Lecture Notes on Probability Theory
Chapters 1 to 6 (circa weeks 1-10)

Fall 2022 course at ETH, Wendelin Werner

Table of Contents

The goal of this course is mostly to study **stochastic processes in discrete time**, i.e. infinite sequences $X_0,\ X_1,\ X_2, \ldots,\ X_n, \ldots$ of random variables (where $0, 1, 2, \ldots$ can be viewed as the discrete time, ie., X_n is the value of X at time n).

One of our main focus of interest will be the **asymptotic behavior** of X_n (or of functions of $(X_1, \ldots, X_n)$) as $n \to \infty$. Key words related to the cases that we will study are: Law of large numbers, series of independent random variables, central limit theorem, martingales and random walks.

The martingale convergence theorems will enable us to discuss the explosion rates of Galton-Watson processes and their now famous R-number.

Warm-up, basic discrete probability

In this warm-up section, we briefly review some ideas and elementary facts from discrete probability, and explain what sort of conceptual issues one runs into when one tries to state "limiting theorems" i.e., results involving an infinite collection of random inputs. This serves as a motivation for the remainder of the lectures and explains why the formalism of measure theory can be very useful here.

Suppose that Ω is a discrete (finite or countable) set. The idea here is that Ω represents the set of all possible outcomes of a random experiment (so y can be something else than just one number). To each point y in Ω, one then assigns the probability $P(y)$ that y is the outcome of this experiment. Here, each $P(y)$ is a real number in $[0,1]$ (a 20% chance would for instance correspond to $P(y) = 0.2$), and one obviously has $\sum_{y\in\Omega} P(y) = 1$. The function P is called a probability measure on Ω.

When A is a subset of Ω (we will sometimes call such subsets "events"), one can define the probability that A happens to be

$$P(A) := \sum_{y\in A} P(y)$$

(so with this notation, $P(\{y\}) = P(y)$). When $P(A) > 0$, one can then also wonder how things look when one has the knowledge that A actually happens. In that case, one can define a new probability measure, which is the *conditional probability measure given* A denoted by P_A to be

$$P_A(y) = \frac{P(y)}{P(A)} \text{ when } y \in A$$

and $P_A(y) = 0$ when $y \notin A$. So, one restricts P to the set A and multiplies the obtained function by $1/P(A)$ so that one again has a probability measure P_A. One sometimes also uses the notation $P(y|A)$ for $P_A(y)$.

This then leads naturally to the following notion: Two events A and B are called *independent* when

$$P(A \cap B) = P(A)P(B).$$

We can note that $\emptyset$ and Ω are independent with any other events, and more generally, one can easily check that if $P(A) = 0$ or $P(A) = 1$, then A is independent of any other set. When $P(A) > 0$, the independence between A and B means that $P_A(B) = P(B)$, i.e., that the occurrence of A does not influence the likelihood of B happening – which corresponds indeed to the intuitive notion of independence.

EXAMPLE 0.-1.1. *Suppose that* Ω *is the* 6×6 *square, i.e.,*

$$\Omega = \{(i,j) \ : \ i \in \{1,\dots,6\}, j \in \{1,\dots,6\}\} = \{1,2,\dots,6\}^2$$

and that P *assigns a probability* $1/36$ *to each of the points.*

We can look at the right half of the square

$$A := \{(i,j) \in \Omega \ : \ i \geq 4\},$$

and the top line

$$B := \{(i,j) \in \Omega \ : \ j = 6\}.$$

The probability of an event is $1/36$ *times the number of elements in the event, so we get that* $P(A) = 18/36 = 1/2$*,* $P(B) = 6/36 = 1/6$*, and* $P(A \cap B) = 3/36 = 1/12$*, so that* A *and* B *are independent.*

In fact the previous example suggests the following formalism. When one has a discrete probability space, then a function X from Ω into $\mathbb{R}$ is called a random variable. It is "a number that one can read as part of the outcome of the experiment". For instance, one can consider the functions $X_1(i,j) = i$ and $X_2(i,j) = j$. One can then also consider other random variables, such as $X(i,j) = i + j$, or $X(i,j) = \max(i,j)$ etc. We can note that the set of possible outcomes of a given random variable X is contained in $X(\Omega)$, which is a subset of $\mathbb{R}$ that is at most countable, because Ω is finite or countable.

The *law* of the random variable X then describes the various probabilities of occurrence of X. In other words, for each $x \in X(\Omega)$, one has

$$P(\{y \in \Omega \ : \ X(y) = x\}) = \sum_{y \in \Omega \ : \ X(y) = x} P(y).$$

One often just writes $P(X = x)$ or $P(\{X = x\})$ for this quantity.

One then says that two random variables X_1 and X_2 are *independent* if for all x_1 and x_2,

$$P(X_1 = x \text{ and } X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$$

(here the left-hand side stands for

$$P(\{y \in \Omega \ : \ X_1(y) = x_1 \text{ and } X_2(y)\})$$

which is in fact the probability of the intersection $\{X_1 = x_1\} \cap \{X_2 = x_2\}$). This intuitively means that what X_1 reveals has no influence on what X_2 reveals.

This is what happens in the example above since for all $(i,j) \in \{1, \dots, 6\}^2$,

$$P(X_1 = i, X_2 = j) = P((i,j)) = 1/36 = (1/6)^2 = P(X_1 = i)P(X_2 = j).$$

This corresponds to the idea that one can view this as the throw of two successive dices. The previous events A and B correspond to the fact that the outcome of the first one is greater or equal to 4 while the event B corresponds to the fact that the outcome of the second one is a 6.

Similarly, one says that k random variables $X_1, \dots, X_k$ defined on Ω are independent, if for all $x_1, \dots, x_k$,

$$P(X_1 = x_1, \dots, , X_k = x_k) = P(X_1 = x_1) \dots P(X_k = x_k).$$

This would correspond to the fact that $X_1, \dots, X_k$ could be viewed as the outcomes of totally independent experiments.

REMARK 0.-1.2. *One says that $(X_1, \dots, X_k)$ are pairwise independent, if for any $i \neq j$, X_i and X_j are independent. This is a weaker condition than to say that $(X_1, \dots, X_k)$ are independent. One can for instance have a look at the example where $k = 3$, where X_1 and X_2 are independent and equal to ± 1 with probability $1/2$, and $X_3 = X_1 X_2$ (see the exercise sheet).*

REMARK 0.-1.3. *There are a number of probability measures on the set of non-negative integers, that then define the laws of random variables that one often uses to model real-life phenomena. This includes the binomial distributions, the geometric distributions and the Poisson distributions. [Personal comment: The most interesting ones in my view are the Poisson distributions!].*

In this course, we will want to study cases where there are "infinitely" (but countably) many random inputs. The simplest example is arguably that of an infinite sequence of (independent) fair coin tosses. The state Ω would then look like $\{0,1\}^{\{1,2,\cdots\}}$. In other words, the outcome would be an element of the space Ω of all infinite sequence $(x_n)_{n \geq 1}$ where each x_n takes the value 0 or 1 (depending if the outcome of the throw number n is a tail or a head). A typical result that we will want to derive will be that (for a sequence of independent fair throws, so that each throw has a probability $1/2$ to be head and $1/2$ to be tail), that with probability one, the proportion of heads

among the first n throws tends to $1/2$ as $n \to \infty$. In other words, if A is the set of infinite sequences in Ω with the property that $(x_1 + \ldots + x_n)/n \to 1/2$ as $n \to \infty$, one wants to prove that $P(A) = 1$.

Here, one encounters a new problem: The set $\Omega = \{0,1\}^{\{1,2,\ldots\}}$ is not countable. So, we are not in the setup described above anymore. In fact, one can see that defining such a random sequence of independent coin tosses is equivalent to defining the Lebesque measure on the interval $[0,1]$, so that one ends up in the realm of measure theory. Let us now give a heuristic explanation to this (this is not meant to be rigorous – we have not yet rigorously defined sequences of independent variables –, it serves more here as a motivation for the formalism that will come up next).

- Suppose that one has been able to define the infinite collection of independent coin tosses, that we represent as a sequence of independent random variables $(X_n)_{n\geq 1}$ where each X_j takes the values 0 and 1 with probability $1/2$. Then, we can construct the number $X = \sum_{n\geq 1} X_n 2^{-n}$. This is the number that can be written in dyadic form as $.X_1X_2X_3\ldots$. For each dyadic interval of the type $[j2^{-n}, (j+1)2^{-n})$, the probability that X belongs to this interval is the probability that $X_1, \ldots, X_n$ take the specific values imposed by the dyadic decomposition of $j2^{-n}$, and this has probability $(1/2)^n$. Taking the disjoint union of j dyadic intervals of length 2^{-n} then shows that for all $j \leq 2^n$,
$$P(X \in [0, j2^{-n})) = j2^{-n},$$
which suggest that X falls "uniformly" in $[0,1]$.
- Conversely, if Y is a number that falls "uniformly" at random in the interval $[0,1]$, then one can write it in its dyadic form $Y = \sum_{n\geq 1} Y_n 2^{-n}$. This defines the sequence $(Y_n)_{n\geq 1}$, and it is an easy exercise (see the exercise sheet) to check that the sequence $(Y_n)_{n\geq 1}$ would behave just like a sequence of independent coin tosses.

So, in a nutshell, as soon as one wants to study sequences of independent random inputs, one ends up having to define probabilities in non-countable spaces Ω, which is one of the things that measure theory is good for. The simplest example of a sequence of independent coin tosses is already equivalent to the definition of the Lebesgue measure.

In particular, when Ω is not countable, one important tricky question is to determine *for which subsets of Ω one can actually observe whether they are true or not.* For instance, in the case of sequence of independent coin tosses where Ω is the set of sequences $(x_n)_{n\geq 1}$ with values in $\{0,1\}$, one will clearly want to say that for each j_0, one can observe whether the j_0-th outcome is a 0 or a 1, i.e., the set of sequences $(x_n)_{n\geq 1}$ with $x_{j_0} = 1$ will be observable, and its probability will be $1/2$. But the set $\mathcal{P}(\Omega)$ of all subsets of Ω is huge (choosing a set in Ω is deciding for each $x \in \Omega$ whether it is inside or outside, so $\mathcal{P}(\Omega)$ is like $\{i,o\}^{\Omega}$), and it will actually in general not be possible to assign a probability to each of these subsets of Ω. Only some class of special subsets will be "accessible". This is the motivation behind the notions of σ-fields in measure theory. In the case of the sequence of independent coin tosses, this set of observable sets will be related to the Borel sets in $[0,1]$.

CHAPTER 1

Abstract generalities: Probability space

1.1. Some basic measure theory background

The mathematical formalism of probability theory is the same as that of measure theory, but with an important new ingredient, namely the notion and use of *independence*. In this section, we briefly review the usual measure theory aspects.

Intuitively, the first purpose of measure theory is to be able to assign to some subsets A of a given set Ω a quantity $\mu(A)$ in $[0,\infty]$ called the measure of A, that one can think off as its mass. Note that here, the set Ω is allowed to be huge. Two questions have to be asked:

- What is the family $\mathcal{A}$ of sets A for which one can define $\mu(A)$?
- What are the properties that the function μ should then satisfy?

Intuitively, it is for instance pretty clear that if A and B are disjoint sets that are both in $\mathcal{A}$, then $A \cup B$ should be in A as well, and that $\mu(A) + \mu(B) = \mu(A \cup B)$. In the formalism of measure theory, one asks for somewhat more (but all this is of course very intuitive):

(1) The family $\mathcal{A}$ should be a σ-algebra (also called a σ-field):

DEFINITION 1.1.1. *A family $\mathcal{A}$ of subsets of Ω is called a* **σ-algebra** *if it satisfies the following three conditions: (a) $\Omega \in \mathcal{A}$, (b) If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$, and (c) If a sequence $(A_j)_{j\geq 1}$ is in $\mathcal{A}$, then $\cup_{j\geq 1} A_j$ is in $\mathcal{A}$ as well.*

(2) The map $\mu : \mathcal{A} \to [0,\infty)$ should be a measure. Since we will be almost always working with finite measures (i.e. with a finite total mass), we do not bother defining infinite measures here:

DEFINITION 1.1.2. *The map μ is a* **finite measure** *on $(\Omega, \mathcal{A})$ if for any sequence $(A_j)_{j\geq 1}$ of pairwise disjoint sets in $\mathcal{A}$ (i.e., such that for all $j_1 \neq j_2$, $A_{j_1} \cap A_{j_2} = \emptyset$) one has*

$$\mu(\cup_{j\geq 1} A_j) = \sum_{j\geq 1} \mu(A_j)$$

(note that we know that $\cup_{j\geq 1} A_j \in \mathcal{A}$ because $\mathcal{A}$ is a σ-algebra).

REMARK 1.1.3. *We used "sequences" of sets $(A_j)_{j\geq 1}$ in these definitions, but we could have equivalently used "countable families" $(A_j)_{j\in J}$ for J countable, since the order of the A_j does not matter in these definitions.*

DEFINITION 1.1.4. *A* **probability space** *is a triplet $(\Omega, \mathcal{A}, P)$ where $\mathcal{A}$ is a σ-field on Ω, and P is a measure on $\mathcal{A}$ with total mass equal to 1, i.e., $P(\Omega) = 1$. Elements of $\mathcal{A}$ are called* **events** *and the measure P is called a* **probability measure**.

REMARK 1.1.5. *As opposed to the warm-up section, there are no conditions here on how large Ω can be.*

The heuristic interpretation goes as follows:

- The set Ω represents everything that can happen in the system. (There might however be some information in $\omega \in \Omega$ that is actually impossible to access).

- The collection $\mathcal{A}$ is the set of events that a given observer is actually able to detect, i.e., that observer can say if a given event $A \in \mathcal{A}$ holds or not.
- The quantity $P(A)$ is the the probability that this event A holds.

Sometimes, we will use sub-σ-fields of $\mathcal{A}$ – ie. subsets of $\mathcal{A}$ that are also themselves σ-fields. Such a family $\mathcal{A}'$ can be thought off as the collection of events that some observer that does not has access to all the information provided in $\mathcal{A}$ is able to detect.

EXAMPLE 1.1.6. *A typical example occurs for instance when one studies a random "evolution" (i.e., something random that takes place in real-time). Then, for each time t, one can consider a σ-algebra $\mathcal{A}_t$ representing everything that can be observed when looking at the system up to time t. Then, clearly, $\mathcal{A}_t$ is a sub-σ-algebra of $\mathcal{A}_{t'}$ when $t \leq t'$.*

REMARK 1.1.7. *One aspect has to be stressed here: This measure-theoretical framework postulates the (a priori not totally intuitive fact) that if one is able to "observe" separately a sequence of events $(A_j)_{j\geq 1}$ then, one is able to observe them "all at once", ie. for instance to check whether they all hold simultaneously (because $\cap_{j\geq 1} A_j$ is then still in the σ-field).*

DEFINITION 1.1.8. *An event $A \in \mathcal{A}$ with $P(A) = 1$ is said to hold* **almost surely**.

REMARK 1.1.9. *When A and A' are two events in $\mathcal{A}$ with $A \subset A'$, one obviously has $P(A) \leq P(A')$ because $P(A) + P(A' \setminus A) = P(A')$ (note that $A' \setminus A = A' \cap (\Omega \setminus A) \in \mathcal{A}$. Similarly, if $A_1, \ldots, A_n$ is a finite collection of events, then we can define the disjoint events $C_1 = A_1$ and $C_j = A_j \setminus (A_1 \cup \ldots \cup A_{j-1})$ for $j \in \{2, \ldots, n\}$, to see that*

$$P(A_1 \cup \ldots \cup A_n) = P(C_1 \cup \ldots \cup C_n) = \sum_{j=1}^{n} P(C_j) \leq \sum_{j=1}^{n} P(A_j).$$

REMARK 1.1.10. *A simple consequence of the definition of σ-fields and of measures is that (see the exercise sheet) that when $(A_n)_{n\geq 1}$ is a* non-decreasing *sequence of events (so that $A_n \subset A_{n+1}$ for all n), then $\cup_{n\geq 1} A_n$ is also in $\mathcal{A}$ and the probability of this event is the non-decreasing limit of the $P(A_n)$, i.e., $P(\cup_{n\geq 1} A_n) = \lim_{n\to\infty} P(A_n)$. Similarly, if $(A_n)_{n\geq 1}$ is non-increasing, then $P(\cap_{n\geq 1} A_n) = \lim_{n\to\infty} P(A_n)$.*
Further consequences are that:

- *If $(A_j)_{j\in J}$ is a countable collection of events with probability 0, then $\cup_{j\in J} A_j$ also has probability 0.*
- *If $(A_j)_{j\in J}$ is a countable collection of almost sure events, then $\cap_{j\in J} A_j$ holds almost surely too. In other words, when "for all $j \in J$, A_j holds almost surely" implies that "almost surely, A_j holds for all $j \in J$" when J is countable.*

REMARK 1.1.11. *One has to be careful that this interchange of "for all $j \in J$" and "almost surely" does not hold (in general) when J is uncountable. For example, if P is a probability measure on some space such that has no atoms (i.e., $P(\{x\}) = 0$ for all $x \in \Omega$ — one example will be the Lebesgue measure in $[0, 1]$), then: For all given $x \in \Omega$, almost surely, the event $A_x := \Omega \setminus \{x\}$ holds because $P(A_x) = 1 - P(\{x\}) = 1$, but the event $\cap_{x\in\Omega} A_x$ is empty, so that it is* not *the case that almost surely, for all $x \in \Omega$, A_x holds.*

REMARK 1.1.12. *Another immediate consequence of the definition of measures is the so-called first Borel-Cantelli Lemma. Before this let us briefly discuss the following: When $(A_n)_{n\geq 1}$ is a sequence of events, then the event $B_{n_0} := \cup_{n\geq n_0} A_n$ is the event that at least one of the A_n's for $n \geq n_0$ holds. The event $\cap_{n_0\geq 1} B_{n_0}$ is then the event that for each $n_0 \geq 1$, there exists at least one $n \geq n_0$ for which A_n holds, which means that A_n holds for infinitely many n's.*

LEMMA 1.1.13 (Borel-Cantelli, I). *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and that A_n is a sequence of events in $\mathcal{A}$ such that $\sum_{n\geq 1} P(A_n) < \infty$. Then, almost surely, only a finite number of A_n's do occur simultaneously.*

PROOF. Saying that a finite number of the A_n's occur simultaneously means that there exists (a possibly random) finite n_0 such that $B_{n_0} := \cup_{n\geq n_0} A_n$ does not hold. In other words, the goal is to show that

$$P(\cap_{n_0\geq 1} B_{n_0}) = 0.$$

We can also note that $n_0 \mapsto B_{n_0}$ is non-increasing, so that

$$P(\cap_{n_0\geq 1} B_{n_0}) = \lim_{n_0\to\infty} P(B_{n_0}).$$

On the other hand, for each given $n_0 \geq 1$,

$$P(B_{n_0}) = \lim_{p\to\infty} P(\cup_{j=0}^{p} A_{n_0+j}) \leq \lim_{p\to\infty} \sum_{j=0}^{p} P(A_{n_0+j}) = \sum_{m\geq n_0} P(A_m).$$

But since $\sum_{m\geq 1} P(A_m)$ is finite, the tail $\sum_{m\geq n_0} P(A_m)$ tends to 0 as $n_0 \to \infty$, so that the desired statement holds. □

We now turn to one of the slightly tricky things with this measure-theoretical setup: There exists always one σ-field on a set Ω, namely the collection $\mathcal{P}(\Omega)$ of all the subsets of Ω. However, this is often (typically, as soon as Ω is not countable) too big a collection to be then able to define interesting measures on it.

It can appear to be quite a challenge to construct very explicitely the σ-fields that one wants to actually work with. But there is a nice trick that allows to circumvent the difficulties, which is to define the σ-fields that are generated by a subset $\mathcal{U}$ of $\mathcal{P}(\Omega)$. One can start by making the following observation:

LEMMA 1.1.14. *The intersection of any (even uncountable) collection of σ-fields on Ω is still a σ-field.*

PROOF. To see this, one considers any given collection $(\mathcal{A}_i)_{i\in I}$ of σ-fields on Ω. Since Ω is in each of the $\mathcal{A}_i$, it is also in $\cap_i \mathcal{A}_i$ so that the first condition in the definition of σ-fields holds. For the second condition, we take $A \in \cap_i \mathcal{A}_i$, so that for each $i \in I$, $A \in \mathcal{A}_i$ – but since $\mathcal{A}_i$ is a σ-field, then $\Omega \setminus A$ is also in $\mathcal{A}_i$; since this is true for all $i \in I$, we get that $\Omega \setminus A \in \cap_i \mathcal{A}_i$. For the third condition, one uses again the same argument. □

Let us stress that the collection I of σ-fields that one takes the intersection of in the previous lemma does not have to be countable. This allows for any $\mathcal{U} \subset \mathcal{P}(\Omega)$ to do the following:

DEFINITION 1.1.15. *We* define **the σ-field $\sigma(\mathcal{U})$ generated by $\mathcal{U}$** *as the intersection of all σ-fields (on Ω) that contain $\mathcal{U}$.*

This definition makes sense: Indeed, the intersection exists because the set of σ-fields that contain $\mathcal{U}$ is not empty (it contains $\mathcal{P}(\Omega)$), and it is a σ-field (because it is the intersection of σ-fields) and it is the "smallest" σ-field containing $\mathcal{U}$ in the sense that it is a subset of any σ-field that contains $\mathcal{U}$.

An important example is when one has a topology on Ω, which means that one has a family $\mathcal{T}$ of open sets, then one can define the **Borel σ-field** associated to this topology as being $\mathcal{B} := \sigma(\mathcal{T})$. This σ-field is always well-defined. In the case of $[0, 1]$, $\mathbb{R}$ or $\mathbb{R}^d$ for $d \geq 1$ – when one starts with the (usual) collection of open sets for the Euclidean distance, this defines the Borel σ-fields $\mathcal{B}_{[0,1]}$ etc. on those sets that will play an important role in these lectures.

The definition $\sigma(\mathcal{U})$ is perfectly sound mathematically, but it is one of these perplexing cases related to the axiom of choice, because the definition does not necessarily provide a concrete recipy to decide if a given set A is in the σ-field or not. For instance, it is possible (using the axiom of choice) to construct subsets A of $[0,1]$ for which it is impossible to decide whether they are in $\mathcal{B}_{[0,1]}$ or not. A consequence of this, is that it can look very difficult to *directly* define μ as an explicit function on $\mathcal{B}$ (because we do not even know what an element of $\mathcal{B}$ looks like!).

However, there is a way around this, that (in the end) allows us not to worry about these issues. There are two type of results in that direction that will be very useful in these lectures:

- The first type of results uses the concept of π-systems:

  DEFINITION 1.1.16. *We say that $\mathcal{G} \subset \mathcal{P}(\Omega)$ is a* **π-system** *if it contains Ω and is stable under (finite) intersections, i.e., that for any A and B in $\mathcal{G}$, one has $A \cap B \in \mathcal{G}$.*

  REMARK 1.1.17. *Our definition here differs slightly from the usual one, where one does not ask for Ω to belong to $\mathcal{G}$.*

  The rule of thumb will be that **it turns out to be convenient to look at σ-fields generated by π-systems**. A first instance of this is the following lemma:

  LEMMA 1.1.18. *[Characterization on a π-system] Consider two finite measures defined on a σ-field $\mathcal{A} \subset \mathcal{P}(\Omega)$. Suppose that the two measures agree on a π-system $\mathcal{G}$ that generates $\mathcal{A}$ (i.e., $\sigma(\mathcal{G}) = \mathcal{A}$). Then $\mu_1 = \mu_2$ on $\mathcal{A}$.*

  Let us provide a brief sketch of the proof (the detailed proof is discussed the next section). The idea is rather simple: We define

  $$\mathcal{F} := \{A \in \mathcal{A} : \ \mu_1(A) = \mu_2(A)\}.$$

  The main step is to show that $\mathcal{F}$ is a σ-field. Once this is proven, then since it anyway contains $\mathcal{G}$, we deduce that it contains $\sigma(\mathcal{G})$ as well (by definition of $\sigma(\mathcal{G})$), from which it follows that $\mu_1 = \mu_2$ on $\mathcal{A}$. The fact that $\mathcal{F}$ is a σ-field will be a consequence of Dynkin's Lemma, that we separately state and prove in the next section.

  This lemma is very useful. For instance, it is quite easy to check that the sets $\{[0,a], a \in [0,1]\}$, $\{\mathbb{R}\} \cup \{(-\infty, a] : a \in \mathbb{R}\}$ and $\{\mathbb{R}^d\} \cup \{(-\infty, a_1] \times \cdots \times (-\infty, a_d] : a_1, \ldots, a_d \in \mathbb{R}\}$ are π-systems that do generate $\mathcal{B}_{[0,1]}$, $\mathcal{B}_{\mathbb{R}}$ and $\mathcal{B}_{\mathbb{R}^d}$ respectively (see the exercises). Hence, the lemma implies that:

  COROLLARY 1.1.19.
  - *Any two finite measures on $\mathcal{B}_{\mathbb{R}}$ such that $\mu_1((-\infty, a]) = \mu_2((-\infty, a])$ for all $a \in \mathbb{R}$ are necessarily equal.*
  - *Any two finite measures on $\mathcal{B}_{[0,1]}$ such that $\mu_1([0,a]) = \mu_2([0,a]$ for all $a \in [0,1]$ are necessarily equal.*
  - *Any two finite measures on $\mathcal{B}_{\mathbb{R}^d}$ such that $\mu_1(\prod_{j=1}^d (-\infty, a_j]) = \mu_2(\prod_{j=1}^d (-\infty, a_j])$ for all $a_1, \ldots, a_d$ in $\mathbb{R}$ are necessarily equal.*

  Another way to reformulate these statements is to use phrases like "there exists at most one measure such that...". For instance:

  COROLLARY 1.1.20 (Uniqueness of the Lebesgue measure). *There exists at most one measure on $\mathcal{B}_{[0,1]}$ such that for all $a \in [0,1]$, $\mu([0,a]) = a$.*

- The second class of results deals with existence of measures. The main one there is the following:

  THEOREM 1.1.21 (Existence of the Lebesgue measure). *There actually exists a measure on $\mathcal{B}_{[0,1]}$ such that for all $a \in [0,1]$, $\mu([0,a]) = a$.*

As we will see later, this will allow us to construct all the probability spaces that we will want to work with in these lectures.

We will not derive this existence theorem in these lectures, and refer to the measure theory courses for this. We will however now state and prove Dynkin's lemma, as the ideas will be also crucial when we discuss independence – this is the topic of the next section.

1.2. Dynkin systems and Dynkin's Lemma

DEFINITION 1.2.1. *A family $\mathcal{D}$ of subsets of Ω is called a* ***Dynkin system*** *(or a λ-system) if it satisfies the following three properties: (i) $\Omega \in \mathcal{D}$, (ii) If $A \in \mathcal{D}$ then $\Omega \setminus A \in \mathcal{D}$ and (iii) For each sequence $(A_n)_{n\geq 1}$ of pairwise disjoint elements from $\mathcal{D}$, one has $(\cup_{n\geq 1} A_n) \in \mathcal{D}$.*

The only difference with the notion of σ-algebra lies in the difference between (iii) and (c). We can in particular notice that a σ-algebra is always a Dynkin system. It is not difficult to see that the converse is not always true (see the exercises). But one has the following very simple equivalence:

LEMMA 1.2.2. *A subset $\mathcal{D}$ of $\mathcal{P}(\Omega)$ is a σ-field if and only if it is both a Dynkin system and a π-system.*

PROOF. A σ-field is always a Dynkin system and a π-system. We therefore only need to check that when $\mathcal{D}$ is both a π-system and a Dynkin system, then it is also a σ-field – ie., that it satisfies the third condition (c). Let us suppose that $A_1, A_2, \ldots$ are in $\mathcal{D}$. Define for each j, $B_j = A_j \setminus (A_1 \cup \ldots \cup A_{j-1})$. Clearly, we see that for all j, $A_1 \cup \ldots \cup A_j = B_1 \cup \ldots \cup B_j$ and that the B_j's are pairwise disjoint.

We now that $B_1 = A_1$ is in $\mathcal{D}$, and $B_2 = A_2 \setminus A_1 = A_2 \cap (\Omega \setminus A_1)$ is in $\mathcal{D}$ too, because $\mathcal{D}$ is a Dynkin system and a π-system. Let us assume that for some $j \geq 2$, $B_1, \ldots, B_j$ are in $\mathcal{D}$. Then, $C_j := B_1 \cup \ldots \cup B_j \in \mathcal{D}$ as a disjoint union of elements in $\mathcal{D}$, and $B_{j+1} = A_{j+1} \setminus C_j = A_{j+1} \cap (\Omega \setminus C_j)$ is in $\mathcal{D}$ as well. Hence, we can conclude by induction that $(B_j)_{j\geq 1}$ is a pairwise disjoint collection of events in $\mathcal{D}$ so that $\cup_j A_j = \cup_j B_j$ is also $\mathcal{D}$. □

Just as for σ-algebras, one can note that the intersection of any (even uncountable) family of Dynkin systems is always a Dynkin system, so that the following definition makes sense:

DEFINITION 1.2.3. *When $\mathcal{G} \subset \mathcal{P}(\Omega)$, one can define $\lambda(\mathcal{G})$ as the intersection of all Dynkin systems that contain $\mathcal{G}$, and call it* **the Dynkin system generated by $\mathcal{G}$.**

Since a σ-algebra is always a Dynkin system, one always has $\lambda(\mathcal{G}) \subset \sigma(\mathcal{G})$. We are now ready to state the following useful result:

LEMMA 1.2.4 (Dynkin's lemma). *When $\mathcal{C}$ a π-system on Ω, then $\lambda(\mathcal{C}) = \sigma(\mathcal{C})$.*

PROOF OF DYNKIN'S LEMMA. Let $\mathcal{C}$ be a π-system on Ω and let $\mathcal{D} = \lambda(\mathcal{C})$ be the Dynkin system generated by $\mathcal{C}$. We want to show that $\sigma(\mathcal{C}) \subset \mathcal{D}$, and for this, it is sufficient to see that $\mathcal{D}$ is a σ-algebra. Because of the previous lemma, we only have to check that it is stable under pairwise intersections.

