Advanced Probability and Applications (Part I)

Olivier Léveque, IC–LTHI, EPFL

March 4, 2018

Contents

1 $\sigma$-fields and random variables 3
  1.1 $\sigma$-fields .................................................. 3
  1.2 $\sigma$-field generated by a collection of events .......................... 4
  1.3 Sub-$\sigma$-field ................................................... 5
  1.4 Random variables ................................................ 5
  1.5 $\sigma$-field generated by a collection of random variables .......... 7

2 Probability measures and distributions 8
  2.1 Probability measures ........................................... 8
  2.2 Distribution of a random variable ................................ 10
  2.3 Cumulative distribution function .................................. 10
  2.4 Two important classes of random variables ....................... 11
  2.5 The Cantor set and the devil’s staircase .......................... 13

3 Independence 15
  3.1 Independence of two events ................................... 15
  3.2 Independence of two random variables .......................... 15
  3.3 Independence of two sub-$\sigma$-fields ............................ 16
  3.4 Independence of more sub-$\sigma$-fields .......................... 17
  3.5 Do independent random variables really exist? ................... 18

4 Expectation 19
  4.1 Discrete non-negative random variables .......................... 19
  4.2 General definition ............................................... 20
Basic terminology and conventions

- A *discrete* set means either a finite set (in bijection with \{1, \ldots, N\} for some \(N \geq 1\)) or a countable set (in bijection with \(\mathbb{N}\)).

- Capital letters \(X, Y, Z\) refer to random variables, while small letters \(x, y, z\) refer to numbers.

- A number \(x \in \mathbb{R}\) is said to be *non-negative* if \(x \geq 0\), and *positive* if \(x > 0\).

- Likewise, a function \(f : \mathbb{R} \rightarrow \mathbb{R}\) is said to be *non-decreasing* if \(f(x_1) \leq f(x_2)\) as soon as \(x_1 < x_2\), and *increasing* if \(f(x_1) < f(x_2)\) as soon as \(x_1 < x_2\).

And an important remark

A necessary preliminary to Probability Theory is Measure Theory; likewise, a necessary preliminary to Measure Theory is Topology, and it is probably fair to say also that a necessary preliminary to Topology is Set Theory. As we cannot cover everything in these notes, some facts will be stated without proof in order to avoid opening too many Pandora’s boxes. . . Readers interested in gaining a deeper understanding of the field are of course encouraged to search for other more detailed references on the subject.
1 σ-fields and random variables

1.1 σ-fields

In probability, the fundamental set Ω describes the set of all possible outcomes (or realizations) of a given experiment. It might be any set, without any particular structure, such as for example Ω = {1, . . . , 6} representing the outcomes of a die roll, or Ω = [0, 1] representing the outcomes of a concentration measurement of some chemical product. Notice moreover that the set Ω need not be composed of numbers exclusively; it would be for example perfectly valid to consider the set Ω = {banana, apple, orange}.

Given a fundamental set Ω, it is important to describe what information does one have on the system, namely on the outcomes of the experiment. This notion of information is well captured by the mathematical notion of σ-field, which is defined below. Notice that in elementary probability courses, it is generally assumed that the information one has about a system is complete, so that it becomes useless to introduce the concept below.

Definition 1.1. Let Ω be a set. A σ-field (or σ-algebra) on Ω is a collection F of subsets of Ω (or events) satisfying the following properties or axioms:

(i) ∅, Ω ∈ F.

(ii) If A ∈ F, then A ∈ F.

(iii) If A1, . . . , An ∈ F, then ∪jn=1 Aj ∈ F.

(iii′) If (An, n ≥ 1) is a sequence of subsets of Ω and An ∈ F for every n ≥ 1, then ∪∞n=1 An ∈ F.

Using De Morgan’s law: ∩jn=1 Aj = (∪jn=1 Aj)c, the above properties imply that

(iv) If A1, . . . , An ∈ F, then ∩jn=1 Aj ∈ F.

(iv′) If (An, n ≥ 1) is a sequence of subsets of Ω and An ∈ F for every n ≥ 1, then ∩∞n=1 An ∈ F.

(v) Also, if A, B ∈ F, then B \ A = B ∩ Ac ∈ F.

Terminology. The pair (Ω, F) is called a measurable space and the events belonging to F are said to be F-measurable, that is, they are the events that one can decide on whether they happened or not, given the information F. In other words, if one knows the information F, then one is able to tell to which events of F (= subsets of Ω) does the realization of the experiment ω belong.

Example 1.2. For a generic set Ω, the following are always σ-fields:

F0 = {∅, Ω} (= trivial σ-field).

P(Ω) = {all subsets of Ω} (= complete σ-field).

Example 1.3. Let Ω = {1, . . . , 6}. The following are σ-fields on Ω:

F1 = {∅, {1}, {2, . . . , 6}, Ω}.

F2 = {∅, {1, 3, 5}, {2, 4, 6}, Ω}.

F3 = {∅, {1, 2, 3}, {4, 5, 6}, Ω}.

Example 1.4. Let Ω = [0, 1] and I1, . . . , In be a family of disjoint intervals in Ω such that I1 ∪ . . . ∪ In = Ω (I1, . . . , In) is also called a partition of Ω). The following is a σ-field on Ω:

F4 = {∅, [0, 1], . . . , In, I1 ∪ I2, . . . , I1 ∪ I2 ∪ I3, . . . , Ω} (NB: there are 2n events in total in F4)

In the discrete setting (that is, in a σ-field with a finite or countable number of elements), the smallest elements contained in a σ-field are called the atoms of the σ-field. Formally, F ∈ F is an atom of F if for any G ∈ F such that G ⊂ F, it holds that either G = ∅ or G = F. In the above example with Ω = [0, 1], the atoms of F4 are therefore I1, . . . , In. Notice moreover that a σ-field with n atoms has 2^n elements, so that the number of elements of a finite σ-field is always a power of 2.
1.2  \( \sigma \)-field generated by a collection of events

An event carries in general more information than itself. As an example, if one knows whether the result of a die roll is odd (corresponding to the event \( \{1, 3, 5\} \)), then one also knows of course whether the result is even (corresponding to the event \( \{2, 4, 6\} \)). It is therefore convenient to have a mathematical description of the information generated by a single event, or more generally by a family of events.

**Definition 1.5.** Let \( \mathcal{A} = \{A_i, i \in I\} \) be a collection of subsets of \( \Omega \) (where \( I \) need not be finite nor countable). The \( \sigma \)-field generated by \( \mathcal{A} \) is the smallest \( \sigma \)-field on \( \Omega \) containing all the events \( A_i \). It is denoted as \( \sigma(\mathcal{A}) \).

**Remark.** A natural question is whether such a vague definition makes sense. Observe first that there is always at least one \( \sigma \)-field containing \( \mathcal{A} \): it is \( \mathcal{P}(\Omega) \). Then, one can show that an arbitrary intersection of \( \sigma \)-fields is still a \( \sigma \)-field. One can therefore provide the following alternative definition of \( \sigma(\mathcal{A}) \): it is the intersection of all \( \sigma \)-fields containing the collection \( \mathcal{A} \), which is certainly a well-defined object.

**Example.** Let \( \Omega = \{1, \ldots, 6\} \) (cf. Example 1.3).

Let \( \mathcal{A}_1 = \{\{1\}\} \). Then \( \sigma(\mathcal{A}_1) = \mathcal{F}_1 \).

Let \( \mathcal{A}_2 = \{\{1, 3, 5\}\} \). Then \( \sigma(\mathcal{A}_2) = \mathcal{F}_2 \).

Let \( \mathcal{A}_3 = \{\{1, 2, 3\}\} \). Then \( \sigma(\mathcal{A}_3) = \mathcal{F}_3 \).

Let \( \mathcal{A} = \{\{1\}, \ldots, \{6\}\} \). Then \( \sigma(\mathcal{A}) = \mathcal{P}(\Omega) \).

**Exercise.** Let \( \mathcal{A} = \{\{1, 2, 3\}, \{1, 3, 5\}\} \). Compute \( \sigma(\mathcal{A}) \).

**Example.** Let \( \Omega = [0, 1] \) and let \( \mathcal{A}_4 = \{I_1, \ldots, I_n\} \) (cf. Example 1.4). Then \( \sigma(\mathcal{A}_4) = \mathcal{F}_4 \). This is a particular instance of the fact that in the discrete case, a \( \sigma \)-field is always generated by the collection of its atoms.

**Borel \( \sigma \)-field on \([0, 1]\).** A very important example of generated \( \sigma \)-field on \( \Omega = [0, 1] \) is the following:

\[
\mathcal{B}([0, 1]) = \sigma(\{\{0\}, \{1\}, [a, b]: a, b \in [0, 1], a < b\})
\]

is the Borel \( \sigma \)-field on \([0, 1]\) and elements of \( \mathcal{B}([0, 1]) \) are called the Borel subsets of \([0, 1]\). As surprising as it may be, it turns out that \( \mathcal{B}([0, 1]) \neq \mathcal{P}([0, 1]) \) [without proof], which generates some difficulties from the theoretical point of view. Nevertheless, it is quite difficult to construct explicit examples of subsets of \([0, 1]\) which are not in \( \mathcal{B}([0, 1]) \). Notice indeed that

a) All singletons belong to \( \mathcal{B}([0, 1]) \). Indeed, for any \( 0 < x < 1 \), \( \{x\} = \bigcap_{n \geq 1} \left[x - \frac{1}{n}, x + \frac{1}{n}\right] \) belongs to \( \mathcal{B}([0, 1]) \), by the property seen above and the fact that the Borel \( \sigma \)-field is by definition the smallest \( \sigma \)-field containing all open intervals.

b) Therefore, all closed intervals, being unions of open intervals and singletons, also belong to \( \mathcal{B}([0, 1]) \).

c) Likewise, all countable intersections of open intervals \( \mathcal{B}([0, 1]) \), as well as all countable unions of closed intervals belong to \( \mathcal{B}([0, 1]) \).

d) The story goes on with countable unions of countable intersections of open intervals, etc. Even though the list is quite long, not all the subsets of \([0, 1]\) are part of \( \mathcal{B}([0, 1]) \), as mentioned above.

