LECTURE NOTES[1]
ENEE 620
RANDOM PROCESSES IN COMMUNICATION AND CONTROL

Armand M. Makowski [2]

today

[2] Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742. E-mail: armand@umd.edu. Phone: (301) 405-6844

Chapter 0

Notation, conventions and terminology

In this preliminary chapter we briefly present the notation, terminology and convention to be used throughout this text.

0.1 Usual mathematical symbols

Throughout, we use $\mathbb{N}$ to denote the set $\{0, 1, \ldots\}$ of all non-negative integers, and write $\mathbb{N}_0$ to denote the set $\{1, 2, \ldots\}$ of all positive integers. We also write $\mathbb{R}$ to denote the set of all real numbers, while the notation $\mathbb{R}_+$ is reserved to represent the set $\{x \in \mathbb{R} : \ x \geq 0\}$ of all non-negative numbers. We introduce the *extended* real line to be the set $\mathbb{R}$ augmented with $\pm\infty$, namely $\overline{\mathbb{R}} = [-\infty, +\infty] = \mathbb{R} \cup \{-\infty, +\infty\}$, and we write $\overline{\mathbb{R}}_+$ to denote the *extended* positive real line, namely $\overline{\mathbb{R}}_+ = \mathbb{R} \cup \{+\infty\}$.

0.2 Countability vs. uncountability

A set S is said to be *countable* if there is a *one-to-one* (or *injective*) mapping $f : S \to \mathbb{N}_0$ – In other words, the set $f(S)$ is a subset of $\mathbb{N}_0$. The countable set S is said to be *finite* (resp. *countably infinite*) if $|f(S)| < \infty$ (resp. $|f(S)| = \infty$). We refer to a set that is *not* countable as being *uncountable*. When $|f(S)| < \infty$, say $|f(S)| = N$ for some positive integer N, the elements of S can always be labelled so that $S = \{s_1, \ldots, s_N\}$. When $|f(S)| = \infty$, the elements of S can always be labelled so that $S = \{s_1, \ldots, s_n, \ldots\}$ – Such labelings are not unique.

0.3 Displayed equations

0.4 Set theory

This section presents a brief review of some notions of Set Theory: We use $\emptyset$ to denote the empty set. Throughout, with S an arbitrary *non-empty* set, let $\mathcal{P}(S)$ denote the collection of all subsets of S (including the empty set) – We also refer to $\mathcal{P}(S)$ as the *power set* of S (sometimes also denoted 2^S).

Elementary set-theoretic operations With E and F subsets of S, we write $E \subseteq F$ when every element of E is also an element of F, and refer to this situation by saying that E is contained in F or that E is a subset of F (resp. F is a superset of E).

The union and intersection of the subsets E and F are subsets of S which are denoted $E \cup F$ and $E \cap F$, respectively. They are defined by

$$E \cup F \equiv \{s \in S : \ s \in E \text{ or } s \in F\}$$

and

$$E \cap F \equiv \{s \in S : \ s \in E \text{ and } s \in F\}$$

We also define the following basic operations:

(i) the complement E^c of E (in S):

$$E^c \equiv \{s \in S : \ s \notin E\} .$$

(ii) the (set) difference $E - F$:

$$E - F \equiv E \cap F^c = \{s \in S : \ s \notin E\} .$$

(iii) the symmetric difference $E \Delta F$:

$$E \Delta F \equiv (E - F) \cup (F - E) = (E \cap F^c) \cup (E^c \cap F) .$$

De Morgan's Laws Let I denote an arbitrary index set. With $\{E_i, \ i \in I\}$ a collection of subsets of S, we have

$$(\cup_{i \in I} E_i)^c = \cap_{i \in I} E_i^c$$

and

$$(\cap_{i \in I} E_i)^c = \cup_{i \in I} E_i^c .$$

Distributivity Let I denote an arbitrary index set. With $\{E_i,\ i \in I\}$ a collection of subsets of S and a subset F of S, we have

$(\cup_{i\in I} E_i) \cap F = \cup_{i\in I} (E_i \cap F)$ [Set intersection is distributive over set union]

and

$(\cap_{i\in I} E_i) \cup F = \cap_{i\in I} (E_i \cup F)$ [Set union is distributive over set intersection]

0.5 Collections of sets

Since subsets of S are *elements* of the power set $\mathcal{P}(S)$, we can think of a ollection of subsets of S as a subset of $\mathcal{P}(S)$. With this in mind we have the following

Subsets If $\mathcal{H}_1$ and $\mathcal{H}_2$ are collections of subsets of S, we write $\mathcal{H}_1 \subseteq \mathcal{H}_2$ to express the fact that every subset of S that belongs to $\mathcal{H}_1$ also belongs to $\mathcal{H}_2$. We then say that $\mathcal{H}_1$ is a *subset* of $\mathcal{H}_2$, or conversely that $\mathcal{H}_2$ is a *superset* of $\mathcal{H}_1$.

Intersections and unions If $\{\mathcal{H}_i,\ i \in I\}$ is a non-empty family of collections of subsets of S, i.e., $\mathcal{H}_i \subseteq \mathcal{P}(S)$ for each i in I, then their *intersection* $\cap_{i\in I}\mathcal{H}_i$ is the collection of subsets of S given by

$$\cap_{i\in I}\mathcal{H}_i \equiv \{E \in \mathcal{P}(S) : \ E \in \mathcal{H}_i,\ i \in I\}.$$

In other words, the collection $\cap_{i\in I}\mathcal{H}_i$ comprises all the subsets of S that belong simultaneously to *each* of the collections $\{\mathcal{H}_i,\ i \in I\}$. In this definition the index set I can be taken to be arbitrary.

0.6 Cartesian products

Let S_a and S_b be two arbitrary sets (possibly identical). The *Cartesian product* of S_a and S_b, denoted $S_a \times S_b$, is the set of *ordered pairs* defined by

$$S_a \times S_b \equiv \{(s_a, s_b) : \ s_a \in S_a, s_b \in S_b\}.$$

We refer to S_a and S_b as the *factors* of the Cartesian product $S_a \times S_b$. If S_c is a third set (possibly identical to either S_a or S_b), we *identify* $(S_a \times S_b) \times S_c$ with $S_a \times (S_b \times S_c)$ in the obvious manner and write $S_a \times S_b \times S_c$ for either set. The generalization to more than two factors is straightforward.

In particular, it is customary to write the Cartesian product of p copies of the same set S_a as $S_a \times \ldots \times S_a$ or simply as S_a^p.

Chapter 1

Modeling random experiments: The Kolmogorov model

A *random experiment* $\mathcal{E}$ is understood as an activity with the following characteristics: It typically has multiple possible outcomes, and the outcome of a realization of the experiment is revealed only after the experiment has been realized. Classical examples include the throw of a dice, the price of a commodity at the end of a trading day on some stock exchange, the temperature taken at noon on January 1 at the top of the Empire State Building, etc.

In these notes we use a widely accepted approach to modeling random experiments that is based on the measure-theoretic formalism proposed by Kolmogorov: According to this approach, a random experiment $\mathcal{E}$ is modeled through a *probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ where

- The set Ω lists all (elementary) outcomes (also known as samples) generated by the experiment $\mathcal{E}$; it is known as the *sample space* for the experiment.

- Events are collections of elementary outcomes, and so are subsets of Ω. The collection of events to which likelihood of occurrence can be assigned is a collection $\mathcal{F}$ of events on Ω. In many cases of interest one is forced for mathematical reasons to take $\mathcal{F}$ to be strictly smaller than the collection of all subsets of Ω.

- The "likelihood" of occurrence of events is assigned only to the events in $\mathcal{F}$, and is given by means a *probability measure* $\mathbb{P}$ defined on $\mathcal{F}$.

These objects will be given precise mathematical meanings in what follows.

1.1 Fields and σ-fields

Throughout, with S a non-empty set, let $\mathcal{S}$ denote a non-empty collection of subsets of S, so that $\mathcal{S} \subseteq \mathcal{P}(S)$.

Definition 1.1.1

The collection $\mathcal{S}$ is said to be a *field* (also known as an *algebra* in some literature) on S if the conditions (F1)-(F3) hold where

(F1) $\emptyset \in \mathcal{S}$.

(F2) Closed under complementarity: If $E \in \mathcal{S}$, then $E^c \in \mathcal{S}$.

(F3) Closed under union: If $E \in \mathcal{S}$ and $F \in \mathcal{S}$, then $E \cup F \in \mathcal{S}$.

The De Morgan's Laws have straightforward implications: The conditions (F1) and (F2) automatically imply that S is an element of the field $\mathcal{S}$. Furthermore, (F2) and (F3) automatically imply

(F4) Closed under intersection: If $E \in \mathcal{S}$ and $F \in \mathcal{S}$, then $E \cap F \in \mathcal{S}$

(F5) Closed under differences: If $E \in \mathcal{S}$ and $F \in \mathcal{S}$, then $E - F \in \mathcal{S}$, $F - E \in \mathcal{S}$ and $E \Delta F \in \mathcal{S}$

Note that (F3) implies (is in fact equivalent to) the seemingly more general statement

(F3b) Closed under finite union: For each $n = 1, 2, \ldots$, if $E_1 \in \mathcal{S}, \ldots, E_n \in \mathcal{S}$, then $\cup_{i=1}^n E_i \in \mathcal{S}$.

while (F4) implies (is in fact equivalent to) the seemingly more general statement

(F4b) Closed under finite intersection: For each $n = 1, 2, \ldots$, if $E_1 \in \mathcal{S}, \ldots, E_n \in \mathcal{S}$, then $\cap_{i=1}^n E_i \in \mathcal{S}$.

For technical reasons that will soon become apparent a stronger notion is needed.

Definition 1.1.2

The non-empty collection of $\mathcal{S}$ of subsets of S is a *σ-field* (also known as a *σ-algebra*) on S if

(F1) $\emptyset \in \mathcal{S}$.

(F2) Closed under complementarity: If $E \in \mathcal{S}$, then $E^c \in \mathcal{S}$.

(F6) Closed under countable union: With I a countable index set, if $E_i \in \mathcal{S}$ for each $i \in I$, then $\cup_{i \in I} E_i \in \mathcal{S}$.

Any σ-field is always a field since the additional property (F6) surmises (F3b) (which is itself equivalent to (F3)) – Just take I to be finite. Again, using De Morgan's Laws we conclude under (F1) and (F2) that (F6) is equivalent to the following statement:

(F6b) Closed under countable intersection: With I a countable index set, if $E_i \in \mathcal{S}$ for each $i \in I$, then $\cap_{i \in I} E_i \in \mathcal{S}$.

Any set S always carries at least two σ-fields, namely the *trivial* σ-field $\{\emptyset, S\}$ and the *full* σ-field $\mathcal{P}(S)$. With an arbitrary set S and a σ-field $\mathcal{S}$ on S, it is customary to refer to the pair $(S, \mathcal{S})$ as a *measurable* space. This is meant to suggest that it is now possible to "measure" the sets in $\mathcal{S}$ by means of a *measure* defined on $\mathcal{S}$, an idea formalized in the next section.

1.2 Additivity and measures

Let $\mathcal{S}$ denote a non-empty collection of subsets of some non-empty set S. Measuring the sets in $\mathcal{S}$ means that a notion of "size" (also referred to as "length" or "volume" or "weight" depending on the context) can be associated with such sets. This is done through a *set function* which maps any set S in $\mathcal{S}$ to a non-negative (possibly infinite) value $\mu[S]$. Of course we expect such a set function to satisfy some natural properties. Additivity is the most obvious one as it reflects the natural idea that the size of an object can be evaluated as the sum of the sizes of its "non-overlapping" components; this is formalized next.

Definition 1.2.1

With arbitrary index set I, the subsets $\{E_i, i \in I\}$ of S are said to be *pairwise disjoint*, or simply *disjoint*, if

$$E_i \cap E_j = \emptyset, \qquad \begin{matrix} i \neq j \\ i, j \in I. \end{matrix}$$

We start with the weakest form of additivity known as *finite* additivity.

Definition 1.2.2

Given a collection $\mathcal{S}$ of subsets of S, a set function $\mu : \mathcal{S} \to [0, \infty]$ is *finitely additive*, or simply *additive*, on $\mathcal{S}$ if for any *finite* collection $\{E_i,\ i \in I\}$ of elements in $\mathcal{S}$ we have

$$\mu\left[\cup_{i \in I} E_i\right] = \sum_{i \in I} \mu\left[E_i\right]$$

whenever the sets $\{E_i,\ i \in I\}$ are *disjoint*, and their union $\cup_{i \in I} E_i$ belongs to $\mathcal{S}$.

The natural setting for this definition is for $\mathcal{S}$ to be a *field* on S since then the union set $\cup_{i \in I} E_i$ automatically belongs to $\mathcal{S}$ when the sets in the finite collection $\{E_i,\ i \in I\}$ are elements of the field $\mathcal{S}$.

In order to deal with situations where the sample space is countably infinite or uncountable, we extend the definition of a finitely additive set function in very much the same way that we extended the notion of a field to that of a σ-field – This is done by allowing the additive evaluation of unions of countably many, not just finitely many, disjoint events.

Definition 1.2.3

Given a collection $\mathcal{S}$ of subsets of S, a set function $\mu : \mathcal{S} \to [0, \infty]$ is *countably additive*, or simply *σ-additive*, on $\mathcal{S}$ if for any *countable* collection $\{E_i,\ i \in I\}$ of elements in $\mathcal{S}$ we have

$$\mu\left[\cup_{i \in I} E_i\right] = \sum_{i \in I} \mu\left[E_i\right]$$

whenever the sets $\{E_i,\ i \in I\}$ are *disjoint*, and their union $\cup_{i \in I} E_i$ belongs to $\mathcal{S}$.

This time the natural setting for this definition is for $\mathcal{S}$ to be a *σ-field* on S since then the set $\cup_{i \in I} E_i$ automatically belongs to $\mathcal{S}$ when the countably many sets $\{E_i,\ i \in I\}$ are elements of the σ-field $\mathcal{S}$. On the other hand, according to Definition 1.2.3 a countably additive set function $\mu : \mathcal{S} \to [0, \infty]$ when defined on a field $\mathcal{S}$ is automatically finitely additive there.

Definition 1.2.4

Let S be an arbitrary non-empty set equipped with a σ-field $\mathcal{S}$. A σ-additive measure μ on $\mathcal{S}$ is a set function $\mu : \mathcal{S} \to [0, \infty]$ which satisfies the properties (M1)-(M2) where

(M1) $\mu\left[\emptyset\right] = 0$.

(M2) σ-additivity: For any countable collection $\{E_i,\ i \in I\}$ of disjoint subsets in $\mathcal{S}$, we have

$$\mu\left[\cup_{i\in I} E_i\right] = \sum_{i\in I} \mu\left[E_i\right].$$

A σ-additive measure is often referred simply as a measure; this terminology always assumes that its domain of definition $\mathcal{S}$ is a σ-field. Obviously any measure is also finitely additive. The qualifier "on $\mathcal{S}$" is usually dropped once it is clear from the context what is the σ-field $\mathcal{S}$ on S being used throughout the discussion. However, sometimes the qualifier "on $(S, \mathcal{S})$" is added when there might be ambiguity as to the measurable space being considered.

A measure is said to be *finite* if $\mu\left[S\right] < \infty$, in which case (M1) is automatically satisfied as a consequence of (M2) since $\mu\left[S\right] = \mu\left[S\right] + \mu\left[\emptyset\right]$ by additivity on account of the obvious relations $S = S \cup \emptyset$ and $S \cap \emptyset = \emptyset$.

With a σ-field $\mathcal{S}$ on an non-empty set S and a measure $\mu : \mathcal{S} \to [0, +\infty]$ defined on $\mathcal{S}$, it is customary to call the triple $(S, \mathcal{S}, \mu)$ a *measure space*.

1.3 Probability measures

Specializing Definition 1.2.4 we obtain the notion of a *probability measure*, a notion that will occupy a central place in further developments.

Definition 1.3.1

Let S be an arbitrary non-empty set equipped with a σ-field $\mathcal{S}$. A *probability measure* $\mu : \mathcal{S} \to \mathbb{R}_+$ on $\mathcal{S}$ is a finite measure on $\mathcal{S}$ with $\mu\left[S\right] = 1$.

Collecting earlier definitions and remarks we readily see that the set function $\mu : \mathcal{S} \to \mathbb{R}_+$ is a probability measure on $(S, \mathcal{S})$ (where $\mathcal{S}$ is a σ-field) if and only if it satisfies the following properties:

(P1) $\mu[S] = 1$.

(P2) σ-additivity: For any countable collection $\{E_i,\ i \in I\}$ of disjoint subsets in $\mathcal{S}$, we have

$$\mu\left[\cup_{i\in I} E_i\right] = \sum_{i\in I} \mu\left[E_i\right].$$

As mentioned earlier, the condition $\mu[S] = 1$ implies $\mu[\emptyset] = 0$. Moreover, if E is any event in the σ-field $\mathcal{S}$, then its complement E^c is also in the σ-field $\mathcal{S}$ with $E \cup E^c = S$, whence

$$\mu\left[E\right] + \mu\left[E^c\right] = \mu\left[S\right] = 1$$

by additivity. It follows that

$$\mu\left[E^c\right] = 1 - \mu\left[E\right], \quad E \in \mathcal{S}$$

and

$$0 \leq \mu\left[E\right] \leq 1, \quad E \in \mathcal{S}.$$

In other words, $\{\mu\left[E\right],\ E \in \mathcal{S}\} \subseteq [0,1]$ and a probability measure is a set function $\mu : \mathcal{S} \to [0,1]$, note merely $\mu : \mathcal{S} \to \mathbb{R}_+$!

1.4 Probability models

As likelihood assignments are implemented through probability measures, we are now ready to introduce the basic model that we will adopt in the study of random phenomena (with the usual change of notation).

Definition 1.4.1

A *probability model* for the random experiment $\mathcal{E}$ is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where the set Ω is the sample space for the experiment, $\mathcal{F}$ is a σ-field of events on Ω and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ (or simply on $\mathcal{F}$).

We refer to $(\Omega, \mathcal{F}, \mathbb{P})$ as a *probability space* (or as a *probability triple*). An event E in $\mathcal{F}$ such that $\mathbb{P}\left[E\right] = 1$ is called a *certain* event, whereas an event E in $\mathcal{F}$ such that $\mathbb{P}\left[E\right] = 0$ is called a *null* event. Next we present simple, yet useful, consequences of the definitions (F1)-(F5) and (P1)-(P2); proofs are elementary and left to the interested reader as exercises [Exercise 1.8] – Some have already been given.

Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, with events E and F in $\mathcal{F}$, we have

(i) Complementarity:

$$\mathbb{P}\left[E^c\right] = 1 - \mathbb{P}\left[E\right].$$

(ii) Generalizing additivity:

$$\mathbb{P}\left[E \cup F\right] = \mathbb{P}\left[E\right] + \mathbb{P}\left[F\right] - \mathbb{P}\left[E \cap F\right]$$

so that

$$\mathbb{P}\left[E \cup F\right] \leq \mathbb{P}\left[E\right] + \mathbb{P}\left[F\right].$$

(iii) Monotonicity (I):

$$\mathbb{P}[F] = \mathbb{P}[F - E] + \mathbb{P}[E], \quad E \subseteq F.$$

(iv) Monotonicity (II):

$$\mathbb{P}[E] \leq \mathbb{P}[F], \quad E \subseteq F.$$

(v) Monotonicity (III):

$$0 \leq \mathbb{P}[E] \leq 1.$$

1.5 Discrete probability models and pmfs

In many applications a major question is concerned with determining the probability measure $\mathbb{P}$ that captures the salient features of the experiment $\mathcal{E}$ under consideration once its sample space Ω has been identified. This requires that the σ-field $\mathcal{F}$ of events be judiciously chosen.

A situation of particular importance arises when Ω is *countable*, in which case it is customary to take $\mathcal{F} = \mathcal{P}(\Omega)$ – This choice reflects the natural desire to assign the likelihood of occurrence to the *individual* outcomes $\{\{\omega\}, \, \omega \in \Omega\}$ (so that anticipating on the material of Section 1.7 we must have $\sigma\left(\{\{\omega\}, \, \omega \in \Omega\}\right) = \mathcal{P}(\Omega)$ [Exercise 1.14]). We refer to such models as *discrete probability models*.

As we now argue, specifying $\mathbb{P}$ on $(\Omega, \mathcal{P}(\Omega))$ is *equivalent* to specifying

$$\{\mathbb{P}[\{\omega\}], \, \omega \in \Omega\}. \tag{1.1}$$

Indeed, if $\mathbb{P}$ has been specified on $(\Omega, \mathcal{P}(\Omega))$, then obviously the values (1.1) are known since $\{\omega\}$ is (an event) in $\mathcal{P}(\Omega)$ for every sample ω in Ω. Conversely, if the values $\{\mathbb{P}[\{\omega\}], \, \omega \in \Omega\}$ were only available, then the obvious relation

$$E = \cup_{\omega \in E} \{\omega\}, \quad E \in \mathcal{P}(\Omega)$$

implies

$$\mathbb{P}[E] = \sum_{\omega \in E} \mathbb{P}[\{\omega\}], \quad E \in \mathcal{P}(\Omega)$$

by the σ-additivity of $\mathbb{P}$ since every subset of the countable sample space Ω is necessarily countable. This shows that the values $\{\mathbb{P}[\{\omega\}], \, \omega \in \Omega\}$ indeed uniquely specify $\mathbb{P}$ on the whole σ-field $\mathcal{P}(\Omega)$, an observation which leads to the following elementary fact.

Fact 1.5.1 *With Ω a countable set, any probability measure $\mathbb{P}$ on the σ-field $\mathcal{P}(\Omega)$ can be uniquely represented by a collection $\{p(\omega),\ \omega \in \Omega\}$ satisfying*

$$0 \leq p(\omega) \leq 1, \quad \omega \in \Omega \quad \textit{and} \quad \sum_{\omega \in \Omega} p(\omega) = 1, \tag{1.2}$$

with the identification $\mathbb{P}\left[\{\omega\}\right] = p(\omega)$ for each ω in Ω. We necessarily have

$$\mathbb{P}\left[E\right] = \sum_{\omega \in E} p(\omega), \quad E \in \mathcal{P}(\Omega). \tag{1.3}$$

We often refer to a collection $\{p(\omega),\ \omega \in \Omega\}$ satisfying (1.2) as a *probability mass function* (pmf) on Ω, written $(p(\omega),\ \omega \in \Omega)$.

1.6 Uniform probability assignments

Let Ω be an arbitrary set to be used as the sample space of a probabilistic experiment $\mathcal{E}$ whose outcomes are known or believed to be *equally* likely to occur – In many literatures an outcome of Ω so selected is said to be selected *at random*. We will avoid this usage and use instead the more accurate terminology whereby the outcomes are *uniformly* generated. A natural question is how to construct the corresponding probability measure $\mathbb{P}$, hereafter referred to as the *uniform* probability measure. Obviously such a construction also requires that we simultaneously identify the appropriate σ-field $\mathcal{F}$ of events on which $\mathbb{P}$ is defined.

In a first step it may seem intuitively reasonable to start building this uniform probability measure by assigning the *same* probability of occurrence to *all* outcomes. This would necessarily require that $\{\{\omega\},\ \omega \in \Omega\} \subseteq \mathcal{F}$, i.e.,

$$\{\omega\} \in \mathcal{F}, \quad \omega \in \Omega \tag{1.4}$$

with

$$\mathbb{P}\left[\{\omega\}\right] = p, \quad \omega \in \Omega \tag{1.5}$$

for some p in $[0,1]$. Again, anticipating on the material of Section 1.7 we must have $\sigma\left(\{\{\omega\},\ \omega \in \Omega\}\right) \subseteq \mathcal{F}$.

Under the requirement (1.4) any *countable* subset E of Ω must belong to the σ-field $\mathcal{F}$ as a consequence of the decomposition $E = \cup_{\omega \in E}\{\omega\}$. By σ-additivity we then conclude that

$$\mathbb{P}\left[E\right] = \sum_{\omega \in E} \mathbb{P}\left[\{\omega\}\right], \quad \begin{array}{c} E \subseteq \Omega \\ \text{Countable.} \end{array} \tag{1.6}$$

Several cases arise:

Finite case ($|\Omega| < \infty$) The set Ω contains a *finite* number of elements, labelled $\omega_1, \ldots, \omega_N$ for some finite N, so $\Omega = \{\omega_1, \ldots, \omega_N\}$. As pointed out in Section 1.5 the requirement (1.4) leads to $\mathcal{F} = \mathcal{P}(\Omega)$. Using (1.5) into (1.6) we get

$$\mathbb{P}[E] = \sum_{\omega \in E} \mathbb{P}[\{\omega\}] = |E|p, \quad E \in \mathcal{P}(\Omega)$$

whence $p = |\Omega|^{-1}$ upon taking $E = \Omega$ in this last relation. It immediately follows that

$$\mathbb{P}[E] = \frac{|E|}{|\Omega|}, \quad E \in \mathcal{P}(\Omega). \tag{1.7}$$

■

Much of elementary Probability Theory is concerned with computing such probabilities through combinatorial arguments that help evaluate the size of various subsets (e.g., E) of a discrete set (e.g., Ω).

Countably infinite case ($|\Omega| = \infty$) The set Ω contains *countably infinite* many elements, say $\Omega = \{\omega_n,\ n = 1, 2, \ldots\}$ for some labeling $\mathbb{N}_0 \to \Omega : n \to \omega_n$. Again, as pointed out in Section 1.5 the requirement (1.4) leads to $\mathcal{F} = \mathcal{P}(\Omega)$. The same argument as above, based on (1.5) and (1.6), shows that

$$\mathbb{P}[E] = |E|p \leq 1, \quad \begin{array}{c} E \in \mathcal{P}(\Omega) \\ |E| < \infty. \end{array} \tag{1.8}$$

Now it is always possible to select a sequence $\{E_n,\ n = 1, 2, \ldots\}$ of subsets of Ω such that $|E_n| = n$ for all $n = 1, 2, \ldots$ – Indeed, with the labeling introduced earlier, just take $E_n = \{\omega_1, \ldots, \omega_n\}$ in which case $|E_n| = n$. Applying (1.8) with $E = E_n$ we conclude that $p \leq n^{-1}$ for *all* $n = 1, 2, \ldots$, whence $p = 0$. A contradiction immediately arises: Indeed, by virtue of (1.6) (with $E = \Omega$), we get $\mathbb{P}[\Omega] = \sum_{\omega \in \Omega} p = 0$, and yet we must have $\mathbb{P}[\Omega] = 1$ because $\mathbb{P}$ is a probability measure. In other words, on a discrete sample set Ω with $|\Omega| = \infty$ it is *not* possible to construct a probability measure that satisfies the uniformity constraint (1.5). ■

Uncountably infinite case When Ω is uncountable, the same arguments as above will still show that $p = 0$, and the conclusion

$$\mathbb{P}[E] = 0, \quad \begin{array}{c} E \subseteq \Omega \\ \text{Countable} \end{array}$$

again follows from (1.6) by σ-additivity.

What can we say concerning $\mathbb{P}[E]$ if the subset E is *not* countable? While the decomposition $E = \cup_{\omega \in E}\{\omega\}$ always holds for any subset E of Ω, there is no guarantee that

$$\mathbb{P}[E] = \sum_{\omega \in E} \mathbb{P}[\{\omega\}], \qquad \begin{array}{c} E \subseteq \Omega \\ \text{Uncountable} \end{array}$$

since $\mathbb{P}$ is only required to be σ-additive. In fact, were this last relationship indeed hold for *every* subset of Ω (including Ω) we would have to conclude that $\mathbb{P}[E] = 0$ for *every* subset of Ω (including Ω, hence already a contradiction) This would certainly define a measure on $\mathcal{P}(\Omega)$, namely the zero measure, definitely not a probability measure on $\mathcal{P}(\Omega)$.

This discussion strongly suggests that when defining probability measures on non-countable sample spaces Ω, under the uniformity constraint (1.4) it may not be feasible to take $\mathcal{F} = \mathcal{P}(\Omega)$ – This will be further discussed in Chapter 5. This can be traced back to the fact that the σ-additivity of $\mathbb{P}$ on $\mathcal{P}(\Omega)$ imposes too many constraints – They cannot all be simultaneously satisfied if $\mathbb{P}$ is to be defined on $\mathcal{P}(\Omega)$, thereby forcing a reduction of $\mathcal{P}(\Omega)$ to a strictly smaller σ-field!

However, the analysis so far does not preclude the possibility of constructing a probability measure $\mathbb{P}$ which reflects constraints naturally associated with uniform selection other than (1.4) and (1.5) – This is illustrated on two examples, namely infinitely many coin tosses of a fair coin in Section **??** and selecting a point at random in the interval $[0, 1]$ in Section **??**. However, such constructions will have to be done on a σ-field strictly smaller than $\mathcal{P}(\Omega)$. ■

The reader may wonder as to why the conclusions reached in the countably infinite and uncountable cases differ: In the countable case the equivalence embedded in Fact 1.5.1 *forces* the construction of the desired uniform probability measure to pass through the constraints (1.4)-(1.5). While this leads to a complete characterization, namely (1.7), when Ω contains a finite number of samples, the situation is quite different in the countably infinite case: The constraints (1.4)-(1.5) are now incompatible with Ω being countably infinite. In the non-countable case, the construction of the desired uniform probability measure cannot pass through the constraints (1.4)-(1.5) – They are too weak as will be illustrated in Section **??**, leaving open the possibility that additional constraints reflecting uniformity could possibly be added to characterize the desired probability measure.

1.7 Generating σ-fields

On a number of occasions it will be helpful to consider the smallest σ-field containing a given collection of subsets of a non-empty set S. The following elementary fact provides the basis for this notion; its proof is left as an exercise [Exercise 1.11].

Fact 1.7.1 *If $\{\mathcal{F}_i,\ i \in I\}$ is a non-empty family of σ-fields on S (with I arbitrary, countable or not), then the intersection $\cap_{i \in I} \mathcal{S}_i$ is a σ-field on the set S.*

Using Fact 1.7.1 it is a simple matter to define the desired concept by leveraging the following easy result [Exercise 1.12].

Lemma 1.7.1 *Let $\mathcal{G}$ denote a non-empty collection of subsets of S. There exists a unique σ-field on S, denoted $\sigma\left(\mathcal{G}\right)$, with the following properties:*
(i) The σ-field $\sigma\left(\mathcal{G}\right)$ contains $\mathcal{G}$;
(ii) The σ-field $\sigma\left(\mathcal{G}\right)$ is minimal in the sense that any other σ-field $\mathcal{S}$ on S containing $\mathcal{G}$ must necessarily contain $\sigma\left(\mathcal{G}\right)$, i.e., $\sigma\left(\mathcal{G}\right) \subseteq \mathcal{S}$.

We refer to the σ-field $\sigma\left(\mathcal{G}\right)$, whose existence is established in Lemma 1.7.1, as the σ-field on S *generated* by $\mathcal{G}$. In fact $\sigma\left(\mathcal{G}\right)$ coincides with the σ-field $\cap_{i \in I} \mathcal{G}_i$ where $\{\mathcal{G}_i,\ i \in I\}$ denotes the family of all the σ-fields on S containing the collection $\mathcal{G}$ – Note that $\{\mathcal{G}_i,\ i \in I\}$ is never empty as it contains the power set $\mathcal{P}(S)$.

Definition 1.7.1

Let $\mathcal{G}$ and $\mathcal{S}$ be two collections of subsets of S with $\mathcal{G} \subseteq \mathcal{S}$. If $\mathcal{S}$ is a σ-field on S with $\mathcal{S} = \sigma\left(\mathcal{G}\right)$, we say that $\mathcal{G}$ *generates* the σ-field $\mathcal{S}$, or equivalently, that $\mathcal{G}$ is a *generating family* (or a *generator*) for $\mathcal{S}$.

The following fact is elementary [Exercise 1.13].

Fact 1.7.2 *If $\mathcal{G}_1$ and $\mathcal{G}_2$ are two collections of subsets of S such that $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\sigma\left(\mathcal{G}_1\right) \subseteq \sigma\left(\mathcal{G}_2\right)$. Moreover, if $\mathcal{G}_2$ is already a σ-field, then $\sigma\left(\mathcal{G}_2\right) = \mathcal{G}_2$ and $\sigma\left(\mathcal{G}_1\right) \subseteq \mathcal{G}_2$.*

1.8 Exercises

Ex. 1.1 Let $\mathcal{S}$ be a σ-field on some non-empty set S with a finite number of elements, i.e., $|\mathcal{S}| < \infty$.

a. Let $\mathcal{S}^\star$ denote the collection of all non-empty elements of $\mathcal{S}$ which do not contain another non-empty element of $\mathcal{S}$. Explain how $\mathcal{S}$ can be generated from

$\mathcal{S}^\star$ – We can think of $\mathcal{S}^\star$ as the "atoms" of $\mathcal{S}$ [**HINT:** Remember that $\mathcal{S}$ is a σ-field on S].

b. Using Part **a** show that we necessarily have $|\mathcal{S}| = 2^m$ with $|\mathcal{S}^\star| = m$.

c. Claim: Any σ-field on a non-empty finite set S necessarily has 2^m subsets of S in it for some positive integer m.

Ex. 1.2 Let $\mathcal{H}$ be a field on some set S. For any additive set function $\mu : \mathcal{H} \rightarrow [0, \infty]$ show that $\mu[\emptyset] = 0$ as soon as there exists H in $\mathcal{H}$ such that $\mu[H] < \infty$.

Ex. 1.3 In Definition 1.2.2 show that it suffices to check that the simpler pairwise conditions

$$\mu\left[E \cup F\right] = \mu\left[E\right] + \mu\left[F\right], \quad \begin{array}{c} E, F \in \mathcal{S} \\ E \cap F = \emptyset \end{array}$$

hold.

Ex. 1.4 Let S denote a countable set. With $\mathcal{F} = \mathcal{P}(S)$, define the set function $\mu : \mathcal{F} \rightarrow \mathbb{R}_+$ by

$$\mu\left[E\right] = |E|, \quad E \in \mathcal{F}$$

where $|E|$ denotes the number of elements in E. Show that the set function $\mu : \mathcal{F} \rightarrow \mathbb{R}_+$ is a measure on $\mathcal{F}$ – It is known as the *counting measure*.

Ex. 1.5 Let S be a countably infinite set, say $S = \mathbb{N}$. Define the collection $\mathcal{F}$ of subsets of S to be $\mathcal{F} \equiv \{F \subseteq S : \text{ Either } |F| < \infty \text{ or } |F^c| < \infty \}$.

a. Show that $\mathcal{F}$ is an algebra on S. Is it a σ-algebra on S? Explain.

b. Define the mapping $\mu : \mathcal{F} \rightarrow \mathbb{R}_+$ by

$$\mu\left[E\right] \equiv \begin{cases} 0 & \text{if } |E| < \infty \\ \\ 1 & \text{if } |E^c| < \infty. \end{cases}$$

Show that μ is finitely additive. Is μ also σ-additive on $\mathcal{F}$? – Prove or give a counterexample!

Ex. 1.6 Let S be an uncountable set, say $S = \mathbb{R}$. Define the collection $\mathcal{F}$ of subsets of S to be $\mathcal{F} \equiv \{E \subseteq S : E \text{ is countable or } E^c \text{ is countable}\}$ where countable means here either finite or countably infinite.

a. Show that $\mathcal{F}$ is an algebra on S.

b. Is $\mathcal{F}$ a σ-algebra on S? Prove or give a counterexample!

c. Define the set function $\mu : \mathcal{F} \rightarrow \mathbb{R}_+$ by

$$\mu\left[E\right] \equiv \begin{cases} 0 & \text{if } E \text{ is countable} \\ \\ 1 & \text{if } E^c \text{ is countable.} \end{cases}$$

Is this set function $\mu : \mathcal{F} \to \mathbb{R}_+$ σ-additive on $\mathcal{F}$? Prove or give a counterexample!

Ex. 1.7 Let S be a countable set. With $c > 0$ show that there exists a unique measure $\mu_c : \mathcal{P}(S) \to [0, +\infty]$ such that $\mu[\{s\}] = c$ for all s in S. Give an expression for $\mu[E]$ for E in $\mathcal{P}(S)$.

Ex. 1.8 Give proofs to the elementary properties (i)-(v) of probability models given in Section 1.4.

Ex. 1.9 Additional elementary properties of a probability measure: Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, with events E, F and G, we have

$$\mathbb{P}[E \cap F] \geq \mathbb{P}[E] + \mathbb{P}[F] - 1,$$

and

$$\mathbb{P}[E \Delta F] = \mathbb{P}[E] + \mathbb{P}[F] - 2\mathbb{P}[E \cap F]$$

where $E\Delta F$ denotes the symmetric difference of E and F (defined as $E\Delta F = (E \cap F^c) \cup (E^c \cap F)$). Furthermore, the following “triangle inequality’

$$\mathbb{P}[E \Delta G] \leq \mathbb{P}[E \Delta F] + \mathbb{P}[F \Delta G]$$

holds.

Ex. 1.10 Let S denote a finite set, say $S = \{1, \ldots, n\}$ for some positive integer n. The random experiment $\mathcal{E}$ consists in selecting uniformly an ordered pair A and B of (possibly empty) subsets of S.

a. Construct an probability model for this experiment – Clearly specify the sample space Ω, the σ-field $\mathcal{F}$ of events and the probability assignment $\mathbb{P}$. Using the ideas developed in Section 1.6, compute the following probabilities:

b. For any subset B of S, compute the probability of the event $\mathcal{I}_B$ given by

$$\mathcal{I}_B \equiv \{(A, B) \in \mathcal{P}(S) \times \mathcal{P}(S) : \; A \subseteq B\}.$$

c. Compute the probability of the event $\mathcal{I}$ given by

$$\mathcal{I} \equiv \{(A, B) \in \mathcal{P}(S) \times \mathcal{P}(S) : \; A \subseteq B\}$$

[**HINT:** Note that $\mathcal{I} = \cup_{B \subseteq S} \mathcal{I}_B$]. See Exercise 2.28 for another take on this problem using conditional probabilities.

Ex. 1.11 Prove Fact 1.7.1.

Ex. 1.12 Prove Lemma 1.7.1.

Ex. 1.13 Prove Fact 1.7.2.

Ex. 1.14 Let S be an arbitrary non-empty set. Let $\mathcal{G}$ denote the collection of all its singletons, namely $\mathcal{G} = \{\{s\}, \ s \in S\}$.

a. If S is a finite set, what is $\sigma(\mathcal{G})$?

b. If S is a countably infinite set, what is $\sigma(\mathcal{G})$?

c. If S is an uncountably infinite set, what is $\sigma(\mathcal{G})$?

Ex. 1.15 With S an arbitrary non-empty set, let $\mathcal{G}$ be a collection of subsets of S. If E denotes a subset of S, define the *trace of* $\mathcal{G}$ *on* E as the collection $\mathcal{G}_E$ of subsets of E given by

$$\mathcal{G}_E \equiv \{G \cap E : \ G \in \mathcal{G}\} .$$

a. Show that $\mathcal{G}_E$ is a field (resp. a σ-field) on E whenever $\mathcal{G}$ is a field (resp. σ-field) on E *regardless* of whether E is an element of $\mathcal{G}$.

b. Show that generating the smallest σ-field and taking a trace are commutative operations, i.e., $\sigma\left(\mathcal{G}_E\right) = \left(\sigma\left(\mathcal{G}\right)\right)_E$.

Ex. 1.16 Let $\{E_i, \ i \in I\}$ denote a countable decomposition of a non-empty set S, i.e., the sets $\{E_i, \ i \in I\}$ are not empty with $\cup_{i \in I} E_i = S$ and $E_i \cap E_j = \emptyset$ for all distinct i and j in I. Discuss the cardinality of the σ-field $\sigma\left(E_i, \ i \in I\right)$.

Chapter 2

Elementary Probability Theory

In Chapter 1 we introduced the notion of a probability model for a random experiment in the sense of Kolmogorov as a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω lists all elementary outcomes of the experiment, and the collection $\mathcal{F}$ identifies the subsets of Ω (or events) whose likelihood can be evaluated by means of the probability measure $\mathbb{P}$ defined on $\mathcal{F}$.

In the present chapter we present some of the most basic concepts often found in textbooks covering elementary Probability Theory. Throughout we are given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ which is held fixed during the discussion.

2.1 Bounding probabilities

With I a *finite* index set, let $\{E_i,\ i \in I\}$ denote any collection of events in $\mathcal{F}$. At this point the reader may wonder how to evaluate the probability of the union $\cup_{i \in I} E_i$ when the events are *not* disjoint (since in that case it is unclear how to invoke σ-additivity).

A formula by Poincaré The next result is attributed to Poincaré and gives a formal answer to this question. It states that

$$\mathbb{P}\left[\cup_{i \in I} E_i\right] = \sum_{k=1}^{|I|} (-1)^{k-1} \left(\sum\nolimits_{j_1 < j_2 < \ldots < j_k} \mathbb{P}\left[\cap_{\ell=1}^{k} E_{j_\ell} \right] \right) \tag{2.1}$$

where for each $k = 1, \ldots, |I|$, the summation $\sum_{j_1 < j_2 < \ldots < j_k}$ is over all ordered k-uples drawn from I. This formula is an expression of the *Inclusion-Exclusion Principle*.

By complementarity this expression is often used in the form

$$\begin{aligned} \mathbb{P}\left[\cap_{i\in I} E_i^c\right] &= 1-\mathbb{P}\left[\cup_{i\in I} E_i\right] \\ &= \sum_{k=0}^{|I|}(-1)^k\left(\sum\nolimits_{j_1<j_2<\ldots<j_k}\mathbb{P}\left[\cap_{\ell=1}^{k} E_{j_\ell}\right]\right) \end{aligned} \tag{2.2}$$

with the understanding that the term corresponding to $k=0$ is set to 1. The expressions (2.1) and (2.2) are easily derived by induction on the size of the cardinality $|I|$ [Exercise 2.1].

While exact, the expressions (2.1) and (2.2) are often too unwieldy to be useful. Instead only the upper and lower bounds given next suffice in many cases; they are also established by induction on the size $|I|$ [Exercise 2.3].

Boole's inequality This bound, also known as the *union bound*, is commonly used in Information Theory and theoretical Computer Science, and states that

$$\mathbb{P}\left[\cup_{i\in I} E_i\right] \le \sum\nolimits_{i\in I}\mathbb{P}\left[E_i\right]. \tag{2.3}$$

The union bound (2.3) also holds when I is countably infinite [Exercise 3.2].

The Bonferroni Inequality This bound gives a *lower* bound on the probability $\mathbb{P}\left[\cup_{i\in I} E_i\right]$ in the form

$$\sum\nolimits_{i\in I}\mathbb{P}\left[E_i\right]-\sum\nolimits_{i,j\in I:\, i<j}\mathbb{P}\left[E_i\cap E_j\right]\le\mathbb{P}\left[\cup_{i\in I} E_i\right]. \tag{2.4}$$

Combining the inequalities (2.3) and (2.4) we get

$$\sum\nolimits_{i\in I}\mathbb{P}\left[E_i\right]-\sum\nolimits_{i,j\in I:\, i<j}\mathbb{P}\left[E_i\cap E_j\right]\le\mathbb{P}\left[\cup_{i\in I} E_i\right]\le\sum\nolimits_{i\in I}\mathbb{P}\left[E_i\right].$$

This opens the door to the possibility that $\mathbb{P}\left[\cup_{i\in I} E_i\right]$ might be well approximated by $\sum_{i\in I}\mathbb{P}\left[E_i\right]$ whenever the term $\sum_{i,j\in I:\, i<j}\mathbb{P}\left[E_i\cap E_j\right]$ is smaller than $\sum_{i\in I}\mathbb{P}\left[E_i\right]$ in a suitable sense. This idea is commonly used in many settings. Sometimes it is more convenient to express these two inequalities in the following equivalent form

$$0\le\sum\nolimits_{i\in I}\mathbb{P}\left[E_i\right]-\mathbb{P}\left[\cup_{i\in I} E_i\right]\le\sum\nolimits_{i,j\in I:\, i<j}\mathbb{P}\left[E_i\cap E_j\right].$$

2.2 Independence

The notion of independence introduced next is key to probabilistic modeling. It is perhaps what makes Probability Theory not just a special case but rather a very rich subarea of Measure Theory.