To see this, we first define, for each given $B \in \mathcal{D}$, the collection

$$\mathcal{D}_B := \{A \in \mathcal{D} : \ A \cap B \in \mathcal{D}\}.$$

We can first note that $\mathcal{D}_B$ is a Dynkin system. Indeed: (i) The fact that $\Omega \in \mathcal{D}_B$ is clear, (ii) When $A \in \mathcal{D}_B$, then we know that $(\Omega \setminus A) \cap B$ is the complement of the disjoint union of $A \cap B$ and of $\Omega \setminus B$ – both these sets are in $\mathcal{D}$, so this disjoint union is in $\mathcal{D}$ as well. (iii) When the sets $A_1, A_2, \cdots \in \mathcal{D}_B$ are pairwise disjoint, then the sets $B \cap A_1, B \cap A_2, \ldots$ are also pairwise disjoint and in $\mathcal{D}$ (by definition of $\mathcal{D}_B$) so that $(\cup_j A_j) \cap B = \cup_j (A_j \cap B)$ is in $\mathcal{D}$ as well. In other words, $\cup_{j\geq 1} A_j \in \mathcal{D}_B$.

We are now going to use the fact that $\mathcal{D}_B$ is a Dynkin system first when B is in $\mathcal{C}$. Since $\mathcal{C}$ is a π-system, we see that $\mathcal{C} \subset \mathcal{D}_B$. Hence, we conclude $\lambda(\mathcal{C}) \subset \mathcal{D}_B$. Since $\mathcal{D}_B \subset \mathcal{D}$ and $\lambda(\mathcal{C}) = \mathcal{D}$, we conclude that $\mathcal{D}_B = \mathcal{D}$. In other words, for all $A \in \mathcal{D}$ and $B \in \mathcal{C}$, we get that $A \cap B \in \mathcal{D}$. We can invert the roles of A and B here: For all $A \in \mathcal{C}$ and $B \in \mathcal{D}$, one has $A \cap B \in \mathcal{C}$.

Let us now consider any B in $\mathcal{D}$. What we have just stated shows that $\mathcal{C} \subset \mathcal{D}_B$, and we know that $\mathcal{D}_B$ is a Dynkin system. Hence, we conclude that $\lambda(\mathcal{C}) \subset \mathcal{D}_B$. In other words, since $\lambda(\mathcal{C}) = \mathcal{D}$, we conclude that for all A, B in $\mathcal{D}$, $A \cap B$ is in $\mathcal{D}$ as well. □

We are now ready to provide the actual proof of the uniqueness lemma:

Proof of Lemma 1.1.18. We first want to check that

$$\mathcal{F} := \{A \in \mathcal{A} : \ \mu_1(A) = \mu_2(A)\}$$

is a Dynkin system. Since $\Omega \in \mathcal{G}$, one has $\mu_1(\Omega) = \mu_2(\Omega)$ and $\Omega \in \mathcal{F}$, so that (i) holds. When $A \in \mathcal{F}$, then since A and $\Omega \setminus A$ are disjoint,

$$\mu_1(\Omega \setminus A) = \mu_1(\Omega) - \mu_1(A) = \mu_2(\Omega) - \mu_2(A) = \mu_2(\Omega \setminus A)$$

so that (ii) holds. Finally, when $(A_j)_{j\geq 1)}$ is a collection of pairwise disjoint sets in $\mathcal{F}$, the properties of measures show that

$$\mu_1(\cup_{j\geq 1} A_j) = \lim_{n\to\infty} (\mu_1(A_1) + \cdots + \mu_1(A_n)) = \lim_{n\to\infty} (\mu_2(A_1) + \cdots + \mu_2(A_n)) = \mu_2(\cup_{j\geq 1} A_j)$$

from which condition (iii) holds.

Now, the assumption in the lemma is that $\mathcal{F}$ contains the π-system $\mathcal{G}$. By Dynkin's lemma we therefore conclude that it also contains $\mathcal{A} = \sigma(\mathcal{G})$, from which it follows that $\mu_1 = \mu_2$ on $\mathcal{A}$. $\square$

1.3. Independence

There is one absolutely crucial new ingredient in probability theory compared to measure theory, namely the concept of *independence*. In some sense, one can argue that most of this course will be exploring various aspects of independence.

We are going to describe two notions of independence – between events and between σ-fields, the latter being in fact a generalization of the former. We start with independence between events:

DEFINITION 1.3.1 (Independence between two events). *We say that A and B in $\mathcal{A}$ are independent if $P(A \cap B) = P(A)P(B)$.*

REMARK 1.3.2. *This notion is very different from the intuition of measure theory where $\mu(A)$ represents the "mass" of A – as here one would say that multiplying two masses gives a mass!*

Just as in the discrete setting, this definition heuristically means that when $P(A) > 0$, then observing that A holds does not influence the likelihood that B holds as well (we will come back to this when we will discuss conditional probabilities).

Note also that if A and B are independent, then $\Omega \setminus A$ and B are also independent, since

$$\begin{aligned} P((\Omega \setminus A) \cap B) &= P(B \setminus (A \cap B)) \\ &= P(B) - P(A \cap B) = P(B) - P(A)P(B) = P(B)(1 - P(A)) = P(B)P(\Omega \setminus B). \end{aligned}$$

Also, Ω is clearly independent from any other event, and $\emptyset$ is independent from any other event.

We now immediately move to the independence between two σ-fields:

DEFINITION 1.3.3 (Independence between two σ-fields). *Two sub-σ-fields $\mathcal{A}_1$ and $\mathcal{A}_2$ of $\mathcal{A}$ are independent if for any $A_1 \in \mathcal{A}_1$ and any $A_2 \in \mathcal{A}_2$, $P(A_1 \cap A_2) = P(A_1)P(A_2)$.*

In other words, whatever event the observer #1 is able to detect is actually independent from whatever event the observer #2 is able to detect.

We can note that $\sigma(\{A\}) = \{\emptyset, A, \Omega \setminus A, \Omega\}$, so that it follows from the above remarks that A and B are independent if and only if the σ-fields generated by $\{A\}$ and $\{B\}$ are independent.

REMARK 1.3.4. *In some textbooks, one defines also independence between families of events (that are not necessarily σ-fields), but this can be a somewhat confusing notion that we will avoid to use here.*

When one wants to check the independence between two σ-fields, one can face a similar problem as in the construction of the measures, namely that one does not know what an event in $\mathcal{A}$ looks like. However, one can circumvent this using the following very useful result:

PROPOSITION 1.3.5 (Checking independence via π-systems). *Consider a probability space $(\Omega, \mathcal{A}, P)$ and two π-systems $\mathcal{G}_1$ and $\mathcal{G}_2$ in $\mathcal{A}$. If for all $A_1 \in \mathcal{G}_1$ and all $A_2 \in \mathcal{G}_2$, the events A_1 and A_2 are independent, then the two σ-fields $\mathcal{A}_1 := \sigma(\mathcal{G}_1)$ and $\mathcal{A}_2 := \sigma(\mathcal{G}_2)$ are independent.*

PROOF. Define

$$\mathcal{U}_2 := \{B_2 \in \mathcal{A}_2 : \ \forall A_1 \in \mathcal{G}_1, P(A_1 \cap B_2) = P(A_1)P(B_2)\}.$$

This is the set of events in $\mathcal{A}_2$ that are independent with any event in the π-system $\mathcal{G}_1$. The assumption in the proposition shows that $\mathcal{G}_2 \subset \mathcal{U}_2$. It is also a very simple exercise to check that $\mathcal{U}_2$ is a Dynkin system – which then readily implies that it contains $\lambda(\mathcal{G}_2)$. But by Dynkin's lemma, $\lambda(\mathcal{G}_2) = \sigma(\mathcal{G}_2) = \mathcal{A}_2$. In other words, any event A_2 is $\mathcal{A}_2$ is independent of any event A_1 in $\mathcal{G}_1$.

Then, we define

$$\mathcal{V}_1 := \{B_1 \in \mathcal{A}_1 : \ \forall B_2 \in \mathcal{A}_2, P(B_1 \cap B_2) = P(B_1)P(B_2)\}.$$

This is the set of events in $\mathcal{A}_1$ that are independent of any event in $\mathcal{A}_2$. We have just seen that this set contains $\mathcal{G}_1$. Again, one can easily check that $\mathcal{V}_1$ is a Dynkin system – which then implies that it is actually equal to $\mathcal{A}_1$ – which proves the proposition. □

We now turn to independence between more than two σ-fields:

DEFINITION 1.3.6 (Independence between a finite collection of σ-fields). *Suppose that $\mathcal{A}_j$ for $1 \leq j \leq n$ is a finite sequence of sub-σ-fields of $\mathcal{A}$. We say that they are independent if for any $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$,*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n).$$

DEFINITION 1.3.7 (Independence between a countable collection of σ-fields). *Suppose that $\mathcal{A}_j$ for $j \geq 1$ are a sequence of sub-σ-fields of $\mathcal{A}$. We say that they are independent if for any $n \geq 1$ and any $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$,*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)\ldots P(A_n).$$

In other words, this is a sequence of independent σ-fields if for any $n \geq 1$, the n σ-fields $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent.

REMARK 1.3.8. *Since $\Omega \in \mathcal{A}_j$ for all j, the independence between a collection of σ-fields, does not depend on the order in which they are enumerated (see exercise sheet).*

REMARK 1.3.9. *This condition (when $n \geq 3$) is stronger than the notion of* pairwise independence.

Again, to check independence between n σ-fields, it is sufficient to check things for generating π-systems:

PROPOSITION 1.3.10 (Checking independence via π-systems, II). *Consider a probability space $(\Omega, \mathcal{A}, P)$ and a finite collection of π-systems $\mathcal{G}_1, \ldots, \mathcal{G}_n$. If for all $A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n$, one has $P(A_1 \cap \ldots \cap A_n) = P(A_1) \times \cdots \times P(A_n)$, then the σ-fields $\mathcal{A}_1 := \sigma(\mathcal{G}_1), \ldots, \mathcal{A}_n := \sigma(\mathcal{G}_n)$ are independent.*

The proof is a slight inductive variation of the proof of the result for two σ-fields.

PROOF. For each $j \in 1, \ldots, n$, define

$$\begin{aligned} \mathcal{U}_j \quad := \quad & \{A_j \in \mathcal{A}_j : \forall A_1 \in \mathcal{G}_1, \ldots, A_{j-1} \in \mathcal{G}_{j-1}, A_{j+1} \in \mathcal{A}_{j+1}, \ldots, A_n \in \mathcal{A}_n, \\ & P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdots P(A_n)\}. \end{aligned}$$

It is again immediate to check that these collections are Dynkin systems.

The assumption in the proposition shows that $\mathcal{G}_n \subset \mathcal{U}_n$. Using Dynkin's lemma just as in the previous proof, we then get that $\mathcal{U}_n = \mathcal{A}_n$.

Then, look at $\mathcal{U}_{n-1}$. Since $\mathcal{U}_n = \mathcal{A}_n$, we see that $\mathcal{G}_{n-1} \subset \mathcal{U}_{n-1}$, and by Dynkin's lemma, we get that $\mathcal{U}_{n-1} = \mathcal{A}_{n-1}$.

Iteratively, we see that for any $j \in \{1, \ldots, n-1\}$, if $\mathcal{U}_{j+1} = \mathcal{A}_{j+1}$, then necessarily $\mathcal{G}_j \subset \mathcal{U}_j$, so that by Dynkin's lemma, $\mathcal{U}_j = \mathcal{A}_j$. So, we get that $\mathcal{U}_j = \mathcal{A}_j$ for $j = n, n-1, \ldots, 1$.

In particular $\mathcal{U}_1 = \mathcal{A}_1$, which is exactly means that for all $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$, one has

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdots P(A_n)\}.$$

□

Let us conclude this section on independence by the second Borel-Cantelli lemma. Let us first discuss the independence between a finite collection $A_1, \ldots, A_n$ of *events*:

DEFINITION 1.3.11. *The events $(A_1, \dots, A_n)$ are independent if for any subset $\{j_1, \dots, j_m\}$ of $\{1, \dots, n\}$, one has*

$$P(A_{j_1} \cap \dots \cap A_{j_m}) = P(A_{j_1})P(A_{j_2}) \dots P(A_{j_m}).$$

One can check that this is equivalent to saying that the n σ-fields generated by the events A_1, A_2, …, A_n are independent (recall that the σ-field generated by A_1 is $\{\emptyset, A_1, \Omega \setminus A_1, \Omega\}$) – see the exercise sheet.

REMARK 1.3.12. *Let us stress that this is a stronger condition that saying that $P(A_1 \cap \dots \cap A_n) = P(A_1) \dots P(A_n)$. This latter condition is for instance always satisfied if one chooses $A_n = \emptyset$, even if A_1 and A_2 are not independent.*

Similarly, a sequence $(A_j)_{j \geq 1}$ of events is said to be a sequence of independent events if for any subset $\{j_1, \dots, j_m\}$ of $\mathbb{N}$, one has

$$P(A_{j_1} \cap \dots \cap A_{j_m}) = P(A_{j_1})P(A_{j_2}) \dots P(A_{j_m}).$$

We will lots of examples of such sequences of events in these lectures.

We are now ready to state and prove the second Borel-Cantelli lemma:

LEMMA 1.3.13. *(Borel-Cantelli, II] If $(A_n)_{n \geq 1}$ is a sequence of independent events such that $\sum_{n \geq 1} P[A_n) = \infty$, then almost surely, there exists infinitely many n's for which A_n occur.*

PROOF. To say that there exists infinitely many n's such that A_n occurs, means that for any n_0, $\cup_{n \geq n_0} A_n$ holds. So, the goal is to prove that

$$P(\cap_{n_0 \geq 1} \cup_{n \geq n_0} A_n) = 1$$

or equivalently that

$$P(\cup_{n_0 \geq 1} \cap_{n \geq n_0} (\Omega \setminus A_n)) = 0.$$

It therefore suffices to check that for each given n_0,

$$P(\cap_{n \geq n_0} (\Omega \setminus A_n)) = 0.$$

But for each $p \geq 1$, $A_{n_0}, \dots, A_{n_0+p}$ are independent, so that

$$P(\cap_{n \geq n_0} (\Omega \setminus A_n)) \leq P(\cap_{n \in [n_0, n_0+p]} (\Omega \setminus A_n)) = \prod_{j=0}^{p} P(\Omega \setminus A_{n_0+j}) = \prod_{j=0}^{p} (1 - P(A_{n_0+j})).$$

But $1 - x \leq \exp(-x)$ when $x \in [0, 1]$ so that this last quantity is bounded by $\exp(-\sum_{j=0}^{p} P(A_{n_0+j}))$, which tends to 0 as $p \to \infty$, so we can conclude that indeed

$$P(\cap_{n \geq n_0} (\Omega \setminus A_n)) = 0.$$

□

CHAPTER 2

Abstract generalities: Random variables

2.1. Measurable functions and generated σ-fields

Suppose that $\mathcal{A}$ and $\mathcal{A}'$ are σ-fields on the sets Ω and Ω' respectively.

DEFINITION 2.1.1. *A function f from Ω into Ω' is said to be measurable (from $(\Omega, \mathcal{A})$ into $(\Omega', \mathcal{A}')$) when for all $A' \in \mathcal{A}'$, the set $f^{-1}(A') := \{\omega \in \Omega,\ f(\omega) \in A'\}$ is in $\mathcal{A}$.*

REMARK 2.1.2. *It is important to specify with respect to which σ-fields one discusses measurability. Sometimes, if it is clear what σ-field one is talking about, one however omits this. For instance, when $\Omega' = \mathbb{R}$, it is usually implicit that $\mathcal{A}' = \mathcal{B}_{\mathbb{R}}$.*

In the probability theory framework, when $(\Omega, \mathcal{A})$ is endowed with a probability measure P, such a measurable function is called a *random variable* with values in Ω'. Often, when $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ one sometimes just says "a random variable", omitting the "with values in $\mathbb{R}$ and with respect to the Borel σ-algebra"). These measurable functions are often denoted by X, Y, ...

Intuitively, X being measurable simply means that $X(\omega)$ is "observable" (in the sense of the σ-field $\mathcal{A}$) – since for any given Borel set $A' \in \mathcal{A}'$, one can "observe" whether $X(\omega) \in A'$ holds or not.

One usually simply writes $\{X \in A'\}$ as a shorthand for $\{\omega \in \Omega,\ :\ X(\omega) \in A'\}$, and simply $P(X \in A')$ for the probability of this event when $A' \in \mathcal{A}'$.

Again, there is a more "checkable" version of this definition that avoids the issue of having to describe explicitely what the elements of $\mathcal{A}'$ look like:

LEMMA 2.1.3. *Let X be a function from Ω into Ω', and let $\mathcal{A}$ and $\mathcal{A}'$ denote σ-fields on Ω and Ω' respectively. If $\mathcal{U}'$ is a subset of $\mathcal{A}'$ such that $\sigma(\mathcal{U}') = \mathcal{A}'$ and if for all $A' \in \mathcal{U}'$, $X^{-1}(A') \in \mathcal{A}$, then X is measurable from $(\Omega, \mathcal{A})$ into $(\Omega', \mathcal{A}')$.*

This has for instance the following immediate corollary:

COROLLARY 2.1.4. *A function X from Ω into $\mathbb{R}$ is measurable from $(\Omega, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$) if (and only if) for all $a \in \mathbb{R}$, $\{X \leq a\} \in \mathcal{A}$.*

PROOF OF THE LEMMA. This has a similar flavor as the results with π-systems, but it is in fact simpler:

One first checks that the family $\mathcal{G}'$ of all subsets A' of Ω' such that $X^{-1}(A') \in \mathcal{A}$ is a σ-algebra: Indeed, $X^{-1}(\Omega') = \Omega$ so that $\Omega' \in \mathcal{G}'$; for all $A' \subset \Omega'$, $X^{-1}(\Omega' \setminus A') = \Omega \setminus X^{-1}(A')$, so that if $A' \in \mathcal{G}'$, then $X^{-1}(\Omega' \setminus A') \in \mathcal{A}$ and therefore $\Omega' \setminus A' \in \mathcal{G}'$; and finally we note that $X^{-1}(\cup_{i\geq 1} A'_i) = \cup_{i\geq 1} X^{-1}(A'_i)$ for any sequence A'_i, from which the third property follows – if the sequence A'_i is in $\mathcal{G}'$, then $X^{-1}(\cup_{i\geq 1} A'_i)$ is in $\mathcal{A}$ as the countable union of elements in $\mathcal{A}$, so that $\cup_i A'_i$ is in $\mathcal{G}'$.

But if this σ-algebra contains $\mathcal{U}'$, then it contains the $\sigma(\mathcal{U}')$ which is $\mathcal{A}'$. Hence, for all $A' \in \mathcal{A}'$, the event $X \in A'$ is in $\mathcal{A}$ – which exactly says that X is measurable. □

An important concept (for instance in order to discuss independence) associated to a random variable is the following:

DEFINITION 2.1.5. *The σ-field $\sigma(X)$ generated by a measurable function X from $(\Omega, \mathcal{A})$ into $(\Omega', \mathcal{A}')$ is the sub-σ-field of $\mathcal{A}$ defined by*

$$\sigma(X) := \{X^{-1}(A') : \ A' \in \mathcal{A}'\}.$$

Heuristically, this is the collection of events that one can check when observing the value of the random function X. It is easy to check that it is a σ-field, so that this definition makes sense: One first notices that $\Omega = X^{-1}(\Omega')$ is in $\sigma(X)$, and that $\Omega \setminus X^{-1}(A') = X^{-1}(\Omega \setminus A')$ for all A', and applying this to $A' \in \mathcal{A}'$ shows that the complement of an event in $\sigma(X)$ is in $\sigma(X)$ as well. Finally, if one considers a sequence $(A'_i)_{i\geq 1}$ in $\mathcal{A}'$, then

$$\cup_{i\geq 1} X^{-1}(A'_i) = X^{-1}(\cup_{i\geq 1} A'_i) \in \sigma(X) -$$

The following fact will turn out to be very useful when we will discuss independence:

LEMMA 2.1.6. *Suppose that X is measurable from $(\Omega, \mathcal{A})$ into $(\Omega', \mathcal{A}')$, and suppose that π' is a generating π-system of $\mathcal{A}'$. Then the collection $X^{-1}(\pi')$ of all sets $X^{-1}(A')$ where $A' \in \pi'$ forms a generating π-system of $\sigma(X)$.*

This has for instance the following useful immediate consequence:

COROLLARY 2.1.7. *When X is a real-valued random variable, then the collection*

$$\{X^{-1}((-\infty, a]) : a \in \mathbb{R}\} \cup \{\Omega\}$$

is a π-system that generates the σ-field $\sigma(X)$.

PROOF OF THE LEMMA. First we notice that the collection of events $X^{-1}(A')$ where $A' \in \pi'$ is a π-system (because $X^{-1}(\Omega') = \Omega$ and $X^{-1}(A') \cap X^{-1}(B') = X^{-1}(A' \cap B')$).

By assumption, this π-system is contained in $\sigma(X)$. Hence, the σ-field $\mathcal{G}$ generated by this π-system is contained in $\sigma(X)$ as well.

We now define $\mathcal{B}'$ to be the collection of events A' in $\mathcal{A}'$ for which $X^{-1}(A') \in \mathcal{G}$. It is again easy to check that $\mathcal{B}'$ is a σ-field, and by construction, it contains π'. It therefore contains $\sigma(\pi') = \mathcal{A}'$. Hence $\mathcal{B}' = \mathcal{A}'$. So, indeed, for all $A' \in \mathcal{A}'$, $X^{-1}(A') \in \mathcal{G}$ which shows that $\sigma(X)$ is contained (and therefore equal to) $\mathcal{G}$. □

We now turn our attention to the case where several random variables are defined on the same probability space.

Suppose now that we are given a probability space $(\Omega, \mathcal{A}, P)$ and a finite family of random variables $(X_1, \ldots, X_n)$ on this space (this means that each X_i looked at separately is a random variable). We can also view $X = (X_1, \ldots, X_n)$ as a function from Ω into $\mathbb{R}^n$. Let us now briefly show that it is measurable from $(\Omega, \mathcal{A})$ into $\mathbb{R}^n$ endowed with the Borel σ-algebra on $\mathbb{R}^n$:

LEMMA 2.1.8. *Suppose that $X_1, \ldots, X_n$ are n functions from Ω into $\mathbb{R}$. The vector $X = (X_1, \ldots, X_n)$ is a measurable function from $(\Omega, \mathcal{A})$ into $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ if and only if each X_i is measurable from $(\Omega, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.*

PROOF. Suppose first that $X = (X_1, \ldots, X_n)$ is a measurable into $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$. We can note that for each $A'_1 \in \mathcal{B}_{\mathbb{R}}$, the set $A' := A'_1 \times \mathbb{R} \times \cdots \times \mathbb{R}$ is in $\mathcal{B}_{\mathbb{R}^n}$, so that $X^{-1}(A') \in \mathcal{A}$. But for this A', $X^{-1}(A') = X_1^{-1}(A'_1)$ – so that X_1 is measurable. Similarly, each X_i is a measurable function from $(\Omega, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Conversely, let us assume that $X_1, \ldots, X_n$ are n (real-valued) random variables. To check that X is a measurable function from $(\Omega, \mathcal{A})$ into $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, it is sufficient to check that for all A' belonging to a generating π-system of $\mathcal{B}_{\mathbb{R}^n}$, $X^{-1}(A') \in \mathcal{A}$. We can choose π' to be the collection of

sets $A'_1 \times \cdots \times A'_n$ for $A'_1, \ldots, A'_n \in \mathcal{B}_{\mathbb{R}}$, which is clearly a generating π-system of $\mathcal{B}_{\mathbb{R}^n}$ – and we can note that for any such $A'_1, \ldots, A'_n$,

$$X^{-1}(A'_1 \times \cdots \times A'_n) = X_1^{-1}(A'_1) \cap \ldots \cap X_n^{-1}(A'_n) \in \mathcal{A},$$

which concludes the proof. □

2.2. Independent random variables

DEFINITION 2.2.1. *Suppose that $X_1, \dots, X_n$ are n random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$ (with values in some measurable spaces). These n random variables are called independent if the corresponding σ-fields $\sigma(X_1), \dots, \sigma(X_n)$ are independent.*

Our previous considerations have the following consequence:

LEMMA 2.2.2. *The n real-valued random variables $(X_1, \dots, X_n)$ are independent if and only if for any $a_1, \dots, a_n$ in $\mathbb{R}$,*

$$P(X_1 \leq a_1, \dots, X_n \leq a_n) = P(X_1 \leq a_1) \times \cdots \times P(X_n \leq a_n).$$

REMARK 2.2.3. *We did not ask here for relations involving also a subset of the collection of n random variables as in the definition of independence between σ-fields. But we can note that when $1 \leq j_1 < j_2 < \cdots < j_k \leq n$, by letting the a_i for $i \notin \{j_1, \dots, j_j\}$ tend to infinity in the previous identity, one actually gets*

$$P(X_{j_1} \leq a_{j_1}, \dots, X_{j_k} \leq a_{j_k}) = P(X_{j_1} \leq a_{j_1}) \times \cdots \times P(X_{j_k} \leq a_{j_k}).$$

PROOF OF THE LEMMA. Suppose first that $(X_1, \dots, X_n)$ are independent, and choose $a_1, \dots, a_n$ in $\mathbb{R}$. Then the events $A_1 = \{X_1 \leq a_1\}, \dots, A_n = \{X_n \leq a_n\}$ are in $\sigma(X_1), \dots, \sigma(X_n)$ respectively, so that

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \ldots P(A_n),$$

which is exactly the desired expression for $P(X_1 \leq a_1, \dots, X_n \leq a_n)$.

Conversely, if we know that for all $a_1, \dots, a_n$ in $\mathbb{R}$, one has

$$P(X_1 \leq a_1, \dots, X_n \leq a_n) = P(X_1 \leq a_1) \times \cdots \times P(X_n \leq a_n)$$

then we have noticed in the above remark that this holds also for $a_1, \dots, a_n \in \mathbb{R} \cup \{+\infty\}$. Using the characterization of independence via π-systems and the fact that the set of events $\{X_j \leq a_j\}$ for $a_j \in \mathbb{R} \cup \{+\infty\}$ is a π-system that generates $\sigma(X_j)$, we get that $\sigma(X_1), \dots, \sigma(X_n)$ are independent. □

We now briefly mention sequences of random variables defined on the same probability space.

DEFINITION 2.2.4. *We say that $(X_n)_{n \geq 1}$ is a sequence of independent random variables if for all $n \geq 1$, the n random variables $(X_1, \dots . X_n)$ are independent.*

In view of the previous lemma, we therefore see that $(X_n)_{n \geq 1}$ is a sequence of real-valued independent random variables if and only if for any $n \geq 1$ and any $a_1, \dots, a_n$ in $\mathbb{R}$,

$$P(X_1 \leq a_1, \dots, X_n \leq a_n) = P(X_1 \leq a_1) \times \cdots \times P(X_n \leq a_n).$$

Such sequences of independent real-valued random variables will be important later in these lectures.

2.3. Laws of (collections of) random variables

To each (real-valued) random variable X on $(\Omega, \mathcal{A}, P)$, one can associate a probability measure μ_X on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, called the **distribution of** X, which is defined by $\mu_X(B) = P(X \in B)$. This is indeed a probability measure, since: $\mu_X(\mathbb{R}) = P(\Omega) = 1$, and for any sequence $(B_i)_{i\geq 1}$ of pair-wise disjoint Borel sets in $\mathcal{B}_\mathbb{R}$, the events $(\{X \in B_i\})_{i\geq 1}$ are clearly pairwise disjoint as well, so that

$$\mu_X\Big(\bigcup_{i\geq 1} B_i\Big) = P\Big(\bigcup_{i\geq 1} \{X \in B_i\}\Big) = \sum_{i\geq 1} P(\{X \in B_i\}) = \sum_{i\geq 1} \mu_X(B_i)\,.$$

The **distribution function** $F(\cdot)$ of a random variable X on $(\Omega, \mathcal{A}, P)$ is the function $F : \mathbb{R} \to [0, 1]$ defined by $F(x) = \mu_X((-\infty, x]) = P(X \leq x)$.

A distribution function always has the following three properties: (i) it is non-decreasing, (ii) it tends to 0 and 1 at $-\infty$ and $+\infty$ respectively, (iii) it is right-continuous.

Indeed: (i) is clear because $(-\infty, a] \subset (-\infty, a']$ as soon as $a \leq a'$. (ii) follows from the fact that $\cap_{n\geq 1}(-\infty, -n] = \emptyset$ and $\cup_{n\geq 1}(-\infty, n] = \mathbb{R}$ (and the properties of measures), and similarly, when $a_n \downarrow a$, then $\cap_{n\geq 1}(-\infty, a_n] = (-\infty, a]$ from which the right-continuity follows.

