \mathbb{E}[X]$. Clearly $Z \in m\mathcal{G}$. Moreover, for any non-negligible $A \in \mathcal{G}$ we have

$$\mathbb{E}[Z|A] = \mathbb{E}[X] = \mathbb{E}[X|A],$$

where in the last equality we used that $\sigma(X)$ and A are independent.

5. Let $Z := \mathbb{E}[X|\mathcal{H}]$ and $Z' := \mathbb{E}[X|\mathcal{G}]$. We have to show that $Z = \mathbb{E}[Z'|\mathcal{H}]$:

   (i) by definition $Z \in m\mathcal{H}$;

   (ii) for any non-negligible $A \in \mathcal{H}$ we have

$$\mathbb{E}[Z|A] = \mathbb{E}[X|A] = \mathbb{E}[Z'|A]$$

   where the last equality holds true because $\mathcal{H} \subset \mathcal{G}$ and thus A is also a non negligible event of $\mathcal{G}$.

6. Set $Z := \mathbb{E}[X|\mathcal{G}]$. We have to prove that $\mathbb{E}[YX|\mathcal{G}] = YZ$:

   (i) by assumption $Y \in m\mathcal{G}$ and so $YZ \in m\mathcal{G}$;

   (ii) by Lemma 1.4.14 with $W = Y\mathbf{1}_A$ we have

$$\mathbb{E}[ZY\mathbf{1}_A] = \mathbb{E}[XY\mathbf{1}_A], \qquad A \in \mathcal{G}.$$

   In particular, $YX, Y\mathbb{E}[X|\mathcal{G}] \in L^1(\Omega, \mathbb{P})$ if $Y \in bm\mathcal{G}$ or, by Remark 1.4.10, if $Y, X \in L^2(\Omega, \mathbb{P})$.

□

The next result is very useful in the theory of stochastic processes, in particular Markov processes.

Lemma 1.4.15 (Freezing lemma)**.** *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be independent σ-algebras and let $X \in m\mathcal{G}$, $f \in m(\mathcal{B} \otimes \mathcal{H})$ such that $f(X, \cdot) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, Then we have*

$$\mathbb{E}[f(X, \cdot)|\mathcal{G}] = \mathbb{E}[f(x, \cdot)]|_{x=X}.$$

For a concise proof, based on Dynkin's Theorem, we refer the reader to [3].

Corollary 1.4.16. *Let X, Y be random vectors such that $X \in m\mathcal{G}$ and $\sigma(Y)$ is independent of $\mathcal{G}$. Let also $f \in m\mathcal{B}$ such that $f(X, Y) \in L^1(\Omega, \mathcal{F}, \mathcal{P})$. Then*

$$\mathbb{E}[f(X, Y)|\mathcal{G}] = \mathbb{E}[f(x, Y)]|_{x=X}.$$

Proof. Clearly, the function $(x, \omega) \mapsto f(x, Y)$ belongs to $m(\mathcal{B} \otimes \mathcal{H})$ with $\mathcal{H} := \sigma(Y)$. Also, by assumption $\mathcal{H}$ and $\mathcal{G}$ are independent. Therefore, the result stems from Theorem 1.4.15. □

We conclude the section with some additional properties, whose proof is basically identical to that of their deterministic counterparts. When they are limits of conditional expectations to another conditional expectation, these should be understood as limits, almost surely, of any selection of representatives.

Proposition 1.4.17. *Let $X \in L^1(\Omega, \mathcal{F}, \mathcal{P})$, $(X_n)_{n\in\mathbb{N}}$ be a sequence of r.v.'s, and $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then:*

1. *(monotone convergence) if $0 \leq X_n \nearrow X$ almost surely, then*
$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}];$$
2. *(Fatou's Lemma) if $X_n \geq 0$ almost surely for any $n \in \mathbb{N}$, then*
$$\mathbb{E}\big[\liminf_{n\to\infty} X_n \big| \mathcal{G}\big] \leq \liminf_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}];$$
3. *(dominated convergence) if $X_n \to X$ almost surely, and if there exists $Y \in L^1(\Omega, \mathcal{F}, \mathcal{P})$ such that $|X_n| \leq Y$ for any $n \in \mathbb{N}$, then*
$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}];$$
4. *(Jensen's inequality) if f is a convex function such that $f(X) \in L^1(\Omega, \mathcal{F}, \mathcal{P})$, then*
$$f\big(\mathbb{E}[X|\mathcal{G}]\big) \leq \mathbb{E}[f(X)|\mathcal{G}].$$
In particular,
$$\|\mathbb{E}[X|\mathcal{G}]\|_p \leq \|X\|_p, \qquad p \geq 1.$$

1.5 Conditional expectation with respect to a random variable

Consider now the particular case when $\mathcal{G} = \sigma(Y)$, with Y random variable. We have the following

Definition 1.5.1. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in m\mathcal{G}$. Then
$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$$
is called the *conditional expectation of X given Y*.

Remark 1.5.2. At a first glance, the notation $\mathbb{E}[X|Y]$ might seems an abuse, as it was already used in Section 1.3 (see Definition 1.3.9) to denote the conditional expectation of X given the partition $\sigma(Y)$, when the latter is countable and contains no negligible events. However, in this case, the notation $\mathbb{E}[X|Y]$ is not ambiguous because $\mathbb{E}[X|p(Y)]$ as defined in Definition 1.3.5 is the only version of $\mathbb{E}[X|\sigma(Y)]$. Furthermore, Remark 1.3.10 yields the representation
$$\mathbb{E}[X|Y] = \mathbb{E}[X|Y = y]|_{y=Y}.$$

In general, by Doob's Theorem, $\mathbb{E}[X|Y]$ always admits a representation of the type $\phi(Y)$. We give the following

Definition 1.5.3. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in m\mathcal{F}$. Any function $\phi \in m\mathcal{B}$ such that
$$\phi(Y) = \mathbb{E}[X|Y]$$
is called a *conditional expectation function of X given Y*, or a *regression function*. The set of all such functions is denoted by $\mathbb{E}[X|Y = \cdot]$.

Proposition 1.5.4. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in m\mathcal{F}$. Then $\mathbb{E}[X|Y = \cdot]$ is not empty and*

$$\phi, \phi' \in \mathbb{E}[X|Y = \cdot] \Longrightarrow \phi = \phi' \quad \mu_Y\text{-a.s.}.$$

Proof. The existence of a regression function stems directly from Doob's Thoerem. Let now $\phi, \phi' \in \mathbb{E}[X|Y = \cdot]$ and assume, by contradiction, that

$$\phi(y) \neq \phi'(y), \qquad y \in H,$$

for a certain $H \in \mathcal{B}$ with $\mu_Y(H) > 0$. Then, there exists $h \in bm\mathcal{B}$ such that

$$\mathbb{E}\big[h\big(\phi(Y)\big)\big] = \int_{\mathbb{R}^d} h\big(\phi(y)\big)\mu_Y(dy) \neq \int_{\mathbb{R}^d} h\big(\phi'(y)\big)\mu_Y(dy) = \mathbb{E}\big[h\big(\phi'(Y)\big)\big].$$

which is a contradiction because $\phi(Y) = \phi'(Y)$ almost surely. □

As a consequence of Proposition 1.5.4, the regression function is univocally determined on the elements that have positive mass under μ_Y. In particular, we have the following

Remark 1.5.5. By Remark 1.5.2, if $p(Y)$ is countable and contains no negligible events, then the function

$$y \mapsto \mathbb{E}[X|Y = y]$$

univocally identifies the elements of $\mathbb{E}[X|Y = \cdot]$ on $Y(\Omega)$. In general, if Y is any random variable and y is such that $\mathbb{P}(Y = y) > 0$, then for any $\phi \in \mathbb{E}[X|Y = \cdot]$ we have

$$\phi(y) = \mathbb{E}[\phi(y)|Y = y] = \mathbb{E}[\phi(Y)|Y = y] = \mathbb{E}\big[\mathbb{E}[X|Y]|Y = y\big] = \mathbb{E}[X|Y = y].$$

Proposition 1.5.4 tells us that $\mathbb{E}[X|Y = \cdot]$ is an equivalence class with respect to the equivalence relation given by the μ_Y-almost sure equality. However, analogously to what we do for $\mathbb{E}[X|Y]$, it is common practice to slightly abuse the notation by treating $\mathbb{E}[X|Y = \cdot]$ as a function. In particular, we will write

$$y \mapsto \mathbb{E}[X|Y = y]$$

to denote an unspecified element of $\mathbb{E}[X|Y = \cdot]$. We stress that, in light of Remark 1.5.5, there is no ambiguity in this notation. Indeed, any element of $\mathbb{E}[X|Y = \cdot]$ computed at y coincides with $\mathbb{E}[X|Y = y]$ as given in Definition 1.1.7, for any y such that $\mathbb{P}(Y = y) > 0$.

We conclude this section with a particular case in which $\mathbb{E}[X|Y = \cdot]$ can be explicitly computed. We recall that, given an absolutely continuous random vector (X, Y) with joint density $\varphi_{(X,Y)}$, the r.v. Y is also absolutely continuous and its density is given by

$$\varphi_Y(y) = \int \varphi_{(X,Y)}(x, y)dx. \tag{1.5.1}$$

Proposition 1.5.6. *Let (X, Y) be an absolutely continuous random vector with joint density $\varphi_{(X,Y)}$. Then, for any $f \in m\mathcal{B}$ such that $f(X, Y) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, we have*

$$\mathbb{E}[f(X, Y)|Y = y] = \int f(x, y)\varphi_{X|Y}(x, y)dx,$$

where the function

$$(x, y) \mapsto \varphi_{X|Y}(x, y) := \begin{cases} \frac{\varphi_{(X,Y)}(x,y)}{\varphi_Y(y)} & \text{if } \varphi_Y(y) > 0, \\ 0, & \text{otherwise} \end{cases} \tag{1.5.2}$$

is called the conditional density of X given Y.

Proof. Set

$$\phi(y) := \int f(x,y)\varphi_{X|Y}(x,y)dx. \tag{1.5.3}$$

We have to prove that $\phi(Y) \in \mathbb{E}[f(X,Y)|Y]$:

(i) obviously $\phi(Y) \in m\sigma(Y)$;

(ii) for any $H \in \mathcal{B}$ we have

$$\mathbb{E}[\phi(Y)\mathbf{1}_{\{Y\in H\}}] = \int \phi(y)\mathbf{1}_H(y)\varphi_{(X,Y)}(x,y)\,(dx\otimes dy)$$

(by Fubini's Theorem, employing $f(X,Y) \in L^1(\Omega,\mathbb{P})$)

$$= \int \phi(y)\mathbf{1}_H(y)\Big(\int \varphi_{(X,Y)}(x,y)dx\Big)dy$$

(by (1.5.3) and (1.5.1))

$$= \int \Big(\int f(z,y)\varphi_{X|Y}(z,y)dz\Big)\mathbf{1}_H(y)\varphi_Y(y)dy$$

(by (1.5.2))

$$= \int \Big(\int f(z,y)\varphi_{(X,Y)}(z,y)dz\Big)\mathbf{1}_H(y)dy$$

(again by Fubini's Theorem)

$$= \int f(z,y)\varphi_{(X,Y)}(z,y)\mathbf{1}_H(y)\,(dz\otimes dy) = \mathbb{E}[f(X,Y)\mathbf{1}_{\{Y\in H\}}].$$

□

Example 1.5.7. We are now able to justify the Heuristics of Example 1.4.1 about the betting game with continuous $[0,6]$-uniform dice. Recalling that

$$A = \{X_1 + X_2 \le 6\},$$

we can write

$$\mathbf{1}_A = f(X_2, X_1),$$

with

$$f(x_2,x_1) = \mathbf{1}_{[0,6]}(x_1+x_2), \qquad g(x_2,x_1) = \mathbf{1}_{[0,6]}(x_1+x_2) - \mathbf{1}_{]6,12]}(x_1+x_2).$$

Now, since (X_2, X_1) is uniformly distributed on $[0,6]^2$, the conditional density of X_2 given X_1 reads as

$$\varphi_{X_2|X_1}(x_2,x_1) = \begin{cases} \frac{1/36}{1/6} = \frac{1}{6} & \text{if } (x_2,x_1)\in[0,6]^2, \\ 0, & \text{otherwise.} \end{cases}$$

Thus one can set

"$\mathbb{P}(A|X_1=5)$" := $\mathbb{E}[\mathbf{1}_A|X_1=5] = \mathbb{E}[f(X_2,X_1)|X_1=5]$

(by Proposition 1.5.6)

$$= \int f(x_2,5)\varphi_{X_2|X_1}(x_2,5)dx_2 = \int \mathbf{1}_{[0,6]}(5+x_2)\,\frac{1}{6}\mathbf{1}_{[0,6]}(x_2)dx_2 = \frac{1}{6}\int \mathbf{1}_{[0,1]}(x_2)dx_2 = \frac{1}{6},$$

which is the result the same result that we reached in (1.4.1) via Heuristic reasoning, based on the independence of the two rolls.

1.6 Exercises

Exercise 1.6.1. Define two random variables X, Y, not independent, such that $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

Chapter 2

Discrete-time stochastic processes

A random random variable with values on a Euclidean space can be thought of as a generalization of the concept of vector. Namely, a random variable associates a vector $x \in \mathbb{R}^d$ to each element $\omega \in \Omega$. In analogy with this, one can think of a stochastic process as of a generalization of the concept of function. To each sample ω we associate a function from an index set I to $\mathbb{R}^d$. When $i \in I$ represents a time-index, a stochastic process represents a random quantity that evolves in time. The time-index sets that are commonly used in stochastic modeling are the intervals of $\mathbb{R}$ and the subsets of $\mathbb{N}_0$. We talk about *continuous-time* stochastic processes in the former case, and about *discrete-time* stochastic processes in the latter. In this chapter, and in general in these notes, we focus on the study of discrete-time processes. However, every time the extension from discrete to continuous indexes does not introduce relevant difficulties, the concepts will be presented for a general index set I.

2.1 Stochastic processes and filtrations

Hereafter, $(\Omega, \mathcal{F})$ denotes a fixed measurable space.