**Remark.** In general, the \( \sigma \)-field generated by a collection of events contains many more elements than the collection itself! The Borel \( \sigma \)-field is a good example. In the finite case, you will observe the same phenomenon while computing \( \sigma(\{\{1, 2, 3\}, \{1, 3, 5\}\}) \) on \( \Omega = \{1, \ldots, 6\} \).

**Remark.** It can be easily checked that the atoms of \( \mathcal{B}([0, 1]) \) are the singletons \( \{x\}, x \in [0, 1] \). Nevertheless, one can check that \( \mathcal{B}([0, 1]) \) is not generated by its atoms (as it is not a discrete \( \sigma \)-field). As a proof of this (exercise), compute what \( \sigma(\{\{x\}, x \in [0, 1]\}) \) is.
Borel σ-field on \( \mathbb{R} \) and \( \mathbb{R}^2 \).

**Definition 1.6.** On the set \( \mathbb{R} \), one defines

\[
B(\mathbb{R}) = \sigma(\{[a,b]: a, b \in \mathbb{R}, a < b\})
\]

The elements of \( B(\mathbb{R}) \) are called Borel sets on \( \mathbb{R} \). Again, notice that \( B(\mathbb{R}) \) is strictly included in \( \mathcal{P}(\mathbb{R}) \).

**Definition 1.7.** On the set \( \mathbb{R}^2 \), one defines

\[
B(\mathbb{R}^2) = \sigma(\{[a,b] \times [c,d]: a, b, c, d \in \mathbb{R}, a < b, c < d\})
\]

Notice that even though \( B(\mathbb{R}^2) \) is generated by rectangles only, it contains all kinds of shapes in \( \mathbb{R}^2 \), including in particular discs and triangles (because every disc and triangle can be seen as a countable union of rectangles). Here again, one sees that the σ-field generated by a collection of events is much larger than the collection of events itself.

Finally, notice that a straightforward generalization of the above definition allows to define \( B(\mathbb{R}^n) \) for arbitrary \( n \). Even more generally, \( B(\Omega) \) can be defined for \( \Omega \) a Hilbert / metric / topological space.

## 1.3 Sub-σ-field

One may have more or less information about a system. In mathematical terms, this translates into the fact that a σ-field contains more or less elements. It is therefore convenient to introduce a (partial) ordering on the ensemble of existing σ-fields, in order to establish a hierarchy of information. This notion of hierarchy is important and will come back when we will be studying stochastic processes that evolve in time.

**Definition 1.8.** Let \( \Omega \) be a set and \( \mathcal{F} \) be a σ-field on \( \Omega \). A sub-σ-field of \( \mathcal{F} \) is a collection \( \mathcal{G} \) of events such that:

(i) If \( A \in \mathcal{G} \), then \( A \in \mathcal{F} \).

(ii) \( \mathcal{G} \) is itself a σ-field.

**Notation.** \( \mathcal{G} \subset \mathcal{F} \).

**Remark.** Let \( \Omega \) be a generic set. The trivial σ-field \( \mathcal{F}_0 = \{\emptyset, \Omega\} \) is a sub-σ-field of any other σ-field on \( \Omega \). Likewise, any σ-field on \( \Omega \) is a sub-σ-field of the complete σ-field \( \mathcal{P}(\Omega) \).

**Example.** Let \( \Omega = \{1, \ldots, 6\} \) (cf. Example 1.3). Notice that \( \mathcal{F}_1 \) is not a sub-σ-field of \( \mathcal{F}_2 \) (even though \( \{1\} \subset \{1, 3, 5\} \)), nor is \( \mathcal{F}_2 \) a sub-σ-field of \( \mathcal{F}_1 \). In general, notice that

1) If \( A \in \mathcal{G} \) and \( \mathcal{G} \subset \mathcal{F} \), then it is true that \( A \in \mathcal{F} \).

but

2) \( A \subset B \) and \( B \in \mathcal{G} \) together do not imply that \( A \in \mathcal{G} \).

**Example.** Let \( \Omega = [0, 1] \) (cf. Example 1.4). Then \( \mathcal{F}_4 \) is a sub-σ-field of \( B([0, 1]) \). Also, if \( \mathcal{F}_5 = \sigma(J_1, \ldots, J_m) \), where \( \{J_1, \ldots, J_m\} \) represents a finer partition of the interval \([0, 1]\) (i.e., each interval \( I \) of \( \mathcal{F}_4 \) is a disjoint union of intervals \( J \)), then \( \mathcal{F}_4 \subset \mathcal{F}_5 \).

## 1.4 Random variables

The notion of random variable is usually introduced in elementary probability courses as a vague concept, essentially characterized by its distribution. In mathematical terms however, random variables do exist prior to their distribution: they are functions from the fundamental set \( \Omega \) to \( \mathbb{R} \) satisfying a measurability property.
Definition 1.9. Let $(\Omega, \mathcal{F})$ be a measurable space. A random variable on $(\Omega, \mathcal{F})$ is a map $X : \Omega \to \mathbb{R}$ satisfying
\begin{equation}
\{ \omega \in \Omega : X(\omega) \in B \} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}) \tag{1}
\end{equation}

Notation. One often simply denotes the set \( \{ \omega \in \Omega : X(\omega) \in B \} = \{ X \in B \} = X^{-1}(B) \): it is called the inverse image of the set \( B \) through the map \( X \) (watch out that \( X \) need not be a bijective function in order for this set to be well defined).

Terminology. The above random variable \( X \) is sometimes called \( \mathcal{F} \)-measurable, in order to emphasize that if one knows the information \( \mathcal{F} \), then one knows the value of \( X \).

Example. If \( \mathcal{F} = \mathcal{P}(\Omega) \), then condition (1) is always satisfied, so every map \( X : \Omega \to \mathbb{R} \) is an \( \mathcal{F} \)-measurable random variable. On the contrary, if \( \mathcal{F} = \{ \emptyset, \Omega \} \), then the only random variables which are \( \mathcal{F} \)-measurable are the maps \( X : \Omega \to \mathbb{R} \) which are constant.

Remark. Condition (1) can be shown to be equivalent to the following condition: [without proof]
\[ \{ \omega \in \Omega : X(\omega) \leq t \} \in \mathcal{F}, \quad \forall t \in \mathbb{R} \]
which is significantly easier to check.

Definition 1.10. Let \( (\Omega, \mathcal{F}) \) be a measurable space and \( A \in \mathcal{F} \) be an event. Then the map \( \Omega \to \mathbb{R} \) defined as
\[ \omega \mapsto 1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases} \]
is a random variable on \( (\Omega, \mathcal{F}) \). It is called the indicator function of the event \( A \).

Example 1.11. Let \( \Omega = \{1, \ldots, 6\} \) and \( \mathcal{F} = \mathcal{P}(\Omega) \) (cf. Example 1.3). Then \( X_1(\omega) = \omega \) and \( X_2(\omega) = 1_{\{1,3,5\}}(\omega) \) are both random variables on \( (\Omega, \mathcal{F}) \). Moreover, \( X_2 \) is \( \mathcal{F}_2 \)-measurable, but notice that \( X_1 \) is neither \( \mathcal{F}_1 \)- nor \( \mathcal{F}_2 \)-measurable.

Example 1.12. Let \( \Omega = [0,1] \) and \( \mathcal{F} = \mathcal{B}([0,1]) \) (cf. Example 1.4). Then \( X_3(\omega) = \omega \) and \( X_4(\omega) = \sum_{j=1}^n x_j 1_{I_j}(\omega) \) are both random variables on \( (\Omega, \mathcal{F}) \). Notice however that only \( X_4 \) is \( \mathcal{F}_4 \)-measurable.

We will need to consider not only random variables, but also functions of random variables. This is why we introduce the following definition.

Definition 1.13. A map \( g : \mathbb{R} \to \mathbb{R} \) such that
\[ \{ x \in \mathbb{R} : g(x) \in B \} \in \mathcal{B}(\mathbb{R}), \quad \forall B \in \mathcal{B}(\mathbb{R}) \]
is called a Borel-measurable function on \( \mathbb{R} \).

Remark. A Borel-measurable function on \( \mathbb{R} \) is therefore nothing but a random variable on the measurable space \( (\mathbb{R}, \mathcal{B}(\mathbb{R})) \).

Notation. Again, one often uses the shorthand notations \( \{ x \in \mathbb{R} : g(x) \in B \} = \{ g \in B \} = g^{-1}(B) \), but this does not mean that \( g \) is invertible!

As it is difficult to construct explicitly sets which are not Borel sets, it is equally difficult to construct functions which are not Borel-measurable. Nevertheless, one often needs to check that a given function is Borel-measurable. A useful criterion for this is the following [without proof].

Proposition 1.14. If \( g : \mathbb{R} \to \mathbb{R} \) is continuous, then it is Borel-measurable.

Finally, let us mention this useful property of functions of random variables.