In fact, it is easy to check that the converse is true:

Proposition 2.3.1 (Lebesgue-Stieltjes). *Any function $F : \mathbb{R} \to [0, 1]$ satisfying these three properties is the distribution function of exactly one probability measure on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$.*

Proof. Let us first prove existence. We start with the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = (0, 1)$, $\mathcal{A} = \mathcal{B}(0, 1)$, and P is the Lebesgue measure on $(0, 1)$. Then, we define the map Y from $(0, 1)$ into $\mathbb{R}$ by

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

We can first note that for all $x \in \mathbb{R}$

$$\{\omega :\ Y(\omega) \leq x\} = \{\omega :\ \omega \leq F(x)\}.$$

This in particular ensures that Y is measurable and that the distribution function of μ_Y is exactly F. So, F is indeed the distribution function of some probability measure μ_Y.

Uniqueness follows from the π-system uniqueness property: If F is the distribution function of two probability measures μ_1 and μ_2, then $\mu_1((-\infty, a]) = \mu_2((-\infty, a])$ for all $a \in \mathbb{R}$, from which it follows that $\mu_1 = \mu_2$ on $\mathcal{B}_\mathbb{R}$. □

In the very same way as for single random variables, we can then define the *law* of $X = (X_1, \ldots, X_n)$ to be the probability measure P_X defined on $\mathcal{B}_{\mathbb{R}^n}$ by $P_X(B) = P(X \in B)$. Just as for single random variables, one can check directly that this indeed defines a probability measure. And, then, using the characterizations of probability measures on $\mathbb{R}^n$ via π-systems, we see that the law of X is fully determined by the knowledge of the map

$$(a_1, \ldots, a_n) \mapsto P(X_1 \leq a_1, \ldots, X_n \leq a_n).$$

In the next chapters, we will often consider sequences $(X_n)_{n\geq 1}$ of random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$. It would be possible to view this as a random sequence in $\mathbb{R}^\mathbb{N}$ (measurable with respect to the appropriate σ-field) but we will actually not need to do this here. Instead, we can simply choose to *define* the law of this sequence X as the knowledge for each given n, of the law of the random vector $(X_1, \ldots, X_n)$. (This definition turns out to be equivalent to the other definitions). In particular, the law of a sequence $(X_n)_{n\geq 1}$ is uniquely determined by the knowledge, for each $n \geq 1$ and each $a_1, \ldots, a_n$, of $P(X_1 \leq a_1, \ldots, X_n \leq a_n)$.

2.4. Functions and limits of random variables

The core message of this section is that *any reasonable function of a finite or countable collection of random variables is still a random variable.* We will go through some examples of this general statement:

(1) Suppose first that $X_1, \dots, X_n$ are real-valued random variables defined on the same probability space. Suppose that f is a measurable function from $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ into $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Then, $Y := f(X_1, \dots, X_n)$ is a random variable too. Indeed, for any $B \in \mathcal{B}_{\mathbb{R}}$, $f^{-1}(B) \in \mathcal{B}_{\mathbb{R}^n}$ and therefore
$$Y^{-1}(B) = (X_1, \dots, X_n)^{-1}(f^{-1}(B)) \in \mathcal{A}.$$

(2) Let us now consider a sequence $(Y_n)_{n\geq 1}$ of random variables. Then the event C that Y_n converges as $n \to \infty$ is measurable (i.e., it is a proper "event"). Indeed, converging in $\mathbb{R}$ is the same as being Cauchy, and therefore
$$C = \{\omega \in \Omega : \forall k \geq 1,\ \exists n_0 \geq 1, \forall n, n' \geq n_0, |Y_n(\omega) - Y'_n(\omega)| \leq 1/k\}.$$
In other words,
$$C = \cap_{k_1} \cup_{n_0} \cap_{n,n' \geq n_0} \{Y_n - Y_{n'} \in [-1/k, 1/k]\}$$
which shows that it is indeed in $\mathcal{A}$ – note that
$$\{Y_n - Y_{n'} \in [-1/k, 1/k]\} = (Y_n - Y_{n'})^{-1}([-1/k, 1/k])$$
is in $\mathcal{A}$ because $Y_n - Y_{n'}$ is measurable.

(3) In the setup of the previous example, when C holds, we can define $Y(\omega)$ to be equal to the limit of the Y_n, and when C does not hold, we can just set $Y = 0$. This function Y is then also a random variable. Indeed, for any $a \in \mathbb{R}$,
$$\{\omega\ :\ Y(\omega) \leq a\} \cap C = C \cap (\cap_{k\geq 1} \cup_{n_0} \cap_{n \geq n_0} \{Y_n \leq a + 1/k\})$$
and
$$\{\omega\ :\ Y(\omega) \leq a\} \cap (\Omega \setminus C) = \emptyset \text{ or } (\Omega \setminus C)$$
depending upon the sign of a.

(4) One can also define the notion of random variables with values in $\mathbb{R} \cup \{-\infty, \infty\}$ – the simplest way is to say that $Y : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$ is such a random variable if $\arctan Y$ is a random variable, where we define $\arctan(+\infty) = \pi/2$ and $\arctan(-\infty) = -\pi/2$. This for instance makes it possible to show results of the following type:

LEMMA 2.4.1. *When $(X_n)_{n\geq 1}$ is a sequence of random variables (defined on a same probability space), then $\limsup_{n\to\infty} X_n$ is a random variable with values in $\mathbb{R} \cup \{-\infty, +\infty\}$.*

PROOF. The definition of $\overline{X} := \limsup_{n\to\infty} X_n$ is that it is the non-increasing limit as $n \to \infty$ of $\sup_{n'\geq n} X_{n'}$. But for each n, $Y_n := \sup_{n'\geq n} X_{n'}$ is the non-decreasing limit of $\max(X_{n'}, \dots, X_{n'+l})$ as $l \to \infty$. The quantity $\arctan(Y_n)$ is then a random variable as limit of random variables, and then $\arctan(\overline{X}) = \lim_{n\to\infty} \arctan(Y_n)$ is a random variable as limit of random variables. □

There are many variations of such results.

2.5. Expectation, variance

One main purpose of measure theory is to define integrals of measurable functions with respect to measures. Let us briefly recall (without full proofs) the basic ideas and results there:

Suppose that $(\Omega, \mathcal{F}, \mu)$ is a measured space, where μ is a finite measure.

- A function of the type $e(\cdot) = \sum_{j=1}^{k} a_j 1_{A_j}(\cdot)$ defined on Ω where $a_j \in \mathbb{R}$ and $A_j \in \mathcal{F}$ is called an elementary function. It is clearly measurable, and we then *define* the integral $\tilde{I}_\mu(e)$ to be $\sum_{j=1}^{k} a_j \mu(A_j)$. [It is a simple exercise to check that this sum does not depend on the choice of $(a_j, A_j)_{j \leq k}$ chosen to represent e. Hint: one can for instance first consider all possible intersections of the sets A_j and then write $e = \sum_{i \leq k'} a'_i 1_{A'_i}$ where the A'_i are disjoint and the a'_i are sums of some of the a_j].
- When f is a measurable *non-negative* function from Ω into $\mathbb{R}$, then, one can *define* the integral $\hat{I}_\mu(f)$ to be the supremum of all $\tilde{I}_\mu(e)$ over all elementary functions e such that $e \leq f$. Note that with this definition, $\hat{I}_\mu(f)$ can be infinite, ie. it can be $+\infty$.
- When f is any measurable function from Ω into $\mathbb{R}$, then we note that $f 1_{\{f \geq 0\}}$ and $-f 1_{\{f<0\}}$ are non-negative and measurable. We say that f is integrable (i.e., in L^1) if both $I_\mu(f 1_{\{f \geq 0\}})$ and $I_\mu(-f 1_{\{f<0\}})$ are finite. When this is the case, one *defines*

$$I_\mu(f) := \hat{I}_\mu(f 1_{\{f \geq 0\}}) - \hat{I}_\mu(-f 1_{\{f<0\}}).$$

This quantity $I_\mu(f)$ is called the integral of f with respect to μ and often denoted by $\int_\Omega f(\omega) d\mu(\omega)$, or $\int f d\mu$.

REMARK 2.5.1. *We see that $\int f d\mu$ is well-defined in $[0, \infty]$ for any measurable non-negative function f, but that when f is not non-negative, we can define $\int f d\mu$ only when $f \in L^1$, i.e., when $\int |f| d\mu < \infty$.*

With such a definition, the various following statements are extremely easy to derive (using facts such as that a bounded monotone sequence of numbers does converge):

(1) Monotonicity. When f, g are two functions in $L^1(\mu)$ such that $f \leq g$, then $\int f d\mu \leq \int g d\mu$. When f, g are two nonnegative measurable functions such that $f \leq g$, then $\int f d\mu \leq \int g d\mu$ (where these quantities are now in $[0, \infty]$).

(2) Linearity. When $a_1, \ldots, a_n$ are constants and $f_1, \ldots, f_n$ are n measurable functions in L^1 (defined on the same measured space), then $af_1 + \ldots + a_n f_n$ is in L^1 as well and $\int (a_1 f_1 + \cdots + a_n f_n) d\mu = a_1 \int f_1 d\mu + \cdots + a_n \int f_n d\mu$.

(3) Jensen's inequality. If μ is a probability measure: When f is a measurable function in $L^1(\mu)$ and φ is a convex function defined on an interval that contains $f(\Omega)$, then

$$\varphi(\int f d\mu) \leq \int \varphi(f) d\mu,$$

where the integral on the right-hand side could possibly take the value $+\infty$.

(4) Monotone convergence. When f_n is a monotone sequence of measurable functions that converges to a limit f, such that each f_n as well as f are in $L^1(\mu)$, then

$$\lim_{n \to \infty} \int f_n d\mu \to \int f d\mu.$$

The same holds if f_n is a non-decreasing sequence of non-negative measurable functions (in which case the various integrals could be infinite, and one considers the limits as elements of $[0, \infty]$).

(5) Fatou's lemma. If f_n is a sequence of non-negative measurable functions that converges to a limit f, then $\int f d\mu \leq \liminf_{n \to \infty} \int f_n d\mu$.

(6) Dominated convergence. When f_n is a sequence of measurable functions that converges to a limit f, and if there exists $g \in L^1$ such that for all $n \geq 1$, $|f_n| \leq g$, then

$$\lim_{n\to\infty} \int f_n d\mu \to \int f d\mu.$$

REVIEW OF THE PROOFS. Let us briefly survey the proofs (without all details – this is not a full substitute for a more detailed measure theory course, but it is quite easy to fill in the missing parts):

- First, the monotonicity for non-negative f and g follow directly from the definition, and then for general f and g, it suffices to substract the inequalities for the nonnegative functions $\int f 1_{f>0} d\mu \geq \int g 1_{g>0} d\mu$ and $\int -g 1_{g<0} d\mu \geq \int -f 1_{f<0} d\mu$.

- Then, one possible next step is to show the monotone convergence theorem in the case of non-decreasing sequences of non-negative functions f_n that converge to a function f. By the monotonicity property, $\int f_n d\mu \leq \int f d\mu$, and the sequence $\int f_n d\mu$ is non-decreasing – we therefore have that $\lim_{n\to\infty} f_n d\mu \leq \int f d\mu$. To prove the reverse inequality, it suffices to show that for any elementary function $e = \sum_{j=1}^k a_j 1_{A_j}$ such that $e \leq f$, one has $\lim_{n\to\infty} \int f_n d\mu \geq \int e d\mu$. To do this, one chooses $\varepsilon > 0$, and notes that since $f_n \to f$, if one defines $E_n = \{f_n \geq (1-\varepsilon)e\}$, then E_n is a non-decreasing family with $\cup E_n = \Omega$. In particular, for all $j \leq k$, $\mu(E_n \cap A_j) \to \mu(A_j)$ as $n \to \infty$. So, we see that by monotonicity,

$$\int f_n d\mu \geq \int f_n 1_{E_n} d\mu \geq \int (1-\varepsilon) e 1_{E_n} d\mu = \sum_{j=1}^k a_j (1-\varepsilon) \mu(A_j \cap E_n) \to (1-\varepsilon) \int e d\mu$$

as $n \to \infty$. Since this is true for all $\varepsilon > 0$, we can conclude.

- This is then the key to unlock the rest: A first useful consequence of this result is that for any nonnegative function f, $\int f d\mu = \lim_{K\to\infty} f 1_{f\leq K}$. A further useful general remark is that any non-negative measurable function f can be approximated by the monotone sequence of non-negative functions $f_n = \varphi_n \circ f$ where $\varphi_n(x) = \sup\{j2^{-n},\ j \geq 0 \text{ and } j2^{-n} \leq x\}$. In other words,

$$f_n(\cdot) = \sum_{j=0}^{\infty} j2^{-n} 1_{\{f(\cdot)\in[j2^{-n},(j+1)2^{-n}]\}}.$$

Then, since $f_n \leq f \leq f_n + 2^{-n}$, one gets that $f_n \to f$. Furthermore, f_n is a non-decreasing sequence of functions so that $\int f_n d\mu$ is also non-decreasing and converges to $\int f d\mu$ (which can be finite or infinite) by the monotone convergence theorem. It is also easy to check (for instance again by the monotone convergence theorem, this time for the sum over j) that for each fixed n,

$$\int f_n d\mu = \sum_{j\geq 0} j2^{-n} \mu(\{\omega :\ f(\omega) \in [j2^{-n}, (j+1)2^{-n}])$$

and the similar fact for $f_n + 2^{-n}$. So one ends up with the concrete bounds

$$\int f_n d\mu \leq \int f d\mu \leq 2^{-n}\mu(\Omega) + \int f_n d\mu.$$

From this type of approximation, one can deduce quite readily the linearity of the integral for nonnegative functions (and positive coeficients) and then the linearity in the general case (by separating each f into $f1_{f>0} - (-f)1_{f<0}$). With these further tools in the bag, the other results follow easily:

- For Jensen's inequality: One can use the monotonicity, the linearity, and the fact that $\varphi(x_0)$ is the supremum of $ax_0 + b$, where a and b range over all values such that the line $ax + b$ lies under the function f).

- For the general mononotone convergence theorem, one can note that $f - f_n$ of $f_n - f$ is a monotone sequence of non-negative functions (either all non-negative or all non-positive), and use the linearity to conclude (details are left to the reader here)

- For Fatou's lemma we can note that $g_n := \inf_{m \geq n} f_m$ is a non-decreasing sequence of non-negative functions that converges to f, so that by monontone convergence, $\int f d\mu = \lim_{n \to \infty} \int g_n d\mu$. But since $g_n \leq f_n$, we immediately get that $\int g_n d\mu \leq \int f_n d\mu$ and the desired result.

- For the dominated convergence theorem (6), by first setting $h_n := f_n - f$, we see that h_n tends to 0 and is bounded by $2|g|$. Then, the sequence $(\sup_{k \geq n} |h_k|)_{n \geq 1}$ is non-increasing, so that $(2|g| - \sup_{k \geq n} |h_k|)_{n \geq 1}$ is a non-decreasing sequence of non-negative functions. Hence by the monotone convergence theorem, the limit of the integrals is the integral of the limit, i.e. $\int 2|g| d\mu$, which means that

$$\int \sup_{k \geq n} |h_k| d\mu \to 0.$$

Since

$$|\int h_n d\mu| \leq \int |h_n| d\mu \leq \int \sup_{k \geq n} |h_k| d\mu$$

we conclude that $\int h_n d\mu \to 0$, so that indeed $\int f_n d\mu \to \int f d\mu$].

□

The Cauchy-Schwarz and Hölder inequalities for integrals are often useful:

PROPOSITION 2.5.2. *When $p, q > 1$ with $1/p + 1/q = 1$, when f and g are non-negative measurable functions such that f^p and g^q are in $L^1(\mu)$, then fg is also in $L^1(\mu)$ and*

$$\int f g d\mu \leq (\int f^p d\mu)^{1/p} (\int g^q d\mu)^{1/q}.$$

In particular, when $p = q = 2$, this is the Cauchy-Schwarz inequality: When f and g are in $L^2(\mu)$ (which means that f^2 and g^2 are in L^1), then $fg \in L^1(\mu)$ and

$$(\int f g d\mu)^2 \leq \int f^2 d\mu \int g^2 d\mu.$$

PROOF. Let us fix $p > 1$ and $q > 1$ so that $1/p + 1/q = 1$, and note that $(p-1)(q-1) = 1$. If the result holds for f and g, then it holds for af and bg as well (for any positive constants a and b). It therefore suffices to prove the result in the case where $\int f^p d\mu = \int g^q d\mu = 1$, i.e., to show that in this case, $\int f g d\mu \leq 1$.

For this, one can first observe that for any non-negative numbers x_0 and y_0, $x_0 y_0 \leq x_0^p/p + y_0^q/q$. This can be seen by tracing the curve $y = x^{p-1}$, and saying that the area rectangle $[0, x_0] \times [0, y_0]$ is smaller than the sum of the area "under the curve" for $x \in [0, x_0]$ and the area "to the left of the curve" (noting that the curve is also $x = y^{q-1}$) for $y \in [0, y_0]$). Therefore, for all $\omega \in \Omega$, $f(\omega) g(\omega) \leq f(\omega)^p/p + g(\omega)^q/q$. By monotonicity, we can integrate both sides with respect to μ, and conclude that $\int f g d\mu \leq 1/p + 1/q = 1$. □

In the context of probability theory (where the measurable functions are random variables), these integrals are called **expectations**. This expectation (which is well-defined when $X \in L^1(P)$) is denoted by $E[X]$ (and is nothing else than $\int_\Omega X(\omega) dP(\omega)$ in the measure theory notation).

When X is a non-negative random variable, one can also define $E[X] = \infty$ when X is not in L^1. So, for instance, when X is not in $L^1(P)$ (one sometimes just writes L^1 instead of $L^1(P)$), $E[|X|]$ is always well-defined as an element in $[0, \infty]$.

Let us list a few useful bits and pieces:

(a) When $X \in L^1$, then for all $a \geq 0$, one cleary has $|X|1_{|X|\geq a} \geq a1_{|X|\geq a}$, so that

$$E[|X|] \geq E[|X|1_{|X|\geq a}] \geq E[a1_{|X|\geq a}] = aP[|X| \geq a].$$

The inequality $aP[|X| \geq a] \leq E[|X|]$ is often refered to as Markov's inequality.

(b) The following criterion turns out to be quite useful:

LEMMA 2.5.3. *A random variable X satisfies $E[|X|] < \infty$ if and only if $\sum_{n\geq 0} P[|X| > n] < \infty$.*

PROOF. This is simply due to the fact that when $x \in [n, n+1)$, well, then $n \leq x < n+1$, so that

$$nP(|X| \in [n, n+1)) \leq E[|X|1_{|X|\in[n,n+1)}] < (n+1)P(|X| \in [n, n+1))$$

Hence, summing this up over all $n \geq 0$, we get

$$\sum_{n\geq 1} nP(|X| \in [n, n+1)) \leq E[|X|] \leq \sum_{n\geq 0}(n+1)P(|X| \in [n, n+1)) = 1 + \sum_{n\geq 1} nP(|X| \in [n, n+1))$$

(where we allow here these sums of non-negative numbers to be infinite). In other words, $E[|X|]$ is finite if and only if $\sum_{n\geq 1} nP(|X| \in [n, n+1))$ is finite.

We now write $n = \sum_{j=1}^{n} 1$, and invert the order of summation in n and j:

$$\sum_{n\geq 1} nP(|X| \in [n, n+1)) = \sum_{j\geq 1}\sum_{n\geq j} P(|X| \in [n, n+1)) = \sum_{j\geq 1} P(|X| \geq j).$$

So, $E[|X|]$ is indeed finite if and only if $\sum_{j\geq 1} P(|X| \geq j)$ is finite. □

REMARK 2.5.4. *Note that one can also apply this result to powers of a random variable Y. For instance, $E[Y^2] < \infty$ if and only if $\sum_j P[Y^2 \geq j] < \infty$, i.e., if $\sum_j P[|Y| \geq \sqrt{j}] < \infty$.*

(c) When X and Y are two random variables that are in L^2, then the product XY is in L^1 because $|XY| \leq X^2 + Y^2$. This leads to the following observation:

LEMMA 2.5.5. *When X and Y are two independent random variables in L^1, then XY is also in L^1 and $E[XY] = E[X]E[Y]$.*

PROOF. Let us first assume that both X and Y are in fact non-negative and bounded (by some constant K). Then obviously XY is non-negative and bounded as well. We can use the same function φ_n as above, and note that

$$\begin{aligned} E[\varphi_n(X)\varphi_n(Y)] &= \sum_{j,j'\geq 0} \frac{jj'}{4^n} P[\varphi_n(X) = j2^{-n}, \varphi_n(Y) = j'2^{-n}] \\ &= \sum_{j,j'\geq 0} \frac{jj'}{4^n} P[\varphi_n(X) = j2^{-n}]P[\varphi_n(Y) = j'2^{-n}] = E[\varphi_n(X)]E[\varphi_n(Y)], \end{aligned}$$

and conclude by the monotone convergence theorem (applied on both sides) that

$$E[XY] = E[X]E[Y].$$

Next, we assume that X and Y are non-negative and in L^1. We can apply the previous result to the truncated variables, so that for all $K > 0$,

$$E[X1_{X\leq K}Y1_{Y\leq K}] = E[X1_{X\leq K}]E[Y1_{Y\leq K}].$$

Again, we can conclude using the monotone convergence theorem, letting $K \to \infty$, that XY is in L^1 and that $E[XY] = E[X]E[Y]$.

Finally, if we only assume that X and Y are in L^1, we can apply the previous result to the four possible combinations of one of $(X1_{X>0}, -X1_{X<0})$ and $(Y1_{Y>0}, Y1_{Y<0})$ to conclude that the lemma holds. □

(d) This last lemma naturally leads to the definition of the **variance** of a random variable X that is in L^2, defined as

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

when $m = E[X]$. We can note that when one expands the square, one gets that

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

This quantity is a number that gives some information on how spread away from m the law of the random variable X is. We see that:

LEMMA 2.5.6. *When X and Y are independent random variables in L^2, then*

$$Var(X+Y) = Var(X) + Var(Y).$$

Indeed, with obvious notations, we note that $m_{X+Y} = m_X + m_Y$ and then $E[(X+Y)^2] = E[X^2] + E[Y^2] + 2E[XY] = E[X^2] + E[Y^2] + 2m_x m_Y$, so that

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

(e) Suppose that X and Y are independent random variables. Then, for any two measurable bounded functions f and g from $\mathbb{R}$ to $\mathbb{R}$, the bounded random variables $f(X)$ and $g(Y)$ are independent as well (since they are measurable with respect to $\sigma(X)$ and $\sigma(Y)$ respectively), and it follows that $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$. The following converse easy statement is sometimes useful:

LEMMA 2.5.7. *If for any bounded continuous functions f and g from $\mathbb{R}$ to $\mathbb{R}$, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$, then X and Y are independent.*

Similar results hold for random vectors, or independence between more than two random variables or vectors.

PROOF. One just has to check that for any reals a and b,

$$P(X \le a, Y \le b) = P(X \le a)P(Y \le b).$$

One just finds decreasing sequences of bounded continuous functions f_n and g_n such that $f_n \to 1_{(-\infty,a]}$ and $g_n \to 1_{(-\infty,b]}$ pointwise (for instance, $f_n - 1_{[-\infty,a]}$ is chosen to be linear on $(a, a+1/n]$ and equal to 1 at $a+$ and to 0 at $a+1/n$), and then one applies the monotone convergence theorem to all terms of the identity

$$E[f_n(X)g_n(Y)] = E[f_n(X)]E[g_n(Y)].$$

□

(f) We have seen that $E[|X|]^2 \le E[X^2]$ (because the difference is $\mathrm{Var}(|X|)$) for any L^2 random variable (this can also be viewed as a special case of the Cauchy-Schwarz inequality when one of the two functions is the constant 1 and the other one is X, or as a consequence of Jensen's inequality for the square function etc.). We conclude this chapter with the following useful observation:

LEMMA 2.5.8 (Paley-Zygmund inequality). *Suppose that X is a non-negative random variable in L^2 with $P(X=0) \neq 1$. Then, for any $\theta \in (0,1)$,*

$$P(X \ge \theta E[X]) \ge (1-\theta)^2 \frac{E[X]^2}{E[X^2]}.$$

For instance, we see that if we know that $E[X^2] < 4E[X]^2$ then

$$P(X \ge E[X]/2) \ge (1/16).$$

REMARK 2.5.9. *The simple intuition behind this result is the following. By scaling, it is sufficient to consider the case where $E[X] = 1$. When $\theta < 1$ is fixed, if the probability that $X < \theta$ is too close to 1, then in order to compensate for this for the expectation $E[X]$ to be equal to 1, then on the remaining event $X \geq \theta$ (that has a small probability), X would have to be quite large in order to average out things so that the integral of X on this event with small probability is still at least $1 - \theta$. But when X is very large, then X^2 is even larger, and therefore $E[X^2]$ will be very large. So, if one has an upper bound on $E[X^2]$, it will impose that $P(X \geq \theta E[X])$ can not be too small.*

PROOF. Let us define $Y = X/E[X]$, so that $E[Y] = 1$. Using Cauchy-Schwarz, we get that

$$1 = E[Y] = E[Y 1_{Y>\theta}] + E[Y 1_{Y \leq \theta}] \leq (E[Y^2] P(Y \geq \theta))^{1/2} + \theta,$$

so that indeed,

$$P(X \geq \theta E[X]) = P(Y \geq \theta) \geq \frac{(1-\theta)^2}{E[Y^2]} = (1-\theta)^2 \frac{E[X]^2}{E[X^2]}.$$

□

CHAPTER 3

Sequences, series and means of independent random variables

3.1. Existence

Let us know explain how, starting from the Lebesgue measure defined on $[0,1]$, one can in fact directly define a sequence of independent random variables with any prescribed law:

PROPOSITION 3.1.1. *For any given sequence of distribution functions $(F_i)_{i\geq 1}$, it is possible to find a probability space $(\Omega, \mathcal{A}, P)$ and a sequence of independent random variables $(Y_i)_{i\geq 1}$ defined on this space such that for each i, the distribution function of Y_i is F_i.*

PROOF. We use the probability space $([0,1], \mathcal{B}_{[0,1]}, P)$ where P is the Lebesgue measure on $[0,1]$.

We know that for any given x, $P[\{x\}] = 0$, so that for any countable set D in $[0,1]$, $P[D] = 0$ too. This holds in particular when D is the set of all dyadic numbers of the type $j/2^n$ for $n \geq 0$ and $j \leq 2^n$. When $x \in [0,1]$ is not such a dyadic number, then there exists a unique sequence $(\varepsilon_n)_{n\geq 1}$ in $\{0,1\}$ such that $x = \sum_{n\geq 1} \varepsilon_n 2^{-n}$. When x is in D, we can simply define ε_n to be equal to 0 for all n. We can then view each ε_n as a random variable defined on $[0,1]$ (one can think of x as ω and of $\varepsilon_j(x)$ as of $X_j(\omega)$).

Clearly, for all $n \geq 1$, $P[\varepsilon_n = 1] = P[\varepsilon_n = 0] = 1/2$, as the set of x's for which $\varepsilon_n = 1$ is the union of 2^{n-1} intervals of length 2^{-n}. Furthermore, for all k and any $(u_1, \ldots, u_k) \in \{0,1\}$,

$$P[\varepsilon_1 = u_1, \ldots, \varepsilon_k = u_k] = 2^{-k} = \prod_{j=1}^{k} P(\varepsilon_{n_j} = u_j),$$

from which it follows that the random variables $(\varepsilon_n)_{n\geq 1}$ are independent.

We can now use any given bijection φ from $\{1,2,3,\ldots\}^2$ into $\{1,2,3,\ldots\}$. Then, we can define, for each $i \geq 1$,

$$X_i := \sum_{j\geq 1} \varepsilon_{\varphi(i,j)} 2^{-j}.$$

We can check that:

- The law of each X_i is the Lebesgue measure on $[0,1]$ – one way to justify this is to say that $P(X_i \in I)$ is equal to the Lebesgue measure of I for any dyadic interval, and that the set of dyadic intervals (i.e., the intervals of the type $[j2^{-n}, (j+1)2^{-n}]$) is a π-system generating $\mathcal{B}_{[0,1]}$.

- The random variables $(X_i)_{i\geq 1}$ are independent: Indeed, for any k, and any collection of dyadic intervals $I_1, \ldots, I_k$,

$$P(X_1 \in I_1, \ldots, X_k \in I_k) = P(X_1 \in I_1) \ldots P(X_k \in I_k)$$

because of the independence of the finite subsets of the (ε_n) involved in these events, and for each fixed i, the collection of events $(X_i \in I)$ when I is a dyadic interval is a π-system that generates $\sigma(X_i)$.

Finally, for each given sequence of distribution functions $(F_n)_{n\geq 1}$, we define for each $i \geq 1$, functions

$$f_i(x) := \sup\{y \in \mathbb{R} \ : \ F_i(y) < x\}$$

and then $Y_i = f_i(X_i)$. Just as in the Lebesgue-Stieljes lemma, we see that for each i, the distribution function of the random variable Y_i is F_i. Furthermore, the random variables $(Y_i)_{i\geq 1}$ are independent, because for each $a_1, \ldots, a_n$,

$$\begin{aligned} P[Y_1 \leq a_1, \ldots, Y_n \leq a_n] &= P[X_1 \leq F_1(a_1), \ldots, X_n \leq F_n(a_n)] \\ &= F_1(a_1) \ldots F_n(a_n) = P[Y_1 \leq a_1] \ldots P[Y_n \leq a_n]. \end{aligned}$$

□

REMARK 3.1.2. *An alternative route to construct families of independent random variables is to use product spaces and product measures (see the exercise sheets).*

REMARK 3.1.3. *We have provided a direct proof here of the fact that the $(X_i)_{i\geq 1}$ are independent. It is however useful to have the following more general results in one's toolbox:*

(1) *Suppose that $(Y_j)_{j\geq 1}$ is a sequence of random variables, and that for some sequence of measurable functions F_n, we know that $F_n(Y_1, \ldots, Y_n)$ converges to some random variable Y. Define the σ-field*

$$\sigma(Y_1, Y_2, \ldots) = \sigma(\cup_{j\geq 1} \sigma(Y_j)).$$

Then, the random variable Y is $\sigma(Y_1, Y_2, \ldots)$-measurable as limits of random variables that are measurable with respect to this σ-field.