Definition 2.1.1. Let I be a non-empty set. A *stochastic process indexed by* I, or simply a *stochastic process*, is a family $(X_i)_{i\in I}$ of random variables defined on $(\Omega, \mathcal{F})$. A *discrete-time stochastic process* is a stochastic process indexed by $I \subset \mathbb{N}_0$.

Remark 2.1.2. If I contains N elements (i.e. I is finite), then a family $(X_i)_{i\in I}$ of functions $X_i : \Omega \to \mathbb{R}^d$ is a stochastic process if and only if the function X given by

$$\omega \mapsto \big(X_1(\omega), \cdots, X_N(\omega)\big)$$

is measurable from $\mathcal{F}$ to $\mathbb{R}^{dN}$.

In these notes the focus is on discrete-time stochastic processes. In the applications, some typical choices for I are

$$\mathbb{N}_0, \quad \mathbb{N}, \quad \{0, \cdots, N\}, \quad \{1, \cdots, N\}.$$

For sake of simplicity, we will often consider $I = \mathbb{N}$ or $I = \mathbb{N}_0$ as a canonical choice. This choice can be made without losing generality, upon index shifting and/or completing the process with null random variables.

A discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ can also be understood (perhaps more intuitively) as a random function

$$X : \Omega \to (\mathbb{R}^d)^{\mathbb{N}}, \qquad \omega \mapsto \big(X_n(\omega)\big)_{n\in\mathbb{N}}, \tag{2.1.1}$$

where $(\mathbb{R}^d)^{\mathbb{N}}$ is called the *trajectory space* and is defined as the set of all the functions $x : \mathbb{N} \to \mathbb{R}^d$, also denoted by $x = (x_n)_{n\in\mathbb{N}}$. In terms of functional relation, representing X as a random function with values on $(\mathbb{R}^d)^{\mathbb{N}}$ or as a sequence of random variables with values on $\mathbb{R}^d$ is perfectly equivalent. However, the latter representation entails a measurability condition, namely

$$X_n \in m\mathcal{F}, \qquad n \in \mathbb{N}. \tag{2.1.2}$$

For reasons that will be clarified in the next section, it is useful to provide $(\mathbb{R}^d)^{\mathbb{N}}$ with a suitable σ-algebra that allows to translate (2.1.2) into a measurability condition for X in (2.1.1). This is achieved by considering the σ-algebra on $(\mathbb{R}^d)^{\mathbb{N}}$ generated by all *finite cylinders* of the type

$$C_n(H) := \big\{x \in (\mathbb{R}^d)^{\mathbb{N}} : (x_1, \cdots, x_n) \in H\big\}.$$

with $n \in \mathbb{N}$ and $H \in \mathcal{B}(\mathbb{R}^{n\times d})$. We set

$$\mathcal{B}^{\mathbb{N}} = (\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}} := \sigma(\mathcal{C}), \qquad \mathcal{C} := \big\{C_n(H) : n \in \mathbb{N}, H \in \mathcal{B}(\mathbb{R}^{n\times d})\big\}. \tag{2.1.3}$$

Proposition 2.1.3. *Let X be a function as in* (2.1.1)*. Then $(X_n)_{n\in\mathbb{N}}$ is a discrete-time stochastic process if and only if $X \in m\big(\mathcal{F}, \mathcal{B}^{\mathbb{N}}\big)$, i.e.*

$$\{X \in H\} \in \mathcal{F}, \qquad H \in \mathcal{B}^{\mathbb{N}}.$$

Proof. By the standard measurability criterion, $X \in m\big(\mathcal{F}, \mathcal{B}^{\mathbb{N}}\big)$ if and only if

$$\{X \in C_n(H)\} = \{(X_1, \cdots, X_n) \in H\} \in \mathcal{F}, \qquad H \in \mathcal{B}^{nd},\ n \in \mathbb{N},$$

which means

$$(X_1, \cdots, X_n) \in m\mathcal{F}, \qquad n \in \mathbb{N}.$$

The latter, in turn, is equivalent to

$$X_n \in m\mathcal{F}, \qquad n \in \mathbb{N}.$$

□

Remark 2.1.4. A result analogous to Proposition 2.1.3 holds true for a generic index set I. However, for a non-countable I, the definition of the product σ-algebra $\mathcal{B}^I$ is more involved and its analysis is more technical. We refer the interested reader to [1] and [3] for a detailed account of the general theory of stochastic processes.

Example 2.1.5. Let us consider a sequence of $N \in \mathbb{N}$ coin-tosses. We first construct the measurable space $(\Omega, \mathcal{F})$ that describes this random phenomenon. For any $n = 1, \cdots, N$ set

$$\Omega_n := \{0, 1\}, \qquad \mathcal{F}_n = \mathcal{P}(\Omega_n) = \big\{\emptyset, \Omega_n, \{0\}, \{1\}\big\}.$$

Here "0" and "1" represent *heads* and *tails*, respectively, as possible outcomes of the n-th toss. We now define

$$\Omega := \Omega_1 \times \cdots \times \Omega_N = \{0, 1\}^N, \qquad \mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_N = \mathcal{P}(\Omega).$$

In practice, a generic element $\omega \in \Omega$ is a sequence of N elements representing "heads" or "tails". For any $n = 1, \cdots, N$ we can now set

$$X_n(\omega) = X_n(\omega_1, \cdots, \omega_N) := \begin{cases} 1 & \text{if } \omega_n = 0 \\ -1 & \text{if } \omega_n = 1 \end{cases}, \qquad \omega \in \Omega. \tag{2.1.4}$$

Obviously, $X_n m\mathcal{F}$ for any $n = 1, \cdots, n$, and thus $X = (X_n)_{n=1,\cdots,N}$ is a discrete-time stochastic process. Here, X_n could represent the gain, after the n-th toss, of a gambler that wins 1 Euro if *heads* occurs and loses 1 Euro if *tails* occurs. Consider now

$$Y_n := \sum_{i=1}^{n} X_i, \qquad Z_n = \sum_{i=n}^{N} X_i, \qquad n = 1, \cdots, N. \tag{2.1.5}$$

Once more, it is clear that $Y_n, Z_n \in m\mathcal{F}$, and thus $Y = (Y_n)_{n=1,\cdots,N}$, $Z = (Z_n)_{n=1,\cdots,N}$ are also discrete-time stochastic processes. Here, Y_n and Z_n could represent the cumulated and the future gains, respectively.

Example 2.1.6. Let us modify the previous example by considering a continuous coin, whose outcome can be any value in $[0, 1]$. A natural modification of the space $(\Omega, \mathcal{F})$ is given by

$$\Omega_n := [0, 1], \qquad \mathcal{F}_n = \mathcal{B}([0, 1])$$

and

$$\Omega := \Omega_1 \times \cdots \times \Omega_N = [0, 1]^N, \qquad \mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_N = \mathcal{B}\big([0, 1]^N\big).$$

For any $n = 1, \cdots, N$ we can now set

$$X_n(\omega) = X_n(\omega_1, \cdots, \omega_N) := \begin{cases} 1 & \text{if } \omega_n \in [0, \frac{1}{2}[\\ -1 & \text{if } \omega_n \in [\frac{1}{2}, 1] \end{cases}, \qquad \omega \in \Omega.$$

Clearly $X_n \in m\mathcal{F}_n$ and thus $X_n \in m\mathcal{F}$. Therefore, $X = (X_n)_{n=1,\cdots,N}$ is still a discrete-time stochastic process, and the same holds for Y, Z defined as in (2.1.5).

In the two previous examples, the concept of stochastic process coincides exactly with the one of random variable, the processes X, Y, Z being indeed $\mathbb{R}^N$-valued random vectors. In general, at this stage, the concept of stochastic process seems to extend the one of random variable only in relation to the choice of the co-domain of a function X defined on Ω. We saw indeed that a stochastic process can be viewed as a measurable function from Ω to the trajectory space. However, the difference between random variable and stochastic process should go beyond that, as the latter is supposed to represent a random phenomenon that unfolds in different times.

To understand this, consider for instance Example 2.1.5. The same measurable space $(\Omega, \mathcal{F})$ could represent a game in which N different coins are tossed simultaneously and the quantities X_n are gained all at once when the outcome of the N-tuple ω is revealed. This is clearly different from a game where the coins are tossed sequentially and the uncertainty of each toss is resolved at different times. In other words, the information arrives gradually to the player, who can observe, at each time, only the quantities that depend on the observable events. In the case of Example 2.1.5, one can assume that the player can see the outcome after each toss: this makes the value X_n observable at time n. We could also assume that the knowledge of the previous outcomes is not forgotten: this makes the value Y_n observable at time n too. By contrast, the value Z_n will not be observable at time n, unless the observer is able to look $N - n$ steps into the future to observe the outcome of the whole sequence.

In order to formalize these concepts, we introduce the following definitions.

Definition 2.1.7. Let $I \subset \mathbb{R}$ a non-empty set. A family $(\mathcal{F}_i)_{i\in I}$ of increasing sub-σ-algebras, namely

$$\mathcal{F}_i \subset \mathcal{F}_j \subset \mathcal{F}, \qquad i, j \in I,\ i \leq j,$$

is called a *filtration* (on $\mathcal{F}$).

Definition 2.1.8. A stochastic process $(X_i)_{i\in I}$ is said *adapted to a filtration* $(\mathcal{F}_i)_{i\in I}$, or simply *adapted* when the reference filtration is obvious from the context, if

$$X_i \in m\mathcal{F}_i, \qquad i \in I.$$

A given filtration $(\mathcal{F}_i)_i$ can be interpreted as the information flow available to the observer. In other words, $\mathcal{F}_i$ contains the events that are observable up to time i. The fact that $(\mathcal{F}_i)_i$ is an increasing family means that the knowledge accumulates in time. With this interpretation in mind, a stochastic process is adapted to a filtration if it is observable by observing the events of the filtration.

Example 2.1.9. Consider the game described in Example 2.1.5. Assuming that the player can witness every coin-toss, it is natural to set

$$\mathcal{F}_n := \Big\{ \big\{\omega = (\omega_1, \cdots, \omega_N) \in \Omega : (\omega_1, \cdots, \omega_n) \in A\big\}, A \in \mathcal{P}\big(\Omega_1 \times \cdots \times \Omega_n\big) \Big\}, \qquad n = 1, \cdots, N. \quad (2.1.6)$$

It is a simple exercise to show that the family $(\mathcal{F}_n)_n$ forms a filtration. It is also straightforward to show that

$$\mathcal{F}_n = \Big\{ \big\{(X_1, \cdots, X_n) \in H\big\}, H \in \mathcal{B}(\mathbb{R}^n) \Big\}, \qquad n = 1, \cdots, N.$$

In particular, setting

$$H = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n-1 \text{ times}} \times H_n \qquad \text{with } H_n \in \mathcal{B}(\mathbb{R})$$

yields

$$\{X_n \in H_n\} = \{(X_1, \cdots, X_n) \in H\} \in \mathcal{F}_n,$$

which is $X_n \in m\mathcal{F}_n$, and thus X is adapted to $(\mathcal{F}_n)_n$. Also, since the latter is an increasing family, we have that

$$X_i \in m\mathcal{F}_n, \qquad i = 1, \ldots, n,$$

and thus the process Y is adapted to $(\mathcal{F}_n)_n$ too. By contrast, it is easy to see that the process Z is not adapted to $(\mathcal{F}_n)_n$. For instance, one can consider Z_{N-1} and observe that

$$\{Z_{N-1} = 2\} = \big\{\omega = (\omega_1, \cdots, \omega_N) \in \Omega : \omega_{N-1} = \omega_N = 1\big\} \notin \mathcal{F}_{N-1},$$

which implies $Z_{N-1} \notin m\mathcal{F}_{N-1}$.

In general, different observers have access to different informations related to the same random phenomenon. In other words, the choice fo the filtration is subjective and depends on the observer.

Example 2.1.10. Let us consider a sequence of $N \in \mathbb{N}$ dice-rolles. We first construct the measurable space $(\Omega, \mathcal{F})$ similarly to what we did in Example 2.1.5. First set

$$\Omega_n := \{1, \dots, 6\}, \qquad n = 1, \cdots, N,$$

with each Ω_n containing the possible outcomes of the n-th roll. Then define

$$\Omega := \Omega_1 \times \cdots \times \Omega_N = \{1, \dots, 6\}^N, \qquad \mathcal{F} = \mathcal{P}(\Omega_1) \otimes \cdots \otimes \mathcal{P}(\Omega_N) = \mathcal{P}(\Omega).$$

A generic element $\omega \in \Omega$ is a sequence of N elements, representing the outcomes of the N rolls. For $n = 1, \cdots, N$, let us consider $\mathcal{F}_n$ as defined in (2.1.6). The family $(\mathcal{F}_n)_n$ forms a filtration and contains all the events observable by observing the outcomes of the first n dice-rolls. Consider now a game in which, at each roll, a gambler wins 1 Euro if the outcome is even and loses 1 Euro otherwise. The gain of the n-th roll is then represented by the random variable

$$X_n(\omega) = X_n(\omega_1, \cdots, \omega_N) := \begin{cases} 1 & \text{if } \omega_n \in \{2, 4, 6\} \\ -1 & \text{if } \omega_n \in \{1, 3, 5\} \end{cases}, \qquad \omega \in \Omega.$$

The non-trivial events of the σ-algebra generated by X_n are

$$\big\{\omega \in \Omega : \omega_n \in \{2, 4, 6\}\big\} \qquad \text{and} \qquad \big\{\omega \in \Omega : \omega_n \in \{1, 3, 5\}\big\},$$

which are both contained in $\mathcal{F}_n$. Therefore, $X = (X_n)_{n=1,\cdots,N}$ is a discrete-time stochastic process adapted to $(\mathcal{F}_n)_n$. Assume now that the gambler cannot see the exact outcome of the dice-rolls: the casino only tells the player whether the outcomes are odd or even. Clearly, in this case, the player does not have access to the information contained in $(\mathcal{F}_n)_n$, meaning that she is not able to observe all the events in $\mathcal{F}_n$ at time n[1]. Instead, the filtration she has access to is the one defined by

$$\mathcal{F}_n^X = \sigma\big(\{X_k = j\} : j = -1, 1 \text{ and } k = 1, \cdots, n\big), \qquad n = 1, \cdots, N.$$

The family $(\mathcal{F}_n^X)_n$ is a filtration and is strictly contained in $(\mathcal{F}_n)_n$. For instance, the event

$$A = \big\{\omega \in \Omega : \omega_2 \in \{2, 4, 6\}\big\} = \{X_2 = 1\}$$

belongs to both $\mathcal{F}_2$ and $\mathcal{F}_2^X$, while the event

$$B = \big\{\omega \in \Omega : \omega_2 = 4\big\}$$

is contained in $\mathcal{F}_2$ but not in $\mathcal{F}_2^X$. The filtration $(\mathcal{F}_n^X)_n$ contains all, and only, the events that are observable by observing the stochastic process X.