Proposition 1.15. If \( X \) is an \( \mathcal{F} \)-measurable random variable and \( g : \mathbb{R} \to \mathbb{R} \) is Borel-measurable, then \( Y = g(X) \) is also an \( \mathcal{F} \)-measurable random variable.
Proof. Let $B \in \mathcal{B}(\mathbb{R})$. Then

$$\{Y \in B\} = \{g(X) \in B\} = \{X \in g^{-1}(B)\} \in \mathcal{F}$$

since $X$ is an $\mathcal{F}$-measurable random variable and $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$ by assumption.

The above proposition is saying no more than the following: assume that knowing the information $\mathcal{F}$ allows you to determine the value of $X$. Then knowing this same information $\mathcal{F}$ also gives you the value of $Y = g(X)$.

### 1.5 $\sigma$-field generated by a collection of random variables

The amount of information contained in a random variable, or more generally in a collection of random variables, is given by the definition below.

**Definition 1.16.** Let $(\Omega, \mathcal{F})$ be a measurable space and $\{X_i, i \in I\}$ be a collection of random variables on $(\Omega, \mathcal{F})$. The $\sigma$-field generated by $X_i, i \in I$, denoted as $\sigma(X_i, i \in I)$, is the smallest $\sigma$-field $\mathcal{G}$ on $\Omega$ such that all the random variables $X_i$ are $\mathcal{G}$-measurable.

**Remark.** Notice that

$$\sigma(X_i, i \in I) = \sigma(\{\{X_i \in B\}, i \in I, B \in \mathcal{B}(\mathbb{R})\})$$

where the right-hand side expression refers to Definition 1.5. It turns out that one also has [without proof]

$$\sigma(X_i, i \in I) = \sigma(\{\{X_i \leq t\}, i \in I, t \in \mathbb{R}\})$$

**Example.** Let $(\Omega, \mathcal{F})$ be a measurable space. If $X_0$ is a constant random variable (i.e. $X_0(\omega) = c \in \mathbb{R}, \forall \omega \in \Omega$), then $\sigma(X_0) = \{\emptyset, \Omega\}$.

**Example.** Let $\Omega = \{1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (cf. Examples 1.3 and 1.11). Then $\sigma(X_1) = \mathcal{P}(\Omega)$ and $\sigma(X_2) = \mathcal{F}_2$.

**Example.** Let $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$ (cf. Examples 1.4 and 1.12). Then $\sigma(X_3) = \mathcal{B}([0, 1])$ and $\sigma(X_4) = \mathcal{F}_4$.

The $\sigma$-field $\sigma(X)$ can be seen as the information carried by the random variable $X$. By definition, a random variable $X$ is always $\sigma(X)$-measurable. Following the proof of Proposition 1.15, one can also show the proposition below.

**Proposition 1.17.** If $X$ is a random variable on a measurable space $(\Omega, \mathcal{F})$ and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then $Y = g(X)$ is a $\sigma(X)$-measurable random variable, which is equivalent to saying that $\sigma(Y) \subset \sigma(X)$: the information carried by $Y$ is in general less than that carried by $X$.

Notice that it can be strictly less: if you think e.g. about the case $Y = X^2$, then the information about the sign of $X$ is lost in $Y$; on the other hand, if the function $g$ is invertible (meaning that one can write $X = g^{-1}(Y)$), then $\sigma(Y) = \sigma(X)$.

A further generalization of Proposition 1.17 is the following: if $g : \mathbb{R}^2 \to \mathbb{R}$ is Borel-measurable and $Y = g(X_1, X_2)$, where $X_1, X_2$ are two random variables, then $Y$ is a $\sigma(X_1, X_2)$-measurable random variable, or put differently, $\sigma(Y) \subset \sigma(X_1, X_2)$. The other inclusion $\sigma(X_1, X_2) \subset \sigma(Y)$ is of course not true in general, as the two random variables $(X_1, X_2)$ carry potentially more information than the single random variable $Y$.

**Final remark.** It turns out that the reciprocal statement of Proposition 1.17 is also true: if $Y$ is a $\sigma(X)$-measurable random variable, then there exists a Borel-measurable function $g : \mathbb{R} \to \mathbb{R}$ such that $Y = g(X)$ [without proof].
2  Probability measures and distributions

2.1  Probability measures

Definition 2.1. Let \((\Omega, \mathcal{F})\) be a measurable space. A probability measure on \((\Omega, \mathcal{F})\) is a map \(\mathbb{P} : \mathcal{F} \rightarrow [0,1]\) satisfying the following axioms:

(i) \(\mathbb{P}(\emptyset) = 0\) and \(\mathbb{P}(\Omega) = 1\).

(ii) If \(A_1, \ldots, A_n \in \mathcal{F}\) are disjoint (i.e. \(A_j \cap A_k = \emptyset\) for all \(1 \leq j \neq k \leq n\)), then \(\mathbb{P}(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} \mathbb{P}(A_j)\).

(ii') If \((A_n, n \geq 1)\) is a collection of disjoint events in \(\mathcal{F}\), then \(\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)\).

The following properties can be further deduced from the above axioms (proofs are left as exercise):

(iii) If \(A_1, \ldots, A_n \in \mathcal{F}\), then \(\mathbb{P}(\bigcup_{j=1}^{n} A_j) \leq \sum_{j=1}^{n} \mathbb{P}(A_j)\).

(iii') If \((A_n, n \geq 1)\) is a collection of events in \(\mathcal{F}\), then \(\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)\).

(iv) If \(A, B \in \mathcal{F}\) and \(A \subseteq B\), then \(\mathbb{P}(A) \leq \mathbb{P}(B)\) and \(\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)\). Also, \(\mathbb{P}(A^c) = 1 - \mathbb{P}(A)\).

(v) If \(A, B \in \mathcal{F}\), then \(\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\). This formula generalizes to the countable union of an arbitrary number of sets: it is called the inclusion-exclusion formula.

(vi) If \((A_n, n \geq 1)\) is a collection of events in \(\mathcal{F}\) such that \(A_n \subseteq A_{n+1}, \forall n \geq 1\), then \(\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mathbb{P}(A_n)\).

(vi') If \((A_n, n \geq 1)\) is a collection of events in \(\mathcal{F}\) such that \(A_n \supseteq A_{n+1}, \forall n \geq 1\), then \(\mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mathbb{P}(A_n)\).

Terminology. The triple \((\Omega, \mathcal{F}, \mathbb{P})\) is called a probability space. Properties (ii), resp. (ii'), are referred to as the additivity, resp. \(\sigma\)-additivity, of probability measures. Properties (iii), resp. (iii'), are referred to as the subadditivity, resp. sub-\(\sigma\)-addivity, of probabilty measures (or more prosaically as the union bound sometimes).

Example. Let \(\Omega = \{1, \ldots, 6\}\) and \(\mathcal{F} = \mathcal{P}(\Omega)\) be the measurable space associated to a die roll. The probability measure associated to a balanced die is defined as

\[
\mathbb{P}_1(\{i\}) = \frac{1}{6}, \forall i \in \{1, \ldots, 6\}
\]

and is extended by additivity to all subsets of \(\Omega\). E.g.,

\[
\mathbb{P}_1(\{1,3,5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}
\]

The probability measure associated to a loaded die is defined as

\[
\mathbb{P}_2(\{6\}) = 1 \quad \text{and} \quad \mathbb{P}_2(\{i\}) = 0, \forall i \in \{1, \ldots, 5\}
\]

and is extended by additivity to all subsets of \(\Omega\), so that for \(A \subseteq \Omega\), \(\mathbb{P}_2(A) = 1\) if \(6 \in A\) and \(\mathbb{P}_2(A) = 0\) otherwise.

In a discrete \(\sigma\)-field, once a probability measure is defined on the atoms of the \(\sigma\)-field, it is always possible to extend it by (\(\sigma\))-additivity to the whole \(\sigma\)-field. In the general case, a similar statement holds true, but the extension procedure is much more complicated.

Example. Let \(\Omega = [0,1]\) and \(\mathcal{F} = \mathcal{B}([0,1])\). One defines the following probability measure on the subintervals of \([0,1]\):

\[
\mathbb{P}(]a, b[) = b - a
\]
Fact. [without proof] Carathéodory’s extension theorem states that \( \mathbb{P} \) can be extended uniquely by \( \sigma \)-additivity to all Borel subsets of \([0, 1]\). It is called the Lebesgue measure on \([0, 1]\) and is sometimes denoted as \( \mathbb{P}(B) = |B| \). Notice that it corresponds also to the uniform distribution on \([0, 1]\).

Example. Let \( \Omega = \mathbb{R} \) and \( \mathcal{F} = \mathcal{B}(\mathbb{R}) \). One defines the following probability measure on open intervals:

\[
\mathbb{P}(]a, b[) = \int_{a}^{b} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \, dx
\]

Such a measure can again be uniquely extended to all Borel subsets of \( \mathbb{R} \): it is called the (normalized) Gaussian measure on \( \mathbb{R} \).

Remarks. - One can also define the following measure on \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\), by setting on open intervals:

\[
\mathbb{P}(]a, b[) = b - a
\]

This measure can be again uniquely extended to all Borel subsets of \( \mathbb{R} \). It is however not a probability measure, as with this definition, one sees (using the above properties) that

\[
\mathbb{P}(\mathbb{R}) = \lim_{n \to \infty} \mathbb{P}(]-n, +n[) = \lim_{n \to \infty} 2n = +\infty
\]

This measure is called the Lebesgue measure on \( \mathbb{R} \) and is again denoted as \( \mathbb{P}(B) = |B| \) for \( B \in \mathcal{B}(\mathbb{R}) \).