(2) *A π-system that generates this $\sigma(Y_1, Y_2, \ldots)$ is the set of events of the type*

$$\{Y_1 \leq a_1, \ldots, Y_n \leq a_n\}$$

where $n \geq 1$ and $a_1, \ldots, a_n$ are in $\mathbb{R} \cup \{+\infty\}$.

(3) *As a consequence, if we now have an array $(Y_j^i)_{i\geq 1, j\geq 1}$ of random variables such that for all $k \geq 1$ and all $n \geq 1$, and all $(a_j^i)_{i\leq k, j\leq n}$ in $\mathbb{R} \cup \{+\infty\}$,*

$$P(\cap_{i\leq k, j\leq n} \{Y_j^i \leq a_j^i\}) = \prod_{i\leq k} P(\cap_{j\leq n} \{Y_j^i \leq a_j^i\}),$$

then the σ-fields $\sigma(Y_i^1, Y_i^1, \ldots)$ for $i \geq 1$ are independent.

So, in the present case of the array $(\varepsilon_{\varphi(i,j)})_{i,j\geq 1}$, we get that $(X_1, X_2, \ldots)$ are independent because each X_i is measurable with respect to $\sigma(\varepsilon_{\varphi(i,1)}, \varepsilon_{\varphi(i,2)}, \ldots)$ and these σ-fields are independent.

3.2. Warm-ups

3.2.1. Use of Borel-Cantelli. As a warm-up to what is going to follow, let us illustrate with a concrete result the fact that when one considers sequences of independent random variables, then the Borel-Cantelli lemmas are quite handy:

Suppose now that $(X_n)_{n\geq 1}$ is a sequence of independent random variables, and than $(a_n)_{n\geq 1}$ is some given sequence of numbers. We see that the sequence of events $\{X_n > a_n\}$ are independent. So, by the Borel-Cantelli lemmas, we see that:

- When $\sum_n P[X_n > a_n] < \infty$, then almost surely there exists a (random) n_0 such that for all $n \geq n_0$, $X_n \leq a_n$-
- When $\sum_n P[X_n > an] = \infty$ then almost surely, $X_n > a_n$ for infinitely many n's.

This type of considerations will be useful in our study of the convergence (or not) of $\sum_{n=1}^N X_n$ as $N \to \infty$ in the next section. Let us state one special result, in the case where the X_n's are identically distributed:

COROLLARY 3.2.1. *Let $(X_n)_{n\geq 1}$ denote a sequence of independent identically distributed random variables. Then:*

- *If $E[|X_1|] < \infty$, then almost surely, $|X_n|/n \to 0$ as $n \to \infty$.*
- *If $E[|X_1|] = \infty$, then almost surely, $|X_n|/n$ does not tend to 0 as $n \to \infty$-*

PROOF. Suppose first that $E[|X_1| < \infty$. Then for any $\varepsilon_0 > 0$, one has $\sum_n P[|X_1|/\varepsilon_0 \geq n] < \infty$. This sum is identical to $\sum_n P[|X_n|/n \geq \varepsilon_0]$. We can then apply the first Borel-Cantelli lemma, and conclude that there almost surely exists n_0 such that for all $n \geq n_0$, $|X_n|/n < \varepsilon_0$.

Let us now consider a fixed positive sequence $\varepsilon_k \to 0$ (for instance, $\varepsilon_k = 1/k$). The same argument shows that almost surely, there exists n_k such that for all $n \geq n_n$, $|X_n|/n < \varepsilon_k$.

The intersection of a countable family of events of probability 1 still has probability 1. Hence, we can exchange the "almost surely" and the "for all k" and state that almost surely, for all k, there exists n_k such that for all $n \geq n_k$, $|X_n|/n \leq \varepsilon_k$. In other words, almost surely, $|X_n|/n \to 0$.

Suppose now that $E[|X_1|] = \infty$. Then, $\sum_n P[|X_n|/n > 1] = \infty$, and by the second Borel-Cantelli lemma, we know that almost surely, $|X_n|/n > 1$ for infinitely many values of n, which shows in particular that almost surely, $|X_n|/n$ does not tend to 0. □

3.2.2. Strong law of large numbers, easy version.

THEOREM 3.2.2 (Easy version of the strong law of large numbers). *Suppose that $(X_n)_{n\geq 1}$ is a sequence of identically distributed independent random variables such that $E[(X_i)^4] < \infty$. Then, almost surely, the sequence $(X_1 + \ldots + X_n)/n$ converges to $E[X_1]$.*

REMARK 3.2.3. *As we will see a bit later, the statement actually still holds when one replaces the condition $E[(X_1)^4] < \infty$ by the weaker condition $E[|X_1|] < \infty$.*

PROOF. Without loss of generality, it suffices to prove the result when $E[X_1] = 0$ (for the general case, one can then apply this result to the sequence $\tilde{X}_i = X_i - E[X_i]$). Let $K = E[(X_1)^4]$. A first remark is that by Cauchy-Schwarz, $E[|X_1|]^4 \leq E[(X_1)^2]^2 \leq E[(X_1)^4] = K$. Let us denote the sum $X_1 + \cdots + X_n$ by S_n. Then,

$$S_n^4 = \sum_{j_1,j_2,j_3,j_4 \leq n} X_{j_1} X_{j_2} X_{j_3} X_{j_4}$$

so that

$$E[(S_n)^4] = \sum_{j_1,j_2,j_3,j_4 \leq n} E[X_{j_1} X_{j_2} X_{j_3} X_{j_4}].$$

Using the independence and the fact that $E[X_i] = 0$, we see that $E[X_{j_1}X_{j_2}X_{j_3}X_{j_4}] = 0$ as soon as one of the indices j_i is different from the other three ones. For instance, if $j_1 \notin \{j_2, j_3, j_4\}$, then

$$E[X_{j_1}X_{j_2}X_{j_3}X_{j_4}] = E[X_1]E[X_{j_2}X_{j_3}X_{j_4}] = 0.$$

So, keeping only the remaining terms, we get that

$$E[(S_n)^4] = \sum_{j=1}^{n} E[(X_j)^4] + 6 \sum_{1 \leq j < j' \leq n} E[(X_j)^2 (X_{j'})^2]$$

(for each $j < j'$, there are six ways to choose j_1, j_2, j_3, j_4 so that j and j' appear twice each). Hence,

$$E[(S_n)^4] = nE[(X_1)^4] + 3n(n-1)E[(X_1)^2]^2 \leq 3n^2 K.$$

In other words,

$$E[(S_n/n)^4] \leq \frac{3K}{n^2}.$$

There are now several (but almost equivalent) ways to deduce from this inequality that $S_n/n \to 0$ almost surely. One way is to note that

$$E[\sum_{n \geq 1} (S_n/n)^4] = \sum_{n \geq 1} E[(S_n/n)^4] \leq 3K \sum_{n \geq 1} n^{-2} < \infty.$$

Hence, $\sum_{n \geq 1} (S_n/n)^4$ is almost surely finite, which implies in particular that (S_n/n) tends to 0 almost surely. □

REMARK 3.2.4. *We can notice that the previous method of proof does not only show that S_n/n tends to $E[X_1]$, but can be used to provide information about the speed of convergence. Indeed, we see that in fact, for all $\varepsilon > 0$, we can replace the last displayed equation of the proof by*

$$E[\sum_{n \geq 1} (S_n)^4/n^{3+\varepsilon}] \leq \sum_{n \geq 1} 3K/n^{1+\varepsilon} \ < \infty,$$

so that by the same argument, $S_n^4/n^{3+\varepsilon} \to 0$ almost surely. In other words, when $E[X_1] = 0$, for all $\delta > 0$, by choosing $\varepsilon = 4\delta$, we get that $S_n/n = o(n^{-1/4+\delta})$ almost surely. This means that in the general case where $E[X_1]$ is not necessarily 0, $(S_n/n - E[X_1]) = o(n^{-1/4+\delta})$.

One could for all integer $k \geq 3$ also apply a similar method of proof under the assumption that $E[(X_1)^{2k}] < \infty$ by expanding $E[(S_n)^{2k}]$. This would then provide a better bound on the rate of convergence of S_n/n. In the case where one knows that all *moments of X_1 are finite (which is for instance the case where the random variables X_1 is itself bounded), we can then let $k \to \infty$, and it is then possible to see that for all $\delta > 0$, $S_n/n^{1/2+\delta} \to 0$ almost surely. All the content of this remark is discussed in the exercise sheets.*

The previous remark shows that the better control one has on the moments of X_1, the better control one gets on the convergence of S_n/n. The following result is a somewhat similar spirit, i.e., that for S_n/n to converge, it is necessary that X_1 is in L^1.

PROPOSITION 3.2.5. *If $(X_n)_{n \geq 1}$ is a sequence of independent identically distributed random variables such that S_n/n converges almost surely, then necessarily $E[|X_1|] < \infty$.*

PROOF. This is a direct consequence of the corollary in the previous section. If S_n/n converges to some limit, then $S_n/(n+1)$ converges to the same limit, so that

$$X_n/n = (S_n/n) - (S_{n-1}/n) \to 0.$$

But the corollary states precisely that for this to happen with positive probability, it is necessary that $E[|X_1|] < \infty$. □

3.3. Kolmogorov's $0-1$ law

Let $(X_n)_{n\geq 1}$ be a sequence of independent random variables defined on a same probability space. We do not assume that they have the same law, but of course, the independence assumption will be crucial here. For each n, we define the σ-field $\mathcal{G}_{n_0}$ defined by the sequence $(X_{n_0+1}, X_{n_0+2}, \dots)$ of random variables. This is the smallest σ-field that contains $\cup_{j\geq 1}\sigma(X_{n_0+j})$.

We also define $\mathcal{G}_\infty := \cap_{n_0}\mathcal{G}_{n_0}$. This is sometimes called the tail σ-algebra of the sequence $(X_n)_{n\geq 1}$. Kolmogorov's $0-1$ law essentially stipulates that this tail σ-algebra contains no probabilistic information:

PROPOSITION 3.3.1 (Kolmogorov's $0-1$ law). *The tail σ-algebra $\mathcal{G}_\infty$ is trivial in the sense that for all $A \in \mathcal{G}_\infty$, one has $P(A) \in \{0, 1\}$.*

PROOF. We first note that a generating π-system of $\mathcal{G}_{n_0}$ is the set of events of the type

$$\{X_{n_0+1} \leq a_1, \dots, X_{n_0+k} \leq a_k\}$$

for $k \geq 1$ and $a_1, \dots, a_k \in \mathbb{R} \cup \{+\infty\}$.

Let us also define $\mathcal{F}_{n_0}$ to be $\sigma(X_1, \dots, X_{n_0})$. A π-system generating $\mathcal{F}_{n_0}$ is the set of events of the type

$$\{X_1 \leq b_1, \dots, X_{n_0} \leq b_{n_0}\}$$

for $b_1, \dots, b_{n_0}$ in $\mathbb{R} \cup \{+\infty\}$.

Note that $\sigma(X_1, X_2, \dots)$ (which is $\mathcal{G}_0$) is generated by the union of all $\mathcal{F}_n$, which is also a π-system.

Here are the steps of the proof:

- For each given n_0, $\mathcal{F}_{n_0}$ and $\mathcal{G}_{n_0}$ are independent. This follows by applying the π-system criterion to the two generating π-systems that we just described, and using the independence of the X_j.
- Since $\mathcal{G}_\infty$ is a subset of each $\mathcal{G}_{n_0}$, this means that for all A in $\mathcal{G}_\infty$ and all $A' \in \cup_{n_0}\mathcal{F}_{n_0}$, $P(A \cap A') = P(A)P(A')$. Using the π-system criterion again (because $\cup_{n_0\geq 1}\mathcal{F}_{n_0}$ is a π-system that generates $\mathcal{G}_0$), this implies that $\mathcal{G}_\infty$ is independent of $\mathcal{G}_0$.
- Finally, since $\mathcal{G}_\infty \subset \mathcal{G}_0$, when $A \in \mathcal{G}_\infty$, we can take $A' = A$ which is also in $\mathcal{G}_0$, so that
$$P(A \cap A) = P(A \cap A') = P(A)P(A') = P(A)^2,$$
which implies that $P(A)$ is either equal to 0 or to 1.

□

This result has some direct interesting consequences:

COROLLARY 3.3.2. *When $(X_n)_{n\geq 1}$ is a sequence of independent random variables, then the event that $\sum_{n=1}^N X_n$ converges as $N \to \infty$ has probability 0 or 1.*

COROLLARY 3.3.3. *When $(X_n)_{n\geq 1}$ is a sequence of independent random variables, then the event that $(\sum_{n=1}^N X_n)/N$ converges as $N \to \infty$ has probability 0 or 1. Furthermore, if it has probability 1, then the limiting random variable is in fact constant – in other words, there then exists a constant C such that almost surely, $(\sum_{n=1}^N X_n)/N \to C$ as $N \to \infty$.*

COROLLARY 3.3.4. *When $(X_n)_{n\geq 1}$ is a sequence of independent random variables, then there exist constants $C_-, C_+ \in \{-\infty, \infty\} \cup \mathbb{R}$, such that almost surely,*

$$\liminf_{N\to\infty} \frac{X_1 + \cdots + X_N}{N} = C_- \text{ and } \limsup_{N\to\infty} \frac{X_1 + \cdots + X_N}{N} = C_+.$$

To prove these corollaries, one just needs to check that the events under consideration are in $\mathcal{G}_{n_0}$ for any given n_0 – and therefore in $\mathcal{G}_\infty$:

- For the first corollary, one can notice that $\sum_{n=1}^{N} X_n$ converges as $n \to \infty$ if and only if $\sum_{n=n_0+1}^{N} X_n$ converges as $n \to \infty$ – so that the event that the sum converges is indeed in $\mathcal{G}_{n_0}$ (for any n_0).
- For the first half of the second corollary: One can notice that $(\sum_{n=1}^{N} X_n)/N$ converges as $N \to \infty$ if and only if $(\sum_{n=n_0+1}^{N} X_n)/N$. Then, if this convergence holds, we can define the limit to be Y, which is a finite random variable. For each given c in $\mathbb{R}$, the event $\{Y \leq c\}$ is also in $\mathcal{G}_{n_0}$ for each given n_0, so that $P(Y \leq c) \in \{0, 1\}$, from which it follows that Y is almost surely constant.
- We leave the similar third corollary proof as an exercise.

3.4. Kolmogorov's three-series theorem

We now establish a criterion (that will turn out to be "optimal") that ensures that a series of independent random variables converge:

THEOREM 3.4.1 (Kolmogorov's three series theorem, I). *Suppose that $(X_n)_{n\geq 1}$ is a sequence of independent random variables. We assume that for some given $a > 0$, if one defines $Y_n := X_n 1_{|X_n|\leq a}$, then the following three series converge (note that this is a condition on the laws of these variables X_n):*
(i) $\sum_{n\geq 1} P[|X_n| > a] < \infty$,
(ii) $\sum_{n=1}^{N} E[Y_n]$ converges as $N \to \infty$,
(iii) $\sum_{n\geq 1} Var[Y_n] < \infty$.
Then, almost surely, the series $S_N := \sum_{n=1}^{N} X_n$ converges as $N \to \infty$ to a (finite) random variable.

REMARK 3.4.2. *We will see a bit later that (i)-(ii)-(iii) is in fact a necessary condition for the convergence of the series. We would like to emphasize that the convergence in (i) and in (iii) are for series of non-negative numbers, whereas the convergence of (ii) (or that of S_N) is a simple convergence (not necessarily an absolute convergence).*

A "close relative" of the previous theorem goes as follows:

THEOREM 3.4.3 (Convergence of series of independent centered L^2 random variables). *Suppose that $(Z_n)_{n\geq 1}$ is a sequence of independent random variables that are all in L^2 and centered (i.e., $E[Z_n] = 0$), with $\sum_{n\geq 1} E[Z_n^2] < \infty$. Then, $S_N := \sum_{n=1}^{N} Z_n$ converges almost surely as $N \to \infty$ to a finite random variable.*

REMARK 3.4.4. *We can note that the conditions in this second theorem immediately imply that the series is Cauchy in L^2 and therefore converges in L^2, because $E((S_{N+p}-S_N)^2] = \sum_{j=1}^{p} E[Z_{N+j}^2]$. So the statement can be also reformulated by saying that when a series of independent random variables converges in L^2, then it also converges almost surely.*

Before turning to the proof of these results, let us state and derive the key lemma that will be used in their proofs. The arguments used here will show up again when we will study martingales in a forthcoming chapter.

LEMMA 3.4.5 (Maximal inequality for sums of L^2 variables). *Suppose that $(V_1, \ldots, V_k)$ are k independent random variables in L^2 such that $E[V_j] = 0$ for $j = 1, \ldots, k$. We then write for each $j \leq k$, $S_j = V_1 + \ldots + V_j$. Then for each $\lambda > 0$,*

$$P[\max_{j\leq k} |S_j| \geq \lambda] \leq E[S_k^2]/\lambda^2.$$

REMARK 3.4.6. *We can note that $E[S_k^2] = \sum_{j=1}^{k} E[V_j^2]$. Also, if we would apply just Markov's inequality to S_k^2, we would get*

$$P[|S_k| \geq \lambda] = P[S_k^2 \geq \lambda^2] \leq E[S_k^2]/\lambda^2.$$

We see that this maximal inequality allows to get the same upper bound for $P(\max(|S_1|, \ldots, |S_k|) \geq \lambda)$ even though random variable $\max(|S_1|, \ldots, |S_k|) \geq |S_k|$.

We now successively prove the maximal inequality, the theorem about series of L^2 variables, and the three-series theorem:

PROOF OF THE MAXIMAL INEQUALITY. The first observation is to note that if $1 \leq j < k$, then if U_j is a $\sigma(V_1, \ldots, V_j)$-measurable random variable in L^2, then V_{j+1} is in L^2 and is independent of U_j, so that

$$E[U_j V_{j+1}] = E[U_j]E[V_{j+1}] = 0.$$

The main trick is to consider a slightly modified sequence $(\Sigma_1, \ldots, \Sigma_k)$ instead of $(S_1, \ldots, S_k)$ defined as follows: If all the values of the latter stay in $(-\lambda, \lambda)$, then we take $\Sigma_j = S_j$ for all $j \leq k$. If however, for some $j \leq k$, $|S_j| \geq \lambda$, then we consider J to be the smallest of the values of j for which $|S_j| \geq \lambda$, and then $(\Sigma_1, \ldots, \Sigma_J) = (S_1, \ldots, S_J)$ while for $j = J+1, \ldots, k$, we freeze $\Sigma_j = S_J$.

For each j, we define A_j to be the event that $\max(|S_1|, \ldots, |S_j|) < \lambda$. This is a $\sigma(V_1, \ldots, V_j)$ measurable event (by observing $S_1, \ldots, S_j$, we know if A_j holds). We then note that

$$\Sigma_k = \sum_{j=1}^{k} 1_{A_{j-1}} V_j.$$

One way to think about it is that one progressively sums all the V_j's, but as soon as the absolute value of this sum exceeds λ, one stops summing. The obtained sum is Σ_k. In particular, we see that $\max(|S_1|, \ldots, |S_k|) \geq \lambda$ if and only if $|\Sigma_k| \geq \lambda$. By Markov's inequality, we know that

$$P[|\Sigma_k| \geq \lambda] \leq E[(\Sigma_k)^2]/\lambda^2,$$

so if we could prove that $E[\Sigma_k^2] \leq E[S_k^2]$, then we are done. But

$$E[\Sigma_k^2] = E[(\sum_{j=1}^{k} 1_{A_{j-1}} V_j)^2]$$

and for all $j' < j$, $1_{A_{j'-1}} V_{j'} 1_{A_{j-1}}$ is a $\sigma(V_1, \ldots, V_{j-1})$ measurable bounded random variable. Hence,

$$E[1_{A_{j'-1}} V_{j'} 1_{A_j} V_j] = 0.$$

So, if we expand the square, all cross-term vanish and

$$E[(\sum_{j=1}^{k} 1_{A_{j-1}} V_j)^2] = \sum_{j=1}^{k} E[1_{A_{j-1}} V_j^2] \leq \sum_{j=1}^{k} E[V_j^2] = E[S_k^2]$$

(in the last equality, we use that $E[V_j V_{j'}] = 0$ as soon as $j \neq j'$), which concludes the proof. □

PROOF OF THE CONVERGENCE OF SERIES OF CENTERED L^2 VARIABLES. Consider a sequence $(Z_n)_{n \geq 1}$ of independent random variables, such that almost surely, $E[Z_n] = 0$ and $\sum_n E[(Z_n)^2] < \infty$. Let $S_n = Z_1 + \ldots + Z_n$.

Let us choose $\varepsilon_l = 1/l$. We want to show that almost surely, there exists at least one value n such that the event

$$W_n := \{\forall n' \geq n,\ |S_{n'} - S_n| < \varepsilon_l\}$$

holds. Indeed, this then shows that for all $n', n'' \geq n$, $|S_{n'} - S_{n''}| < 2\varepsilon_l$. Since this is almost surely true for all $l \geq 1$ (the intersection of countably many events with probability 1 still has probability 1), so that we can conclude that the sequence S_n is almost surely a Cauchy sequence, and therefore almost surely converges.

It clearly suffices to show that $P[W_n] \to 1$ as $n \to \infty$. But

$$1 - P[W_n] = P[\exists n' \geq n,\ |S_{n'} - S_n| \geq \varepsilon_l] = \lim_{k \to \infty} P[\exists j \leq k,\ |S_{n+j} - S_n| \geq \varepsilon_l].$$

This is where we use the maximal inequality. When applied to the sequence $V_1 = Z_{n+1}, \ldots, V_k = Z_{n+k}$, it says that

$$P[\exists j \leq k,\ |S_{n+j} - S_n| \geq \varepsilon_l] \leq \varepsilon_l^{-2} E[(\sum_{j=1}^{k} Z_{n+j})^2] = \varepsilon_l^{-2} \sum_{j=1}^{k} E[Z_{n+j}^2].$$

Letting $k \to \infty$, we finally get that

$$1 - P[W_n] \leq \varepsilon_l^{-2} \sum_{j \geq 1} E[Z_{n+j}^2] = \varepsilon_l^{-2} \sum_{n' \geq n} E[Z_{n'}^2].$$

But since $\sum_{n' \geq 1} E[(Z_{n'})^2] < \infty$, this final tail sum does indeed tend to 0 as $n \to \infty$, so that we are done. □

Proof of the three series theorem. Assume that (i)-(ii)-(iii) holds. By Borel-Cantelli, we know from Condition (i) that almost surely, $|X_n| > a$ holds for only finitely many values of n, so that for some (possibly random) n_0, one has $Y_n = X_n$ for all $n \geq n_0$. In particular, this shows that (up to a zero probability event) the series S_N will converge if and only if $\sum_{n=1}^{N} Y_n$ converges.

By (ii), the series $\sum_{n=1}^{N} Y_n$ converges if and only if $\sum_{n=1}^{N} Z_n$ converges, where $Z_n = Y_n - E[Y_n]$.

Now, the sequence $(Z_n)_{n \geq 1}$ is a sequence of independent centered bounded random variables, that satisfies the conditions of the L^2 series convergence theorem because of (iii), so that $\sum_{n=1}^{N} Z_n$ indeed converges almost surely as $N \to \infty$. □

3.5. Law of large numbers

We now state and prove the stronger version of the law of large numbers. Let us insist on the fact that the setup is here quite different from Kolmogorov's three series theorem (here the random variables are identically distributed) and the object of study is different (it is the mean of the random variables, not their sum) – however, as we shall see, it is possible to view the law of large numbers as a consequence of the three series theorem.

THEOREM 3.5.1 (Stronger version of the strong law of large numbers). *Suppose that $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables such that $E[|X_1|] < \infty$. We define $S_n = X_1 + \ldots + X_n$. Then, almost surely, the sequence S_n/n converges to $E[X_1]$.*

REMARK 3.5.2. *Recall that we have seen that when $E[|X_1|] = \infty$, then almost surely, $|X_n| \geq n$ for infinitely n so that almost surely, S_n/n is not convergent. The condition $E[|X_1|] = \infty$ is therefore optimal.*

There are a number of different ways to derive this result. We opt here for the one that views it as a rather direct consequence of the three-series theorem for series of random variables (or rather the criterion for convergence of sums of L^2 independent variables that we established), using Césaro mean tricks.

Before proving the theorem, let us first state and prove the following very simple fact about (deterministic) series:

LEMMA 3.5.3 (Kronecker's lemma, simple version). *Suppose that $(x_n)_{n\geq 1}$ is a sequence of real numbers such that $\sum_{j=1}^n x_j/j$ converges as $n \to \infty$, then $(x_1 + \cdots + x_n)/n$ tends to 0 as $n \to \infty$.*

PROOF. We assume that the limit of $w_n := \sum_{j=1}^n x_j/j$ exists and we denote it by w. The Césaro mean theorem when applied to this convergent sequence says that $(w_1 + \cdots + w_N)/N \to w$ as $N \to \infty$. But

$$\frac{1}{N}\sum_{n=1}^N w_n = \frac{1}{N}\sum_{n=1}^N (\sum_{j=1}^n \frac{x_j}{j}) = \frac{1}{N}\sum_{j=1}^N \sum_{n=j}^N \frac{x_j}{j} = \frac{1}{N}\sum_{j=1}^N (N-j+1)\frac{x_j}{j} = \frac{N+1}{N}\sum_{j=1}^N \frac{x_j}{j} - \frac{1}{N}\sum_{j=1}^N x_j.$$

Hence, one indeed gets that

$$\frac{1}{N}\sum_{j=1}^N x_j = \frac{N+1}{N} w_N - \frac{1}{N}\sum_{n=1}^N w_n \to w - w = 0.$$

□

As a prelude to the proof, let us outline an idea that does however not quite work in the general case: Notice that it is sufficient to consider the case where $(X_n)_{n\geq 1}$ is a sequence of i.i.d. random variables in L^1 such that $E[X_1] = 0$ (for the general case, one can then just add a constant m to each of the X_n). If one could show, for instance by the three-series theorem, that $\sum_{j=1}^n X_j/j$ converges almost surely as $n \to \infty$, then one could directly conclude via Kronecker's lemma. The convergence on $\sum_{j\geq 1} X_j/j$ does however not quite hold in general (see the remark after the statement of the "three-series theorem part II" in the next section, and the corresponding exercise), which is why we perform the first preliminary steps in the subsequent proof.

PROOF OF THE STRONG LAW OF LARGE NUMBERS. We first note that $\sum_{n\geq 1} P[|X_n| > n] = \sum_{n\geq 1} P[|X_1| > n] \leq E[|X_1|]$ is finite because $E[|X_1|] < \infty$. So, by Borel-Cantelli, the number of values of n for which $|X_n| > n$ is almost surely finite. It is therefore sufficient to show that $(X_1' + \ldots + X_n')/n$ converges almost surely to $E[X_1]$, when $X_n' = X_n 1_{|X_n|\leq n}$.

By dominated convergence, we know that $E[X'_n] = E[X_1 1_{|X_1|\leq n}] \to E[X_1]$ as $n \to \infty$. Taking the Césaro mean, we therefore get that

$$\frac{1}{n}\sum_{j=1}^{n} E[X'_j] \to E[X_1].$$

It is therefore sufficient to prove that $(Y'_1 + \ldots + Y'_n)/n$ converges almost surely to 0 where $Y'_j := X'_j - E[X'_j]$. We will prove this by establishing that the series of independent bounded and centered random variables $\sum_{j=1}^{n} Y_j/j$ converges almost surely as $n \to \infty$ (which allows us to conclude using Kronecker's lemma). We can note that

$$E[(Y'_n)^2] = \mathrm{Var}(X'_n) \leq E[|X_1|^2 1_{|X_1|\leq n}] \leq \sum_{j=1}^{n} j^2 P[|X_1| \in (j-1, j]].$$

Summing over n, we get that for some constant C,

$$\begin{aligned}\sum_{n\geq 1} E[(Y'_n/n)^2] &\leq \sum_{j\geq 1} \big(j^2 P(|X_1| \in (j-1,j])(\sum_{n\geq j} n^{-2})\big) \\ &\leq C\sum_{j\geq 1} jP(|X_1| \in (j-1,j]) \leq CE[|X_1|+1] < \infty.\end{aligned}$$

Hence, we can use the convergence of series of L^2 centered independent random variables theorem to conclude that the series $\sum_{j=1}^{n}(Y'_j/j)$ does converge indeed almost surely as $n \to \infty$. By Kronecker's lemma, we then get that $(Y'_1 + \ldots + Y'_n)/n$ converges almost surely to 0. □

REMARK 3.5.4. *We will give another proof of this law of large numbers using inverse martingales. We will then also show that under the same conditions one also has that* $E[|S_n/n - E[X_1]|] \to 0$*, i.e., that* S_n/n *converges also in* L^1 *to* $E[X_1]$*.*

REMARK 3.5.5. *There exist various extensions or generalizations of the strong law of large numbers. It is for instance possible to replace the condition that the* $(X_n)_{n\geq 1}$ *are independent and identically distributed (and in* L^1*) by the weaker condition that there are pairwise independent and identically distributed (and in* L^1*). This is known as Etemadi's law of large numbers – its proof follows another route than building on series of independent random variables. See again the exercise sheet.*

3.6. The conditions in the three series theorem are necessary

To conclude this chapter, let us return to the study of series of independent random variables, and discuss whether the conditions in the three-series theorem are in fact also necessary. This section is not really central in the construction of these lectures, but it is of course interesting to see that the conditions in Kolmogorov's three series theorem are in fact optimal.