In this section we consider a collection $\{E_i,\ i \in I\}$ of events in $\mathcal{F}$ where I is an *arbitrary* index set, and present the several notions of independence commonly discussed.

Pairwise independence The events $\{E_i,\ i \in I\}$ are said to be *pairwise independent* if the conditions

$$\mathbb{P}\left[E_i \cap E_j\right] = \mathbb{P}\left[E_i\right] \mathbb{P}\left[E_j\right], \qquad \begin{array}{c} i \neq j \\ i, j \in I \end{array}$$

all hold. With I finite, this constitutes a set of $\frac{|I|(|I|-1)}{2}$ conditions. When considering only two events, i.e., $|I| = 2$, this set of conditions reduces to a single condition, in which case the qualifier "pairwise" is dropped and the two events are simply said to be independent.

The terminology may be misleading. Indeed, *if two events are independent, it does not necessarily mean that their outcomes are not influencing each other in any way*; see Exercise 2.9 for an illustration of this point.

Mutual independence (with I finite) The events $\{E_i,\ i \in I\}$ are said to be *mutually independent* if the conditions

$$\mathbb{P}\left[\cap_{j \in J} E_j\right] = \prod_{j \in J} \mathbb{P}\left[E_j\right], \qquad \begin{array}{c} J \subset I \\ |J| \geq 1 \end{array}$$

are *all* satisfied – This represents $2^{|I|} - (|I|+1)$ non-trivial conditions. In Exercises 2.7 and 2.8 situations are given (with $|I| = 3$ so that $2^{|I|} - (|I|+1) = 2^3 - 4 = 4$ conditions) where some of the inequalities are satisfied but others are not. In particular, Exercise 2.7 already shows that pairwise independence may not imply mutual independence.

Mutual independence (with I arbitrary, countable or uncountable) The events $\{E_i,\ i \in I\}$ are said to be *mutually independent* if for each *finite* subset $J \subseteq I$ with $0 < |J| < \infty$, the events $\{E_j,\ j \in J\}$ are mutually independent. It is easy to check that this definition is equivalent to the following requirement [Exercise 2.10].

Fact 2.2.1 *When I is an uncountable index set, the collection $\{E_i,\ i \in I\}$ of events in $\mathcal{F}$ are mutually independent. if and only if the conditions*

$$\mathbb{P}\left[\cap_{k \in K} E_k\right] = \prod_{k \in K} \mathbb{P}\left[E_k\right], \qquad \begin{array}{c} K \subseteq I \\ |K| < \infty \\ |K| \geq 1 \end{array}$$

are all simultaneously satisfied.

Set-theoretic operations preserve independence in the following sense.

Theorem 2.2.1 *Consider a collection $\{E_i,\ i \in I\}$ of events in $\mathcal{F}$ where I is an arbitrary index set. If the events $\{E_i,\ i \in I\}$ are mutually independent, then the following statements hold:*

(i) For every subset $J \subseteq I$, the events $\{E_j,\ j \in J\}$ are mutually independent.

(ii) Taking complements does not affect mutual independence: For any subset $C \subseteq I$ (possibly empty), the events $\{E_i,\ i \in C; E_j^c,\ j \in I - C\}$ are mutually independent.

(iii) Partitioning does not affect mutual independence: The events $\{G_k,\ k \in K\}$ are mutually independent where K is an index set, $\{I_k,\ k \in K\}$ is a partition of I and for each k in K, the event G_k is defined by set-theoretic operations exclusively on the events $\{E_i,\ i \in I_k\}$ – Here set-theoretic operations refer to taking the complement of a set, union and intersection.

Part (i) is trivial, and although Part (ii) is subsumed by Part (iii), we invite the reader to provide a direct proof in Exercise 2.12. The proof of (iii) is more delicate and will not be given here.

The next two sections provide examples where the notion of independence plays a major role.

2.3 Modeling repeated coin tosses

In this section we discuss a class of random experiments associated with a well-known game of chance, namely the repeated throw of a coin. Historically this situation has provided much impetus for the early development of Probability Theory. It illustrates a probabilistic paradigm that recurs in many applications where the random experiment of interest consists of *repeated* random trials (or sub-experiments), each with exactly two possible outcomes, carried out under *identical* and *independent* conditions.

To set the stage, we first describe the random experiment of interest in some more detail: A two-sided coin is tossed repeatedly n times (with n a positive integer). Each toss results in one of two outcomes, say H = “Head” and T = “Tail” – It is often convenient to label the outcomes as $H = 1$ and $T = 0$ or even as $H = 1$ and $T = -1$. Furthermore we assume that the n successive tosses do form *independent* trials, each carried out under *identical* conditions. This implies that the likelihood of occurrence in each trial remains the same throughout the n trials, say p (resp. $1-p$) for any coin toss resulting in H (resp. T) with p a scalar in $[0,1]$.

From now on we use the labeling convention $H=1$ and $T=0$ so that the outcome of the random experiment can be represented by a binary sequence, i.e., a sequence of 0's and 1's, of length n. Put differently, each outcome is a word of length n with entries drawn from $\{0,1\}$. This leads to taking $\Omega=\{0,1\}^n$ for the sample space with generic element ω given by $\omega=(\omega_1,\ldots,\omega_n)$ where ω_i is an element of $\{0,1\}$ for each $i=1,\ldots,n$. Following the approach in Section 1.5, with $\mathcal{F}=\mathcal{P}(\Omega)$ as usual, we will construct the appropriate probability measure $\mathbb{P}$ on $\mathcal{P}(\Omega)$ by identifying a pmf $(p(\omega),\ \omega\in\{0,1\}^n)$ which reflects the probabilistic properties described earlier.

To do so, for each $k=1,\ldots,n$ we define the events

$$H_k \equiv \{\omega' \in \Omega:\ \omega'_k = 1\} \tag{2.5}$$

and

$$T_k \equiv \{\omega' \in \Omega:\ \omega'_k = 0\}. \tag{2.6}$$

The event H_k (resp. T_k) contains the outcomes of the n tosses for which the k^{th} toss results in H (resp. T). The disjoint sets H_k and T_k are complements of each other in Ω since $H_k\cup T_k=\{0,1,\}^n$, hence $T_k=H_k^c$. That the n successive tosses form *independent* trials, each carried out under *identical* conditions, *naturally* translates into the events $\{H_k,\ k=1,\ldots,n\}$ being *mutually independent* with

$$\mathbb{P}\left[H_k\right] = p = 1-\mathbb{P}\left[T_k\right], \quad k=1,\ldots,n. \tag{2.7}$$

Now pick ω in Ω, and introduce the index sets $H(\omega)\equiv\{k=1,\ldots,n:\ \omega_k=1\}$ and $T(\omega)\equiv\{k=1,\ldots,n:\ \omega_k=0\}$. The disjoint index sets $H(\omega)$ and $T(\omega)$ are obviously complement of each other in $\{1,\ldots,n\}$ since $H(\omega)\cup T(\omega)=\{1,\ldots,n\}$. Noting the representation

$$\{\omega\} = \left(\cap_{k\in H(\omega)} H_k\right) \cap \left(\cap_{\ell\in T(\omega)} T_\ell\right),$$

we get

$$\begin{aligned}
\mathbb{P}\left[\{\omega\}\right] &= \mathbb{P}\left[\left(\cap_{k\in H(\omega)} H_k\right) \cap \left(\cap_{\ell\in T(\omega)} T_\ell\right)\right] \\
&= \prod_{k\in H(\omega)} \mathbb{P}\left[H_k\right] \cdot \prod_{\ell\in T(\omega)} \mathbb{P}\left[T_\ell\right] \\
&= p^{|H(\omega)|}\cdot(1-p)^{|T(\omega)|} \qquad (2.8)
\end{aligned}$$

upon invoking Part (ii) of Theorem 2.2.1, namely that taking complements does not change mutual independence. In the last expression, $|H(\omega)|$ (resp. $|T(\omega)|$) denotes the number of trials (or equivalently, coin tosses) in the sample ω that result in H (resp. T).

In conclusion the assumptions that the n successive tosses form independent trials, each carried out under identical conditions, lead to the pmf $(p(\omega),\ \omega \in \{0,1\}^n)$ being given by

$$\begin{aligned} p(\omega) &= p^{|H(\omega)|} \cdot (1-p)^{|T(\omega)|} \\ &= p^{|H(\omega)|} \cdot (1-p)^{n-|H(\omega)|}, \quad \omega \in \{0,1\}^n \end{aligned} \tag{2.9}$$

since $|H(\omega)| + |T(\omega)| = n$. The case $p = \frac{1}{2}$ is often referred to as the fair case and is explored in Exercise 2.18.

2.4 A probabilistic proof of a formula by Euler

The Riemann function $\zeta : (0,\infty) \to [0,+\infty]$ is defined by

$$\zeta(s) \equiv \sum_{m=1}^{\infty} \frac{1}{m^s}, \quad s > 0. \tag{2.10}$$

It is easy to show that $\zeta(s) = \infty$ on the range $0 < s \leq 1$ and that $\zeta(s) < \infty$ for $s > 1$. The following identity was established by Euler.

Theorem 2.4.1 *It holds that*

$$\zeta(s) = \prod_{p \in \mathcal{P}} \frac{1}{1-p^{-s}}, \quad s > 1 \tag{2.11}$$

where $\mathcal{P}$ denotes the set of prime numbers.

We will now provide a probabilistic proof of this remarkable identity, thereby illustrating the power of probabilistic thinking!

Proof. Fix $s > 1$. The basic idea of the proof is to construct a discrete probability model $(\Omega, \mathbb{F}, \mathbb{P}_s)$ tailored to the value $\zeta(s)$ of the Riemann function: Take $\Omega = \mathbb{N}_0$ and $\mathcal{F} = \mathcal{P}(\mathbb{N}_0)$. Following the approach developed in Section 1.5 we define $\mathbb{P}_s$ on $\mathcal{P}(\mathbb{N}_0)$ through the pmf $(p_s(n),\ n = 1,2,\ldots)$ given by

$$p_s(n) = \frac{n^{-s}}{\zeta(s)}, \quad n = 1,2,\ldots$$

From the definition (2.10) of $\zeta(s)$ this definition is well posed and the collection $(p_s(n),\ n = 1,2,\ldots)$ is indeed a pmf on $\mathbb{N}_0$.

Next, for each $k = 1, 2, \ldots$, we introduce the subset M_k of $\mathbb{N}_0$ given by $M_k \equiv \{kn,\ n = 1, 2, \ldots\}$. An elementary calculation then shows that

$$\mathbb{P}_s\left[M_k\right] = \sum_{n=1}^{\infty} p_s(kn) = \sum_{n=1}^{\infty} \frac{(kn)^{-s}}{\zeta(s)} = \frac{1}{k^s}.$$

For each $\ell = 2, 3, \ldots$, if $p_1, \ldots, p_\ell$ are positive integers with *no common* divisors, then it is the case [Part (i) of Exercise 2.21] that

$$\cap_{r=1}^{\ell} M_{p_r} = M_{p_1 p_2 \ldots p_\ell}. \tag{2.12}$$

It then follows that

$$\begin{aligned} \mathbb{P}_s\left[\cap_{r=1}^{\ell} M_{p_r}\right] &= \mathbb{P}_s\left[M_{p_1 p_2 \ldots p_\ell}\right] \\ &= \frac{1}{(p_1 p_2 \ldots p_\ell)^s} \\ &= \prod_{r=1}^{\ell} \mathbb{P}_s\left[M_{p_r}\right]. \qquad (2.13) \end{aligned}$$

As this conclusion obviously holds for *any* collection of prime numbers $p_1, \ldots, p_\ell$, the events $\{M_p,\ p \in \mathcal{P}\}$ are mutually independent; see also Fact 2.2.1.

Now label the prime numbers by increasing size, say $\mathcal{P} = \{p_1, p_2, \ldots\}$ with $p_1 < p_2 < \ldots$. By Part (ii) of Theorem 2.2.1 the mutual independence of the events $\{M_p,\ p \in \mathcal{P}\}$ implies the mutual independence of the complementary events $\{M_p^c,\ p \in \mathcal{P}\}$. Thus, for any $\ell = 1, 2, \ldots$, it holds that

$$\mathbb{P}_s\left[\cap_{r=1}^{\ell} M_{p_r}^c\right] = \prod_{r=1}^{\ell} \mathbb{P}_s\left[M_{p_r}^c\right] = \prod_{r=1}^{\ell} \left(1 - p_r^{-s}\right). \tag{2.14}$$

Let ℓ go to infinity in (2.14): On one hand we get

$$\lim_{\ell\to\infty} \mathbb{P}_s\left[\cap_{r=1}^{\ell} M_{p_r}^c\right] = \lim_{\ell\to\infty} \prod_{r=1}^{\ell} \left(1 - p_r^{-s}\right) = \prod_{r=1}^{\infty} \left(1 - p_r^{-s}\right). \tag{2.15}$$

On the other hand we have $\cap_{p\in\mathcal{P}} M_p^c = \{1\}$ [Part (ii) of Exercise 2.21] and

$$\lim_{\ell\to\infty} \mathbb{P}_s\left[\cap_{r=1}^{\ell} M_{p_r}^c\right] = \mathbb{P}_s\left[\cap_{r=1}^{\infty} M_{p_r}^c\right] = \mathbb{P}_s\left[\{1\}\right] = \zeta(s)^{-1}. \tag{2.16}$$

The first equality expresses the continuity from above of the probability measure $\mathbb{P}_s$ discussed in Lemma 3.1.2 (applied with $E_\ell = \cap_{r=1}^{\ell} M_{p_r}^c$ for all $\ell = 1, 2, \ldots$). The proof of Theorem 2.4.1 is now complete upon combining (2.15) and (2.16). ■

2.5 Conditional probabilities

Conditional probabilities naturally arise when independence does not hold. We begin with a classical definition.

Definition 2.5.1

Consider an event B in $\mathcal{F}$ such that $\mathbb{P}\left[B\right] > 0$. The conditional probability of the event A in $\mathcal{F}$ given B is defined as the ratio

$$\mathbb{P}\left[A|B\right] \equiv \frac{\mathbb{P}\left[A \cap B\right]}{\mathbb{P}\left[B\right]}. \tag{2.17}$$

When $\mathbb{P}\left[B\right] = 0$ it is customary to take $\mathbb{P}\left[A|B\right]$ to be arbitrary in $[0,1]$. However, the relation

$$\mathbb{P}\left[A|B\right]\mathbb{P}\left[B\right] = \mathbb{P}\left[A \cap B\right], \quad A \in \mathcal{F} \tag{2.18}$$

is always true regardless of whether $\mathbb{P}\left[B\right] > 0$ or not: When $\mathbb{P}\left[B\right] > 0$ this is clear from (2.17), while if $\mathbb{P}\left[B\right] = 0$, then $\mathbb{P}\left[A \cap B\right] = 0$ and $\mathbb{P}\left[A|B\right]\mathbb{P}\left[B\right] = 0$, irrespective of the arbitrary value selected for $\mathbb{P}\left[A|B\right]$. The following fact is an easy consequence of the definitions.

Fact 2.5.1 *With $\mathbb{P}\left[B\right] > 0$, the mapping $\mathbb{Q}_B : \mathcal{F} \to [0,1]$ defined by*

$$\mathbb{Q}_B(A) \equiv \frac{\mathbb{P}\left[A \cap B\right]}{\mathbb{P}\left[B\right]}, \quad A \in \mathcal{F}$$

is a probability measure on $\mathcal{F}$.

Pairwise independence can be easily characterized in terms of conditional probabilities.

Fact 2.5.2 *Let A and B be two events in $\mathcal{F}$. Under the condition $\mathbb{P}\left[B\right] > 0$, the events A and B are independent if and only if $\mathbb{P}\left[A|B\right] = \mathbb{P}\left[A\right]$.*

In other words, the events A and B are independent if the *conditional* probability of A given B coincides with its *unconditional* probability $\mathbb{P}\left[A\right]$. This is a simple consequence of (2.18) and of the definition of pairwise independence; its proof is left as an exercise [Exercise 2.30].

We close this chapter with two easy, but important, consequences associated with the notion of conditional probability.

Definition 2.5.2

With I a countable index set, the events $\{B_i,\ i \in I\}$ in $\mathcal{F}$ form an $\mathcal{F}$-measurable *partition* of Ω whenever

$$B_i \cap B_j = \emptyset, \qquad \begin{matrix} i,j \in I \\ i \neq j \end{matrix} \qquad \text{and} \qquad \cup_{i \in I} B_i = \Omega.$$

This definition does not preclude that $\mathbb{P}[B_i] = 0$ for some i in I. However, the second condition yields

$$\sum_{i \in I} \mathbb{P}[B_i] = 1,$$

a fact which implies $\mathbb{P}[B_i] > 0$ for at least one index i in I.

The Law of Total Probability For each A in $\mathcal{F}$, the obvious decomposition $A = \cup_{i \in I}(A \cap B_i)$ yields

$$\begin{aligned} \mathbb{P}[A] &= \sum_{i \in I} \mathbb{P}[A \cap B_i] \\ &= \sum_{i \in I} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i], \quad A \in \mathcal{F}. \end{aligned} \tag{2.19}$$

Put differently,

$$\mathbb{P}[A] = \sum_{i \in I} \mathbb{Q}_{B_i}(A)\mathbb{P}[B_i], \quad A \in \mathcal{F}$$

in the notation used in Fact 2.5.1.

Bayes' rule – From prior probabilities to posterior probabilities Consider any event A in $\mathcal{F}$ such that $\mathbb{P}[A] > 0$. For each k in I, we have

$$\begin{aligned} \mathbb{P}[B_k|A] &= \frac{\mathbb{P}[B_k \cap A]}{\mathbb{P}[A]} \\ &= \frac{\mathbb{P}[A|B_k]\,\mathbb{P}[B_k]}{\mathbb{P}[A]} \\ &= \frac{\mathbb{P}[A|B_k]\,\mathbb{P}[B_k]}{\sum_{i \in I} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i]} \end{aligned} \tag{2.20}$$

upon using the Law of Total Probability to evaluate the denominator. This last relation, which gives the *posterior* probability $\mathbb{P}[B_k|A]$ in terms of the *likelihoods* $\{\mathbb{P}[A|B_i],\ i \in I\}$ and of the *prior probabilities* $\{\mathbb{P}[B_i],\ i \in I\}$, is a celebrated relation known as Bayes' rule or Bayes' law. It plays a central role in many branches of Statistics and Data Science.

2.6 Exercises

Unless otherwise specified a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ is assumed given. Exercises 2.1–2.6 assume the setting of Section 2.1.

Ex. 2.1 Prove Poincaré's formulae (2.1) and (2.2) [**HINT:** The result is true when $|I| = 2$, and then use an induction argument on the size of I].

Ex. 2.2 What happens to Poincaré's formula (2.1) when the events $\{E_i,\ i \in I\}$ are disjoint?

Ex. 2.3 Prove the bounds (2.3) and (2.4) [**HINT:** The results are true when $|I| = 2$, and then use an induction argument on the size of I].

Ex. 2.4 Prove the bound

$$\mathbb{P}\left[\cup_{i\in I} E_i\right] \leq \min_{k\in I}\left(\sum_{i\in I}\mathbb{P}\left[E_i\right] - \sum_{\ell\in I, \ell\neq k}\mathbb{P}\left[E_k \cap E_\ell\right]\right)$$

due to Kounias.

Ex. 2.5 The bounds (2.3) and (2.4) take a particularly simple form when for each $k = 1, 2$, the individual probabilities $\mathbb{P}\left[\cap_{\ell=1}^{k} E_{j_\ell}\right]$ do not depend on the index set $i_1 < \ldots < i_k$. In such situations, show that

$$|I|\mathbb{P}\left[E_1\right] - \frac{|I|(|I|-1)}{2}\mathbb{P}\left[E_1 \cap E_2\right] \leq \mathbb{P}\left[\cup_{i\in I} E_i\right] \leq |I|\mathbb{P}\left[E_1\right].$$

Ex. 2.6 Show that

$$\mathbb{P}\left[\cap_{i\in I} E_i\right] \geq \sum_{i\in I}\mathbb{P}\left[E_i\right] - (|I| - 1).$$

a. First proof: The result is true when $|I| = 2$ by virtue of Exercise 1.9. Then proceed with an induction argument on the size of I.

b. Second proof: Apply the union bound to the collection $\{E_i^c,\ i \in I\}$.

Ex. 2.7 An item is selected uniformly from a set comprising four distinct objects labelled $1, 2, 3, 4$. To model this experiment we take $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ given by the uniform pmf $p(1) = \ldots = p(4) = \frac{1}{4}$. On this probability space define three events, say A, B and C, such that the events A, B and C are pairwise independent but not mutually independent, i.e., $\mathbb{P}\left[A \cap B\right] = \mathbb{P}\left[A\right]\mathbb{P}\left[B\right]$, $\mathbb{P}\left[B \cap C\right] = \mathbb{P}\left[B\right]\mathbb{P}\left[C\right]$ and $\mathbb{P}\left[A \cap C\right] = \mathbb{P}\left[A\right]\mathbb{P}\left[C\right]$ and yet $\mathbb{P}\left[A \cap B \cap C\right] \neq \mathbb{P}\left[A\right]\mathbb{P}\left[B\right]\mathbb{P}\left[C\right]$.

Ex. 2.8 An item is selected uniformly from a set comprising eight distinct objects labelled $1, \ldots, 8$. We model this experiment by taking $\Omega = \{1, \ldots, 8\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ given by the uniform pmf $p(1) = \ldots = p(8) = \frac{1}{8}$. On this probability space define three events, say A, B and C, such that $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$, $\mathbb{P}[A \cap C] = \mathbb{P}[A]\,\mathbb{P}[C]$ and $\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A]\,\mathbb{P}[B]\,\mathbb{P}[C]$, yet $\mathbb{P}[B \cap C] \neq \mathbb{P}[B]\,\mathbb{P}[C]$. In other words, the pairs of events A and B, and A and C are each pairwise independent and the condition $\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A]\,\mathbb{P}[B]\,\mathbb{P}[C]$ holds. However, the events B and C are not pairwise independent – This illustrates that the three events A, B and C are not mutually independent.

Ex. 2.9 Two identical six-facetted dice are cast one after the other under identical and independent conditions, and the outcomes recorded. We model this experiment by taking $\Omega = \{(k,\ell),\ k,\ell = 1,\ldots\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ given by the uniform pmf $p(k,\ell) = \frac{1}{36}$ $(k,\ell = 1,\ldots,6)$. Consider the events A and B given by

$$A \equiv [\text{ The sum of the two outcomes is 7 }]$$

and

$$B \equiv [\text{ The first outcome is 3 }]\,.$$

Show that the events A and B are independent. Although there is intuitively a form of "influence" between the events A and B – After all getting a "3" in the first outcome obviously affects the second outcome that could produce a sum "7", these events are independent according to definition given in Section 2.2 because the realization of one event does not affect the probability of the other.

Ex. 2.10 Prove the equivalence stated in Fact 2.2.1.

Ex. 2.11 Let $\{E_i,\ i \in I\}$ denote a collection of events in $\mathcal{F}$.

a. With I countably infinite, explain why the definition of independence requiring

$$\mathbb{P}[\cap_{j\in J} E_j] = \prod_{j\in J} \mathbb{P}[E_j]\,, \qquad \begin{matrix} J \subseteq I \\ 1 \leq |J| \end{matrix} \tag{2.21}$$

is equivalent to the one given in Section 2.2 (and equivalent to the one given in Fact 2.2.1) – However, note that when J is countably infinite the conditions (2.21) are usually not informative.

b. With I *uncountable*, explain why a definition of independence requiring

$$\mathbb{P}[\cap_{j\in J} E_j] = \prod_{j\in J} \mathbb{P}[E_j]\,, \qquad \begin{matrix} J \subseteq I \\ 1 \leq |J| \end{matrix}$$

makes no mathematical sense.

Ex. 2.12 Prove Part (ii) of Theorem 2.2.1 [**HINT 1:** It suffices to consider the case when I is finite] [**HINT 2:** Proceed by induction on the size of the index set C when evaluating the probabilities

$$\mathbb{P}\left[\left(\cap_{j\in C}E_j^c\right)\cap\left(\cap_{i\in I-C}E_i\right)\right]$$

where $C \subseteq I$. Start with the case $|C| = 1$].

Ex. 2.13 If the events $E_1, \ldots, E_n$ in $\mathcal{F}$ are mutually independent events such that $\mathbb{P}\left[\cup_{i=1}^n E_i\right] = 1$, show that $\mathbb{P}\left[E_k\right] = 1$ for some index $k = 1, \ldots, n$. Is the index k unique?

Ex. 2.14 Let E, F and G denote three events in $\mathcal{F}$ which are **mutually independent**, and assume that $0 < \mathbb{P}\left[E\right], \mathbb{P}\left[F\right] < 1$. Under what conditions are the events $E \cap G$ and $F \cap G$ independent?

Ex. 2.15 Let A, E_1 and E_2 denote three events in $\mathcal{F}$. Assuming that for each $k = 1, 2$, the events A and E_k are independent, show that the events A and $E_1 \cap E_2$ are independent if and only if the events A and $E_1 \cup E_2$ are independent.

Ex. 2.16 Let A denote an event in $\mathcal{F}$ with $0 < \mathbb{P}\left[A\right] < 1$ (in order to avoid trivial situations of limited interest). Define the collection $\mathcal{F}_A$ of events in $\mathcal{F}$ by

$$\mathcal{F}_A = \left\{F \in \mathcal{F} : \ \mathbb{P}\left[F \cap A\right] = \mathbb{P}\left[F\right] \cdot \mathbb{P}\left[A\right]\right\}.$$

a. Show that both Ω and the empty set $\emptyset$ belong to $\mathcal{F}_A$.

b. Show that $\mathcal{F}_A$ is closed under complementarity, i.e., if F is an element of $\mathcal{F}_A$, then so is its complement F^c.

c. Is the family $\mathcal{F}_A$ a σ-field on Ω? Prove or give a counterexample!

Ex. 2.17 Consider the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ where (i) $\Omega = \{1, \ldots, p\}$ for some prime number p; (ii) $\mathcal{F}$ is the power set of Ω; and (iii) the probability assignment $\mathbb{P}$ is uniform in the sense that $\mathbb{P}\left[A\right] = \frac{|A|}{p}$ for every subset A of Ω. Consider now two independent events A and B, neither of which is empty. What can you say concerning these sets?

Ex. 2.18 Consider the situation discussed in Section 2.3. The pmf $(p(\omega),\ \omega \in \{0,1\}^n)$ described at (2.9) was obtained by translating the requirement that the n successive tosses form independent trials, each carried out under identical conditions. This was done by positing the mutual independence of the events $\{H_k,\ k = 1, 2, \ldots, n\}$ defined at (2.5) with probability assignment (2.7).

If the coin is fair, i.e., $p = \frac{1}{2}$, then the expression (2.9) reduces to

$$p(\omega) = 2^{-n}, \quad \omega \in \{0,1\}^n. \tag{2.22}$$

Since $\Omega = \{0,1\}^n$ for this model, hence $\Omega| = 2^n$), we conclude from the discussion in Section 1.6 that the pmf (2.22) corresponds to uniform selection.

Conversely, if we consider the uniform probability assignment (2.22), show that the events $\{H_k,\ k = 1,2,\ldots,n\}$ defined at (2.5) are necessarily mutually independent.

Ex. 2.19 The situation discussed in Exercise 2.18 can be further generalized: Consider n probability triples, each with a finite sample space, say $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1)$, $\ldots$, $(\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$. Consider the measurable space $(\Omega, \mathcal{F})$ where $\Omega \equiv \Omega_1 \times \ldots \times \Omega_n$ and $\mathcal{F} \equiv \mathcal{P}(\Omega)$.

a. Show that there exists a unique probability measure $\mathbb{P}$ on $\mathcal{P}(\Omega)$ such that

$$\mathbb{P}\left[E_1 \times \ldots \times E_n\right] = \prod_{k=1}^{n} \mathbb{P}_k\left[E_k\right], \qquad \begin{array}{l} E_k \in \mathcal{P}(\Omega_k) \\ k = 1,\ldots,n \end{array} \tag{2.23}$$

and give an expression for the probabilities

$$\mathbb{P}\left[E\right], \quad E \in \mathcal{P}(\Omega).$$

[**HINT:** What is the value of $\mathbb{P}\left[\{\omega\}\right]$ for each ω in Ω?].

b. Show that the unique probability measure $\mathbb{P}$ satisfying is uniform on $\mathcal{P}(\Omega)$ if and only if each probability measure $\mathbb{P}_k$ is uniform on $\mathcal{P}(\Omega_k)$, $k = 1,\ldots,n$.

Ex. 2.20 A fair coin is rolled n times under identical and independent conditions, as in Exercise 2.18. We adopt the model discussed in Section 2.3 (with $p = \frac{1}{2}$).

With distinct $i, j = 1,\ldots,n$, define the event E_{ij} as the event where the outcomes of the i^{th} and j^{th} tosses are identical (e.g., both are heads). Show that the $\frac{n(n-1)}{2}$ events $\{E_{ij},\ 1 \leq i < j \leq n\}$ are pairwise independent but not mutually independent!

Ex. 2.21 In the proof of Theorem 2.4.1:

(i) Prove the set equality (2.21) [**HINT:** Prove it first for $\ell = 2$, and use induction on ℓ to establish the general case].

(ii) Show that $\cap_{p \in \mathcal{P}} M_p^c = \{1\}$.

Ex. 2.22 Establish Fact 2.5.1

Ex. 2.23 In Definition 2.5.1 show through examples that both inequalities $\mathbb{P}\left[A\right] < \mathbb{P}\left[A|B\right]$ and $\mathbb{P}\left[A|B\right] < \mathbb{P}\left[A\right]$ are possible – Assume that $\mathbb{P}\left[B\right] > 0$.

Ex. 2.24 Given are three scalars α, β and γ in $(0,1)$. Construct a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ and a pair of events E and F in $\mathcal{F}$ such that $\mathbb{P}[F] = \beta$, $\mathbb{P}[E|F] = \alpha$ and $\mathbb{P}[E|F^c] = \gamma$

Ex. 2.25 Let E and F be events in $\mathcal{F}$ with $\mathbb{P}[E] > 0$ and $\mathbb{P}[F] > 0$. We say that event E is *positively correlated* with event F if $\mathbb{P}[E|F] \geq \mathbb{P}[E]$.

a. Show the equivalence of the following three statements (i)-(iii) below:

(i) Event E is positively correlated with event F

(ii) Event F is positively correlated with event E

(iii) Event E^c is positively correlated with event F^c

b. Construct a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ and three events E, F and G in $\mathcal{F}$ such that $\mathbb{P}[E|F] > \mathbb{P}[E]$ [Event E is (strictly) positively correlated with event F] $\mathbb{P}[F|G] > \mathbb{P}[F]$ [Event E is (strictly) positively correlated with event F] but $\mathbb{P}[E|G] > \mathbb{P}[E]$ [Event E is not positively correlated with event G] [**HINT:** Take $\Omega = \{1,2,3\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and the uniform probability assignment on $\mathcal{F}$].

Ex. 2.26 The decomposition formula for conditional probabilities: Given three events E, F and G in $\mathcal{F}$ such that $\mathbb{P}[F \cap G] > 0$ and $\mathbb{P}[F \cap G^c] > 0$, show that

$$\mathbb{P}[E|F] = \mathbb{P}[G|F]\,\mathbb{P}[E|F \cup G] + \mathbb{P}[G^c|F]\,\mathbb{P}[E|F \cup G^c].$$

Ex. 2.27 Consider events $E_1, \ldots, E_n$ in $\mathcal{F}$ that are disjoint with $\mathbb{P}[E_i] > 0$ for all $i = 1, \ldots, n$. For any event E in $\mathcal{F}$ show that the bounds

$$\min_{i=1,\ldots,n} \mathbb{P}[F|E_i] \leq \mathbb{P}[E| \cup_i^n E_i] \leq \max_{i=1,\ldots,n} \mathbb{P}[E|E_i]$$

hold [**HINT:** Let $\alpha_1, \ldots, \alpha_n$ be non-negative scalars such that $\alpha_1 + \ldots + \alpha_n = 1$. Then, for any $x_1, \ldots, x_n$ in $\mathbb{R}$ the inequalities

$$\min_{i=1,\ldots,n} x_i \leq \sum_{i=1}^n \alpha_i x_i \leq \max_{i=1,\ldots,n} x_i$$

are satisfied].

Ex. 2.28 We return to Exercise 1.10: Let S denote a finite set, say $S = \{1, \ldots, n\}$ for some positive integer n. The random experiment $\mathcal{E}$ consists in selecting uniformly an ordered pair A and B of (possibly empty) subsets of S. Using the model developed there, evaluate $\mathbb{P}[\mathcal{I}]$ by first evaluating the conditional probabilities

$$\mathbb{P}[\mathcal{I}|B], \quad B \subseteq S,$$

and then using the Law of Total Probability.

Ex. 2.29 The following fact is useful when modeling situations associated with sequential decision making: Given a probability model $(\Omega, \mathcal{F}, \mathbb{P})$, with events $A_1, \ldots, A_n$ in $\mathcal{F}$, show that

$$\mathbb{P}\left[A_1 \cap \ldots \cap A_n\right] = \mathbb{P}\left[A_1\right] \cdot \prod_{i=2}^{n} \mathbb{P}\left[A_i | A_1 \cap \ldots \cap A_{i-1}\right].$$

[**HINT:** The observation

$$\mathbb{P}\left[A_1 \cap \ldots \cap A_n\right] = \mathbb{P}\left[A_n | A_1 \cap \ldots \cap A_{n-1}\right] \cdot \mathbb{P}\left[A_1 \cap \ldots \cap A_{n-1}\right]$$

suggests a proof by induction on n].

Ex. 2.30 Establish Fact 2.5.2

Ex. 2.31 Your cupboard contains six cups and six saucers. There are two blue cups and two blue saucers, two red cups and two red saucers, two white cups and two white saucers. As you are setting the table for a dinner party, you randomly assign cups to saucers.

a. Construct a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ that would model this situation.

b. Compute the probability that each of the six cups is assigned to a saucer of the same color!

c. Compute the probability that none of the six cups is assigned to a saucer of the same color!

We close with several urn problems:

Ex. 2.32 An urn contains R red balls and B blue balls (with $R \geq 1$ and $B \geq 1$). A ball is drawn at random from the urn and then put aside. A second ball is then drawn again at random from the urn (which now contains one less ball).

a. Construct a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ that would model this situation.

b. Consider the event E and F informally defined as $E =$ [The first ball drawn is red] and $F =$ [The second ball drawn is red]. Are they independent?

Ex. 2.33 Consider two urns, say U_1 and U_2, each of which contains contains R red balls and B blue balls. A ball is drawn at random from urn U_1, and put in urn U_2. Urn U_2 is then well stirred and shaken, and a ball is drawn at random from urn U_2.

a. Describe a complete probability model $(\Omega, \mathcal{F}, \mathbb{P})$ to model this situation.

b. Compute the probability that the ball drawn from urn U_2 is red.

Ex. 2.34 There are three urns, say U_1, U_2 and U_3. Urn U_1 contains R_1 red balls and B_1 blue balls, urn U_2 contains R_2 red balls and B_2 blue balls, and urn U_3 contains R_3 red balls and B_3 blue balls. An urn is selected at random, and then a ball is selected at random from it.

a. Describe a complete probability model $(\Omega, \mathcal{F}, \mathbb{P})$ to model this situation.

b. Compute the probability that the selected ball came from urn U_1 if the ball selected is red.

Ex. 2.35 An urn contains B blue balls and R red balls. They are removed one by one at random and not replaced until the urn is empty.

a. Construct a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ that would model this situation.

b. Compute the probability that the first red ball drawn is the $(k+1)$-th ball drawn – What is the range of k?

c. Compute the probability that the last ball drawn is red.

Ex. 2.36 Consider N urns (with $N \geq 2$), say $U_1, \ldots, U_N$, each of which initially contains R red balls and B blue balls. Each of the urns has been well stirred and shaken! A ball is drawn at random from urn U_1, and put in urn U_2 which is then well stirred and shaken! Then, a ball is drawn at random from urn U_2 and put in urn U_3. The process is repeated until a ball is finally drawn at random from last urn U_N.

a. Describe a complete probability model $(\Omega, \mathcal{F}, \mathbb{P})$ to model this situation.

b. Compute the probability that the ball selected from urn U_1 is red.

Ex. 2.37 There are two urns, U_1 and U_2, each containing $N-1$ blue balls and one (1) red ball.

a. From each urn, n balls are randomly selected without replacement (with $1 \leq n \leq N$), and the selections at urns U_1 and U_2 are carried out independently from urn to urn:

Construct a probability model $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ that would model this situation. Compute the probability $P^\star(E)$ of selecting at least one red ball among the $2n$ balls that have been drawn.

b. Combine all $2N$ balls into a third urn U_3 and let $2n$ balls be randomly selected from urn U_3 without replacement:

Construct a probability model $(\Omega_\star, \mathcal{F}_\star, \mathbb{P}_\star)$ that would model this second situation. Compute the probability $P_\star(E)$ of selecting at least one red ball among the $2n$ balls drawn from urn U_3.

c. Compare $P^\star(E)$ and $P_\star(E)$. Is it surprising?

In Exercises 2.38–2.40 we consider an urn containing n balls, some of which are red with the remaining ones being blue. Although the exact composition of the urn is *not* known a priori, it is believed that the values $0, 1, \ldots, n$ for the number of red balls are *equally likely*.

Ex. 2.38 First draw a ball at random from the urn and record its color.

Construct a probability model $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ to compute the probability that k red balls ($k = 0, 1, \ldots, n$) were initially in the urn given that the ball drawn was red.

Ex. 2.39 First draw a ball at random from the urn, throw it away and then draw a second ball at random from the remaining $n - 1$ balls.

Construct a probability model $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ to compute the probability that k red balls ($k = 0, 1, \ldots, n$) were initially in the urn given that the two balls drawn were red and blue in that order? Is the probability model $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ appropriate to answer this question? If not, explain!

Ex. 2.40 First draw a ball at random from the urn, throw it away and then draw a second ball at random from the remaining $n - 1$ balls. This time, compute the probability that k red balls ($k = 0, 1, \ldots, n$) were initially in the urn given that the two balls drawn were red and blue, but the order is unknown.

Chapter 3

Limits of probabilities vs. probabilities of limiting events

In many applications we are confronted with the following situation: On some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, there is a need to examine the probabilistic behavior of a sequence of events $\{E_n,\ n = 1, 2, \ldots\}$ in $\mathcal{F}$ as n becomes large. Specifically, we are interested in understanding how the sequence of probabilities $\{\mathbb{P}\left[E_n\right],\ n = 1, 2, \ldots\}$ behaves – Does this sequence converge and if so, what is the limit?

A typical example arise when trying to assess the performance of statistical procedures. In such a situation, the event E_n describes an event where the procedure provides an acceptable level of performance when n samples are used. Special attention is then given to situations where either $\lim_{n\to\infty} \mathbb{P}\left[E_n\right] = 0$ or $\lim_{n\to\infty} \mathbb{P}\left[E_n\right] = 1$.

In the present chapter we start developing some understanding of these issues by presenting some basic facts concerning the limiting behavior of these probabilities. In particular, we present conditions under which the limit $\lim_{n\to\infty} \mathbb{P}\left[E_n\right]$, exists, and identify this limit.

3.1 Monotonicity and σ-additivity

Consider a sequence $\{E_n,\ n = 1, 2, \ldots\}$ of events in $\mathcal{F}$.

Lemma 3.1.1 *(Continuity from below) If the sequence is monotone increasing in the sense that $E_n \subseteq E_{n+1}$ for all $n = 1, 2, \ldots$, then $\lim_{n\to\infty} \mathbb{P}\left[E_n\right] = \mathbb{P}\left[\cup_{n=1}^{\infty} E_n\right]$.*

Proof. Note the relation

$$\cup_{n=1}^{\infty} E_n = \cup_{m=1}^{\infty} F_m$$

where

$$F_m \equiv E_m - E_{m-1}, \quad m = 1, 2, \ldots$$

(under the convention $E_0 = \emptyset$). The events $\{F_m,\ m = 1, 2, \ldots\}$ being pairwise disjoint, we get

$$\begin{aligned}
\mathbb{P}\left[\cup_{n=1}^{\infty} E_n\right] &= \mathbb{P}\left[\cup_{m=1}^{\infty} F_m\right] \\
&= \sum_{m=1}^{\infty} \mathbb{P}\left[F_m\right] \quad \text{[By the } \sigma\text{-additivity of } \mathbb{P}] \\
&= \sum_{m=1}^{\infty} \left(\mathbb{P}\left[E_m\right] - \mathbb{P}\left[E_{m-1}\right]\right) \\
&= \lim_{m\to\infty} \left(\sum_{k=1}^{m} \left(\mathbb{P}\left[E_k\right] - \mathbb{P}\left[E_{k-1}\right]\right) \right) \\
&= \lim_{m\to\infty} \left(\mathbb{P}\left[E_m\right] - \mathbb{P}\left[E_0\right]\right) = \lim_{m\to\infty} \mathbb{P}\left[E_m\right]. \qquad (3.1)
\end{aligned}$$

■

This result can be interpreted as a *continuity* result for $\mathbb{P}$ in the following sense: If we *define* $\lim_{n\to\infty} E_n \equiv \cup_{n=1}^{\infty} E_n$, then Lemma 3.1.1 states that $\lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n\to\infty} E_n]$. The analog of Lemma 3.1.1 for sequences of events which are monotone decreasing is given next.

Lemma 3.1.2 *(Continuity from above) If the sequence is monotone decreasing in the sense that $E_{n+1} \subseteq E_n$ for all $n = 1, 2, \ldots$, then $\lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}[\cap_{n=1}^{\infty} E_n]$.*

This result can also be recast as a continuity result for $\mathbb{P}$: This time, if we *define* $\lim_{n\to\infty} E_n \equiv \cap_{i=1}^{\infty} E_n$, then $\lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n\to\infty} E_n]$ by virtue of Lemma 3.1.2. The proof of this result is similar to the one given for Lemma 3.1.1. In fact, these two results are equivalent once we observe that a sequence $\{E_n,\ n = 1, 2, \ldots\}$ is monotone increasing (resp. decreasing) if and only if its complementary sequence $\{E_n^c,\ n = 1, 2, \ldots\}$ is monotone decreasing (resp. increasing).

We stress that in Lemma 3.1.1 and Lemma 3.1.2 the existence of the limit $\lim_{n\to\infty} \mathbb{P}[E_n]$ is trivially guaranteed by the monotonicity of the sequence $\{\mathbb{P}[E_n],\ n = 1, 2, \ldots\}$. The added information provided by these results is an identification of the limit as the *probability* of the well-defined events $\cup_{n=1}^{\infty} E_n$ and $\cap_{n=1}^{\infty} E_n$, respectively.

We close this section by pointing out an inherent equivalence between Lemma 3.1.1 and Lemma 3.1.2, on one the hand, and σ-additivity of the underlying probability measure on the other; see Proposition 5.1.1 for a formal statement of this equivalence in the context of arbitrary measures.