- We see here that defining first \( \mathbb{P} \) on the singletons \( \{x\} \) (which are the atoms of \( \mathcal{B}(\mathbb{R}) \)) instead of the open intervals \( ]a, b[ \) would not be a good idea, as we would have \( \mathbb{P}(\{x\}) = 0 \), \( \forall x \in \mathbb{R} \) for both the Gaussian measure and the Lebesgue measure on \( \mathbb{R} \), although these are clearly different.

Definition 2.2. Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. An event \( A \in \mathcal{F} \) is said to be negligible if \( \mathbb{P}(A) = 0 \), resp. almost sure (abbreviated a.s.) if \( \mathbb{P}(A) = 1 \).

Remark. The wording “almost sure” is far from ideal, but has been commonly agreed upon.

It should be emphasized that a negligible event need not be empty, nor need an almost sure event be equal to the whole space \( \Omega \). Here are examples:

- In the probability space of a loaded die (see above), the set \( \{1, 2, 3, 4, 5\} \) is a negligible event, while the singleton \( \{6\} \) is an almost sure event.

- In the probability space \(([0, 1], \mathcal{B}([0, 1]), \mathbb{P} = \text{Lebesgue measure})\), any singleton \( \{x\} \) is negligible.

Here is moreover a general statement that can be made about negligible and almost sure sets.

Proposition 2.3.

- Let \((A_n, n \geq 1)\) be a collection of negligible events in \( \mathcal{F} \). Then \( \bigcup_{n \geq 1} A_n \) is also negligible.

- Let \((B_n, n \geq 1)\) be a collection of almost sure events in \( \mathcal{F} \). Then \( \bigcap_{n \geq 1} B_n \) is also almost sure.

Proof. By the sub-\( \sigma \)-additivity property (property (iii') above),

\[
\mathbb{P}\left( \bigcup_{n \geq 1} A_n \right) \leq \sum_{n \geq 1} \mathbb{P}(A_n) = \sum_{n \geq 1} 0 = 0
\]

which proves the first claim. The second claim is a consequence of the first one: consider \( A_n = B_n^c \); then \( \mathbb{P}(A_n) = \mathbb{P}(B_n^c) = 0 \) by assumption and

\[
\mathbb{P}\left( \bigcap_{n \geq 1} B_n \right) = 1 - \mathbb{P}\left( \bigcup_{n \geq 1} A_n \right) \geq 1 - \sum_{n \geq 1} \mathbb{P}(A_n) = 1 - 0 = 1
\]

As a consequence, any countable set in \([0, 1]\) is negligible with respect to the Lebesgue measure. In particular, \( \mathbb{Q} \cap [0, 1] \) is negligible! Perhaps more surprisingly, there exist also uncountable sets in \([0, 1]\) which are negligible with respect to the Lebesgue measure (see below).
2.2 Distribution of a random variable

Definition 2.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X$ be a random variable defined on this probability space. The distribution of $X$ is the map $\mu_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined as

$$\mu_X(B) = \mathbb{P}(\{X \in B\}), \quad B \in \mathcal{B}(\mathbb{R})$$

Remark. The fact that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ implies that $\mu_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The triple $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ forms therefore a new probability space.

Notation. If a random variable $X$ has distribution $\mu$, this is denoted as $X \sim \mu$. Likewise, if two random variables $X$ and $Y$ share the same distribution $\mu$, then they are said to be identically distributed and this is denoted as $X \sim Y \sim \mu$.

Example 2.5. The probability space describing two independent (and balanced) die rolls is $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and

$$\mathbb{P}(\{(i,j)\}) = \frac{1}{36}, \quad \forall (i,j) \in \Omega$$

Let $X_1(i,j) = i$ be the result of the first die, and $Y(i,j) = i + j$ be the sum of the two dice. Then

$$\mu_{X_1}(\{i\}) = \mathbb{P}(\{X_1 = i\}) = \mathbb{P}(\{(i,1), \ldots, (i,6)\}) = \frac{6}{36} = \frac{1}{6}, \quad \forall i \in \{1, \ldots, 6\}$$

and

$$\mu_Y(\{2\}) = \mathbb{P}(\{Y = 2\}) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36}, \quad \mu_Y(\{3\}) = \mathbb{P}(\{Y = 3\}) = \mathbb{P}(\{(1,2), (2,1)\}) = \frac{1}{18}$$

More generally:

$$\mu_Y(\{i\}) = \frac{6 - |7 - i|}{36}, \quad i \in \{2, \ldots, 12\}$$

2.3 Cumulative distribution function

Definition 2.6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X$ be a random variable defined on this probability space. The cumulative distribution function (cdf) of $X$ is the map $F_X : \mathbb{R} \to [0,1]$ defined as

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

Fact. The knowledge of $F_X$ is equivalent to the knowledge of $\mu_X$ [without proof].

From the properties of probability measures, one deduces that the cdf of a random variable satisfies the following properties:

(i) $\lim_{t \to -\infty} F_X(t) = 0$, $\lim_{t \to +\infty} F_X(t) = 1$.
(ii) $F_X$ is non-decreasing, i.e. $F_X(s) \leq F_X(t)$ for all $s < t$.
(iii) $F_X$ is right-continuous on $\mathbb{R}$, i.e. $\lim_{t \uparrow t} F_X(t + \varepsilon) = F_X(t)$, for all $t \in \mathbb{R}$.

Indeed:

(i) $\lim_{n \to +\infty} F_X(t) = \lim_{n \to +\infty} F_X(n) = \lim_{n \to +\infty} \mathbb{P}(\{X \leq n\}) = 1$, as the sequence of events $\{X \leq n\}$ is an increasing sequence with $\cup_{n \geq 1} \{X \leq n\} = \Omega$. The result then follows from the use of property (vi) listed above for probability measures. A similar reasoning shows that $\lim_{t \to -\infty} F_X(t) = \lim_{n \to +\infty} F_X(-n) = \lim_{n \to +\infty} \mathbb{P}(\{X \leq -n\}) = 0$, by the fact that $\{X \leq -n\}$ is this time a decreasing sequence of events with $\cap_{n \geq 1} \{X \leq -n\} = \emptyset$ and the use of property (vi') of probability measures.

(ii) If $s \leq t$, then $\{X \leq s\} \subset \{X \leq t\}$, so $F_X(s) = \mathbb{P}(\{X \leq s\}) \leq \mathbb{P}(\{X \leq t\}) = F_X(t)$. 

10
(iii) For any \( t \in \mathbb{R} \), we have \( \lim_{\varepsilon \downarrow 0} F_X(t + \varepsilon) = \lim_{n \to \infty} F_X(t + \frac{1}{n}) = \lim_{n \to \infty} \mathbb{P}(\{X \leq t + \frac{1}{n}\}) = \mathbb{P}(\{X \leq t\}) = F_X(t) \), as again the sequence \( \{X \leq t + \frac{1}{n}\} \) is a decreasing sequence of events with \( \cap_{n \geq 1} \{X \leq t + \frac{1}{n}\} = \{X \leq t\} \).

**Important remarks.**
- Any function \( F : \mathbb{R} \to \mathbb{R} \) satisfying the above properties (i), (ii) and (iii) is the cdf of a random variable [without proof].
- One cannot show that the cdf of a random variable is left-continuous (and therefore continuous) in general. Indeed, repeating the above argument, we obtain: for any \( t \in \mathbb{R} \), \( \lim_{\varepsilon \downarrow 0} F_X(t - \varepsilon) = \lim_{n \to \infty} F_X(t - \frac{1}{n}) = \lim_{n \to \infty} \mathbb{P}(\{X \leq t - \frac{1}{n}\}) = \mathbb{P}(\{X < t\}) \), as \( \{X \leq t - \frac{1}{n}\} \) is a decreasing sequence of events with \( \cup_{n \geq 1} \{X \leq t - \frac{1}{n}\} = \{X < t\} \). But \( \mathbb{P}(\{X < t\}) \neq F_X(t) \) in general. It is wrong in particular for discrete random variables (see below).
- Any cdf \( F_X \) has at most a countable number of jumps on the real line [without proof]. If \( F_X \) has a jump of size \( p \in [0, 1] \) at \( t \in \mathbb{R} \), this actually means that \( \mathbb{P}(\{X = t\}) = F_X(t) - \lim_{\varepsilon \downarrow 0} F_X(t - \varepsilon) = p \).
This implies in particular that a discrete random variable cannot take more than a countable number of values.

### 2.4 Two important classes of random variables

**Discrete random variables.**

**Definition 2.7.** \( X \) is a discrete random variable if it takes values in a discrete (i.e., finite or countable) subset \( C \) of \( \mathbb{R} \), that is, \( X(\omega) \in C \) for every \( \omega \in \Omega \).

The distribution of a discrete random variable is entirely characterized by the numbers \( p_x = \mathbb{P}(\{X = x\}) \), where \( x \in C \). Notice that \( 0 \leq p_x \leq 1 \) for all \( x \in C \) and that \( \sum_{x \in C} p_x = \mathbb{P}(\{X \in C\}) = 1 \). The sequence of numbers \((p_x, x \in C)\) is sometimes called the probability mass function (pmf) of the random variable \( X \). It should not be confused with the probability density function (pdf) defined below for continuous random variables only.

One further has:

\[
\mu_X(B) = \mathbb{P}(\{X \in B\}) = \sum_{x \in C \cap B} p_x, \quad \forall B \in \mathcal{B}(\mathbb{R})
\]

and

\[
F_X(t) = \mathbb{P}(\{X \leq t\}) = \sum_{x \in C, x \leq t} p_x, \quad \forall t \in \mathbb{R}
\]

is a step function.