[1] In this case she will never be able to observe some events in $\mathcal{F}_n$

The previous example introduces the following

Definition 2.1.11. Let $I \subset \mathbb{R}$ a non-empty set and let $X = (X_i)_{i\in I}$ be a stochastic process. The family $(\mathcal{F}_i^X)_{i\in I}$ defined by

$$\mathcal{F}_i^X := \sigma\big(\{X_j \in H\} : H \in \mathcal{B}, j \leq i\big), \qquad n \in I,$$

is called the *natural filtration of* X.

The natural filtration of X is the smallest filtration that makes X adapted. It can be interpreted as the information flow available to an observer that can only observe the stochastic process X.

We conclude this section with a notion of observability that is stronger than adaptedness, namely *predictability*. Intuitively, a discrete-time stochastic process is *predictable* if it can be observed at least one step prior its realization.

Definition 2.1.12. A discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is said *predictable* with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ if

$$X_n \in \mathcal{F}_{n-1}, \qquad n \in \mathbb{N}.$$

As we shall see the class of predictable processes plays an important role in the mathematical theory of stochastic processes (see Theorem **??** below), as well as in many applications.

Example 2.1.13. Let $X = (X_n)_{n\in\mathbb{N}_0}$ be a discrete-time real-valued stochastic process representing the evolution of the market price of a given financial asset. Precisely, considering the time-grid

$$0 = t_0 < t_1 < \cdots < t_n < \cdots,$$

the value X_n represents the observed price at time t_n. Assume now that we want to define an investment strategy in this asset, namely a sequence of stochastic scalar quantities $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ with the following meaning: α_n represents the amount of shares of the asset held in the time-period $[t_{n-1}, t_n]$. Assuming that we can observe at any time t_n the current value of the asset, our reference filtration will be given by $(\mathcal{F}_n^X)_{n\in\mathbb{N}_0}$. Therefore, since the quantity α_n is decided at the beginning of the period $[t_{n-1}, t_n]$, at time t_{n-1}, its value will only depend on the prices observed up to time t_{n-1}, and thus only on the events in $\mathcal{F}_{n-1}^X$. In mathematical terms, this is translated as $\alpha_n \in m\mathcal{F}_{n-1}^X$, i.e. α is predictable with respect to the natural filtration of X. On the other hand, the value of the investment at the end of the period $[t_{n-1}, t_n]$, at time t_n, is given by $V_n = \alpha_n X_n$, which is measurable w.r.t. to $\mathcal{F}_n^X$ but not w.r.t. $\mathcal{F}_{n-1}^X$. Thus $(V_n)_{n\in\mathbb{N}}$ is "only" adapted to $(\mathcal{F}_n^X)_{n\in\mathbb{N}_0}$.

2.2 Law of a stochastic process

The notions introduced in the previous section pertain to the description of the uncertainty, in that they allow to describe the evolution of uncertain quantities dependent on a given phenomenon, in relation to the information flow available to the observer. In order to quantify the uncertainty related to these quantities, we equip the measurable space $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}$. In this way we will be able to assign a probability to some specific sets of trajectories of a stochastic process. In other words, in analogy with the framework of random variables, the probability measure $\mathbb{P}$ induces a *law* (or a *distribution*) for a stochastic process defined on $(\Omega, \mathcal{F})$.

Throughout the rest of this chapter, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space.

Definition 2.2.1. Let $X = (X_n)_{n\in\mathbb{N}}$ be a discrete-time stochastic process with values on $\mathbb{R}^d$. The probability measure μ_X on $(\mathbb{R}^d)^{\mathbb{N}}$ defined by

$$\mu(H) := \mathbb{P}(\{X \in H\}), \qquad H \in \mathcal{B}^{\mathbb{N}} = \big(\mathcal{B}(\mathbb{R}^d)\big)^{\mathbb{N}} \tag{2.2.1}$$

is called the *law (or distribution) of X.*

Remark 2.2.2. Definition 2.2.1 is well posed in light of Proposition 2.1.3, which ensures that the event

$$\{X \in H\} = \big\{\omega \in \Omega : (X_n(\omega))_{n\in\mathbb{N}} \in H\big\} \in \mathcal{B}^{\mathbb{N}},$$

and that μ as defined in (2.2.1) is a probability measure.

The next result provides us with a useful characterization of the law of a discrete-time stochastic process. As it turns out, the latter is univocally identified by the family of all the finite-dimensional marginal distributions.

Proposition 2.2.3. *Let μ and ν be two probability measures on $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$ that coincide on $\mathcal{C}$, i.e.*

$$\mu\big(C_n(H)\big) = \nu\big(C_n(H)\big), \qquad n \in \mathbb{N}, \quad H \in \mathcal{B}(\mathbb{R}^{nd}),$$

then $\mu = \nu$.

In particular, the law of a discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is univocally determined by the finite-dimensional distributions $\mu_{(X_1,\cdots,X_N)}$, $N \in \mathbb{N}$.

Proof. Let us recall that $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$ is the σ-algebra defined generated by $\mathcal{C}$, the latter being the class of finite-dimensional cylinders (see definition (2.1.3)). It is easy to verify that $\mathcal{C}$ is an algebra. In particular, $\mathcal{C}$ is closed with respect to finite intersections and generates $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$. Therefore, by Dynkin's Theorem, $\mu = \nu$.

Furthermore, if X is a stochastic process with law μ, then

$$\mu\big(C_n(H)\big) = \mathbb{P}\big(\{X \in C_n(H)\}\big) = \mathbb{P}\big(\{(X_1,\cdots,X_n) \in H\}\big) = \mu_{(X_1,\cdots,X_N)}(H), \qquad n \in \mathbb{N}, \quad H \in \mathcal{B}(\mathbb{R}^{nd}).$$

□

Remark 2.2.4. If $X = (X_n)_{n\in\mathbb{N}}$ is such that $X_n = 0$ almost surely for any $n > N$ with $N \in \mathbb{N}$, then μ_X is univocally identified by $\mu_{(X_1,\cdots,X_N)}$. In this case, with a slide abuse of notation we write $\mu_X = \mu_{(X_1,\cdots,X_N)}$.

Indeed, assuming that X takes values on $\mathbb{R}^d$, for any $n \in \mathbb{N}$ with $n > N$ and $H \in \mathcal{B}(\mathbb{R}^{nd})$ we have

$$\mu_{(X_1,\cdots,X_n)}(H) = \mathbb{P}\big(\{(X_1,\cdots,X_n) \in H\}\big)$$

($X_{N+1} = \cdots = X_n = 0$ almost surely)

$$\begin{aligned} &= \mathbb{P}\big(\{(X_1,\cdots,X_n) \in H\} \cap \{X_{N+1} = 0\} \cap \cdots \cap \{X_n = 0\}\big) \\ &= \mathbb{P}\big(\{(X_1,\cdots,X_N) \in H_0\}\big) = \mu_{(X_1,\cdots,X_N)}(H_0), \end{aligned}$$

where

$$H_0 := \{x \in \mathbb{R}^{Nd} : (x, \mathbf{0}) \in H\}, \qquad \mathbf{0} = (0,\cdots,0) \in \mathbb{R}^{(n-N)d}.$$

Example 2.2.5. Let $N \in \mathbb{N}$ and $(\Omega, \mathcal{F}, \mathbb{P})$ be given by

$$\Omega := \{0,1\}^N, \qquad \mathcal{F} := \mathcal{P}(\Omega), \qquad \mathbb{P} := \mathrm{Unif}_\Omega.$$

In relation to the sequence of N coin tosses discussed in Example 2.1.5 and Example 2.1.9, this choice of the probability $\mathbb{P}$ reflects the fact that the tosses are independent and that the coin is balanced. Indeed, setting

$$E_n^0 := \{\omega \in \Omega : \omega_n = 0\}, \qquad E_n^1 := \{\omega \in \Omega : \omega_n = 1\}, \qquad n = 1, \cdots, N,$$

we have

$$\mathbb{P}(E_n^0) = \mathbb{P}(E_n^1) = \frac{2^{N-1}}{2^N} = \frac{1}{2}, \qquad n = 1, \cdots, N, \tag{2.2.2}$$

and

$$\mathbb{P}(\{\omega\}) = \mathbb{P}\big(\{(\omega_1, \cdots, \omega_N)\}\big) = \frac{1}{2^N} = \frac{1}{2} \times \cdots \times \frac{1}{2} = \mathbb{P}(E_1^{\omega_1}) \times \cdots \times \mathbb{P}(E_N^{\omega_N}), \qquad \omega \in \Omega. \tag{2.2.3}$$

Recall now the stochastic process $X = (X_n)_{n=1,\cdots,N}$ defined in (2.1.4). By (2.2.2) we obtain

$$\mu_{X_n} = \frac{1}{2}\delta_{\{-1\}} + \frac{1}{2}\delta_{\{1\}}, \qquad n = 1, \cdots, N.$$

Furthermore, (2.2.3) yields

$$\mu_X = \mu_{(X_1,\ldots,X_N)} = \mu_{X_1} \otimes \cdots \otimes \mu_{X_N} = \Big(\frac{1}{2}\delta_{\{-1\}} + \frac{1}{2}\delta_{\{1\}}\Big)^{\otimes N},$$

which is X is a sequence of independent random variables.

Example 2.2.6. Let $X = (X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Then, in light of Proposition 2.2.3 and Remark 1.2.12, the law of X is identified by

$$\mu_{(X_1,\ldots,X_n)} = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}, \qquad n \in \mathbb{N}. \tag{2.2.4}$$

Having introduced the law of a stochastic process allows us to define a notion of weak uniqueness for stochastic processes, which extends the one we have for random variables.

Definition 2.2.7. Two discrete-time stochastic processes $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ are said *equal in law* (or *in distribution*) if $\mu_X = \mu_Y$. In this case we write

$$X \overset{\mathrm{d}}{=} Y \qquad \text{or} \qquad X \sim Y.$$

Remark 2.2.8. In light of Propostion 2.2.3, $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ are equal in law if the share the same finite dimensional distributions, i.e. $\mu_{(X_1,\cdots,X_n)} = \mu_{(Y_1,\cdots,Y_n)}$ for any $n \in \mathbb{N}$.

We note that stating $X \sim Y$ is way stronger than stating the uniqueness of all the single-time marginal distributions, namely $X_n \sim Y_n$ for any n, as it is shown in the following example

Example 2.2.9. Let

$$\Omega = \mathbb{R}^2, \quad \mathcal{F} = \mathcal{B}, \qquad \mathbb{P} = \mathcal{N}_{0,I_2},$$

and set

$$X_1(\omega) = \omega_1, \qquad X_2(\omega) = \omega_2, \qquad \omega \in \Omega.$$

Also set

$$Y_1 = Y_2 \equiv X_1.$$

We have $\mu_{X_1} = \mu_{X_2} = \mu_{Y_1} = \mu_{Y_2}$, but $\mu_{(X_1,X_2)} \neq \mu_{(Y_1,Y_2)}$. Indeed,

$$\mu_{(X_1,X_2)}(A) = 0 \neq 1 = \mu_{(Y_1,Y_2)}(A), \qquad A := \{(x,y) \in \mathbb{R}^2 : x = y\}.$$

Like for random variables, we can introduce a strong notion of uniqueness. We first observe that, if $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ are discrete-time stochastic processes indexed by $\mathbb{N}$, we have

$$\{X = Y\} = \big\{\omega \in \Omega : X_n(\omega) = Y_n(\omega) \text{ for any } n \in \mathbb{N}\big\} = \bigcap_{n\in\mathbb{N}} \underbrace{\{X_n = Y_n\}}_{\in\mathcal{F}} \in \mathcal{F}. \tag{2.2.5}$$

Therefore, it is well-posed the following

Definition 2.2.10. Two discrete-time stochastic processes $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ are said *indistinguishable*, and we also write $X = Y$ $\mathbb{P}$-*almost surely* (or simply *almost surely*, if the choice of the probability measure $\mathbb{P}$ is clear from the context), i.e.

$$\mathbb{P}\big(\{X = Y\}\big) = 1. \tag{2.2.6}$$

In this case we can also say that Y is a *modification* of X, and viceversa.

Remark 2.2.11. By (2.2.5), we have that X and Y are indistinguishable if and only if $X_n = Y_n$ almost surely for any $n \in \mathbb{N}$, namely if

$$\mathbb{P}(\{X_n = Y_n\}) = 1, \qquad n \in \mathbb{N}. \tag{2.2.7}$$

This is true because X and Y are discrete-time stochastic processes. In particular, we note that (2.2.7) implies (2.2.6) because the intersection in (2.2.5) is countable. In the general theory of stochastic processes, the concept of *modifications* is given by (2.2.7) and is weaker than the concept of indistinguishability, the latter given by (2.2.6). When X and Y are indexed by a non-countable set, the equivalence between the two notions is restored in some particular cases, for instance when the trajectories of X and Y are continuous.