3.2 Limsup, liminf and limits – Continuity of probability measures

In analogy with the convergence of sequences on $\mathbb{R}$, these continuity results for monotone sequences of events can be extended to arbitrary sequences of events as follows: Let $\{E_n,\ n = 1, 2, \ldots\}$ be an arbitrary collection of events in $\mathcal{F}$. Define

$$\limsup_{n\to\infty} E_n \equiv \cap_{n=1}^{\infty} \left(\cup_{m=n}^{\infty} E_m \right) = \cap_{n=1}^{\infty} \overline{E}_n$$

with

$$\overline{E}_n = \cup_{m=n}^{\infty} E_m, \quad n = 1, 2, \ldots$$

Similarly,

$$\liminf_{n\to\infty} E_n \equiv \cup_{n=1}^{\infty} \left(\cap_{m=n}^{\infty} E_m \right) = \cup_{n=1}^{\infty} \underline{E}_n$$

with

$$\underline{E}_n = \cap_{m=n}^{\infty} E_m, \quad n = 1, 2, \ldots$$

The sets $\limsup_{n\to\infty} E_n$ and $\liminf_{n\to\infty} E_n$ are well defined and always exist; we refer to them as the *limit sup* and *limit inf*, respectively, of the collection $\{E_n,\ n = 1, 2, \ldots\}$ in analogy with similar notions for real-valued sequences.

We have the memnonic notation

$$\limsup_{n\to\infty} E_n = [\ E_n \text{ infinitely often (i.o.)}\ ]$$

and

$$\liminf_{n\to\infty} E_n = [\ \text{Eventually all } E_n\ ] .$$

Both $\limsup_{n\to\infty} E_n$ and $\liminf_{n\to\infty} E_n$ are events in $\mathcal{F}$, and the inclusion

$$\liminf_{n\to\infty} E_n \subseteq \limsup_{n\to\infty} E_n \tag{3.2}$$

holds [Exercise 3.1]

Definition 3.2.1

The collection $\{E_n,\ n = 1, 2, \ldots\}$ of events is said to *converge* if the condition

$$\limsup_{n\to\infty} E_n = \liminf_{n\to\infty} E_n \tag{3.3}$$

holds, in which case we define its limit, denoted $\lim_{n\to\infty} E_n$, to be the event at (3.3).

With this definition of set continuity, we have the following continuity property for probability measures.

Proposition 3.2.1 *If the collection $\{E_n,\ n = 1, 2, \ldots\}$ of events in $\mathcal{F}$ converges according to Definition 3.2.1, then we have the continuity result*

$$\lim_{n\to\infty} \mathbb{P}\left[E_n\right] = \mathbb{P}\left[\lim_{n\to\infty} E_n\right].$$

Proposition 3.2.1 requires *no* monotonicity assumption on the collection $\{E_n,\ n = 1, 2, \ldots\}$, only the convergence condition (3.3), in contrast with both Lemma 3.1.2 and Lemma 3.1.2 which are in fact special cases of this result [Exercise 3.3]. A different take to Proposition 3.2.1 can be found in Exercise 3.6.

Proof. For each $n = 1, 2, \ldots$ the monotone inclusions $\underline{E}_n \subseteq \underline{E}_{n+1}$ and $\overline{E}_{n+1} \subseteq \overline{E}_n$ flow from the definitions. The continuity of $\mathbb{P}$ on monotone sequences implies $\lim_{n\to\infty} \mathbb{P}\left[\underline{E}_n\right] = \mathbb{P}\left[\liminf_{n\to\infty} E_n\right]$ by Lemma 3.1.1 and $\lim_{n\to\infty} \mathbb{P}\left[\overline{E}_n\right] = \mathbb{P}\left[\limsup_{n\to\infty} E_n\right]$ by Lemma 3.1.2. Under the assumed convergence condition (3.3) satisfied by the collection of events $\{E_n,\ n = 1, 2, \ldots\}$, the equality

$$\lim_{n\to\infty} \mathbb{P}\left[\underline{E}_n\right] = \lim_{n\to\infty} \mathbb{P}\left[\overline{E}_n\right] \tag{3.4}$$

then follows. On the other hand, we have

$$\mathbb{P}\left[\underline{E}_n\right] \le \mathbb{P}\left[E_n\right] \le \mathbb{P}\left[\overline{E}_n\right], \quad n = 1, 2, \ldots$$

by virtue of the obvious inclusions $\underline{E}_n \subseteq E_n \subseteq \overline{E}_n$. Let n go to infinity in this last chain of inequalities: A standard sandwich argument yields the desired result as we make use of (3.4). ■

3.3 Borel-Cantelli Lemmas

The Borel-Cantelli Lemmas given next constitute an example of a *zero-one law*; they play a central role is the derivation of various convergence results. Recall that if $\{E_n,\ n = 1, 2, \ldots\}$ is a collection of events in $\mathcal{F}$, then

$$\limsup_{n\to\infty} E_n = [\ E_n \text{ i.o.}\ ] = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} E_m.$$

Lemma 3.3.1 *If* $\{E_n,\ n = 1, 2, \ldots\}$ *is a collection of events in* $\mathcal{F}$ *such that*

$$\sum_{n=1}^{\infty} \mathbb{P}\left[E_n\right] < \infty,$$

then it is always the case that $\mathbb{P}\left[\ E_n \text{ i.o.}\ \right] = 0$.

Proof. Obviously,

$$\begin{aligned}
&\mathbb{P}\left[\ E_n \text{ i.o.}\ \right] \\
&\quad = \mathbb{P}\left[\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} E_m\right] \\
&\quad = \lim_{n\to\infty} \mathbb{P}\left[\cup_{m=n}^{\infty} E_m\right] \quad \text{[By monotonicity in } n \text{ and Lemma 3.1.2]} \\
&\quad = \lim_{n\to\infty} \left(\lim_{k\to\infty} \mathbb{P}\left[\cup_{m=n}^{n+k} E_m\right] \right) \quad \text{[By monotonicity in } k \text{ and Lemma 3.1.1]} \\
&\quad \leq \lim_{n\to\infty} \left(\lim_{k\to\infty} \sum_{m=n}^{n+k} \mathbb{P}\left[E_m\right] \right) \quad \text{[By the union bound on } \mathbb{P}\left[\cup_{m=n}^{n+k} E_m\right]\text{]} \\
&\quad \leq \lim_{n\to\infty} \left(\sum_{m=n}^{\infty} \mathbb{P}\left[E_m\right] \right).
\end{aligned}$$

The result follows since $\lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}\left[E_m\right] = 0$ under the summability condition $\sum_{n=1}^{\infty} \mathbb{P}\left[E_n\right] < \infty$. ■

It is natural to wonder what happens to the conclusion of Lemma 3.3.2 when the condition $\sum_{n=1}^{\infty} \mathbb{P}\left[E_n\right] = \infty$ holds instead. If we add independence, then the following result holds.

Lemma 3.3.2 *When the events* $\{E_n,\ n = 1, 2, \ldots\}$ *in* $\mathcal{F}$ *are mutually independent, then* $\mathbb{P}\left[\ E_n \text{ i.o.}\ \right] = 1$ *under the condition*

$$\sum_{n=1}^{\infty} \mathbb{P}\left[E_n\right] = \infty.$$

Proof. Our point of departure is the observation that

$$[E_n \text{ i.o. }]^c = \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} E_m^c.$$

By arguments similar to the one given in the proof of Lemma 3.3.1 we get

$$\begin{aligned}
&1 - \mathbb{P}[\, E_n \text{ i.o. }] \\
&\quad = \mathbb{P}\left[\cup_{n=1}^{\infty} \cap_{m=n}^{\infty} E_m^c\right] \\
&\quad = \lim_{n\to\infty} \mathbb{P}\left[\cap_{m=n}^{\infty} E_m^c\right] \quad \text{[By monotonicity in } n \text{ and Lemma 3.1.1]} \\
&\quad = \lim_{n\to\infty} \left(\lim_{k\to\infty} \mathbb{P}\left[\cap_{m=n}^{n+k} E_m^c\right]\right) \quad \text{[By monotonicity in } k \text{ and Lemma 3.1.2].}
\end{aligned}$$

For each $n = 1, 2, \ldots$ and $k = 1, 2, \ldots$, we see that

$$\begin{aligned}
\mathbb{P}\left[\cap_{m=n}^{n+k} E_m^c\right] &= \prod_{m=n}^{n+k} \mathbb{P}\left[E_m^c\right] \quad \text{[By mutual independence]} \\
&= \prod_{m=n}^{n+k} \left(1 - \mathbb{P}\left[E_m\right]\right) \\
&\le \prod_{m=n}^{n+k} e^{-\mathbb{P}[E_m]} \quad \text{[Because } 1 - x \le e^{-x},\ x \ge 0] \\
&= e^{-\sum_{m=n}^{n+k} \mathbb{P}[E_m]}. \qquad (3.5)
\end{aligned}$$

Thus, $\lim_{k\to\infty} \mathbb{P}\left[\cap_{m=n}^{n+k} E_m^c\right] = 0$ since $\sum_{n=1}^{\infty} \mathbb{P}[E_n] = \infty$, and the desired conclusion follows. ■

Without the assumption of independence the conclusion of Lemma 3.3.2 may not hold as the following example shows.

Counterexample 3.3.1 Consider the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ (see Chapter 5) where $\Omega = [0, 1]$, $\mathcal{F}$ is the Borel σ-field on $[0, 1]$ and $\mathbb{P}$ is Lebesgue measure λ on $\mathcal{F}$. As explained in Chapter 4, the Borel σ-field on $[0, 1]$ contains all closed intervals of the form $[a, b]$ with $0 \le a \le b \le 1$ and $\lambda([a, b]) = b - a$. Now consider the collection of events $\{E_n,\ n = 1, 2, \ldots\}$ with $E_n = [0, n^{-1}]$ for each $n = 1, 2, \ldots$. It is easy to check that $[E_n \text{ i.o. }] = \cap_{n=1}^{\infty} E_n = \{0\}$ [See Exercise 3.3]. Therefore, $\mathbb{P}[E_n \text{ i.o. }] = 0$, and yet $\sum_{n=1} \mathbb{P}[E_n] = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$. ■

We refer the reader to Exercise 3.9 for a more general version of this counterexample.

3.4 Exercises

All exercises assume an underlying probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 3.1 Show the validity of (3.2).

Ex. 3.2 Show that the union bound (2.3) also holds when I is countable, finite or not [**HINT:** The union bound in the countably infinite case follows from the fact that it holds for any finite I and an easy application of Lemma 3.1.1].

Ex. 3.3 On the way to show that Proposition 3.2.1 subsumes both Lemma 3.1.2 and Lemma 3.1.2, consider a monotone sequence of events $\{E_n,\ n = 1, 2, \ldots\}$ in $\mathcal{F}$.

a. Under the monotone increasing assumption of Lemma 3.1.1, show that $\limsup_{n\to\infty} E_n = \liminf_{n\to\infty} E_n = \cup_{n=1}^{\infty} E_n$.

b. Under the monotone decreasing assumption of Lemma 3.1.2, show that $\limsup_{n\to\infty} E_n = \liminf_{n\to\infty} E_n = \cap_{n=1}^{\infty} E_n$.

Ex. 3.4 Let $\{E_n,\ n = 1, 2, \ldots\}$ and $\{F_n,\ n = 1, 2, \ldots\}$ be two collections of events in $\mathcal{F}$.

a. Show that

$$\limsup_{n\to\infty} \left(E_n \cap F_n\right) \subseteq \left(\limsup_{n\to\infty} E_n\right) \cap \left(\limsup_{n\to\infty} E_n\right)$$

and

$$\left(\limsup_{n\to\infty} E_n\right) \cup \left(\limsup_{n\to\infty} F_n\right) \subseteq \limsup_{n\to\infty} \left(E_n \cup F_n\right)$$

b. Give examples where the inclusions are strict and where they hold as equality.

Ex. 3.5 Repeat Exercise 3.4 for the liminf operation.

Ex. 3.6 A different route to Proposition 3.2.1:

a. With $\{E_n,\ n = 1, 2, \ldots\}$ an arbitrary collection of events in $\mathcal{F}$, show the validity of the inequalities

$$\mathbb{P}\left[\liminf_{n\to\infty} E_n\right] \leq \liminf_{n\to\infty} \mathbb{P}\left[E_n\right]$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left[E_n\right] \leq \mathbb{P}\left[\limsup_{n\to\infty} E_n\right].$$

b. Use Part **a** to construct another proof of Proposition 3.2.1. In particular show that the result holds under a condition weaker than the set-theoretic condition (3.3), namely

$$\mathbb{P}\left[\left(\limsup_{n\to\infty} E_n\right) \Delta \left(\liminf_{n\to\infty} E_n\right)\right] = 0.$$

.

Ex. 3.7 Let $\{E_-, E_+, E_n,\ n = 1, 2, \ldots\}$ be a collection of events in $\mathcal{F}$ such that $E_- \subseteq E_n \subseteq E_+$ for all $n = 1, 2, \ldots$. If $\mathbb{P}[E_-] = \mathbb{P}[E_+]$, show that

$$\mathbb{P}\left[\liminf_{n\to\infty} E_n\right] = \mathbb{P}\left[\limsup_{n\to\infty} E_n\right] = \mathbb{P}[E_-] = \mathbb{P}[E_+].$$

.

Ex. 3.8 Consider a sequence of events $\{E_n,\ n = 1, 2, \ldots\}$ such that $\mathbb{P}[E_n] = 1$ for all $n = 1, 2, \ldots$

a. Show that

$$\mathbb{P}[\cap_{j\in J} E_j] = 1 \quad \text{and} \quad \mathbb{P}[\cup_{j\in J} E_j] = 1, \qquad \begin{array}{c} J \subseteq \{1, 2, \ldots\} \\ 1 \leq |J| < \infty. \end{array}$$

b. What can you say concerning the value of $\mathbb{P}[\ E_n \text{ i.o.}\ ]$? Can the answer be obtained by making use of the Borel-Cantelli Lemmas?

Ex. 3.9 Generalizing Counterexample 3.3.1: Consider a monotone decreasing sequence of events $\{E_n,\ n = 1, 2, \ldots\}$ Assume that $0 < \mathbb{P}[E_n] < 1$ for all $n = 1, 2, \ldots$.

a. Under such circumstances show that the events $\{E_n,\ n = 1, 2, \ldots\}$ cannot be mutually independent.

b. It follows by Lemma 3.1.2 that $\mathbb{P}[\ E_n \text{ i.o.}\ ] = \mathbb{P}[\cap_{n=1}^{\infty} E_n] = \lim_{n\to\infty} \mathbb{P}[E_n]$. Use this fact to conclude that the entire range $0 \leq \mathbb{P}[\ E_n \text{ i.o.}\ ] < 1$ of values is possible under the condition $\sum_{n=12}^{\infty} \mathbb{P}[E_n] = \infty$ when the events $\{E_n,\ n = 1, 2, \ldots\}$ fail to be independent.

Chapter 4

Measurable mappings: A tale of σ-fields

Many situations of interest naturally require that the sample space be *uncountable* – For instance, for some models it is appropriate for the sample space to be $\mathbb{R}^p$ (or a subset thereof); see Chapter 6 for some concrete examples. Unfortunately in such cases determining the appropriate σ-field of events on which to define the probability measures is technically more delicate: In a nutshell the required σ-additivity imposes too many constraints if the probability measure is to be defined on the entire power set of the sample space. This precludes that the power set of the sample space be used as the σ-field (as we did for the countable case in Section 1.5).

In the present chapter we start to address the challenge of constructing σ-fields on uncountable sample spaces such as (subsets of) $\mathbb{R}^p$. The basic idea is to tie the definition to the underlying topological properties of the sample space. This arises from the need to assign a likelihood of occurrence to certain subsets of the sample space, said subsets being building blocks of the standard topology on $\mathbb{R}^p$. This leads to the important notion of the Borel σ-field on $\mathbb{R}^p$ and to the related concept of Borel measurability.

The narrative continues in Chapter 5 and in Chapter 6: In Chapter 5 we introduce an approach due to C. Carathéodory (originally developed to define Lebesgue measure), and explain how it can be leveraged to construct the desired measures. Chapter 6 then presents a number of extension results that can be used in the context of some important applications.

4.1 Measurable mappings

We begin by discussing the notion of *measurability* of a mapping: To fix the notation, let S_a and S_b be arbitrary sets (but possibly identical). For any mapping $g : S_a \to S_b$, recall that

$$g^{-1}(E_b) \equiv \{s_a \in S_a : \ g(s_a) \in E_b\}\,, \quad E_b \in \mathcal{P}(S_b).$$

With $\mathcal{H}_b$ denoting a collection of subsets of S_b, it will be natural to extend this notation to collections of subsets by writing

$$g^{-1}(\mathcal{H}_b) \equiv \left\{g^{-1}(E_b) : \ E_b \in \mathcal{H}_b\right\}.$$

Consider now the situation where the sets S_a and S_b are equipped with σ-fields $\mathcal{S}_a$ and $\mathcal{S}_b$, respectively – Thus, the pairs $(S_a, \mathcal{S}_a)$ and $(S_b, \mathcal{S}_b)$ are measurable spaces. In cases where $S_a = S_b \equiv S$ we could in principle have distinct σ-fields $\mathcal{S}_a$ and $\mathcal{S}_b$ on S.

Definition 4.1.1

When the sets S_a and S_b are equipped with σ-fields $\mathcal{S}_a$ and $\mathcal{S}_b$, respectively, the mapping $g : S_a \to S_b$ is said to be $(\mathcal{S}_a, \mathcal{S}_b)$-*measurable* if the conditions

$$g^{-1}(E_b) \in \mathcal{S}_a, \quad E_b \in \mathcal{S}_b \tag{4.1}$$

all hold.

The conditions (4.1) can be rewritten more compactly as

$$g^{-1}(\mathcal{S}_b) \subseteq \mathcal{S}_a. \tag{4.2}$$

Adding a third set S_c equipped with a σ-field $\mathcal{S}_c$, we consider now a situation where there are now three measurable spaces $(S_a, \mathcal{S}_a)$, $(S_b, \mathcal{S}_b)$ and $(S_c, \mathcal{S}_c)$. With mappings $g : S_a \to S_b$ and $h : S_b \to S_c$, we associate the *composition* mapping $h \circ g : S_a \to S_c$ given by

$$(h \circ g)\,(s_a) \equiv h(g(s_a)), \quad s_a \in S_a.$$

The measurability of the composition mapping is straightforward.

Fact 4.1.1 *If the mapping $g : S_a \to S_b$ is $(\mathcal{S}_a, \mathcal{S}_b)$-measurable and the mapping $h : S_b \to S_c$ is $(\mathcal{S}_b, \mathcal{S}_c)$-measurable, then the composition mapping $h \circ g : S_a \to S_c$ is itself $(\mathcal{S}_a, \mathcal{S}_c)$-measurable.*

Proof. The conclusion follows from the elementary set-theoretic fact

$$(h \circ g)^{-1}(E_c) = g^{-1}(h^{-1}(E_c)), \quad E_c \in \mathcal{P}(S_c) \tag{4.3}$$

when coupled with the $(\mathcal{S}_b, \mathcal{S}_c)$-measurability of h and the $(\mathcal{S}_a, \mathcal{S}_b)$-measurability of g. Details are left to the interested reader [Exercise 4.2]. ■

Lemma 4.1.1 given next is key to showing that the measurability of a mapping can often be determined by checking a *reduced* set of conditions.

Lemma 4.1.1 *Let $\mathcal{H}_b$ be a collection of subsets of S_b. For any mapping $g : S_a \to S_b$, the following statements hold:*

(i) If $\mathcal{H}_b$ is a σ-field on S_b, then the collection $g^{-1}(\mathcal{H}_b)$ is a σ-field on S_a;

(ii) More generally, we always have

$$g^{-1}(\sigma(\mathcal{H}_b)) = \sigma\left(g^{-1}(\mathcal{H}_b)\right). \tag{4.4}$$

Proof. Claim (i): We leave it as an exercise [Exercise 4.1] to check that the collection $g^{-1}(\mathcal{H}_b)$ is a σ-field on S_a when $\mathcal{H}_b$ is a σ-field on S_b.

Claim (ii): We now turn to establishing (4.4): By Part (i) applied to the σ-field $\sigma(\mathcal{H}_b)$, the collection $g^{-1}(\sigma(\mathcal{H}_b))$ is a σ-field which contains $g^{-1}(\mathcal{H}_b)$, and the inclusion $\sigma\left(g^{-1}(\mathcal{H}_b)\right) \subseteq g^{-1}(\sigma(\mathcal{H}_b))$ is straightforward by virtue of Fact 1.7.2.

To establish the reverse inclusion $g^{-1}(\sigma(\mathcal{H}_b)) \subseteq \sigma\left(g^{-1}(\mathcal{H}_b)\right)$, consider the collection $\mathcal{H}^\star_{b,g}$ of subsets given by

$$\mathcal{H}^\star_{b,g} \equiv \left\{E_b \in \mathcal{P}(S_b) : \ g^{-1}(E_b) \in \sigma\left(g^{-1}(\mathcal{H}_b)\right)\right\}. \tag{4.5}$$

It is easy to check that $\mathcal{H}^\star_{b,g}$ is a σ-field on S_b [Exercise 4.1] containing $\mathcal{H}_b$. Therefore, $\mathcal{H}^\star_{b,g}$ contains $\sigma(\mathcal{H}_b)$ and we get

$$g^{-1}(\sigma(\mathcal{H}_b)) \subseteq g^{-1}\left(\mathcal{H}^\star_{b,g}\right) \subseteq \sigma\left(g^{-1}(\mathcal{H}_b)\right)$$

where the last inclusion follows by the definition of $\mathcal{H}^\star_{b,g}$. This completes the proof of (4.4). ■

We now use Lemma 4.1.1 to obtain an equivalent definition of measurability.

Lemma 4.1.2 *If the σ-field $\mathcal{S}_b$ on S_b is generated by the collection $\mathcal{H}_b$ of subsets of S_b, i.e., $\mathcal{S}_b = \sigma(\mathcal{H}_b)$, then the mapping $g : S_a \to S_b$ is $(\mathcal{S}_a, \mathcal{S}_b)$-measurable if and only if the conditions*

$$g^{-1}(E_b) \in \mathcal{S}_a, \quad E_b \in \mathcal{H}_b \tag{4.6}$$

all hold.

In the same way that the conditions (4.2) are equivalent to (4.1), we can write the conditions (4.6) in the equivalent form

$$g^{-1}(\mathcal{H}_b) \subseteq \mathcal{S}_a. \tag{4.7}$$

The equivalence stated in Lemma 4.1.2 is *operationally* useful in that only the reduced *subset* (4.6) of conditions (associated with the *generator* $\mathcal{H}_b$ for $\mathcal{S}_b$) needs to be checked instead of the entire set (4.1) – An important example will be discussed shortly in the next section.

Proof. The condition (4.2) obviously implies (4.7) since $\mathcal{H}_b \subseteq \mathcal{S}_b$. Conversely, assume that the mapping $g : S_a \to S_b$ satisfies (4.7): The equality $g^{-1}(\mathcal{S}_b) = g^{-1}(\sigma(\mathcal{H}_b))$ obviously holds since $\mathcal{S}_b = \sigma(\mathcal{H}_b)$ by assumption, while the equality $g^{-1}(\sigma(\mathcal{H}_b)) = \sigma\left(g^{-1}(\mathcal{H}_b)\right)$ follows from Lemma 4.1.1 – Combining these two equalities gives $g^{-1}(\mathcal{S}_b) = \sigma\left(g^{-1}(\mathcal{H}_b)\right)$. Finally, under condition (4.7) we conclude that $\sigma\left(g^{-1}(\mathcal{H}_b)\right) \subseteq \mathcal{S}_a$ as we use the fact that $\mathcal{S}_a$ is itself a σ-field; see Fact 1.7.2. This complete the proof of Lemma 4.1.2. ■

4.2 The Borel σ-field on $\mathbb{R}$

We refer to a subset I of $\mathbb{R}$ of the form (a, b) (with $a \leq b$ in $\mathbb{R}$) as a *bounded open interval*. Let $\mathcal{I}(\mathbb{R})$ denote the collection of all bounded open intervals of $\mathbb{R}$.

As can be seen from the discussion in Section **??**, it is quite natural to consider assigning a measure to such intervals. This requires at minimum that we consider the σ-field generated by $\mathcal{I}(\mathbb{R})$ as we do next.

Definition 4.2.1

The *Borel* σ-field on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-field on $\mathbb{R}$ containing all bounded open intervals of $\mathbb{R}$, i.e., $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{I}(\mathbb{R}))$.

The Borel σ-field $\mathcal{B}(\mathbb{R})$ can be generated in many different ways. To see this consider the following collections of subsets of $\mathbb{R}$:

The bounded open intervals

$$\mathcal{H}_0(\mathbb{R}) \equiv \left\{ (a,b), \quad \begin{matrix} a \leq b \\ a, b \in \mathbb{R} \end{matrix} \right\} = \mathcal{I}(\mathbb{R}).$$

The bounded closed intervals

$$\mathcal{H}_1(\mathbb{R}) \equiv \left\{ [a,b], \quad \begin{matrix} a \leq b \\ a, b \in \mathbb{R} \end{matrix} \right\}.$$

The bounded open-closed intervals

$$\mathcal{H}_2(\mathbb{R}) \equiv \left\{ (a,b], \quad \begin{matrix} a \leq b \\ a, b \in \mathbb{R} \end{matrix} \right\}.$$

The bounded closed-open intervals

$$\mathcal{H}_3(\mathbb{R}) \equiv \left\{ [a,b), \quad \begin{matrix} a \leq b \\ a, b \in \mathbb{R} \end{matrix} \right\}.$$

The open semi-intervals

$$\mathcal{H}_4(\mathbb{R}) \equiv \{(-\infty, a), \ a \in \mathbb{R}\}.$$

The closed semi-intervals

$$\mathcal{H}_5(\mathbb{R}) \equiv \{(-\infty, a], \ a \in \mathbb{R}\}.$$

The open semi-intervals

$$\mathcal{H}_6(\mathbb{R}) \equiv \{(a, +\infty) : \ a \in \mathbb{R}\}.$$

A key observation is contained in the following result.

Lemma 4.2.1 *With the notation above, it holds that*

$$\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{H}_k(\mathbb{R})), \quad k = 0, 1, \ldots, 6. \tag{4.8}$$

The approximation arguments given in the proof of Lemma 4.2.1 can also be used in the multi-dimensional setting of Section 4.4.

Proof. Fix a and b in $\mathbb{R}$ with $a \leq b$. The set-theoretic facts $[a,b] = \cap_{n=1}^\infty (a - \frac{1}{n}, b + \frac{1}{n})$ and $(a,b] = \cap_{n=1}^\infty (a, b + \frac{1}{n})$ readily imply $\mathcal{H}_1 \subseteq \sigma(\mathcal{H}_0)$ and $\mathcal{H}_2 \subseteq \sigma(\mathcal{H}_0)$. On the other hand we also have

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

with $n(a,b) = \lceil 2(b-a)^{-1} \rceil$. These two equalities readily imply $\mathcal{H}_0 \subseteq \sigma(\mathcal{H}_1)$ and $\mathcal{H}_0 \subseteq \sigma(\mathcal{H}_2)$, respectively. It immediately follows that $\sigma(\mathcal{H}_0) = \sigma(\mathcal{H}_1)$ and $\sigma(\mathcal{H}_0) = \sigma(\mathcal{H}_2)$. A similar argument also shows that $\sigma(\mathcal{H}_0) = \sigma(\mathcal{H}_3)$, and we conclude that $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{H}_k)$ for $k = 0, 1, 2, 3$.

In the same vein, upon noting that $(-\infty, a] = \cap_{n=1}^\infty (-\infty, a + \frac{1}{n})$ and $(-\infty, a) = \cup_{n=1}^\infty (-\infty, a - \frac{1}{n}]$, we conclude that $\mathcal{H}_5 \subseteq \sigma(\mathcal{H}_4)$ and $\mathcal{H}_4 \subseteq \sigma(\mathcal{H}_5)$, respectively, and the equality $\sigma(\mathcal{H}_4) = \sigma(\mathcal{H}_5)$ follows. The fact that $\sigma(\mathcal{H}_6) = \sigma(\mathcal{H}_5)$ is an easy consequence of the fact that the complement of any subset in $\mathcal{H}_6$ is a subset in $\mathcal{H}_5$, and vice versa.

Next, the inclusion $\mathcal{H}_4 \subseteq \sigma(\mathcal{H}_2)$ holds since $(-\infty, a] = \cup_{n=0}^\infty (a - (n+1), a - n]$, hence $\sigma(\mathcal{H}_4) \subseteq \sigma(\mathcal{H}_2) = \mathcal{B}(\mathbb{R})$. To get the reverse inclusion, start with the observation that $(a,b) = (-\infty, b) \cap (-\infty, a]^c$. This shows that $\mathcal{H}_0 \subseteq \sigma(\mathcal{H}_4) = \sigma(\mathcal{H}_5)$, hence $\sigma(\mathcal{H}_0) = \mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{H}_4)$. The equality $\sigma(\mathcal{H}_0) = \sigma(\mathcal{H}_4)$ follows, and the proof of (4.8) for $k = 4, 5$ is complete.

■

In spite of its seemingly simple definition, namely $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{I}(\mathbb{R}))$, the Borel σ-field constitutes an extremely large and rather unwieldy collection of objects. To make this point even more apparent we now provide a characterization of $\mathcal{B}(\mathbb{R})$ in terms of a generator much larger than the ones appearing in Lemma 4.2.1.

We set the stage with a well-known definition from the standard topology on $\mathbb{R}$:

Definition 4.2.2

A subset U of $\mathbb{R}$ is *open* if for every x in U, there exists a bounded open interval I_x (in $\mathcal{I}(\mathbb{R})$) containing x (i.e., $x \in I_x$) and contained in U (i.e., $I_x \subseteq U$). A set F is said to be *closed* if its complement F^c (in $\mathbb{R}$) is open.

Let $\mathcal{O}(\mathbb{R})$ denote the collection of all open subsets of $\mathbb{R}$. It is elementary to check that bounded open intervals and all open semi-intervals in $\mathcal{H}_4(\mathbb{R})$ are open sets, and that bounded closed intervals in $\mathcal{H}_1(\mathbb{R})$ and all closed semi-intervals in $\mathcal{H}_5(\mathbb{R})$ are closed sets. However the bounded open-closed intervals in $\mathcal{H}_2(\mathbb{R})$ and closed-open intervals in $\mathcal{H}_3(\mathbb{R})$ are neither open nor closed.

The key technical point that highlights the importance of bounded open intervals as building blocks for the usual topology on $\mathbb{R}$ is given next; see [**?**] for a proof.

Fact 4.2.1 *Any open subset U in $\mathbb{R}$ can be expressed as the union of a countable collection of non-overlapping open intervals, i.e., there exists a countable collection $\{J_i, i \in I\}$ of open intervals of $\mathbb{R}$ such that*

$$U = \cup_{i\in I} J_i \quad \textit{with} \quad J_k \cap J_\ell = \emptyset, \quad \begin{array}{c} k \neq \ell \\ k, \ell \in I. \end{array} \tag{4.9}$$

Fact 4.2.1 leads easily to the following characterization of $\mathcal{B}(\mathbb{R})$ in terms of open sets.

Lemma 4.2.2 *The smallest σ-field on $\mathbb{R}$ containing all open subsets of $\mathbb{R}$ coincides with the Borel σ-field $\mathcal{B}(\mathbb{R})$ on $\mathbb{R}$, i.e., $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{O}(\mathbb{R}))$.*

Proof. As pointed earlier, we already have $\mathcal{I}(\mathbb{R}) \subseteq \mathcal{O}(\mathbb{R})$, hence $\sigma(\mathcal{I}(\mathbb{R})) \subseteq \sigma(\mathcal{O}(\mathbb{R}))$, and the inclusion $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{O}(\mathbb{R}))$ holds. To obtain the reverse inclusion, note that $\mathcal{O}(\mathbb{R}) \subseteq \sigma(\mathcal{I}(\mathbb{R}))$ by Fact 4.2.1, hence $\sigma(\mathcal{O}(\mathbb{R})) \subseteq \sigma(\mathcal{I}(\mathbb{R}))$. In other words, $\sigma(\mathcal{O}(\mathbb{R})) \subseteq \mathcal{B}(\mathbb{R})$, and the proof is complete. ■

In short, the collection of all open sets on $\mathbb{R}$ generates the Borel σ-field $\mathcal{B}(\mathbb{R})$, thereby highlighting its connection with the standard topology on $\mathbb{R}$.

4.3 Cartesian products

Before discussing the multi-dimensional case we introduce some facts concerning cartesian products: Let S_a and S_b be two arbitrary sets (possibly identical). If $\mathcal{H}_a$ and $\mathcal{H}_b$ are collections of subsets of S_a and S_b, respectively, it is natural to set

$$\mathcal{H}_a \times \mathcal{H}_b \equiv \{E_a \times E_b : \ E_a \in \mathcal{H}_a, E_b \in \mathcal{H}_b\}.$$

A set $E_a \times E_b$ in $\mathcal{H}_a \times \mathcal{H}_b$ is called a *rectangle* with sides E_a in S_a and E_b in S_b. These notions generalize to more than two factors in an obvious manner.

In general the collection $\mathcal{H}_a \times \mathcal{H}_b$ is *not* a σ-field (resp. a field) on the Cartesian product $S_a \times S_b$ *even* if each of the collections $\mathcal{H}_a$ and $\mathcal{H}_b$ is itself a σ-field (resp. a field) [Exercise 4.4]. The next result shows how generators on the individual factors give rise to a natural notion of measurability on Cartesian products.

Lemma 4.3.1 *Let S_a and S_b be two arbitrary sets. If $\mathcal{H}_a$ and $\mathcal{H}_b$ are collections of subsets of S_a and S_b, respectively, then it holds that*

$$\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right) = \sigma\left(\sigma\left(\mathcal{H}_a\right) \times \sigma\left(\mathcal{H}_b\right)\right). \tag{4.10}$$

On the basis of (4.10) it is customary to write

$$\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right) = \sigma\left(\mathcal{H}_a\right) \otimes \sigma\left(\mathcal{H}_b\right).$$

Proof. As the inclusion $\mathcal{H}_a \times \mathcal{H}_b \subseteq \sigma\left(\mathcal{H}_a\right) \times \sigma\left(\mathcal{H}_b\right)$ obviously holds, we immediately get the inclusion $\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right) \subseteq \sigma\left(\sigma\left(\mathcal{H}_a\right) \times \sigma\left(\mathcal{H}_b\right)\right)$. To establish the reverse inclusion

$$\sigma\left(\sigma\left(\mathcal{H}_a\right) \times \sigma\left(\mathcal{H}_b\right)\right) \subseteq \sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right), \tag{4.11}$$

we proceed as follows: Define the collections

$$\mathcal{H}_a^\star \equiv \left\{E_a \in \mathcal{P}(S_a) : \ E_a \times S_b \in \sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right)\right\} \tag{4.12}$$

and

$$\mathcal{H}_b^\star \equiv \left\{E_b \in \mathcal{P}(S_b) : \ S_a \times E_b \in \sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right)\right\}. \tag{4.13}$$

It is a simple matter to check that $\mathcal{H}_a^\star$ and $\mathcal{H}_b^\star$ are σ-fields on S_a and S_b, respectively [Exercise 4.5].

Pick an arbitrary subset E of $S_a \times S_b$ that belongs to $\mathcal{H}_a^\star \times \mathcal{H}_b^\star$. Thus, $E = E_a \times E_b$ with E_a in $\mathcal{H}_a^\star$ and E_b in $\mathcal{H}_b^\star$. By definition both cartesian products $E_a \times S_b$ and $S_a \times E_b$ belong to the σ-field $\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right)$, hence their intersection also belongs to $\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right)$. However, as we note that $(E_a \times S_b) \cap (S_a \times E_b) = E_a \times E_b$, we conclude that E is also an element of $\sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right)$. Put differently, we have just shown that

$$\mathcal{H}_a^\star \times \mathcal{H}_b^\star \subseteq \sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right). \tag{4.14}$$

Obviously $\mathcal{H}_a \subseteq \mathcal{H}_a^\star$, hence $\sigma\left(\mathcal{H}_a\right) \subseteq \mathcal{H}_a^\star$ since $\mathcal{H}_a^\star$ is σ-field on S_a [Exercise 4.5]. We similarly have $\sigma\left(\mathcal{H}_b\right) \subseteq \mathcal{H}_b^\star$. It follows from (4.14) that

$$\sigma\left(\mathcal{H}_a\right) \times \sigma\left(\mathcal{H}_b\right) \subseteq \sigma\left(\mathcal{H}_a \times \mathcal{H}_b\right).$$

and the conclusion (4.11) follows. The proof of (4.10) is now complete. ■

The discussion easily generalizes to more than two factors. For instance, consider sets S_a, S_b and S_c and let $\mathcal{S}_a$, $\mathcal{S}_b$ and $\mathcal{S}_c$ be collections of subsets of S_a, S_b and S_c, respectively. Applying Lemma 4.3.1 repeatedly we get

$$\begin{aligned}
\sigma\left(\mathcal{S}_a \times (\mathcal{S}_b \times \mathcal{S}_c)\right) &= \sigma\left(\sigma\left(\mathcal{S}_a\right) \times \sigma\left(\mathcal{S}_b \times \mathcal{S}_c\right)\right) \\
&= \sigma\left(\sigma\left(\mathcal{S}_a\right) \times \sigma\left(\sigma\left(\mathcal{S}_b\right) \times \sigma\left(\mathcal{S}_c\right)\right)\right) \\
&= \sigma\left(\sigma\left(\mathcal{S}_a\right) \times \left(\sigma\left(\mathcal{S}_b\right) \times \sigma\left(\mathcal{S}_c\right)\right)\right), \qquad (4.15)
\end{aligned}$$

while similar arguments (applied to $\mathcal{S}_a \times (\mathcal{S}_b \times \mathcal{S}_c)$) show that

$$\sigma\left(\mathcal{S}_a \times (\mathcal{S}_b \times \mathcal{S}_c)\right) = \sigma\left(\left(\sigma\left(\mathcal{S}_a\right) \times \sigma\left(\mathcal{S}_b\right)\right) \times \sigma\left(\mathcal{S}_c\right)\right). \qquad (4.16)$$

In the same vein that $\mathcal{S}_a \times (\mathcal{S}_b \times \mathcal{S}_c)$ is *identified* with $\mathcal{S}_a \times (\mathcal{S}_b \times \mathcal{S}_c)$, a fact summarily written $\mathcal{S}_a \times \mathcal{S}_b \times \mathcal{S}_c$, we identify $\sigma\left(\sigma\left(\mathcal{S}_a\right) \times \left(\sigma\left(\mathcal{S}_b\right) \times \sigma\left(\mathcal{S}_c\right)\right)\right)$ with $\sigma\left(\left(\sigma\left(\mathcal{S}_a\right) \times \sigma\left(\mathcal{S}_b\right)\right) \times \sigma\left(\mathcal{S}_c\right)\right)$ and write $\sigma\left(\sigma\left(\mathcal{S}_a\right) \times \sigma\left(\mathcal{S}_b\right) \times \sigma\left(\mathcal{S}_c\right)\right)$. Combining these conventions we shall write

$$\sigma\left(\mathcal{S}_a \times \mathcal{S}_b \times \mathcal{S}_c\right) = \sigma\left(\mathcal{S}_a\right) \otimes \sigma\left(\mathcal{S}_b\right) \otimes \sigma\left(\mathcal{S}_c\right).$$

These conventions and notation readily generalize to situations with multiple factors.

4.4 The Borel σ-fields on $\mathbb{R}^p$ $(p = 2, 3, \ldots)$

As we turn to the multi-dimensional case, let p denote a fixed positive integer. In higher dimensions it also quite natural to consider assigning a measure to certain subsets of $\mathbb{R}^p$, say subsets of the form

$$B_1 \times \ldots \times B_p,$$

where for each $k = 1, \ldots, p$, the set B_k is a subset in one of the collections $\mathcal{H}_0(\mathbb{R}), \ldots, \mathcal{H}_5(\mathbb{R})$ introduced in Section 4.2. Specializing this definition to $\mathcal{H}_0(\mathbb{R})$ we get the following definitions.

Definition 4.4.1

An *bounded open rectangle* R in $\mathbb{R}^p$ is a product set of the form $I_1 \times \ldots \times I_p$ where for each $k = 1, \ldots, p$, the factor set I_k is a bounded open interval (a_k, b_k) (with $a_k \leq b_k$ in $\mathbb{R}$). In other words,

$$R = \{(y_1, \ldots, y_p) \in \mathbb{R}^p : \ a_k < y_k < b_k, \ k = 1, 2, \ldots, p\}.$$

Let $\mathcal{R}_{\text{Open}}\left(\mathbb{R}^p\right)$ denote the collection of all bounded open rectangles in $\mathbb{R}^p$. Obviously we have $\mathcal{R}_{\text{Open}}\left(\mathbb{R}^p\right) = \mathcal{H}_0\left(\mathbb{R}\right) \times \ldots \times \mathcal{H}_0\left(\mathbb{R}\right) = \mathcal{H}_0\left(\mathbb{R}\right)^p$, or in a slightly different notation, $\mathcal{R}_{\text{Open}}\left(\mathbb{R}^p\right) = \mathcal{I}\left(\mathbb{R}\right) \times \ldots \times \mathcal{I}\left(\mathbb{R}\right) = \mathcal{I}\left(\mathbb{R}\right)^p$, the latter clearly showing that $\mathcal{R}_{\text{Open}}\left(\mathbb{R}^p\right)$ is a natural multi-dimensional generalization of $\mathcal{I}\left(\mathbb{R}\right)$. In analogy with Definition 4.2.1 given for $p = 1$ we now introduce the notion of a Borel σ-field on $\mathbb{R}^p$.

Definition 4.4.2

The *Borel* σ-field on $\mathbb{R}^p$, denoted $\mathcal{B}(\mathbb{R}^p)$, is the smallest σ-field on $\mathbb{R}^p$ containing all bounded open rectangles in $\mathbb{R}^p$, i.e., $\mathcal{B}(\mathbb{R}^p) \equiv \sigma\left(\mathcal{R}_{\text{Open}}\left(\mathbb{R}^p\right)\right)$.

Not too surprisingly, the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$ is related to the usual topology on $\mathbb{R}^p$ (as the Borel σ-field $\mathcal{B}(\mathbb{R})$ was to the usual topology on $\mathbb{R}$). To clarify this connection further consider next the usual notion of an open set in $\mathbb{R}^p$ whose definition is analogous to Definition 4.2.1 given for $p = 1$.

Definition 4.4.3

A subset U of $\mathbb{R}^p$ is *open* if for every $x = (x_1, \ldots, x_p)$ in U, there exists a bounded open rectangle R_x (in $\mathcal{R}_{\text{Open}}(\mathbb{R}^p)$) containing x (i.e., $x \in R_x$) and contained in U (i.e., $R_x \subseteq U$). A set F is said to be *closed* if its complement F^c (in $\mathbb{R}^p$) is open.

In the notation of Definition 4.4.1, the set U is open in $\mathbb{R}^p$ if for every $x = (x_1, \ldots, x_p)$ in U and each $k = 1, \ldots, p$, there exist scalars $a_k(x)$ and $b_k(x)$ such that $a_k(x) < x_k < b_k(x)$ and

$$\{(y_1, \ldots, y_p) \in \mathbb{R}^p : \ a_k < y_k < b_k, \ k = 1, \ldots\} \subseteq U.$$

Let $\mathcal{O}(\mathbb{R}^p)$ denote the collection of all open sets in $\mathbb{R}^p$.

In the scalar case, according to Fact 4.2.1 the bounded open intervals are the building blocks of the standard topology on $\mathbb{R}$. A similar situation holds in higher dimensions in that open rectangles in $\mathbb{R}^p$ are now the building blocks of the standard topology on $\mathbb{R}^p$. This is the message of the following well-known fact from topology [**?**].