**Example.** A binomial random variable \( X \) with parameters \( n \geq 1 \) and \( p \in [0, 1] \) (denoted as \( X \sim \text{Bi}(n, p) \)) takes values in \( \{0, \ldots, n\} \) and is characterized by the numbers

\[
p_k = \mathbb{P}(\{X = k\}) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \{0, \ldots, n\}
\]

where \( \binom{n}{k} = \frac{n!}{k!(n-k)!} \) are the binomial coefficients.

**Continuous random variables.**

**Definition 2.8.** \( X \) is a continuous random variable if \( \mathbb{P}(\{X \in B\}) = 0 \) whenever \( B \in \mathcal{B}(\mathbb{R}) \) is such that \( |B| = 0 \) (remember that \( |B| \) is the Lebesgue measure of \( B \)).

In particular, this implies that if \( X \) is a continuous random variable, then \( \mathbb{P}(\{X = x\}) = 0 \ \forall x \in \mathbb{R} \) (as \( |\{x\}| = 0 \ \forall x \in \mathbb{R} \)).
**Fact.** [without proof] If $X$ is a continuous random variable according to the above definition, then there exists a Borel-measurable function $p_X : \mathbb{R} \to \mathbb{R}$, called the *probability density function* (pdf) of $X$, such that $p_X(x) \geq 0 \forall x \in \mathbb{R}$, $\int_{\mathbb{R}} p_X(x) \, dx = 1$ and

$$
\mu_X(B) = \mathbb{P}(\{X \in B\}) = \int_B p_X(x) \, dx, \quad \forall B \in \mathcal{B}(\mathbb{R})
$$

Moreover,

$$
F_X(t) = \mathbb{P}(\{X \leq t\}) = \int_{-\infty}^t p_X(x) \, dx, \quad \forall t \in \mathbb{R}
$$

is a continuous and differentiable function (whose derivative is $F'_X(t) = p_X(t)$).

**Important remarks.**
- $p_X(x) \neq \mathbb{P}(\{X = x\})$, simply because $\mathbb{P}(\{X = x\}) = 0$ for all $x \in \mathbb{R}$.
- $p_X(x) \geq 0$, but as this quantity is not a probability, it is perfectly possible that $p_X(x) > 1$ for some values of $x$. The only requirement is that the integral of $p_X(x)$ over $\mathbb{R}$ is equal to 1.

**Example.** A Gaussian random variable $X$ with mean $\mu$ and variance $\sigma^2$ (denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$) takes values in $\mathbb{R}$ and has pdf

$$
p_X(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}
$$

so in particular, $p_X(\mu) = \frac{1}{\sqrt{2\pi}\sigma^2} > 1$ if $\sigma < \frac{1}{\sqrt{2\pi}}$.

**Remark.** One could think that the only existing distributions are either discrete or continuous, or a combination of these. It turns out that life is more complicated than that! Some distributions are neither discrete, nor continuous. A famous example is the distribution whose cdf is the *devil’s staircase*, as we shall see below.

**Change of variables.** Let $X$ be a generic random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $f : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function and $Y = f(X)$. Assume we know $F_X$ the cdf of $X$; what can we say on $F_Y$ the cdf of $Y$? Not much in general, but assume for example that $f$ is increasing, i.e., that $f(x_1) < f(x_2)$ as soon as $x_1 \leq x_2$. Then $f$ is invertible, so

$$
F_Y(t) = \mathbb{P}(\{Y \leq t\}) = \mathbb{P}(\{f(X) \leq t\}) = \mathbb{P}(\{X \leq f^{-1}(t)\}) = F_X(f^{-1}(t))
$$

If in addition $X$ is a continuous random variable with pdf $p_X$ and $f$ is differentiable (with $f'(x) > 0$ for all $x \in \mathbb{R}$, as $f$ is increasing), then $Y$ is a continuous random variable also and

$$
p_Y(t) = \frac{dF_Y(t)}{dt} = \frac{dF_X(f^{-1}(t))}{dt} \frac{df^{-1}(t)}{dt} = p_X(f^{-1}(t)) \frac{df^{-1}(t)}{f'(f^{-1}(t))}
$$

Setting $x = f^{-1}(t)$ in the above relation, we get the more natural formula $p_X(x) = p_Y(f(x)) f'(x)$.

Similar reasonings allow to deal with $f$ decreasing or more general cases.

**Remark.** In the case where $X$ is a continuous random variable and $f$ is assumed to be non-decreasing only (i.e., $f(x_1) \leq f(x_2)$ if $x_1 < x_2$), notice that $f$ is not necessarily invertible in this case, so that the above formulas do not hold. Also, $Y = f(X)$ need not be a continuous random variable in this case (consider for example the case where $f(x) = 1_{\{x \geq 0\}}$; $Y$ is a then discrete random variable taking values in $\{0, 1\}$ only).

---

1 This fact is known as the Radon-Nikodym theorem. It has many different formulations and important applications in probability theory.
2.5 The Cantor set and the devil’s staircase

The Cantor set is a subset $C$ of $[0, 1]$ obtained by removing recursively “middle intervals” as follows:

Mathematically, for $n \geq 1$, let

$$A_n = \bigcup_{a_1, \ldots, a_{n-1} \in \{0, 2\}} \left[ \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{1}{3^n}, \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{2}{3^n} \right],$$

be the set of (open) intervals removed at stage $n$. In particular, $A_1 = \left( \frac{1}{3}, \frac{2}{3} \right], A_2 = \left( \frac{1}{9}, \frac{2}{9} \right] \cup \left( \frac{7}{9}, \frac{8}{9} \right], \text{ etc.}$

Notice also that $A_n$ and $A_m$ are disjoint for every $n \neq m$. The Cantor set $C$ is then defined as

$$C = [0, 1] \setminus \bigcup_{n \geq 1} A_n$$

This set has strange properties: first observe that the Lebesgue measure of each $A_n$ is given by

$$|A_n| = \frac{2^{n-1}}{3^n}$$

(each set $A_n$ is indeed made of $2^{n-1}$ disjoint intervals, each of length $\frac{1}{3^n}$)

so that, using the fact that the $A_n$ are disjoint as well as the formula for geometric series, we obtain:

$$|C| = 1 - \sum_{n \geq 1} \frac{2^{n-1}}{3^n} = 1 - \frac{1}{2} \left( \frac{1}{1 - \frac{2}{3}} - 1 \right) = 1 - \frac{1}{2} (3 - 1) = 0$$

The Cantor set $C$ has therefore Lebesgue measure 0. Surprisingly, $C$ is also uncountable. This can be seen as follows: any number $x \in [0, 1]$ can be written using its binary decomposition:

$$x = \sum_{n \geq 1} \frac{b_n}{2^n}, \quad \text{where} \quad b_n \in \{0, 1\} \quad (2)$$

Likewise, any number $x \in [0, 1]$ can also be written using its ternary decomposition:

$$x = \sum_{n \geq 1} \frac{a_n}{3^n}, \quad \text{where} \quad a_n \in \{0, 1, 2\}$$

It is a fact [without proof] that any $x \in C$ can be written as

$$x = \sum_{n \geq 1} \frac{a_n}{3^n}, \quad \text{where} \quad a_n \in \{0, 2\} \quad (3)$$

i.e., $x \in C$ if and only if $a_n \neq 1$ for every $n \geq 1$. Comparing formulas (2) and (3), we see that the sets $[0, 1]$ and $C$ are in bijection with each other (the bijection being $b_n = 0 \leftrightarrow a_n = 0$ and $b_n = 1 \leftrightarrow a_n = 2$), proving that $C$ is uncountable, because $[0, 1]$ is (the proof that the set $[0, 1]$ is uncountable is by the way also due to Cantor and is called the diagonalization argument).
Let us now turn to the devil’s staircase. This strange cdf has the following shape:

It can be defined recursively as follows: \( F(t) = \frac{1}{2} \) for \( t \in ]\frac{1}{3}, \frac{2}{3}[ \), \( F(t) = \frac{1}{4} \) for \( t \in ]\frac{1}{9}, \frac{2}{9}[ \), \( F(t) = \frac{3}{4} \) for \( t \in ]\frac{7}{9}, \frac{8}{9}[ \), etc. Formally, one can define \( F \) on the sets \( A_n \) as follows:

\[
F(t) = \frac{n-1}{2n} + \frac{1}{2^n} \sum_{k=1}^{n-1} \frac{a_k}{2^k} + \frac{1}{3^n} \sum_{k=1}^{n-1} \frac{a_k}{3^k} + \frac{2}{3^n} \quad \text{for} \quad t \in \bigcup_{n \geq 1} A_n
\]

It is then a fact that \( F \) can be extended by continuity to all \( t \in [0, 1] \) [without proof, but the picture above should convince you].

Notice now its strange properties: \( F(0) = 0, F(1) = 1 \), \( F \) is non-decreasing on \([0, 1]\), and on any set \( A_n \), \( F \) is flat, so that \( F'(t) = 0 \) for all \( t \in \bigcup_{n \geq 1} A_n \), which is the complement of \( C \) on \([0, 1]\). This is saying more precisely that the set where \( F \) is flat has full Lebesgue measure on the interval \([0, 1]\), so that the function \( F \) is almost flat. Moreover, we just said above that \( F \) is continuous on the interval \([0, 1]\).

If you think for a while, all these properties seem to contradict each other, but actually, they don’t! At the beginning of the 20th century, the work of Cantor led to a revolution in mathematics…

The next question is: where to classify the devil’s staircase, i.e., is it the cdf of a continuous or of a discrete random variable?