We have the following

Proposition 2.2.12. *Let $X = (X_n)_{n\in\mathbb{N}}$ and $Y = (Y_n)_{n\in\mathbb{N}}$ be indistinguishable stochastic processes, then they are equal in law.*

Proof. For any $n \in \mathbb{N}$, we have

$$\mathbb{P}\big(\underbrace{\{(X_1,\dots,X_n) = (Y_1,\dots,Y_n)\}}_{\supset\{X=Y\}}\big) = 1.$$

Thus the two random vectors $(X_1,\dots,X_n)$ and $(Y_1,\dots,Y_n)$ are equal almost surely. Therefore, they are equal in law, i.e. $\mu_{(X_1,\dots,X_n)} = \mu_{(Y_1,\dots,Y_n)}$. Therefore, Proposition 2.2.3 yields $\mu_X = \nu_X$. □

Remark 2.2.13. Clearly, the previous statement cannot be inverted. Also note that two processes can be equal in law even if they are defined on different probability spaces.

After discussing the uniqueness between stochastic processes, we now discuss the problem of the existence of a stochastic process with a pre-assigned distribution. In analogy with random variables, we have the following

Proposition 2.2.14 (Canonical construction)**.** *Let μ be a probability measure on $\mathcal{B}^{\mathbb{N}} = (\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$. There exists a discrete-time stochastic process $X = (X_n)_{\in\mathbb{N}}$ such that $\mu_X = \mu$.*

Proof. Set

$$\Omega := (\mathbb{R}^d)^{\mathbb{N}}, \qquad \mathcal{F} := \mathcal{B}^{\mathbb{N}}, \qquad \mathbb{P} := \mu,$$

and define the canonical stochastic process

$$X_n(\omega) := \omega_n, \qquad \omega \in \Omega, \quad n \in \mathbb{N}.$$

□

Remark 2.2.15. The natural filtration of the process $X = (X_n)_{\in\mathbb{N}}$ constructed via canonical construction, as in the proof of Proposition 2.2.14, is given by

$$\mathcal{F}_n^X := \sigma\big(\{C_n(H) : H \in \mathbb{R}^{nd}\}\big).$$

At first glance it might seem like Proposition 2.2.14 solves the problem of the existence. A deeper look shows instead that this alone is not satisfactory. The problem is that we do not know how to define, constructively, a probability measure μ on $\mathcal{B}^{\mathbb{N}}$ with meaningful statistical properties. In the context of standard random variables, namely in the finite-dimensional case, we know a few constructive ways to define a distribution on $\mathcal{B}(\mathbb{R}^d)$, such as the definition of a density, of a cumulative distribution function or of a characteristic function. In the context of stochastic processes, the trajectory space is infinite-dimensional and we do not have the previous tools at our disposal. However, Proposition 2.2.3 tells us that a probability μ on $\mathcal{B}^{\mathbb{N}}$ is univocally determined by its finite-dimensional marginals μ_n, $n \in \mathbb{N}$, given by

$$\mu_n(H) := \mu\big(C_n(H)\big), \qquad H \in \mathcal{B}^N. \tag{2.2.8}$$

Therefore, it is natural to wonder if it is possible to construct μ starting from a family of pre-assigned finite dimensional distributions μ_n.

For instance, in Example 2.2.6 we considered a sequence $X = (X_n)_{n\in\mathbb{N}}$ of independent random variables, each one with distribution ν_n. Denoting by μ the law of X, the independence property (2.2.4) reads as

$$\mu_n = \nu_1 \otimes \cdots \otimes \nu_n, \qquad n \in \mathbb{N}.$$

However, the existence of a μ with these marginals, and thus the existence of a sequence of independent random variables, so far has only been assumed. As we shall see, this result is true, though its proof is surprisingly non-trivial.

Question: given a family $(\mu_n)_{n\in\mathbb{N}}$ with each μ_n being an arbitrary distribution on $\mathcal{B}(\mathbb{R}^{nd})$, is there a probability μ on $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$ with marginals μ_n?

Answer: if and only if the family $(\mu_n)_{n\in\mathbb{N}}$ is consistent.

We now clarify the consistency requirement. For any $H \in \mathcal{B}(\mathbb{R}^{nd})$ and $n \in \mathbb{N}$, we have

$$C_n(H) = C_{n+1}(H \times \mathbb{R}^d).$$

Therefore, if μ is a probability on $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$ and $(\mu_n)_{n\in\mathbb{N}}$ is given by (2.2.8), the following consistency condition has to be satisfied:

$$\mu_{n+1}(H\times\mathbb{R}^d)=\mu_n(H),\qquad H\in\mathcal{B}(\mathbb{R}^{nd})\qquad n\in\mathbb{N}. \tag{2.2.9}$$

Example 2.2.16. Consider

$$\mu_1=\mathcal{N}_{0,1}\quad\text{(distribution on }\mathbb{R}\text{)},\qquad\qquad \mu_2=\delta_{(0,0)}\quad\text{(distribution on }\mathbb{R}^2\text{)}.$$

Then

$$\mu_1(\{0\})=0\neq 1=\mu_2(\{0\}\times\mathbb{R}).$$

Therefore, μ_1 and μ_2 are not consistent.

Example 2.2.17. Let $(\nu_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $\mathcal{B}(\mathbb{R}^d)$. The family defined by

$$\mu_n=\nu_1\otimes\cdots\otimes\nu_n,\qquad n\in\mathbb{N},$$

satisfies (2.2.9).

The next result shows that (2.2.9) is also a sufficient condition in order for μ_n, $n\in\mathbb{N}$, to be the marginals of a probability μ on $\mathcal{B}^{\mathbb{N}}$.

Theorem 2.2.18 (Kolmogorov Extension Theorem)**.** *Let $(\mu_n)_{n\in\mathbb{N}}$ be a family such that each μ_n is a distribution on $\mathbb{R}^{nd}$, consistent in the sense of* (2.2.9)*. There exists a unique probability measure μ on $(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}}$ such that*

$$\mu\big(C_n(H)\big)=\mu_n(H),\qquad H\in(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}},\ n\in\mathbb{N}.$$

The proof of Kolmogorov Extension Theorem strongly relies on Carathéodory's extension theorem. We provide a sketch of the proof, avoiding the technical details. For a complete proof we refer the reader to [1] or [3]. Critically, K.E.T. remains true in a more general context with the proof being substantially unchanged. Namely, one can replace the index set $\mathbb{N}$ with a generic set I (possibly uncountable), upon suitably modifying the set $\mathcal{C}$ of finite cylinders and adapting the consistency condition (2.2.9), and also replace the space $\mathbb{R}^d$ with a generic separable and complete metric space. For this reason, K.E.T. is a cornerstone of the general theory of stochastic processes, beyond the discrete-time setting.

Proof of Theorem 2.2.18 (sketch). Consider the function $\bar{\mu}:\mathcal{C}\to[0,1]$ defined by

$$\bar{\mu}\big(C_n(H)\big):=\mu_n(H),\qquad H\in(\mathcal{B}(\mathbb{R}^d))^{\mathbb{N}},\ n\in\mathbb{N}.$$

Critically, $\bar{\mu}$ is well defined in light of the consistency property (2.2.9). Indeed, for any $n\in\mathbb{N}$, iterating (2.2.9) one has

$$\mu\big(C_n(H)\big)=\mu\big(C_{n+m}(H\times\mathbb{R}^{md})\big),\qquad m\in\mathbb{N},\ H\in\mathcal{B}(\mathbb{R}^{nd}).$$

Since $\mathcal{C}$ is an algebra, if we can show that $\bar{\mu}$ has the properties of a probability measure, then the statement stems from Carathéodory's extension theorem.

Clearly, we have

$$\bar{\mu}\big((\mathbb{R}^d)^{\mathbb{N}}\big)=\bar{\mu}\big(C_1(\mathbb{R}^d)\big)=\mu_1(\mathbb{R}^d)=1.$$

Also, for any $n \in \mathbb{N}$, $m \in \mathbb{N}_0$ and $H_1 \in \mathcal{B}^{nd}$, $H_2 \in \mathcal{B}^{(n+m)d}$ we have

$$C_n(H_1) \cap C_{n+m}(H_2) = \emptyset \iff (H_1 \times \mathbb{R}^{md}) \cap H_2) = \emptyset.$$

Thus we have

$$\bar{\mu}\big(C_n(H_1) \uplus C_{n+m}(H_2)\big) = \bar{\mu}\big(\underbrace{C_{n+m}(H_1 \times \mathbb{R}^{md}) \uplus C_{n+m}(H_2)}_{=C_{n+m}((H_1\times\mathbb{R}^{md})\uplus H_2)}\big) = \mu_{n+m}\big((H_1 \times \mathbb{R}^{md}) \uplus H_2\big)$$

(since μ_{n+m} is a measure)

$$= \mu_{n+m}\big(H_1 \times \mathbb{R}^{md}\big) + \mu_{n+m}(H_2) = \bar{\mu}\big(\underbrace{C_{n+m}(H_1 \times \mathbb{R}^{md})}_{=C_n(H_1)}\big) + \bar{\mu}\big(C_{n+m}(H_2)\big).$$

Thus $\bar{\mu}$ is also finitely additive. To conclude, we need to prove σ-additivity. Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of pair-wise disjoint elements in $\mathcal{C}$ such that

$$F := \bigcup_{n\in\mathbb{N}} F_n \in \mathcal{C}.$$

We have

$$\bar{\mu}(F) = \bar{\mu}\Big(\biguplus_{n=1}^{N} F_n \uplus D_N\Big), \qquad D_N = F \setminus \biguplus_{n=1}^{N} F_n.$$

As $\mathcal{C}$ is an algebra, $D_N \in \mathcal{C}$ and thus the finite additivity of $\bar{\mu}$ yields

$$\bar{\mu}(F) = \bar{\mu}\Big(\biguplus_{n=1}^{N} F_n\Big) + \bar{\mu}(D_N) = \sum_{n=1}^{N} \bar{\mu}\big(F_n\big) + \bar{\mu}(D_N).$$

Therefore, in order to prove that $\bar{\mu}$ is σ-additive, and thus conclude the proof, we need to show that

$$\bar{\mu}(D_N) \longrightarrow 0 \qquad \text{as } N \to \infty.$$

This is achieved by noting that $D_N \searrow \emptyset$ and by employing a compactness argument. We omit the details for brevity. □

Corollary 2.2.19. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a family such that each μ_n is a distribution on $\mathbb{R}^{nd}$, consistent in the sense of* (2.2.9)*. There exists a stochastic process $X = (X_n)_{\mathbb{N}}$ such that*

$$\mu_{(X_1,\cdots,X_n)} = \mu_n, \qquad n \in \mathbb{N}.$$

Proof. It is a direct consequence of Theorem 2.2.18 and Proposition 2.2.14. □

As an application of this result, we are finally able to prove the existence of a sequence of independent random variables with pre-assigned distributions.

Proposition 2.2.20. *For any sequence $(\nu_n)_{n\in\mathbb{N}}$ of distributions on $\mathcal{B}(\mathbb{R}^d)$, there exists a sequence $(X_n)_{n\in\mathbb{N}}$ of independent random variables such that*

$$\mu_{X_n} = \nu_n, \qquad n \in \mathbb{N}. \tag{2.2.10}$$

Proof. By Corollary 2.2.19, there exists a stochastic process $X = (X_n)_{\mathbb{N}}$ such that

$$\mu_{(X_1,\cdots,X_n)} = \nu_1 \otimes \cdots \otimes \nu_n, \qquad n \in \mathbb{N}.$$

Thus (2.2.10) is satisfied and so

$$\mu_{(X_1,\cdots,X_n)} = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}, \qquad n \in \mathbb{N}.$$

The latter means (see Remark 1.2.12) that X is a sequence of independent random variables. □

2.3 Stopping times

Assume that we are interested in a certain occurrence[2] that takes place in a dynamic random phenomenon. This could be, for instance: a gambler going bankrupt while playing the roulette at Las Vegas, according to a certain strategy; or, the occurrence of a tail in a sequence of coin tosses; or, the number of people infected by a virus reaching a certain threshold, or, the price of a financial asset dropping below a certain value. We might be also interested in the time (or stage, in the discrete setting) τ at which these occurrences take place. Clearly, τ would be a random time, as its value depends on the outcome of the random phenomenon. Additionally, we might want to request that an observer is able to tell whether the occurrence took place at time t based on the information available to her up to time t, or, more succinctly, that the occurrence of interest can be observed as it happens. As the events known by the observer at time t are those contained in $\mathcal{F}_t$, the latter request mathematically translates into $\{\tau \leq t\} \in \mathcal{F}_t$ for any time t. When a random times enjoys this property, it will be called a stopping time.

Definition 2.3.1. Let $I \subset \mathbb{R}$ be a non-empty set. A function $\tau : \Omega \to I \cup \{+\infty\}$ is called a *stopping time with respect to a filtration* $(\mathcal{F}_i)_{i\in I}$ if

$$\{\tau \leq i\} \in \mathcal{F}_i, \qquad i \in I.$$

Remark 2.3.2. A stopping time is a random variable. Indeed, by definition of filtration, $\mathcal{F}_i \subset \mathcal{F}$ for any $i \in I$, and thus

$$\{\tau \leq x\} = \bigcup_{i\in I\cap\mathbb{Q}, i\leq x} \{\tau \leq i\} \in \mathcal{F}, \qquad x \in \mathbb{R},$$

and

$$\{\tau < +\infty\} = \bigcup_{i\in I\cap\mathbb{Q}} \{\tau \leq i\} \in \mathcal{F}.$$

Example 2.3.3. Deterministic times, i.e. $\tau \equiv i$ with $i \in I$, are stopping times with respect to any filtration.