Fact 4.4.1 *For any open set U in $\mathbb{R}^p$ there exists a countable family of bounded open rectangles $\{R_i, \ i \in I\}$ in $\mathcal{R}_{\text{Open}}(\mathbb{R}^p)$ with countable I such that $U = \cup_{i \in I} R_i$.*

Fact 4.4.1 easily leads to a characterization of $\mathcal{B}(\mathbb{R}^p)$ in terms of the usual topology on $\mathbb{R}^p$; the proof, left as an easy exercise, mimics that of Lemma 4.2.2 (with $\mathcal{I}(\mathbb{R})$ replaced by $\mathcal{R}_{\mathrm{Open}}(\mathbb{R}^p)$ and leveraging this time Fact 4.4.1) [Exercise 4.6].

Lemma 4.4.1 *The smallest σ-field on $\mathbb{R}^p$ containing all open subsets of $\mathbb{R}^p$ coincides with the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$ on $\mathbb{R}^p$, i.e., $\mathcal{B}(\mathbb{R}^p) \equiv \sigma(\mathcal{O}(\mathbb{R}^p))$.*

As in the scalar case discussed in Section 4.2 the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$ can be generated in many different ways; we leave it to the reader to explore the appropriate multi-dimensional generalization of Lemma 4.2.1. However, for reasons that will soon become apparent in later chapters, there is one generating family that occupies a central place in developing the notion of probability distribution functions for random variables [Chapter **??**]: Let $\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p)$ denote the collection of all *closed southwest rectangles*, namely

$$\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p) \equiv \left\{ J_1 \times \ldots \times J_p, \quad \begin{array}{c} J_k = (-\infty, a_k] \\ a_k \in \mathbb{R} \\ k = 1, \ldots, p \end{array} \right\}.$$

This family is the p-dimensional analog of the one-dimensional family $\mathcal{H}_5(\mathbb{R})$, as can be seen from its representation as the p-fold Cartesian product

$$\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p) = \mathcal{H}_5(\mathbb{R}) \times \ldots \times \mathcal{H}_5(\mathbb{R}) = \mathcal{H}_5(\mathbb{R})^p.$$

Leveraging the ideas of Section 4.3 we obtain the following alternate representation of the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$.

Lemma 4.4.2 *The representation $\mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p))$ holds.*

Proof. By Lemma 4.2.1 we already have $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{H}_0(\mathbb{R})) = \sigma(\mathcal{H}_5(\mathbb{R}))$. Using these facts and the representation $\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p) = \mathcal{H}_5(\mathbb{R}) \times \ldots \times \mathcal{H}_5(\mathbb{R})$, we readily get

$$\begin{aligned}
&\sigma(\mathcal{R}_{\mathrm{SW-Closed}}(\mathbb{R}^p)) \\
&\quad = \sigma(\mathcal{H}_5(\mathbb{R}) \times \ldots \times \mathcal{H}_5(\mathbb{R})) \\
&\quad = \sigma(\sigma(\mathcal{H}_5(\mathbb{R})) \times \ldots \times \sigma(\mathcal{H}_5(\mathbb{R}))) \quad \text{[By Lemma 4.3.1]} \\
&\quad = \sigma(\sigma(\mathcal{H}_0(\mathbb{R})) \times \ldots \times \sigma(\mathcal{H}_0(\mathbb{R}))) \quad \text{[By Lemma 4.2.1]} \\
&\quad = \sigma(\mathcal{H}_0(\mathbb{R}) \times \ldots \times \mathcal{H}_0(\mathbb{R})) \quad \text{[By Lemma 4.3.1]} \\
&\quad = \sigma(\mathcal{R}_{\mathrm{Open}}(\mathbb{R}^p)) \\
&\quad = \sigma(\mathcal{B}(\mathbb{R}^p)) \quad \text{[By Definition 4.4.2].}
\end{aligned} \tag{4.17}$$

■

An inspection of the proof of Lemma 4.4.2 leads to the following observation.

Lemma 4.4.3 *The representation*

$$\mathcal{B}(\mathbb{R}^p) = \sigma\left(\mathcal{B}(\mathbb{R}) \times \ldots \times \mathcal{B}(\mathbb{R})\right) = \mathcal{B}(\mathbb{R}) \otimes \ldots \otimes \mathcal{B}(\mathbb{R}) \tag{4.18}$$

holds.

4.5 Borel mappings

We specialize the definitions of Section 4.1 to the situation when the domain of the mapping is a measurable space $(S, \mathcal{S})$ and its range space is $\mathbb{R}^p$ for some positive integer p: Thus, in Definition 4.1.1 we write $(S, \mathcal{S})$ for $(S_a, \mathcal{S}_a)$, $S_b = \mathbb{R}^p$ and it is understood (unless specified otherwise) that $\mathcal{S}_b = \mathcal{B}(\mathbb{R}^p)$.

Definition 4.5.1

A mapping $g : S \to \mathbb{R}^p$ defined on a measurable space $(S, \mathcal{S})$ is a *Borel* mapping if it is an $(\mathcal{S}, \mathcal{B}(\mathbb{R}^p))$-measurable mapping in the sense of Definition 4.1.1, namely that the conditions

$$g^{-1}(B) \in \mathcal{S}, \quad B \in \mathcal{B}(\mathbb{R}^p) \tag{4.19}$$

all hold.

With $(S_c, \mathcal{S}_c) = (\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$ for some positive integer q, Fact 4.1.1 takes the following form.

Fact 4.5.1 *If $g : S \to \mathbb{R}^p$ and $h : \mathbb{R}^p \to \mathbb{R}^q$ are Borel mappings, then the composition mapping $h \circ g : S \to \mathbb{R}^q$ is also a Borel mapping.*

In this restricted context Lemma 4.1.2 leads to the following fact which is crucial for understanding the importance of probability distributions.

Lemma 4.5.1 *Let $\mathcal{H}$ denote any collection of subsets of $\mathbb{R}^p$ which generates the Borel σ-field $\mathcal{B}(\mathbb{R}^p)$, i.e., $\mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{H})$. The mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if and only if the conditions*

$$g^{-1}(E) \in \mathcal{S}, \quad E \in \mathcal{H} \tag{4.20}$$

all hold.

We close this section by investigating how the measurability of one-dimensional mappings informs the measurability of vector-valued mappings. We begin by noting that any mapping $g : S \to \mathbb{R}^p$ can also be viewed as a p-tuple of mappings $g_1, \ldots, g_p : S \to \mathbb{R}$ where for each $k = 1, \ldots, p$, the mapping $g_k : S \to \mathbb{R}$ picks up the k^{th} coordinate of $g(s)$ so that

$$g(s) = (g_1(s), \ldots, g_p(s)), \quad s \in S.$$

Lemma 4.5.2 *The mapping $g : S \to \mathbb{R}^p$ is a Borel mapping if and only if the mappings $g_1, \ldots, g_p : S \to \mathbb{R}$ are all Borel mappings.*

Proof. We begin with an easy observation: Consider the rectangle R given by

$$R \equiv B_1 \times B_2 \times \ldots \times B_p \tag{4.21}$$

where $B_1, B_2, \ldots, B_p$ are subsets of $\mathbb{R}$. It is elementary to see that

$$\begin{aligned} g^{-1}(R) &= \{s \in S : \ g(s) \in B\} \\ &= \{s \in S : \ g_\ell(s) \in B_\ell, \ \ell = 1, \ldots, p\} \\ &= \cap_{\ell=1}^p \{s \in S : \ g_\ell(s) \in B_\ell\} \\ &= \cap_{\ell=1}^p g_\ell^{-1}(B_\ell). \end{aligned} \tag{4.22}$$

First assume that the mapping $g : S \to \mathbb{R}^p$ is a Borel mapping according to Definition 4.5.1. Fix $k = 1, \ldots, p$ and use (4.22) with $B_\ell = \mathbb{R}$ for $\ell = 1, \ldots, p$ whenever $\ell \neq k$ and $B_k = (-\infty, a_k]$ for some a_k in $\mathbb{R}$. It is plain that a set so constructed is also a Borel subset of $\mathbb{R}^p$ – See the proof of Lemma 4.4.2 where it is shown (among other things) that $\sigma\left(\sigma\left(\mathcal{H}_5(\mathbb{R})\right) \times \ldots \times \sigma\left(\mathcal{H}_5(\mathbb{R})\right)\right)$ coincides with $\mathcal{B}\left(\mathbb{R}^p\right)$. It is also plain that $g^{-1}(R) = g_k^{-1}((-\infty, a_k])$ since for eack $\ell \neq k$ we have $g_\ell^{-1}(B_\ell) = g_\ell^{-1}(\mathbb{R}) = S$. But $g^{-1}(R)$ being an element of $\mathcal{S}$ by the Borel measurabiity of g, it follows that $g_k^{-1}((-\infty, a_k])$ is also an element of $\mathcal{S}$. Therefore, a_k being arbitrary we conclude that $g_k^{-1}(\mathcal{H}_5(\mathbb{R})) \subseteq \mathcal{S}$, hence $g_k : S \to \mathbb{R}$ is Borel measurable by virtue of Lemma 4.5.1 and the fact that $\mathcal{B}(\mathbb{R}) = \sigma\left(\mathcal{H}_5(\mathbb{R})\right)$ by Lemma 4.2.1.

Conversely, assume that the mappings $g_1, \ldots, g_p : S \to \mathbb{R}$ are all Borel mappings. By Lemma 4.4.3, the subset R defined at (4.21) will be a Borel subset of $\mathbb{R}^p$ whenever the sets $B_1, \ldots, B_p$ are all Borel subsets in $\mathbb{R}$. In that case the assumed Borel measurability of the component mappings implies that the sets $g_1^{-1}(B_1), \ldots, g_p^{-1}(B_p)$ all belongs to $\mathcal{S}$, whence $\cap_{\ell=1}^p g_\ell^{-1}(B_\ell)$ also belongs to $\mathcal{S}$ and we conclude that $g^{-1}(R)$ is an element of $\mathcal{S}$. ■

4.6 Extended Borel mappings and limits

Sometimes the notion of a Borel mapping defined in Section 4.5 will fail to cover important situations that arise in applications when the mapping can assume the values $\pm\infty$. Recall that the extended real line is defined as $\overline{\mathbb{R}} = [-\infty, \infty] = \mathbb{R} \cup \{-\infty, +\infty\}$.

Definition 4.6.1

Consider a measurable space $(S, \mathcal{S})$. A mapping $g : S \to \overline{\mathbb{R}}$ is said to be an *extended Borel* mapping if

$$g^{-1}(B) \in \mathcal{S}, \quad B \in \mathcal{B}(\overline{\mathbb{R}})$$

where the extended Borel σ-field $\mathcal{B}(\overline{\mathbb{R}})$ on $\overline{\mathbb{R}}$ is defined as

$$\mathcal{B}(\overline{\mathbb{R}}) \equiv \sigma\left(\mathcal{B}(\mathbb{R}), \{-\infty\}, \{+\infty\}\right). \tag{4.23}$$

The Borel measurability of the mapping $g : S \to \mathcal{B}(\overline{\mathbb{R}})$ according to Definition 4.6.1 is easily checked [Exercise 4.9] to be *equivalent* to the conditions

$$\{s \in S : \ g(s) \in (-\infty, a]\} \in \mathcal{S}, \quad a \in \mathbb{R} \tag{4.24}$$

and

$$S_{\pm\infty} \equiv \{s \in S : \ g(s) = \pm\infty\} \in \mathcal{S}. \tag{4.25}$$

Furthermore, any Borel mapping $g : S \to \mathbb{R}$ is necessarily an extended Borel mapping [Exercise 4.10]

Lemma 4.6.1 *Consider a sequence of extended Borel mappings $\{g_n, \ n = 1, 2, \ldots\}$ which are all defined on the same measurable space $(S, \mathcal{S})$. The following mappings $S \to [-\infty, +\infty]$ derived from the sequence $\{g_n, \ n = 1, 2, \ldots\}$ are all extended Borel mappings:*
(i) The maximum mappings $S \to \overline{\mathbb{R}}$ defined by

$$s \to \max_{m=1,\ldots,n} g_m(s), \qquad \begin{matrix} n = 1, 2, \ldots \\ s \in S \end{matrix}$$

(ii) The minimum mappings $S \to \overline{\mathbb{R}}$ defined by

$$s \to \min_{m=1,\ldots,n} g_m(s), \qquad \begin{matrix} n = 1, 2, \ldots \\ s \in S \end{matrix}$$

(iii) The supremum mapping $S \to \overline{\mathbb{R}}$ defined by

$$s \to \sup_{m \geq 1} g_m(s), \quad s \in S.$$

(iv) The infimum mapping $S \to \overline{\mathbb{R}}$ defined by

$$s \to \inf_{m \geq 1} g_m(s), \quad s \in S.$$

(v) The limsup mapping $S \to \overline{\mathbb{R}}$ defined by

$$s \to \limsup_{n \to \infty} g_n(s), \quad s \in S.$$

(vi) The liminf mapping $S \to \overline{\mathbb{R}}$ defined by

$$s \to \liminf_{n \to \infty} g_n(s), \quad s \in S.$$

Proof. Fix $n = 1, 2, \ldots$. For each a in $\mathbb{R}$, we note that

$$\left\{ s \in S : \max_{m=1,\ldots,n} g_m(s) \leq a \right\} = \cap_{m=1}^{n} \{ s \in S : \; g_m(s) \leq a \} \in \mathcal{S}$$

and

$$\left\{ s \in S : \min_{m=1,\ldots,n} g_m(s) > a \right\} = \cap_{m=1}^{n} \{ s \in S : \; g_m(s) > a \} \in \mathcal{S}$$

since for each $m = 1, 2, \ldots$, the mapping $g_m : S \to \mathbb{R}$ is an extended Borel mapping with $\{ s \in S : \; g_m(s) \leq a \}$ and $\{ s \in S : \; g_m(s) > a \}$ both being subsets in $\mathcal{S}$. The extended Borel measurability of the maximum and minimum mappings follows from Lemma 4.5.1 upon noting that the closed unbounded intervals $\mathcal{H}_5(\mathbb{R}) = \{ (-\infty, a], \; a \in \mathbb{R} \}$ generate the σ-field $\mathcal{B}(\mathbb{R})$.

[Exercise 4.11]. The Borel measurability of the limsup and liminf mappings is now straightforward; the details of the proof are left to the interested reader. ■

It is a simple matter to check [Exercise 4.11] that

$$S^{\star} \equiv \left\{ s \in S : \liminf_{n \to \infty} g_n(s) = \limsup_{n \to \infty} g_n(s) \right\} \in \mathcal{S}$$

and that on the set $S^{\star}$, the limit $\lim_{n \to \infty} g_n(s)$ exists (possibly as an element in $\overline{\mathbb{R}}$), and is the common value assumed by $\liminf_{n \to \infty} g_n$ and $\limsup_{n \to \infty} g_n$.

4.7 Exercises

Ex. 4.1 In the proof of Lemma 4.1.1 show that

a. the collection $g^{-1}(\mathcal{S}_b)$ is a σ-field on S_a if $\mathcal{S}_b$ is a σ-field on S_b.

b. the collection $\mathcal{H}_b$ defined at (4.5) is a σ-field on S_b.

Ex. 4.2 Prove the validity of (4.3) and fill in the details of the proof of Fact 4.1.1 .

Ex. 4.3 Consider the setting of Definition 4.1.1: Let $\mu_a : \mathcal{S}_a \to [0, +\infty]$ a measure defined on $\mathcal{S}_a$, and define the set function $\mu_b : \mathcal{S}_b \to [0, +\infty]$ by

$$\mu_b \left[E_b\right] \equiv \mu_a \left[g^{-1}\left(E_b\right)\right], \quad E_b \in \mathcal{S}_b.$$

Show that the set function $\mu_b : \mathcal{S}_b \to [0, +\infty]$ is a measure defined on $\mathcal{S}_b$. It is a probability measure if μ_a is a probability measure.

Ex. 4.4 Let E_a and E_b be *strict non-empty* subsets of the sets S_a and S_b, respectively. Show that the complement of $E_a \times E_b$ in $S_a \times S_b$ is usually *not* a rectangle with sides in S_a and S_b, i.e., $E_a \times E_b$ cannot be written as $G_a \times G_b$ with G_a and G_b subsets of S_a and S_b.

Use this fact to conclude that if $\mathcal{H}_a$ and $\mathcal{H}_b$ are collections of subsets of S_a and S_b, repectively, then the collection $\mathcal{H}_a \times \mathcal{H}_b$ cannot be a field (σ-field) on $S_a \times S_b$ even if $\mathcal{H}_a$ and $\mathcal{H}_b$ are fields (σ-fields) on S_a and S_b, respectively.

Ex. 4.5 Show that the collections $\mathcal{H}_a^\star$ and $\mathcal{H}_b^\star$ defined at (4.12) and (4.13), respectively, are σ-fields.

Ex. 4.6 Give a proof of Lemma 4.4.1.

Ex. 4.7 Usually the topology on $\mathbb{R}^p$ is characterized in terms of open balls rather than open bounded rectangles – This allows for natural generalizations to general metric spaces. With this in mind, with $r > 0$ and x in $\mathbb{R}^p$, define the *open ball centered at x of radius r* to be the subset $B_r(x)$ given by

$$B_r(x) \equiv \{y \in \mathbb{R}^p : \ \|x - y\| < r\}$$

where $\|z\| = \sqrt{\sum_{i=1}^n z_i^2}$ for every $z = (z_1, \ldots, z_n)$ in $\mathbb{R}^p$. Let $\mathcal{H}_{\text{Open-Ball}}(\mathbb{R}^p)$ denote the collection of all such open balls, namely

$$\mathcal{H}_{\text{Open-Ball}}(\mathbb{R}^p) \equiv \{B_r(x) : \ r > 0, \ x \in \mathbb{R}^p\}.$$

Show that $\mathcal{B}(\mathbb{R}^p) = \sigma\left(\mathcal{H}_{\text{Open-Ball}}(\mathbb{R}^p)\right)$ **[HINT:** Use Fact 4.4.1].

Ex. 4.8 Most (if not all) mappings $\mathbb{R}^p \to \mathbb{R}^q$ encountered in applications are Borel mappings. In particular, any continuous mapping $\mathbb{R}^p \to \mathbb{R}^q$ can be shown to be a Borel mapping! [**HINT:** Use the fact that a mapping $g : \mathbb{R}^p \to \mathbb{R}^q$ is continuous if and only if $g^{-1}\left(\mathcal{O}(\mathbb{R}^q)\right) \subseteq \mathcal{O}(\mathbb{R}^p)$].

Ex. 4.9 Consider the mapping $g : S \to \overline{\mathbb{R}}$ defined on the same measurable space $(S, \mathcal{S})$. Show that this mapping is an extended Borel mapping if and only conditions (4.24) and (4.25) all hold.

Ex. 4.10 Show the validity of the following statements:

a. The collection $\mathcal{B}(\overline{\mathbb{R}})$ of subsets of $\overline{\mathbb{R}}$ defined by (4.23) is a σ-field on $\overline{\mathbb{R}}$.

b. Show that the σ-field $\mathcal{B}(\overline{\mathbb{R}})$ is also generated by the collections $\{[-\infty, a],\ a \in \mathbb{R}\}$ and $\{(a, +\infty),\ a \in \mathbb{R}\}$ of subsets of $\overline{\mathbb{R}}$.

c. Any Borel mapping $g : S \to \mathbb{R}$ is necessarily an (extended) Borel mapping $S \to \overline{\mathbb{R}}$.

Ex. 4.11 Consider the extended Borel mappings $g, h : S \to \overline{\mathbb{R}}$ defined on the same measurable space $(S, \mathcal{S})$.

a. Show that the mapping $S \to \overline{\mathbb{R}} : s \to -g(s)$ is also an extended Borel mapping.

b. Show that the set $S^\star \equiv \{s \in S :\ g(s) = h(s)\}$ belongs to $\mathcal{S}$.

Chapter 5

Constructing (probability) measures: Carathéodory at work

As already mentioned in Chapter 1, determining the appropriate σ-field of events on which to define probability measures is technically more delicate when the sample space is uncountable, say $\mathbb{R}^p$ (or subsets thereof). In such commonly encountered situations, the required σ-additivity of the probability measure precludes the power set of the sample space be used as the σ-field (as we did for the countable case in Section 1.5).

In Chapter 4 we discussed how assigning a measure to certain "natural" subsets of $\mathbb{R}^p$, say intervals in $\mathbb{R}$ or more generally "rectangles" in $\mathbb{R}^p$, leads to the notion of Borel σ-fields. We now explore whether the assignments on these generating families (intimately associated with the usual topology on these sample spaces) can indeed be "extended" to a full measure that is well defined on the generated σ-field. As we shall see shortly the existence of such extensions will be guaranteed with the help of ideas introduced by Carathéodory.

5.1 Facts concerning general measures

In this section we summarize some useful facts concerning general measures. They are analogous to properties that were discussed earlier for probability measures; see Chapter 1. Let $(S, \mathcal{S}, \mu)$ be a measure space as described in Section 1.2 with the understanding that $\mathcal{S}$ is a σ-field on the non-empty set S and the set function $\mu : \mathcal{S} \to [0, +\infty]$ is a measure defined on $\mathcal{S}$.

Proposition 5.1.1 *Consider a finitely additive set function $\mu : \mathcal{S} \to [0, +\infty]$ defined on some σ-field $\mathcal{S}$ carried by the set S; see Definition 1.2.2. The properties* *(i)-(iii) listed below are equivalent where*

(i) The set function $\mu : \mathcal{S} \to [0, +\infty]$ is continuous from below on $\mathcal{S}$: For every monotone increasing collection $\{E_n,\ n = 1, 2, \ldots\}$ of subsets in $\mathcal{S}$ (i.e., $E_n \subseteq E_{n+1}$ for all $n = 1, 2, \ldots$), it holds that $\lim_{n\to+\infty} \mu[E_n] = \mu[\cup_{n=1}^{\infty} E_n]$.

(ii) The set function $\mu : \mathcal{S} \to [0, +\infty]$ is continuous from above on $\mathcal{S}$: For every monotone decreasing collection $\{E_n,\ n = 1, 2, \ldots\}$ of subsets in $\mathcal{S}$ (i.e., $E_{n+1} \subseteq E_n$ for all $n = 1, 2, \ldots$), it holds that $\lim_{n\to+\infty} \mu[E_n] = \mu[\cap_{n=1}^{\infty} E_n]$ provided there exists some index $n^\star$ such that $\mu[E_{n^\star}] < +\infty$.

(iii) The set function $\mu : \mathcal{S} \to [0, +\infty]$ is σ-additive on $\mathcal{S}$: For every disjoint collection $\{E_n,\ n = 1, 2, \ldots\}$ of subsets in $\mathcal{S}$, it holds that

$$\mu[\cup_{n=1}^{\infty} E_n] = \sum_{n=1}^{\infty} \mu[E_n]. \tag{5.1}$$

The need for the finiteness condition in Part (ii) is explored in Exercise 5.1.

Proof. It is easy to see (say by complementarity and finiteness condition) that (i) and (ii) are equivalent, and that (iii) implies (i) and (ii) – The arguments are essentially the same as the ones given in Section 3.1.

Therefore it remains to show that (i) implies (iii): Thus, assume that (i) holds and consider a collection $\{E_n,\ n = 1, 2, \ldots\}$ of disjoint subsets in $\mathcal{S}$. We introduce the monotone increasing collection $\{F_n,\ n = 1, 2, \ldots\}$ of subsets in $\mathcal{S}$ given by

$$F_n \equiv \cup_{m=1}^{n} E_m, \quad n = 1, 2, \ldots$$

The additivity of μ on $\mathcal{S}$ yields

$$\mu[F_n] = \sum_{m=1}^{n} \mu[E_m], \quad n = 1, 2, \ldots \tag{5.2}$$

Let n go to infinity in this last relation: We get

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mu[E_m] = \sum_{m=1}^{\infty} \mu[E_m].$$

On the other end, using (i) with the sequence $\{F_n,\ n = 1, 2, \ldots\}$ we find

$$\lim_{n\to\infty} \mu[F_n] = \mu[\cup_{n=1}^{\infty} F_n] = \mu[\cup_{n=1}^{\infty} E_n]$$

as we note that $\cup_{n=1}^{\infty} F_n = \cup_{n=1}^{\infty} E_n$. This establishes (5.1) and (iii) holds. ■

Lemma 5.1.1 *For any measure* $\mu : \mathcal{S} \to [0, +\infty]$ *defined on* $\mathcal{S}$ *the following properties hold:*

(i) Monotonicity: For any sets E *and* F *in* $\mathcal{S}$*, it holds that*

$$\mu[E] \le \mu[F], \quad E \subseteq F \tag{5.3}$$

(ii) Sub-additivity: For any countable collection $\{E_i,\ i \in I\}$ *in* $\mathcal{S}$*, it holds that*

$$\mu[\cup_{i\in I} E_i] \le \sum_{i\in I} \mu[E_i]. \tag{5.4}$$

The inequality (5.4) is the form the union bound (2.4) takes in this more general context.

Proof. For any subsets E and F in $\mathcal{S}$ the decomposition $E \cup F = E \cup (F - E)$ holds with $E \cap (F - E) = \emptyset$, hence

$$\mu[E \cup F] = \mu[E] + \mu[F - E] \tag{5.5}$$

by the additivity of μ. When $E \subseteq F$, then $E \cup F = F$ and (5.5) becomes

$$\mu[F] = \mu[E] + \mu[F - E],$$

so that (5.3) is an immediate consequence of the fact that $\mu[F - E] \ge 0$. This establishes Claim (i).

We now turn to Claim (ii): For arbitrary subsets E and F in $\mathcal{S}$, we note from (5.5) that $\mu[F - E] \le \mu[F]$ by the monotonicity of μ and the inequality

$$\mu[E \cup F] \le \mu[E] + \mu[F]$$

follows. This establishes Claim (ii) when I has two elements. An easy proof by induction gives the result when I is finite but arbitrary.

Consider now the case where I is countably infinite, say $I = \{1, 2, \ldots\}$ without loss of generality. The union bound for finite collections (which we have just established) already implies

$$\mu[\cup_{k=1}^n E_k] \le \sum_{k=1}^n \mu[E_k], \quad n = 1, 2, \ldots$$

Let n go to infinity in this inequality: We note that the subsets $\{\cup_{k=1}^n E_k,\ n = 1, 2, \ldots\}$ form a sequence of non-decreasing sets. By Claim (i) of Proposition 5.1.1, namely the continuity from below of μ, we conclude that

$$\lim_{n\to\infty} \mu[\cup_{k=1}^n E_k] = \mu[\cup_{k=1}^\infty E_k].$$

On the other hand, $\lim_{n\to+\infty}\sum_{k=1}^{n}\mu\left[E_k\right]=\sum_{k=1}^{\infty}\mu\left[E_k\right]$. Combining these observations completes the proof of Claim (ii). ■

There are important situations where $\mu\left[S\right]=+\infty$ but much of measure theory can still be developed as in the finite case through localization arguments. This arises when the measure $\mu:\mathcal{S}\to[0,+\infty]$ is *σ-finite* in the following sense:

Definition 5.1.1

Given a collection $\mathcal{S}$ of subsets of S, a set function $\mu:\mathcal{S}\to[0,+\infty]$ is *σ-finite* There exists a sequence of sets $\{E_n,\ n=1,2,\ldots\}$ in the σ-field $\mathcal{S}$, said sequence being monotone increasing, i.e., $E_n\subset E_{n+1}$ for all $n=1,2,\ldots$, which "exhausts" S in that $\cup_{n=1}^{\infty}E_n=S$ and for which $\mu\left[E_n\right]<+\infty$ for all $n=1,2,\ldots$.

It might be surprising at first but subsets of a set of measure zero are not necessarily themselves measurable (i.e., members of the σ-field where the measure is defined), in which case they are not of measure zero since no measure can be assigned to them. The next definition formally addresses this issue.

Definition 5.1.2

A measure space $(S,\mathcal{S},\mu)$ is said to be *complete* (or simply that the measure μ is complete on $\mathcal{S}$) if whenever a set E in $\mathcal{S}$ has zero measure under μ, i.e., $\mu\left[E\right]=0$, then every subset E' of E is also in $\mathcal{S}$ (in which case we automatically have $\mu\left[E'\right]=0$).

There are strategies to construct a complete measure space from a measure space $(S,\mathcal{S},\mu)$; one such approach is described in Exercise 5.4.

5.2 The extension problem

Let S denote a non-empty set, and consider two collections $\mathcal{H}_1$ and $\mathcal{H}_2$ of subsets of S.

Definition 5.2.1

With $\mathcal{H}_1\subseteq\mathcal{H}_2$, the set functions $\mu_1:\mathcal{H}_1\to[0,+\infty]$ and $\mu_2:\mathcal{H}_2\to[0,+\infty]$ are said to *agree or coincide on* $\mathcal{H}_1$, written $\mu_1=\mu_2$ on $\mathcal{H}_1$, if

$$\mu_1\left[F\right]=\mu_2\left[F\right],\quad F\in\mathcal{H}_1.$$

It is customary to say that μ_1 is a *restriction* of μ_2 on $\mathcal{H}_1$ and that μ_2 is an *extension* of μ_1 to $\mathcal{H}_2$ if $\mu_1 = \mu_2$ on $\mathcal{H}_1$.

The extension problem can be formulated as follows: Find extensions (in the sense of Definition 5.2.1) of a set function $\mu_1 : \mathcal{H}_1 \to [0, +\infty]$ which preserve some of its properties. Here we are concerned with extending a σ-additive set function into a measure: The set function $\mu_1 : \mathcal{H}_1 \to [0, +\infty]$ is known to be σ-additive on the collection $\mathcal{H}_1$ which has minimal structure and is *not* a σ-field. We seek to extend the set function $\mu_1 : \mathcal{H}_1 \to [0, +\infty]$ into a measure $\mu_2 : \mathcal{H}_2 \to [0, +\infty]$ where $\mathcal{H}_2$ is a σ-field which contains *at the very least* the σ-field $\sigma(\mathcal{H}_1)$ generated by $\mathcal{H}_1$.

The main difficulty in carrying out such extensions lies in the following observation: Even when $\mathcal{H}_2 = \sigma(\mathcal{H}_1)$, no obvious constructive way exists for identifying sets which are in $\mathcal{H}_2$ but not in $\mathcal{H}_1$; only the *existence* of $\sigma(\mathcal{H}_1)$ as the smallest σ-field containing $\mathcal{H}_1$ can be asserted; see Section 1.7. As a result it is hard to imagine how to construct the extension to μ_1 to sets in $\sigma(\mathcal{H}_1)$ that are not in $\mathcal{H}_1$. Instead we shall settle in the main for establishing the *existence* of such extensions. As shown by C. Carathéodory, there is a way out of this difficulty by invoking the notion of outer measure and a related idea of measurability.

5.3 Outer measures

Let S be a non-empty set.

Definition 5.3.1

An *outer measure* on S is a set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ that satisfies the following properties:

(OM1) $\mu^\star[\emptyset] = 0$

(OM2) Monotonicity: If $E \in \mathcal{P}(S)$ and $F \in \mathcal{P}(S)$ such that $E \subseteq F$, then $\mu^\star[E] \leq \mu^\star[F]$.

(OM3) Countable subadditivity: With I a countable index set, if $E_i \in \mathcal{P}(S)$ for each i in I, then

$$\mu^\star[\cup_{i \in I} E_i] \leq \sum_{i \in I} \mu^\star[E_i]. \tag{5.6}$$

The condition (5.6) can be rephrased as saying that the union bound holds for the set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$. Therefore, a measure defined on the entire power set of S is an outer measure since a measure always satisfies union bounds; see Lemma 5.1.1.

Following Carathéodory, with an outer measure $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ we associate the following notion of $\mu^\star$*-measurability*.

Definition 5.3.2

A subset E of S is $\mu^\star$-measurable if

$$\mu^\star[F] = \mu^\star[F \cap E] + \mu^\star[F \cap E^c], \quad F \in \mathcal{P}(S). \tag{5.7}$$

Let $\mathcal{M}(\mu^\star)$ denote the collection of all subsets of S which are $\mu^\star$-measurable.

For any pair of subsets E and F of S, the sets $F \cap E$ and $F \cap E^c$ are disjoint with $F = (F \cap E) \cup (F \cap E)$. Therefore, by (OM3) it is always the case that

$$\mu^\star[F] \leq \mu^\star[F \cap E] + \mu^\star[F \cap E^c], \quad F \in \mathcal{P}(S). \tag{5.8}$$

Thus, to establish the $\mu^\star$-measurability of the subset E it suffices to show the reverse inequality, namely

$$\mu^\star[F] \geq \mu^\star[F \cap E] + \mu^\star[F \cap E^c], \quad F \in \mathcal{P}(S). \tag{5.9}$$

The next result, due to Carathéodory, is remarkable for its generality; in particular it shows how an outer measure always induces a measure on a σ-field!

Theorem 5.3.1 *If the set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ is an outer measure on S, then $\mathcal{M}(\mu^\star)$ is a σ-field on S, and the restriction of $\mu^\star$ to $\mathcal{M}(\mu^\star)$ is a measure.*

While it is plain from (5.7) that the empty set is contained in $\mathcal{M}(\mu^\star)$ and that $\mathcal{M}(\mu^\star)$ is closed under complementation, establishing that $\mathcal{M}(\mu^\star)$ is closed under countable union is more involved. We refer the reader to References [**?**, **?**, **?**, **?**] for a complete proof that the collection $\mathcal{M}(\mu^\star)$ is indeed a σ-field on S and that the restriction of $\mu^\star$ to $\mathcal{M}(\mu^\star)$ is a measure. This measure is *complete* in the sense of Definition 5.1.2.

Lemma 5.3.1 *For any outer measure $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$, the measure space $(S, \mathcal{M}(\mu^\star), \mu^\star)$ is complete.*

Proof. Consider a set E in $\mathcal{M}(\mu^\star)$ with $\mu^\star[E] = 0$, and let E' be any subset of E. For any subset F of S, we note that $F \cap E' \subseteq E' \subseteq E$, hence $\mu^\star[F \cap E'] \leq \mu^\star[E]$ by (OM2). Starting with (5.8) we get

$$\begin{aligned} \mu^\star[F] &\leq \mu^\star\left[F \cap E'\right] + \mu^\star\left[F \cap (E')^c\right] \\ &\leq \mu^\star[E] + \mu^\star\left[F \cap (E')^c\right] \\ &= \mu^\star\left[F \cap (E')^c\right] \\ &\leq \mu^\star[F] \end{aligned} \tag{5.10}$$

where the last inequality is again a consequence of (OM2) since $F \cap (E')^c \subseteq F$. It follows that

$$\mu^\star[F] = \mu^\star\left[F \cap E'\right] + \mu^\star\left[F \cap (E')^c\right], \quad F \in \mathcal{P}(S),$$

and the set E' is therefore $\mu^\star$-measurable. Using (OM2) again we conclude that $\mu^\star[E'] = 0$ since $0 \leq \mu^\star[E'] \leq \mu^\star[E] = 0$. ■

Going back to the extension problem formulated in Section 5.2 we now present a way to exploit Theorem 5.3.1 by associating an outer measure with most set functions.

5.4 Induced outer measures

Outer measures on S are quite easy to find through a construction which associates an outer measure on S with (almost) any set function defined on a collection of subsets of S. The following terminology will simplify the presentation.

Definition 5.4.1

Let $\mathcal{H}$ denote a collection of subsets of S. For any subset E of S, the collection $\{E_i,\ i \in I\}$ of sets in $\mathcal{H}$ is called a *countable covering* of E by sets in $\mathcal{H}$ if I is a countable index set and the covering condition

$$E \subseteq \cup_{i \in I} E_i$$

holds. We refer to such a covering of E as a *countable $\mathcal{H}$-covering* of E.

Let $\mathcal{H}_E$ denote the collection of all countable $\mathcal{H}$-coverings of E; the collection $\mathcal{H}_E$ may be empty for some set E.

Definition 5.4.2

Let $\mathcal{H}$ denote a collection of subsets of S. With any set function $\nu : \mathcal{H} \to [0, +\infty]$ we associate the set function $\mu^\star_\nu : \mathcal{P}(S) \to [0, +\infty]$ given by

$$\mu^\star_\nu [E] \equiv \inf_{\{E_i,\ i \in I\} \in \mathcal{H}_E} \left(\sum_{i \in I} \nu [E_i] \right), \quad E \in \mathcal{P}(S) \tag{5.11}$$

with the usual convention that the infimum in (5.11) is set to $+\infty$ if the collection $\mathcal{H}_E$ is empty.

The properties of set functions defined by (5.11) are presented next.

Theorem 5.4.1 *If the collection $\mathcal{H}$ of subsets of S contains the empty set $\emptyset$ and the set function $\nu : \mathcal{H} \to [0, +\infty]$ has the property that $\nu [\emptyset] = 0$, then the set function $\mu^\star_\nu : \mathcal{P}(S) \to [0, +\infty]$ defined by (5.11) is an outer measure on S known as the outer measure induced by ν.*

Proof. Obviously, since $\mathcal{H}$ contains the empty set $\emptyset$, we have $\mu^\star_\nu [\emptyset] = 0$ under the assumption $\nu [\emptyset] = 0$, hence (OM1) holds.

To show that (OM2) holds, pick subsets E and F in $\mathcal{P}(S)$ such that $E \subseteq F$. A countable $\mathcal{H}$-covering of E is also a countable $\mathcal{H}$-covering of F, hence $\mathcal{H}_F \subseteq \mathcal{H}_E$, and the conclusion $\mu^\star_\nu [E] \leq \mu^\star_\nu [F]$ follows.

As we turn to (OM3), let $\{E_i,\ i \in I\}$ be a countable collection of subsets of S. We need to show that (5.6) holds. If $\mu^\star_\nu [E_i] = +\infty$ for some i in I, then (5.6) automatically holds.

Next consider the situation where $\mu^\star_\nu [E_i] < +\infty$ for all i in I: For each i in I, let the collection $\{F_{i|k},\ k \in K_i\}$ of sets in $\mathcal{H}$ be a countable $\mathcal{H}$-covering of E_i. Obviously $\cup_{i \in I} E_i \subseteq \cup_{i \in I} \left(\cup_{k \in K_i} F_{i|k} \right)$, so that the *countable* collection $\{F_{i|k},\ i \in I,\ k \in K_i\}$ is a countable $\mathcal{H}$-covering of $\cup_{i \in I} E_i$, and the inequality

$$\mu^\star_\nu [\cup_{i \in I} E_i] \leq \sum_{i \in I} \left(\sum_{k \in K_i} \nu \left[F_{i|k} \right] \right)$$

obtains by definition. For each i in I, the finiteness of $\mu^\star_\nu [E_i]$ implies that the collection $\{F_{i|k},\ k \in K_i\}$ can always be selected so that

$$\sum_{k \in K_i} \nu \left[F_{i|k} \right] \leq \mu^\star_\nu [E_i] + \varepsilon \cdot a_i$$

where $\varepsilon > 0$ for some $a_i > 0$. It is always possible to select the scalars $\{a_i,\ i \in I\}$ so that $\sum_{i\in I} a_i < +\infty$. Combining these facts easily leads to

$$\mu^\star_\nu\left[\cup_{i\in I} E_i\right] \leq \sum_{i\in I}\left(\mu^\star_\nu\left[E_i\right] + \varepsilon \dot{a}_i\right) \leq \sum_{i\in I} \mu^\star_\nu\left[E_i\right] + \varepsilon\left(\sum_{i\in I} a_i\right),$$

and $\varepsilon > 0$ being arbitrary we obtain (5.6) upon letting $\varepsilon > 0$ go to zero. ■

5.5 Combining Theorem 5.3.1 and Theorem 5.4.1

With a non-empty set S, let $\mathcal{H}$ denote a collection of subsets of S which contains the empty set $\emptyset$, and let the set function $\nu : \mathcal{H} \to [0, +\infty]$ have the property that $\nu\left[\emptyset\right] = 0$ As before, let $\mu^\star_\nu : \mathcal{P}(S) \to [0, +\infty]$ denote the outer measure on S induced by ν. Applying Theorem 5.3.1 to the outer measure $\mu^\star_\nu$, we conclude that the collection $\mathcal{M}(\mu^\star_\nu)$ is a σ-field on S, and that the restriction of $\mu^\star_\nu$ to $\mathcal{M}(\mu^\star_\nu)$ is a measure – This measure is known as the *Carathéodory measure* induced by ν.

A solution to the problem of extending the σ-additive set function $\nu : \mathcal{H} \to [0, +\infty]$ to a measure $\mu : \sigma\left(\mathcal{H}\right) \to [0, +\infty]$ is now in sight provided the following issues can be resolved:

(i) Does the inclusion

$$\mathcal{H} \subseteq \mathcal{M}(\mu^\star_\nu) \tag{5.12}$$

hold? After all there is no guarantee that the sets in $\mathcal{H}$ are indeed $\mu^\star_\nu$-measurable. Note that (5.12) would imply $\sigma(\mathcal{H}) \subseteq \mathcal{M}(\mu^\star_\nu)$.

(ii) Can we conclude that the conditions

$$\mu^\star_\nu(E) = \nu(E), \quad E \in \mathcal{H} \tag{5.13}$$

hold? This last condition complements (5.12) and expresses the fact that the measure $\mu^\star_\nu$ is indeed an extension of ν.

We first address condition (5.13). Pick E in $\mathcal{H}$: We always have $\mu^\star_\nu(E) \leq \nu(E)$ since E is trivially a countable $\mathcal{H}$-covering of itself. The validity of the reverse inequality $\nu(E) \leq \mu^\star_\nu(E)$ is equivalent to requiring

$$\nu\left[E\right] \leq \sum_{i\in I} \nu\left[E_i\right] \tag{5.14}$$

for *every* countable $\mathcal{H}$-covering $\{E_i,\ i \in I\}$ of E. Although this condition is *reminiscent* of the union bound, a word of caution is in order here: Even if the set function $\nu : \mathcal{H} \to [0, +\infty]$ were σ-additive on $\mathcal{H}$ (as is necessary for an extension to exist), it would satisfy both monotonicity and the union bound property *on* $\mathcal{H}$ (see comment following Lemma 5.1.1). With this in mind it is tempting to argue that the following chain of inequalities

$$\nu [E] \leq \nu [\cup_{i \in I} E_i] \leq \sum_{i \in I} \nu [E_i]$$

holds, the first inequality by virtue of the covering condition $E \subseteq \cup_{i \in I} E_i$ and the second inequality by a union bound argument, in which case (5.14) would automatically hold. However, there is *no* guarantee that $\cup_{i \in I} E_i$ belongs to $\mathcal{H}$ (which typically is not going to be a σ-field), and both steps may fail to be valid. If this were the case then (5.14) would automatically hold – See the proof of Theorem 6.1.1 where $\mathcal{H}$ is a σ-field and the aforementioned issue disappears. Therefore, if condition (5.14) were to hold, additional conditions are needed.

We now turn to condition (5.13). In many (important) applications the collection $\mathcal{H}$ is not even a field of subsets of S. Yet, weaker structural properties can be imposed on the collection of sets $\mathcal{H}$ to ensure the validity of condition (5.13) – One such notion is introduced in the next section.

5.6 Semi-rings, etc

The notion introduced next expands on the approach originally used for defining Lesbegue's measure on $\mathbb{R}^p$.

Definition 5.6.1

The collection $\mathcal{H}$ of subsets of S is said to be a *semi-ring* on H if the conditions (SR1)-(SR3) hold where

(SR1) $\emptyset \in \mathcal{H}$.

(SR2) Closed under intersection: If $E \in \mathcal{H}$ and $F \in \mathcal{H}$, then $E \cap F \in \mathcal{H}$.

(SR3) Relative complements as a finite disjoint union: If $E \in \mathcal{H}$ and $F \in \mathcal{H}$, then there exists a finite collection $F_1, \ldots, F_n$ of disjoint sets in $\mathcal{H}$ such that the representation $E - F = \cup_{i=1}^{n} F_i$ holds.

Again (SR2) implies (is in fact equivalent) to the seemingly more general statement

(SR2b) Closed under finite intersection: For each $n = 1, 2, \ldots$, if $E_1 \in \mathcal{H}, \ldots, E_n \in \mathcal{H}$, then $\cap_{i=1}^n E_i \in \mathcal{H}$.