Let us first try to see it as the cdf of a continuous variable. We have seen that \( F'(t) = 0 \) for any \( t \notin C \). Therefore, if \( F \) were to admit a pdf, this pdf would be equal to 0 almost everywhere on \([0, 1]\). But then, such a function cannot integrate to 1 on the interval \([0, 1]\). \( F \) is therefore not the cdf of a continuous random variable (even though it is itself a continuous function).

Let us now try to view \( F \) as the cdf of a discrete random variable and look for the corresponding pmf. From the definition of \( F \), it is clear that the pmf assigns no weight to elements \( t \notin C \). Using the symmetry of the function \( F \), one could then perhaps argue that the pmf should be the uniform distribution on \( C \). But as we have seen above, \( C \) is uncountable, so in particular infinite. Such a uniform discrete distribution on \( C \) does therefore not exist!

One may still argue that \( F \) is the cdf of the uniform distribution on \( C \), as well as \( F(t) = t \) is the cdf of the uniform distribution on \([0, 1]\). This raises the question: what does that mean to pick a point uniformly in \( C \)? The same question is equally valid with \( C \) replaced by \([0, 1]\), actually…
3 Independence

The notion of independence is a central notion in probability. It is usually defined for events and random variables in elementary probability courses. Nevertheless, as it will become clear below, the independence between \( \sigma \)-fields turns out to be the most natural concept (remembering that a \( \sigma \)-field is related to the amount of information one has on a system).

In the following subsections, all events, random variables and sub-\( \sigma \)-fields are defined in a common probability space \( (\Omega,F,P) \).

3.1 Independence of two events

**Definition 3.1.** Two events \( A_1, A_2 \in \mathcal{F} \) are independent if \( P(A_1 \cap A_2) = P(A_1)P(A_2) \).

**Remark.** Although this definition is quite standard, a more explanatory definition of independence of two events is given by using conditional probabilities: we say that \( A_1 \) and \( A_2 \) are independent if \( P(A_1|A_2) = P(A_1) \), which is saying that the realization of event \( A_2 \) has no influence on the probability that \( A_1 \) happens. Using the formula for the conditional probability \( P(A_1|A_2) = P(A_1 \cap A_2)/P(A_2) \), we then recover the above definition. We will come back later to conditional probability, which is a central concept in probability theory.

**Notation.** \( A_1 \perp A_2 \).

**Proposition 3.2.** If two events \( A_1, A_2 \in \mathcal{F} \) are independent, then it also holds that

\[
P(A_1 \cap A_2^c) = P(A_1)P(A_2^c), \quad P(A_1^c \cap A_2) = P(A_1^c)P(A_2) \quad \text{and} \quad P(A_1^c \cap A_2^c) = P(A_1^c)P(A_2^c)
\]

**Proof.** We show here the first equality (noticing that the other two can be proved in a similar way):

\[
P(A_1 \cap A_2^c) = P(A_1 \setminus (A_1 \cap A_2)) = P(A_1) - P(A_1 \cap A_2) = P(A_1) - P(A_1)P(A_2) = P(A_1)P(1 - P(A_2)) = P(A_1)P(A_2^c)
\]

Notice that the above proposition says actually something very natural. Let us assume for example that one rolls a balanced die with four faces. Then the events \{the outcome is 1 or 2\} and \{the outcome is even\} are independent; more precisely, the different informations associated with these events are. So the events \{the outcome is 1 or 2\} and \{the outcome is odd\} are also independent. This will motivate the extension of the definition of independence to \( \sigma \)-fields below.

3.2 Independence of two random variables

**Definition 3.3.** Two \( \mathcal{F} \)-measurable random variables \( X_1, X_2 \) are independent if

\[
P(\{X_1 \in B_1, X_2 \in B_2\}) = P(\{X_1 \in B_1\})P(\{X_2 \in B_2\}), \quad \forall B_1, B_2 \in \mathcal{B}(\mathbb{R})
\]

**Notation.** \( X_1 \perp X_2 \).

**Example.** Let \( X_0(\omega) = c \in \mathbb{R}, \forall \omega \in \Omega \) be a constant random variable. According to the above definition, \( X_0 \) is independent of any other random variable defined on \( (\Omega,F,P) \).

The above general definition with Borel sets has an interesting direct consequence.

**Proposition 3.4.** Let \( f_1, f_2 : \mathbb{R} \rightarrow \mathbb{R} \) be two Borel-measurable functions. If \( X_1, X_2 \) are independent random variables, then \( Y_1 = f_1(X_1) \) and \( Y_2 = f_2(X_2) \) are also independent random variables.
Proof. From the assumption made, we have for every \( B_1, B_2 \in \mathcal{B}(\mathbb{R}) \):

\[
\mathbb{P}(\{Y_1 \in B_1, Y_2 \in B_2\}) = \mathbb{P}(\{f_1(X_1) \in B_1, f_2(X_2) \in B_2\}) = \mathbb{P}(\{X_1 \in f_1^{-1}(B_1), X_2 \in f_2^{-1}(B_2)\})
\]

\[
= \mathbb{P}(\{X_1 \in f_1^{-1}(B_1)\}) \mathbb{P}(\{X_2 \in f_2^{-1}(B_2)\}) = \mathbb{P}(\{f_1(X_1) \in B_1\}) \mathbb{P}(\{f_2(X_2) \in B_2\})
\]

\[
= \mathbb{P}(\{Y_1 \in B_1\}) \mathbb{P}(\{Y_2 \in B_2\})
\]

Notice again that \( f_1, f_2 \) need not be invertible for the above equalities to hold: \( f_i^{-1}(B_i) \) is just a notation for the preimage of \( B_i \).

Checking the independence relation for all Borel sets \( B_1 \) and \( B_2 \) can be painful... Luckily, the following proposition helps [without proof].

**Proposition 3.5.** \( X_1, X_2 \) are independent if and only if

\[
\mathbb{P}(\{X_1 \leq t_1, X_2 \leq t_2\}) = \mathbb{P}(\{X_1 \leq t_1\}) \mathbb{P}(\{X_2 \leq t_2\}), \quad \forall t_1, t_2 \in \mathbb{R}
\]

The above welcome simplification is connected to the fact that the knowledge of the distribution \( \mu_X \) of a random variable \( X \) is equivalent to that of its cdf \( F_X \).

Further simplifications of this definition occur in the two following situations [again, without proofs]:

- Assume \( X_1, X_2 \) are two discrete random variables, taking values in a common countable set \( C \). Then \( X_1, X_2 \) are independent if and only if

\[
\mathbb{P}(\{X_1 = x_1, X_2 = x_2\}) = \mathbb{P}(\{X_1 = x_1\}) \mathbb{P}(\{X_2 = x_2\}), \quad \forall x_1, x_2 \in C
\]

**Example.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be the probability space describing two independent die rolls in Example 2.5 and let \( X_1(i, j) = i \) and \( X_2(i, j) = j \). One verifies below that these two random variables are indeed independent. It was already shown that \( \mathbb{P}(\{X_1 = i\}) = \frac{1}{6}, \forall i \in \{1, \ldots, 6\}. \) Likewise, \( \mathbb{P}(\{X_2 = j\}) = \frac{1}{6}, \forall j \in \{1, \ldots, 6\} \) and

\[
\mathbb{P}(\{X_1 = i, X_2 = j\}) = \mathbb{P}(\{(i, j)\}) = \frac{1}{36} = \mathbb{P}(\{X_1 = i\}) \mathbb{P}(\{X_2 = j\}), \quad \forall (i, j) \in \Omega
\]

so \( X_1 \) and \( X_2 \) are independent.

- Assume now \( X_1, X_2 \) are jointly continuous random variables, that is, there exists a Borel-measurable function \( p_{X_1, X_2} : \mathbb{R}^2 \to \mathbb{R}_+ \) (=joint pdf) such that

\[
\mathbb{P}(\{(X_1, X_2) \in B\}) = \int_B p_{X_1, X_2}(x_1, x_2) \, dx_1 \, dx_2, \quad \forall B \in \mathcal{B}(\mathbb{R}^2)
\]

Then \( X_1, X_2 \) are independent if and only if the function \( p_{X_1, X_2} \) can be factorized as follows:

\[
p_{X_1, X_2}(x_1, x_2) = p_{X_1}(x_1) p_{X_2}(x_2), \quad \forall (x_1, x_2) \in \mathbb{R}^2
\]

### 3.3 Independence of two sub-\( \sigma \)-fields

The above two definitions of independence for events and random variables can actually be seen as particular instances of a more general definition, concerning the independence of sub-\( \sigma \)-fields, that is to say, the independence of two different types of information one may have on a system.

**Definition 3.6.** Two sub-\( \sigma \)-fields \( \mathcal{G}_1, \mathcal{G}_2 \) of \( \mathcal{F} \) are independent if

\[
\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2), \quad \forall A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2
\]

\(^2\)This can always be assumed for discrete random variables: indeed, if \( X_1 \in C_1 \) and \( X_2 \in C_2 \), then simply consider \( C = C_1 \cup C_2 \), which is also countable.
Notation. $G_1 \perp G_2$.

One can readily check (using in particular Proposition 3.2 for the first line) that
- $A_1, A_2$ are independent according to Definition 3.3 if and only if $\sigma(A_1), \sigma(A_2)$ are independent according to Definition 3.6.
- $X_1, X_2$ are independent according to Definition 3.4 if and only if $\sigma(X_1), \sigma(X_2)$ are independent according to Definition 3.6.

### 3.4 Independence of more sub-$\sigma$-fields

The notion of independence of more than two $\sigma$-fields generalizes easily as follows.