THEOREM 3.6.1 (Three series theorem, part II). *Suppose that $(X_n)_{n\geq 1}$ is a sequence of independent random variables such that almost surely, $S_N := \sum_{n=1}^N X_n$ converges as $N \to \infty$ to a (finite) random variable. Then, for all $a > 0$, the following three series converge: (i) $\sum_{n\geq 1} P[|X_n| > a] < \infty$, and if $Y_n := X_n 1_{|X_n|\leq a}$, then (ii) $\sum_{n=1}^N E[Y_n]$ converges as $N \to \infty$ and (iii) $\sum_{n\geq 1} Var[Y_n] < \infty$.*

We can note that by the $0-1$ law, the convergence of S_N has probability 0 or 1. So, the condition in the theorem could be weakened to "if the probability that S_N converges in $\mathbb{R}$ is positive".

REMARK 3.6.2. *One side consequence of this result is that if $(X_n)_{n\geq 1}$ is a sequence of independent random variables, the fact that the three conditions (i), (ii), (iii) holds for some $a > 0$ implies that they actually hold for all a (one can first deduce the convergence of the stochastic series using the part I of the three series theorem, and then deduce the convergence for all a by this part II).*

REMARK 3.6.3. *This result can for instance be used as follows: If one of the three series described in (i), (ii) or (iii) does diverge, then we know that almost surely, S_N does not converge as $N \to \infty$. This for instance allows to construct an example of i.i.d. random variables $(U_n)_{n\geq 1}$ in L^1 with $E[U_1] = 0$ such that $\sum_{n\geq 1} U_n/n$ does almost surely not converge (one applies the above result for $X_n = U_n/n$ by checking for instance that the second series diverges). See the exercise sheet.*

Before proving the theorem, let us first show the following proposition, which is in a way focusing on the third condition above, or can be viewed as exploring how the L^2 condition for series of independent bounded centered random variables is also necessary:

PROPOSITION 3.6.4. *Suppose that $(Z_n)_{n\geq 1}$ is a sequence of independent random variables such that (a) there exists $a > 0$ such that almost surely, for all $n \geq 1$, $|Z_n| \leq a$, (b) for all $n \geq 1$, $E[Z_n] = 0$ and (c) $\sum_{n\geq 1} E[Z_n^2] = \infty$, then almost surely, $\sum_{n=1}^N Z_n$ does not converge as $N \to \infty$.*

REMARK 3.6.5. *Another way to phrase this result is that if $(Z_n)_{n\geq 1}$ is a sequence of independent* bounded *random variables with $E[Z_n] = 0$ such that $\sum_{n=1}^N Z_n$ converges almost surely, then necessarily $\sum_{n\geq 1} E[Z_n^2] < \infty$.*

PROOF. Let $S_n = Z_1 + \ldots + Z_n$. We note that

$$E[(S_n)^4] = E[((\sum_{j=1}^n Z_j)^4] = \sum_{j_1,j_2,j_3,j_4\leq n} E[Z_{j_1} Z_{j_2} Z_{j_3} Z_{j_4}].$$

We can use the very same argument as in the proof of the "easy" version of the law of large numbers (noting that $E[Z_{j_1} Z_{j_2} Z_{j_3} Z_{j_4}] = 0$ as soon as one of the indices among $j_1, \ldots, j_4$ appears only once because of the independence between the centered variables Z_j). This shows that

$$E[(S_n)^4] = \sum_{j=1}^n E[Z_j^4] + 6 \sum_{j<j'\leq n} E[Z_j^2] E[Z_{j'}^2]$$

(the factor 3 comes from the numbers of ways to choose j_1, j_2, j_3, j_4 so that two of them are j and two of them are j') so that

$$E[(S_n)^4] \leq (\sum_{j=1}^{n} E[Z_j^2])(a^2 + 3\sum_{j=1}^{n} E[Z_j^2]).$$

For n_1 large enough, we know that $a^2 \leq \sum_{j=1}^{n_1} E[Z_j^2]$ (because of the condition (iii)), so that that for such a choice of n_1,

$$E[S_{n_1}^4] \leq 4E[S_{n_1}^2]^2.$$

By the Paley-Zygmund inequality applied to $X = S_{n_1}^2$, this implies that

$$P(S_{n_1}^2 \geq E[S_{n_1}^2]/4) \geq 1/16.$$

In particular, since $E[S_{n_1}^2] = \sum_{j=1}^{n_1} E[Z_j^2] \geq a$,

$$P(|S_{n_1}| \geq a/2) \geq 1/16.$$

Once this n_1 is chosen, we can now apply the same result to the sequence $(Z_{n_1+n})_{n\geq 1}$ instead of the sequence $(Z_n)_{n\geq 1}$, and we get that there exists n_2 wuch that

$$P[|S_{n_2} - S_{n_1}| \geq a/2] \geq 1/16.$$

We now then actually choose iteratively an increasing deterministic sequence $(n_k)_{k\geq 1}$ such that for all k,

$$P[|S_{n_{k+1}} - S_{n_k}| \geq a/2] \geq 1/16.$$

We can finally note that the random variables $(S_{n_{k+1}} - S_{n_k})_{k\geq 1}$ are independent, so that by the second Borel-Cantelli lemma, we can conclude that almost surely, $|S_{n_{k+1}} - S_{n_k}| \geq a/2$ for infinitely many values of k, which indeed implies that almost surely, S_n does not converge as $n \to \infty$. □

We can now prove the more general result:

Proof of the theorem. Let us suppose that S_N converges almost surely. This implies in particular that $X_n \to 0$ almost surely, so that for any given $a > 0$, there almost surely exists n_0 such that $|X_n| \leq a$ for all $n \geq n_0$. The second Borel-Cantelli lemma then shows that necessarily $\sum_n P[|X_n| > a] < \infty$ (since otherwise, we would get that $|X_n| > a$ for infinitely many n's). So, condition (i) necessarily holds. By the first Borel-Cantelli lemma, this then shows that $Y_n = X_n$ for all but finitely many n's, which in turn implies that $\sum_{n=1}^{N} Y_n$ does also converge almost surely as $N \to \infty$.

We can now use the following little trick: Let us consider on some (new) probability space, a collection $(Y_1', Y_1'', Y_2', Y_2'', \ldots)$ of independent random variables such that for all n, Y_n' and Y_n'' both have the same law as Y_n. Then, define $Z_n = Y_n' - Y_n''$ and

$$\tilde{S}_n = (Y_1' + \ldots + Y_n') - (Y_1'' + \cdots + Y_n'') = Z_1 + \ldots + Z_n.$$

We know that with probability 1, both $Y_1' + \cdots + Y_n'$ and $Y_1'' + \cdots + Y_n''$ do converge almost surely as $n \to \infty$, so that $\tilde{S}_n$ does almost surely converge as well. We note that $E[Z_j] = E[Y_j'] - E[Y_j''] = 0$, that

$$E[(Z_j)^2] = \mathrm{Var}(Y_j' - Y_j'') = \mathrm{Var}(Y_j') + \mathrm{Var}(Y_j'') = 2\mathrm{Var}(Y_j)$$

and that for each j, $|Z_j| \leq 2a$ almost surely. We can then apply the previous lemma to deduce that $\sum_{n\geq 1} \mathrm{Var}(Y_n' - Y_n'') = 2\sum_{n\geq 1} \mathrm{Var}(Y_n)$ is finite.

We can then apply the actual three-series theorem to the sequence of independent random variables $(Y_n - E[Y_n])_{n\geq 1}$ (there are all bounded and centered, and we have just proved that the sum of the variances is finite) to deduce that almost surely, the sum $\sum_{n=1}^{N}(Y_n - E[Y_n])$ converges as

$N \to \infty$. But since $\sum_{n=1}^{N} Y_n$ also converges almost surely, we finally conclude that the deterministic series $\sum_{n=1}^{N} E[Y_n]$ converges as $N \to \infty$. $\square$

CHAPTER 4

Conditional expectation, martingales and their (a.s.) convergence

4.1. Conditional expectation

4.1.1. Warm-up. One motivation for what follows goes as follows: Consider a probability space $(\Omega, \mathcal{A}, P)$ and a non-decreasing sequence $(\mathcal{F}_n)_{n\geq 0}$ of sub-σ-fields of $\mathcal{A}$. Intuitively, one can for instance think of $\mathcal{F}_n$ to be the collection of events that are "observable" if one knows all what has happened up to time n.

In what follows, we will fix n and write $\mathcal{G} = \mathcal{F}_n$. One general intuitive idea is that given the information about what has happened up to time n, one should be able to construct the conditional probability given what one has observed so far of any event $B \in \mathcal{A}$. In some sense, it is the "new updated" probability that B holds. There are two ways to try to make this idea work:

- The "ambitious" route that we will not follow, is to try to define the "conditional probability measure given $\mathcal{G}$", which would be a $\mathcal{G}$-measurable function from Ω into the set of probability measure on $(\Omega, \mathcal{A})$.

  In the case of a countable state space, this is actually not difficult to do – indeed, $\mathcal{G}$ then corresponds to the decomposition of Ω into at most countable many disjoint sets $(\Omega_j)_{j\in J}$ (i.e., in $\mathcal{G}$, one can basically only observe in which Ω_j the random outcome is). Then, for each j such that $P(\Omega_j) > 0$, one can define the conditional probability measure P_j given Ω_j to be $P_j(B) = P(B|\Omega_j) = P(B \cap \Omega_j)/P(\Omega_j)$ for all $B \in \mathcal{A}$. We then define the conditional probability given $\mathcal{G}$ to be the map that to each $\omega \in \Omega_j$ associates $P_\omega := P_j$. We can note already that P_j is not well-defined when $P(\Omega_j) = 0$, i.e., this map $\omega \mapsto P_\omega$ is defined up to zero-probability events only.

  In the general case, it is a somewhat more tricky business to define such conditional probability measures (to start with, one would have to define the σ-field on the space of probability measures on $\mathcal{A}$, to make sense of the $\mathcal{F}_n$ measurability of the conditional probability measure).
- The "pragmatic" route, which is the one that we will describe here is to slightly reduce our ambition here, and instead construct, for each given $B \in \mathcal{A}$ separately, the conditional probability of B given $\mathcal{G}$ as a real-valued random variable that is measurable with respect to $\mathcal{G}$ (so that one knows this random number given the information at time n), and that intuitively speaking represents the "new updated probability" of B occuring, given the information that is available up to time n. So in some sense, one knows what happened before time n and averages out only over the randomness that occurs after time n. As opposed to the "ambitious" route, we do not worry about whethere looking "at all B's simultaneously" one defines a conditional measure. We only define, for each given fixed B, a random variable $Y := P(B|\mathcal{G})$.

  In the case of the discrete state-space discussed above, then Y will be a random variable that is constant and equal to $P(B|\Omega_j)$ on each Ω_j such that $P(\Omega_j) \neq 0$ (as we will see later, on the other Ω_j's where $P(\Omega_j) = 0$, we do not really care, as the random variable will be defined up to "almost sure identity"). We can note that this random variable Y satisfies the following properties: (a) It is $\mathcal{G}$-measurable, it takes its values in $[0, 1]$, and

(b) for any $A \in \mathcal{G}$ (so that $\mathcal{G}$ is the disjoint union of some Ω_j's),

$$E[Y1_A] = \sum_{j\ :\Omega_j \subset A} E[Y1_{\Omega_j}] = \sum_{j\ :\Omega_j \subset A} P(B \cap \Omega_j) = P(B \cap A).$$

In the general case, where the state space is not necessarily countable, this random variable Y should have property (a), and it is natural to expect that it should also satisfy the property (b') that for all $A \in \mathcal{G}$,

$$E[Y1_A] = P(B \cap A).$$

In some sense, this equation should be understood by computing $P(B \cap A)$ by first integrating out the randomness that occurs after time n, when one knows all the information available in $\mathcal{G}$ – which gives rise to $Y1_A$, and then integrating out the randomness that occurs up to time n.

As we shall see in the next sections, the properties (a) and (b') do completely characterize the random variable Y (up to its values on zero-probability events). So, it will indeed be possible to define these conditional probabilites $P(B|\mathcal{G})$ for each B separately.

A very slightly more general setup, is for each random variable $X \in L^1$ (that now replaces 1_B), to define the conditional expectation of X given the sub-σ-algebra $\mathcal{G}$ of $\mathcal{A}$. This will now be the "new updated expected value" of X given $\mathcal{G}$. The next sections will be devoted to the definition and properties of the conditional expectation of a random variable $X \in L^1$, given a sub-σ-algebra $\mathcal{G}$ of $\mathcal{A}$.

4.1.2. Definition. After this warm-up, we are ready for the actual formal mathematical definitions and constructions!

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space. Throughout this section we will suppose that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{A}$. We can recall that when a random variable is $\mathcal{G}$-measurable, then it is automatically also $\mathcal{A}$-measurable.

Throughout this section X will denote an $\mathcal{A}$-measurable real-valued random variable such that $E[|X|] < \infty$.

Definition 4.1.1. *We say that Y is a version of the conditional expectation of X given $\mathcal{G}$ if Y is an L^1 random variable that is measurable with respect to $\mathcal{G}$ such that for any $A \in \mathcal{G}$, $E[X1_A] = E[Y1_A]$.*

Remark 4.1.2. *A big warning: A conditional expectation is in fact a random variable!*

Remark 4.1.3. *This property corresponds to the intuitive idea that Y is the meaned out value over all possible outcomes of X that are available given the observation one does in $\mathcal{G}$.*

Let us first check that this definition and notation makes sense, i.e., that:

Lemma 4.1.4. *If Y and Y' are two versions of the conditional expectation of X given $\mathcal{G}$, then $Y = Y'$ almost surely.*

Proof. We can choose the event $A := \{Y > Y'\}$ which is $\mathcal{G}$ measurable, so that

$$E[(Y - Y')1_{Y-Y'>0}] = E[Y1_A] - E[Y'1_A] = E[X1_A] - E[X1_A] = 0.$$

Since $Y - Y' > 0$ on A, we get that $(Y - Y')1_A$ is a non-negative random variable with zero expectation – which is therefore almost surely equal to 0. The only way for this to hold is that $P(A) = 0$.

Inverting the roles of Y and Y', we similarly see that $P(Y' > Y) = 0$, so that we can indeed conclude that $Y = Y'$ almost surely. □

REMARK 4.1.5. *We therefore see that the conditional expectation is defined up to the equivalence relation given by the almost sure identity between random variables. We therefore also write that "$Y = E[X|\mathcal{G}]$ almost surely" to say that Y is a version of the conditional expectation of X given $\mathcal{G}$. We can note that replacing random variables by the equivalence class for this equivalence relation is the standard procedure when one discusses spaces of measurable functions, for instance L^2 spaces. For instance, when one wants to endow an L^2 with its Hilbert space structure, one needs that $E[X^2] = 0$ if and only $X = 0$ almost surely, i.e., if and only if the equivalence class of X is that of 0.*

Let us stress that at this point, we have not proved that such random variables exist! We will postpone the statement and proof of this fact to the next section. Before that, let us provide two equivalent definitions of the conditional expectation:

Suppose that π is a generating π-system of $\mathcal{G}$. Then, one has the more checkable version of the definition:

LEMMA 4.1.6. *If Y is a random variable in $L^1(\mathcal{G})$ such that for all $A \in \pi$, $E[Y1_A] = E[X1_A]$, then $Y = E[X|\mathcal{A}]$ a.s.*

PROOF. It is again a direct consequence of Dynkin's lemma: Let

$$\mathcal{U} := \{A \in \mathcal{G} : \ E[1_A Y] = E[1_A X]\}.$$

It is immediate to check that it is a Dynkin system, and by assumption, it contains π. It therefore contains $\sigma(\pi) = \mathcal{G}$, which in turn implies that $Y = E[X|\mathcal{G}]$. □

At the other end, we have that:

LEMMA 4.1.7. *If $Y = E[X|\mathcal{G}]$ a.s., then for all bounded $\mathcal{G}$-measurable random variable Z, $E[YZ] = E[XZ]$.*

PROOF. It suffices to prove this for non-negative bounded random variable Z (the general case then follows by decomposing Z into $Z1_{Z>0} - (-Z)1_{Z<0}$). When Z is a bounded non-negative random variable, we can consider its dyadic approximations $Z_n = \varphi_n(Z)$. For each n and j, the event $\{Z_n = j2^{-n}\}$ is $\mathcal{G}$-measurable, so that

$$j2^{-m}E[X1_{Z_n=j2^{-n}}] = j2^{-n}E[Y1_{Z_n=j2^{-n}}].$$

Summing over j, we get that $E[XZ_n] = E[YZ_n]$ and then finally, letting $n \to \infty$, we get by dominated convergence that $E[XZ] = E[YZ]$. □

We conclude this first section with the following three remarks that will be used in the forthcoming sections:

(1) If Y is a version of the conditional expectation of X given $\mathcal{G}$, then $E[|Y|] \leq E[|X|]$. [Indeed: the event $A = \{Y > 0\}$ is $\mathcal{G}$-measurable, so that

$$E[|Y|] = E[1_A Y] - E[1_{\Omega\setminus A}Y] = E[1_A X] - E[1_{\Omega\setminus A}X] \leq E[1_A|X|] + E[1_{\Omega\setminus A}|X|] = E[|X|]$$

which proves the claim].

(2) If $X \geq 0$ almost surely and $Y = E[X|\mathcal{G}]$ almost surely, then $Y \geq 0$ almost surely. [Indeed, we can choose $A = \{Y < 0\}$, which is $\mathcal{G}$-measurable, and note that $E[Y1_{Y<0}] = E[X1_A] \geq 0$, from which it follows that $Y \geq 0$ almost surely].

(3) Suppose that X' is another L^1 random variable measurable with respect to $\mathcal{A}$ such that $X \geq X'$. Suppose that $Y = E[X|\mathcal{G}]$ a.s. and $Y' = E[X'|\mathcal{G}]$ a.s., then $Y' \geq Y$ a.s. [Indeed, this implies that $Y - Y'$ is a version of the conditional expectation of $X - X'$ given $\mathcal{G}$ and we can apply the previous observation].

4.1.3. Construction. The goal of this section is to prove the existence of conditional expectation:

PROPOSITION 4.1.8. *For all $X \in L^1$ and all sub-σ-field $\mathcal{G}$ of $\mathcal{A}$, there exists Y such that $Y = E[X|\mathcal{G}]$ almost surely.*

PROOF. The proof will proceed in two steps:

(1) We first assume that X is in L^2, and we will use the fact that L^2 is actually a Hilbert space.

Here is first some heuristic: Let us recall that when U is an L^2 random variable, the value of m that minimizes $E[(U-m)^2]$ is $m = E[U]$ [Indeed, $E[(U - E[U] + x)^2] - x^2 = \mathrm{Var}(U - E[U] + x) = \mathrm{Var}(U)$]. Hence, it seems natural to guess that $E[X|\mathcal{G}]$ will be the $\mathcal{G}$-measurable Z random variable that minimizes $E[(X-Z)^2]$, which we know is the "orthogonal projection" of X on the subspace of L^2 fucntions that are measurable with respect to $\mathcal{G}$.

So, we consider the following two Hilbert spaces: $H_1 := L^2(\Omega, \mathcal{A}, P)$ and $H_2 := L^2(\Omega, \mathcal{G}, P)$. Recall that elements of such Hilbert spaces are equivalence classes of random variables modulo the equivalence relation $X \sim X'$ when $X = X'$ almost surely. The scalar product between the equivalence classes $\tilde{X}_1$ and $\tilde{X}_2$ of X_1 and X_2 is then $E[X_1 X_2]$. H_2 is a sub-space of H_1, so for each $\tilde{X}$ in H_1, it is possible to define its orthogonal projection $\tilde{Y}$ onto H_2, which is characterized by the fact that $\tilde{Y} \in H_2$ and that $\tilde{X} - \tilde{Y}$ is orthogonal to H_2, i.e., that for all $\tilde{Z} \in H_2$, $E[Z(X-Y)] = 0$ (where X, Y and Z are random variables in the corresponding equivalence classes).

In particular, choosing $Z = 1_A$, we get that $E[Y 1_A] = E[X 1_A]$ for all $A \in \mathcal{G}$, which shows that any element Y in the equivalence class of $\tilde{Y}$ is a version of $E[X|\mathcal{G}]$. This therefore completes the construction of the conditional expectation when X is in L^2.

(2) Next we consider the case where X is in L^1 and non-negative. For any positive integer K, we define the random variable $X_K := X 1_{|X|<K}$, which is bounded and therefore in L^2. We can therefore define for each K, a version Y_K of the conditional expectation of X given $\mathcal{G}$. Since $X_{K+1} \geq X_K$, it follows that $Y_{K+1} \geq Y_K$ almost surely. Hence, $(Y_K)_{K \geq 1}$ is a non-decreasing sequence of non-negative random variables that therefore converges to some radom variable Y with values in $[0, \infty]$. For any $A \in \mathcal{G}$, we know that $E[Y_K 1_A] = E[X_K 1_K]$, and letting $K \to \infty$, we can apply the monotone convergence theorem to conclude that $E[Y 1_A] = E[X 1_A]$. When $A = \Omega$, we see that $E[Y] = E[X] < \infty$, so that Y is indeed a version of the conditional expectation of X given $\mathcal{G}$ – i.e., the conditional expectation of X given $\mathcal{G}$ exists.

Finally when X is in L^1 but can also be negative, we simply define $Y = Y_+ - Y_-$ where $Y_+ = E[X 1_{X>0}|\mathcal{G}]$ a.s. and $Y_- = E[-X 1_{x<0}|\mathcal{G}]$ a.s., and we check that Y is a version of the conditional expectation of X given $\mathcal{G}$.

□

Hence, from now on, by a very slight abuse of notations, we can treat $E[X|\mathcal{G}]$ as a well-defined random variable (while it is in fact defined "up to zero probability events" – but this has no effect on any identity that one then writes and involves $E[X|\mathcal{G}]$ because one will either consider expectations of this random variable, or almost sure properties). We can therefore view $X \mapsto E[X|\mathcal{G}]$ as a map from $L^1(\Omega, \mathcal{A}, P)$ into $L^1(\Omega, \mathcal{G}, P)$ (because $E[|E[X|\mathcal{G}]|] \leq E[|X|]$).

We can note that the definition of conditional expectations shows immediately that this map in linear.

REMARK 4.1.9. *There exists also a direct justification of the existence of the conditional expectation using the Radon-Nikodym Theorem – but the hands-on construction that we provided is maybe more intuitive from a probabilistic perspective. Here is the outline of the Radon-Nikodym based construction:*

We first suppose that X is in L^1 and is almost surely non-negative. We define the finite measure μ on $(\Omega, \mathcal{G})$ by

$$\mu(A) = E[1_A X]$$

for all $A \in \mathcal{G}$ (note that this is also a measure on $(\Omega, \mathcal{A})$ but we won't need this here) – the fact that this is a measure follows from instance from the monotone convergence theorem). This measure is absolutely continuous with respect to the probability measure P (on $(\Omega, \mathcal{G})$) since $\mu(A) = 0$ as soon as $P(A) = 0$. The Radon-Nikodym theorem then shows the existence of a random variable $Y \in L^1(\Omega, \mathcal{G}, P)$ such that $d\mu/dP = Y$, i.e., such that for all $A \in \mathcal{G}$, $\mu(A) = E[1_A Y]$. This random variable Y is therefore a version of the conditional expectation of X given $\mathcal{G}$.

The general case follows by decomposing X into the difference between $X1_{X>0}$ and $-X1_{X<0}$.

4.1.4. Simple properties. Let us now list and prove a few simple properties of such conditional expectations (for notational convenience, we will sometimes write m_X for the expectation of X):

(1) If $\mathcal{G}$ is a trivial σ-field (where all events have probability in $\{0, 1\}$, then $E[X|\mathcal{G}] = E[X]$ almost surely. [Indeed, if we write (to avoid notational confusion) $m_X = E[X]$, then for all A with $P(A) = 0$, one anyway has $E[Y1_A] = 0 = E[m_X 1_A]$ and for A with $P(A) = 1$, one anyway has $E[X1_A] = E[X] = E[m_X 1_A]$].

(2) If $\mathcal{G}$ is the entire σ-field $\mathcal{A}$, and $Y = E[X|\mathcal{G}] = X$ almost surely. [Indeed, one obviously has $E[X1_A] = E[X1_A]$ and X is $\mathcal{A}$ measurable].

(3) If the σ-field $\mathcal{H}$ is independent of $\sigma(X)$, then $E[X|\mathcal{H}] = E[X]$ almost surely. [Indeed, the independence between X and the random variable 1_A when $A \in \mathcal{H}$ ensures that $E[X1_A] = E[X]E[1_A] = E[m_X 1_A]$].

(4) When Z is a bounded $\mathcal{G}$-measurable random variable, then $E[ZX|\mathcal{G}] = ZE[X|\mathcal{G}]$. [Indeed, for all $A \in \mathcal{G}$, the random variable $Z1_A$ is also a bounded $\mathcal{G}$-measurable random variable so that

$$E[1_A ZX] = E[1_A ZE[X|\mathcal{G}]],$$

and since $ZE[X|\mathcal{G}]$ is in L^1 and $\mathcal{G}$-measurable as well, we can conclude that it is almost surely equal to $E[XZ|\mathcal{G}]$].

(5) If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$ and if $Y = E[X|\mathcal{G}]$ almost surely, then $E[Y|\mathcal{H}] = E[X|\mathcal{H}]$ almost surely. [Indeed, for all $A \in \mathcal{H}$, it is also in $\mathcal{G}$, so that

$$E[1_A Y] = E[1_A X] = E[1_A E[X|\mathcal{G}]]$$

from which the statement follows].

(6) If $\mathcal{G}$ and $\mathcal{H}$ are two sub-σ-algebras of $\mathcal{A}$ and if $\sigma(\sigma(X) \cup \mathcal{G})$ is independent of $\mathcal{H}$, then

$$E[X|\sigma(\mathcal{G} \cup \mathcal{H})] = E[X|\mathcal{G}]$$

almost surely. [Note that this generalizes the statement (3) – since when $\mathcal{G}$ is the trivial σ-field, $E[X|\mathcal{G}] = E[X]$ almost surely. To prove this more general statement, we can note that a π-system generating $\sigma(\mathcal{G} \cup \mathcal{H})$ is the collection of sets $A \cap A'$ where $A \in \mathcal{G}$ and $A' \in \mathcal{H}$, and to see that for such a set,

$$\begin{aligned} E[1_{A\cap A'} X] &= E[1_{A'} 1_A X] = E[1_{A'}]E[1_A X] \\ &= E[1_{A'}]E[1_A E[X|\mathcal{G}] = E[1_{A'} 1_A E[X|\mathcal{G}]] = E[1_{A\cap A'} E[X|\mathcal{G}]], \end{aligned}$$

using the independence of $\mathcal{H}$ with $\sigma(X, \mathcal{G})$ twice and the fact that $1_A E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable and therefore independent of $1_{A'}$.]

4.1.5. Generalizations of the properties of the expectation. We have just seen that in the special case where $\mathcal{G}$ is the trivial σ-field, then $E[X|\mathcal{G}] = E[X]$ almost surely. So, in some sense, the conditional expectation is a generalization of the usual expactation. As we are now going to see, all the convergence type theorems (monotone convergence, dominated convergence) as well as Jensen's inequality can be generalized very easily to conditional expectations. All the proofs will be direct consequences of the corresponding results for "usual" expectations and the definition of conditional expectation.

For instance (we always work here in some probability space $(\Omega, \mathcal{A}, P)$, and $\mathcal{G}$ is a sub-σ-field of $\mathcal{A}$):

PROPOSITION 4.1.10 (Monotone convergence for conditional expectations). *Suppose that X_n is a non-decreasing family of non-negative random variables that are all in L^1 and that converge to X that is also in L^1. Suppose that $\mathcal{G}$ is a sub-σ-field of $\mathcal{A}$. Then almost surely, $E[X_n|\mathcal{G}]$ converges to $E[X|\mathcal{G}]$.*

PROOF. The sequence of non-negative random variables $E[X_n|\mathcal{G}]$ is (almost surely) not decreasing by monotonicity of the conditional expectation, so that is almost surely converges to some non-negative $\mathcal{G}$ measurable random variable Y (on the event where convergence does not hold, one can just decide that $Y = 0$). For all $A \in \mathcal{G}$, the sequences $(1_A X_n)_{n\geq 1}$ and $(1_A E[X_n|\mathcal{G}])_{n\geq 1}$ are almost surely non-decreasing sequences of non-negative random variables, so that by the usual monotone convergence theorem,

$$E[1_A X_n] \to E[1_A X] \text{ and } E[1_A E[X_n|\mathcal{G}]] \to E[1_A Y]$$

Since $E[1_A X_n] = E[1_A E[X_n|\mathcal{G}]$ for each $n \geq 1$, we conclude that the limits are the same, i.e., that $E[1_A X] = E[1_A Y]$. In the special case where $A = \Omega$, this ensures that $Y \in L^1$ and concludes the proof. □

Next comes the dominated convergence theorem:

PROPOSITION 4.1.11 (Dominated convergence theorem for conditional expectations). *Suppose that X_n is a sequence of random variables that converges almost surely to X, and that there exists a random variable Z in L^1 such that for all $n \geq 1$, $|X_n| \leq Z$ almost surely. Then $E[X_n|\mathcal{G}]$ converges almost surely to $E[X|\mathcal{G}]$.*

PROOF. It suffices to check that $E[X_n - X|\mathcal{G}] \to 0$, and since

$$|E[X_n - X|\mathcal{G}]| \leq E[|X_n - X|\ |\mathcal{G}] \leq E[\sup_{k\geq n} |X_k - X|\ |\mathcal{G}],$$

it suffices to check that for $X'_n = \sup_{k\geq n} |X_k - X|$, $E[X'_n|\mathcal{G}] \to 0$ almost surely.