There is no guarantee that the set $S = \emptyset^c$ is an element of $\mathcal{H}$ since a semi-ring may not be closed under taking complements – Only the weaker requirement (SR3) needs to be satisfied. However, in some literature condition (SR3) is replaced by the following requirement:

(SR3b) Closed under complements: If $E \in \mathcal{H}$, then there exists a finite collection $F_1, \ldots, F_n$ of disjoint sets in $\mathcal{H}$ such that the representation $E^c = \cup_{i=1}^n F_i$ holds.

As discussed below, this condition is stronger than (SR3).

Fact 5.6.1 *If the collection $\mathcal{H}$ of subsets of S satisfies (SR1), (SR2) and (SR3b), then (SR3) holds as well and $\mathcal{H}$ is a semi-ring.*

Proof. We need only show that (SR3) holds under (SR1), (SR2) and (SR3b): Thus, pick sets E and F in $\mathcal{H}$. Under (SR3b), there exists a finite collection $F_1, \ldots, F_n$ of disjoint sets in $\mathcal{H}$ such that the representation $F^c = \cup_{i=1}^n F_i$ holds. Therefore,

$$E - F = E \cap F^c = E \cap \left(\cup_{i=1}^n F_i\right) = \cup_{i=1}^n \left(E \cap F_i\right)$$

and (SR3) holds because the sets $E \cap F_1, \ldots, E \cap F_n$ are in $\mathcal{H}$ under (SR2); they are obviously disjoint since the sets $F_1, \ldots, F_n$ are disjoint under (SR3b). ■

However, under (SR1) and (SR2) the conditions (SR3) and (SR3b) are *equivalent* when S belongs to $\mathcal{S}$: In that case (SR3) with $E = S$ and F arbitrary in $\mathcal{H}$ yield $S - F = S \cap F^c = F^c = \cup_{i=1}^n F_i$ where the sets $F_1, \ldots, F_n$ are disjoint in $\mathcal{H}$. Thus, F^c can be represented as a finite union of disjoint sets in $\mathcal{H}$ and (SR3b) holds.

Even under the stronger condition (SR3b) (applied to $E = \emptyset$) it is not possible in general to ascertain that S belongs to $\mathcal{S}$, but only that $S = \cup_{i=1}^n F_i$ for disjoint sets $F_1, \ldots, F_n$ in $\mathcal{H}$. To smooth out this technical point, some authors augment the notion of semi-ring by requiring that S be included in the semi-ring, in which case such collections are called *semi-fields*.

A key fact associated with semi-rings is given next and already sheds some light as to their usefulness in the Carathédory program outlined in Section 5.5.

Lemma 5.6.1 *If the collection $\mathcal{H}$ of subsets of S is a semi-ring on S, let $\mathcal{H}^\star$ denote the collection of all unions of finite disjoint collections of sets in $\mathcal{H}$. Then, $\mathcal{H}^\star$ is a field of subsets of S containing $\mathcal{H}$. In fact it is the smallest field of subsets of S containing $\mathcal{H}$, i.e., if $\mathcal{G}$ is a field of subsets of S containing $\mathcal{H}$, then $\mathcal{H}^\star \subseteq \mathcal{G}$.*

5.7 Exercises

Ex. 5.1 Consider the following measure space $(S, \mathcal{S}, \mu)$ where $S = \mathbb{N}_0$, $\mathcal{S} = \mathcal{P}(\mathbb{N}_0)$ and the measure $\mu : \mathcal{S}, \to [0, +\infty]$ is the counting measure; see Exercise 1.4. Consider the monotone decreasing sequence of sets $\{E_n,\ n = 1, 2, \ldots\}$ where $E_n \equiv \{k \in \mathbb{N}_0 :\ k \geq n\}$ for each $n = 1, 2, \ldots$. Show that $\lim_{n \to +\infty} \mu\left[E_n\right] = \infty$ while $\mu\left[\cup_{n=1}^{\infty} E_n\right] = 0$. Does this contradict Part (ii) of Proposition 5.1.1?

Ex. 5.2 Show that any measure defined on the entire power set of S is necessarily an outer measure.

Ex. 5.3 Show that a finitely additive outer measure $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ is necessarily a measure on $\mathcal{P}(S)$.

Ex. 5.4 Completing a probability space: Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{N}$ denote the collection of all null events (under $\mathbb{P}$), i.e., $\mathcal{N} \equiv \{N \in \mathcal{F} :\ \mathbb{P}\left[N\right] = 0\}$. Consider now the collection $\mathcal{N}^\star$ of all subsets of Ω that are subsets of $\mathbb{P}$-null events, i.e.,

$$\mathcal{N}^\star \equiv \{M \in \mathcal{P}(\Omega) :\ M \subseteq N \text{ for some } N \in \mathcal{N}\ \} = \cup_{N \in \mathcal{N}} \mathcal{P}(N).$$

Subsets in $\mathcal{N}^\star$ are not necessarily events in $\mathcal{F}$.

Show that the collection $\mathcal{F}^\star \equiv \{E \cup M :\ E \in \mathcal{F},\ M \in \mathcal{N}^\star\}$ is also a σ-field on Ω (which contains $\mathcal{F}$).

b. Define the set function $\mathbb{P}^\star : \mathcal{F}^\star \to [0, 1]$ by

$$\mathbb{P}^\star\left[E^\star\right] \equiv \mathbb{P}\left[E\right], \qquad \begin{array}{c} E^\star = E \cup M \\ E \in \mathcal{F},\ M \in \mathcal{N}^\star. \end{array}$$

Show that this definition is well posed in the following sense: If $E^\star$ admits the two representations $E_1 \cup M_1$ and $E_2 \cup M_2$ with $E_k \in \mathcal{F}$ and $M_k \in \mathcal{N}^\star$, $k = 1, 2$, then $\mathbb{P}\left[E_1\right] = \mathbb{P}\left[E_2\right]$, thereby yielding an unambiguous value for $\mathbb{P}^\star\left[E^\star\right]$ [**HINT:** Make use of the following observation: If for each $k = 1, 2$, $M_k \subseteq N_k$ where N_k is an element of $\mathcal{N}$, then the equality $E_1 \cup M_1 = E_2 \cup M_2$ implies the inclusions $E_1 \subseteq M_2 \cup N_2$ and $E_2 \subseteq M_1 \cup N_1$].

c. Show that the set function $\mathbb{P}^\star : \mathcal{F}^\star \to [0, 1]$ (which is well defined as per Part **b**) is a probability measure on $\mathcal{F}^\star$ which coincides with $\mathbb{P}$ on $\mathcal{F}$.

d. Show that the probability measure $\mathbb{P}^\star$ is complete on $\mathcal{F}^\star$ in the sense that if $\mathbb{P}^\star\left[E^\star\right] = 0$ for some $E^\star$ in $\mathcal{F}^\star$, then for any subset $E^{\star\star}$ of $E^\star$, it holds that $E^{\star\star}$ belongs to $\mathcal{F}^\star$ with $\mathbb{P}^\star\left[E^{\star\star}\right] = 0$.

Ex. 5.5 Consider the set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ defined by

$$\mu^\star(E) \equiv \begin{cases} |E| & \text{if } E \text{ is finite} \\ \\ +\infty & \text{if } E \text{ is not finite.} \end{cases}$$

a. Show that this set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ is an outer measure on S.
b. Determine the $\mu^\star$-measurable sets of S.

Ex. 5.6 Define the set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ by $\mu^\star [\emptyset] = 0$ and $\mu^\star [E] = +\infty$ for $E \neq \emptyset$.
a. Show that this set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ is an outer measure on S.
b. Determine the $\mu^\star$-measurable sets of S.

Ex. 5.7 With S a non-countable set, define the set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ by $\mu^\star [E] = 0$ if E is countable and $\mu^\star [E] = 1$ if E is not countable.
a. Show that this set function $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ is an outer measure on S.
b. Determine the $\mu^\star$-measurable sets of S.

Ex. 5.8 Start with an outer measure $\mu^\star : \mathcal{P}(S) \to [0, +\infty]$ on S. For a given subset G of S, define the set function $\nu_G^\star : \mathcal{P}(S) \to [0, +\infty]$ given by

$$\nu_G^\star [E] \equiv \mu^\star [E \cap G], \quad G \in \mathcal{P}(S).$$

a. Show that this set function $\nu_G^\star : \mathcal{P}(S) \to [0, +\infty]$ is an outer measure on S.
b. Determine the relation between the $\mu^\star$-measurable sets and $\mu_G^\star$-measurable sets.

Ex. 5.9 With $\mathcal{H} \equiv \{\emptyset, S, \{s\},\ s \in S\}$, consider the set function $\nu : \mathcal{H} \to [0, +\infty]$ given by

$$\nu [E] \equiv \begin{cases} 0 & \text{if } E = \emptyset \\ +\infty & \text{if } E = S \\ 1 & \text{if } E \neq \emptyset,\, E \neq S. \end{cases}$$

Describe the outer measure $\mu_\nu^\star : \mathcal{P}(S) \to [0, +\infty]$ induced by ν.

Ex. 5.10 Assume S to be uncountable. With $\mathcal{H} \equiv \{\emptyset, S, \{s\},\ s \in S\}$, consider the set function $\nu : \mathcal{H} \to [0, +\infty]$ given by

$$\nu [E] \equiv \begin{cases} 1 & \text{if } E = S \\ 0 & \text{if } E \neq S. \end{cases}$$

Describe the outer measure $\mu_\nu^\star : \mathcal{P}(S) \to [0, +\infty]$ induced by ν.

Ex. 5.11 Assume that $\mathcal{H}$ is a σ-field on S and the set function $\nu : \mathcal{H} \to [0, +\infty]$ is a measure on $\mathcal{H}$.

a. Show that $\mu^\star$ and ν agree on $\mathcal{H}$.

b. Show that every set in $\mathcal{H}$ is $\mu^\star_\nu$-measurable.

c. These two facts imply that $\mu^\star_\nu$ is an extension of ν on $\mathcal{M}(\mu^\star_\nu)$ which may be larger than the initial σ-field $\mathcal{H}$. Give an example when this will happen!

Ex. 5.12 Given a non-empty family $\{\mathcal{S}_i,\ i \in I\}$ of σ-fields (resp. fields) on some arbitrary set S, show that the collection $\cap_{i\in I}\mathcal{S}_i$ is a σ-field (resp. field) on S.

Chapter 6

Constructing (probability) measures: Extension results and examples

In Chapter 4 we discussed how assigning a measure to certain "natural" subsets of $\mathbb{R}^p$, say intervals in $\mathbb{R}$ or more generally rectangles in $\mathbb{R}^p$, leads to the notion of Borel σ-fields. In Chapter 5 we explored whether the assignments on these generating families (intimately associated with the usual topology on these sample spaces) can indeed be "extended" to a full measure that is well defined on the generated σ-field. The existence of such extensions was guaranteed with the help of ideas introduced by Carathéodory. In the current chapter we present several examples of extension results, apply them on various examples.

6.1 Extension results

In this section we build on the ideas developed in Section 5.5. and present various extension results under progressively weaker conditions; when the proofs are not given they can found in the cited references.

The setting is as in Section 5.5: With a non-empty set S, let $\mathcal{H}$ denote a collection of subsets of S which contains the empty set $\emptyset$, and let the set function $\nu : \mathcal{H} \to [0, +\infty]$ have the property that $\nu\,[\emptyset] = 0$. With $\mu^\star_\nu : \mathcal{P}(S) \to [0, +\infty]$ denoting the outer measure on S induced by ν, Theorem 5.3.1 applied to $\mu^\star_\nu$ yields that the collection $\mathcal{M}(\mu^\star_\nu)$ is a σ-field on S, and that the restriction of $\mu^\star_\nu$ to $\mathcal{M}(\mu^\star_\nu)$ is a measure, the so-called Carathéodory measure induced by ν.

We proceed with a number of extension results under increasingly weaker conditions on the collection $\mathcal{H}$ of subsets of S over which the set function $\nu : \mathcal{H} \to$

$[0,+\infty]$ is originally defined. Obviously, for an extension to a measure to exists it is *necessary* for the set function $\nu : \mathcal{H} \to [0,+\infty]$ to be σ-additive on $\mathcal{H}$.

The first result confirms the fact that the Carathéodory measure induced by *any* measure is indeed an extension of that measure.

Theorem 6.1.1 *Assume that $\mathcal{H}$ is a σ-field of subsets of S and that the set function $\nu : \mathcal{H} \to [0,+\infty]$ is a measure on $\mathcal{H}$. Then, the inclusion*

$$\mathcal{H} \subseteq \mathcal{M}(\mu^\star_\nu) \tag{6.1}$$

holds and the Carathéodory measure $\mu^\star_\nu : \mathcal{M}(\mu^\star_\nu) \to [0,+\infty]$ induced by ν coincides with ν on $\mathcal{H}$, i.e.,

$$\nu(F) = \mu^\star_\nu(F), \quad F \in \mathcal{H}. \tag{6.2}$$

In other words, the measure $\mu^\star_\nu$ extends the measure ν to the larger σ-field $\mathcal{M}(\mu^\star_\nu)$. The proof of Theorem 6.1.1 is available in Section 6.6. Note that $\mu^\star_\nu$ is *complete* on $\mathcal{M}(\mu^\star_\nu)$ while the initial measure ν may *not* have been complete on $\mathcal{H}$, therefore creating the possibility that the inclusion (6.1) is strict.

The next result assumes only that the collection $\mathcal{H}$ is a field on S; a proof is available in the references [**?**][Thm. 11.2, p. 164] and [**?**][Thm. 1.14, p. 31].

Theorem 6.1.2 *Assume that $\mathcal{H}$ is a field of subsets of S and that the set function $\nu : \mathcal{H} \to [0,+\infty]$ is σ-additive on $\mathcal{H}$ with $\nu[\emptyset] = 0$. Then there exists a measure $\nu_{\text{Ext}} : \sigma(\mathcal{H}) \to [0,+\infty]$ which is an extension of the set function $\nu : \mathcal{H} \to [0,+\infty]$ to the smallest σ-field $\sigma(\mathcal{H})$ generated by $\mathcal{H}$: The inclusion*

$$\sigma(\mathcal{H}) \subseteq \mathcal{M}(\mu^\star_\nu) \tag{6.3}$$

holds and

$$\nu_{\text{Ext}}(F) = \mu^\star_\nu(F), \quad F \in \sigma(\mathcal{H}). \tag{6.4}$$

Furthermore, if the set function $\nu : \mathcal{H} \to [0,+\infty]$ is σ-finite on $\mathcal{H}$, then ν_{Ext} is the unique extension of ν to the σ-field $\sigma(\mathcal{H})$ and it is σ-finite on $\sigma(\mathcal{H})$.

In many situations of interest the collection $\mathcal{H}$ is not even a field of subsets of S. Instead the collection $\mathcal{H}$ has a weaker structure; in particular there are important applications where $\mathcal{H}$ is only a semi-ring on S as defined in Section 5.6. [**?**][Thm. 11.3, p. 164]. A proof is given in Section 6.7.

Theorem 6.1.3 *Assume that $\mathcal{H}$ is a semi-ring of subsets of S and that the set function $\nu : \mathcal{H} \to [0, +\infty]$ is both finitely additive and countably subadditive on $\mathcal{H}$ with $\nu[\emptyset] = 0$. Then there exists a measure $\nu_{\mathrm{Ext}} : \sigma(\mathcal{H}) \to [0, +\infty]$ which is an extension of the set function $\nu : \mathcal{H} \to [0, +\infty]$ to the smallest σ-field $\sigma(\mathcal{H})$ generated by $\mathcal{H}$: The inclusion*

$$\sigma(\mathcal{H}) \subseteq \mathcal{M}(\mu^\star_\nu) \tag{6.5}$$

holds and

$$\nu_{\mathrm{Ext}}(F) = \mu^\star_\nu(F), \quad F \in \sigma(\mathcal{H}). \tag{6.6}$$

Furthermore, if the set function $\nu : \mathcal{H} \to [0, +\infty]$ is σ-finite on $\mathcal{H}$, then ν_{Ext} is the unique extension of ν to the σ-field $\sigma(\mathcal{H})$ and it is σ-finite on $\sigma(\mathcal{H})$.

6.2 Example 1 – Infinite coin tossings and its generalization

Consider a random experiment $\mathcal{E}$ modeled by the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ where the sample space Ω contains a finite number N of distinct elements, say $s_1, \ldots, s_N$, the σ-field $\mathcal{F}$ is the power set $\mathcal{P}(\Omega)$, and the probability measure $\mathbb{P}$ is described through the probability mass function $\boldsymbol{p} = (p(s_1, \ldots, p(s_N))$. See Section 1.5.

This experiment is repeated infinitely many times under "identical and independent conditions." Let the resulting random experiment be denoted by $\mathcal{E}_\infty$. We now explore how to build an appropriate probability triple $(\Omega_\infty, \mathcal{F}_\infty, \mathbb{P}_\infty)$ to model experiment $\mathcal{E}_\infty$.

The sample space For ease of notation, let $\Omega_1, \ldots, \Omega_k, \ldots$ denote identical copies of Ω. It is appropriate to take Ω_∞ to be the Cartesian product

$$\Omega_\infty \equiv \times_{k=1}^\infty \Omega_k = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n \times \ldots$$

(sometimes also denoted $\Omega^{\mathbb{N}_0}$). The set Ω_∞ is the collection of all infinite length words drawn from the alphabet $\{s_1, \ldots, s_N\}$, and its generic element ω_∞ is of the form

$$\omega_\infty = (\omega_1, \omega_2, \ldots, \omega_n, \ldots), \qquad \begin{matrix} \omega_k \in \Omega = \{s_1, \ldots, s_N\} \\ k = 1, 2, \ldots \end{matrix}$$

The σ-field of events To construct $\mathcal{F}_\infty$ we proceed as follows: For each $n = 1, 2, \ldots$ let $\mathcal{F}_n$ denote the collection of subsets of Ω_∞ defined by

$$\mathcal{F}_n \equiv \{B_{1,2,\ldots,n} \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots : B_{1,2,\ldots,n} \in \mathcal{P}(\Omega_1 \times \ldots \Omega_n)\}.$$

It is a simple matter to check that $\mathcal{F}_n$ is a σ-field on Ω_∞. Let $\mathcal{F}^\star_\infty$ denote the collection of subsets of Ω_∞ given by

$$\mathcal{F}^\star_\infty \equiv \cup_{n=1}^\infty \mathcal{F}_n .$$

While the collection $\mathcal{F}^\star_\infty$ is a field on Ω_∞, it is not a σ-field on Ω_∞ [Exercise 6.1]. This leaves us no choice but to introduce the smallest σ-field containing $\mathcal{F}^\star_\infty$, namely

$$\mathcal{F}_\infty \equiv \sigma\left(\mathcal{F}^\star_\infty\right).$$

The probability measure To define $\mathbb{P}_\infty$ on $\mathcal{F}_\infty$, we first define it on the field $\mathcal{F}^\star_\infty$ and then invoke Theorem 6.1.2 to ensure its extension on $\mathcal{F}_\infty$.

Pick E in $\mathcal{F}^\star_\infty$. As this set must be in $\mathcal{F}_n$ for some $n = 1, 2, \ldots$, it is therefore of the form

$$E = B_{1,2,\ldots,n} \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots \tag{6.7}$$

for some set $B_{1,2,\ldots,n}$ in $\mathcal{P}(\Omega_1 \times \ldots \Omega_n)$. In particular, take $B_{1,2,\ldots,n} = \{(\omega_1, \ldots, \omega_n)\}$ with $\omega_1, \ldots, \omega_n$ arbitrary in Ω, in which case it is appropriate to take

$$\mathbb{P}_\infty\left[\{(\omega_1, \ldots, \omega_n)\} \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots\right] = p(\omega_1) \ldots p(\omega_n) = \prod_{k=1}^n p(\omega_k)$$

to reflect the fact that the experiment is repeated under "identical and independent conditions" as in the coin tossing experiment of Section 2.3. This a modeling assumption reflecting the conditions under which the experiment is carried out! Using the additivity of $\mathbb{P}_\infty$ on $\mathcal{F}_n$, it is now straightforward to evaluate the probability of the sets (6.7), namely

$$\mathbb{P}_\infty\left[E\right] = \sum_{(\omega_1,\ldots,\omega_n) \in B_{1,2,\ldots,n}} p(\omega_1) \ldots p(\omega_n).$$

It is easy to check that $\mathbb{P}_\infty$ is well defined on $\mathcal{F}^\star_\infty$ in the following sense: If the event E in $\mathcal{F}_n$ were viewed as an event in $\mathcal{F}_m$ for some $m > n$, then (6.7) become

$$E = B_{1,2,\ldots,m} \times \Omega_{m+1} \times \Omega_{m+2} \times \ldots \tag{6.8}$$

for some set $B_{1,2,\ldots,m}$ in $\mathcal{P}(\Omega_1 \times \ldots \Omega_m)$ related to $B_{1,2,\ldots,n}$ through

$$B_{1,2,\ldots,m} = B_{1,2,\ldots,n} \times \Omega_{n+1} \times \ldots \times \Omega_m .$$

It is then elementary to check that

$$\sum_{(\omega_1,\ldots,\omega_m) \in B_{1,2,\ldots,m}} p(\omega_1) \ldots p(\omega_m) = \sum_{(\omega_1,\ldots,\omega_n) \in B_{1,2,\ldots,n}} p(\omega_1) \ldots p(\omega_n) \tag{6.9}$$

since

$$\sum_{\omega_k \in \Omega_k} p(\omega_k) = 1, \quad k = n+1, \ldots, m.$$

With $B_1, \ldots, B_n$ subsets of Ω, we easily check that the events $\tilde{E}_1, \ldots, \tilde{E}_n$ given by

$$\tilde{E}_k \equiv \Omega_1 \times \ldots \Omega_{k-1} \times B_k \times \Omega_{k+1} \times \ldots \times \Omega_n \times \ldots, \quad k = 1, \ldots, n$$

are mutually independent under $\mathbb{P}_\infty$!

Example 1 – Tossing a coin infinitely often When $\Omega = \{0, 1\}$, the model discussed here captures the situation when a coin is tossed infinitely often under identical and independent conditions – See the discussion in Section 2.3 where the coin was tossed a finite number of times. In that setting, $\Omega_\infty = \{0, 1\}^{\mathbb{N}_0}$ and with p (resp. $1 - p$) denoting the probability that the outcome of an individual toss is Head (= 1) (resp. Tail (= 0)), the previous calculations become

$$\begin{aligned} &\mathbb{P}_\infty \left[\{(\omega_1, \ldots, \omega_n)\} \times \{0, 1\} \times \{0, 1\} \times \ldots \right] \\ &= \ p^{\sum_{k=1}^n \omega_k} (1 - p)^{\sum_{k=1}^n (1 - \omega_k)} \end{aligned}$$

for every $n = 1, 2, \ldots$ and every $(\omega_1, \ldots, \omega_n)$ in $\{0, 1\}^n$. As expected we recover the model presented in Section 2.3.

6.3 Example 2 – Borel measures on $\mathcal{B}(\mathbb{R})$

The Borel σ-field $\mathcal{B}(\mathbb{R})$ supports an important class of measures – The terminology will not surprise you!

Definition 6.3.1

A measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$ is called a Borel measure.

With any Borel measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$, we associate the mapping $F_\mu : \mathbb{R} \to \mathbb{R}$ given by

$$F_\mu(x) \equiv \mu \left[(-\infty, x] \right], \quad x \in \mathbb{R}. \tag{6.10}$$

The mapping $F_\mu : \mathbb{R} \to \mathbb{R}$ is monotone increasing and right-continuous [Exercise 6.2], and the additivity of μ yields

$$F_\mu(b) = F_\mu(a) + \mu \left[(a, b] \right], \quad \begin{array}{c} a < b \\ a, b \in \mathbb{R}. \end{array}$$

Furthermore, if the Borel measure μ is finite (i.e., $\mu[\mathbb{R}] < \infty$, then

$$\mu[(a,b]] = F_\mu(b) - F_\mu(a), \qquad \begin{matrix} a < b \\ a,b \in \mathbb{R}. \end{matrix} \tag{6.11}$$

It is natural to wonder whether this process can be reversed: Consider a mapping $F : \mathbb{R} \to \mathbb{R}$ which is monotone increasing and right-continuous. Monotonicity guarantees the existence of left limits at every point, i.e., $\lim_{y\uparrow x} F(x) = F(x-)$ for every x in $\mathbb{R}$, and the limits $\lim_{x\to -\infty} F(x) \equiv F(-\infty)$ and $\lim_{x\to +\infty} F(x) \equiv F(+\infty)$ are both well defined, possibly infinite. Does there exist a measure $\mu_F : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$ such that

$$\mu_F[(a,b]] = F(b) - F(a), \qquad \begin{matrix} a < b \\ a,b \in \mathbb{R} \end{matrix}$$

as we take our cue from (6.11). The case where $F(x) = x$ for all x in $\mathbb{R}$ would correspond to the usual "length" measure.

To answer this question, consider the collection $\mathcal{H}$ of subsets of $\mathbb{R}$ given by

$$\mathcal{H} \equiv \mathcal{H}_2(\mathbb{R}) \cup \mathcal{H}_5(\mathbb{R}) \cup \mathcal{H}_6(\mathbb{R})$$

where the collections $\mathcal{H}_2(\mathbb{R})$, $\mathcal{H}_5(\mathbb{R})$ and $\mathcal{H}_6(\mathbb{R})$ were introduced in Section 4.2 – Thus, subsets in $\mathcal{H}$ are either of the form $(a,b]$ or $(-\infty, b]$ or $(a, +\infty)$ with $-\infty \leq a < b < +\infty$. It is easy to check that $\mathcal{H}$ is a semi-ring [Exercise 6.3].

With the mapping $F : \mathbb{R} \to \mathbb{R}$ we associate the set function $\nu_F : \mathcal{H} \to [0, +\infty]$ given by

$$\nu_F[(a,b]] \equiv F(b) - F(a), \quad -\infty \leq a \leq b < +\infty$$

and

$$\nu_F[(a,\infty)] \equiv F(+\infty) - F(a), \quad -\infty \leq a$$

6.4 Example 3 – Product measures

6.5 Examples

Examples of semi-rings include

On $\mathbb{R}$:

$$\mathcal{H}_2(\mathbb{R}) = \left\{ (a,b], \begin{matrix} a \leq b \\ a,b \in \mathbb{R} \end{matrix} \right\}$$

and

$$\mathcal{H}_2^\star(\mathbb{R}) = \mathcal{H}_2(\mathbb{R}) \cup \{(-+\infty, b],\ b \in \mathbb{R}\} \cup \{(a, +\infty),\ a \in \mathbb{R}\}$$

On $\mathbb{R}^p$:

$$\mathcal{H}_2(\mathbb{R}) \times \ldots \times \mathcal{H}_2(\mathbb{R})$$

and

$$\mathcal{H}_2^\star(\mathbb{R}) \times \ldots \times \mathcal{H}_2^\star(\mathbb{R})$$

On $\{0,1\}^{\mathbb{N}_0}$

6.6 A proof of Theorem 6.1.1

Start with the definition of the set function $\mu_\nu^\star : \mathcal{P}(S) \to [0,+\infty]$: It is given by

$$\mu_\nu^\star[E] \equiv \inf_{\{E_i,\ i\in I\}\in\mathcal{H}_E} \left(\sum_{i\in I} \nu[E_i] \right), \quad E \in \mathcal{P}(S) \tag{6.12}$$

where $\mathcal{H}_E$ denotes the collection of all countable $\mathcal{H}$-coverings of E.

Pick E in $\mathcal{H}$. Obviously, E is a countable $\mathcal{H}$-covering of itself, hence $\mu_\nu^\star[E] \leq \nu[E]$. To prove that $\nu[E] \leq \mu_\nu^\star[E]$ we proceed as follows: Let $\{E_i,\ i \in I\}$ be a countable $\mathcal{H}$-covering of E. Since $E \subseteq \cup_{i\in I} E_i$, we obviously have $E = (\cup_{i\in I} E_i) \cap E = \cup_{i\in I}(E_i \cap E)$. But ν is a measure on the σ-field $\mathcal{H}$, hence

$$\nu[E] = \nu[\cup_{i\in I}(E_i \cap E)] \leq \sum_{i\in I} \nu[E_i \cap E]$$

by the union bound, while

$$\sum_{i\in I} \nu[E_i \cap E] \leq \sum_{i\in I} \nu[E_i]$$

by the monotonicity of the measure ν on $\mathcal{H}$. Combining these inequalities we conclude that

$$\nu[E] \leq \sum_{i\in I} \nu[E_i].$$

and the conclusion $\nu[E] \leq \mu_\nu^\star[E]$ follows since this inequality is valid for *any* countable $\mathcal{H}$-covering of E.

In order to show the inclusion $\mathcal{H} \subseteq \mathcal{M}(\mu_\nu^\star)$, we need to show that any subset E in $\mathcal{H}$ is $\mu_\nu^\star$-measurable, namely that

$$\mu_\nu^\star[F] = \mu_\nu^\star[F \cap E] + \mu_\nu^\star[F \cap E^c], \quad F \in \mathcal{P}(S).$$

As per the discussion following Definition 5.3.2, it suffices to show that

$$\mu_\nu^\star[F] \geq \mu_\nu^\star[F \cap E] + \mu_\nu^\star[F \cap E^c], \quad F \in \mathcal{P}(S). \tag{6.13}$$

Pick F in $\mathcal{P}(S)$. By definition, with $\mathcal{H}_F$ denoting the collection of all countable $\mathcal{H}$-coverings of F, we get

$$\begin{aligned} \mu^\star_\nu[F] &= \inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i]\right) \\ &= \inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\left(\nu[F_i\cap E]+\nu[F_i\cap E^c]\right)\right) \\ &= \inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i\cap E]+\sum_{i\in I}\nu[F_i\cap E^c]\right) \qquad (6.14)\end{aligned}$$

where the following observation was used: For each i in I, F_i is in $\mathcal{H}$ and so is E, therefore $F_i\cap E$ and $F_i\cap E^c$ are disjoint sets which both are in the σ-field $\mathcal{H}$ with $F_i=(F_i\cap E)\cup(F_i\cap E^c)$. It then follows by additivity that $\nu[F_i]=\nu[F_i\cap E]+\nu[F_i\cap E^c]$.

Next, standard properties of the infimum operation yield

$$\begin{aligned} \mu^\star_\nu[F] &\geq \inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i\cap E]\right) \\ &\qquad + \inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i\cap E^c]\right). \qquad (6.15)\end{aligned}$$

If $\{F_i,\ i\in I\}$ is a countable $\mathcal{H}$-covering of F, then because E is in the σ-field $\mathcal{H}$, it is plain that $\{F_i\cap E,\ i\in I\}$ is a countable $\mathcal{H}$-covering of $F\cap E$, hence

$$\inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i\cap E]\right)\geq\mu^\star_\nu[F\cap E]. \qquad (6.16)$$

A similar argument, with E^c replacing E, leads to

$$\inf_{\{F_i,\, i\in I\}\in\mathcal{H}_F}\left(\sum_{i\in I}\nu[F_i\cap E^c]\right)\geq\mu^\star_\nu[F\cap E^c] \qquad (6.17)$$

Combining the inequalities (6.15), (6.16) and (6.17) we conclude that (6.13) holds.
■

6.7 A proof of Theorem 6.1.3

A natural definition We start by defining the set function $\mu : \mathcal{H}^\star \to [0, +\infty]$: Pick any set E in the field $\mathcal{H}^\star$. By Lemma 5.6.1 membership of E in the field $\mathcal{H}^\star$ implies that it can be represented as a finite union of disjoint sets $\{E_i,\ i \in I\}$ in $\mathcal{H}$, i.e., $E = \cup_{i\in I} E_i$. Additivity suggests that we define

$$\mu[E] \equiv \sum_{i\in I} \nu[E_i]\,. \tag{6.18}$$

A well-posed definition This definition is well posed and independent of the representation used for E: Indeed, let E admit another representation as a finite union of disjoint sets $\{G_j,\ j \in J\}$ in $\mathcal{H}$ such that $E = \cup_{j\in J} G_j$. Note the obvious set equalities

$$E_i = E \cap E_i = \cup_{j\in J}\left(E_i \cap G_j\right), \quad i \in I$$

where for each i in I and j in J the intersections $E_i \cap G_j$ are elements of $\mathcal{H}$. Therefore,

$$\begin{aligned}
\sum_{i\in I} \nu[E_i] &= \sum_{i\in I} \nu[E \cap E_i] \\
&= \sum_{i\in I}\left(\sum_{j\in J} \nu[E_i \cap G_j]\right) \\
&= \sum_{j\in J}\left(\sum_{i\in I} \nu[E_i \cap G_j]\right) \\
&= \sum_{j\in J} \nu[G_j] \qquad (6.19)
\end{aligned}$$

as we repeatedly use the assumed additivity of ν on $\mathcal{H}$.

Uniqueness Let $\tilde{\mu} : \mathcal{H}^\star \to [0, +\infty]$ denote another extension of ν which is additive on $\mathcal{H}$. If E is any element in $\mathcal{H}^\star$, then there exists $\{E_i,\ i \in I\}$ disjoint sets in $\mathcal{H}$ such that $E = \cup_{i\in I} E_i$. By the additivity of $\tilde{\mu}$ on $\mathcal{H}^\star$, hence on $\mathcal{H}$, we get

$$\tilde{\mu}[E] = \sum_{i\in I} \tilde{\mu}[E_i] = \sum_{i\in I} \nu[E_i] = \mu[E] \tag{6.20}$$

where the second equality follows from the fact that $\tilde{\mu}$ is an extensions of ν. This shows that the extensions μ and $\tilde{\mu}$ coincide on $\mathcal{H}$.

Additivity From the definition (6.18) it is easy to see that the set function $\mu : \mathcal{H}^\star \to [0, +\infty]$ is additive on the field $\mathcal{H}^\star$.

Next, if we assume that the set function $\nu : \mathcal{H} \to [0, +\infty]$ is σ-additive on $\mathcal{H}$, we now show that the set function $\mu : \mathcal{H}^\star \to [0, +\infty]$ is itself σ-additive on $\mathcal{H}^\star$: Let $\{E_i,\ i \in I\}$ denote a *countable* collection of disjoint elements of $\mathcal{H}^\star$, and assume that $E = \cup_{i \in I} E_i$ is itself an element of $\mathcal{H}^\star$. We need to show then that

$$\mu[E] = \sum_{i \in I} \mu[E_i].$$

Since E is an element of $\mathcal{H}^\star$, it admits a representation of the form $E = \cup_{j \in J} F_j$ for some *finite* union of disjoint sets $\{F_j,\ j \in J\}$ in $\mathcal{H}$, i.e., $E = \cup_{j \in J} F_j$. Note that

$$E_i = E_i \cap E = E_i \cap (\cup_{j \in J} F_j) = \cup_{j \in J} (E_i \cap F_j), \quad i \in I \tag{6.21}$$

with the sets $\{E_i \cap F_j,\ i \in I,\ j \in J\}$ *all* being disjoint sets in $\mathcal{H}$. Using these facts we get

$$\begin{aligned} \sum_{i \in I} \mu[E_i] &= \sum_{i \in I} \mu[E_i \cap E] \\ &= \sum_{i \in I} \left(\sum_{j \in J} \mu[E_i \cap F_j] \right) \\ &= \sum_{j \in J} \left(\sum_{i \in I} \mu[E_i \cap F_j] \right) \\ &= \sum_{j \in J} \mu[F_j] \\ &= \mu[E]. \end{aligned} \tag{6.22}$$

For each j in J, the equality $\sum_{i \in I} \mu[E_i \cap F_j] = \mu[F_j]$ is a consequence of the σ-additivity of ν on $\mathcal{H}$. This completes the proof of Theorem 6.1.3. ■

6.8 Exercises

Ex. 6.1 In the discussion of Section 6.2 show the following facts:

a. For each $n = 1, 2, \ldots$, the collection $\mathcal{F}_n$ is a σ-field on Ω_∞.

b. The collection $\mathcal{F}^\star_\infty \equiv \cup_{n=1}^\infty \mathcal{F}_n$ is a field on Ω_∞.

c. Although the collection $\mathcal{F}_\infty^\star$ is a field on Ω_∞, it is *not* a σ-field on Ω_∞ [**HINT:** Take the countable collection $\{E_n,\ n = 1, 2, \ldots\}$ given by

$$E_n \equiv \{\omega\} \times \ldots \times \{\omega\} \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots, \quad n = 1, 2, \ldots$$

for some ω in Ω. For each $n = 1, 2, \ldots$, the set E_n belongs to $\mathcal{F}_n$, hence to $\mathcal{F}_\infty$. Identify the set $\cap_{n=1}^\infty E_n$ and determine whether it belongs to $\mathcal{F}_\infty^\star$.]

Ex. 6.2 Consider the mapping $F_\mu : \mathbb{R} \to \mathbb{R}$ associated via (6.10) with a Borel measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$. Show that this mapping is monotone increasing and right-continuous with left limits.

Ex. 6.3 With the collections $\mathcal{H}_2(\mathbb{R})$, $\mathcal{H}_5(\mathbb{R})$ and $\mathcal{H}_6(\mathbb{R})$ introduced in Section 4.2, check that $\mathcal{H} \equiv \mathcal{H}_2(\mathbb{R}) \cup \mathcal{H}_5(\mathbb{R}) \cup \mathcal{H}_6(\mathbb{R})$ is a semi-ring.

Chapter 7

Random variables and their distributions

So far we have been concerned with modeling the full random experiment $\mathcal{E}$, and this has led us to introduce the notion of a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. However, in many settings there is interest not in the full model itself but rather in various numerical characteristics associated with the experiment. This is formalized through the notion of *random variable* (rv) and of its *probability distribution*, notions which we introduce next and which we study in some generality in Chapter 7. The discussion will be specialized to discrete rvs in Chapter 8 and to (absolutely) continuous rvs in Chapter 9.

7.1 Random variables

Throughout we assume given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ which is held fixed during the discussion. Also let p be an arbitrary positive integer.

Definition 7.1.1

A mapping $X : \Omega \to \mathbb{R}^p$ is a *random variable* (rv) defined on $(\Omega, \mathcal{F})$ if the conditions

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad B \in \mathcal{B}(\mathbb{R}^p) \tag{7.1}$$

all hold.

In other words, the mapping $X : \Omega \to \mathbb{R}^p$ is a rv if it is a Borel mapping $X : \Omega \to \mathbb{R}^p$ in the sense of Definition 4.5.1 with $S = \Omega$ (the sample space) and $\mathcal{S} = \mathcal{F}$

(the σ-field of events). It is customary to write $[X \in B]$ in lieu of $X^{-1}(B)$ and $\mathbb{P}[X \in B]$ for $\mathbb{P}[[X \in B]]$.

As in Section 4.5, the rv $X : \Omega \to \mathbb{R}^p$ can also be viewed as a p-tuple of mappings $X_1, \ldots, X_p : \Omega \to \mathbb{R}$ where for each $k = 1, \ldots, p$, the mapping $X_k : \Omega \to \mathbb{R}$ picks up the k^{th} coordinate of X so that

$$X(\omega) = (X_1(\omega), \ldots, X_p(\omega)), \quad \omega \in \Omega.$$

Translating Lemma 4.5.2 to the setting of Definition 7.1.1 we conclude that the mapping $X : \Omega \to \mathbb{R}^p$ is then a rv if and only if each of the component mappings $X_1 : \Omega \to \mathbb{R}, \ldots, X_p : \Omega \to \mathbb{R}$ is a rv. This requires that the conditions

$$\{\omega \in \Omega : \ X_k(\omega) \le x_k, \ k = 1, \ldots, p\} \in \mathcal{F}, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p \tag{7.2}$$

all be satisfied. Sometimes it is convenient to rewrite them in equivalent form as either

$$\cap_{k=1}^p [X_k \le x_k] \in \mathcal{F}, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p. \tag{7.3}$$

or

$$[X \in R(x)] \in \mathcal{F}, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p \tag{7.4}$$

where

$$R(x) \equiv (-\infty, x_1] \times \ldots \times (-\infty, x_p].$$

7.2 Probability distribution functions

Consider an $\mathbb{R}^p$-valued rv $X : \Omega \to \mathbb{R}^p$ as given in Definition 7.1.1. Thus far, this is a *deterministic* object. We attach to it *probabilistic* content by taking into account the probability measure $\mathbb{P}$ under which the likelihood of events for the experiment $\mathcal{E}$ is evaluated.

Definition 7.2.1

The probability distribution (function) of the rv X (under $\mathbb{P}$) is the mapping $F_X : \mathbb{R}^p \to [0, 1]$ defined by

$$\begin{aligned} F_X(x) &\equiv \mathbb{P}[X \in (-\infty, x_1] \times \ldots \times (-\infty, x_p]] \\ &= \mathbb{P}[X_1 \le x_1, \ldots, X_p \le x_p], \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p. \end{aligned} \tag{7.5}$$

with the notation $X = (X_1, \ldots, X_p)$.

This definition is well posed in view of the equivalent conditions (7.2)-(7.4). Note that the the values

$$\{\mathbb{P}\left[X \in B\right],\ B \in \mathcal{B}(\mathbb{R}^p)\} \tag{7.6}$$

constitute the entire probabilistic information concerning the rv X (under $\mathbb{P}$). However it turns out that there is *as much* probabilistic information in the probability distribution $F_X : \mathbb{R}^p \to [0,1]$ as in the entire collection (7.6). This is quite fortunate since the probability distribution is a mapping $\mathbb{R}^p \to [0,1]$ whereas (7.6) describes a *set* function $\mathbb{B}(\mathbb{R}^p) \to [0,1]$.

In fact, knowledge of $F_X : \mathbb{R}^p \to \mathbb{R}$ *uniquely* determine the values (7.6): Indeed, it is a simple matter to check that the set function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \to [0,1]$ defined by

$$\mathbb{P}_X\left[B\right] \equiv \mathbb{P}\left[X \in B\right], \quad B \in \mathcal{B}(\mathbb{R}^p) \tag{7.7}$$

is a probability measure on $\mathcal{B}(\mathbb{R}^p)$ [Exercise 7.1]. The probability distribution $F_X : \mathbb{R}^p \to \mathbb{R}$ thus specifies $\mathbb{P}_X$ on the collection of all semi-infinite boxes

$$\left\{ \prod_{k=1}^{p} (-\infty, x_k]\,, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p \right\}.$$

and therefore can be uniquely extended to $\mathcal{B}(\mathbb{R}^p)$ as a consequence of Carathéodory's Theorem **??**.

The fact that the set function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \to [0,1]$ defined by (7.7) is a probability measure on $\mathcal{B}(\mathbb{R}^p)$ suggests the following useful interpretation: The probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ was selected as a model for the underlying random experiment $\mathcal{E}$. The rv $X : \Omega \to \mathbb{R}^p$ can be viewed as itself inducing a random experiment, denoted $\mathcal{E}_X$, whose elementary outcomes are the values $\{X(\omega), \omega \in \Omega\}$ – After all, if the outcome ω in $\mathcal{E}$ can only be known if the experiment $\mathcal{E}$ is realized, then outcome $X(\omega)$ of the experiment $\mathcal{E}_X$ will be known only after ω has been observed and the numerical value $X(\omega)$ evaluated.

With this in mind it is natural to think of the triple $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)$ as a natural probability model associated with the random experiment $\mathcal{E}_X$. If there is interest only in this associated experiment (and not in the underlying experiment $\mathcal{E}$), we need only focus on the triple $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)$ since the probability measure $\mathbb{P}_X$ carries *all* the probabilistic information related to it. Working with $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)$ instead of with $(\Omega, \mathcal{F}, \mathbb{P})$ often affords an advantageous model reduction. Furthermore, the equivalence between $\mathbb{P}_X$ and F_X means that for many purposes it will suffice to learn about the properties of the probability distribution function F_X.