**Definition 3.7.** Let $\{G_1, \ldots, G_n\}$ be a finite collection of sub-$\sigma$-fields of $F$. This collection is independent if
\[
P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdots P(A_n), \quad \forall A_1, \ldots, A_n \in G_n
\]

A finite collection of events $\{A_1, \ldots, A_n\}$ is declared to be independent if $\{\sigma(A_1), \ldots, \sigma(A_n)\}$ is independent, which is equivalent to saying that
\[
P(A_1^* \cap \ldots \cap A_n^*) = P(A_1^*) \cdots P(A_n^*)
\]
where $A_i^*$ = either $A_i$ or $A_i^c$, $i \in \{1, \ldots, n\}$. Notice that for $n > 2$, verifying only that
\[
P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdots P(A_n)
\]
does not suffice to guarantee independence of the whole collection of events (i.e., Proposition 3.2 does not generalize to the case $n > 2$). One can actually show that independence of $\{A_1, \ldots, A_n\}$ holds if and only if
\[
P \left( \bigcap_{j \in S} A_j \right) = \prod_{j \in S} P(A_j), \quad \forall S \subset \{1, \ldots, n\}
\]

Notice also that pairwise independence (i.e., $P(A_j \cap A_k) = P(A_j)P(A_k)$ for every $j \neq k$) is not enough to ensure the independence of the whole collection.

A finite collection of random variables $\{X_1, \ldots, X_n\}$ is similarly declared to be independent if $\{\sigma(X_1), \ldots, \sigma(X_n)\}$ is independent, which is equivalent to saying that
\[
P(\{X_1 \in B_1, \ldots, X_n \in B_n\}) = P(\{X_1 \in B_1\}) \cdots P(\{X_n \in B_n\}), \quad \forall B_1, \ldots, B_n \in B(\mathbb{R})
\]
and all the simplifications seen above apply similarly.

Finally, one can further generalize independence to an arbitrary (i.e., not necessarily countable) collection of sub-$\sigma$-fields of $F$.

**Definition 3.8.** Let $\{G_i, i \in I\}$ be an arbitrary collection of sub-$\sigma$-fields of $F$. This collection is independent if any finite subcollection $\{G_{i_1}, \ldots, G_{i_m}\}$ is independent.

Infinite collections of sub-$\sigma$-fields or random variables occur in various contexts, but most prominently when dealing with stochastic processes, as we shall see during this course.
3.5 Do independent random variables really exist?

An innocent sentence such as “Let $X_1, X_2, X_3, \ldots$ be an infinite collection of independent and identically distributed (i.i.d.) random variables…” immediately raises a question: does there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which these random variables could be defined altogether? The answer is (fortunately for the remainder of this course…): yes, but as we will see, the set $\Omega$ needed to fit all these independent random variables is quite large (to say the least!).

- Let us start by exploring what $\Omega$ is needed for only two i.i.d. random variables $X_1$ and $X_2$, each distributed according to the same distribution $\mu$ on $\mathbb{R}$, say. In this case, the set $\Omega = \mathbb{R}^2$ suffices. Indeed, for $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$, let us first define $X_1(\omega) = \omega_1$ and $X_2(\omega) = \omega_2$. Now, what about $\mathcal{F}$ and $\mathbb{P}$? For $\mathcal{F}$, we will consider the $\sigma$-field generated by the rectangles of the form

$$[a_1, b_1] \times [a_2, b_2],$$

with $a_1 < b_1$ and $a_2 < b_2$.

which is nothing but the already encountered Borel $\sigma$-field $\mathcal{B}(\mathbb{R}^2)$. As for $\mathbb{P}$, we set it to be given by

$$\mathbb{P}([a_1, b_1] \times [a_2, b_2]) = \mu([a_1, b_1]) \mu([a_2, b_2])$$

on the rectangles. Caratheodory’s extension theorem then ensures that $\mathbb{P}$ can be uniquely extended to $\mathcal{B}(\mathbb{R}^2)$. With these definitions in hand, one can check that for every $B_1, B_2 \in \mathcal{B}(\mathbb{R})$, one has

$$\mathbb{P}(\{X_1 \in B_1, X_2 \in B_2\}) = \mathbb{P}(\{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1 \in B_1, \omega_2 \in B_2\}) = \mathbb{P}(B_1 \times B_2) = \mu(B_1) \cdot \mu(B_2)$$

while

$$\mathbb{P}(\{X_1 \in B_1\}) = \mathbb{P}(\{(\omega_1, \omega_2) \in \mathbb{R}^2 : \omega_1 \in B_1\}) = \mathbb{P}(B_1 \times \mathbb{R}) = \mu(B_1) \cdot \mu(\mathbb{R}) = \mu(B_1)$$

and similarly for $B_2$, proving the claim that $X_1$ and $X_2$ are independent.

- For an infinite (yet countable, but this can be further generalized) collection of random variables, things are slightly more complicated, but the basic principle remains the same. First, the set $\Omega$ needed in this case becomes

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots) : \omega_n \in \mathbb{R}, \forall n \geq 1\} = \mathbb{R}^\mathbb{N}$$

which can be viewed either as the set of infinite sequences of real numbers, or equivalently as the set of functions from $\mathbb{N}$ to $\mathbb{R}$. We then define

$$X_n(\omega) = \omega_n, \quad \forall n \geq 1$$

Now comes the trouble: what about $\mathcal{F}$ and $\mathbb{P}$? For $\mathcal{F}$, we take as before the $\sigma$-field generated by the “rectangles” in $\mathbb{R}^\mathbb{N}$, which are of the form:

$$[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \ldots$$

where $n$ is now an arbitrary positive integer. Notice that the $\sigma$-field $\mathcal{F}$ is quite large! Nevertheless, defining $\mathbb{P}$ on the above rectangles remains simple:

$$\mathbb{P}([a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \times \mathbb{R} \times \mathbb{R} \times \ldots) = \mu([a_1, b_1]) \cdot \mu([a_2, b_2]) \cdots \mu([a_n, b_n])$$

and Caratheodory’s extension theorem ensures again that $\mathbb{P}$ can be uniquely extended to $\mathcal{F}$. It is then quite easy to see that with all these definitions,

$$\mathbb{P}(\{X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\}) = \mathbb{P}(\{X_1 \in B_1\}) \cdot \mathbb{P}(\{X_2 \in B_2\}) \cdots \mathbb{P}(\{X_n \in B_n\})$$

for any fixed $n \geq 1$, which was our aim.
4 Expectation

4.1 Discrete non-negative random variables

Let us first define what the expectation (or mean) of a random variable is in the “simple” case where $X$ is a discrete non-negative random variable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X$ be a discrete non-negative random variable defined on this space with values in a countable set $C$ (notice that $C \subset \mathbb{R}_+$, as $X$ is non-negative by assumption). Let also $(p_x, x \in C)$ denote its pmf (recall that $p_x = P(\{X = x\})$, so that $0 \leq p_x \leq 1$ for every $x \in C$ and $\sum_{x \in C} p_x = 1$).

Definition 4.1. The expectation or expected value of $X$ is defined as follows:

$$E(X) = \sum_{x \in C} x P(\{X = x\}) = \sum_{x \in C} x p_x$$

Being a countable sum of non-negative terms (as all $x \in C$ are non-negative), the above expectation can therefore take values in the interval $[0, +\infty]$, with $+\infty$ included (in case the series diverges). In order to illustrate this, let us consider some particular cases.

Example 4.2. A simple example first. Let $X$ be a Bernoulli random variable with parameter $0 < p < 1$, that is, $P(\{X = 1\}) = p = 1 - P(\{X = 0\})$. Then

$$E(X) = p \cdot 1 + (1 - p) \cdot 0 = p$$

So the expected value of $X$ is $p$. But what does that exactly mean, as $X$ never takes the value $p$? The answer will come in a few chapters.

Example 4.3 (known as Saint-Petersburg’s paradox). Here is a more puzzling example. Let us consider an i.i.d. sequence of random variables $(X_n, n \geq 1)$ with $P(\{X_1 = +1\}) = P(\{X_1 = -1\}) = \frac{1}{2}$, and define the random variables

$$T = \inf\{n \geq 1 : X_n = +1\} \quad \text{and} \quad G = 2^T$$

This models the following game: you toss a coin multiple times, until “heads” ($X_n = +1$) comes out for the first time. Denote this time $T$. Then your gain is $2^T$. How much money would you be ready to pay to be allowed to play such a game? In general, the answer to such a question is the expected gain of the game. Let us compute this expected gain in the present case:

$$E(G) = E(2^T) = \sum_{n \geq 1} 2^n P(\{T = n\}) = \sum_{n \geq 1} 2^n P(\{X_1 = \ldots = X_{n-1} = -1, X_n = +1\}) = \sum_{n \geq 1} 2^n \frac{1}{2^n} = +\infty$$

where we have used the independence of the $X$’s. The expected gain being infinite, this is potentially saying that you would be ready to pay any finite amount of money to play such a game, and still believe you would be on the winning side, as you would pay less than your expected gain in any case. But would you, really? Probably no...

There are two ways to resolve this paradox:

1) A first idea is the following: the expected gain is infinite, because one assumes that you are ready to potentially toss a coin indefinitely, which is of course not the case in practice. If we modify the game by saying that there is a time horizon $N$ after which, assuming you had only obtained “tails” until then, you are out of the game and do not gain anything, then your expected gain becomes:

$$E(G) = \sum_{n=1}^{N} 2^n \frac{1}{2^n} + 0 = N$$
This seems already like a much more reasonable amount to invest on this game! Still, it is not clear that one would be willing to invest an amount of one million in order to get the right to toss a coin a million times, as most of the times, “heads” would come up after 1, 2, 3 or 4 tosses...