The sequence $(2Z - X'_n)_{n\geq 0}$ is a non-decreasing sequence of non-negative random variables that converges to $2Z$ which is in L^1, so that we can apply the monotone convergence theorem for conditional expectations, that says that

$$E[2Z - X'_n|\mathcal{G}] \to E[2Z|\mathcal{G}]$$

almost surely. By linearity of the conditional expectation, we deduce that indeed $E[X'_n|\mathcal{G}] \to 0$. □

Finally, let us state and prove the conditional version of Jensen's inequality:

PROPOSITION 4.1.12 (Jensen's inequality for conditional expectations). *If X is a random variable in L^1, if φ is a convex function on an interval I of $\mathbb{R}$ that contains $X(\Omega)$, and if $\varphi(X) \in L^1$, then*

$$\varphi(E[X|\mathcal{G}]) \leq E[\varphi(X)|\mathcal{G}]$$

almost surely.

PROOF. A first remark is that by monotonicity, it is easy to check that $E[X|\mathcal{A}] \in I$ almost surely, so that $\varphi(E[X|\mathcal{G}])$ is well-defined and that it is equal to the supremum of all $aE[X|\mathcal{G}] + b$ over all a and b such that $ax + b \leq \varphi(x)$ in I. It therefore suffices to show that for all any such a and b,

$$aE[X|\mathcal{G}] + b \leq E[\varphi(X)|\mathcal{G}].$$

But by linearity and monotonicity, for such a and b,

$$aE[X|\mathcal{G}] + b \leq E[aX + b|\mathcal{G}] \leq E[\varphi(X)|\mathcal{G}]$$

which proves this fact. □

4.2. Martingales, super-martingales, almost sure convergence criterion

4.2.1. Definition. We are now going to introduce a class of processes $(M_n)_{n\geq 0}$ (called martingales) that does generalize sums of independent random variables with zero expectation. The intuitive difference is that the law of the increment $(M_{n+1} - M_n)$ between time n and $n+1$ will be allowed to depend on the information available at time n – but whatever has happened before, the conditional expectation of this increment will be 0.

As an illustration, one can think of a player that at each step can choose to play (or not play) at some game in a casino (and there are several choices available), where all games are "fair" with zero average gain. Then, at time n, the player can decide to play or not, and also decide about the amount of money played and the type of game played, depending on what has happened before (or on the phone calls received from the outside world) etc. So, in this case, it is not the case anymore that the increment $M_{n+1} - M_n$ of the "wealth M_n of the player" between time n and $n+1$ is necessarily independent of $M_0, \ldots, M_n$.

To formalize this, one starts with a probability space $(\Omega, \mathcal{A}, P)$, and one first considers a filtration $(\mathcal{F}_n)_{n\geq 0}$ i.e., a non-decreasing family of sub-σ-fields on $\mathcal{A}$. One usually interprets n as the time, and $\mathcal{F}_n$ as the information available at time n.

When one has such a filtration, it is also natural to introduce $\mathcal{F}_\infty$ to be the filtration generated by $\cup_{n\geq 0}\mathcal{F}_n$ that is the information available "at infinite time" – we note that this σ-field can be smaller than $\mathcal{A}$ and that $\cup_{n\geq 0}\mathcal{F}_n$ is a generating π-system of $\mathcal{F}_\infty$.

Definition 4.2.1. *We say that the family of random variables $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ if*

(a) For each $n \geq 0$, M_n is an L^1 random variable that is $\mathcal{F}_n$-measurable (we sometimes write $M_n \in L^1(\mathcal{F}_n)$,

(b) For each $n \geq 0$, $M_n = E[M_{n+1}|\mathcal{F}_n]$ almost surely.

Note that since M_n is measurable with respect to $\mathcal{F}_n$, property (b) is equivalent to $E[M_{n+1} - M_n|\mathcal{F}_n] = 0$ a.s.

Remark 4.2.2. *The definition of a martingale is always with respect to some filtration. It is therefore a property of $((M_n)_{n\geq 0}, \mathcal{F}_{n\geq 0})$ and not of $(M_n)_{n\geq 0}$ alone.*

However, whenever $(M_n)_{n\geq 0}$ is a martingale with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$, one can define for each $n \geq 0$, the σ-field $\mathcal{A}_n := \sigma(M_0, \ldots, M_n)$. The filtration $(\mathcal{A}_n)_{n\geq 0}$ is called the "own" filtration of $(M_n)_{n\geq 0}$ – one can note that $\mathcal{A}_n \subset \mathcal{F}_n$, because one knows that $M_0, \ldots, M_n$ are all measurable with respect to $\mathcal{F}_n$. It is easy to see that $(M_n)_{n\geq 0}$ is also a martingale with respect to this new filtration.

Indeed, for each $n \geq 0$, M_n is in L^1, it is measurable with respect to $\mathcal{A}_n$, and

$$E[M_{n+1} - M_n|\mathcal{A}_n] = E[E[M_{n+1} - M_n|\mathcal{F}_n]|\mathcal{A}_n] = E[0|\mathcal{A}_n] = 0$$

almost surely. So, a martingale with respect to some filtration is always a martingale in its own filtration.

Definition 4.2.3. *We say that the family of random variables $(M_n)_{n\geq 0}$ is a super-martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ if it satisfies the same property (a) as martingales and if*

(b') For each $n \geq 0$, $M_n \geq E[M_{n+1}|\mathcal{F}_n]$ almost surely.

Simlarly, if instead, for each $n \geq 0$, $M_n \geq E[M_{n+1}|\mathcal{F}_n]$ a.s., then $(M_n)_{n\geq 0}$ is called a *submartingale* with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$.

Clearly, $(M_n)_{n\geq 0}$ is a submartingale with respect to some filtration if and only if $(-M_n)_{n\geq 0}$ is a supermartingale with respect to that filtration – and if $(M_n)_{n\geq 0}$ is both a submartingale and a supermartingale with respect to some filtration, then it is a martingale with respect to the same filtration.

REMARK 4.2.4. *So, the intuition is that for a submartingale, the average gain at each game is non-negative, while for a supermartingale is non-positive. "Submartingales tend to increase" while "supermartingales tend to decrease".*

Note that if $(M_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$, then for each $n \geq 0$,

$$M_n = E[M_{n+1}|\mathcal{F}_n] = E[E[M_{n+2}|\mathcal{F}_{n+1}]|\mathcal{F}_n] = E[M_{n+2}|\mathcal{F}_n]$$

almost surely, and by induction using the same argument, for each $n \geq 0$ and for each $k \geq 1$,

$$M_n = E[M_{n+k}|\mathcal{F}_n]$$

almost surely.

PROPOSITION 4.2.5. *Suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ and that φ is a convex function on an interval I such that almost surely, for all $n \geq 0$, $M_n \in I$ and $\varphi(M_n) \in L^1$. Then, $(\varphi(M_n))_{n\geq 0}$ is a submartingale with respect to this filtration.*

REMARK 4.2.6. *On the event (of zero probability) where $M_n \notin I$, one can for instance decide that $\varphi(M_n) = 0$, so that $\varphi(M_n)$ is a proper random variable.*

The proposition therefore shows in particular that:
- If $(M_n)_{n\geq 0}$ is a martingale then $(|M_n|)_{n\geq 0}$ is a submartingale.
- If $(M_n)_{n\geq 0}$ is a martingale that is in L^p for some $p > 1$, then $(|M_n|^p)_{n\geq 0}$ is a submartingale.

PROOF. It is a direct consequence of Jensen's inequality for conditional expectations. Indeed, we know by assumption that $Y_n := \varphi(X_n)$ is in L^1 and $\mathcal{F}_n$-measurable, and we have

$$\varphi(X_n) = \varphi(E[X_{n+1}|\mathcal{F}_n]) \leq E[\varphi(X_{n+1})|\mathcal{F}_n]$$

almost surely. □

REMARK 4.2.7. *Clearly, if $(X_n)_{n\geq 0}$ are independent random variables in L^1 such that for all $n \geq 0$, $E[X_n] = 0$, then $S_n := \sum_{j\leq n} X_j$ (with $S_0 = 0$) is a martingale with respect to the filtration $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. The question of the convergence of martingales is therefore a more general question than that of series of independent L^1 random variables with mean 0.*

We will note later that in fact, our strategy of proof of convergence for series of independent centered L^2 random variables can be adapted rather directly to derive a convergence criterion for martingales in L^2. However, we will start by deriving another (better) criterion that builds on somewhat different ideas.

REMARK 4.2.8 (Discrete stochastic calculus). *The main idea of our main result on martingales (the martingale convergence theorem that we will describe in the next section) relies on a very simple idea: Suppose that $(M_n)_{n\geq 0}$ is a discrete martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$ and that $(H_n)_{n\geq 0}$ is a bounded adapted process with respect to that filtration, i.e. that for each $n \geq 0$, H_n is a bounded $\mathcal{F}_n$-measurable random variable.*

Then, we can define a new process $(I_n)_{n\geq 0}$ by $I_0 = 0$ and for all $n \geq 1$,

$$I_n = \sum_{j=0}^{n-1} H_j(M_{j+1} - M_j).$$

One intuitive way to think about it is that at each time n, there is the opportunity to play a game, that will lead to an increment of $M_{n+1} - M_n$ (for which the conditional expected gain is zero, so the game is "fair") – but one can choose "how much one bets" (and to choose this, one can use all the information available in $\mathcal{F}_n$). So if $H_n = 2$, the gain or loss would be $I_{n+1} - I_n = 2(M_{n+1} - M_n)$, if $H_n = 0$, one does not play and nothing happens between time n and $n + 1$ so that $I_{n+1} = I_n$, while if $H_n = -1$, $I_{n+1} = I_n - (M_{n+1} - M_n)$.

The important simple observation is that this process is then also a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ *because each* I_n *is clearly in* L^1 *and* $\mathcal{F}_n$ *measurable, and*

$$E[I_{n+1} - I_n|\mathcal{F}_n] = E[H_n(M_{n+1} - M_n)|\mathcal{F}_n] = H_n E[M_{n+1} - M_n|\mathcal{F}_n] = 0$$

almost surely.

Similarly, if $(M_n)_{n\geq 0}$ *is a supermartingale and* H_n *is bounded, adapted and non-negative, then*

$$E[I_{n+1} - I_n|\mathcal{F}_n] = E[H_n(M_{n+1} - M_n)|\mathcal{F}_n] = H_n E[M_{n+1} - M_n|\mathcal{F}_n] \leq 0$$

almost surely, so that $(I_n)_{n\geq 0}$ *is also a supermartingale.*

We will use such constructions for a variety of processes $(H_n)_{n\geq 0}$*.*

4.2.2. The (super)-martingale convergence theorem. We are now going to prove a criterion that ensures that a martingale, submartingale or supermartingale converges. We start this section on a new page in order to highlight its importance!

DEFINITION 4.2.9. *We say that a family $(X_j)_{j\in J}$ of random variables is bounded in L^1 if $\sup_{j\in J} E[|X_j|] < \infty$.*

THEOREM 4.2.10. *If $(M_n)_{n\geq 0}$ is a supermartingale (with respect to some filtration) that is bounded in L^1, then M_n converges almost surely as $n \to \infty$.*

REMARK 4.2.11. *This immediately implies that a submartingale that is bounded in L^1 converges almost surely (since $-M_n$ converges a.s.), and that a martingale that is bounded in L^1 converges almost surely (since a martingale is always a supermartingale).*

REMARK 4.2.12. *The limit M_∞ is $\mathcal{F}_\infty$ then measurable and it is necessarily in L^1 since*

$$E[|M_\infty|] \leq \lim_{n\to\infty} E[|M_n|] \leq \sup_{n\geq 1} E[|M_n|] < \infty.$$

An immediate corollary is that:

COROLLARY 4.2.13. *If $(M_n)_{n\geq 0}$ is a non-negative supermartingale with respect to some filtration, then M_n converges almost surely to some finite random variable M_∞.*

Of course, this also shows in particular that:

COROLLARY 4.2.14. *A non-negative martingale does converge almost surely.*

PROOF OF THE COROLLARIES. This is simply due to the fact that when $(M_n)_{n\geq 0}$ is a non-negative supermartingale, then

$$E[|M_n|] = E[M_n] = E[E[M_n|\mathcal{F}_0]] \leq E[M_0]$$

so that $(M_n)_{n\geq 0}$ is necessarily bounded in L^1. □

We now turn to the proof of the theorem. Let us first make a few comments, and state the key lemma. The main observation is that a sequence of numbers $(x_n)_{n\geq 0}$ does converge in $[-\infty, \infty]$ if and only if for any pair of rational numbers $a < b$, the number $N_{a\to b}$ of its "upcrossings" from below b to above a is finite. More precisely, we can define, for each $a < b$, the sequence $\sigma_1 = \min\{n \geq 0 :\ x_n < a\}$ and for each $j \geq 1$,

$$\tau_j := \min\{n > \sigma_j :\ x_n > b\} \text{ and } \sigma_{j+1} := \min\{n > \tau_j :\ x_n < a\}$$

with the convention that $\min \emptyset = +\infty$. The intervals $[\sigma_j, \tau_j)$ are called the upcrossing intervals, and the total number of upcrossings $N_{a\to b}$ is the maximum value of j such that $\tau_j < \infty$. We can similarly define

$$N_{a\to b;n} := \max\{j \geq 1 :\ \tau_j \leq n\}$$

which is the total number of completed upcrossings before time n (this time with the convention that $\max \emptyset = 0$).

The reason for which the finiteness of the $N_{a\to b}$ is related to the convergence of x_n goes as follows:

- The set $[-\infty, \infty]$ is compact, so that x_n has at least one accumulation point in $[-\infty, \infty]$, and it converges if and only if it has exactly one accumumation point. [If it had (at least) to distinct accumulation points, then one could find two (finite) rationals $a < b$ that are (strictly) inbetween these two accumulation points, and then it is easy to check that $N_{a\to b}$ has to be infinite, since $M_n > b$ for infinitely many n's, and $M_n < a$ for infinitely many n's.]

- If the sequence had only one accumulation point, then necessarily, for any pair of rationals $a < b$, $N_{a\to b} < \infty$ because this accumulation point can not be both in $[-\infty, a]$ and in $[b, \infty]$.

The key lemma to prove the theorem is that:

LEMMA 4.2.15 (Bound on upcrossings). *If $(M_n)_{n\geq 0}$ is a supermartingale, then for all $a < b$, if $N_{a\to b;n_0}$ denotes the number of completed upcrossings from below a to above b by $(M_n)_{n\leq n_0}$ (before time n_0), then*

$$E[N_{a\to b;n_0}] \leq \frac{1}{b-a} E[(a - M_{n_0})1_{a>M_{n_0}}].$$

Before proving this lemma, let us explain why it implies the theorem:

PROOF OF THE CONVERGENCE THEOREM. We assume that $(M_n)_{n\geq 0}$ is a supermartingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ such that $K := \sup_n E[|M_n|] < \infty$. The inequality in the lemma in particular implies that for all pair of rationals $a < b$ and all $n_0 \geq 1$,

$$E[N_{a\to b;n_0}] \leq \frac{1}{b-a}(|a| + E[|M_{n_0}|]) \leq \frac{1}{b-a}(|a| + K).$$

Hence, since the non-decreasing sequence $N_{a\to b;n_0}$ almost surely converges to $N_{a\to b}$ (in $\mathbb{N}\cup\{\infty\}$), we get from the monotone convergence theorem (for non-negative random variables) that $E[N_{a\to b}] \leq (|a| + K)/(b - a)$, so that in particular, $N_{a\to b}$ is finite almost surely.

Hence (exchanging the a.s. and the "for all a, b"), almost surely, for any pair of rational numbers $a < b$, the completed number of upcrossings from below a to above b is finite, which implies by the above remark that almost surely, the sequence M_n does converges to some limit M_∞ in $[-\infty, +\infty]$. We note that $|M_n| \to |M_\infty|$ almost surely (as limits in $\mathbb{R} \cup \{+\infty\}$), so that by Fatou's Lemma,

$$E[|M_\infty|] = E[\lim_{n\to\infty} |M_n|] \leq \liminf_{n\to\infty} E[|M_n|] \leq K.$$

Hence, M_n does indeed converge almost surely towards a finite random variable. □

PROOF OF THE LEMMA. For each $n < n_0$, we let A_n denote the event that "n is part of an upcrossing interval", i.e., that for some $j \geq 1$,

$$\sigma_j \leq n < \tau_j.$$

Clearly, to determine if this is the case, it suffices to look at the sequence of numbers $M_0, \ldots, M_n$, so that A_n is $\sigma(M_0, \ldots, M_n)$ measurable (and therefore $\mathcal{F}_n$-measurable). Hence,

$$E[(M_{n+1} - M_n)1_{A_n}] = E[E[(M_{n+1} - M_n)1_{A_n}|\mathcal{F}_n]] = E[1_{A_n}E[(M_{n+1} - M_n)|\mathcal{F}_n]] \leq 0$$

because $(M_n)_{n\geq 0}$ is a supermartingale (one way to interpret this is to say that we use 1_{A_n} as our process H_n in order to construct a new supermartingale (I_n), just like in the final remark of the previous section). If we sum this up from $n = 0$ to $n_0 - 1$, we get

$$E[\sum_{n=0}^{n_0-1} (M_{n+1} - M_n)1_{A_n}] \leq 0.$$

We can regroup the terms in the sum by upcrossing intervals (this corresponds to write out what I_n is), and we then get (writing N for $N_{a\to b;n_0}$)

$$\sum_{n=0}^{n_0-1} (M_{n+1} - M_n)1_{A_n} = \sum_{j=1}^{N} (M_{\tau_j} - M_{\sigma_j}) + 1_{\sigma_{N+1}\leq n_0}(M_{n_0} - M_{\sigma_{N+1}}).$$

Since $M_{\tau_j} - M_{\sigma_j} > b - a$ for all $j \leq N$, and $M_{n_0} - M_{\sigma_{N+1}} \geq M_{n_0} - a$ if $n_0 \geq \sigma_{N+1}$, we get that

$$(b - a)E[N] + E[1_{\sigma_{N+1}\leq n_0}(M_{n_0} - a)] \leq 0,$$

which immediately shows that

$$(b-a)E[N] \leq E[(a-M_{n_0})1_{\sigma_{N+1}\leq n_0}] \leq E[(a-M_{n_0})1_{a>M_{n_0}}1_{\sigma_{N+1}\leq n_0}] \leq E[(a-M_{n_0})1_{a>M_{n_0}}]$$

and proves the lemma. □

REMARK 4.2.16. *We can recall that when X is a random variable in L^p for $p>1$, then by Jensen's inequality*

$$E[|X|] \leq E[|X|^p]^{1/p}$$

so that if a family $(X_i)_{i\in I}$ of random variables is bounded in L^p, then since

$$\sup_{i\in I} E[|X_i|] \leq \sup_{i\in I} E[|X_i|^p]^{1/p} < \infty,$$

it is also bounded in L^1. Hence, a martingale that is bounded in L^p does converge almost surely.

One can mention that the particular case $p=2$ applied to the very particular martingales that are series of independent centered random variables in L^2 gives the same result as the one we proved in the chapter on series of independent random variables. We will come back to L^p martingales for $p>1$ later in the lectures.

4.3. An example: Galton-Watson processes

Let us now provide examples of martingales that feel quite different to the series of independent random variables. These are related to the so-called Galton-Watson processes that models a random genealogical tree. We will then see what the previous martingale convergence theorem tells us for those particular examples.

The general Galton-Watson process is defined heuristically as follows. One is given a probability measure $\mathcal{L}$ on $\mathbb{N}$ and we suppose that $\mathcal{L}$ is not the Dirac mass at 1. This law represents the distribution of the number of children of a given individual A. We then start with one individual and make the assumption that the number of children of different individuals are all independent – we then define X_n to be the number of descendants of A at the n-generation. So, $X_0 = 1$, X_1 is the number of children of A, X_2 the number of grand-children etc.

Of course, this also models the spread of an epidemic, and the quantity R that we will see in a couple of paragraphs is nowadays quite famous!

To model this properly, we can for instance start with a collection of independent identically distributed random variables $(\xi_{n,j})_{n\geq 0, j\geq 1}$ that all have the law $\mathcal{L}$. The idea is to use these random variables iteratively with respect to n: If there are at least j descendants of A at the n-generation, then $\xi_{n,j}$ will be used to determine the number of children of the j-th individual of the n-th generation. In other words, we define $X_0 = 1$ and then iteratively for all $n \geq 0$,

$$X_{n+1} := \sum_{j: j\geq 1, j\leq X_n} \xi_{n,j}$$

(and if $X_n = 0$, then $X_{n+1} = 0$ because there is no term in the sum!).

Definition 4.3.1. *The collection $(X_n)_{n\geq 0}$ is called the Galton-Watson process with offspring distribution $\mathcal{L}$.*

It is then natural to define the filtration

$$\mathcal{F}_n := \sigma(\cup_{k<n} \cup_{j\geq 0} \sigma(\xi_{k,j})).$$

Then, the random variable X_n is indeed $\mathcal{F}_n$-measurable.

Let us first consider the case where the offspring distribution is in L^1, i.e., when $E[\xi] < \infty$ when ξ follows $\mathcal{L}$. We will denote this expectation by R. The Galton-Watson process is said to be subcritical, critical or supercritical when $R < 1$, $R = 1$ or $R > 1$ respectively.

Lemma 4.3.2. *When $(X_n)_{n\geq 1}$ is a Galton-Watson process $(X_n)_{n\geq 0}$ with integrable offspring distribution, then $(M_n := X_n/R^n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ (and therefore with respect to its own filtration).*

Proof. Indeed, we can check by induction that for all $n \geq 0$, X_n (and therefore M_n is in L^1 because

$$E[X_{n+1}] = E[\sum_{j\geq 1} 1_{j\leq X_n}\xi_{n,j}] = \sum_{j\geq 1} E[1_{j\leq X_n}\xi_{n.j}] = \sum_{j\geq 1} mE[1_{j\leq X_n}] = RE[X_n].$$

Furthermore,

$$\begin{aligned} E[X_{n+1}|\mathcal{F}_n] &= E[\sum_{j\geq 1} 1_{j\leq X_n}\xi_{n,j}|\mathcal{F}_n] = \sum_{j\geq 1} E[1_{j\leq X_n}\xi_{n,j}|\mathcal{F}_n] \\ &= \sum_{j\geq 1} 1_{j\leq X_n} E[\xi_{n,j}|\mathcal{F}_n] = \sum_{j\geq 1} 1_{j\leq X_n} E[\xi_{n,j}] = RX_n \end{aligned}$$

almost surely (here we successively used the monotone convergence theorem for conditional expectations, the fact that X_n was $\mathcal{F}_n$ measurable, the fact that $\xi_{n,j}$ is independent of $\mathcal{F}_n$ and finally the fact that $E[\xi_{n,j}] = R$). Hence, $E[M_{n+1}|\mathcal{F}_n] = M_n$ almost surely. □

Let us first see what the consequences of this convergence theorem are for critical or subcritical Galton-Watson processes:

- Then, since $(M_n)_{n\geq 0}$ is a non-negative martingale, the process $(X_n)_{n\geq 0}$ is a non-negative supermartingale. Hence, X_n almost surely converges to some finite random variable X_∞.
- Since X_n is a sequence of integers, this means that X_∞ is also integer-valued and that almost surely, there exists n_0 such that for all $n \geq n_0$, $X_n = X_\infty$.
- But we can notice that by the Borel-Cantelli lemma, for all $k \geq 1$, the set of n's such that $\sum_{j=1}^{k} \xi_{n,j} \neq k$ is almost surely infinite (because these events are independent and all have positive probability – this is where the assumption that $\mathcal{L}$ is not the Dirac mass at 1 is used), which shows that almost surely, $X_\infty \neq k$.
- So the only possibility is that $X_\infty = 0$ almost surely, i.e. that there exists n_0 such that $X_n = 0$ for all $n \geq n_0$.

We can therefore conclude with the following statement:

COROLLARY 4.3.3. *Subcritical and critical Galton-Watson genealogical trees is almost surely die out.*

This corollary on critical Galton-Watson processes can actually also be proved by rather direct means (we will come back to this later in these lectures) without using martingales,

However, the following corresponding result (that directly follows from the convergence theorem for the non-negative martingale $(M_n)_{n\geq 0}$) for supercritical Galton-Watson processes typically builds on martingales!

COROLLARY 4.3.4. *If $(X_n)_{n\geq 0}$ is a supercritical Galton-Watson process (i.e., with offspring distribution such that $R := E[\xi] > 1$), then X_n/R^n does almost surely converge to a finite limit M_∞.*

This raises the question of whether M_∞ is almost surely equal to 0 or not. Note that when M_∞ is not almost surely equal to 0, then, on the event where $\{M_\infty \neq 0\}$, one has

$$X_n \sim M_\infty R^n$$

as $n \to \infty$, i.e., the exponential growth of X_n. We will come back to this in the next chapter!

4.4. Inverse martingales

We illustrate the fact that this martingale convergence theorem and the ideas in its proof were quite powerful by showing how to deduce the strong law of large numbers from the upcrossing lemma in its proof.

4.4.1. Convergence of inverse martingales. We will start by defining a new class of processes, called *inverse martingales*.

We suppose now that $(\mathcal{A}_r)_{r\leq 0}$ is a filtration indexed by the *negative integers*. In other words, each $\mathcal{A}_r$ is a sub-σ-field on $\mathcal{A}$, and for each $r < 0$, $\mathcal{A}_r \subset \mathcal{A}_{r+1}$.

So, if one move "backwards" in time by taking $r = -n$ for $n \geq 0$, then the larger n is, the more information is lost. Another way to think about it is that we are looking at an evolution from time $-\infty$ to the present time 0, and that for all $r \leq 0$, $\mathcal{A}_r$ is the information available if one knows everything that happened up to time r.

We can then still define martingales with respect to such "inverse" filtrations – these are called "inverse martingales".

DEFINITION 4.4.1. *The process $(M_r)_{r\leq 0}$ is an inverse martingale with respect to $(\mathcal{A}_r)_{r\leq 0}$ if for all $r \leq 0$, M_r is in L^1 and if*

$$E[M_r|\mathcal{A}_{r-1}] = M_{r-1}$$

almost surely.

Again, there are two ways to think about them: either by looking at the time r coming from $-\infty$ to 0 (and in this case, one is just like in the usual martingale setting, for all $r < 0$, $M_r = E[M_{r+1}|\mathcal{A}_r]$ almost surely), or by looking at the evolution of M_{-n} when n goes from 0 to infinity, in which case, $M_{-(n+1)}$ is obtained by "partially averging out" the values of M_{-n} (corresponding to the idea that information gets lost).

We can note that if $(M_r)_{r\leq 0}$ is an inverse martingale with respect to $(\mathcal{A}_r)_{r\leq 0}$, then for all $r \leq -2$,

$$M_r = E[M_{r+1}|\mathcal{A}_r] = E[E[M_{r+2}|\mathcal{A}_{r+1}]|\mathcal{A}_r] = E[M_{r+2}|\mathcal{A}_r] = \cdots = E[M_0|\mathcal{A}_r]$$

almost surely.

Conversely, if X is any $\mathcal{A}_0$-measurable random variable in L^1, then the process $(E[X|\mathcal{A}_r])_{r\leq 0}$ is an inverse martingale with respect to $(\mathcal{A}_r)_{r\leq 0}$. So, a process is an inverse martingale if and only if it is of that particular form.

We now state a first convergence theorem for inverse martingales:

THEOREM 4.4.2. *An inverse martingale $(M_r)_{r\leq 0}$ does converge almost surely to some finite limit $M_{-\infty}$ as $r \to -\infty$.*

PROOF. For each $r_0 < 0$, one can define just as for usual martingales, the number $N(a \to b; [r_0, 0])$ to be the total number of upcrossings from below a to above b by the finite sequence $(M_{r_0}, M_{r_0+1}, \ldots, M_{-1}, M_0)$. Exactly as in Lemma 4.2.15, one then has

$$E[N(a \to b; [r_0, 0])] \leq \frac{1}{b-a} E[(a - M_0)1_{a>M_0}]$$

Similarly, $N(a \to b; [-n_0, 0])$ is non-decreasing with respect to n_0, and it therefore has a limit $N(a, b) := N(a \to b; (-\infty, 0])$ as $n_0 \to \infty$, which is the "total" number of upcrossings from below a to above b by the inverse martingale M.

The key point here is that, as opposed to the case of usual martingales, the bound on the right-hand side of the last displayed equation does anyway not depend on $r_0 = -n_0$, and is anyway

finite. So, we conclude by the monotone convergence theorem that $E[N(a \to b)]$ is finite, so that in particular, $N(a \to b)$ is almost surely finite.

We can then wrap up the proof as for the martingale convergence theorem: Almost surely, $N(a \to b)$ is finite for all pair of rational numbers $a < b$, from which it follows that almost surely, the sequence M_{-n_0} has at most one accumulation point $M_{-\infty}$ in $\mathbb{R} \cup \{-\infty, +\infty\}$ (so it almost surely converges to $M_{-\infty}$ as $n_0 \to \infty$. But $E[|M_{-n_0}|] \leq E[|M_0|] < \infty$, so that by Fatou's lemma, $E[|M_{-\infty}|] \leq E[|M_0|] < \infty$, and $M_{-\infty}$ is almost surely finite. □

4.4.2. Law of large numbers via inverse martingales. Let us now show how to deduce part of the strong law of large numbers from this statement: Consider a sequence of independent identically distributed random variables $X_1, \ldots, X_n, \ldots$ such that $X_1 \in L^1$. We let $S_n = X_1 + \ldots + X_n$ for $n \geq 1$ and $S_0 = 0$. We can then define for each $n \geq 0$,

$$\mathcal{A}_{-n} := \sigma(S_n, X_{n+1}, X_{n+2}, \ldots).$$

Clearly, since $S_{n+1} = S_n + X_{n+1}$, we see that $\mathcal{A}_{-n-1} \subset \mathcal{A}_{-n}$, so that $(\mathcal{A}_r)_{r \leq 0}$ is an inverse filtration.