Example 2.3.4. Consider the sequence of coin tosses in Example 2.1.5. Recalling that

$$X_n(\omega) = X_n(\omega_1, \cdots, \omega_N) := \begin{cases} 1 & \text{if } \omega_n = 0 \\ -1 & \text{if } \omega_n = 1 \end{cases}, \qquad \omega \in \{0,1\}^N,$$

set

$$\tau(\omega) := \min\{n = 1, \cdots, N : X_n(\omega) = 1\}, \qquad \omega \in \{0,1\}^N,\ \omega \neq (1, \cdots, 1),$$

together with $\tau(1, \cdots, 1) := +\infty$. The value of τ represents the "first time" (1st toss, 2nd toss, etc.) a *head* occurs in the sequence. Consider now the natural filtration $(\mathcal{F}_n^X)_{n\in\mathbb{N}}$ of X, which contains the events available to an observer that has access to the outcome of the nth toss in real time (at time n). Such an observer must be able to tell, at each time n, if a *head* has occurred in the first nth tosses. Indeed, one can write

$$\{\tau \leq n\} = \{X_1 = 1\} \cup \cdots \cup \{X_n = 1\} \in \mathcal{F}_n^X,$$

which confirms that τ is a stopping time with respect to $(\mathcal{F}_n^X)_{n\in\mathbb{N}}$. Let us now modify τ as

$$\nu(\omega) := \min\{n = 1, \cdots, N : (X_n, \cdots, X_N)(\omega) = (1, \cdots, 1)\}, \qquad \omega \in \Omega,\ \omega \neq (1, \cdots, 1),$$

[2]The use of the word *event* instead of *occurence* would be missleading here, as the former is a technical term to indicate an element of $\mathcal{F}$.

This would be the first occurrence of *head* in a subsequence that contains only heads all the way until the end. An observer who only knows, at time $n < N$, the outcomes of the first n tosses, is unable to tell whether such a subsequence has already started or not. Indeed, we have

$$\{\nu \leq n\} = \{X_n = 1\} \cap \cdots \cap \{X_N = 1\} \notin \mathcal{F}_n^X,$$

and thus ν is not a stopping time with respect to $(\mathcal{F}_n^X)_{n\in\mathbb{N}}$.

Proposition 2.3.5. *A random variable $\tau : \Omega \to \mathbb{N} \cup \{+\infty\}$ is a stopping time with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if and only if*

$$\{\tau = n\}, \{\tau < n\}, \{\tau \geq n\}, \{\tau > n\} \in \mathcal{F}_n, \qquad n \in \mathbb{N}. \tag{2.3.1}$$

Proof. It is enough to prove that τ is a stopping time if and only if $\{\tau = n\} \in \mathcal{F}_n$ for any $n \in \mathbb{N}$, as the other events in (2.3.1) can be obtained by the latter and $\{\tau \leq n\}$ through finite unions, intersections and complements, and since $\mathcal{F}_n$ is, by definition, a σ-algebra.

If $\{\tau = n\} \in \mathcal{F}_n$ for any $n \in \mathbb{N}$, then for any $m \in \mathbb{N}$ we have

$$\{\tau \leq m\} = \bigcup_{n=1,\cdots,m} \underbrace{\{\tau = n\}}_{\in\mathcal{F}_n \subset \mathcal{F}_m} \in \mathcal{F}_m.$$

Let now τ be a stopping time. Then, for any $n \in \mathbb{N}$ we have

$$\{\tau = n\} = \underbrace{\{\tau \leq n\}}_{\in\mathcal{F}_n} \setminus \underbrace{\{\tau \leq n-1\}}_{\in\mathcal{F}_{n-1}\subset\mathcal{F}_n} \in \mathcal{F}_n.$$

□

Remark 2.3.6. It is interest to observe that the above characterization of stopping times does not hold in the continuous-time case. To fix the ideas, let $I = [0, 1]$ and τ be a stopping time with respect to a filtration $(\mathcal{F}_t)_{t\in[0,1]}$. Then, for any $t \in]0, 1]$ we have

$$\{\tau < t\} = \bigcup_{n\in\mathbb{N}} \underbrace{\{\tau \leq t - 1/n\}}_{\in\mathcal{F}_{t-\frac{1}{n}}\subset\mathcal{F}_t} \in \mathcal{F}_t.$$

However, if $\{\tau < s\} \in \mathcal{F}_s$ for any $s \in]0, 1]$, then we can write

$$\{\tau \leq t\} = \bigcap_{n\in\mathbb{N}} \underbrace{\{\tau < t + 1/n\}}_{\in\mathcal{F}_{t+\frac{1}{n}}}, \qquad t \in]0, 1].$$

As, in general, $\mathcal{F}_{t+\frac{1}{n}} \not\subset \mathcal{F}_t$ for any $n \in \mathbb{N}$, one cannot conclude that τ is a stopping time, unless the filtration is assumed to be *right-continuous*, which is

$$\bigcap_{\varepsilon>0} \mathcal{F}_{t+\varepsilon} \in \mathcal{F}_t.$$

When we have an adapted stochastic process X, stopping times can be defined to keep track of certain occurrences in relation to the trajectories of X. For instance, one can define τ as the time at which X enters a certain Borel set of $\mathbb{R}^d$.

Definition 2.3.7. Let $X = (X_n)_{n\in\mathbb{N}}$ be a discrete-time stochastic process and $H \in \mathcal{B}$. The random time τ_H defined as

$$\tau_H := \begin{cases} \min I_H, & \text{if } H \neq \emptyset \\ +\infty, & \text{otherwise} \end{cases}, \qquad I_H := \{n \in \mathbb{N} : X_n \in H\}, \tag{2.3.2}$$

is called the *H-hitting time of X*.

Proposition 2.3.8. *Let $X = (X_n)_{n\in\mathbb{N}}$ be a discrete-time stochastic process, adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, and let $H \in \mathcal{B}$. The H-hitting time of X is a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

Proof. For any $n \in \mathbb{N}$ we have

$$\{\tau_H \leq n\} = \bigcup_{m=1}^{n} \underbrace{\{X_m \in H\}}_{\substack{\in \mathcal{F}_m \subset \mathcal{F}_n \\ (X \text{ is adapted})}} \in \mathcal{F}_n.$$

□

Corollary 2.3.9. *Let $H \in \mathcal{B}$. The H-hitting time of a discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is a stopping time with respect to the natural filtration $(\mathcal{F}^X_n)_{n\in\mathbb{N}}$.*

Remark 2.3.10. Hitting times of Borel sets can also be defined for a continuous-time stochastic process, pretty much in the same way as in (2.3.2), up to replacing min with inf. However, when it comes to proving that τ_H is a stopping time, the continuous-time case presents additional challenges. First of all, one has to limit the choice of Borel sets to sets that have certain topological properties, i.e. open and closed sets. Also, one must assume the trajectories of X to posses some minimal regularity, i.e. continuous or right(left)-continuous. These type of limitations are need in order to reduce uncountable unions

$$\{\tau_H \leq t\} = \bigcup_{s\leq t} \{X_s \in H\}$$

into countable ones. Furthermore, in the case of hitting times of open sets, one has to resort to the right-continuity assumption on the underlying filtration in order to assert that τ_H is a stopping time.

Stopping times are stable with respect to the basic algebraic transformations.

Proposition 2.3.11. *Let τ, ν be stopping times with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$. Then, the random variables*

$$\tau + \nu, \quad \tau\nu, \quad \tau \wedge \nu, \quad \tau \vee \nu \tag{2.3.3}$$

are stopping times with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$. In particular, $\tau + n$, $n\tau$, $\tau \wedge n$ and $\tau \vee n$ are stopping times for any $n \in \mathbb{N}_0$.

Proof. First note that all the random variables in (2.3.3) take values on $\mathbb{N}_0 \cup \{+\infty\}$. We have

$$\{\tau + \nu = n\} = \bigcup_{i=0}^{n} \{\tau = i\} \cap \{\nu = n - i\} \in \mathcal{F}_n, \qquad n \in \mathbb{N}_0,$$

and thus, by Proposition 2.3.5, $\tau + \nu$ is a stopping time. An analogous argument can be applied to show that $\tau\nu$ is a stopping time. Also, for any $n \in \mathbb{N}$, we have

$$\{\tau \wedge \nu \leq n\} = \{\tau \leq n\} \cup \{\nu \leq n\} \in \mathcal{F}_n,$$

$$\{\tau \vee \nu \leq n\} = \{\tau \leq n\} \cap \{\nu \leq n\} \in \mathcal{F}_n,$$

and thus $\tau \wedge \nu$ and $\tau \vee \nu$ are stopping times.

□

Remark 2.3.12. The conclusion of Proposition 2.3.11 remains true, with identical proof, if the index-set is of the form $I = \mathbb{N}_0 \cap [n, +\infty[$ with $n \in \mathbb{N}$. On the other hand, in the continuous-time case the situation is more complex. For instance, if I is an interval of $\mathbb{R}$: $\tau \wedge \nu$ and $\tau \vee \nu$ are stopping times, and the proof is the same as in discrete-time; $\tau + \nu$ is a stopping time under the additional assumption of right-continuity on the filtration; $\tau\nu$ is not necessarily a stopping time.

Remark 2.3.13. Note that, if τ, ν are stopping times, then $\tau - \nu$ is not, in general, a stopping time.

Given an adapted stochastic process X and a finite stopping time τ, it makes sense to compute X_t at $t = \tau$. Precisely, we can set

$$X_\tau(\omega) := X_{\tau(\omega)}(\omega), \qquad \omega \in \Omega. \tag{2.3.4}$$

Proposition-Definition 2.3.14. *Let $X = (X_n)_{n\in\mathbb{N}}$ be a discrete-time stochastic process, adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, and τ be a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Then:*

(i) if τ is finite, the function X_τ defined by (2.3.4) *is a random variable.*

(ii) $X_{\tau\wedge\cdot} := (X_{\tau\wedge n})_{n\in\mathbb{N}}$ is a stochastic process adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, which is called process stopped at τ.

Proof. Let $H \in \mathcal{B}$ be arbitrarily chosen. If τ is finite, we have

$$\{X_\tau \in H\} = \bigcup_{m\in\mathbb{N}} \Big(\underbrace{\{\tau = m\} \cap \{X_m \in H\}}_{\in \mathcal{F}_m \subset \mathcal{F}} \Big) \in \mathcal{F},$$

where we used that τ is a stopping time and X is adapted. In general, for any $n \in \mathbb{N}$, we have

$$\{X_{\tau\wedge n} \in H\} = \bigcup_{m\in\mathbb{N}} \Big(\{\tau \wedge n = m\} \cap \{X_m \in H\} \Big) = \bigcup_{m=1}^{n} \Big(\underbrace{\{\tau \wedge n = m\} \cap \{X_m \in H\}}_{\in \mathcal{F}_m \subset \mathcal{F}_n} \Big) \in \mathcal{F}_n,$$

where we used that, by Propostion 2.3.11, $\tau \wedge n$ is a stopping time. □

The stopped process $X_{\tau\wedge\cdot}$ can by intuitively described as the process whose trajectories coincide with those of X up to time τ, and remain constantly equal to X_τ at any successive time.

Example 2.3.15. Consider the hitting time $\tau = \tau_{\{x\}}$, with $x \in \mathbb{R}^d$, of a discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$. We have

$$X_{\tau\wedge n}(\omega) = \begin{cases} X_n, & \text{if } n \leq \tau(\omega) \\ x, & \text{if } n > \tau(\omega) \end{cases}, \qquad \omega \in \Omega.$$

2.4 Exercises

Exercise 2.4.1 (Successive return times)**.** Consider a discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, and let $x \in \mathbb{R}^d$. Set $\tau_0 := \tau_{\{x\}}$, and

$$\tau_n := \begin{cases} \min I_{x,n}, & \text{if } I_{x,n} \neq \emptyset \\ +\infty, & \text{otherwise} \end{cases}, \qquad I_{x,n} := \{m > \tau_{n-1} : X_m = x\}, \qquad n \in \mathbb{N}.$$

The times τ_n, $n \in \mathbb{N}$, are called *return times of* X *at* x. Show that they are stopping times with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

Chapter 3

Martingales

Throughout this chapter, in all definitions, propositions and examples, we assume the random variables and filtrations to be defined on a reference probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In particular, we will adopt the shortened notation L^p in place of $L^p(\Omega, \mathcal{F}, \mathbb{P})$, unless a different probability space is considered.

3.1 Definitions and basic properties

Definition 3.1.1. Let $I \subset \mathbb{R}$ be a non-empty set and $X = (X_i)_{i \in I}$ be a stochastic process with values on $\mathbb{R}^d$. We say that X is a *martingale with respect to a filtration* $(\mathcal{F}_i)_{i \in I}$ if the following conditions hold:

(i) [*integrability*] $X_i \in L^1$ for any $i \in I$;

(ii) [*martingale property*] We have

$$\mathbb{E}[X_j | \mathcal{F}_i] = X_i, \qquad i, j \in I,\ j > i. \tag{3.1.1}$$

Condition (i) above is a technical assumption, which is needed in order for the conditional expectations in (3.1.10) to make sense. Condition (ii) is instead the "core" assumption, which can be broadly stated as follows: *the conditional expectation of the future values, given all the information available at the present, coincides with the present value.*

Remark 3.1.2. If a stochastic process X is a martingale with respect to a filtration $(\mathcal{F}_i)_{i \in I}$, then it is adapted to this filtration. This is a straightforward consequence of the condition (i) in Definition 1.4.2. Also, any linear combination of two martingales is again a martingale. This stems from the fact that L^1 is a vector space, and form the linearity of the conditional expected value (Proposition 1.4.13-1.).

Lemma 3.1.3. *Let X be a martingale with respect to a filtration $(\mathcal{F}_i)_{i \in I}$ and assume $X_j \in L^p$ for some $j \in I$ and $p \geq 1$. Then, $X_i \in L^p$ for any $i \leq j$. Furthermore, if $p \geq 2$, then*

$$\mathbb{E}\big[\langle X_i, X_j \rangle\big] = \mathbb{E}[|X_i|^2], \qquad i \leq j. \tag{3.1.2}$$

Proof. Let $i \leq j$ be fixed. By the martingale property and conditional Jensen's inequality (Proposition 1.4.17), we obtain

$$\mathbb{E}[|X_i|^p] = \mathbb{E}\big[|\mathbb{E}[X_j|\mathcal{F}_i]|^p\big] \leq \mathbb{E}\big[\mathbb{E}[|X_j|^p|\mathcal{F}_i]\big] = \mathbb{E}[|X_j|^p] < +\infty.$$

Assume now $p \geq 2$. To ease notation, let us assume that X takes values on $\mathbb{R}$. We have

$$\mathbb{E}[X_i X_j] = \mathbb{E}\big[\mathbb{E}[X_i X_j|\mathcal{F}_i]\big] = \mathbb{E}\big[X_i \underbrace{\mathbb{E}[X_j|\mathcal{F}_i]}_{=X_i}\big] = \mathbb{E}[X_i^2],$$

where we employed the total probability formula in the first equality, and Proposition 1.4.13-6. in the second equality, owing to the fact that $X_j, X_i \in L^p \subset L^2$. □

In the discrete-time case, we can give the following useful characterization of martingales.