In what follows, unless stated otherwise all rvs are defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

7.3 Marginalization

The following situation arises in many settings: Consider k distinct rvs $X_1 : \Omega \to \mathbb{R}^{p_1}, \ldots, X_k : \Omega \to \mathbb{R}^{p_k}$ with $p_1, \ldots, p_k$ positive integers. We can alternatively view this collection of rvs as a single rv $X : \Omega \to \mathbb{R}^p$ given by

$$X = (X_1, \ldots, X_k)$$

with $p = p_1 + \ldots + p_k$ – In this notation we implicitly assume that all vectors are row vectors. As usual we have

$$\mathbb{P}\left[X \in B\right] = \mathbb{P}\left[X_1 \in B_1, \ldots, X_k \in B_k\right], \quad \begin{array}{c} B_\ell \in \mathcal{B}(\mathbb{R}^{p_\ell}) \\ \ell = 1, \ldots, k \end{array}$$

when taking rectangles of the form

$$B = B_1 \times \ldots \times B_k.$$

In particular, by taking $B_\ell = (-\infty, x_\ell]$ with arbitrary x_ℓ in $\mathbb{R}^{p_\ell}$ for each $\ell = 1, \ldots, k$, we conclude that the probability distribution function of the rv $X : \Omega \to \mathbb{R}^p$ (or equivalently, the joint probability distribution function of the rvs $X_1, \ldots, X_k$) is given by

$$\begin{aligned} & F_{(X_1,\ldots,X_k)}(x_1, \ldots, x_k) \\ = \ & \mathbb{P}\left[X_1 \le x_1, \ldots, X_k \le x_k\right], \quad \begin{array}{c} x_\ell \in \mathbb{R}^{p_\ell} \\ \ell = 1, \ldots, k. \end{array} \end{aligned} \tag{7.8}$$

Now, pick any subset $J \subseteq \{1, \ldots, k\}$ with $1 \le |J|$ and note that

$$[X_1 \le x_1, \ldots, X_k \le x_k] = \left(\cap_{\ell \in J} [X_\ell \le x_\ell]\right) \cap \left(\cap_{\ell \in J^c} [X_\ell \le x_\ell]\right).$$

for each $x = (x_1, \ldots, x_k)$ in $\mathbb{R}^p$. Next, let the coordinates x_ℓ each go (componentwise) monotonically to $+\infty$ for each $\ell \in J^c$ (where J^c is the complement of J with respect to the index set $\{1, \ldots, k\}$). It is elementary to check that

$$\begin{aligned} & \lim_{x_\ell \to \infty,\ \ell \in J^c} F_{(X_1,\ldots,X_k)}(x_1, \ldots, x_k) \\ = \ & \lim_{x_\ell \to \infty,\ \ell \in J^c} \mathbb{P}\left[X_1 \le x_1, \ldots, X_k \le x_k\right] \\ = \ & \lim_{x_\ell \to \infty,\ \ell \in J^c} \mathbb{P}\left[\left(\cap_{\ell \in J} [X_\ell \le x_\ell]\right) \cap \left(\cap_{\ell \in J^c} [X_\ell \le x_\ell]\right)\right] \\ = \ & \mathbb{P}\left[X_\ell \le x_\ell,\ \ell \in J\right] \\ = \ & F_{(X_\ell,\ \ell \in J)}(x_\ell,\ \ell \in J), \quad \begin{array}{c} x_\ell \in \mathbb{R}^{p_\ell} \\ \ell \in J. \end{array} \end{aligned} \tag{7.9}$$

This is an easy consequence of Lemma 3.1.1 when combined with the observation that

$$\lim_{x_\ell \to \infty,\ \ell \in J^c} \cap_{\ell \in J^c} [X_\ell \leq x_\ell] = \cap_{\ell \in J^c} [X_\ell \in \mathbb{R}] = \Omega.$$

The passage from $F_{(X_1,\ldots,X_k)}$ to $F_{(X_\ell,\ \ell \in J)}$ is known as *marginalization*, and is implemented by setting (the components of) $x_\ell = +\infty$ in $F_{(X_1,\ldots,X_k)}$ for each ℓ in J^c.

Through marginalization the joint probability distribution of the $\mathbb{R}^p$-valued rv $X = (X_1, \ldots, X_k)$ determines the probability distribution of any subset of components of X. However, we stress that the converse is not true in general – The marginalization process *cannot* be reversed unless additional assumptions are in place, the most common one being the mutual independence of the rvs $\{X_1, \ldots, X_k\}$; see Section 7.9. Put simply, in general knowledge of the individual probability distributions of the rvs $\{X_1, \ldots, X_k\}$ will not be sufficient to reconstruct the probability distribution of the concatenated rv $X = (X_1, \ldots, X_k)$.

7.4 Poperties of probability distribution functions ($p = 1$)

It is easy to see that the following properties hold when $p = 1$.

Proposition 7.4.1 *Given a rv $X : \Omega \to \mathbb{R}$ with probability distribution function $F_X : \mathbb{R} \to [0,1]$ under $\mathbb{P}$, the following properties hold:*

(i) Monotonicity:

$$F_X(x) \leq F_X(y), \qquad \begin{array}{c} x < y \\ x, y \in \mathbb{R}. \end{array}$$

(ii) Right-continuity:

$$\lim_{y \downarrow x} F_X(y) = F_X(x), \quad x \in \mathbb{R}.$$

(iii) Existence of a left limit:

$$\lim_{y \uparrow x} F_X(y) = F_X(x-) \quad \textit{with} \quad \mathbb{P}[X = x] = F_X(y) - F_X(x-), \quad x \in \mathbb{R}.$$

(iv) Behavior at infinity: Monotonically we have $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

Proof. (i) The monotonicity of F_X is inherited from that of $\mathbb{P}$ once we note that with x and y in $\mathbb{R}$, we have $[X \leq x] \subseteq [X \leq y]$ as soon as $x < y$, whence

$$\mathbb{P}[X \leq y] = \mathbb{P}[X \leq x] + \mathbb{P}[x < X \leq y]$$

or equivalently,

$$F_X(y) - F_X(x) = \mathbb{P}[x < X \leq y] \geq 0.$$

(ii) Pick x in $\mathbb{R}$, and let $\{y_n,\ n = 1, 2, \ldots\}$ denote a decreasing sequence in $\mathbb{R}$ such that $x < y_n$ for each $n = 1, 2, \ldots$ with $\lim_{n\to\infty} y_n = x$. By comments in (i) we have

$$F_X(y_n) - F_X(x) = \mathbb{P}[x < X \leq y_n], \quad n = 1, 2, \ldots$$

The sets $\{[x < X \leq y_n],\ n = 1, 2, \ldots\}$ form a decreasing sequence of events with $\cap_{n=1}^{\infty}[x < X \leq y_n] = \emptyset$. The conclusion $\lim_{n\to\infty} \mathbb{P}[x < X \leq y_n] = 0$ follows from Lemma 3.1.2. This last limit being independent of the sequence used, we have $\lim_{n\to\infty} F_X(y_n) = F_X(x)$ as desired.

(iii) Similarly, pick x in $\mathbb{R}$, and let $\{y_n,\ n = 1, 2, \ldots\}$ denote an increasing sequence in $\mathbb{R}$ such that $y_n < x$ for each $n = 1, 2, \ldots$ with $\lim_{n\to\infty} y_n = x$. Again, by comments in (i) we have

$$F_X(x) - F_X(y_n) = \mathbb{P}[y_n < X \leq x], \quad n = 1, 2, \ldots$$

The sets $\{[y_n < X \leq x],\ n = 1, 2, \ldots\}$ form a decreasing sequence of events with $\cap_{n=1}^{\infty}[y_n < X \leq x] = [X = x]$. This time, using Lemma 3.1.2, we get $\lim_{n\to\infty} \mathbb{P}[y_n < X \leq x] = \mathbb{P}[X = x]$. The limit $\mathbb{P}[X = x]$ being independent of the sequence used, we conclude that the limit $\lim_{n\to\infty} F_X(y_n)$ exists and is independent of the sequence used. The desired result follows.

(iv) Finally, let $\{y_n,\ n = 1, 2, \ldots\}$ denote an increasing sequence in $\mathbb{R}$ with $\lim_{n\to\infty} x_n = \infty$ monotonically. It is plain that the events $\{[X \leq x_n],\ n = 1, 2, \ldots\}$ form a increasing sequence of events with

$$\cup_{n=1}^{\infty}[X \leq y_n] = [X \in \mathbb{R}] = \Omega$$

and applying Lemma 3.1.1 yields the desired result.

Similarly, let $\{x_n,\ n = 1, 2, \ldots\}$ denote a decreasing sequence in $\mathbb{R}$ with $\lim_{n\to\infty} x_n = -\infty$ monotonically. The events $\{[X \leq y_n],\ n = 1, 2, \ldots\}$ form an increasing sequence of events with $\cap_{n=1}^{\infty}[X \leq x_n] = \emptyset$, and the desired result is obtained by applying Lemma 3.1.1. ■

Similar arguments lead to the following useful fact which complements Claim (iii) of Proposition 7.4.1 [Exercise 7.3].

Fact 7.4.1 *Given a rv $X : \Omega \to \mathbb{R}$ with probability distribution function $F_X : \mathbb{R} \to [0,1]$ under $\mathbb{P}$, it holds that*

$$\mathbb{P}[X < x] = F_X(x-), \quad x \in \mathbb{R}.$$

Definition 7.4.1

Let $\mathcal{C}(F_X)$ denote the set of points in $\mathbb{R}$ where $F_X : \mathbb{R} \to [0,1]$ is continuous, i.e.,

$$\mathcal{C}(F_X) = \{x \in \mathbb{R} : \ F_X(x-) = F_X(x)\}.$$

The complement $\mathcal{C}(F_X)^c$ of $\mathcal{C}(F_X)$ in $\mathbb{R}$ consists of the points where $F_X : \mathbb{R} \to [0,1]$ is not continuous. It is customary to call $\mathcal{C}(F_X)$ (resp. $\mathcal{C}(F_X)^c$) the set of continuity (resp. discontinuity) points of the probability distribution function of the rv X.

Lemma 7.4.1 *For any rv $X : \Omega \to \mathbb{R}$, its probability distribution function $F_X : \mathbb{R} \to [0,1]$ has the property that $\mathcal{C}(F_X)^c$ is a countable subset of $\mathbb{R}$.*

Proof. For each $n = 1, 2, \ldots$, let $\mathcal{D}_n$ denote the collection of points of discontinuity in $\mathcal{C}(F_X)^c$ whose discontinuity jump lies in the interval $(\frac{1}{n+1}, \frac{1}{n}]$, i.e.,

$$\mathcal{D}_n \equiv \left\{ x \in \mathcal{C}(F_X)^c : \ \frac{1}{n+1} < F_X(x) - F_X(x-) \le \frac{1}{n} \right\}.$$

Noting that

$$|\mathcal{D}_n| \cdot \frac{1}{n+1} \le \sum_{x \in \mathcal{D}_n} (F_X(x) - F_X(x-)) \le 1,$$

it follows that $|\mathcal{D}_n| \le n+1$. The desired result is now immediate since $\mathcal{C}(F_X)^c = \cup_{n=1}^{\infty} \mathcal{D}_n$. ■

7.5 Poperties of probability distribution functions ($p > 1$)

The case $p \ge 1$ is more involved: As the quantity $\mathbb{P}[x_k < X_k \le y_k]$ can be expressed *solely* in terms of $F_X : \mathbb{R}^p \to [0,1]$, it provides a constraint that a probability distribution function must satisfy! For instance with $p = 2$, it is easy to check that

$$\begin{aligned}
& \mathbb{P}\left[a < X_1 \le b, \alpha < X_2 \le \beta\right] \\
&= \mathbb{P}\left[X_1 \le b, \alpha < X_2 \le \beta\right] - \mathbb{P}\left[X_1 \le a, \alpha < X_2 \le \beta\right] \\
&= \left(\mathbb{P}\left[X_1 \le b, X_2 \le \beta\right] - \mathbb{P}\left[X_1 \le b, X_2 \le \alpha\right]\right) \\
&\quad - \left(\mathbb{P}\left[X_1 \le a, X_2 \le \beta\right] - \mathbb{P}\left[X_1 \le a, X_2 \le \alpha\right]\right) \\
&= \left(F_{(X_1,X_2)}(b,\beta) - F_{(X_1,X_2)}(b,\alpha)\right) \\
&\quad - \left(F_{(X_1,X_2)}(a,\beta) - F_{(X_1,X_2)}(a,\alpha)\right), \quad \begin{array}{c} a < b \\ \alpha < \beta \end{array}
\end{aligned} \tag{7.10}$$

We list below the properties that characterize the probability distribution

Proposition 7.5.1 *Given a rv $X : \Omega \to \mathbb{R}^p$ with probability distribution function $F_X : \mathbb{R}^p \to [0,1]$, the following properties hold:*

(i) Monotonicity:

(ii) Right-continuity: With the understanding that $y_k \downarrow x_k$ for each $k = 1, \ldots p$, we have

$$\lim_{y \downarrow x} F_X(y) = F_X(x), \quad x \in \mathbb{R}^p$$

(iii) Existence of a left limit: With the understanding that $y_k \uparrow x_k$ for each $k = 1, \ldots, p$, we have

$$\lim_{y \uparrow x} F_X(y) = F_X(x-) \quad \text{with} \quad \mathbb{P}\left[X = x\right] = F_X(y) - F_X(x-), \quad x \in \mathbb{R}^p$$

(iv) Behavior at infinity: Monotonically we have

$$\lim_{\min(x_k,\ k=1,\ldots,p) \to -\infty} F_X(x) = 0$$

and

$$\lim_{\min(x_k,\ k=1,\ldots,p) \to \infty} F_X(x) = 1.$$

7.6 Probability distribution functions ($p = 1$)

For $p = 1$, we turn the four properties established in Proposition 7.4.1 into a *definition* and introduce the concept of a probability distribution (function) with no reference to a measurable mapping defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 7.6.1

A probability distribution (function) on $\mathbb{R}$ is any mapping $F : \mathbb{R} \to [0, 1]$ such that

(i) Monotonicity:

$$F(x) \leq F(y), \qquad \begin{matrix} x < y \\ x, y \in \mathbb{R}. \end{matrix}$$

(ii) Right-continuity:

$$\lim_{y \downarrow x} F(y) = F(x), \quad x \in \mathbb{R}.$$

(iii) Existence of left limits:

$$\lim_{y \uparrow x} F(y) = F(x-) \quad x \in \mathbb{R}.$$

(iv) Behavior at infinity: Monotonically, we have

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1.$$

Obviously, if $X : \Omega \to \mathbb{R}$ is a rv, then its probability distribution function $F_X : \mathbb{R} \to [0, 1]$ is a probability distribution function in the sense of Definition 7.6.1. The converse given next shows that any probability distribution function in the sense of Definition 7.6.1 can always be understood as the probability distribution of a rv defined on some probability triple as defined in the sense of Definition 7.2.1. To present this construction we need the following notion that formalizes how to "invert" an arbitrary monotone increasing mapping $\mathbb{R} \to \mathbb{R}_+$.

Definition 7.6.2

Consider a mapping $F : \mathbb{R} \to \mathbb{R}_+$ which is monotone non-decreasing, i.e., $F(x) \leq F(y)$ whenever $x < y$ in $\mathbb{R}$. The *generalized inverse* associated with F is the mapping $F^{\leftarrow} : \mathbb{R}_+ \to [-\infty, +\infty]$ given by

$$F^{\leftarrow}(u) \equiv \inf \left(x \in \mathbb{R} : \ u \leq F(x) \right), \quad u \geq 0 \tag{7.11}$$

with $F^{\leftarrow}(u) = +\infty$ if the set $\{x \in \mathbb{R} : \ u \leq F(x)\}$ is empty.

Lemma 7.6.1 *For any probability distribution function $F : \mathbb{R} \to [0,1]$, there exists a probability triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ and a rv $X^\star : \Omega^\star \to \mathbb{R}$ defined on it such that its probability distribution under $\mathbb{P}^\star$ coincides with F, i.e.,*

$$\mathbb{P}^\star [X^\star \leq x] = F(x), \quad x \in \mathbb{R}.$$

This is the basis of Monte-Carlo simulation. There exists a multi-dimensional analog to this fact to be discussed later on.

Proof. Take $\Omega^\star = [0,1]$, $\mathcal{F}^\star = \mathcal{B}([0,1])$ and $\mathbb{P}^\star = \lambda$. Define the rv $X^\star : \Omega^\star \to \mathbb{R}$ by setting

$$X^\star(\omega^\star) = F^-(\omega^\star), \quad \omega^\star \in [0,1]$$

where $F^\leftarrow : [0,1] \to [-\infty, \infty]$ is the generalized inverse of F defined at (7.11). It is easy to check that

$$\mathbb{P}^\star [X^\star \leq x] = F(x), \quad x \in \mathbb{R}$$

and the probability distribution of the rv $X^\star$ under $\mathbb{P}^\star$ is indeed the probability distribution function $F : \mathbb{R} \to [0,1]$. ■

7.7 Probability distribution functions ($p \geq 1$)

7.8 Functions of rvs

Consider a rv $X : \Omega \to \mathbb{R}^p$. For any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}^q$ for some positive integer q. define the mapping $Y : \Omega \to \mathbb{R}^q$ by composing the rv $X : \Omega \to \mathbb{R}^p$ with g, namely

$$Y(\omega) = g(X(\omega)), \quad \omega \in \Omega. \tag{7.12}$$

We know that $Y : \Omega \to \mathbb{R}^q$ is a rv. A natural question is how to determine the probability distribution function $F_Y : \mathbb{R}^p \to [0,1]$ of the rv Y in terms of the probability distribution function $F_X : \mathbb{R}^p \to [0,1]$ of the rv X. The basic idea is contained in the following observation: For any Borel subset B in $\mathbb{R}^q$, it holds that

$$\begin{aligned} \mathbb{P}[Y \in B] &= \mathbb{P}[g(X) \in B] \\ &= \mathbb{P}\left[X \in g^{-1}(B)\right], \quad B \in \mathbb{R}^q. \end{aligned} \tag{7.13}$$

It immediately follows that

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left[Y \le y\right] \\ &= \mathbb{P}\left[g(X) \le y\right] = \mathbb{P}\left[g^{-1}\left((-\infty, y]\right)\right], \quad y \in \mathbb{R}^q \end{aligned} \tag{7.14}$$

as we take $B = (-\infty, y]$ in (7.8). From this last expression it is plain that in general there is no simple relationship between the probability distribution function of Y and the probability distribution function of X. In order to make progress additional assumptions are needed. Here is one example, but others will be discussed in the next two chapters.

Fact 7.8.1 *Assume* $p = q = 1$.

(i) If the mapping Borel mapping $\mathbb{R} \to \mathbb{R}$ *is strictly monotone increasing, i.e.,* $g(x) < g(y)$ *whenever* $x < y$ *in* $\mathbb{R}$*, then*

$$F_Y(y) = F_X(g^{-1}(y)), \quad y \in \mathbb{R} \quad y \in \mathbb{R}$$

(ii) If the mapping Borel mapping $\mathbb{R} \to \mathbb{R}$ *is strictly monotone decreasing, i.e.,* $g(y) < g(x)$ *whenever* $x < y$ *in* $\mathbb{R}$*, then*

$$F_Y(y) = 1 - F_X(g^{-1}(y)-), \quad y \in \mathbb{R}.$$

Proof. The proof is elementary: Use (7.14) with the following observation based on the strict monotonicity of g: For Claim (i) we have $Y \le y$ if and only if $X \le g^{-1}(y)$, and for Claim (ii) we have $Y \le y$ if and only if $g^{-1}(y) \le X$. ■

7.9 Independence of rvs

Consider a collection of rvs $\{X_i,\ i \in I\}$ which are all defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that for each i in I, the rv X_i is a $\mathbb{R}^{p_i}$-valued rv for some positive integer p_i.

Definition 7.9.1

With *I finite*, the rvs $\{X_i,\ i \in I\}$ are *mutually independent* if for any selection B_i in $\mathcal{B}(\mathbb{R}^{p_i})$ for each i in I, the events

$$\{[X_i \in B_i],\ i \in I\}$$

are mutually independent.

Applying the definitions given in Section 2.2, we see that the rvs $\{X_i,\ i \in I\}$ are mutually independent according to Definition 7.9.1 if the conditions

$$\mathbb{P}\left[X_j \in B_j,\ j \in J\right] = \prod_{j \in J} \mathbb{P}\left[X_j \in B_j\right], \qquad \begin{array}{c} B_j \in \mathcal{B}(\mathbb{R}^{p_j}),\ j \in J \\ J \subseteq I \\ 1 \leq |J| \leq |I| \end{array} \tag{7.15}$$

all hold. It is now easy to see that the rvs $\{X_i,\ i \in I\}$ are mutually independent if and only if the *smaller* set of conditions

$$\mathbb{P}\left[X_i \in B_i,\ i \in I\right] = \prod_{i \in I} \mathbb{P}\left[X_i \in B_i\right], \qquad \begin{array}{c} B_i \in \mathcal{B}(\mathbb{R}^{p_i}) \\ i \in I \end{array} \tag{7.16}$$

hold. Indeed, while (7.15) implies (7.16), it is easy to see that (7.16) implies (7.15) – Just take $B_j = \mathbb{R}^{p_j})$ for j in $I - J$!

Definition 7.9.2

More generally, with I arbitrary (and possibly uncountable), the rvs $\{X_i,\ i \in I\}$ are mutually independent if for every finite subset $J \subseteq I$, the rvs $\{X_j,\ j \in J\}$ are mutually independent.

In view of the previous comments, it is plain that the rvs $\{X_i,\ i \in I\}$ are mutually independent if and only if

$$\mathbb{P}\left[X_i \in B_i,\ i \in I\right] = \prod_{i \in I} \mathbb{P}\left[X_i \in B_i\right], \qquad \begin{array}{c} B_i \in \mathcal{B}(\mathbb{R}^{p_i}) \\ i \in I \end{array} \tag{7.17}$$

hold.

With k some fixed integer, in what follows consider a collection $\{X_1, \ldots, X_k\}$ of rvs which are *all* defined on the *same* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. For each $i = 1, \ldots, k$, the rv X_i is a $\mathbb{R}^{p_i}$-valued rv for some positive integer p_i. Again we concatenate these k rvs into a single $\mathbb{R}^p$-valued rv, denoted $(X_1, \ldots, X_k)$, where $p = p_1 + \ldots + p_k$.

The following characterization of the mutual independence of a finite number of rvs is useful.

Lemma 7.9.1 *The rvs $\{X_1, \ldots, X_k\}$ are mutually independent if and only if*

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i), \qquad \begin{array}{c} x_i \in \mathbb{R}^{p_i} \\ i = 1, \ldots, k \end{array} \tag{7.18}$$

where for each $i = 1, \ldots, n$, $F_{X_i} : \mathbb{R}^{p_i} \to [0,1]$ is the probability distribution function of the rv X_i, while $F_{X_1,\ldots,X_k} : \mathbb{R}^p \to [0,1]$ is the probability distribution function of the $\mathbb{R}^p$-valued rv $(X_1, \ldots, X_k)$.

Proof. If the rvs $\{X_1, \ldots, X_k\}$ are mutually independent, then (7.16) holds. For every $i = 1, \ldots, k$, use $B_i = (-\infty, x_i]$ with x_i in $\mathbb{R}^{p_i}$, and (7.16) becomes

$$\mathbb{P}\left[X_i \leq x_i,\ i = 1, \ldots, k\right] = \prod_{i=1}^{k} \mathbb{P}\left[X_i \leq x_i\right], \qquad \begin{array}{c} x_i \in \mathbb{R}^{p_i} \\ i = 1, \ldots, k \end{array} \tag{7.19}$$

This shows that (7.18) indeed holds.

Conversely, assume that (7.18) holds, or equivalently, that

$$\mathbb{P}\left[X_i \leq x_i,\ i = 1, \ldots, k\right] = \prod_{i=1}^{k} \mathbb{P}\left[X_i \leq x_i\right], \qquad \begin{array}{c} x_i \in \mathbb{R}^{p_i} \\ i = 1, \ldots, k \end{array} \tag{7.20}$$

■

7.10 Taking limits

Tailoring Definition 4.6.1 to the context of probability models we introduce the notion of an *extended* rv.

Definition 7.10.1

A mapping $X : \Omega \to \overline{\mathbb{R}}$ is said to be an *extended* rv if it an *extended Borel* mapping in the sense of Definition 4.6.1, i.e., we have

$$X^{-1}(B) \in \mathcal{F}, \quad B \in \mathcal{B}(\overline{\mathbb{R}})$$

where the extended Borel σ-field $\mathcal{B}(\overline{\mathbb{R}})$ on $\overline{\mathbb{R}}$ is defined by (4.23).

Consider a sequence $\{X_n,\ n = 1, 2, \ldots\}$ of of extended rvs which are *all* defined on the *same* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Using Lemma 4.6.1 we readily conclude that the following mappings $\Omega \to [-\infty, \infty]$ are rvs in the extended sense:

The maximum mappings $S \to \overline{\mathbb{R}}$ defined by

$$s \to \max_{m=1,\ldots,n} X_m(\omega), \qquad \begin{array}{c} n = 1, 2, \ldots \\ \omega \in \Omega \end{array}$$

The minimum mappings $S \to \overline{\mathbb{R}}$ defined by

$$s \to \min_{m=1,\ldots,n} X_m(\omega), \qquad \begin{array}{c} n = 1, 2, \ldots \\ \omega \in \Omega \end{array}$$

The supremum mapping $\Omega \to [-\infty, \infty]$ defined by

$$\omega \to \sup_{n \geq 1} X_m(\omega), \quad \omega \in \Omega$$

The infimum mapping $\Omega \to [-\infty, \infty]$ defined by

$$\omega \to \inf_{n \geq 1} X_m(\omega), \quad \omega \in \Omega.$$

The limsup mapping $\Omega \to [-\infty, \infty]$ defined by

$$\omega \to \limsup_{n \to \infty} X_n(\omega), \quad \omega \in \Omega.$$

The liminf mapping $\Omega \to [-\infty, \infty]$ defined by

$$\omega \to \liminf_{n \to \infty} X_n(\omega), \quad \omega \in \Omega.$$

It follows that

$$\Omega^\star \equiv \left[\liminf_{n \to \infty} X_n = \limsup_{n \to \infty} X_n \right] \in \mathcal{F}$$

and on $\Omega^\star$, $\lim_{n \to \infty} X_n$ exists (possibly as an element in $[-\infty, \infty]$), and is the common value assumed by $\liminf_{n \to \infty} X_n$ and $\limsup_{n \to \infty} X_n$.

When $\mathbb{P}[\Omega^\star] = 1$ it is customary to say that the sequence $\{X_n,\ n = 1, 2, \ldots\}$ *converges almost surely* (a.s.) (under $\mathbb{P}$), written

$$\lim_{n \to \infty} X_n \quad \mathbb{P}\text{-a.s.}$$

In that case, for any rv $X : \Omega \to \mathbb{R}$ such that

$$X(\omega) = \lim_{n \to \infty} X_n(\omega), \quad \omega \in \Omega^\star$$

we shall write

$$\lim_{n \to \infty} X_n = X \quad \mathbb{P}\text{-a.s.}$$

Such a rv X always exists when $\mathbb{P}[\Omega^\star] = 1$ but is *not* unique. Existence is immediate since we can always take

$$X(\omega) \equiv \begin{cases} \liminf_{n \to \infty} X_n(\omega) = \limsup_{n \to \infty} X_n(\omega) & \text{if } \omega \in \Omega^\star \\ \\ Z(\omega) & \text{if } \omega \notin \Omega^\star \end{cases}$$

where $Z : \Omega \to \mathbb{R}$ is some arbitrary rv, and non-uniqueness is obvious.

7.11 Exercises

Ex. 7.1 Show that the set function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \to [0,1]$ defined by (7.7) is a probability measure on $\mathcal{B}(\mathbb{R}^p)$. See also Exercise 4.3.

Ex. 7.2 Consider a pair of rvs $X, Y : \Omega \to \mathbb{R}$. Give direct arguments to show that the following mappings $\Omega \to \mathbb{R}$ are rvs:

a. $U = |X|$.

b. $V = \max(X, Y)$.

c. $W = \min(X, Y)$.

d. $Z = \alpha X + \beta Y$ with α and β arbitrary in $\mathbb{R}$.

Ex. 7.3 Give a proof of Fact 7.4.1.

Ex. 7.4 Consider a mapping $F : \mathbb{R} \to \mathbb{R}_+$ which is monotone non-decreasing, i.e., $F(x) \leq F(y)$ whenever $x < y$ in $\mathbb{R}$. Recall the definition (7.11) of the generalized inverse associated with F. Assume that F is a probability distribution function $F : \mathbb{R} \to [0,1]$,

a. What is the value of $F^{\leftarrow}(u)$ when $F(x-) \leq u < F(x)$ for some x in $\mathbb{R}$ (which is a point of discontinuity for F)?

b. Find the generalized inverse associated with

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 10 \\ 1 & \text{if } 10 \leq x \end{cases}$$

with $0 < p < 1$. Draw the graph of $F^{\leftarrow} : \mathbb{R}_+ \to [-\infty, \infty]$. Compute $F^{\leftarrow}(F(x)$ for all x in $\mathbb{R}$.

c. Find the generalized inverse associated with

$$F(x) = 1 - e^{-\lambda x^+}, \quad x \in \mathbb{R}$$

with $\lambda > 0$ and $x^+ = \max(0, x)$ for all x in $\mathbb{R}$. Compute $F^{\leftarrow}(F(x)$ for all x in $\mathbb{R}$.

Ex. 7.5 Let $F_1, \ldots, F_n$ denote probability distribution functions $\mathbb{R} \to [0,1]$. Determine which of the following mappings $G : \mathbb{R} \to \mathbb{R}$ defined below is also a probability distribution function:

a. With $\alpha_1, \ldots, \alpha_n$ in $(0,1)$ such that $\alpha_1 + \ldots + \alpha_n = 1$, the convex combination

$$G(x) = \sum_{k=1}^{n} \alpha_k F_k(x), \quad x \in \mathbb{R}$$

b. The product

$$G(x) = F_1(x) \dots F_n(x), \quad x \in \mathbb{R}.$$

c. The product

$$G(x) = F_1(x-)F_1(x), \quad x \in \mathbb{R}.$$

d. With $0 < u < 1$,

$$G(x) = 1 - u^{F_1(x)}, \quad x \in \mathbb{R}.$$

e. With $r_1 > 0, \dots, r_n > 0$,

$$G(x) = \prod_{k=1}^{n} F_k(x)^{r_k}, \quad x \in \mathbb{R}.$$

f. With $r_1 > 0, \dots, r_n > 0$,

$$G(x) = 1 - \prod_{k=1}^{n} \left(1 - F_k(x)^{r_k}\right), \quad x \in \mathbb{R}.$$

Ex. 7.6 Let F denote a probability distribution function $\mathbb{R} \to [0, 1]$. Determine which of the following mappings $\mathbb{R} \to \mathbb{R}$ defined below is also a probability distribution function:

$$H(x) = F(x) + (1 - F(x)) \log(1 - F(x)), \quad x \in \mathbb{R}$$

and

$$K(x) = -(1 - F(x))\, e + e^{1-F(x)}, \quad x \in \mathbb{R}.$$

Ex. 7.7 For $k = 1, 2$, consider the rv $X_k : \Omega_k \to \mathbb{R}^p$ defined on the probability triple $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$. Assume the rvs X_1 and X_2 to have the same probability distribution under $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively, written $(X_1, \mathbb{P}_1) = (X_2, \mathbb{P}_2)$. If the probability triples $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ are identical, we write $X_1 =_{st} X_2$. For any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}^q$, show that the rvs $g(X_1)$ and $g(X_2)$ have the same probability distribution under $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

Exercises 7.8–7.8 deal with the notion of a symmetric rv: A rv $X : \Omega \to \mathbb{R}^p$ is said to have a *symmetric* probability distribution (or more simply to be a symmetric random variable) if the rvs X and $-X$ have the same probability distribution (under $\mathbb{P}$), i.e., $X =_{st} -X$.

Ex. 7.8 Consider a symmetric rv $X : \Omega \to \mathbb{R}^p$. Give necessary and sufficient conditions on $F_X : \mathbb{R} \to [0,1]$ for the rv X to have a symmetric probability distribution.

Ex. 7.9 Assume the rv $X : \Omega \to \mathbb{R}^p$ defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ to be a symmetric rv. For any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}^q$, show that the rv $g(X) : \Omega \to \mathbb{R}^q$ is also a symmetric rv if the mapping is odd symmetric, i.e., $g(-x) = -g(x)$ for all x in $\mathbb{R}^p$.

Ex. 7.10 Consider a symmetric rv $X : \Omega \to \mathbb{R}$. Fix $a > 0$. With the rv X, we associate the rv $Y_a : \Omega \to \mathbb{R}$ given by

$$Y_a \equiv \begin{cases} X & \text{if } |X| \leq a \\ \\ -X & \text{if } a < |X|. \end{cases}$$

If X has a symmetric probability distribution, show that the rv Y_a has the same distribution as the rv X. This problem is often formulated with $X \sim \mathrm{N}(0,1)$ but the result holds more generally and requires very little computation. Again the power of probabilistic thinking at work!

Ex. 7.11 This problem deals with joint probability distribution functions.

a. Consider two rvs $X, Y : \Omega \to \mathbb{R}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and as usual let $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ denote their joint probability distribution function (under $\mathbb{P}$). If $X = Y$ a.s. (under $\mathbb{P}$), show that

$$F_{X,Y}(x,y) = H(\min(x,y)), \quad x, y \in \mathbb{R}$$

for some mapping $H : \mathbb{R} \to [0,1]$. Identify this mapping!

Next you are told that the function $F : \mathbb{R}^2 \to [0,1]$ is of the form

$$F(x,y) = K(\min(x,y)), \quad x, y \in \mathbb{R}$$

for some mapping $K : \mathbb{R} \to [0,1]$.

b. What properties should the mapping $K : \mathbb{R} \to [0,1]$ exhibit in order for the function $F : \mathbb{R}^2 \to [0,1]$ to be the joint probability distribution of a pair of rvs U and V defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$?

c. If the function $F : \mathbb{R}^2 \to [0,1]$ is indeed the joint probability distribution of a pair of rvs U and V defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, discuss whether the rvs U and V can be independent under $\mathbb{P}$?

d. Under the conditions obtained in Part **b**, is it always the case that $U = V$ a.s.? Explain [**HINT:** Note that $[U < V] = \cup_{x \in \mathbb{Q}} [U < xV]$, compute $\mathbb{P}[U < xV]$ and use the union bound].

Ex. 7.12 Consider a rv $X : \Omega \to \mathbb{R}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, Show that if $\mathbb{P}\left[X > 0\right] > 0$, then there exists $\delta > 0$ such that $\mathbb{P}\left[X \geq \delta\right] > 0$ [**HINT:** Combine the continuity from below of $\mathbb{P}$ with the fact that $\cup_{n=1}^{\infty} A_n = [X, 0]$ where $A_n \equiv [X \geq \frac{1}{n}]$ for each $n = 1, 2, \ldots$].

Chapter 8

Discrete random variables

A particularly important class of rvs is the class of *discrete* rvs. They are explored in this chapter.

8.1 Discrete distributions

Definition 8.1.1

A rv $X : \Omega \to \mathbb{R}^p$ is a discrete rv if there exists a countable subset $S_X \subseteq \mathbb{R}^p$ such that

$$\mathbb{P}\left[X \in S_X\right] = 1.$$

We refer to the countable S_X entering this definition as the *support* of the discrete rv X. It is often more convenient to characterize the distributional properties of the rv X through its *probability mass function* (pmf) $\boldsymbol{p}_X \equiv (p_X(x),\ x \in S_X)$ given by

$$p_X(x) = \mathbb{P}\left[X = x\right], \quad x \in S_X.$$

The importance of the pmf of a discrete rv is easily understood from the following easy fact.

Fact 8.1.1 *For any discrete rv $X : \Omega \to \mathbb{R}^p$ with support S, it holds that*

$$\begin{aligned} \mathbb{P}\left[X \in B\right] &= \sum_{x \in B \cap S_X} \mathbb{P}\left[X = x\right] \\ &= \sum_{x \in B \cap S_X} p_X(x), \quad B \in \mathcal{B}(\mathbb{R}^p). \qquad (8.1) \end{aligned}$$

Proof. Pick B in $\mathcal{B}(\mathbb{R}^p)$. Under Definition 8.1.1, write $S_X = \cup_{x \in S_X} \{x\}$ and note that $\mathbb{P}\left[X \notin S_X\right] = 0$. With this in mind we have

$$\begin{aligned} \mathbb{P}\left[X \in B\right] &= \mathbb{P}\left[X \in B, X \in S_X\right] + \mathbb{P}\left[X \in B, X \notin S_X\right] \\ &= \mathbb{P}\left[X \in B, X \in \cup_{x \in S_X} \{x\}\right] \\ &= \sum_{x \in S_X} \mathbb{P}\left[X \in B, X = x\right] \qquad (8.2) \end{aligned}$$

by σ-additivity since S_X is countable. This complete the proof of (8.1). ■

Note that

$$0 \leq p_X(x) \leq 1, \quad x \in S_X \quad \text{and} \quad \sum_{x \in S_X} p_X(x) = 1.$$

This observation and Fact 8.1.1 together lead to the following definition:

Definition 8.1.2

With S a countable subset of $\mathbb{R}^p$, a pmf with support on S is any collection $\boldsymbol{p} = (p(x),\ x \in S)$ such that

$$0 \leq p(x) \leq 1, \quad x \in S \quad \text{and} \quad \sum_{x \in S} p(x) = 1.$$

Lemma 7.6.1 has the following analog for pmfs; its proof is elementary and does not require the use of the generalized inverse of a monotone increasing function.

Lemma 8.1.1 *With S a countable subset of $\mathbb{R}^p$, for any pmf $\boldsymbol{p} = (p(x),\ x \in S)$ with support on S, there exists a probability triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ and a rv $X^\star : \Omega^\star \to \mathbb{R}$ defined on it such that*

$$\mathbb{P}^\star\left[X^\star = x\right] = p(x), \quad x \in S.$$

It follows that $X^\star$ is a discrete rv with support S and pmf $\boldsymbol{p} = (p(x),\ x \in S)$.

Proof. Take $\Omega^\star = S$, $\mathcal{F}^\star = \mathcal{P}(S)$ and as was done in Section 1.5, define the probability measure $\mathbb{P}^\star$ on $\mathcal{P}(S)$ by setting

$$\mathbb{P}^\star\left[E\right] = \sum_{\omega^\star \in E} p(\omega^\star), \quad E \in \mathcal{P}(S).$$

The rv $X^\star : \Omega^\star \to \mathbb{R}^p$ defined by

$$X^\star(\omega^\star) = \omega^\star, \quad \omega^\star \in \Omega^\star$$

is clearly a discrete rv supported by S. For each x in S we have $[X^\star = x] = \{x\}$ so that

$$\mathbb{P}^\star \left[X^\star = x\right] = p(x), \quad x \in S$$

and $\boldsymbol{p}$ is indeed the pmf of $X^\star$ under $\mathbb{P}^\star$. ■

8.2 Marginalization

We revisit the process of marginalization in the context of discrete rvs; the setup is that of Section 7.3: With positive integers $p_1, \ldots, p_k$, consider k distinct rvs $X_1 : \Omega \to \mathbb{R}^{p_1}, \ldots, X_k : \Omega \to \mathbb{R}^{p_k}$. Again we view this collection of rvs as a single rv $X : \Omega \to \mathbb{R}^p$ given by $X = (X_1, \ldots, X_k)$ with $p = p_1 + \ldots + p_k$ – In this notation we implicitly assume that all vectors are row vectors.

The first observation is straightforward, and states that the rv $X : \Omega \to \mathbb{R}^p$ is a discrete rv if and only if for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a discrete rv. More precisely:

Lemma 8.2.1 *(i) If for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a discrete rv with support $S_{X_\ell} \subseteq \mathbb{R}^{p_\ell}$ and pmf $\boldsymbol{p}_{X_\ell}$, then the rv $X : \Omega \to \mathbb{R}^p$ is necessarily a discrete rv whose support $S_X \subseteq \mathbb{R}^p$ satisfies the inclusion*

$$S_X \subseteq \prod_{\ell=1}^{k} S_{X_\ell}. \tag{8.3}$$

(ii) Conversely, if the rv $X : \Omega \to \mathbb{R}^p$ is a discrete rv with support $S_X \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p}_X$, then for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a discrete rv with support $S_{X_\ell} \subseteq \mathbb{R}^{p_\ell}$ satisfying the inclusion

$$S_{X_\ell} \subseteq \mathrm{Proj}_{\mathbb{R}^{p_\ell}}\left(S_X\right). \tag{8.4}$$

The pmf $\boldsymbol{p}_{X_\ell}$ is given by

$$\boldsymbol{p}_{X_\ell}(x_\ell) = \sum\nolimits^{\star\ell} p_X(x_1, \ldots, x_{\ell-1}, x_\ell, x_{\ell+1}, \ldots, x_k), \quad x_\ell \in S_{X_\ell} \tag{8.5}$$

with the summation $\sum^{\star\ell}$ taken over the set countable set $S_{-\ell}$ given by

$$S_{-\ell} \equiv \left\{x_1 \in S_{X_1}, \ldots, x_{\ell-1} \in S_{X_{\ell-1}}, x_{\ell+1} \in S_{X_{\ell+1}}, \ldots, x_k \in S_{X_k}\right\}.$$

Proof.

■

8.3 Independence of rvs

We now discuss the notion of independence for discrete rvs by specializing the results obtained in Section 7.9: Consider a collection of discrete rvs $\{X_i,\ i \in I\}$ which are all defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that for each i in I, with some positive integer p_i, the rv $X_i \to \mathbb{R}^{p_i}$ is a discrete rv with support $S_{X_i} \subseteq \mathbb{R}^{p_i}$ and pmf $\boldsymbol{p}_{X_i}$.

The following characterization of the mutual independence of a finite number of rvs is useful.

Lemma 8.3.1 *If for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a discrete rv with support $S_{X_\ell} \subseteq \mathbb{R}^{p_\ell}$ and pmf $\boldsymbol{p}_{X_\ell}$, then the rvs $\{X_1, \ldots, X_k\}$ are mutually independent if and only if*

$$p_{(X_1,\ldots,X_k)}(x_1, \ldots, x_k) = \prod_{i=1}^{k} p_{X_i}(x_i), \qquad \begin{array}{c} x_i \in S_{X_i} \\ i = 1, \ldots, k \end{array} \tag{8.6}$$

where $\boldsymbol{p}_{(X_1,\ldots,X_k)}$ is the pmf of the discrete $\mathbb{R}^p$-valued rv $(X_1, \ldots, X_k)$ with support S_X given by

$$S_X = \prod_{\ell=1}^{k} S_{X_\ell}. \tag{8.7}$$

Note the contrast with the situation encountered in Lemma 8.2.1 where only the inclusion (8.3) could be asserted.

Proof. The mutual independence of the rvs $X_1, \ldots, X_k$ implies

$$\mathbb{P}\left[X_1 = x_1, \ldots, X_k = x_k\right] = \prod_{\ell=1}^{k} \mathbb{P}\left[X_\ell = x_\ell\right], \qquad \begin{array}{c} x_\ell \in \mathbb{R}^{p_\ell} \\ \ell = 1, \ldots, k \end{array}$$

as we make use of (7.16) with the singletons $B_1 = \{x_1\}, \ldots, B_k = \{x_k\}$. It is also immediate that (8.7) holds.