2) Another (and better) way out is the following: considering the expected gain as the fair price to pay for being allowed to play the game does not reflect one’s aversion to risk. Rather than paying \( x = \mathbb{E}(G) \), many people would probably be ready to pay the amount \( y \) satisfying

\[ \sqrt{y} = \mathbb{E}(\sqrt{G}) \]

or even more risk-averse people would only consider paying the amount \( z \) satisfying

\[ \log_2(z) = \mathbb{E}(\log_2 G) \]

One can check that \( y = \frac{1}{3 - 2 \sqrt{2}} \approx 5.83 \) and that \( z = 4 \). The square root and log functions are called utility functions. These are generally taken to be increasing and concave, with their particular shape reflecting one’s risk aversion.

Remark. Why did we restrict ourselves here to non-negative random variables? The advantage of this restriction is that in this case, the expectation is always well defined, even if possibly infinite. In the general case, the sum \( \sum_{x \in C} x p_x \) may contain positive and negative terms, and may therefore “truly” diverge, making it impossible to define what \( \mathbb{E}(X) \) is. More on this in the next section.

### 4.2 General definition

From the point of view of measure theory, random variables are maps from \( \Omega \) to \( \mathbb{R} \). Correspondingly, the expectation of a random variable \( X \) is the Lebesgue integral of the map \( X \), that is, the “area under the curve \( \omega \mapsto X(\omega) \)”, where the horizontal axis is measured with the probability measure \( \mathbb{P} \).

Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space and \( X \) be a random variable defined on this probability space. The expectation of \( X \), denoted as \( \mathbb{E}(X) \), will be defined in three steps.

**Step 1.** Assume first that \( X \) is a discrete non-negative (and by definition, \( \mathcal{F} \)-measurable) random variable. We already saw the definition of \( \mathbb{E}(X) \) above, but let us observe here that such a random variable \( X \) may be written as

\[ X(\omega) = \sum_{i \geq 1} x_i 1_{A_i}(\omega), \quad \omega \in \Omega \]

where \( x_i \geq 0 \) are distinct and \( A_i = \{ \omega \in \Omega : X(\omega) = x_i \} \in \mathcal{F} \) are disjoint. The expectation of \( X \) is then defined as

\[ \mathbb{E}(X) = \sum_{i \geq 1} x_i \mathbb{P}(A_i) = \sum_{i \geq 1} x_i \mathbb{P}(\{X = x_i\}) \]

which corresponds to the definition seen above. It is therefore again the case that \( \mathbb{E}(X) \in [0, +\infty] \). Notice moreover that in the particular case where \( X = 1_A \) with \( A \in \mathcal{F} \) (which is nothing but a Bernoulli random variable), one has \( \mathbb{E}(X) = \mathbb{P}(A) \).

**Step 2.** Assume now that \( X \) is a generic \( \mathcal{F} \)-measurable non-negative random variable, i.e., \( X(\omega) \geq 0 \), \( \forall \omega \in \Omega \). Let us define the following sequence of discrete random variables:

\[ X_n(\omega) = \sum_{i \geq 1} \frac{i - 1}{2^n} 1_{A_i^{(n)}}(\omega) \quad \text{where} \quad A_i^{(n)} = \left\{ \omega \in \Omega : \frac{i - 1}{2^n} < X(\omega) \leq \frac{i}{2^n} \right\} \]

Notice that \( x_i = \frac{i - 1}{2^n} \geq 0 \) and that \( A_i^{(n)} \in \mathcal{F} \), since \( X \) is \( \mathcal{F} \)-measurable. So according to Step 1, one has for each \( n \)

\[ \mathbb{E}(X_n) = \sum_{i \geq 1} \frac{i - 1}{2^n} \mathbb{P}\left(A_i^{(n)}\right) = \sum_{i \geq 1} \frac{i - 1}{2^n} \mathbb{P}\left(\left\{\frac{i - 1}{2^n} < X \leq \frac{i}{2^n}\right\}\right) \in [0, +\infty]. \]
It should be observed that \((X_n, n \in \mathbb{N})\) is actually an non-decreasing sequence of non-negative “staircases”, that is,

\[ 0 \leq X_n(\omega) \leq X_{n+1}(\omega), \quad \forall n \]

The staircase gets indeed refined at each step, as the size of the steps is divided by 2 from \(n\) to \(n+1\). Likewise, one easily sees that \(E(X_n) \leq E(X_{n+1})\) for all \(n\), so \((E(X_n), n \in \mathbb{N})\) is an non-decreasing sequence, that therefore converges (possibly to \(+\infty\)). One defines

\[
E(X) = \lim_{n \to \infty} E(X_n) = \lim_{n \to \infty} \sum_{i \geq 1} \frac{i-1}{2^n} P\left( \left\{ \frac{i-1}{2^n} < X \leq \frac{i}{2^n} \right\} \right) \in [0, +\infty]
\]

**Step 3.** Finally, consider a generic \(\mathcal{F}\)-measurable random variable \(X\). One defines its **positive and negative parts**:

\[
X^+(\omega) = \max(0, X(\omega)), \quad X^-(\omega) = \max(0, -X(\omega))
\]

Notice that both \(X^+(\omega) \geq 0\) and \(X^-(\omega) \geq 0\), and that

\[
X^+(\omega) - X^-(\omega) = X(\omega), \quad X^+(\omega) + X^-(\omega) = |X(\omega)|
\]

In measure theory, one does not want to deal with ill-defined quantities such as \(\infty - \infty\). One therefore defines \(E(X)\) only when \(E(|X|) = E(X^+) + E(X^-) < +\infty\), using the formula:

\[
E(X) = E(X^+) - E(X^-)
\]

**Simplified expressions in two important particular cases.** [without proofs]

Let \(X\) be a random variable and \(g : \mathbb{R} \to \mathbb{R}\) be a Borel-measurable function.

- If \(X\) is a discrete random variable with values in \(C\) and \(E(|g(X)|) = \sum_{x \in C} |g(x)| P(\{X = x\}) < +\infty\), then
  \[
  E(g(X)) = \sum_{x \in C} g(x) P(\{X = x\})
  \]

- If \(X\) is a continuous random variable with pdf \(p_X\) and \(E(|g(X)|) = \int_{\mathbb{R}} |g(x)| p_X(x) \, dx < +\infty\), then
  \[
  E(g(X)) = \int_{\mathbb{R}} g(x) p_X(x) \, dx
  \]

**Terminology.** - If \(E(|X|) < \infty\), then \(X\) is said to be an **integrable** random variable.
- If \(E(X^2) < \infty\), then \(X\) is said to be a **square-integrable** random variable.
- If there exists \(c > 0\) such that \(|X(\omega)| \leq c\), \(\forall \omega \in \Omega\), then \(X\) is said to be a **bounded** random variable.
- If \(E(X) = 0\), then \(X\) is said to be a **centered** random variable.

One has the following series of implications:

\[
X \text{ is bounded } \Rightarrow X \text{ is square-integrable } \Rightarrow X \text{ is integrable}
\]

\[
X \text{ is integrable and } Y \text{ is bounded } \Rightarrow XY \text{ is integrable}
\]

\[
X, Y \text{ are both square-integrable } \Rightarrow XY \text{ is integrable}
\]

The fact that any bounded random variable is integrable follows from the simple fact that \(E(|X|) \leq C\) if \(|X(\omega)| \leq C\) for all \(\omega \in \Omega\). Any bounded random variable \(X\) is therefore also square-integrable (as \(X^2\) is bounded if \(X\) is bounded). Likewise, if \(X\) is integrable and \(|Y(\omega)| \leq C\) for all \(\omega \in \Omega\), then \(|X(\omega) Y(\omega)| \leq C |X(\omega)|\), so \(XY\) is also integrable. The other implications follow from Cauchy-Schwarz’ inequality (see next section).
Basic properties. [without proofs]

Linearity. If $c \in \mathbb{R}$ is a constant and $X, Y$ are integrable, then

$$
E(cX) = cE(X) \quad \text{and} \quad E(X + Y) = E(X) + E(Y)
$$

Remark. As simple as the above statement looks, it is actually not so easy to prove, essentially because there are many different “simple” representations of the random variable $X + Y$.

Positivity. If $X$ is integrable and non-negative, then $E(X) \geq 0$.

“Strict” positivity. If $X$ is integrable, non-negative and $E(X) = 0$, then $\mathbb{P}(|X| = 0) = 1$. One cannot indeed guarantee in this case that $X(\omega) = 0$ for all $\omega \in \Omega$, but just that the event $\{\omega \in \Omega : X(\omega) = 0\}$ is almost sure. Another way to say this is: “$X = 0$ almost surely”, often abbreviated as “$X = 0$ a.s.”

Monotonicity. If $X, Y$ are integrable and $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$, then $E(X) \geq E(Y)$.

Variance, covariance and independence.

Definition 4.4. Let $X, Y$ be two square-integrable random variables. The variance of $X$ is defined as

$$
\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 \geq 0
$$

and the covariance of $X$ and $Y$ is defined as

$$
\text{Cov}(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).
$$

Terminology. If Cov($X,Y$) = 0, then $X$ and $Y$ are said to be uncorrelated.

Facts. [without proofs] Let $c \in \mathbb{R}$ be a constant and $X,Y$ be square-integrable random variables.

a) $\text{Var}(cX) = c^2 \text{Var}(X)$.

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

In addition, if $X, Y$ are independent, then

c) $\text{Cov}(X,Y) = 0$, i.e. $E(XY) = E(X)E(Y)$ (but the reciprocal statement is wrong).

d) $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.