Proposition 3.1.4. *A discrete-time stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is a martingale with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if and only if*

(i) $X_n \in L^1$ for any $n \in \mathbb{N}$;

(ii) we have

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n, \qquad n \in \mathbb{N}. \tag{3.1.3}$$

Proof. Obviously, if X is a martingale, then $X_n \in L^1$ holds by definition and (3.1.3) stems from the martingale property with $i = n$ and $j = n + 1$.

To prove the converse implication, we proceed by induction. Set $n, m \in \mathbb{N}$ with $m > 0$, and assume that

$$\mathbb{E}[X_m|\mathcal{F}_n] = X_n. \tag{3.1.4}$$

Applying the tower property of the conditional expectation (Proposition 1.4.13-5.), the assumption (3.1.3) and the inductive hypothesis (3.1.4), in this order, yields

$$\mathbb{E}[X_{m+1}|\mathcal{F}_n] = \mathbb{E}\big[\mathbb{E}[X_{m+1}|\mathcal{F}_m]\big|\mathcal{F}_n\big] = \mathbb{E}[X_m|\mathcal{F}_n] = X_n.$$

This, together with the basic step (3.1.3), proves that (3.1.4) holds true for any $n, m \in \mathbb{N}$ with $m > n$, which is the martingale property. □

Remark 3.1.5. If a stochastic process X is a martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$, then it has constant expectation, namely

$$\mathbb{E}[X_i] = \mathbb{E}[X_j], \qquad i, j \in I.$$

This trivially follows from taking the expectation in both sides of (3.1.10) and by applying the total probability formula (1.4.7). In particular, if $X = (X_n)_{n\in\mathbb{N}}$ is a discrete-time martingale, then

$$\mathbb{E}[X_n] = \mathbb{E}[X_1], \qquad n \in \mathbb{N}.$$

Though this is certainly an important property of martingales, the reader shall not give in to the temptation of thinking of martingales as stochastic processes with constant expectation. As it is shown in the next example, the latter is only a necessary condition for martingality, but not a sufficient one.

Example 3.1.6 (Independent sequence)**.** Let $Y = (Y_n)_{n\in\mathbb{N}}$ be and independent sequence of independent summable random variables. Then,

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n^Y] = \mathbb{E}[Y_{n+1}].$$

However, the RHS above differs from Y_n unless the latter is almost-surely constant. In particular, one could further assume, together with the independence, that $\mathbb{E}[Y_n] = \mu$ for any $n \in \mathbb{N}$. In this case, one would obtain a stochastic process with constant expectation that is not a martingale.

The previous example shows that the martingale property is incompatible with independence, except for the trivial case of a constant stochastic process. The next example, however, shows how we can construct a martingale starting from a sequence of independent random variables with null expectation.

Example 3.1.7. [Fair game] Let $Y = (Y_n)_{n\in\mathbb{N}}$ be and independent sequence of independent summable random variables such that $\mathbb{E}[Y_n] = 0$ for any $n \in \mathbb{N}$. Let $x \in \mathbb{R}$ and set

$$X_n := x + \sum_{i=1}^{n} Y_i, \qquad n \in \mathbb{N}_0.$$

The stochastic process $X = (X_n)_{n\in\mathbb{N}_0}$ is a martingale, with expectation y, with respect to $(\mathcal{F}_n^Y)_{n\in\mathbb{N}}$, where we set $\mathcal{F}_0^Y := \{\emptyset, \Omega\}$. Indeed, for any $n \in \mathbb{N}_0$ we have $X_{n+1} = X_n + Y_{n+1}$, and thus, by the linearity of the conditional expected value,

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n^Y] = \mathbb{E}[X_n|\mathcal{F}_n^Y] + \mathbb{E}[Y_{n+1}|\mathcal{F}_n^Y] = X_n + \underbrace{\mathbb{E}[Y_{n+1}]}_{=0}. \tag{3.1.5}$$

The second equality above stems from $X_n \in m\mathcal{F}_n^Y$ and from the fact that Y_{n+1} is independent of $\mathcal{F}_n^Y$. Intuitively, we can think of Y as the *gains* associated to a certain sequence of independent bets. The condition $\mathbb{E}[Y_n] = 0$ reflects the fact that the single bets are fair games. In this view, X can be regarded as the *cumulated gain* of a dynamic fair game.

Remark 3.1.8. A closer look at the first equality in (3.1.5) reveals that a discrete-time process $X = (X_n)_{n\in\mathbb{N}_0}$ has the martingale property with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ if and only if it is adapted and

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = 0, \qquad n \in \mathbb{N}, \tag{3.1.6}$$

where

$$Y_{n+1} := X_{n+1} - X_n \tag{3.1.7}$$

Clearly, the process X in Example 3.1.7, where the increments Y_n are independent random variables with null expectation and $\mathcal{F}_n = \mathcal{F}_n^Y$, is sufficient for (3.1.6). In the next proposition we show that, for square-integrable processes, a necessary condition for the martingale property to hold (w.r.t. any filtration) is that the increments are pairwise uncorrelated. As a remarkable consequence, the limit theorems for sums of uncorrelated random variables can be applied to martingales.

Proposition 3.1.9. *Let X be a discrete-time martingale with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$. If $X_m \in L^2$ for some $m \in \mathbb{N}$, then the increments $(Y_n)_{n=1,\cdots,m}$ defined by* (3.1.7) *are pairwise uncorrelated and*

$$\mathbb{E}[X_m^2] = \mathbb{E}[X_0^2] + \sum_{n=1}^{m} \mathbb{E}[Y_n^2]. \tag{3.1.8}$$

Proof. By Lemma 3.1.3, we have $Y_n \in L^2$ for any $n = 1, \cdots, m$. For any $1 \le n' \le n < m$, we have

$$\begin{aligned}
\mathbb{E}[(Y_{n+1} - \mathbb{E}[Y_{n+1}])(Y_{n'} - \mathbb{E}[Y_{n'}])] &= \mathbb{E}[Y_{n+1}Y_{n'}] && \text{(the increments have null expectation)} \\
&= \mathbb{E}\big[\mathbb{E}[Y_{n+1}Y_{n'}|\mathcal{F}_n]\big] && \text{(by (1.4.8), as } Y_{n'} \in m\mathcal{F}_n,\ Y_{n+1}Y_{n'} \in L^2\ ) \\
&= \mathbb{E}\big[Y_{n'}\mathbb{E}[Y_{n+1}|\mathcal{F}_n]\big] = 0. && \text{(by (3.1.6)).}
\end{aligned}$$

Owing once more to the fact that martingales have constant expectations, the formula for the variance of a sum of uncorrelated random variables yields

$$\mathbb{E}[(X_m - X_0)^2] = \sum_{n=1}^{m} \mathbb{E}[Y_n^2].$$

Finally, (3.1.2) yields

$$\mathbb{E}[(X_m - X_0)^2] = \mathbb{E}[X_m^2] - \mathbb{E}[X_0^2],$$

which proves (3.1.8) and completes the proof. □

The martingale property is clearly dependent on the choice of the filtration. It is then natural to question the stability of this property with respect to a change of filtration. In particular, the following questions might arise:

Question: if X is a martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$, does it remain a martingale with respect to a smaller filtration?

Answer: yes, so long as it remains adapted.

Question: does it remain a martingale with respect to a larger filtration?

Answer: not in general.

Proposition 3.1.10. *Let X be a martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$. Let $(\mathcal{G}_i)_{i\in I}$ be another filtration such that:*

(i) X is adapted to $(\mathcal{G}_i)_{i\in I}$;

(ii) $\mathcal{G}_i \subset \mathcal{F}_i$ for any $i \in I$.

Then X is a martingale with respect to $(\mathcal{G}_i)_{i\in I}$. In particular, X is a martingale with respect to its natural filtration $(\mathcal{F}_i^X)_{i\in I}$.

Proof. For any $i, j \in I$ with $i < j$ we have,

$$\mathbb{E}[X_j|\mathcal{G}_i] = \mathbb{E}\big[\underbrace{\mathbb{E}[X_j|\mathcal{F}_i]}_{=X_i}|\mathcal{G}_i\big] = X_i,$$

where the first and the last inequalities stem from the tower property (1.4.6) and from (1.4.5), respectively. □

Example 3.1.11. Proposition 3.1.10, together with Example 3.1.7, shows that a sequence $Y = (Y_n)_{n\in\mathbb{N}}$ of non-constant independent random variables cannot be a martingale with respect to any filtration.

Example 3.1.12. Let X be the martingale constructed in Example 3.1.7. Proposition 3.1.10 tells us that X remains a martingale with respect to $(\mathcal{F}_n^X)_{n\in\mathbb{N}_0}$. We now show how X can lose the martingale property by enlarging the filtration. For any $n \in \mathbb{N}_0$ set $\mathcal{F}_n := \mathcal{F}_{n+1}^Y$: clearly, $\mathcal{F}_n \supset \mathcal{F}_n^Y$. We have

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_{n+1},$$

which differs from X_n unless Y_{n+1} is almost surely equal to 0.

Remark 3.1.13. By utilizing the tower property as in the proof of Proposition 3.1.10 one can show that, if X is a martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$, then

$$\mathbb{E}[X_j|X_i] = X_i, \qquad i,j \in I,\ i<j. \tag{3.1.9}$$

Not that (3.1.9) alone does not imply the martingale property. In order to obtain that, one also needs $\mathbb{E}[X_j|\mathcal{F}_i] = \mathbb{E}[X_j|X_i]$. The latter, when $\mathcal{F}_i = \mathcal{F}_i^X$, intuitively states the independence of the future conditional expectations from the past.

Martingales are stochastic processes whose value at a given time coincides with the conditional expectation of future values given the current information. We now introduce a class of stochastic processes whose values under(over)estimate the future conditional expectations.

Definition 3.1.14. Let $I \subset \mathbb{R}$ be a non-empty set and $X = (X_i)_{i\in I}$ a stochastic process. We say that X is a *sub(super)-martingale with respect to a filtration* $(\mathcal{F}_i)_{i\in I}$ if

(i) $X_i \in L^1(\Omega, \mathcal{F}_i, \mathbb{P})$ for any $i \in I$; in particular, X is adapted.

(ii) We have

$$\mathbb{E}[X_j|\mathcal{F}_i] \underset{(\leq)}{\geq} X_i, \qquad i,j \in I,\ j>i. \tag{3.1.10}$$

Remark 3.1.15. While adaptedness is included in the martingale property, the sub and super-martingale property (3.1.10) does not imply adaptedness. For this reason, the latter is explicitly requested in condition (i).

Remark 3.1.16. A stochastic process X is a martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$ if and only if it is both a sub and a super-martingale. Also, if X is a martingale, by the conditional Jensen's inequality (Proposition 1.4.17-4.), we have

$$|X_i| = |\mathbb{E}[X_j|\mathcal{F}_i]| \leq \mathbb{E}\big[|X_j|\big|\mathcal{F}_i\big], \qquad i<j,$$

and thus $|X|$ is a sub-martingale.

Remark 3.1.17. If a stochastic process X is a sub(super)-martingale with respect to a filtration $(\mathcal{F}_i)_{i\in I}$, then it has non-decreasing (non-increasing) expectation, namely

$$\mathbb{E}[X_i] \underset{(\geq)}{\leq} \mathbb{E}[X_j], \qquad i,j \in I,\ i<j.$$

Example 3.1.18. Let Y, X be as in Example 3.1.7, except this time we assume $\mathbb{E}[Y_n] = \mu \in \mathbb{R}$ for any $n \in \mathbb{N}$. Then, (3.1.5) reads as

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n^Y] = X_n + \mu, \qquad n \in \mathbb{N}.$$

The latter shows that X is a sub(super)-martingale if and only if $\mu \geq (\leq) 0$.

3.2 Examples of martingales and sub(super)-martingales

Example 3.2.1 (Levy's martingale)**.** Let $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ be a random variable, $I \subset \mathbb{R}$ be a non-empty set and $(\mathcal{F}_i)_{i\in I}$ be a filtration. The process defined by

$$X_i := \mathbb{E}[Y|\mathcal{F}_i], \qquad i \in I,$$

is a martingale. The martingale property stems directly from the tower property, i.e.

$$\mathbb{E}[X_j|\mathcal{F}_i] = \mathbb{E}\big[\mathbb{E}[Y|\mathcal{F}_j]\big|\mathcal{F}_i\big] = \mathbb{E}[Y|\mathcal{F}_i] = X_i, \qquad i < j.$$

Example 3.2.2 (Quadratic variation)**.** Let $X = (X_n)_{n\in\mathbb{N}_0}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$, such that $X_n \in L^2$ for any $n \in \mathbb{N}$. Setting

$$\langle X\rangle_n := \sum_{i=1}^{n} \mathbb{E}[Y_i^2|\mathcal{F}_{i-1}], \qquad n \in \mathbb{N}_0, \tag{3.2.1}$$

the stochastic process defined by

$$Z_n := X_n^2 - \langle X\rangle_n - X_0^2, \qquad n \in \mathbb{N}_0,$$

is a martingale with null expectation. Note that the stochastic process $\langle X\rangle$ is predictable (Definition 2.1.12). Therefore, we have

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_n^2 + Y_{n+1}^2 + 2X_nY_{n+1}|\mathcal{F}_n] - \langle X\rangle_{n+1} - X_0^2$$

(by (3.2.1))

$$= \mathbb{E}[X_n^2|\mathcal{F}_n] + 2\mathbb{E}[X_nY_{n+1}|\mathcal{F}_n] - \langle X\rangle_n - X_0^2$$

(by the properties 2 and 6-(b) in Proposition 1.4.13)

$$= \underbrace{X_n^2 - \langle X\rangle_n}_{=Z_n} + 2X_n \underbrace{\mathbb{E}[Y_{n+1}|\mathcal{F}_n]}_{=0(\text{by (3.1.6)})} - X_0^2.$$

The predictable stochastic process $\langle X\rangle$ is called *quadratic variation of* X. In particular, if X is the *fair game* martingale in Example 3.1.7, then $\langle X\rangle_n = \sum_{i=1}^n \mathbb{E}[Y_i^2]$ for any $n \in \mathbb{N}$.