Conversely, assume that (8.6)–(8.7) hold. As already noted earlier, e.g., in the proof of Fact 8.1.1, we have

$$\mathbb{P}\left[X_\ell \in S^c_{X_\ell}\right] = 0, \quad \ell = 1, \ldots, k.$$

It follows that For any choice B_1 in $\mathcal{B}(R^{p_1})$, ..., B_k in $\mathcal{B}(R^{p_k})$, this last fact allows us to write

$$\begin{aligned} & \mathbb{P}\left[X_1 \in B_1, \ldots, X_k \in B_k\right] \\ & \quad = \mathbb{P}\left[X_1 \in B_1 \cap S_{X_1}, \ldots, X_k \in B_k \cap S_{X_k}\right] \end{aligned} \tag{8.8}$$

It then follows that

$$\begin{aligned} & \mathbb{P}\left[X_1 \in B_1, \ldots, X_k \in B_k\right] \\ & \quad = \sum_{x_1 \in B_1 \cap S_{X_1}} \cdots \sum_{x_k \in B_1 \cap S_{X_k}} \mathbb{P}\left[X_1 \in B_1, \ldots, X_k \in B_k\right] \\ & \quad = \sum_{x_1 \in B_1 \cap S_{X_1}} \cdots \sum_{x_k \in B_1 \cap S_{X_k}} \mathbb{P}\left[X_1 = x_1, \ldots, X_k = x_k\right] \end{aligned} \tag{8.9}$$

■

Through marginalization the joint probability distribution of the $\mathbb{R}^p$-valued rv $X = (X_1, \ldots, X_k)$ determines the probability distribution of any subset of components of X. However, we stress that the converse is not true in general as the marginalization process cannot be reversed unless additional assumptions are in place, the most common one being the mutual independence of the rvs $\{X_1, \ldots, X_k\}$; see Section 7.9. Put simply, in general knowledge of the individual probability distributions of the rvs $\{X_1, \ldots, X_k\}$ will not be sufficient to reconstruct the probability distribution of the concatenated rv $X = (X_1, \ldots, X_k)$.

8.4 Functions of discrete rvs

We return to Section 7.8 where functions of rvs, not necessarily discrete, were discussed: Consider a rv $X : \Omega \to \mathbb{R}^p$. For any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}^q$, we introduced the rv $Y : \Omega \to \mathbb{R}^q$ defined at (7.12) by composing the rv $X : \Omega \to \mathbb{R}^p$ with g. For any Borel subset B in $\mathbb{R}^q$, the relationships

$$\mathbb{P}\left[Y \in B\right] = \mathbb{P}\left[g(X) \in B\right] = \mathbb{P}\left[X \in g^{-1}(B)\right], \quad B \in \mathbb{R}^q \tag{8.10}$$

were shown to hold.

Fact 8.4.1 *Consider a discrete rv $X : \Omega \to \mathbb{R}^p$ with support $S_X \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p}_X = (p_X(x),\ x \in S_X)$. Then, for any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}^q$, the rv $Y : \Omega \to \mathbb{R}^q$ defined at (7.12) is a discrete rv with support $S_Y \equiv \{g(x) :\ x \in S_X\}$ and pmf $\boldsymbol{p}_Y = (p_Y(Y),\ y \in S_Y)$ determined through*

$$p_Y(y) = \sum_{x \in S_X :\ g(x)=y} p_X(x), \quad y \in S_Y. \tag{8.11}$$

Proof. Under the assumptions on the rv X, the set $S_Y \equiv \{g(x) :\ x \in S_X\}$ is a countable subset of $\mathbb{R}^q$, hence a Borel subset of $\mathbb{R}^q$. From (8.10) we also note that

$$\begin{aligned}
\mathbb{P}\left[Y \in S_Y\right] &= \mathbb{P}\left[X \in g^{-1}(S_Y)\right] \\
&= \mathbb{P}\left[X \in g^{-1}(S_Y) \cap S_X\right] + \mathbb{P}\left[X \in g^{-1}(S_Y) \cap S_X^c\right] \\
&= \mathbb{P}\left[X \in g^{-1}(S_Y) \cap S_X\right] \\
&= \mathbb{P}\left[X \in S_X\right] = 1 \qquad (8.12)
\end{aligned}$$

since $S_X \subseteq g^{-1}(S_Y)$, and the rv Y is therefore a discrete rv with support S_Y.

To determine the pmf of the rv Y pick y in S_Y and consider the set $B_y \equiv \{x \in S_X :\ g(x) = y\}$ – Note that $B_y = g^{-1}(\{y\}) \cap S_X$. Using Fact 8.1.1 with B_y we then conclude from (8.10) that

$$\begin{aligned}
\mathbb{P}\left[Y = y\right] &= \mathbb{P}\left[X \in g^{-1}(\{y\})\right] \\
&= \mathbb{P}\left[X \in g^{-1}(\{y\}) \cap S_X\right] \\
&= \sum_{x \in B_y} \mathbb{P}\left[X = x\right] \\
&= \sum_{x \in S_X :\ g(x)=y} \mathbb{P}\left[X = x\right]
\end{aligned}$$

and the relation (8.11) is established. ■

Note that a non-discrete rv $X : \Omega \to \mathbb{R}^p$ can also give rise to a discrete rv when composed with a mapping as discussed here, the obvious case occurring when the mapping $g : \mathbb{R}^p \to \mathbb{R}^q$ itself has a countable range, i.e., the set $g(\mathbb{R}^p)$ is countable.

In Sections 8.5–8.9 we present well-known discrete rvs and their probability distributions through their pmfs. Unless mentioned otherwise we

8.5 Uniform rvs

Discrete uniform rvs are characterized by a finite range in $\mathbb{Z}$, say $\{a, a+1, \ldots, b-1, b\}$ with $a \leq b$ in $\mathbb{Z}$. The rv $U_{a,b} : \Omega \to \mathbb{R}$ is said to be a *discrete uniform* rv over the range $\{a, a+1, \ldots, b-1, b\}$, written $X \sim \mathcal{U}(\{a, a+1, \ldots, b-1, b\})$, if a *Bernoulli* rv with parameter p ($0 \leq p \leq 1$), written $B(p) \sim \mathrm{Ber(p)}$, if

$$\mathbb{P}\left[U_{a,b} = z\right] = \frac{1}{b-a+1}, \quad z = a, a+1, \ldots, b_1, b. \tag{8.13}$$

8.6 Bernoulli rvs

Bernoulli rvs arise naturally in the modeling of coin tossing, and are the simplest of discrete rvs. With p in $[0,1]$ the rv $B(p) : \Omega \to \mathbb{R}$ is said to be a *Bernoulli* rv with parameter p ($0 \leq p \leq 1$), written $B(p) \sim \mathrm{Ber(p)}$, if

$$\mathbb{P}\left[B(p) = 1\right] = 1 - \mathbb{P}\left[B(p) = 0\right] = p. \tag{8.14}$$

This is a discrete rv with support $S = \{0, 1\}$ and pmf given by

$$p(1) = p \quad \text{and} \quad p(0) = 1 - p. \tag{8.15}$$

In some contexts binary rvs taking the symmetric values ±1, known as Walsh rvs, are more appropriate: With p in $[0,1]$ the rv $W(p) : \Omega \to \mathbb{R}$ is said to be a Walsh rv with parameter p ($0 \leq p \leq 1$), written $W(p) \sim \mathrm{Walsh(p)}$, if

$$\mathbb{P}\left[W(p) = 1\right] = 1 - \mathbb{P}\left[W(p) = -1\right] = p. \tag{8.16}$$

This is a discrete rv with support $S = \{-1, 1\}$ and pmf given by

$$p(1) = p \quad \text{and} \quad p(-1) = 1 - p. \tag{8.17}$$

The Bernoulli rv and the corresponding Wlash rv are easily related to each other through the relations

$$W(p) = 2B(p) - 1 \quad \text{and} \quad B(p) = \frac{1 + W(p)}{2}.$$

8.7 Binomial rvs

Binomial rvs are discrete rvs whose pmfs are parametrized by a positive integer n and a probability parameter p in $[0,1]$. A rv X is said to be a *Binomial* rv with parameters $n = 1, 2, \ldots$ and p ($0 \leq p \leq 1$), written $X \sim \mathrm{Bin(n,p)}$, if

$$\mathbb{P}\left[X = x\right] = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{8.18}$$

This is a discrete rv with support $S = \{0, 1, \ldots, n\}$.

Binomial rvs naturally occur as follows: Consider n *mutually independent* Bernoulli rvs $X_1(p), \ldots, X_n(p)$ with parameter p which are defined on the *same* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ – A natural setting is the model introduced in Section 2.3. Their sum $S_n(p)$ is the rv given by

$$S_n(p) \equiv \sum_{k=1}^{n} X_k(p). \tag{8.19}$$

This rv is a discrete rv with support $S = \{0, 1, \ldots, n\}$. With $x = 0, 1, \ldots, n$, the event $S_n(p) = x$ can occur in exactly $\binom{n}{x}$ ways, where x of the n Bernoulli rvs assume the value 1 while the remaining $n - x$ take the value 0. Each of this situation is occurring with probability $p^x(1-p)^{n-x}$ as we invoke the fact that the n Bernoulli rvs $X_1(p), \ldots, X_n(p)$ are mutually independent. With this in mind we get

$$\begin{aligned} \mathbb{P}[S_n(p) = x] &= \mathbb{P}\left[\sum_{k=1}^{n} X_k(p) = x\right] \\ &= \binom{n}{x}\mathbb{P}\left[\begin{array}{c} X_k(p) = 1, \ k = 1, \ldots, x \\ X_\ell(p) = 0, \ \ell = x+1, \ldots, n \end{array}\right] \\ &= \binom{n}{x} p^x (1-p)^{n-x}. \end{aligned} \tag{8.20}$$

8.8 Poisson rvs

A rv $X : \Omega \to \mathbb{N}$ is said to be a *Poisson* rv with parameter $\lambda > 0$, written $X \sim \text{Poi}(\lambda)$, if

$$\mathbb{P}[X = x] = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots$$

The fact that

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = \left(\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}\right) e^{-\lambda} = e^{\lambda} \cdot e^{-\lambda} = 1$$

shows the family $\boldsymbol{p}_\lambda = (p_\lambda(x), \ x = 0, 1, \ldots)$ given by

$$p_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots \tag{8.21}$$

is a pmf with support $\mathbb{N}$. We refer to this pmf as the Poisson pmf with parameter λ.

The Poisson pmf naturally emerges through the following limiting process: Let $\{p_n, \ n = 1, 2, \ldots\}$ denote a collection of scalars in $(0, 1)$ Fix $x = 0, 1, \ldots$ and consider the rv $S_n(p_n)$ defined at (8.19). For each $n = x, x+1, \ldots$, we observe that

$$\begin{aligned}\mathbb{P}\left[S_n(p_n) = x\right] &= \binom{n}{x} \cdot p_n^x \cdot (1-p_n)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \cdot p_n^x \cdot (1-p_n)^{n-x} \\ &= \frac{n(n-1)\ldots(n-x+1)}{x!} p_n^x \cdot (1-p_n)^{n-x} \\ &= \frac{1}{x!} \cdot \left(\prod_{\ell=0}^{x-1} (n-\ell)p_n\right) \cdot (1-p_n)^{n-x} \\ &= \frac{(np_n)^x}{k!} \cdot \left(\prod_{\ell=0}^{x-1} \left(1 - \frac{\ell}{n}\right)\right) \cdot (1-p_n)^{n-x}. \qquad (8.22)\end{aligned}$$

Assume that $\lim_{n\to\infty} np_n = \lambda$: It is easy to see that $\lim_{n\to\infty} \frac{(np_n)^x}{k!} = \frac{\lambda^x}{x!}$ and that $\lim_{n\to\infty} \prod_{\ell=0}^{x-1} \left(1 - \frac{\ell}{n}\right) = 1$. On the other hand, standard facts lead to

$$\lim_{n\to\infty} (1-p_n)^{n-x} = \lim_{n\to\infty} \left(\left(1 - \frac{np_n}{n}\right)^n\right)^{\frac{n-x}{n}} = e^{-\lambda}. \qquad (8.23)$$

Letting n go to infinity in (8.22) we get

$$\lim_{n\to\infty} \mathbb{P}\left[S_n(p_n) = x\right] = \frac{\lambda^x}{x!} e^{-\lambda}$$

and the Poisson pmf emerges!

8.9 Geometric rvs

A rv $X : \Omega \to \mathbb{N}$ is said to be a *geometric* rv with parameter p ($0 \leq p \leq 1$), written $X \sim \text{Geo}(p)$, if

$$\mathbb{P}\left[X = x\right] = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

The fact that

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{x=0}^{\infty} (1-p)^x = \frac{1}{1-(1-p)} = 1$$

shows that the family $\boldsymbol{p}_p = (p_p(x),\ x = 0, 1, \ldots)$ given by

$$p_p(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots \qquad (8.24)$$

is a pmf with support $\mathbb{N}_0$. We refer to this pmf as the geometric pmf with parameter p.

Sometimes it is more appropriate to consider a related pmf, namely the pmf $\boldsymbol{p}_p^\star = \left(p_p^\star(x),\ x = 0, 1, \ldots\right)$ given by

$$p_p^\star(x) = p(1-p)^x, \quad x = 0, 1, \ldots \tag{8.25}$$

with support $\mathbb{N}$.

Geometric rvs occur naturally in the context of the game of chance where a two-sided coin is tossed infinitely many times under identical and independent conditions. If we assume that on a single toss the likelihood of head (resp. tail) is p (resp. $1-p$ with $0 < p < 1$), then the probability model $(\Omega, \mathcal{F}, \mathbb{P})$ developed in Section 6.2 adequately model this random experiment – Here we take $\Omega = \{0,1\}^{\mathbb{N}_0}$ with the usual understanding that the outcomes Head and Tail are encoded as 1 and 0, respectively. In that model consider the mapping $X : \Omega \to \mathbb{N} \cup \{+\infty\}$ given by

$$X(\omega) \equiv \left\{ \begin{array}{c} \text{The number of tosses before} \\ \text{the first Head appears in the sample } \omega \end{array} \right\}, \quad \omega \in \Omega$$

with $X(\omega) = +\infty$ if Heads never appears in the in the sample ω. It is easy to show that

$$\mathbb{P}\left[X = x\right] = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

8.10 Exercises

Unless specified otherwise, all rvs are assumed to be defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 8.1 Consider a rv $X : \Omega \to \mathbb{R}$ which is uniformly distributed over the interval $(-a, a)$ for some $a > 0$.

a. Give its probability distribution function $F_X : \mathbb{R} \to [0,1]$.

b. Find the probability distribution function $F_{X^+} : \mathbb{R} \to [0,1]$ of the rv $X^+ = \max(0, X)$. Is the rv X^+ a discrete rv?

Ex. 8.2 This is a continuation of Exercise 7.8: Consider a symmetric rv $X : \Omega \to \mathbb{R}^p$. Specialize your answer obtained in Exercise 7.8 to the case when X is a discrete rv with support $S \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p} = (p(x),\ x \in S)$. In particular, show that a discrete rv with support $S \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p} = (p(x),\ x \in S)$ is symmetric if and only if $S = -S$ and $p(x) = p(-x)$ for all x in S.

Ex. 8.3 Consider two independent discrete rvs X and Y which are identically distributed; their common $\boldsymbol{p} = (p(z),\ z \in \mathbb{Z})$ has support $S = \mathbb{Z}$.

a. Compute $\mathbb{P}[X = Y]$ explicitly in terms of $(p(z),\ z = 0, \pm 1, \pm 2, \ldots)$.

b. Without doing any calculations show that $\mathbb{P}[X < Y] = \mathbb{P}[Y < X]$.

c. Using Parts **a** and **b** show that $\mathbb{P}[X < Y] = \frac{1}{2}\left(1 - \sum_{z \in \mathbb{Z}} p(z)^2\right)$.

Ex. 8.4 Consider a collection $\{X_n,\ n = 1, 2, \ldots\}$ of discrete rvs with X_n distributed uniformly over the set $\{1, \ldots, n\}$ for each $n = 1, 2, \ldots$. For any positive integer k compute the probability that an infinite number of the rvs will assume the value k, namely $\mathbb{P}[X_n = k \text{ i.o.}]$.

Ex. 8.5 Let $X_1, \ldots, X_n$ be n discrete rvs $\Omega \to \mathbb{N}$ defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. They are assumed to be mutually independent. Define the sum rv $S_n \equiv X_1 + \ldots + X_n$.

a. Compute the pmf of the rv S_n if for all $k = 1, \ldots, n$, $X_k \sim \text{Ber}(p)$ for some $0 < p < 1$.

b. Compute the pmf of the rv S_n if for all $k = 1, \ldots, n$, $X_k \sim \text{Bin}(n_k, p)$ for some $0 < p < 1$ and positive integer n_k. Can you use Part **a** to conclude without having to do any calculations?

c. Compute the pmf of the rv S_n if for all $k = 1, \ldots, n$, $X_k \sim \text{Poi}(\lambda_k)$ for some $\lambda_k > 0$.

Ex. 8.6 This problem arises in the context of Eschenauer-Gligor random key predistribution scheme: Let P and K be two positive integers such that $K < P$. Given is a pool of P distinct keys, labelled $1, 2, \ldots, P$ – Write $\mathcal{P} \equiv \{1, \ldots, P\}$. Each of n devices selects uniformly at random exactly K keys from the key pool $\mathcal{P}$, said selections being mutually independent. For each $i = 1, \ldots, n$, let Γ_i denote the random set of K keys selected by device i. According to the Eschenauer-Gligor scheme, two devices that can communicate wirelessly will be able to do so securely if they share at least one key in common.

a. Construct a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ to study this situation.

b. Compute

$$\mathbb{P}[\Gamma_i = S], \qquad \begin{array}{c} S \subseteq \mathcal{P} \\ |S| = K. \end{array}$$

Now define the binary rvs

$$\chi_{ij} = \mathbf{1}[\Gamma_i \cap \Gamma_j \neq \emptyset], \qquad \begin{array}{c} i \neq j \\ i, j = 1, \ldots, n. \end{array}$$

Note that $\chi_{ij} = 1$ (resp. $\chi_{ij} = 0$) means that nodes i and j have a key in common (resp. do not have a key in common) in their key rings.

c. Compute the probabilities

$$\mathbb{P}\left[\chi_{ij} = 1\right] = \mathbb{P}\left[\Gamma_i \cap \Gamma_j = \emptyset\right], \qquad \begin{array}{c} i \neq j \\ i,j = 1,\ldots,n. \end{array}$$

d. Do the rvs $\{\chi_{1j},\ j = 2,\ldots,n\}$ form a collection of mutually independent rvs?

e. Are the rvs χ_{12}, χ_{23} and χ_{31} mutually independent?

Ex. 8.7 This problem deals with the following random experiment: A coin is tossed infinitely many times under identical and independent conditions. It is assumed that on a single toss the likelihood of head is p (with $0 < p < 1$). To model this experiment use the probability model $(\Omega, \mathcal{F}, \mathbb{P})$ developed in Section 6.2 (where $\Omega = \{0,1\}^{\mathbb{N}_0}$).

a. Define the mapping $X : \Omega \to \mathbb{N} \cup \{+\infty\}$ given by

$$X(\omega) = \left\{ \begin{array}{c} \text{The number of tosses before} \\ \text{the first Head appears in the sample } \omega \end{array} \right\}, \quad \omega \in \Omega$$

with $X(\omega) = +\infty$ if Heads never appears in the in the sample ω. Explain why the mapping $X : \Omega \to \mathbb{R}$ so defined is indeed a rv. Is it a discrete rv?

c. Find the pmf of this rv, i.e., $\{\mathbb{P}\left[X = m\right],\ m = 1,2,\ldots\}$.

d. On the probability triple used here (and discussed in Section 6.2), is it possible to define a rv $Y : \Omega \to \mathbb{R}$ which is *not* a discrete rv? In the affirmative give an example.

Ex. 8.8 A rv $X : \Omega \to \mathbb{R}$ is said to have a *symmetric* probability distribution (or more simply to be a symmetric random variable) if the rvs X and $-X$ have the same probability distribution (under $\mathbb{P}$), i.e., $X =_{st} -X$. Specialize your answer to the case when X is a discrete rv with support $S \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p} = (p(x),\ x \in S)$.

Ex. 8.9 Let X and Y be two independent Poisson rvs $\Omega \to \mathbb{R}$, say $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ with $\lambda, \mu > 0$. Show that the rv $Z = X + Y$ is also a Poisson rv and identify its parameter, i.e., $Z \sim \text{Poi}(\nu)$ for some $\nu > 0$. Generalize to K mutually independent Poisson rvs $X_1, \ldots, X_K$ with

$$X_k =_{st} \text{Poi}(\lambda_k) \qquad \begin{array}{c} \lambda_k > 0 \\ k = 1,\ldots,K \end{array}$$

Carefully explain your reasoning.

Ex. 8.10 Let N be a Poisson rv, and let $\{B_n,\ n = 1, 2, \ldots\}$ be a collection of Bernoulli rvs with

$$\mathbb{P}\left[B_n = 1\right] = 1 - \mathbb{P}\left[B_n = 0\right] = p, \quad \begin{array}{c} n = 1, 2, \ldots \\ 0 < p < 1. \end{array}$$

If the rvs $\{N, B_n,\ n = 1, 2 \ldots\}$ are mutually independent, show that (i) the rvs X and Y defined by

$$X := \sum_{i=1}^{N} B_i \quad \text{and} \quad Y := \sum_{i=1}^{N} (1 - B_i)$$

are independent, and that (ii) the rvs X and Y are Poisson rvs with parameters λp and $\lambda(1-p)$, respectively. Can you use this result to provide an alternative solution to Exercise 8.9. Explain! Again a case of probabilistic reasoning at work!

Ex. 8.11 Consider the discrete rv $Z : \Omega \to \mathbb{Z}$ whose pmf $\boldsymbol{p}_Z = (p_Z(z),\ z \in \mathbb{Z})$ (under $\mathbb{P}$) is given by

$$p_Z(z) = Cq^{|z|}, \quad z \in \mathbb{Z}$$

for some $C > 0$ and $0 < q < 1$.

a. Determine the value of C as a function of q.

Ex. 8.12 Let B and X be two independent rvs $\Omega \to \mathbb{N}$ with $B \sim \mathrm{Ber}(\frac{1}{2})$ and $X \sim \mathrm{Geo}(p)$ with $0 < p < 1$, i.e., $\mathbb{P}\left[B = 1\right] = p$ and $\mathbb{P}\left[B = 0\right] = 1 - p$, while $\mathbb{P}\left[X = \ell\right] = p(1-p)^{\ell-1}$ for $\ell = 1, 2, \ldots$. Define the rv $Y : \Omega \to [0, +\infty)$ given by

$$Y \equiv B \cdot X + (1 - B) \cdot \frac{1}{X} = \begin{cases} X & \text{if } B = 1 \\ \\ X^{-1} & \text{if } B = 0 \end{cases}$$

a. Determine its support S_Y of the discrete rv Y and find its pmf $\boldsymbol{p}_Y$.

b. Introduce the discrete rv $Z : \Omega \to [0, +\infty)$ given by $Z \equiv Y^{-1}$. Determine the support S_Z of the discrete rv Z. and find its pmf $\boldsymbol{p}_Z$.

Ex. 8.13 Let $X_1, \ldots, X_n$ be n discrete rvs $\Omega \to \mathbb{N}$ defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. They are assumed to be mutually independent and to be geometrically distributed in the sense that for each $k = 1, \ldots, n$, we have $X_k \sim \mathrm{Geo}(p_k)$ for some $0 < p_k < 1$ (not necessarily identical).

a. For each $k = 1, 2, \ldots, n$, compute $\mathbb{P}\left[X_k > x\right]$ for each $x = 0, 1, \ldots$.

b. Find the pmf of the rv $V_n \equiv \min\{X_1, \ldots, X_n\}$ **[HINT:** Compute $\mathbb{P}\left[V_n > x\right]$ for each $x = 0, 1, \ldots$ and identify the pmf!].

Ex. 8.14 Let $P, U_1, \ldots, U_n$ be $n+1$ mutually independent rvs defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that the rvs $U_1, \ldots, U_n$ are uniformly distributed on the interval $[0,1]$, and that the rv P is simple rv of the form

$$P = \sum_{i \in I} p_k \mathbf{1}\left[A_k\right]$$

for some finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ and scalars $\{p_i,\ i \in I\}$ in $[0,1]$ all distinct. Define the discrete rvs $X_1, \ldots, X_n$ to be

$$X_k \equiv \mathbf{1}\left[U_k \leq P\right], \quad k = 1, 2, \ldots$$

a. Assume first that $|I| = 1$. What is the common pmf of the rvs $X_1, \ldots, X_n$? Are the rvs $X_1, \ldots, X_n$ pairwise independent? Are they mutually independent?

Assume next that $|I| \geq 2$ with $\{p_i,\ i \in I\} \subset (0,1)$.

b. Find the common pmf of the rvs $X_1, \ldots, X_n$.

c. Are the rvs $X_1, \ldots, X_n$ pairwise independent? Are they mutually independent?

Chapter 9

(Absolutely) continuous random variables

A particularly important class of rvs is the class of *discrete* rvs. They are explored in this chapter.

9.1 Continuous distributions

Definition 9.1.1

A rv $X : \Omega \to \mathbb{R}^p$ is a continuous rv if such that

$$.$$

9.2 Marginalization

9.3 Exercises

Unless specified otherwise, all rvs are assumed to be defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Chapter 10

Mathematical expectations: Definitions

The probability distribution function of a rv $X : \Omega \to \mathbb{R}$ is a complicated object – For all intent and purposes, it is an infinite-dimensional object since it needs to be specified at every point x in $\mathbb{R}$. Yet much information concerning the probabilistic behavior of the rv can already be gleaned from lower-dimensional measures associated with its probability distribution. In the frequentist context, such quantities can be viewed as averages. In this chapter we make sense of them through the notion of *expected value* or *expectation* of a rv. This requires us to appeal to Lebesgue integration (and its generalization) as developed in the context of Measure Theory. This is developed in the next section under the following algebraic conventions: No meaning is attributed to $\infty - \infty$. Furthermore, we shall make the following conventions:

$$0 \cdot (\pm\infty) = \pm 0,$$

$$c \pm \infty = \pm\infty, \quad c \in \mathbb{R}$$

and

$$c \cdot (\pm\infty) = \mathrm{sgn}(c) \cdot (\pm\infty), \quad \begin{array}{c} c \neq 0 \\ c \in \mathbb{R} \end{array}$$

10.1 Natural requirements

Throughout the discussion we assume given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ on which all rvs are defined. Whenever possible, with any rv $X : \Omega \to \mathbb{R}$ we seek to associate a (possibly infinite) scalar in $[-\infty, \infty]$, denoted $\mathbb{E}[X]$; this value can be interpreted as an *average value* for X as *weighted* by its probability distribution

F_X. We shall refer to $\mathbb{E}\left[X\right]$, when it exists, as the *expectation* of X. This definition for the expectation operator is guided by the following requirements.

R1. Expectation is determined solely by the probability distribution of X – The expectation $\mathbb{E}\left[X\right]$ should be determined *solely* by the probability distribution $F_X : \mathbb{R} \to \mathbb{R}$: Thus, if $X' : \Omega' \to \mathbb{R}$ is another rv (possibly defined on some different probability triple $(\Omega', \mathcal{F}', \mathbb{P}')$) with distribution $F_{Y'} : \mathbb{R} \to \mathbb{R}$ (under $\mathbb{P}'$), then the distributional equality $F_X = F_{Y'}$ implies $\mathbb{E}\left[X\right] = \mathbb{E}'[X']$ (when it exists). Put differently, the existence of $\mathbb{E}\left[X\right]$ is determined, and its value computable, on the basis of F_X alone.

The definition of the quantity $\mathbb{E}\left[X\right]$ does *not* depend on the *type* of distribution of the rv X, say discrete or absolute continuous, but does coincide with the usual definitions given in elementary courses in Probability Theory. The first step towards realizing this requirement will follow from the next requirements.

R2. Expectation generalizes probabilities – The expectation of the indicator function of an event A in $\mathcal{F}$ should coincide with its probability under $\mathbb{P}$, namely if $X = \mathbf{1}\left[A\right]$, then

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\mathbf{1}\left[A\right]\right] = \mathbb{P}\left[A\right], \quad A \in \mathcal{F}.$$

R3. Expectations for non-negative rvs – The expectation of non-negative rvs is *always* well defined (although it could be infinite) with $\mathbb{E}\left[X\right] \geq 0$ whenever $X \geq 0$.

R4. Linearity – The expectation operator is *linear* in the following sense: Consider rvs $X, Y : \Omega \to \mathbb{R}$ defined on the *same* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. If their expectations exist, then for any scalars a and b, the equality

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right]$$

holds whenever the expression $a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right]$ is well defined. In particular this will happen when $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[Y\right]$ are both finite. When at least one of the expectations $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[Y\right]$ is infinite, this requirement may put conditions on the sign of a and b for the right-hand side to be well defined.

A definition of $\mathbb{E}\left[X\right]$ that meets the requirements **R1-R4** is given through a three-stage process discussed in the next sections:

- Step 1: For indicator rvs and for simple rvs.

- Step 2: For non-negative rvs through an approximation argument (via a limiting process) in terms of simple rvs (to be defined next).
- Step 3: For arbitrary rvs through a decomposition in positive and negative parts.

As will become shortly apparent, this three-step definition is quite concrete in spite of the many mathematical details (which can be omitted in first reading) that need to be considered. We shall see that the expectation operation so constructed has a couple of useful by-products:

Monotonicity The operation that associates an expectation with a rv is *monotone* in the following sense: If two rvs $X, Y : \Omega \to \mathbb{R}$ are ordered in the sense that $X \leq Y$, then $Y - X \geq 0$. If in addition both expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are well defined and finite, then $\mathbb{E}[Y - X] \geq 0$ by **R3** while $\mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X]$ by linearity **R4**. As a result, $\mathbb{E}[X] \leq \mathbb{E}[X]$. It turns out that a somewhat stronger result holds when the expectations are not finite.

Interchange of limits and expectations Consider a sequence of rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ all defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for each ω in an event $\Omega_\star$ in $\mathcal{F}$ with $\mathbb{P}[\Omega_\star] = 1$. Furthermore assume that the expectations of the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ are all well defined. Under certain conditions we shall see that the first limit below exists, and that the *interchange* of limit and expectation

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right] (= \mathbb{E}[X])$$

is valid.

10.2 Simple rvs

Simple rvs to be defined shortly are the building blocks of this construction. First a couple of definitions and some terminology.

Definition 10.2.1

With I an index set, an *$\mathcal{F}$-partition* of Ω is a collection $\{A_i,\ i \in I\}$ of events in $\mathcal{F}$ such that

$$A_i \cap A_j = \emptyset, \quad \begin{matrix} i \neq j \\ i, j \in I \end{matrix} \quad \text{and} \quad \cup_{i\in I} A_i = \Omega.$$

Such an $\mathcal{F}$-partition is said to be *finite* (resp. *countable*) if the index set I is finite (resp. countable). In most cases of interest, the events $\{A_i,\ i \in I\}$ are non-empty.

Definition 10.2.2

A rv $X : \Omega \to \mathbb{R}$ is said to be a *simple* rv if it is of the form

$$X = \sum_{i \in I} a_i \mathbf{1}\left[A_i\right] \tag{10.1}$$

for some finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ and a collection $\{a_i,\ i \in I\}$ of scalars in $\mathbb{R}$.

A simple rv X is well defined due to the fact that $\{A_i,\ i \in I\}$ is an $\mathcal{F}$-partition: Indeed, for each ω in Ω, there exists exactly one index i in I such that ω belongs to A_i, in which case $X(\omega) = a_i$. In this definition, some of the events in the partition could be empty and the scalars values $\{a_i,\ i \in I\}$ are not necessarily all distinct of each other. Thus, the representation (10.2.2) of a simple rv is *not* necessarily unique. However, in many arguments there is no loss of generality in assuming the values $\{a_k,\ k \in I\}$ to be *distinct* scalars and the events $\{A_k,\ k \in I\}$ forming the $\mathcal{F}$-partition to be all non-empty, in which case $\{X(\omega),\ \omega \in \Omega\} = \{a_k,\ k \in I\}$ with

$$A_k = \left[X = a_k\right], \quad k \in I.$$

We refer to this representation as the *generic* representation of the simple rv. It is easy to see that it is *unique*.

Here are some easy facts concerning simple rvs; the proofs are left as exercises.

Fact 10.2.1 *If $X, Y : \Omega \to \mathbb{R}$ are simple rvs, then the rvs $X + Y$ and cX (with scalar c) are also simple rvs.*

A number of proofs (including that of Fact 10.2.1) will rely on the following simple observation (mentionned here for easy reference): Assume that $X : \Omega \to \mathbb{R}$ is a simple rv with finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ and collection $\{a_i,\ i \in I\}$ of scalars in $\mathbb{R}$, and that $Y : \Omega \to \mathbb{R}$ is a simple rv with finite $\mathcal{F}$-partition $\{B_j,\ j \in J\}$ and collection $\{b_j,\ j \in J\}$ of scalars in $\mathbb{R}$. It is plain that

$$X = \sum_{i \in I} \sum_{j \in J} a_i \mathbf{1}\left[A_i \cap B_j\right] \text{ and } Y = \sum_{i \in I} \sum_{j \in J} b_j \mathbf{1}\left[A_i \cap B_j\right].$$

In other words, the rv X (resp. Y) can be interpreted as a simple rv with finite $\mathcal{F}$-partition $\{C_{i,j},\ (i,j) \in I \times J\}$ and collection $\{a_{i,j},\ (i,j) \in I \times J\}$ (resp.

$\{b_{i,j},\ (i,j) \in I \times J\}$) of scalars in $\mathbb{R}$ where for each pair (i,j) in $I \times J$, we have defined $C_{i,j} = A_i \cap B_j$ with scalars $a_{i,j} = a_i$ and $b_{i,j} = b_j$. Put differently, when considering two simple rvs, there is no loss of generality in assuming that they are constructed on the same finite $\mathcal{F}$-partition.

10.3 Approximating with simple rvs

The next definition is central to the definition of expectation presented in later sections.

Definition 10.3.1

The sequence of rvs $\{X_n,\ n = 1,2,\ldots\}$ is called a *(monotone) staircase approximation from below* for the rv $X : \Omega \to \mathbb{R}$ if for each $n = 1,2,\ldots$, the rv $X_n : \Omega \to \mathbb{R}$ is a simple variable such that

(i) The sequence is pointwise non-decreasing in the sense that for every ω in Ω, the sequence $\{X_n(\omega),\ n = 1,2,\ldots\}$ is non-decreasing with

$$X_n(\omega) \leq X_{n+1}(\omega) \leq X(\omega), \quad n = 1,2,\ldots \tag{10.2}$$

(ii) The sequence converges pointwise with

$$\lim_{n\to\infty} X_n(\omega) = X(\omega), \quad \omega \in \Omega. \tag{10.3}$$

The existence of the limit (10.3) is ensured by the monotonicity (10.2).

Throughout we shall drop the qualifiers monotone and from below. For the purpose of defining expectations the key observation concerning simple rvs is contained in the following lemma which deals with *non-negative* rvs.

Lemma 10.3.1 *For any non-negative rv $X : \Omega \to \mathbb{R}_+$, there always exists a staircase approximation $\{X_n,\ n = 1,2,\ldots\}$ of X made of simple non-negative rvs $\Omega \to \mathbb{R}_+$ with*

$$X_n = g_n(X), \quad n = 1,2\ldots$$

for some Borel mapping $g_n : \mathbb{R} \to \mathbb{R}_+$.

The sequence $\{X_n,\ n = 1,2,\ldots\}$ whose existence is announced in Lemma 10.3.1 is *not* unique as can be seen by a careful examination of the proof below [Exercise 10.2].

Proof. For each $n = 1, 2, \ldots$, consider the rv $X_n : \Omega \to \mathbb{R}_+$ given by

$$X_n = \begin{cases} k2^{-n} & \text{if} \quad \begin{array}{l} k2^{-n} \le X < (k+1)2^{-n}, \\ k = 0, 1, \ldots, 4^n - 1 \end{array} \\ & \\ 0 & \text{if} \quad X \ge 2^n. \end{cases} \tag{10.4}$$

The reader will readily check that the rv X_n is a simple rv associated with the $\mathcal{F}$-partition $\{A_{n,k},\ k = 0, 1, \ldots, 4^n - 1, 4^n\}$ given by

$$A_{n,k} = \begin{cases} [k2^{-n} \le X < (k+1)2^{-n}] & \text{if } k = 0, 1, \ldots, 4^n - 1 \\ & \\ [X \ge 2^n] & \text{if } k = 4^n \end{cases} \tag{10.5}$$

and associated values $\{a_{n,k},\ k = 0, 1, \ldots, 4^n - 1, 4^n\}$ given by

$$a_{n,k} = \begin{cases} k2^{-n} & \text{if } k = 0, 1, \ldots, 4^n - 1 \\ & \\ 0 & \text{if } k = 4^n. \end{cases}$$

The partition (10.5) is an $\mathcal{F}$-partition by virtue of the fact that X is a rv.

Parts (i) and (ii) are immediate consequence of the following observation (whose proof is left as an exercise): Fix x arbitrary in $\mathbb{R}$, and set

$$k_n(x) \equiv \lfloor x2^n \rfloor \quad \text{and} \quad x_n \equiv k_n(x)2^{-n}, \quad n = 1, 2, \ldots$$

As we note that $2k_n(x) \le x2^{n+1}$, it is a simple matter to check that $2k_n(x) \le k_{n+1}(x)$, whence $x_n \le x_{n+1} \le x$ with $x_n \le x < x_n + 2^{-n}$. The sequence $\{x_n,\ n = 1, 2, \ldots\}$ is therefore monotone increasing with $\lim_{n\to\infty} x_n = x$. Obviously we have $x_n \ge 0$ for all $n = 1, 2, \ldots$ if $x \ge 0$. ■

Note that for each $n = 1, 2, \ldots$, the rv X_n in Lemma 10.3.1 can be defined as

$$X_n = \begin{cases} \lfloor X2^n \rfloor 2^{-n} & \text{if } 0 \le X < 2^n \\ & \\ 0 & \text{otherwise.} \end{cases} \tag{10.6}$$

Therefore, it is indeed the case that $X_n = g_n(X)$ with Borel mapping $g_n : \mathbb{R} \to \mathbb{R}_+$ given by

$$g_n(x) = \begin{cases} \lfloor x2^n \rfloor 2^{-n} & \text{if } 0 \le x < 2^n \\ & \\ 0 & \text{otherwise.} \end{cases} \tag{10.7}$$

10.4 Defining the expectation of a rv

We are now ready to define the expectation of a rv $X : \Omega \to \mathbb{R}$. This will be done according to the three steps announced earlier.

Step 1 – Simple rvs

Consider a simple rv $X : \Omega \to \mathbb{R}$ of the form (10.1) for some finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ with associated collection $\{a_i,\ i \in I\}$ of scalars in $\mathbb{R}$. We define its expectation $\mathbb{E}[X]$ by

$$\mathbb{E}[X] \equiv \sum_{i \in I} a_i \mathbb{P}[A_i] . \tag{10.8}$$

It follows immediately that if $X \equiv c$ for some scalar c in $\mathbb{R}$, then $\mathbb{E}[X] = x$. Furthermore, given the requirements laid down in Section 10.1 the definition (10.8) is the only definition possible: Indeed, we must have

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\sum_{i \in I} a_i \mathbf{1}[A_i]\right] \\ &= \sum_{i \in I} a_i \mathbb{E}[\mathbf{1}[A_i]] \quad \text{[By linearity } \mathbf{R4}\text{]} \\ &= \sum_{i \in I} a_i \mathbb{P}[A_i] \quad \text{[By } \mathbf{R2}\text{]} \end{aligned} \tag{10.9}$$

The definition (10.8) does *not* depend on the particular representation used for the simple rv X, and is therefore well posed [Exercise 10.2].

Linearity and monotonicity of expectation both hold on the class of simple rvs. This follows by an easy use of the observation concluding Section 10.2, and is left as an easy exercise.

Lemma 10.4.1 *Assume the rvs $X, Y : \Omega \to \mathbb{R}$ to be simple rvs. The following holds:*

(i) Linearity: For arbitrary scalars a and b in $\mathbb{R}$,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

where the rv $aX + bY$ is also a simple rv by Fact 10.2.1;

(ii) Monotonicity: If $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Step 2 – Non-negative rvs

Consider a non-negative rv $X : \Omega \to \mathbb{R}_+$, and let $\{X_n,\ n = 1, 2, \ldots\}$ denote any collection of simple non-negative rvs which form a staircase approximation of X: We define $\mathbb{E}[X]$ by the limiting process

$$\mathbb{E}[X] \equiv \lim_{n\to\infty} \mathbb{E}[X_n]. \tag{10.10}$$

The expectation $\mathbb{E}[X]$ defined at (10.10) always exists as an element in $[0, +\infty]$; this is a consequence of the fact that the sequence $\{\mathbb{E}[X_n],\ n = 1, 2, \ldots\}$ is non-decreasing in $\mathbb{R}_+$ by Part (ii) of Lemma 10.4.1 (since the sequence $\{X_n,\ n = 1, 2, \ldots\}$) is monotone non-decreasing).

At this point the reader may wonder whether this definition is *independent* of the staircase approximation sequence $\{X_n,\ n = 1, 2, \ldots\}$ being used in (10.10). Before showing that it is indeed the case, we prove the following fact whose proof is given in Section 10.6.

Lemma 10.4.2 *Consider a non-negative rv $X : \Omega \to \mathbb{R}_+$, and let $\{X_n,\ n = 1, 2, \ldots\}$ denote any collection of simple non-negative rvs which form a staircase approximation for X. For any simple non-negative rv $Y : \Omega \to \mathbb{R}_+$ such that $Y \leq X$, it holds that*

$$\mathbb{E}[Y] \leq \lim_{n\to\infty} \mathbb{E}[X_n]. \tag{10.11}$$

This observation has the following consequence:

Lemma 10.4.3 *Consider a non-negative rv X, and let $\{X_{1,n},\ n = 1, 2, \ldots\}$ and $\{X_{2,n},\ n = 1, 2, \ldots\}$ be collections of simple non-negative rvs which form a staircase approximation of X, i.e., for each $k = 1, 2$, we have $X_{k,n} \leq X_{k,n+1} \leq X$ for $n = 1, 2, \ldots$ with $\lim_{n\to\infty} X_{k,n} = X$ pointwise. It holds that*

$$\lim_{n\to\infty} \mathbb{E}[X_{1,n}] = \lim_{n\to\infty} \mathbb{E}[X_{2,n}]. \tag{10.12}$$

This indeed shows that the definition (10.10) of $\mathbb{E}[X]$ is independent of the staircase approximation sequence for X being used.

Proof. Fix $k = 1, 2$ and note that $X_{k,n} \leq X$ for all $n = 1, 2, \ldots$. For $\ell = 1, 2$ with $\ell \neq k$, Lemma 10.4.2 yields $\mathbb{E}[X_{\ell,m}] \leq \lim_{n\to\infty} \mathbb{E}[X_{k,n}]$ for each $m = 1, 2, \ldots$ (when applied with $Y = X_{\ell,m}$). It immediately follows that

$$\lim_{m\to\infty} \mathbb{E}[X_{\ell,m}] \leq \lim_{n\to\infty} \mathbb{E}[X_{k,n}].$$

Exchanging the role of k and ℓ in this last inequality we conclude to the validity of (10.12). ■

Step 3 – The general case

Setting $X^+ = \max(0, X)$ and $X^- = \max(0, -X)$, we recall the decompositions

$$X = X^+ - X^- \quad \text{and} \quad |X| = X^+ + X^-. \tag{10.13}$$

We define

$$\mathbb{E}[X] \equiv \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right] \tag{10.14}$$

with the understanding that at least one of the terms $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ is finite. This definition is indeed the natural definition if one is to expect the expectation to have a chance to be linear.