We then *define* for each $n \geq 0$,

$$M_{-n} := E[X_1 | \mathcal{A}_{-n}]$$

(note also that $M_0 = X_1$ a.s.). This is clearly an inverse martingale, so that M_{-n} does converge almost surely as $n \to \infty$. But one the other hand, since $\sigma(S_n, X_1)$ is independent of $\sigma(X_{n+1}, \ldots)$, we see that for all $n \geq 1$,

$$M_{-n} = E[X_1 | \sigma(S_n, X_{n+1}, \ldots)] = E[X_1 | \sigma(S_n)]$$

almost surely.

We also note that by symmetry, for each $j \leq n$ and each $a \in \mathbb{R}$,

$$E[X_1 1_{S_n \leq a}] = E[X_j 1_{S_n \leq a}]$$

from which it follows that

$$E[X_1 | \sigma(S_n)] = E[X_j | \sigma(S_n)]$$

almost surely. Summing this up over j, we get that

$$nE[X_1 | \sigma(S_n)] = E[\sum_{j=1}^{n} X_j | \sigma(S_n)] = E[S_n | \sigma(S_n)].$$

So, we get that $S_n = nM_{-n}$, i.e., that $M_{-n} = S_n/n$.

We have therefore provided here another proof of the fact that S_n/n does almost surely converge to a finite random variable. This random variable is furthermore constant because of Kolmogorov's $0-1$ law. In the next chapter on uniform integrability, we will see how to deduce directly (without using our results on series of independent random variables or Kronecker's lemma) that the constant has to be $E[X_1]$ and also the fact (that we had not proven in the proof based on series of independent random variables) that $E[|S_n/n - E[X_1]|] \to 0$.

CHAPTER 5

Uniformly integrable martingales, optional stopping theorem

5.1. Uniform integrability

5.1.1. Different types of convergence of random variables. In this section, we will discuss/review the different ways in which one can say that a sequence $(X_n)_{n\geq 1}$ of random variables defined on a same probability space does converge to a random variable X as $n \to \infty$.

It is worth stressing that all the notions that we will discuss here are quite different from the notion of *convergence in law* that we will discuss later, and that deals with convergence of a sequence of laws of random variables, and not of their "realization" in a probability space.

The following types of convergence can be defined:

- The almost sure convergence, which is the one that we have (almost exclusively) used so far. One says that $X_n \to X$ almost surely, if the event $\{\omega \in \Omega : \ X_n \to X\}$ has probability 1.
- Convergence in probability: One says that X_n converges in probability to X if for any $\epsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.
- Convergence in an L^p space for $p \in [1, \infty)$. We say that X_n converges in L^p if $X \in L^p$ and if $E[|X_n - X|^p] \to 0$ as $n \to \infty$.

Let us immediately make a few remarks:

- The convergence in L^p corresponds to a convergence in a metric space. In other words, one views the random variable as an element in the space of measurable functions from Ω into $\mathbb{R}$, endowed with the L^p norm $\|X\|_p := E[|X|^p]^{1/p}$ and the convergence in L^p is the convergence in this space (recall also that when one looks at L^p spaces, one in fact considers equivalence classes of random variables with respect to the equivalence relation $X \sim Y$ if $X = Y$ almost surely).
- The same holds for the convergence in probability. A choice for the corresponding distance would be
$$d(X, Y) := E[\min(1, |Y - X|)]$$
(one could also take the equivalent distance $E[(X - Y)/(1 + |X - Y|)]$). In other words, X_n converges in probability to X if and only if $d(X_n, X) \to 0$.
- The almost sure convergence is of different type – it does not correspond to a notion of convergence in a metric space.

The convergence in probability is the weakest of these notions and it is implied by each one of the others. The diagram for $p' > p \geq 1$ is:

$$\text{CV in } L^{p'} \Rightarrow \text{CV in } L^p \Rightarrow \text{CV in } L^1 \Rightarrow \text{CV in probability} \Leftarrow \text{a.s. CV}$$

[To check the first three implications, one can note that for all $\varepsilon > 0$,

$$E[|X_n - X|^{p'}]^{1/p'} \geq E[|X_n - X|^p]^{1/p} \geq E[|X_n - X|] \geq \varepsilon P(|X_n - X| \geq \varepsilon)$$

because of Jensen's inequality and Markov's inequality. And if $X_n \to X$ almost surely, then for all $\varepsilon > 0$, almost surely,

$$\cup_{n_0 \geq 1} \cap_{n \geq n_0} \{|X_n - X| < \varepsilon\}$$

so that

$$P(|X_n - X| < \varepsilon) \geq P(\cap_{n \geq n_0} \{|X_n - X| < \varepsilon\}) \to 1$$

as $n_0 \to \infty$ from which the convergence in probability holds]. It is easy to see that none of these implications is an equivalence. In particular, convergence in probability does *not* imply almost sure convergence.

EXAMPLE 5.1.1 (The flying saucepans). *One can have the following "flying saucepans" example in mind: Consider the space* $[0,1]$ *endowed with the Borel* σ*-field and the Lebesgue probability measure. Then, for each* $k \geq 0$ *and* $j \in \{0, 1, \ldots, 2^k - 1\}$*, define*

$$f_{2^k+j}(x) = 1_{[j2^{-k},(j+1)2^{-k}]}.$$

Then f_n *tends in probability to* 0 *because when* $n = 2^k + j$*,* $\lambda[\{x : f_n(x) \neq 0\}] = 2^{-k} \to 0$ *as* $n \to \infty$*, but for each* x*, one can find infinitely many* n*'s for which* $f_n(x) = 1$*.*

The Borel-Cantelli lemma allows to derive the following:

PROPOSITION 5.1.2. *If* $(X_n)_{n \geq 1}$ *converges in probability to* X*, then there exists a deterministic sequence* $n_k \to \infty$*. such that* X_{n_k} *converges almost surely to* X*.*

REMARK 5.1.3. *For instance, in the flying saucepans example, one can take* $n_k = 2^k$*, and then* $f_{n_k} = 1_{[0,2^{-k}]}$*.*

PROOF. Convergence in probability means that for any fixed $\varepsilon > 0$, $P[|X_n - X| > \varepsilon] \to 0$ as $n \to \infty$. Let $n_0 = 0$ and then define iteratively, for each $k \geq 1$, n_k to be the first $n > n_{k-1}$ for which

$$P(|X_n - X| > 2^{-k}) \leq 2^{-k}$$

(we know that such an n_k exists because $P[|X_n - X| > 2^{-k}] \to 0$ as $n \to \infty$. For this choice of sequence $(n_k)_{k \geq 1}$, we see that

$$\sum_{k \geq 1} P(|X_{n_k} - X| > 2^{-k}) < \infty$$

so that by the first Borel-Cantelli lemma, we know that almost surely, for all large enough k, $|X_{n_k} - X| \leq 2^{-k}$, which implies the desired almost sure convergence. □

The core of the following two sections will be devoted to the following question: *What is missing for a sequence that converges in probability to converge in* L^1*?*

We will then see what consequences this has for martingales – in the previous chapter, we have established a criterion that ensures that ensures that a martingale $(M_n)_{n \geq 0}$ does converge almost surely to some random variable M_∞, and we will see under what conditions it also converges in L^1 (so that in particular, $E[M_\infty] = E[M_0]$).

5.1.2. Uniform integrability. The answer to the previous question is actually quite simple. If X_n tends to be closer and closer to X on a larger and larger portion of the probability space, then for $E[|X - X_n|]$ not to tend to 0, the contribution to this integral in the (very small) remaining part of the probability space has not to vanish, and this can only be dued to the fact that on that part, X_N becomes very large. In other heuristic words, the "enemies" here are higher and higher very thin "spikes" whose integral would not vanish.

EXAMPLE 5.1.4 (Taking off saucepans). *The very simple example of random variables that tend to* 0 *almost surely but not in* L^1 *would be to consider the Lebesgue measure on* $[0,1]$ *and the functions* $f_n = 2^n 1_{[0,2^{-n}]}$*. Then,* f_n *clearly tends to* 0 *almost surely, but* $E[f_n] = 1$*, so* f_n *does not tend to* 0 *in* L^1*.*

Let us formalize the property that a collection of random variables does not have higher and higher spikes:

DEFINITION 5.1.5. *A family of random variables $(X_i)_{i\in I}$ is called uniformly integrable, if*

$$\sup_{i\in I} E[|X_i|1_{|X_i|>K}] \to 0$$

as $A \to \infty$.

Equivalently, a family of random variables $(X_i)_{i\in I}$ is uniformly integrable, if for any $\varepsilon > 0$, one can find K large enough so that for all $i \in I$, $E[|X_i|1_{|X_i|>K}] \leq \varepsilon$.

REMARK 5.1.6. *A UI family is necessarily bounded in L^1. Indeed, the condition implies in particular that when K is large enough, then for all $i \in I$,*

$$E[|X_i|1_{|X_i|>K}] \leq 1,$$

so that

$$E[|X_i|] \leq K + 1.$$

REMARK 5.1.7. *A family with just one random variable Z in L^1 is clearly UI since by dominated convergence, $E[|Z|1_{|Z|>K}] \to 0$ as $K \to \infty$. A finite family $(Z_1, \ldots, Z_n)$ of random variables in L^1 is also UI since*

$$\max_{j\leq n} E[|Z_j|1_{|Z_j|>K}] \to 0$$

as $K \to \infty$ (as the max of n converging sequences). More generally, a family $(X_i)_{i\in I}$ with the property that there exists Z in L^1 such that for all $i \in I$, $|X_i| \leq Z$ almost surely is uniformly integrable, because for all $i \in I$, $E[|X_i|1_{|X_i|>K}] \leq E[|Z|1_{|Z|>K}]$.

A useful equivalent description of UI is the following:

LEMMA 5.1.8 (Equivalent formulation of UI). *A family of random variables $(X_i)_{i\in I}$ is uniformly integrable if and only if it is bounded in L^1 and if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $A \in \mathcal{A}$ with $P[A] \leq \delta$ and all $i \in I$, one has $E[|X_i|1_A] \leq \varepsilon$.*

PROOF. Suppose that the family $(X_i)_{i\in I}$ is UI. We have already noted that it is then bounded in L^1. We now choose $\varepsilon > 0$. We can then find K so that for all $i \in I$,

$$E[|X_i|1_{|X_i|>K}] \leq \varepsilon/2.$$

Let us now choose any measurable set A with $P[A] \leq \varepsilon/(2K)$. Then, for all $i \in I$,

$$E[|X_i|1_A] \leq E[|X_i|1_{\{|X_i|>K\}}] + E[|X_i|1_{A\cap\{|X_i|\leq K\}}] \leq (\varepsilon/2) + KP(A) \leq \varepsilon$$

so that the desired statement holds for $\delta = \varepsilon/(2K)$.

Let us now conversely suppose that $(X_i)_{i\in I}$ is bounded in L^1 and that for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $A \in \mathcal{A}$ with $P(A) \leq \delta$, and for all $i \in I$, $E[|X_i|1_A] \leq \varepsilon$.

Then, for any $\varepsilon > 0$, we choose $K := \sup_i E[|X_i|]/\delta$, and we then see by Markov's inequality that for all $i \in I$,

$$P(|X_i| \geq K) \leq E[|X_i|]/K \leq \delta.$$

We can therefore apply the above property to $A = \{|X_i| \geq K\}$, and we get that for all $i \in I$,

$$E[|X_i|1_{|X_i|\geq K}] \leq \varepsilon,$$

so that $(X_i)_{i\in I}$ is indeed UI. □

An example of an application of this lemma is the following (of course, this could also have been derived directly):

COROLLARY 5.1.9. *If two families $(X_i)_{i\in I}$ and $(Y_j)_{j\in J}$ are UI, then the family $(X_i+Y_j)_{(i,j)\in I\times J}$ is UI.*

PROOF. We first note that this new family is bounded in L^1 (Since $|X_i+Y_j| \leq |X_i|+|Y_j|$) and that for each $\varepsilon > 0$, one can find δ so that for all A with $P[A] \leq \delta$ and all $i \in I$ and $j \in J$,

$$E[|X_i|1_A] \leq \varepsilon/2 \text{ and } E[|Y_j|1_A] \leq \varepsilon/2.$$

Then we just write

$$E[|X_i + Y_j|1_A] \leq E[|X_i|1_A] + E[|Y_j|1_A] \leq \varepsilon.$$

□

We can now state the main result of this section:

PROPOSITION 5.1.10. *A sequence of random variables $(X_n)_{n\geq 1}$ in L^1 converges in L^1 if and only if it converges in probability and is uniformly integrable.*

REMARK 5.1.11. *In some sense, this statement contains all the limiting theorems for integral. For instance, the dominated convergence theorem is due to the previous observation that a sequence of random variables that is dominated by an L^1 random variable is UI.*

PROOF. Let us first suppose that X_n converges in L^1 to X. We have already seen that it then converges in probability, and since we know that the limit X is also in L^1, it will be sufficient to show that $(Y_n)_{n\geq 1}$ is UI, where $Y_n := X_n - X$.

Let us choose $\varepsilon > 0$. Since $E[|Y_n|] \to 0$, we can find n_0 such that for all $n \geq n_0$, $E[|Y_n|] < \varepsilon$. On the other hand, We can then look at the finitely many L^1 random variables $Y_1, \ldots, Y_{n_0-1}$ and since this family is UI, we see that for large enough K,

$$E[|Y_n|1_{|Y_n|>K}] < \varepsilon$$

also for $n < n_0$. Hence, for this choice of K, we get that for all $n \geq 1$, $E[|Y_n|1_{|Y_n|>K}] \leq \varepsilon$ which shows that $(Y_n)_{n\geq 1}$ is indeed UI.

Conversely, let us suppose that the sequence $(X_n)_{n\geq 1}$ converges in probability to X and that it is UI. Since uniform integrability implies boundedness in L^1, it follows from Fatou's lemma that X is also in L^1. The previous corollary then shows that the sequence $(Y_n := X_n - X)_{n\geq 1}$ is UI. Our goal is to show that $E[|Y_n|] \to 0$.

Suppose that $\varepsilon > 0$. We first choose K such that for all n, $E[|Y_n|1_{|Y_n|\geq K}] \leq \varepsilon/2$. But

$$E[|Y_n|] \leq E[|Y_n||1_{|Y_n|<K}] + E[|Y_n|1_{|Y_n|\geq K}].$$

When $n \to \infty$, we know that for all $\varepsilon > 0$,

$$E[|Y_n|1_{|Y_n|<K}] \leq (\varepsilon/4) + KP(|Y_n| \geq \varepsilon/4)$$

which is bounded by $\varepsilon/2$ for all large enough n (because $P(|Y_n| \geq \varepsilon/4) \to 0$). Hence, we conclude that for all large enough n, $E[|Y_n|] \leq \varepsilon$, which shows the desired convergence in L^1. □

5.1.3. Two (very!) useful uniform integrability criteria. The following two criteria are good to have in our toolbox:

PROPOSITION 5.1.12 (Boundedness in L^p for $p > 1$ implies UI). *A family of random variables $(X_i)_{i\in I}$ such that for some $p > 1$,*

$$\sup_{i\in I} E[|X_i|^p] < \infty$$

is uniformly integrable.

PROOF. Suppose that for some $p > 1$, $\sup_{i\in I} E[|X_i|^p] \leq C < \infty$. Then,

$$E[|X_i|1_{|X_i|>K}] \leq E[\frac{|X_i|^{p-1}}{K^{p-1}}|X_i|1_{|X_i|>K}] \leq \frac{1}{K^{p-1}}E[|X_i|^p] \leq \frac{C}{K^{p-1}},$$

so that $(X_i)_{i\in I}$ is indeed uniformly integrable. □

PROPOSITION 5.1.13 (Collections of conditional expectations are UI). *Suppose that X is in L^1. Let $(\mathcal{A}_i)_{i\in I}$ denote a collection of sub-σ-fields of $\mathcal{A}$. The family of random variables $(E[X|\mathcal{A}_i])_{i\in I}$ is UI.*

PROOF. Let $Y_i = E[|X||\mathcal{A}_i]$. Since $|E[X|\mathcal{A}_i]| \leq Y_i$ almost surely, it in fact suffices to show that the family $(Y_i)_{i\in I}$ is UI.

Let $\varepsilon > 0$. Since $|X|$ is in L^1, it follows that one can find $\delta > 0$ such that for all $A \in \mathcal{A}$ with $P[A] < \delta$, one has $E[|X|1_A] \leq \varepsilon$ (this can be viewed as a special case of the previous lemma, but can also be checked directly just as in its proof, noting that

$$E[|X|1_A] \leq E[|X|1_{|X|>K}] + KP[A],$$

and choosing first K so that the first term is smaller than $\varepsilon/2$, and then δ so that the second one is smaller than $\varepsilon/2$ as well).

By Markov's inequality, $P[|Y_i| > K] \leq E[|Y_i|]/K \leq E[|X|]/K$. Furthermore, the event $A := \{|Y_i| \geq K\}$ is in $\mathcal{A}_i$, so that

$$E[|Y_i|1_{|Y_i|>K}] = E[|X|1_{|Y_i|>K}] = E[|X|1_A] \leq \varepsilon$$

as soon as $K > E[|X|]/\delta$ (for all $i \in I$). So, one indeed has that for all K large enough,

$$\sup_{i\in I} E[|Y_i|1_{|Y_i|>K}] \leq \varepsilon.$$

□

5.2. Two examples of consequences of these UI criteria

5.2.1. Consequence for inverse martingales. We are going to use the second criterion for uniform integrability here.

Recall that when $(M_r)_{r\leq 0}$ is an inverse martingale with respect to an inverse filtration $(\mathcal{A}_r)_{r\leq 0}$, then $M_0 \in L^1(\mathcal{A}_0)$ and for all $r \leq 0$,

$$M_r = E[M_0|\mathcal{A}_r]$$

almost surely. So, by the previous proposition, an inverse martingale is always UI.

On the other hand, using the number of upcrossings considerations, we have argued that an inverse martingale converges almost surely, i.e., that for some $M_r \to M_{-\infty}$ almost surely as $r \to -\infty$.

Hence, the sequence M_{-n} converges a.s. and is UI – it therefore also converges in L^1, which implies in particular that $E[M_{-\infty}] = E[M_0]$.

We can actually deduce one further statement here: Let $\mathcal{A}_{-\infty} := \cap_{r\leq 0}\mathcal{A}_r$. We note that for each $r_0 \leq 0$, $M_{-\infty}$ is measurable with respect to $\mathcal{A}_{r_0}$ (because all M_r for $r \leq r_0$ are), so that $M_{-\infty}$ is also measurable with respect to $\mathcal{A}_{-\infty}$. Let us then choose any $A \in \mathcal{A}_{-\infty}$. Then,

$$E[M_{-\infty}1_A] = \lim_{r\to-\infty} E[M_r 1_A] = \lim_{r\to-\infty} E[M_0 1_A] = E[M_0 1_A].$$

We can therefore conclude that $M_{-\infty} = E[M_0|\mathcal{A}_{-\infty}]$ almost surely.

Let us summarize all the results on inverse martingales as a theorem:

THEOREM 5.2.1 (Summary of results on inverse martingales). *If $(M_r)_{r\leq 0}$ is an inverse martingale with respect to the inverse filtration $(\mathcal{A}_r)_{r\leq 0}$, then M_r converges almost surely and in L^1 to $M_{-\infty} := E[M_0|\mathcal{A}_{-\infty}]$ where $\mathcal{A}_{-\infty} := \cap_{r\leq 0}\mathcal{A}_r$.*

Let us now explain how one can apply this result for the law of large numbers: We can recall (from the end of the previous chapter) that if $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables that are in L^1, then if one defines $S_0 = 0$ and $S_n = X_1 + \ldots + X_n$ for $n \geq 1$, and

$$\mathcal{A}_{-n} := \sigma(S_n, X_{n+1}, X_{n+2}, \ldots),$$

then the process $M_{-n} := E[X_1|\mathcal{A}_{-n}]$ is an inverse martingale, and that for symmetry reasons, $M_{-n} = S_n/n$. We have already pointed out that this particular inverse martingale does converge almost surely to some constant random variable (hte convergence is due to the upcrossing argument, and the fact that the limit if a constant is due to Kolmogorov's $0-1$ law). We can now add that this convergence holds also in L^1, which implies that this constant has to be $E[X_1]$, because $E[M_{-\infty}] = E[M_0]$. We can then finally sum it also up in the form of the following theorem:

THEOREM 5.2.2 (Law of large numbers, strongest version). *If $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables in L^1, then the sequence $(X_1 + \cdots + X_n)/n$ does converge almost surely and in L^1 to $E[X_1]$.*

5.2.2. Explosion of supercritical Galton-Watson processes. Let us immediately note that one can deduce the following statement for martingales from the previous UI criterion: Suppose that a martingale $(M_n)_{n\geq 0}$ is *bounded in L^p* for some $p > 1$. Then:

(1) By the previous criterion, $(M_n)_{n\geq 0}$ is UI.
(2) UI implies boundedness in L^1, so that the convergence theorem for martingales implies that M_n converges almost surely as $n \to \infty$ to some random variable M_∞ which is also in L^1.
(3) Combining (1) and (2) (since almost sure convergence and uniform integrability combined do imply convergence in L^1), we get that M_n converges to M_∞ also in L^1, and that $E[M_\infty] = \lim_{n\to\infty} E[M_n] = E[M_0]$.

We will come back to these type of general statements in the section on UI martingales and in the coming chapter on L^2 martingales, but we can already illustrate how this leads to interesting results for supercritical Galton-Watson processes.

We consider here a Galton-Watson process $(X_n)_{n\geq 0}$ with reproduction law $\mathcal{L}$ when (if ξ has law $\mathcal{L}$) $R := E[\xi]$ is finite and greater than 1. Recall that the process $(M_n := X_n/R^n)_{n\geq 0}$ is a non-negative martingale, that therefore converges almost surely to some finite random variable M_∞. At this point, we however did not show that $P(M_\infty = 0) \neq 1$.

We are now going to assume that the reproduction law $\mathcal{L}$ has a finite second moment, i.e., that $E[\xi^2] < \infty$. We now iteratively show that X_n is also in L^2 and actually compute $E[X_n^2]$: Using the same notations as in the section where we introduced the Galton-Watson processes, we have

$$\begin{aligned}
E[X_{n+1}^2|\mathcal{F}_n] &= E[\sum_{j,j'\geq 1} 1_{j\leq X_n} 1_{j'\leq X_n} \xi_{n,j}\xi_{n,j'}|\mathcal{F}_n] \\
&= \sum_{j,j'\geq 1} E[1_{j\leq X_n} 1_{j'\leq X_n} \xi_{n,j}\xi_{n,j'}|\mathcal{F}_n] \\
&= \sum_{j,j'\geq 1} 1_{j\leq X_n} 1_{j'\leq X_n} E[\xi_{n,j}\xi_{n,j'}|\mathcal{F}_n] \\
&= \sum_{j,j'\geq 1, j\neq j'} 1_{j\leq X_n} 1_{j'\leq X_n} E[R^2] + \sum_{j\geq 1} E1_{j\leq X_n} E[\xi^2] \\
&= R^2 X_n(X_n - 1) + E[\xi^2]X_n \\
&= R^2 X_n^2 + X_n \mathrm{Var}(\xi)
\end{aligned}$$

so that (recalling that $E[X_n] = R^n$ and $M_n = X_n/R^n$), we get that

$$E[M_{n+1}^2] = E[M_n^2] + \frac{\mathrm{Var}(\xi)}{R^{n+2}}$$

and therefore, since $M_0 = 1$, we get inductively that for all $n \geq 1$,

$$E[M_n^2] = 1 + \mathrm{Var}(\xi) \sum_{j=1}^{n} \frac{1}{R^{j+1}}.$$

In particular, since $R > 1$, we see that $(M_n)_{n\geq 0}$ is bounded in L^2. We can therefore conclude that it is UI, and therefore (as explained above) that M_n also converges in L^1 to M_∞.

Since for all $n \geq 0$, $E[M_n] = E[M_0] = 1$, this shows in particular that $E[M_\infty] = \lim_{n\to\infty} E[M_n] = 1$, and therefore that $P(M_\infty > 0) > 0$. We have therefore shown that:

PROPOSITION 5.2.3. *If $(X_n)_{n\geq 0}$ is a supercritical Galton-Watson process such that the reproduction law has a finite second moment, then (if $R = E[\xi]$), X_n/R^n converges almost surely and in L^1 to a random variable M_∞. On the set of non-zero probability for which $M_\infty \neq 0$, one therefore has $X_n \sim M_\infty R^n$ as $n \to \infty$.*

REMARK 5.2.4. *It is possible (see towards the end of the course about Markov chains) to show that in this case (i.e. with finite second moments for the reproduction law), the event $\{M_\infty = 0\}$ is equal to the event that the genealogy dies out, i.e., $\cup_{n\geq 1}\{X_n = 0\}$.*

Interestingly, it is actually possible to find reproduction laws $\mathcal{L}$ for which $E[\xi] \in (1, \infty)$ such that $M_\infty = 0$ almost surely (of course, because of the above proposition, they would have to satisfy $E[\xi^2] = \infty$), so that for those laws, one has $X_n = o(R^n)$ almost surely. However, the previous proposition easily implies the following result that shows that for every $\varepsilon > 0$, on a set of positive probability, M_n explodes at least like $(R - \varepsilon)^n$:

PROPOSITION 5.2.5. *If $(X_n)_{n\geq 0}$ is a supercritical Galton-Watson process such that the reproduction law has a finite second moment, then (if $R = E[\xi]$), for all $\varepsilon > 0$,*

$$P[X_n/(R-\varepsilon)^n \to \infty] > 0.$$

PROOF. Since $E[\xi 1_{\xi \leq K}] \to E[\xi] = R$ as $K \to \infty$, we can choose K large enough such that

$$E[\xi 1_{\xi \leq K}] > (R - \varepsilon).$$

Then, for all n and j, we can define

$$\tilde{\xi}_{n,j} := \xi_{n,j} 1_{\xi_{n.j} \leq K}$$

and we call $\tilde{\mathcal{L}}$ the law of these random variables.

We can then use these random variables to construct a Galton-Watson process $(\tilde{X}_n)_{n\geq 0}$ with reproduction law $\tilde{\mathcal{L}}$. It follows immediately from the fact that $\tilde{\xi}_{n,j} \leq \xi_{n,j}$ for all n and j that for all $n \geq 0$, $\tilde{X}_n \leq X_n$.

On the other hand, we know that $E[\tilde{\xi}^2] \leq K^2 < \infty$, so that by the previous proposition, $\lim_{n\to\infty} X_n/E[\tilde{\xi}]^n$ is positive with a positive probability. Since $E[\tilde{\xi}] > R - \varepsilon$, we conclude that on this event, $X_n/(R-\varepsilon)^n \to \infty$. □

5.3. UI martingales and the optional stopping theorem

5.3.1. UI martingales. Before discussing questions about martingales, let us make the following observation:

LEMMA 5.3.1. *If a sequence of L^1 random variables $(X_n)_{n\geq 0}$ converges in L^1 to some limit X, then for all sub-σ-field $\mathcal{G}$ of $\mathcal{A}$, the sequence $E[X_n|\mathcal{G}]$ converges in L^1 to $E[X|\mathcal{G}]$.*

One way to interpret this result it to say that the mapping $X \mapsto E[X|\mathcal{G}]$ is continuous from $L^1(\mathcal{A})$ into $L^1(\mathcal{G})$.

PROOF. This is simply due to the fact that

$$E[|E[X_n|\mathcal{G}] - E[X|\mathcal{G}]|] \leq E[E[|X_n - X|\ |\mathcal{G}] = E[|X_n - X|] \to 0$$

as $n \to \infty$. □

Let us now list some observations before stating the main theorem on uniformly integrable martingales. We assume here that $(\mathcal{F}_n)_{n\geq 0}$ is some filtration in some probability space $(\Omega, \mathcal{A}, P)$ and that $\mathcal{F}_\infty$ is the σ-field generated by $\cup_n \mathcal{F}_n$.

- Suppose that a martingale $(M_n)_{n\geq 0}$ (with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$) is uniformly integrable. Then, we have seen that it is bounded in L^1 (simply because $\sup_n E[|M_n|] \leq \sup_n E[|M_n|1_{|M_n|\geq K}] + K$ which is finite when K is large enough). By the martingale convergence theorem, M_n therefore converges almost surely to some finite $\mathcal{F}_\infty$-measurable limit M_∞ (it is $\mathcal{F}_\infty$-measurable as limit of $\mathcal{F}_\infty$-measurable variables). Since $(M_n)_{n\geq 0}$ is also UI, we can conclude that it converges also in L^1 to M_∞.
- Next, suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$ that does converge in L^1 to some random variable M_∞.