Example 3.2.3. Let $X = (X_i)_{i\in I}$ be a stochastic process, and f be a convex (concave) function such that $\mathbb{E}[|f(X_i)|] < +\infty$ for any $i \in \mathcal{I}$. Then $(f(X_i))_{i\in I}$ is a sub(super)-martingale if either:

(i) X is a martingale;

(ii) X is a sub(super)-martingale and f is increasing (decreasing).

Notable cases are the functions $f(x) = x^2, e^x, |x|$ (see also Remark 3.1.16).

Example 3.2.4 (Polya's Urn)**.** Consider an urn containing an initial number of $N = R_0 + B_0$ balls, with R_0, B_0 denoting the number of red and black balls, respectively. An infinite sequence of draws takes place according to the following mechanism: after each draw, the ball is put back into the urn together with an

additional ball of the same color. Denoting by R_n the number of red balls after the n-th draw, the process $X = (X_n)_{n\in\mathbb{N}_0}$ denoting the fraction of red balls, i.e.

$$X_n := \frac{R_n}{N+n}, \qquad n \in \mathbb{N}_0,$$

is a martingale with respect to its natural filtration, i.e. $\mathcal{F}_n = \mathcal{F}_n^X = \mathcal{F}_n^R$. Indeed, we have

$$\begin{aligned}\mathbb{E}[X_{n+1}|\mathcal{F}_n] &= \mathbb{E}[X_{n+1}\mathbf{1}_{\{R_{n+1}=R_n+1\}}|\mathcal{F}_n] + \mathbb{E}[X_{n+1}\mathbf{1}_{\{R_{n+1}=R_n\}}|\mathcal{F}_n] \\ &= \mathbb{E}[\frac{R_n+1}{N+n+1}\mathbf{1}_{\{R_{n+1}=R_n+1\}}|\mathcal{F}_n] + \mathbb{E}[\frac{R_n}{N+n+1}\mathbf{1}_{\{R_{n+1}=R_n\}}|\mathcal{F}_n]\end{aligned}$$

($R_n \in m\mathcal{F}_n$)

$$= \frac{R_n+1}{N+n+1}\mathbb{P}(R_{n+1} = R_n+1|\mathcal{F}_n) + \frac{R_n}{N+n+1}\mathbb{P}(R_{n+1} = R_n|\mathcal{F}_n).$$

Now, it can be proved that (see Exercise **??**) that

$$\mathbb{P}(R_{n+1} = R_n+1|\mathcal{F}_n) = \mathbb{P}(R_{n+1} = R_n+1|R_n) = \frac{R_n}{N} = X_n.$$

Therefore, we obtain

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \frac{R_n+1}{N+n+1}X_n + \frac{R_n}{N+n+1}(1-X_n) = \frac{R_n}{N+n} = X_n.$$

3.3 Martingales and predictable processes

In this section we present two fundamental results of the theory of discrete-time stochastic processes, which link the class of martingales (and sub/super-martingales) to that of predictable processes (see Definition 2.1.12).

Theorem 3.3.1 (Doob's decomposition theorem)**.** *Let $X = (X_n)_{n\in\mathbb{N}_0}$ be a stochastic process, adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$ and such that $X_n \in L^1$ for any $n \in \mathbb{N}_0$. Then:*

(i) There exists a pair $(M, A) = (M_n, A_n)_{n\in\mathbb{N}_0}$ such that M is a martingale and A is a predictable process, such that

$$X = M + A, \qquad \text{almost surely,} \tag{3.3.1}$$

and

$$M_0 = X_0, \quad A_0 = 0, \qquad \text{almost surely.} \tag{3.3.2}$$

We call such a pair a Doob decomposition of X.

(ii) A second pair (M', A') with M adapted and A predictable is also a Doob decomposition of X if and only if $M = M'$ and $A = A'$ almost surely.

(iii) X is a sub-(super-)martingale if and only if A is almost surely non-decreasing (non-increasing), namely

$$\mathbb{P}(\{A_n \leq A_{n+1} \text{ for any } n \in \mathbb{N}_0\}) = 1. \tag{3.3.3}$$

$$(\geq)$$

Proof. We first prove (ii). Let (M, A) be a Doob decomposition of X. By (3.3.1) we have

$$X_{n+1} - X_n = M_{n+1} - M_n + A_{n+1} - A_n, \qquad n \in \mathbb{N}_0 \tag{3.3.4}$$

and, by conditioning both sides, we obtain the recursion

$$A_{n+1} = \mathbb{E}[X_{n+1}|\mathcal{F}_n] - X_n + A_n, \qquad n \in \mathbb{N}_0. \tag{3.3.5}$$

Here we employed Remark 3.1.8 and the fact that A is predictable. Also, plugging (3.3.5) into (3.3.4) yields the recursion

$$M_{n+1} = M_n + X_{n+1} - \mathbb{E}[X_{n+1}|\mathcal{F}_n], \qquad n \in \mathbb{N}_0. \tag{3.3.6}$$

Now, the recursions (3.3.5)-(3.3.6) imply almost sure uniqueness. Indeed, let (M', A') be another Doob decomposition of X. Then, by (3.3.2), $A'_0 = A_0$ and $M'_0 = M_0$. Furthermore, (M', A') has to satisfy (3.3.5)-(3.3.6), and thus

$$\begin{aligned} A_{n+1} - A'_{n+1} &= A_n - A'_n, \\ M_{n+1} - M'_{n+1} &= M_n - M'_n, \end{aligned}$$

for any $n \in \mathbb{N}_0$. Therefore, by induction, we simply obtain

$$A_n = A'_n, \quad M_n = M'_n, \qquad n \in \mathbb{N}_0,$$

and thus, by Remark 2.2.11, $M = M'$ and $A = A'$ almost surely.

Let now (M', A') be a pair with M' adapted and A' predictable. Then, if $M = M'$ and $A = A'$ almost surely, we have that M' is also a martingale and (3.3.1)-(3.3.2) hold true. Thus (M', A') is also a Doob decomposition of X.

We now prove (i): let the pair (M, A) be defined through $(M_0, A_0) = (X_0, 0)$ and the recursions (3.3.5)-(3.3.6). Then, (3.3.1) can be simply verified, by writing M and A as

$$\begin{aligned} M_n &= M_0 + \sum_{i=1}^{n} (M_i - M_{i-1}) = M_0 + \sum_{i=1}^{n} \big(X_i - \mathbb{E}[X_i|\mathcal{F}_{i-1}]\big), \\ A_n &= \sum_{i=1}^{n} (A_i - A_{i-1}) = \sum_{i=1}^{n} \big(\mathbb{E}[X_i|\mathcal{F}_{i-1}] - X_{i-1}\big), \end{aligned} \tag{3.3.7}$$

for any $n \in \mathbb{N}_0$. The fact that M is a martingale and A is predictable also follows easily from (3.3.7).

We finally prove (iii). By (3.3.5) we have that X is a sub-(super)martingale if and only if

$$\begin{gathered} A_{n+1} - A_n \geq 0, \qquad n \in \mathbb{N}_0. \\ (\leq) \end{gathered}$$

An argument analogous to (2.2.5) shows that the latter is equivalent to (3.3.3). □

Definition 3.3.2 (α-transform of X)**.** Let $X = (X_n)_{n\in\mathbb{N}_0}$ and $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ be two stochastic processes. The stochastic process $G(\alpha, X) = (G(\alpha, X)_n)_{n\in\mathbb{N}_0}$ given by

$$G(\alpha, X)_n := \sum_{i=1}^{n} \alpha_i (X_i - X_{i-1}), \qquad n \in \mathbb{N}_0,$$

is called the *α-transform of X*.

The following result roughly states that the α-transform of a martingale is a null-mean martingale if α is bounded and predictable, and, viceversa, a process is a martingale if its α-transform has constant, null, mean value for any bounded predictable process α.

Proposition 3.3.3. *Let $X = (X_n)_{n\in\mathbb{N}_0}$ be a stochastic processes adapted to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$, and such that $X_n \in L^1$ for any $n \in \mathbb{N}_0$. Then we have:*

(i) If $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ is a bounded predictable process, then $G(\alpha, X)$ is a martingale with

$$\mathbb{E}[G(\alpha, X)_n] = 0, \qquad n \in \mathbb{N}_0; \tag{3.3.8}$$

(ii) If (3.3.8) holds true for any $\alpha = (\alpha_n)_{n\in\mathbb{N}}$ bounded predictable process, then X is a martingale.

Proof. Part (i). As $X_n \in L^1$ and α is bounded, we clearly have $G(\alpha, X)_n \in L^1$. Furthermore, as α is predictable (thus adapted) and X is adapted, then $G(\alpha, X)$ is adapted. Now, by definition of $G(\alpha, X)$, we have

$$G(\alpha, X)_{n+1} = G(\alpha, X)_n + \alpha_{n+1}(X_{n+1} - X_n), \qquad n \in \mathbb{N}_0.$$

Therefore, by the linearity of the conditional expectation, we obtain

$$\mathbb{E}[G(\alpha, X)_{n+1}|\mathcal{F}_n] = \mathbb{E}[G(\alpha, X)_n|\mathcal{F}_n] + \mathbb{E}[\underbrace{\alpha_{n+1}}_{\in bm\mathcal{F}_n}(X_{n+1} - X_n)|\mathcal{F}_n]$$

by properties (1.4.5)-(1.4.8) of the conditional expectation

$$= G(\alpha, X)_n + \alpha_n\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] = G(\alpha, X)_n, \qquad n \in \mathbb{N}_0,$$

where the last inequality stems from Remark 3.1.8. Finally, (3.3.8) is obvious as martingales have constant expectation and $G(\alpha, X)_0 = 0$ by definition.

Part (ii). Fix $n \in \mathbb{N}_0$ and $A \in \mathcal{F}_n$. Let α be defined as

$$\alpha_m := \begin{cases} \mathbf{1}_A & \text{if } m = n+1 \\ 0 & \text{if } m \in \mathbb{N}_0 \setminus \{n+1\} \end{cases}.$$

Then we have

$$G(\alpha, X)_{n+1} = \alpha_{n+1}(X_{n+1} - X_n) = \mathbf{1}_A(X_{n+1} - X_n),$$

and thus, by (3.3.8),

$$\mathbb{E}[\mathbf{1}_A(X_{n+1} - X_n)] = 0.$$

Since $A \in \mathcal{F}_n$ is arbitrary, Remark 1.4.3 yields $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_n|\mathcal{F}_n]$. This concludes the proof.

□

3.4 Martingales and stopping times

3.5 Maximal inequalities and L^2 convergence

3.6 Limit theorems

3.7 Exercises

Exercise 3.7.1. Let $X = (X_i)_{i\in I}$ be an adapted stochastic process. Show that X is a martingale if and only if the process given by $\widetilde{X}_n := X_n - X_0$ is a martingale with null expectation.

Exercise 3.7.2. Let $X = (X_i)_{i\in I}$ be a non-negative sub-martingale and f be a convex increasing function. Prove that $(f(X_i))_{i\in I}$ is a sub-martingale. *Hint: the definitions of conditional expectation and of sub-martingale need to be extended.*

Chapter 4

Markov processes

Chapter 5

Stochastic Gradient Descent Method

The *stochastic gradient descent (SGD)* method is an iterative optimization algorithm that became very popular due to its widespread application to the training of learning models. SGD is part of the family of so-called stochastic approximation methods. It can be thought of as a probabilistic extension of the well-known *deterministic gradient descent* method, hereafter simply *gradient descent (GD)*, introduced by Cauchy in 1847. The original idea of SGD was introduced in 1951 by Robbins and Monro. Subsequently, many different variants and refinements were proposed by several authors, especially in the last 3 decades, mostly prompt by the increasing demand of efficient and robust algorithms for the minimization of the loss function associated to complex learning models, such as neural networks. Therefore, SGD is arguably to be understood as a family of methods rather than a specific algorithm.

In Machine Learning, the advantages of SGD methods over GD are multiple:

- They typically allow for faster computation compared to GD. While the latter requires the computation of the gradient of the penalty function at all the points of the training set, SGD requires the computation of the gradient only at a "small" subset of points (in theory one single point);

- The intrinsic randomness of the method makes it easier for the iterates to escape from "bad" regions in which deterministic methods are naturally attracted to strictly local minima. Although mathematical proofs are scarce and only cover particular cases, there is concrete empirical evidence of the fact that SGD methods can reach global minima even in non-convex optimization problems;

- Compared to GD, SGD methods would typically converge to local minima with lower curvature, which translates into better generalization capability of the trained learning model. In other words, overfitting phenomena are easier to avoid by training a model with SGD.

The rest of the chapter is structured as follows: in Section 5.1 we describe the deterministic gradient descent algorithm, in Section 5.2 we describe a simple SGD algorithm for which it is possible to establish the convergence rate under fairly simple assumptions. In Section 5.3 we discuss some applications of SGD to Machine Learning problems, highlighting the advantages over GD. In Section 5.4 we outline some refined versions of SGD that are commonly employed in the training of learning models.