  For all $n_0 \leq n$, we know that $M_{n_0} = E[M_n|\mathcal{F}_{n_0}]$ almost surely. Since M_n converges in L^1 to M_∞, the previous lemma shows that for all n_0, $E[M_n|\mathcal{F}_{n_0}] \to E[M_\infty|\mathcal{F}_{n_0}]$ in L^1 as $n \to \infty$. But since $E[M_n|\mathcal{F}_{n_0}] = M_{n_0}$ almost surely for all $n \geq n_0$, we conclude that for all $n_0 \geq 0$,

  $$M_{n_0} = E[M_\infty|\mathcal{F}_{n_0}]$$

  almost surely.
- Finally, let us suppose that X is a random variable in L^1, and that for all $n \geq 0$, $M_n := E[X|\mathcal{F}_n]$ almost surely. Then, we know by the tower property for conditional expectations that $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$. Furthermore, by our second criteria for uniform integrability, we know that this martingale $(M_n)_{n\geq 0}$ is UI.

We have therefore shown the following theorem:

THEOREM 5.3.2. *Let $(M_n)_{n\geq 0}$ be a collection of random variables defined in some filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_n)_{n\geq 0}, P)$. The following three statements are equivalent:*

1. *$(M_n)_{n\geq 0}$ is a uniformly integrable martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.*
2. *$(M_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ that converges in L^1 to some random variable M_∞.*
3. *There exists a random variable in $L^1(\mathcal{A})$ such that for all $n \geq 0$, $M_n = E[X|\mathcal{F}_n]$ almost surely.*

Furthermore, if any one of these statements holds (and therefore the other two as well), then M_n converges also almost surely to M_∞, and $M_\infty = E[X|\mathcal{F}_\infty]$ almost surely.

REMARK 5.3.3. *The situation is therefore a little bit similar to the case of inverse martingales, where we have seen that if $(\mathcal{F}_r)_{r\leq 0}$ was an inverse filtration, then when $X \in L^1$, $E[X|\mathcal{F}_r]$ converges almost surely and in L^1 to $E[X|\mathcal{F}_{-\infty}]$ as $r \to \infty$. Here, we have that when $(\mathcal{F}_n)_{n\geq 0}$ is a (forward)*

filtration and $X \in L^1$, then $E[X|\mathcal{F}_n]$ converges almost surely and in L^1 to $E[X|\mathcal{F}_\infty]$ as $n \to \infty$. So in some sense, the conditional expectation is also "continuous" with respect to (monotone sequences of) σ-fields.

One particular feature of the convergence in L^1 of martingales is that since $E[M_n] = E[M_0]$ for all n, if the martingale M_n converges in L^1 to M_∞, then $E[M_\infty] = E[M_0]$. This turns often out to be very useful. We have seen this already in the case of the supercritical Galton-Watson processes, and it will be also important in the numerous applications of the optional stopping theorem that we will discuss in the next section.

5.3.2. The optional stopping theorem. Throughout this section, we consider a filterered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_n)_{n\geq 0}, P)$.

DEFINITION 5.3.4. *A mapping T from Ω into $\mathbb{N} \cup \{+\infty\}$ is called a stopping time (with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$) if for all $n \geq 0$, $\{T = n\} \in \mathcal{F}_n$. The stopping time is called finite if furthermore, $T < \infty$ almost surely.*

Note that if T is a stopping time, then

$$\{T \leq n\} = \cup_{j=0}^{n} \{T = j\} \in \mathcal{F}_n$$

and that $\{T > n\} = \Omega \setminus \{T \leq n\}$ is in $\mathcal{F}_n$ as well.

Suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ and that T is a stopping time with respect to this same filtration. We then defined the *stopped process* M^T by

$$M_n^T := M_{\min(n,T)}$$

for all $n \geq 0$ [To avoid notational confusion, in the remained of this chapter, we will write $(x)^p$ for the p-th power of x instead of x^p]. It is then very easy to see that:

LEMMA 5.3.5. *The process $(M_n^T)_{n\geq 0}$ is also a martingale with respect to the same filtration.*

PROOF. Indeed, for all $n \geq 0$,

$$M_n^T = \sum_{j=0}^{n} 1_{T=j} M_j + 1_{T>n} M_n$$

is clearly $\mathcal{F}_n$ measurable and in L^1, and

$$E[(M_{n+1}^T - M_n^T)|\mathcal{F}_n] = E[1_{T>n}(M_{n+1} - M_n)|\mathcal{F}_n] = 1_{T>n} E[M_{n+1} - M_n|\mathcal{F}_n] = 0$$

almost surely. □

As a consequence, we see that:

COROLLARY 5.3.6. *If $(M_n)_{n\geq 0}$ and T is a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$, then for all $n \geq 0$, one has $E[M_{\min(n,T)}] = E[M_0]$.*

When the stopping time T is almost surely finite, then M_n^T clearly converges almost surely to M_T as $n \to \infty$. Note that since almost surely,

$$M_T = \sum_{n\geq 0} 1_{T=n} M_n,$$

M_T is indeed a random variable (and is in fact measurable with respect to $\mathcal{F}_\infty$). So, we can immediately conclude that:

THEOREM 5.3.7 (Optional stopping theorem). *If T is a finite stopping time such that the stopped martingale $(M_n^T)_{n\geq 0}$ is UI, then M_n^T converges in L^1 to M_T, and in particular, $E[M_T] = E[M_0]$.*

Note that we did not build here upon any martingale convergence result to derive this result. The fact that M_n^T converges almost surely to M_T is obvious because T is finite (so that the martingale convergence theorem is not used here), and then the UI assumption implies that it also converges in L^1.

This result is very simple but very useful, as the next section and the exercise sheets will show. In practise, to check that the stopped martingale is UI, one often uses the boundedness in L^p criterion for some $p > 1$.

REMARK 5.3.8. *When T is a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$, it is often useful (see the exercise sheets) to define as follows the σ-field denoted by $\mathcal{F}_T$ and that can be interpreted as "the information about what happened until T": A set A is in $\mathcal{F}_T$ if and only if it is in $\mathcal{F}_\infty$, and if for any $n \geq 0$,*

$$A \cap \{T = n\} \in \mathcal{F}_n.$$

It is easy to see that this is a σ-field. Stopping times and their corresponding σ-fields will show up again in these lectures in the context of Markov chains.

5.3.3. Application to simple random walk exit times. Let us now illustrate how one can use the optional stopping theorem to deduce results about the simplest of processes, namely the simple random walk in $\mathbb{Z}$.

We consider a sequence of independent identically distributed random variables $(\varepsilon_j)_{j\geq 1}$ with $P(\varepsilon_j = 1) = P(\varepsilon_j = -1) = 1/2$, we define $S_n = \varepsilon_1 + \cdots + \varepsilon_n$ for $n \geq 1$ and $S_0 = 0$, and $\mathcal{F}_n := \sigma(\cup_{j\leq n}\sigma(\varepsilon_j))$.

We define for each $x \in \mathbb{Z}$, the random time

$$T_x = \min\{n \geq 0 : \ S_n = x\}$$

and for all $a < 0 < b$, the exit time of the interval (a, b) by

$$T_{a,b} = \min(T_a, T_b).$$

We now make the following remarks:

- Almost surely, all the times T_x for $x \in \mathbb{Z}$ are finite. There are numerous ways to prove this. One way is to note that by symmetry, the events $\{\limsup_{n\to\infty} S_n = +\infty\}$ and $\{\liminf_{n\to\infty} S_n = -\infty\}$ have the same probability, and by Kolmogorov's $0-1$ law, this probability is either 0 or 1. Hence, if $P(\limsup_{n\to\infty} |S_n| = \infty) = 1$, then both previous events hold almost surely,

  On the other hand, for all $K \geq 1$, it is easy to see that

  $$P(\limsup_{n\to\infty} |S_n| < K) = 0.$$

  For instance, one can apply the Borel-Cantelli lemma to the events

  $$A_j := \{\varepsilon_{2Kj+1} = \varepsilon_{2Kj+2} = \ldots = \varepsilon_{2Kj+2K} = 1\}$$

  to see that almost surely A_j holds for infinitely many j's. But if A_j holds then either $S_{2Kj} \leq -K$ or $S_{2Kj+2K} \geq K$, so that almost surely, $|S_{2Kj}| \geq K$ for infinitely many values of j.

  Hence, we get that $P(\limsup_{n\to\infty} |S_N| < \infty) = 0$ which completes the proof.
- All the times T_x and $T_{a,b}$ are stopping times with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ – indeed, for all $x \in \mathbb{Z}$ and all $n \geq 0$,

  $$\{T_x = n\} = \{S_n = x\} \cap (\cap_{j<n}\{S_j \neq x\}) \in \mathcal{F}_n$$

  (and for all $a < 0 < b$, $\{T_{a,b} \leq n\} = \{T_a \leq n\} \cup \{T_b \leq n\} \in \mathcal{F}_n$).

- The simple random walk $(S_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ because for all $n \geq 0$,

$$E[S_{n+1} - S_n|\mathcal{F}_n] = E[\varepsilon_{n+1}|\mathcal{F}_n] = E[\varepsilon_{n+1}] = 0$$

almost surely.
- Now comes an obervation that should be the **big warning** when one does try to apply the optional stopping without checking the uniform integrability condition: Consider the finite stopping time $\tau := T_{-1}$. Then the stopped walk $(S_n^\tau)_{n\geq 0}$ is a martingale. It converges almost surely to $S_\tau = -1$ as $n \to \infty$ because $\tau < \infty$ almost surely. We also know that $S_0 = 0$. So clearly,

$$E[S_\tau] = 1 \neq 0 = E[S_0]$$

and the conclusion of the optional stopping theorem does not hold in this case. So, this actually proves that $(S_n^T)_{n\geq 0}$ is *not UI*.

It is actually a typical example of a non-UI sequence of random variables. The heuristic explanation is that when n is large and $\tau > n$, then (in order to avoid -1 until time n), the "typical" behaviour (on this event of small probability) will be that S_n is very large.
- However, the martingales $(S_n^\sigma)_{n\geq 0}$ when $\sigma := T_{a,b}$ for $a < 0 < b$ are clearly UI (they are actually bounded), since

$$|S_n^\sigma| \leq \max(-a, b).$$

So, we can apply the optional stopping theorem here so that

$$E[S_{T_{a,b}}] = E[S_0] = 0.$$

But

$$E[S_{T_{a,b}}] = E[S_{T_{a,b}} 1_{T_a<T_b}] + E[S_{T_{a,b}} 1_{T_b<T_b}] = aP(T_a < T_b) + bP(T_b < T_a).$$

Since $P(T_b < T_a) = 1 - P(T_a < T_b)$, we conclude that

$$P(T_a < T_b) = \frac{b}{b-a}.$$

The optional stopping theorem therefore enables us to compute the probability that the simple random walk hits b before a.
- When $(S_n)_{n\geq 0}$ is a simple random walk, one can actually a whole family of other martingales. The first one to consider here is the *quadratic martingale*: For all $n \geq 0$, we define

$$Q_n := (S_n)^2 - n.$$

For each $n \geq 0$, Q_n is a bounded $\mathcal{F}_n$-measurable random variable, and

$$E[(S_{n+1})^2|\mathcal{F}_n] = E[(S_n + \varepsilon_{n+1})^2|\mathcal{F}_n] = (S_n)^2 + 2S_n E[\varepsilon_{n+1}|\mathcal{F}_n] + E[(\varepsilon_{n+1})^2] = (S_n)^2 + 1$$

almost surely – the process $(Q_n)_{n\geq 0}$ is therefore also a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.

If $\sigma = T_{a,b}$ for $a < 0 < b$, then $(Q_n^\sigma)_{n\geq 0}$ is also a martingale, and it converges almost surely to Q_σ as $n \to \infty$. Instead of checking that this stopped martingale is UI in order to conclude that $E[Q_\sigma] = E[Q_0] = 0$, we instead use the following argument: For each $n \geq 0$, we know that $E[Q_n^\sigma] = E[Q_0] = 0$, so that

$$E[(S_{\min(n,\sigma)})^2] = E[\min(n, \sigma)].$$

Now, by dominated convergence, the left-hand side converges to $E[(S_\sigma)^2]$ as $n \to \infty$, and by monotone convergence (for non-negative random variables), the right-hand side converges to $E[\sigma]$. We can therefore conclude that σ is in L^1, and that $E[\sigma] = E[(S_\sigma)^2]$.

But by the previous description of the probability that $T_a < T_b$, we get that

$$E[(S_\sigma)^2] = \frac{(a)^2 b}{b-a} + \frac{(b)^2(-a)}{b-a} = -ab,$$

and we can therefore conclude that

$$E[T_{a,b}] = |ab|.$$

- A very rich class of further martingales are the so-called *exponential martingales*. For each $\lambda \in \mathbb{R}$, we define

$$M(\lambda)_n := \exp(\lambda S_n)/(\cosh \lambda)^n.$$

For each λ, the process $(M(\lambda)_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$, because for each $n \geq 0$,

$$E[\exp(\lambda S_{n+1})|\mathcal{F}_n] = E[\exp(\lambda S_n)\exp(\lambda \varepsilon_{n+1})|\mathcal{F}_n] = \exp(\lambda S_n)E[\exp(\lambda \varepsilon_{n+1})] = \exp(\lambda S_n)\cosh\lambda$$

almost surely (where we have used that $\exp(\lambda S_n)$ is a bounded $\mathcal{F}_n$-measurable random variable, and ε_{n+1} is independent of $\mathcal{F}_n$), so that indeed $E[M(\lambda)_{n+1}|\mathcal{F}_n] = M(\lambda)_n$ almost surely.

We can then apply the optional stopping theorem to the martingales $(M(\lambda)_n)_{n\geq 0}$ stopped at T_b or $T_{a,b}$ for $a < 0 < b$. For example (we leave other cases to the exercise sheets), when $\lambda \geq 0$, then for all $n \geq 0$,

$$|M(\lambda)_n^{T_b}| \leq \exp(\lambda b)$$

so that $(M(\lambda)_n^{T_b})_{n\geq 0}$ is UI. So, we can apply the optional stopping theorem, and since $S_{T_b} = b$ almost surely and $M(\lambda)_0 = 1$, we get that

$$E[\exp(\lambda b)(\cosh(\lambda))^{-T_b}] = 1,$$

which we can rewrite as

$$E[(\cosh(\lambda))^{-T_b}] = \exp(-\lambda b).$$

This allows to get an explicit expression for $E[\exp(-uT_b)]$ for all $u \geq 0$ (by writing $\cosh(\lambda) = \exp(u)$).

CHAPTER 6

L^p martingales for $p > 1$, Doob's inequalities

We have aleady seen in the previous chapter that if a martingale was bounded in L^p for given some $p > 1$, then it was uniformly integrable and therefore, it converges almost surely and in L^1 to some limiting random variable. To prove this, we did build on (a) the almost sure convergence theorem for martingales that are bounded in L^1, and (b) the general considerations on uniformly integrable families.

In the present "stand-alone" mini-chapter, we are going to provide a different approach to the convergence of martingales that are bounded in L^2, using a generalization of the ideas that we used in our study of series of independent random variables.

In particular, we will show that the maximal inequality that was the key in the study of series of independent random variables generalizes nicely to the case of martingales. Doob's L^2 inequality is also a very useful result on its own right (for possible further developments, such as martingales in continuous time).

We will also use the maximal inequality to derive some new results about the convergence of L^p martingales for $p > 1$.

We choose to write it as a seperate chapter to stress the fact that we mostly do not build here on our previous results (i.e. on the almost sure convergence theorem or on UI considerations).

6.1. Stand-alone analysis of L^2 martingales, part I

We start with the following simple observation that is specific to the case of L^2 martingales:

Suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$, and that M_n is in L^2 for each $n \geq 0$ (this is what we call "an L^2-martingale").

Then, we can write for each $n \geq 1$, $X_n := M_n - M_{n-1}$, so that $M_n = M_0 + X_1 + \cdots + X_n$. The random variables $(X_n)_{n\geq 1}$ are not necessarily independent, but one still has that for all $n' < n$,

$$E[X_{n'} X_n] = E[X_{n'} E[X_n | \mathcal{F}_{n-1}]] = 0$$

(because $X_n - E[X_n|\mathcal{F}_{n-1}]$ is orthogonal to $L^2(\mathcal{F}_{n-1}$, and $X_{n'}$ is in that space). Similarly, for all event A in $\mathcal{F}_{n-1}$,

$$E[1_A X_{n'} X_n] = 0.$$

If we expand $E[(\sum_{n=1}^N X_n)^2]$, all cross-terms disappear, so that we are left with

$$E[(M_n - M_0)^2] = \sum_{n=1}^{N} E[X_n^2].$$

So, many features of the sums of independent centered L^2 variables do still hold for L^2 martingales.

In particular, we see that the martingale $(M_n)_{n\geq 0}$ is bounded in L^2 if and only if

$$E[M_0^2] + \sum_{n\geq 0} E[(M_{n+1} - M_n)^2] < \infty.$$

This makes it possible to prove directly the following fact:

PROPOSITION 6.1.1. *If the martingale $(M_n)_{n\geq 0}$ is bounded in L^2, then it converges in L^2.*

REMARK 6.1.2. *This result is new compared to our previous results. So far, we have proved that a martingale that is bounded in L^2 does converge almost surely and in L^1.*

PROOF. We just note that

$$E[(M_{n+k} - M_n)^2] = \sum_{j=0}^{k-1} E[(M_{n+j+1} - M_{n+j})^2] \leq \sum_{j\geq 0} E[(M_{n+j+1} - M_{n+j})^2]$$

which tends to 0 as $n \to \infty$ (uniformly with respect to k) if the martingale is bounded in L^2. Hence, $(M_n)_{n\geq 0}$ is Cauchy in the space of L^2 random variables, and since this space is complete, it converges in that space. □

If we combine this proposition with the almost sure convergence theorem (since boundedness in L^2 implies boundedness in L^1, we get the following "ultimate" statement about convergence of L^2 martingales.

PROPOSITION 6.1.3. *If an L^2 martingale $(M_n)_{n\geq 0}$ is bounded in L^2, i.e., if $\sup_{n\geq 0} E[(M_n)^2] < \infty$, then M_n converges in L^2 and almost surely to some limit M_∞.*

REMARK 6.1.4. *We are going to see an alternative proof of the almost sure convergence of martingales that are bounded in L^2 via Doob's inequalities in the next section.*

6.2. Doob's inequalities for martingales

6.2.1. The maximal inequality. Suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$. We define

$$M_n^* = \max_{j\in\{0,\dots,n\}} |M_j|.$$

Then:

PROPOSITION 6.2.1 (Doob's maximal inequality). *For all $\lambda > 0$ and $n \geq 0$,*

$$\lambda P(M_n^* \geq \lambda) \leq E[|M_n| 1_{M_n^* \geq \lambda}] \leq E[|M_n|].$$

PROOF. We will use the same technique of proof than for the maximal inequality for sums if independent random variables. The idea of "freezing as soon as M_j is outside of $(-\lambda, \lambda)$" can be nicely encapsulated using the notion of stopping times, so we will use this formalism:

Let us define

$$T := \min\{n \geq 0 : \ |M_n| \geq \lambda\}$$

with the convention $\min \emptyset = \infty$. This is clearly a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$.

We note that $\{T \leq n\}$ if and only if $M_n^* \geq \lambda$, and that whenever T is finite, $|M_T| \geq \lambda$. Hence,

$$\lambda P(M_n^* \geq \lambda) = \sum_{j=0}^{n} \lambda P(T = j) \leq \sum_{j=0}^{n} E[|M_j| 1_{T=j}].$$

But we know that $(|M_j|)_{j\geq 0}$ is a submartingale and that $\{T = j\} \in \mathcal{F}_j$, so for all $j \leq n$,

$$E[|M_j| 1_{T=j}] \leq E[|M_n| 1_{T=j}].$$

Hence,

$$\lambda P(M_n^* \geq \lambda) \leq \sum_{j=0}^{n} E[|M_n| 1_{T=j}] = E[|M_n| 1_{T\leq n}],$$

which proves the claim. □

REMARK 6.2.2. *Almost the same proof can be used to show the following somewhat stronger result – here we use the notation* $x^+ := x1_{x>0}$*:*

PROPOSITION 6.2.3 (The submartingale version). *If* $(X_n)_{n\geq 0}$ *is a submartingale, then for all* $\lambda > 0$ *and* $n \geq 0$*,*

$$\lambda P(\max(X_0^+, \dots, X_n^+) \geq \lambda) \leq E[X_n 1_{\max(X_0^+,\dots,X_n^+)\geq\lambda}] \leq E[X_n^+].$$

The previous proposition is then the particular case when $X_n = |M_n|$*, and the maximal inequality for sums* S_n *of independent centered* L^2 *variables that we proved in the series of independent random variables chapter is the special case where one then considers the submartingale* $(S_n)^2$*.*

PROOF. One defines this time

$$T := \min\{n \geq 0 : \ X_n \geq \lambda\},$$

and then notes that $\{T \leq n\}$ if and only if $\max(X_0^+, \dots, X_n^+) \geq \lambda$, and that

$$\lambda P(T \leq n) \leq \sum_{j=0}^{n} E[1_{T=j} X_j] \leq \sum_{j=0}^{n} E[1_{T=j} X_n] \leq E[1_{T\leq n} X_n] \leq E[X_n^+].$$

□

6.2.2. Doob's L^2 inequality. We now consider the case where $(M_n)_{n\geq 0}$ is a martingale in L^2, i.e., such that for all $n \geq 0$, $E[M_n^2] < \infty$. Then:

PROPOSITION 6.2.4 (Doob's L^2 inequality). *For all* $n \geq 0$*,* $E[(M_n^*)^2] \leq 4E[M_n^2]$*.*

REMARK 6.2.5. *This is quite a remarkable inequality, as it holds for any* L^2 *martingale!*

REMARK 6.2.6. *Again, there is a submartingale version of this inequality that is treated in the exercise sheet.*

PROOF. This inequality is in fact a rather direct proof of the maximal inequality. The trick is to recall that by Fubini, for any non-negative random variable U,

$$E[U^2] = E[2\int_0^\infty 1_{u\leq U} u du] = 2\int_0^\infty uP(U \geq u)du.$$

So, applying this to the random variable M_n^*, and using the maximal inequality, we get

$$\begin{aligned} E[(M_n^*)^2] &= \int_0^\infty 2uP(M_n^* \geq u)du \\ &\leq 2\int_0^\infty E[|M_n| \times 1_{M_n^*\geq u}]du = 2E[|M_n| \int_0^\infty 1_{M_n^*\geq u} du] = 2E[|M_n| \times M_n^*]. \end{aligned}$$

But by Cauchy-Schwarz,

$$E[|M_n| \times M_n^*] \leq E[(M_n)^2]^{1/2} E[(M_n^*)^2]^{1/2}.$$

So we conclude that when $E[(M_n^*)^2] \neq 0$,

$$E[(M_n^*)^2]^{1/2} \leq 2E[(M_n)^2]^{1/2}$$

which proves the claim. □

REMARK 6.2.7. *This allows to revisit also the proof of the fact that a martingale bounded in* L^2 *does converge in* L^2 *in the following way (building on the almost sure convergence theorem) – this is of course much less direct than the proof based on orthogonality of increments, but as we shall see in the next section, it has the advantage that it can be easily generalized to* L^p *martingales for* $p > 1$*:*

Since $(M_n)_{n\geq 0}$ is bounded in L^2, it is also bounded in L^1 and therefore M_n converges almost surely to some random variable M_∞.

Since $\sup_n E[(M_n)^2] < \infty$, it follows by Fatou's lemma that M_∞ is also in L^2.

Let $M_\infty^* = \sup_{n\geq 0} |M_n|$. Monotone convergence and Doob's maximal inequality shows that

$$E[(M_\infty^*)^2] = \lim_{n\to\infty} E[(M_n^*)^2] \leq \limsup_{n\to\infty} 4E[(M_n)^2] = 4\sup_{n\geq 0} E[(M_n)^2] < \infty$$

(because the martingale is bounded in L^2). Hence, noting that $|M_n - M_\infty|^2 \leq (|M_n| + |M_\infty|)^2 \leq 4(M_\infty^*)^2$ almost surely, we can conclude by dominated convergence that $E[(M_n - M_\infty)^2] \to 0$, which proves the convergence in L^2.

6.2.3. Doob's L^p inequalities, convergence of L^p martingales. Suppose now that $p > 1$, and that $(M_n)_{n\geq 0}$ is a martingale with respect to some filtration, with the property that $M_n \in L^p$ for all $n \geq 0$. Then, we have a similar inequality relating the expected value of $(M_n^*)^p$ and that of $|M_n|^p$:

PROPOSITION 6.2.8 (Doob's L^p inequality). *For any $n \geq 0$, $E[(M_n^*)^p] \leq (p/(p-1))^p E[|M_n|^p]$.*

REMARK 6.2.9. *The particular case $p = 2$ is Doob's L^2 inequality that we have seen a few paragraphs ago. We can also note that as $p \to 1+$, the constant $(p/(p-1))^p$ tends to ∞, which illustrates that* there is no Doob's L^1 inequality.

This proposition is proved just like the L^2 inequality, just replacing the Cauchy-Schwarz inequality by Hölder's inequality:

PROOF. One notes that for any non-negative random variable U that is in L^p,

$$E[U^p] = E[p\int_0^\infty 1_{u\leq U} u^{p-1} du] = p\int_0^\infty u^{p-2} u P(U \geq u) du.$$

So, applying this to the random variable M_n^*, and using the maximal inequality, we get

$$E[(M_n^*)^p] \leq p\int_0^\infty u^{p-2} E[|M_n| 1_{u\leq M_n^*} du] \leq \frac{p}{p-1}\, E[|M_n|(M_n^*)^{p-1}] du.$$

But Hölder's inequality gives (for $1/q + 1/p = 1$ – note that then $(p-1)q = p$)

$$E[|M_n|(M_n^*)^{p-1}] \leq E[|M_n|^p]]^{1/p} E[(M_n^*)^{(p-1)q}]^{1/q} = E[|M_n|^p]^{1/p} E[(M_n^*)^p]]^{1/q}.$$

So we conclude that when $E[(M_n^*)^p] \neq 0$,

$$E[(M_n^*)^p]^{1-1/q} \leq \frac{p}{p-1} E[|M_n|^p]^{1/p}$$

which proves the claim (recall that $1 - 1/q = 1/p$). □

We then get the following corollary (just as in the remark in the L^2 case):

COROLLARY 6.2.10. *If a martingale is bounded in L^p for some $p > 1$, then it converges almost surely and in L^p as $n \to \infty$.*

PROOF. Being bounded in L^p implies being bounded in L^1, so that the martingale M_n converges almost surely to some M_∞ that is in L^1. But M_n^* is a non-decreasing non-negative sequence that converges almost surely to $M_\infty^* = \sup_{n\geq 0} |M_n|$, and by the monotone convergence theorem for non-negative variables,

$$E[(M_\infty^*)^p] = \sup_n E[(M_n^*)^p] \leq (p/p-1)^p \sup_n E[|M_n^p|] < \infty.$$

Since $|M_n - M_\infty|^p \leq 2^p (M_n^*)^p$, we can conclude by the dominated convergence theorem that $E[|M_n - M_\infty|^p] \to 0$ as $n \to \infty$. □

6.3. Stand-alone analysis of L^2 martingales, part II

Let us state once again the results about convergence of L^2 martingales that we have established so far.

PROPOSITION 6.3.1 (A convergence criteria for L^2 martingales). *Suppose that $(M_n)_{n\geq 0}$ is a martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$ such that all M_n are in L^2 and that $\sup_{n\geq 0} E[M_n^2] < \infty$. Then M_N converges almost surely and in L^2 as $N \to \infty$ to a finite random variable*

REMARK 6.3.2. *Our proof (so far) is a combination of the simple L^2 convergence argument from the begining of this chapter, and the (more involved) almost sure convergence theorem.*

We now provide an alternative proof of the almost sure convergence part, based on Doob's L^2 inequality:

ALMOST SURE CONVERGENCE VIA DOOB'S INEQUALITY. We first recall that

$$E[(M_n)^2] = E[(M_0)^2] + \sum_{j=0}^{n-1} E[(M_{j+1} - M_j)^2].$$

So, the boundedness in L^2 implies that

$$\sum_{j\geq 0} E[(M_{j+1} - M_j)^2] < \infty.$$

So, one can find an increasing sequence $n_k \to \infty$, such that

$$\sum_{j\geq n_k} E[(M_{j+1} - M_j)^2] < 4^{-(k+1)}.$$

For each k, we can apply Doob's inequality to the martingale $(M_{n_k+j})_{j\geq 0}$ (which is a martingale with respect to $(\mathcal{F}_{n_k+j})_{j\geq 0}$, which shows that

$$E[(\sup_{n\geq n_k} (M_n - M_{n_k})^2] \leq 4 \sup_{n\geq n_k} E[(M_n - M_{n_k})^2] < 4^{-k}.$$

Hence, by Markov inequality, for all $n \geq n_k$,

$$P(\sup_{n\geq n_k} (M_n - M_{n_k})^2 \geq 2^{-k}) \leq 2^{-k}.$$

Therefore, by Borel-Cantelli, we conclude that almost surely, there exists k_0 such that for all $k \geq k_0$, for all $n \geq n_k$,

$$\sup_{n\geq n_k} |M_n - M_{n_k}| \leq 2^{-k/2},$$

which also implies that for all $n, n' \geq n_k$,

$$\sup_{n,n'\geq n_k} |M_n - M_{n'}| \leq 2 \times 2^{-k/2}.$$

Hence, $(M_n)_{n\geq 0}$ is almost surely a Cauchy sequence, and it therefore almost surely converges. □

Summary

This concludes our series of chapters devoted to martingales. Here is a summary of all our convergence results for martingales:

$$\begin{aligned} \text{Bounded in } L^1 &\Rightarrow \text{CV almost surely} \\ \text{UI} &\Rightarrow \text{CV almost surely and in } L^1 \\ \text{Bounded in } L^p \text{ for } p>1 &\Rightarrow \text{CV almost surely and in } L^p \end{aligned}$$

and:

The optional stopping theorem is a special subcase of the UI $\Rightarrow L^1$ CV.