5.1 Gradient Descent Method

The goal is to find the minima of a function

$$f : D \to \mathbb{R}, \qquad D \subset \mathbb{R}^d,$$

which enjoys suitable regularity properties. To start familiarizing with the ideas, we can first assume that f is differentiable at any internal point of its domain D. This assumption will be later considerably relaxed. After fixing an initial point in D, say $w^{(0)} = 0$, the point $w^{(1)}$ is obtained by moving away from $w^{(0)}$ in the opposite direction of the gradient of f at $w^{(0)}$, with step-size $\eta_n > 0$. The same procedure is then iterated for N times. Precisely, we set

$$\begin{aligned} &w^{(0)} := 0, \\ &w^{(n)} := w^{(n-1)} - \eta_n \nabla f(w^{(n-1)}), \qquad n = 1, \cdots, N. \end{aligned} \tag{5.1.1}$$

We refer to the point $w^{(n)}$ as to the n-th iterate, while the constants η_n are called learning rates. After the N-th step, different choices of output, hereafter $\overline{w}^{(N)}$, can be made. One obvious choice consists in selecting either the last vector, i.e. $\overline{w}^{(N)} := w^{(N)}$, or the minimizer, i.e. $\overline{w}^{(N)} := \arg\min_{n=0,\cdots,N} f(w^{(n)})$. A less obvious choice is given by the arithmetic average

$$\overline{w}^{(N)} := \frac{1}{N+1} \sum_{n=0}^{N} w^{(n)}. \tag{5.1.2}$$

This last option turns out to be particularly useful when considering the stochastic version of the method, as well as to relax the regularity assumptions on f. The basic version of the GD method can then be summarized as follows.

GD minimization algorithm for f
Input: integer $N > 0$, positive sequence $\eta_1, \ldots, \eta_N$
Initialize: $w^{(0)} = 0$
Iterate: for any $n = 1, \cdots, N$:
$w^{(n)} = w^{(n-1)} - \eta \nabla f(w^{(n-1)})$
Output: $\overline{w}^{(N)} := w^{(N)}$, or $\overline{w}^{(N)} := \arg\min_{n=0,\cdots,N} f(w^{(n)})$, or $\overline{w}^{(N)} = \frac{1}{N+1} \sum_{n=0}^{N} w^{(n)}$

Definition 5.1.1. Let $D \subset \mathbb{R}^d$ and $\rho > 0$. A function $f : D \to \mathbb{R}$ is called *ρ-Lipschitz continuous* if

$$|f(x) - f(y)| \leq \rho |x - y|, \qquad x, y \in D.$$

Theorem 5.1.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, differentiable and ρ-Lipschitz continuous function, and let $R > 0$, $N \in \mathbb{N}$. Setting*

$$m := \min_{|w| \leq R} f, \qquad \eta := \frac{R}{\rho \sqrt{N+1}}, \tag{5.1.3}$$

and $\overline{w}^{(N)}$ as in (5.1.2)*, we have*

$$f(\overline{w}^{(N)}) - m \leq \frac{R\rho}{\sqrt{N+1}}. \tag{5.1.4}$$

Before proving this theorem we highlight some of its important consequences.

Remark 5.1.3. The point $\overline{w}^{(N)}$ in estimate (5.1.4) is doubly dependent on N: not only through the number of steps but also through the choice of η in (5.1.3). Nevertheless, in the sequel we prefer not to make the latter dependence explicit to avoid burdening the notation.

Remark 5.1.4. Assume that f has a global minimum in $w^* \in \mathbb{R}^d$. Then, Theorem 5.1.2 ensures the convergence of $f(\overline{w}^{(N)})$ to the minimum $f(w^*)$, provided that $R \geq |w^*|$. Indeed, (5.1.4) together with $f(\overline{w}^{(N)}) - f(w^*) \geq 0$ yields

$$|f(\overline{w}^{(N)}) - f(w^*)| \leq \frac{R\,\rho}{\sqrt{N+1}}, \qquad N \in \mathbb{N}.$$

In particular, one can approximate $f(w^*)$ with arbitrary precision by setting N suitably large. Precisely, for any $\epsilon > 0$ we have

$$|f(\overline{w}^{(N)}) - f(w^*)| \leq \epsilon$$

if $N \geq \frac{R^2\rho^2}{\epsilon^2} - 1$.

Remark 5.1.5. If f has a global minimum in w^*, it is not granted that the sequence $\left(\overline{w}^{(N)}\right)_N$ converges to w^*, unless w^* is the unique minimizer.

The proof of Theorem 5.1.2 relies on the following

Lemma 5.1.6. *Let $v_1, \cdots v_N, w^*$ a sequence of vectors in $\mathbb{R}^d$, and $\eta > 0$. Setting*

$$\begin{aligned} w^{(0)} &= 0, \\ w^{(n)} &= w^{(n-1)} - \eta\, v_n, \qquad n \in \mathbb{N}, \end{aligned}$$

we have

$$\sum_{n=0}^{N} \langle v_{n+1}, w^{(n)} - w^* \rangle \leq \frac{|w^*|^2}{2\eta} + \frac{\eta}{2} \sum_{n=0}^{N} |v_{n+1}|^2. \tag{5.1.5}$$

In particular, we have

$$\frac{1}{N+1} \sum_{n=0}^{N} \langle v_{n+1}, w^{(n)} - w^* \rangle \leq \frac{R\rho}{\sqrt{N+1}} \tag{5.1.6}$$

for any $R, \rho > 0$ such that:

$$\eta = \frac{R}{\rho\sqrt{N+1}} \tag{5.1.7}$$

and

$$|w^*| \leq R, \qquad |v_n| \leq \rho, \quad n = 1, \cdots, N+1. \tag{5.1.8}$$

Proof. A direct computation shows

$$\begin{aligned} \langle v_{n+1}, w^{(n)} - w^* \rangle &= \frac{1}{2\eta}\Big(|w^{(n)} - w^*|^2 + \eta^2 |v_{n+1}|^2 - |w^{(n)} - w^* - \eta v_{n+1}|^2 \Big) \\ &= \frac{1}{2\eta}\Big(|w^{(n)} - w^*|^2 - |w^{(n+1)} - w^*|^2 \Big) + \frac{\eta}{2} |v_{n+1}|^2. \end{aligned}$$

Adding up w.r.t. n yields

$$\sum_{n=0}^{N} \langle v_{n+1}, w^{(n)} - w^* \rangle = \frac{1}{2\eta} \sum_{n=0}^{N} \Big(|w^{(n)} - w^*|^2 - |w^{(n+1)} - w^*|^2 \Big) + \frac{\eta}{2} \sum_{n=0}^{N} |v_{n+1}|^2$$

(the first one is a telescopic sum, and $w^{(0)} = 0$)

$$= \frac{1}{2\eta}\Big(|w^*|^2 - |w^{(N)} - w^*|^2\Big) + \frac{\eta}{2}\sum_{n=0}^{N} |v_{n+1}|^2,$$

which proves (5.1.5). To prove (5.1.6) it is enough to observe that, if (5.1.8) holds true, then

$$\frac{|w^*|^2}{2\eta} + \frac{\eta}{2}\sum_{n=0}^{N} |v_{n+1}|^2 \leq \frac{R^2}{2\eta} + \frac{\eta\,(N+1)\rho^2}{2}$$

(if (5.1.7) also holds)

$$= R\rho\,(N+1).$$

□

Proof of Theorem 5.1.2. Recalling that f is convex, Jensen's inequality yields

$$f(\overline{w}^{(N)}) = f\bigg(\frac{1}{N+1}\sum_{n=0}^{N} w^{(n)}\bigg) \leq \frac{1}{N+1}\sum_{n=0}^{N} f(w^{(n)}).$$

Therefore, for any $w^* \in \arg\min_{|w|\leq B}$, we obtain

$$f(\overline{w}^{(N)}) - m = f(\overline{w}^{(N)}) - f(w^*) \leq \frac{1}{N+1}\sum_{n=0}^{N} \big(f(w^{(n)}) - f(w^*)\big)$$

(f is convex)

$$\leq \frac{1}{N+1}\sum_{n=0}^{N} \big\langle \underbrace{\nabla f(w^{(n)})}_{=v_{n+1}}, w^{(n)} - w^* \big\rangle.$$

Finally, (5.1.4) stems from (5.1.8) in Lemma 5.1.6. □

5.2 Stochastic Gradient Descent (SGD) Method

We now present a stochastic variation of the GD method presented in the previous section. The basic idea is to replace the gradient $\nabla f(w^{(n-1)})$ in (5.1.1) with a random variable $\mathbf{v}_n$ whose expected value, conditioned to the information up to time n, coincides with $\nabla f(w^{(n)})$.

Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$, and an adapted stochastic process $(\mathbf{v}_n)_{n\in\mathbb{N}}$. Let also $\eta > 0$ and set

$$\mathbf{w}^{(0)} := 0, \tag{5.2.1}$$

$$\mathbf{w}^{(n)} := \mathbf{w}^{(n-1)} - \eta\mathbf{v}_n, \qquad n \in \mathbb{N}. \tag{5.2.2}$$

Fix now $N \in \mathbb{N}$. The process $(\mathbf{v}_n)_n$ will be required to satisfy the following

Assumption 5.2.1. For any $n = 1, \cdots, N+1$,

$$\mathbb{E}_{n-1}[\mathbf{v}_n] = \nabla f(\mathbf{w}^{(n-1)}) \quad \mathbb{P}\text{-a.s.}.$$

The output of the algorithm is set once more as

$$\overline{\mathbf{w}}^{(N)} := \frac{1}{N+1}\sum_{n=0}^{N}\mathbf{w}^{(n)}.$$

To sum up, the basic algorithm of the SGD method reads as

SGD minimization algorithm for f
Input: real $\eta > 0$, integer $N > 0$
Initialize: $\mathbf{w}^{(0)} = 0$
Iterate: for any $n = 1, \cdots, N$
sample from a r.v. $\mathbf{v}_n$ such that $\mathbb{E}_{n-1}[\mathbf{v}_t] = \nabla f(\mathbf{w}^{(n-1)})$ $\mathbb{P}$-a.s.
$\mathbf{w}^{(n)} = w^{(n-1)} - \eta\,\mathbf{v}_n$
Output: $\overline{\mathbf{w}}^{(N)} = \frac{1}{N+1}\sum_{n=0}^{N}\mathbf{w}^{(n)}$

Before presenting a convergence result for the SGD algorithm, analogous the one in Theorem 5.1.2 for deterministic SG, we present some examples.

Theorem 5.2.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, and $R > 0$, $N \in \mathbb{N}$. Let also Assumption 5.2.1 be in force and $\rho > 0$ such that*

$$|\mathbf{v}_n| \le \rho, \qquad n = 1, \cdots, N. \tag{5.2.3}$$

Setting

$$m := \min_{|w|\le R} f, \qquad \eta := \frac{R}{\rho\sqrt{N+1}},$$

we have

$$\mathbb{E}\big[f(\overline{\mathbf{w}}^{(N)})\big] - m \le \frac{R\,\rho}{\sqrt{N+1}}.$$

Proof. By applying Lemma 5.1.6 to the random sequence $\mathbf{v}_n$, $n = 1, \cdots, N+1$, and by the monotony of the expected value, we obtain

$$\frac{1}{N+1}\mathbb{E}\bigg[\sum_{n=0}^{N}\langle \mathbf{v}_{n+1}, \mathbf{w}^{(n)} - w^*\rangle\bigg] \le \frac{R\rho}{\sqrt{N+1}}. \tag{5.2.4}$$

Furthermore, the same argument used in the proof of Theorem 5.1.2, together with the monotonicity of the expected value, yields

$$\mathbb{E}\big[f(\overline{w}^{(N)})\big] - m \le \frac{1}{N+1}\mathbb{E}\bigg[\sum_{n=0}^{N}\big\langle \nabla f(\mathbf{w}^{(n)}), \mathbf{w}^{(n)} - w^*\big\rangle\bigg]$$

(by Assumption 5.2.1)

$$= \frac{1}{N+1}\mathbb{E}\bigg[\sum_{n=0}^{N}\big\langle \mathbb{E}_n\big[\mathbf{v}_{n+1}\big], \mathbf{w}^{(n)} - w^*\big\rangle\bigg]$$

((5.2.2) and (5.2.3) imply $\mathbf{w}^{(n)} \in bm\mathcal{F}_n$)

$$= \frac{1}{N+1}\mathbb{E}\bigg[\mathbb{E}_n\bigg[\sum_{n=0}^{N}\big\langle \mathbf{v}_{n+1}, \mathbf{w}^{(n)} - w^*\big\rangle\bigg]\bigg]$$

(by the total probability formula)

$$= \frac{1}{N+1}\mathbb{E}\Big[\sum_{n=0}^{N}\big\langle \mathbf{v}_{n+1}, \mathbf{w}^{(n)} - w^*\big\rangle\Big].$$

This, together with (5.2.4), concludes the proof. □

5.3 Applications to Machine Learning

5.4 Advanced methods

Appendix A

A primer on Probability Theory with Measure Theory

Theorem A.0.1 (Radon-Nicodym Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space, and ν be another σ-finite measure on $(\Omega, \mathcal{F})$ such that $\nu << \mu$. Then, there exists $L \in m\mathcal{F}^+$ such that:*

(i) $L \in L^1(\Omega, \mathcal{F}, \mu)$;

(i) $\nu(A) = \int_A L \, d\mu$ for any $A \in \mathcal{F}$.

Furthermore, if $L' \in m\mathcal{F}^+$ is another r.v. that satisfies $(i), (ii)$, then $L = L'$ μ-almost everywhere, and thus also ν-almost everywhere.

Definition A.0.2. Under the assumptions of Theorem A.0.1, the set of random variables L satisfying (i)-(ii) is a non-empty equivalence class w.r.t. the " $\overset{\text{a.s.}}{=}$ " equivalence relation, which is called *Radon-Nikodym derivative of ν w.r.t. μ*. It is denoted by

$$\frac{d\nu}{d\mu}, \qquad \text{or by} \quad \frac{d\nu}{d\mu}\Big|_{\mathcal{F}}$$

if we want to stress the dependence on the σ-algebra $\mathcal{F}$.

Bibliography

[1] O. Kallenberg and O. Kallenberg, *Foundations of modern probability*, vol. 2, Springer, 1997.

[2] A. Pascucci, *PDE and martingale methods in option pricing*, Springer Science & Business Media, 2011.

[3] ——, *Teoria Della Probabilità*, Springer, 2020.