There are four possible cases: (i) If both $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are finite, then $\mathbb{E}[|X|] = \mathbb{E}[X^+] + \mathbb{E}[X^-] < \infty$; (ii) If $\mathbb{E}[X^+] = \infty$ with $\mathbb{E}[X^-]$ finite, then $\mathbb{E}[X] = \infty$; (iii) If $\mathbb{E}[X^-] = \infty$ with $\mathbb{E}[X^+]$ finite, then $\mathbb{E}[X] = -\infty$. in both these cases $\mathbb{E}[|X|] = \mathbb{E}[X^+] + \mathbb{E}[X^-] = \infty$. (iv) Finally, if $\mathbb{E}[X^+] = \mathbb{E}[X^-] = \infty$, then $\mathbb{E}[X]$ cannot be defined, yet $\mathbb{E}[|X|] = \infty$. We summarize these observations through the following definition:

Definition 10.4.1

The expectation $\mathbb{E}[X]$ of the rv X is said to *exist* if

$$\min\left(\mathbb{E}\left[X^+\right], \mathbb{E}\left[X^-\right]\right) < \infty.$$

It will be *finite* if the stronger condition $\mathbb{E}[X^+] + \mathbb{E}[X^-] < \infty$ holds, in which case the rv X is said to be *integrable*.

10.5 The expectation of a rv depends only on its probability distribution

As we now show the expectation operation defined in Section 10.4 does satisfy the requirement **R1**.

Lemma 10.5.1 *For any rv $X : \Omega \to \mathbb{R}$, whenever it exists, the expectation $\mathbb{E}[X]$ is fully determined by the probability distribution F_X of the rv X.*

Proof. We start with the case where X is a non-negative rv. Let $\{X_n,\ n = 1, 2, \ldots\}$ denote the collection of simple non-negative rvs introduced in the proof of Lemma 10.3.1 to show the existence of a staircase approximation to X. For each $n = 1, 2, \ldots$ we note that

$$\begin{aligned}
\mathbb{E}\left[X_n\right] &= \sum_{k=0}^{4^n-1} k2^{-n}\mathbb{P}\left[k2^{-n} \leq X < (k+1)2^{-n}\right] \\
&= \sum_{k=0}^{4^n-1} k2^{-n}\left(F_X((k+1)2^{-n}-) - F_X(k2^{-n}-)\right),
\end{aligned}$$

and $\mathbb{E}\left[X_n\right]$ indeed depends only on F_X. The definitional equality $\mathbb{E}\left[X\right] = \lim_{n\to\infty}\mathbb{E}\left[X_n\right]$ thus implies that $\mathbb{E}\left[X\right]$ depends only on the probability distribution F_X of X.

For the general case, introduce the non-negative rvs X^+ and X^-. By the first part of the proof, we conclude that the expectations $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$, while possibly infinite, are determined by the probability distributions F_{X^+} and F_{X^-}, respectively. However, it is plain that the probability distribution of X^+ is determined by that of X since

$$\mathbb{P}\left[X^+ \leq x\right] = \begin{cases} F_X(x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{10.15}$$

by elementary calculations. Therefore, the expectation $\mathbb{E}\left[X^+\right]$ depends only on the probability distribution F_X of X. A similar argument shows that the expectation $\mathbb{E}\left[X^-\right]$ depends only on the probability distribution F_X of X, and the desired conclusion concerning $\mathbb{E}\left[X\right]$ (when well defined) follows ■

10.6 A proof of Lemma 10.4.2

The expectations $\mathbb{E}\left[Y\right]$ and $\mathbb{E}\left[X_n\right]$, $n = 1, 2, \ldots$, are well defined since they involve simple rvs, and $\lim_{n\to\infty}\mathbb{E}\left[X_n\right]$ is well defined by monotonicity (see Part (ii) of Lemma 10.4.1).

Pick $\varepsilon > 0$, and fix $n = 1, 2, \ldots$: Define the event $A_n(\varepsilon) \equiv [X_n \geq Y - \varepsilon]$. The rv X_n being non-negative it is plain that

$$\begin{aligned}
X_n &\geq X_n \cdot \mathbf{1}\left[A_n(\varepsilon)\right] \\
&\geq (Y - \varepsilon) \cdot \mathbf{1}\left[A_n(\varepsilon)\right]
\end{aligned}$$

$$
\begin{aligned}
&= Y \cdot \mathbf{1}[A_n(\varepsilon)] - \varepsilon \cdot \mathbf{1}[A_n(\varepsilon)] \\
&= Y - Y \cdot \mathbf{1}[A_n(\varepsilon)^c] - \varepsilon \cdot \mathbf{1}[A_n(\varepsilon)] \\
&\geq Y - C \cdot \mathbf{1}[A_n(\varepsilon)^c] - \varepsilon \qquad (10.16)
\end{aligned}
$$

where $C = \max_{k \in K} c_k$ if the simple rv Y has the representation $Y = \sum_{k\in} c_k \mathbf{1}[C_k]$ with $\mathcal{F}$-partition $\{C_k,\ k \in K\}$ and non-negative associated scalars $\{c_k,\ k \in K\}$.

Take expectations in (10.16) and note that all the rvs involved, namely X_n, Y and $\mathbf{1}[A_n(\varepsilon)^c]$, are all simple rvs. Therefore, using Lemma 10.4.1 repeatedly, we readily conclude that

$$
\mathbb{E}[X_n] \geq \mathbb{E}[Y] - C \cdot \mathbb{P}[A_n(\varepsilon)^c] - \varepsilon. \qquad (10.17)
$$

We now let n go to infinity in the inequality (10.17): The sequence $\{A_n(\varepsilon),\ n = 1, 2, \ldots\}$ is a monotonically increasing sequence of events (with $A_n(\varepsilon) \subseteq A_{n+1}(\varepsilon)$ for all $n = 1, 2, \ldots$ since $X_n \leq X_{n+1}$) such that

$$
\cup_{n=1,2,\ldots} A_n(\varepsilon) = \Omega.
$$

Using continuity from below of Lemma 3.1.1 we get $\lim_{n\to\infty} \mathbb{P}[A_n(\varepsilon)] = \mathbb{P}[\Omega] = 1$, or equivalently, $\lim_{n\to\infty} \mathbb{P}[A_n(\varepsilon)^c] = 0$, and the conclusion $\lim_{n\to\infty} \mathbb{E}[X_n] \geq \mathbb{E}[Y] - \varepsilon$ obtains. The desired result (10.11) follows upon observing that $\varepsilon > 0$ is arbitrary. ■

10.7 An alternate definition of $\mathbb{E}[X]$ when the rv X is non-negative

In this section we return to the case of non-negative rvs: Let $\{X_n,\ n = 1, 2, \ldots\}$ denote a(ny) staircase approximation for the non-negative rv $X : \Omega \to \mathbb{R}_+$. By Part (ii) of Lemma 10.4.1, the sequence $\{\mathbb{E}[X_n],\ n = 1, 2, \ldots\}$ being non-decreasing, the definitional limit $\lim_{n\to\infty} \mathbb{E}[X_n]$ exists with

$$
\mathbb{E}[X] \equiv \lim_{n\to\infty} \mathbb{E}[X_n] = \sup_{n=1,2,\ldots} \mathbb{E}[X_n]. \qquad (10.18)
$$

Since $\lim_{n\to\infty} X_n = X$ monotonically, it is also the case that

$$
X = \lim_{n\to\infty} X_n = \sup_{n=1,2,\ldots} X_n. \qquad (10.19)
$$

Combining these two simple observations we can rewrite (10.18) as an interchange of the supremum and integration operations, namely

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\sup_{n=1,2,\ldots} X_n\right] = \sup_{n=1,2,\ldots} \mathbb{E}\left[X_n\right]. \tag{10.20}$$

However, Lemma 10.4.3 implies that the quantity $\sup_{n=1,2,\ldots} \mathbb{E}\left[X_n\right]$ is independent of the sequence $\{X_n,\ n = 1, 2, \ldots\}$ used as a staircase approximation for the rv X. As a way to understand why this may occur consider the following arguments: With the non-negative rv $X : \Omega \to \mathbb{R}_+$, we associate the set $\mathcal{S}(X)$ of simple non-negative rvs which are bounded above by X, namely

$$\mathcal{S}(X) \equiv \left\{ Y : \Omega \to \mathbb{R}_+ : \begin{array}{c} \text{Simple non-negative rv} \\ Y \leq X \end{array} \right\}.$$

Lemma 10.3.1 ensures that $\mathcal{S}(X)$ is not empty, and the equality

$$X = \sup_{Y \in \mathcal{S}(X)} Y \tag{10.21}$$

is easily seen to hold [Exercise 10.4]. Comparing (10.19) and (10.21) it is then not unreasonable to expect (10.20) to generalize in the form of the following interchange:

Lemma 10.7.1 *For any non-negative rv $X : \Omega \to \mathbb{R}_+$, it holds that*

$$\mathbb{E}\left[\sup_{Y \in \mathcal{S}(X)} Y\right] = \sup_{Y \in \mathcal{S}(X)} \mathbb{E}\left[Y\right] \tag{10.22}$$

Proof. Set

$$\mathtt{Sup}(X) \equiv \sup\left\{\mathbb{E}\left[Y\right] :\ Y \in \mathcal{S}(X)\right\}.$$

Note that $\mathtt{Sup}(X)$ is well defined by Step 1 since the rvs in $\mathcal{S}(X)$ are all non-negative and simple. It is plain that (10.22) is equivalent to

$$\mathtt{Sup}(X) = \mathbb{E}\left[\sup_{Y \in \mathcal{S}(X)} Y\right] (= \mathbb{E}\left[X\right]) \tag{10.23}$$

where the last equality is a direct consequence of the observation (10.21)

Let $\{X_n,\ n = 1, 2, \ldots\}$ denote any collection of simple non-negative rvs which form a staircase approximation for X. Obviously, for each $n = 1, 2, \ldots$, the rv X_n is an element of $\mathcal{S}(X)$, whence $\mathbb{E}[X_n] \leq \mathtt{Sup}(X)$, and the conclusion $\mathbb{E}[X] \leq \mathtt{Sup}(X)$ follows by virtue of (10.18).

To establish the reverse inequality we proceed as follows: By the very definition of $\mathtt{Sup}(X)$ as a supremum, there exists a sequence of simple non-negative rvs $\{Y_n,\ n = 1, 2, \ldots\}$ in $\mathcal{S}(X)$ such that $\lim_{n\to\infty} \mathbb{E}[Y_n] = \mathtt{Sup}(X)$. The rvs $\{Y_n,\ n = 1, 2, \ldots\}$ may not form a monotone sequence, but the simple non-negative rvs $\{Z_n,\ n = 1, 2, \ldots\}$ defined by

$$Z_n = \max\left(X_n, \max\left(Y_1, \ldots, Y_n\right)\right), \quad n = 1, 2, \ldots$$

form a non-decreasing sequence in $\mathcal{S}(X)$. Noting that $X_n \leq Z_n \leq X$ for each $n = 1, 2, \ldots$, we conclude that $\lim_{n\to\infty} Z_n = X$ by the fact that the rvs $\{X_n,\ n = 1, 2, \ldots\}$ form a staircase approximation for X. Thus, the rvs $\{Z_n,\ n = 1, 2, \ldots\}$ also form a staircase approximation for X, and we have $\lim_{n\to\infty} \mathbb{E}[Z_n] = \mathbb{E}[X]$ by definition (10.10) (and Lemmas 10.4.3).

From the definitions we conclude that $\mathbb{E}[Y_n] \leq \mathbb{E}[Z_n]$ (since $Y_n \leq Z_n$) for each $n = 1, 2, \ldots$ with $\lim_{n\to\infty} \mathbb{E}[Y_n] = \mathtt{Sup}(X)$ by construction while we already have $\lim_{n\to\infty} \mathbb{E}[Z_n] = \mathbb{E}[X]$. Therefore, we have $\mathtt{Sup}(X) \leq \mathbb{E}[X]$ upon using the fact that $\lim_{n\to\infty} \mathbb{E}[Y_n] \leq \lim_{n\to\infty} \mathbb{E}[Z_n]$. This complete the proof of 10.23. ■

Lemma 10.7.1 provides an alternate definition for the expectation of non-negative rvs, namely

$$\mathbb{E}[X] \equiv \sup_{Y \in \mathcal{S}(X)} \mathbb{E}[Y] \tag{10.24}$$

for any non-negative rv X. While compact, this alternate definition is not constructive and therefore lacks any operational meaning for evaluating expectations.

10.8 Exercises

All rvs are defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 10.1 Prove Fact 10.2.1.

Ex. 10.2 By revisiting the proof of Lemma 10.3.1 show that the staircase approximation $\{X_n,\ n = 1, 2, \ldots\}$ whose existence is discussed there is not unique, i.e., show an alternative construction.

Ex. 10.3 Consider the simple rv $X : \Omega \to \mathbb{R}$, and assume it admits two representations, namely for $k = 1, 2$, it holds that

$$X = \sum_{i_k \in I_k} a_{k,i_k} \mathbf{1}\left[A_{i_k}\right]$$

for some finite $\mathcal{F}$-partition $\{A_{i_k},\ i_k \in I_k\}$ with associated collection $\{a_{k,i_k},\ i_k \in I_k\}$ of scalars in $\mathbb{R}$. Show by a direct argument that we necessarily have

$$\sum_{i_1 \in I_1} a_{1,i_1} \mathbb{P}\left[A_{i_1}\right] = \sum_{i_2 \in I_2} a_{2,i_2} \mathbb{P}\left[A_{i_2}\right]$$

so that the expectation $\mathbb{E}\left[X\right]$ of the simple rv X defined at (10.8) is independent of its representation.

Ex. 10.4 Establish (10.21).

Ex. 10.5 Let the rv $X : \Omega \to \mathbb{R}$ be a symmetric rv, i.e., $X =_{st} -X$.

a. Can $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$ assume different values?

b. Give conditions under which $\mathbb{E}\left[X\right]$ is well defined and finite, and show that $\mathbb{E}\left[X\right] = 0$ in that case.

c. Give an example of a symmetric rv X for which $\mathbb{E}\left[X\right]$ is not well defined.

Ex. 10.6 With positive scalar $M > 0$, the rv $X : \Omega \to \mathbb{R}$ is said to be M-bounded if $|X| \leq M$, i.e., $|X(\omega)| \leq M$ for all ω in Ω. Show that the expectation of an M-bounded rv X always exists and is finite with $|\mathbb{E}\left[X\right]| \leq M$.

Ex. 10.7 Consider two rvs $X, X' : \Omega \to \mathbb{R}$ with the property that $\mathbb{P}\left[X \neq X'\right] = 0$. Show that $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[X'\right]$ are both well defined simultaneously, in which case $\mathbb{E}\left[X\right] = \mathbb{E}\left[X'\right]$ (finite or not), or neither is well defined [**HINT:** Use Lemma 10.5.1].

Ex. 10.8 Let $X : \Omega \to \mathbb{R}$ be a rv with finite expectation, i.e., $\mathbb{E}\left[|X|\right] < \infty$.

a. If $X \geq 0$, show that $\lim_{n\to\infty} n\mathbb{P}\left[X \geq n\right] = 0$ (so that $\lim_{n\to\infty} n\mathbb{P}\left[X > n\right] = 0$ as well). [**HINT:** If $X \geq 0$, recall that the value $\mathbb{E}\left[X\right]$ does not depend on the approximating staircase sequence used in defining the expectation!]

b. What happens to this statement when X can take both positive or negative values?

c. If $\mathbb{E}\left[|X|^r\right] < \infty$ for some $r > 0$, show that $\lim_{n\to\infty} n^r\mathbb{P}\left[|X| > n\right] = 0$.

Ex. 10.9 Let $X : \Omega \to \mathbb{R}$ be a discrete rv such that $\mathbb{P}[X \in \mathbb{N}] = 1$. Using the fact that $X = \sum_{n=0}^{\infty} \mathbf{1}[X > n]$ and the Monotone Convergence Theorem shows that

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}[X > n]$$

regardless of whether $\mathbb{E}[X] < \infty$ or not. Compare with the proof suggested in Exercise **??**.

Ex. 10.10 Let $X : \Omega \to \mathbb{R}$ be a discrete rv such that $\mathbb{P}[X \in \mathbb{N}] = 1$. Show that $\mathbb{E}[X]$ can also be evaluated as

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}[X > n] = \sum_{n=1}^{\infty} \mathbb{P}[X \geq n]$$

regardless of whether $\mathbb{E}[X] < \infty$ or not. [**HINT:** Note that $\mathbb{P}[X = n] = \mathbb{P}[X \geq n] - \mathbb{P}[X \geq n+1]$ for each $n = 0, 1, \ldots$].

Ex. 10.11 Let $X : \Omega \to \mathbb{R}$ be a rv with finite expectation, i.e., $\mathbb{E}[|X|] < \infty$.

a. If $X \geq 0$, show that $\lim_{n\to\infty} n\mathbb{P}[X \geq n] = 0$ (so that $\lim_{n\to\infty} n\mathbb{P}[X > n] = 0$ as well). [**HINT:** If $X \geq 0$, recall that the value $\mathbb{E}[X]$ does not depend on the approximating staircase sequence used in defining the expectation!]

b. What happens to this statement when X can take both positive or negative values?

Ex. 10.12 With a rv $\xi : \Omega \to \mathbb{R}$, we define the rvs $X, Y, Z : \Omega \to \mathbb{R}$ given by $X \equiv \sin(\xi)$, $Y \equiv \frac{\xi}{1+\xi^2}$ and $Z \equiv \xi \cdot \cos(\xi)$.

a. For each of these three rvs, determine whether the expectation exists and whether it is finite if *no* additional assumption is imposed on the probability distribution function of ξ. In each case justify your answer!

In what follows, assume ξ to be a *symmetric* rv under $\mathbb{P}$ in the sense that the rvs ξ and $-\xi$ have the same probability distribution under $\mathbb{P}$.

b. Evaluate $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ – This can be done without any calculations and without knowing anything more about the rvs!

c. Give an example that shows that $\mathbb{E}[Z]$ may not always exist. Give an **additional** condition on ξ to ensure that the expectation $\mathbb{E}[Z]$ can be evaluated and find its value.

Chapter 11

Mathematical expectations (I): Basic properties

The present and next chapters are devoted to a discussion of useful properties of the expectation operator introduced in Chapter 10. The *basic* properties discussed in Chapter 11 are easy consequences of the three step definition of expectation given in Section 10.4. Proofs have been included for the sake of completeness; they are straightforward, albeit at times tedious, and can be omitted in a first reading.

Throughout we are given rvs all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

11.1 Basic properties (I)

A. Mutiplying by a constant

If $\mathbb{E}[X]$ exists, then for each c in $\mathbb{R}$, $\mathbb{E}[cX]$ also exists and it holds that $\mathbb{E}[cX] = c\mathbb{E}[X]$.

The desired conclusion is clearly true for simple rvs – See the first part of Lemma 10.4.1 with $a = c$ and $b = 0$.

If X is a non-negative rv, let the rvs $\{X_n,\ n = 1, 2, \ldots\}$ be the simple non-negative rvs associated with X in Lemma 10.3.1, so that $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$ by definition. For each c in $\mathbb{R}$, the rvs $\{cX_n,\ n = 1, 2, \ldots\}$ are also simple rvs and $\mathbb{E}[cX_n] = c\mathbb{E}[X_n]$ for all $n = 1, 2, \ldots$ by the first part of the proof.

If $c \geq 0$ the rvs $\{cX_n,\ n = 1, 2, \ldots\}$ are non-negative and non-decreasing with $\lim_{n\to\infty} cX_n = cX$ pointwise. Therefore, using the definition for non-negative rvs, we get

$$\begin{aligned}\mathbb{E}\left[cX\right] &= \lim_{n\to\infty}\mathbb{E}\left[cX_n\right] \\ &= \lim_{n\to\infty}c\mathbb{E}\left[X_n\right] \\ &= c\lim_{n\to\infty}\mathbb{E}\left[X_n\right] = c\mathbb{E}\left[X\right].\end{aligned}$$

If $c < 0$, then $(cX)^+ = 0$ and $(cX)^- = -cX = |c|X$. As a result, $\mathbb{E}\left[(cX)^+\right] = 0$ and $\mathbb{E}\left[(cX)^-\right] = \mathbb{E}\left[|c|X\right] = |c|\mathbb{E}\left[X\right]$ by the earlier part of the proof, and we conclude $\mathbb{E}\left[X\right] = -\mathbb{E}\left[(cX)^-\right] = -|c|\mathbb{E}\left[X\right] = c\mathbb{E}\left[X\right]$ as desired.

For the general case, first consider $c > 0$. Noting that $(cX)^+ = cX^+$ and $(cX)^- = cX^-$, we obtain $\mathbb{E}\left[(cX)^+\right] = \mathbb{E}\left[cX^+\right] = c\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[(cX)^-\right] = \mathbb{E}\left[cX^-\right] = c\mathbb{E}\left[X^-\right]$ by the result for non-negative rvs. Therefore, $\mathbb{E}\left[cX\right]$ is well defined as soon as $\mathbb{E}\left[X\right]$ is well defined with $\mathbb{E}\left[cX\right] = \mathbb{E}\left[(cX)^+\right] - \mathbb{E}\left[(cX)^-\right] = c\mathbb{E}\left[X^+\right] - c\mathbb{E}\left[X^-\right] = c\mathbb{E}\left[X\right]$. The case $c < 0$ is handled *mutatis mutandi* and is left to the interested reader. ■

B. Monotonicity

If $X \leq Y$, then $\mathbb{E}\left[X\right] \leq \mathbb{E}\left[Y\right]$ as soon as both $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[Y\right]$ exist (possibly infinite). In particular, (i) if $-\infty < \mathbb{E}\left[X\right]$, then $-\infty < \mathbb{E}\left[Y\right]$ and $\mathbb{E}\left[X\right] \leq \mathbb{E}\left[Y\right]$, or (ii) if $\mathbb{E}\left[Y\right] < \infty$, then $\mathbb{E}\left[X\right] < \infty$ and $\mathbb{E}\left[X\right] \leq \mathbb{E}\left[Y\right]$.

We start with X and Y being both simple rvs, in which case the desired result is simply Part (ii) of Lemma 10.4.1.

Next we consider the case when the rvs X and Y are non-negative rvs satisfying $X \leq Y$. Let rvs $\{X_n,\ n = 1, 2, \ldots\}$ (resp. $\{Y_n,\ n = 1, 2, \ldots\}$) form a staircase approximation for the rv X (resp. Y). The rvs $\{\max(X_n, Y_n),\ n = 1, 2, \ldots\}$ are simple non-negative rv which also form a staircase approximation for the rv Y: Indeed, since $\lim_{n\to\infty} X_n = X$ and $\lim_{n\to\infty} Y_n = Y$ by construction, it holds that

$$\lim_{n\to\infty}\max(X_n, Y_n) = \max(\lim_{n\to\infty}X_n, \lim_{n\to\infty}Y_n) = \max(X, Y) = Y,$$

while the monotonicity of the staircase approximations for X and Y also yields $\max(X_n, Y_n) \leq \max(X_{n+1}, Y_{n+1})$ for each $n = 1, 2, \ldots$. In particular, $\mathbb{E}\left[Y\right] = \lim_{n\to\infty}\mathbb{E}\left[\max(X_n, Y_n)\right]$ by the usual construction (independent of the staircase approximation used). However, we note that $X_n \leq \max(X_n, Y_n)$ for each $n = 1, 2, \ldots$, whence $\mathbb{E}\left[X_n\right] \leq \mathbb{E}\left[\max(X_n, Y_n)\right]$ by the first part of the proof. Letting

n go to infinity we readily conclude that $\mathbb{E}[X] \leq \mathbb{E}[Y]$. The desired result holds for non-negative rvs.

Finally we turn to arbitrary rvs X and Y such that $X \leq Y$. By direct inspection we have $X^+ \leq Y^+$ and $Y^- \leq X^-$, whence $\mathbb{E}[X^+] \leq \mathbb{E}[Y^+]$ and $\mathbb{E}[Y^-] \leq \mathbb{E}[X^-]$ since monotonicity of expectations was shown to always hold for non-negative rvs. The desired conclusion

$$\mathbb{E}[X] = \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right] \leq \mathbb{E}\left[Y^+\right] - \mathbb{E}\left[Y^-\right] = \mathbb{E}[Y]$$

follows (under the usual caveat that the expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are well defined). ■

C. Taking absolute values

If $\mathbb{E}[X]$ exists, then $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.

Note that $-|X| \leq X \leq |X|$ and apply Property **B** twice, once to $-|X| \leq X$ and once to $X \leq |X|$. ■

D. Localization

If $\mathbb{E}[X]$ exists, then $\mathbb{E}[X\mathbf{1}[A]]$ exists for any event A in $\mathcal{F}$. Furthermore, if $\mathbb{E}[X]$ is finite, then $\mathbb{E}[X\mathbf{1}[A]]$ is also finite.

For any A in $\mathcal{F}$, introduce the rv $X_A = X\mathbf{1}[A]$. We have $0 \leq X_A^\pm = X^\pm \mathbf{1}[A]$ by direct inspection so that $X_A^\pm \leq X^\pm$. Obviously $\mathbb{E}\left[X_A^\pm\right] \leq \mathbb{E}[X^\pm]$ by Property **B**, whence $\min\left(\mathbb{E}\left[X_A^+\right], \mathbb{E}\left[X_A^-\right]\right) \leq \min\left(\mathbb{E}[X^+], \mathbb{E}[X^-]\right)$ and $\mathbb{E}\left[X_A^+\right] + \mathbb{E}\left[X_A^-\right] \leq \mathbb{E}[X^+] + \mathbb{E}[X^-]$. The conclusions are now straightforward from Definition 10.4.1 as $\min\left(\mathbb{E}[X^+], \mathbb{E}[X^-]\right) < \infty$ (resp. $\mathbb{E}[X^+] + \mathbb{E}[X^-] < \infty$) implies $\min\left(\mathbb{E}\left[X_A^+\right], \mathbb{E}\left[X_A^-\right]\right) < \infty$ (resp. $\mathbb{E}\left[X_A^+\right] + \mathbb{E}\left[X_A^-\right] < \infty$). ■

E. Adding rvs

We have $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ if (i) the rvs X and Y are non-negative or (ii) if $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are both finite.

We start with X and Y being both simple rvs, in which case the desired result, namely $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ was already established as Part (i) of Lemma 10.4.1 with $a = b = 1$.

Next we consider the case when the rvs X and Y are non-negative rvs, and let the rvs $\{X_n,\ n = 1, 2, \ldots\}$ (resp. $\{Y_n,\ n = 1, 2, \ldots\}$) form a staircase approximation for the rv X (resp. Y). The simple rvs $\{X_n + Y_n,\ n = 1, 2, \ldots\}$ also form a staircase approximation for the rv $X+Y$, hence $\mathbb{E}[X + Y] = \lim_{n\to\infty} \mathbb{E}[X_n + Y_n]$. By the first part of the proof we have $\mathbb{E}[X_n + Y_n] = \mathbb{E}[X_n] + \mathbb{E}[Y_n]$ for each $n = 1, 2, \ldots$ and letting n go to infinity in these equalities we conclude that $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ as we recall that $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$, and $\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[Y_n]$ by construction.

Finally, we turn to the case when X and Y are arbitrary rvs with $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ both finite, or equivalently, $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$. Decomposing each of the rvs X, Y and $X + Y$ according to (10.13) we find

$$X + Y = \left(X^+ - X^-\right) + \left(Y^+ - Y^-\right)$$

as well as

$$X + Y = (X + Y)^+ - (X + Y)^- .$$

Combining these two expressions for the sum $X + Y$ and rearranging terms, we obtain

$$X^+ + Y^+ + (X + Y)^- = X^- + Y^- + (X + Y)^+ .$$

Taking expectations on both sides of this last relationship between non-negative rvs yields

$$\mathbb{E}\left[X^+ + Y^+ + (X + Y)^-\right] = \mathbb{E}\left[X^- + Y^- + (X + Y)^+\right] \tag{11.1}$$

where each of these expectations can be expressed as

$$\mathbb{E}\left[X^+ + Y^+ + (X + Y)^-\right] = \mathbb{E}\left[X^+\right] + \mathbb{E}\left[Y^+\right] + \mathbb{E}\left[(X + Y)^-\right]$$

and

$$\mathbb{E}\left[X^- + Y^- + (X + Y)^+\right] = \mathbb{E}\left[X^-\right] + \mathbb{E}\left[Y^-\right] + \mathbb{E}\left[(X + Y)^+\right]$$

upon using the first part of the proof – Indeed *all* the rvs involved are non-negative, so all the expectations exist and additivity holds.

Returning to (11.1) we conclude to the equality

$$\begin{aligned} &\mathbb{E}\left[X^+\right] + \mathbb{E}\left[Y^+\right] + \mathbb{E}\left[(X + Y)^-\right] \\ &= \ \mathbb{E}\left[X^-\right] + \mathbb{E}\left[Y^-\right] + \mathbb{E}\left[(X + Y)^+\right] . \end{aligned} \tag{11.2}$$

Note that $|X+Y| \le |X|+|Y|$, whence $\mathbb{E}[|X+Y|] \le \mathbb{E}[|X|]+\mathbb{E}[|Y|]$ by Property **B** and again by the first part of the proof. Therefore, under the assumption that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$, we have $\mathbb{E}[|X+Y|] < \infty$ with both $\mathbb{E}\left[(X+Y)^+\right]$ and $\mathbb{E}\left[(X+Y)^-\right]$ being finite. The representation

$$\mathbb{E}[X+Y] = \mathbb{E}\left[(X+Y)^+\right] - \mathbb{E}\left[(X+Y)^-\right]$$

thus holds with $\mathbb{E}[X+Y]$ finite, and (11.2) can now be rewritten as

$$\begin{aligned} &\mathbb{E}\left[(X+Y)^+\right] - \mathbb{E}\left[(X+Y)^-\right] \\ &\quad = \ \mathbb{E}\left[X^+\right] + \mathbb{E}\left[Y^+\right] - \mathbb{E}\left[X^-\right] - \mathbb{E}\left[Y^-\right]. \end{aligned}$$

In other words, $\mathbb{E}[X+Y] = \mathbb{E}[X]+\mathbb{E}[Y]$ as desired. ■

Extensions are discussed in Exercises 11.2 and 11.3.

11.2 Basic properties (II)

The next group of properties will make use of the following notion: We consider situations where a property P may or not hold for every sample ω in Ω. We shall say that property P holds *almost surely* (under $\mathbb{P}$) if the event

$$\{\omega \in \Omega : \text{ Property } P \text{ holds at } \omega\}$$

has probability one. We shall often write P holds a.s. or P holds $\mathbb{P}$-a.s. when we wish to emphasize the fact that relevant probabilities are evaluated under $\mathbb{P}$. For instance, for rvs $X, Y : \Omega \to \mathbb{R}$, we write $X = Y$ a.s. (resp. $X \le Y$ a.s.) to express the fact that $\mathbb{P}[X=Y]=1$ (resp. $\mathbb{P}[X \le Y]=1$).

F.

If $X = 0$ a.s., then $\mathbb{E}[X]$ is well defined with $\mathbb{E}[X]=0$.

First assume the rv X to be simple with $X = \sum_{i \in I} a_i \mathbf{1}[A_i]$ for some finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ and a collection $\{a_i,\ i \in I\}$ of scalars in $\mathbb{R}$. The condition $X = 0$ a.s. implies $\mathbb{P}[A_i]=0$ whenever $a_i \neq 0$ [Exercise 11.4], whence $\mathbb{E}[X] = \sum_{i \in I} a_i \mathbb{P}[A_i] = 0$.

If $X \ge 0$, then any staircase approximation $\{X_n,\ n=1,2,\ldots\}$ satisfies $0 \le X_n \le X$ for all $n = 1,2,\ldots$, and the constraint $X=0$ a.s. implies $X_n = 0$ a.s. for all $n=1,2,\ldots$, whence $\mathbb{E}[X_n]=0$ by the first part of the proof. Therefore, $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] = 0$ by the definition of $\mathbb{E}[X]$.

For arbitrary rv X, note that $X^{\pm} = 0$ a.s. if $X = 0$ a.s., whence $\mathbb{E}[X^{\pm}] = 0$ and $\mathbb{E}[X] = \mathbb{E}[X^{+}] - \mathbb{E}[X^{-}] = 0$. ■

Another proof of this result uses the fact outlined in Exercise 10.7: Recall that $\mathbb{E}[X] = c$ if $X \equiv c$ for some scalar c, whence $\mathbb{E}[X] = c$ if $X = c$ a.s.

G. Almost sure (a.s.) equality

If $X = Y$ a.s. with $\mathbb{E}[|X|] < \infty$, then $\mathbb{E}[|Y|] < \infty$ and $\mathbb{E}[X] = \mathbb{E}[Y]$.

Write

$$E \equiv \{\omega \in \Omega : \ X(\omega) = Y(\omega)\}.$$

Recall that $X = X\mathbf{1}[E] + X\mathbf{1}[E^c]$ and $Y = Y\mathbf{1}[E] + Y\mathbf{1}[E^c]$, so that

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X\mathbf{1}[E] + X\mathbf{1}[E^c]] \\ &= \mathbb{E}[X\mathbf{1}[E]] + \mathbb{E}[X\mathbf{1}[E^c]] \quad \text{[By Property }\mathbf{D}\text{ and Property }\mathbf{E}] \\ &= \mathbb{E}[Y\mathbf{1}[E]] + \mathbb{E}[X\mathbf{1}[E^c]] \quad [\text{Since } X = Y \text{ on } E] \\ &= \mathbb{E}[Y\mathbf{1}[E]] \quad \text{[By Property }\mathbf{F}] \end{aligned}$$

as we note that $X\mathbf{1}[E^c] = 0$ a.s. But it is also the case that $Y\mathbf{1}[E^c] = 0$ a.s., hence $\mathbb{E}[Y\mathbf{1}[E^c]] = 0$ by Property **F** and we conclude that $\mathbb{E}[X] = \mathbb{E}[Y\mathbf{1}[E]] + \mathbb{E}[Y\mathbf{1}[E^c]]$. These arguments applied to $|X|$ and $|Y|$ (instead of X and Y) also show that $\mathbb{E}[|X|] = \mathbb{E}[|Y|\mathbf{1}[E]] + \mathbb{E}[|Y|\mathbf{1}[E^c]] = \mathbb{E}[|Y|]$ by Property **E**, hence $\mathbb{E}[|Y|] < \infty$. It follows that $\mathbb{E}[X] = \mathbb{E}[Y\mathbf{1}[E]] + \mathbb{E}[Y\mathbf{1}[E^c]] = \mathbb{E}[Y]$ by Property **D** and Property **E**. ■

Another proof of this result is outlined in Exercise 10.7.

H.

If $X \geq 0$ with $\mathbb{E}[X] = 0$, then $X = 0$ a.s.

Consider the sets $E \equiv \{\omega \in \Omega : \ X(\omega) > 0\}$ and

$$E_n \equiv \left\{\omega \in \Omega : \ X(\omega) \geq \frac{1}{n}\right\}, \quad n = 1, 2, \ldots$$

These events clearly belong to $\mathcal{F}$. We need to establish that $\mathbb{P}[E] = 0$.

For each $n = 1, 2, \ldots$, define the rv $X_n \equiv X\mathbf{1}[E_n]$. It is plain that $0 \leq X_n \leq X$ so that $0 \leq \mathbb{E}[X_n] \leq \mathbb{E}[X]$ by Property **B**. Fix $n = 1, 2, \ldots$ The assumption

$\mathbb{E}[X] = 0$ implies $\mathbb{E}[X_n] = 0$, and the obvious inequalities $0 \leq \frac{1}{n}\mathbf{1}[E_n] \leq X_n$ then yield

$$0 \leq \frac{1}{n}\mathbb{P}[E_n] \leq \mathbb{E}[X_n] = 0$$

upon using Property **B** again, whence $\mathbb{P}[E_n] = 0$. Finally, the sequence of events $\{E_n,\ n = 1, 2, \ldots\}$ being increasing with $E = \cup_{n=1}^{\infty} E_n$, it follows that $\mathbb{P}[E] = \lim_{n\to\infty} \mathbb{P}[E_n] = 0$ by invoking continuity from below of Lemma 3.1.1. ■

I.

Assume $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$. If $\mathbb{E}[X\mathbf{1}[A]] \leq \mathbb{E}[Y\mathbf{1}[A]]$ for all A in $\mathcal{F}$, then $X \leq Y$ a.s.

By Property **D**, the condition $\mathbb{E}[|X|] < \infty$ (resp. $\mathbb{E}[|Y|] < \infty$) implies the finiteness of $\mathbb{E}[X\mathbf{1}[A]]$ (resp. $\mathbb{E}[Y\mathbf{1}[A]]$) for all A in $\mathcal{F}$. Define the event B by

$$B \equiv \{\omega \in \Omega :\ Y(\omega) < X(\omega)\}.$$

It is plain that $Y\mathbf{1}[B] \leq X\mathbf{1}[B]$ whence $\mathbb{E}[Y\mathbf{1}[B]] \leq \mathbb{E}[X\mathbf{1}[B]]$ by Property **B**, while $\mathbb{E}[X\mathbf{1}[B]] \leq \mathbb{E}[Y\mathbf{1}[B]]$ by assumption, and the conclusion $\mathbb{E}[X\mathbf{1}[B]] = \mathbb{E}[Y\mathbf{1}[B]]$ follows, or equivalently, $\mathbb{E}[(X - Y)\mathbf{1}[B]] = 0$. But $(X - Y)\mathbf{1}[B] \geq 0$ and Property **H** yields $(X - Y)\mathbf{1}[B] = 0$ a.s.

With $A \equiv [(X - Y)\mathbf{1}[B] = 0]$, pick ω in A so that $(X(\omega) - Y(\omega))\mathbf{1}[B](\omega) = 0$. If ω also lies in B, then $\mathbf{1}[B](\omega) = 1$ and the equality $X(\omega) = Y(\omega)$ follows. On the other hand, we also have $X(\omega) - Y(\omega) > 0$ by the definition of B, and a contradiction occurs. Thus, $A \cap B = \emptyset$ or equivalently, $A \subseteq B^c$, whence $\mathbb{P}[A] \leq \mathbb{P}[B^c]$ with $\mathbb{P}[A] = 1$. In fine, $\mathbb{P}[B^c] = 1$ and the conclusion $X \leq Y$ a.s. follows. ■

J. Extended rvs

For any extended rv $X : \Omega \to [-\infty, \infty]$, the condition $\mathbb{E}[|X|] < \infty$ implies $|X| < \infty$ a.s.

With $A \equiv \{\omega \in \Omega :\ |X(\omega)| = \infty\}$, assume that $\mathbb{P}[A] > 0$. Then, by Property **D** we have $\mathbb{E}[|X|\mathbf{1}[A]] \leq \mathbb{E}[|X|]$. But $\infty \cdot \mathbb{P}[A] \leq \mathbb{E}[|X|\mathbf{1}[A]]$ while $\mathbb{E}[|X|] < \infty$ by assumption, and a contradiction follows. Thus, $\mathbb{P}[A] = 0$ necessarily. ■

11.3 Simple variables vs. discrete rvs

The notion of simple rv is a set-theoretic one, as it requires only the existence of the measurable space $(\Omega, \mathcal{F})$ on which it is defined. On the other hand, defining discrete rvs requires the existence of a probability measure $\mathbb{P}$ on the underlying measurable space $(\Omega, \mathcal{F})$. While a simple rv is always a discrete rv (with finite support), a discrete rv (even with finite support) is not necessarily a simple rv. This is made clear by the following example.

Example 11.3.1 Take $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$ and with $a < b$ in $[0,1]$, define the probability measure $\mathbb{P}$ on $\mathcal{F}$ by setting

$$\mathbb{P}[E] = \frac{|E \cap \{a,b\}|}{2}, \quad E \in \mathcal{F}.$$

The rv $X : \Omega \to \mathbb{R} : \omega \to \omega$ is not a simple rv since $X(\Omega) = [0,1]$ but it is definitely a discrete rv with support $S = \{a,b\}$ since $\mathbb{P}[X \in S] = 1$ by the definition of $\mathbb{P}$.

The next result shows that the evaluation of the expectation of a discrete rv can be carried out by using the expression presented in the elementary treatment of Probability Theory.

Proposition 11.3.1 *If $X : \Omega \to \mathbb{R}$ is a discrete rv with (countable) support S, then*

$$\mathbb{E}[X] = \sum_{x \in S} x \cdot \mathbb{P}[X = x] \tag{11.3}$$

if either the rv X is a.s. non-negative (with $S \subseteq \mathbb{R}_+$) or if the absolute summability condition

$$\sum_{x \in S} |x| \cdot \mathbb{P}[X = x] < \infty \tag{11.4}$$

holds.

Proof. Assume first that the set S contains only finitely many elements. The rv $X^\star : \Omega \to \mathbb{R}$ defined by

$$X^\star \equiv \sum_{x \in S} x \cdot \mathbf{1}[X = x] + 0 \cdot \mathbf{1}[X \notin S]$$

is a simple rv with expectation given by

$$\mathbb{E}[X^\star] = \sum_{x \in S} x \cdot \mathbb{P}[X = x].$$

Note that $X = X^\star$ on $[X \in S]$, hence $X = X^\star$ a.s.. Using Property **G**, we conclude that $\mathbb{E}[|X|] < \infty$ since $\mathbb{E}[|X^\star|] < \infty$ and the equality $\mathbb{E}[X] = \mathbb{E}[X^\star]$ follows, hence

$$\mathbb{E}[X] = \sum_{x \in S} x \cdot \mathbb{P}[X = x] .$$

Assume now that S is countably infinite. If $S \subseteq \mathbb{R}_+$, then $X = X\mathbf{1}[X \in S]$ a.s. with $X\mathbf{1}[X \in S] \geq 0$. Introduce the simple non-negative rvs $\{X_n,\ n = 1, 2, \ldots\}$ given by

$$X_n \equiv \sum_{\ell=1}^{n} x_\ell \cdot \mathbf{1}[X = x_\ell], \quad n = 1, 2, \ldots$$

where $\{x_\ell,\ \ell = 1, 2, \ldots\}$ is a labeling of S. By the first part of the proof we have

$$\mathbb{E}[X_n] = \sum_{\ell=1}^{n} x_\ell \cdot \mathbb{P}[X = x_\ell], \quad n = 1, 2, \ldots$$

The sequence $\{X_n,\ n = 1, 2, \ldots\}$ is monotone increasing with $\lim_{n \to \infty} X_n = X\mathbf{1}[X \in S]$, and is therefore a staircase approximation for the rv $X\mathbf{1}[X \in S]$ (albeit not necessarily the one presented in Lemma 10.3.1). The constructive definition of $\mathbb{E}[X\mathbf{1}[X \in S]]$ then yields

$$\begin{aligned}
\mathbb{E}[X\mathbf{1}[X \in S]] &= \lim_{n \to \infty} \mathbb{E}[X_n] \\
&= \lim_{n \to \infty} \left(\sum_{\ell=1}^{n} x_\ell \cdot \mathbb{P}[X = x_\ell] \right) \\
&= \sum_{\ell=1}^{\infty} x_\ell \cdot \mathbb{P}[X = x_\ell] \\
&= \sum_{x \in S} x \cdot \mathbb{P}[X = x]
\end{aligned}$$

where in the equality before last the series converges (possibly to $+\infty$) by monotonicity since $S \subseteq \mathbb{R}_+$. Using Property **G**, if $\mathbb{E}[X\mathbf{1}[X \in S]] < \infty$, then $\mathbb{E}[|X|] < \infty$ and we conclude

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}[X \in S]] = \sum_{x \in S} x \cdot \mathbb{P}[X = x] .$$

Finally, for an arbitrary discrete rv X, it is plain that X^+ and X^- are both discrete rvs with $\mathbb{P}[X^\pm \in S_\pm] = 1$ where $S^+ = \{x \in S : x \geq 0\}$ and $S^- =$

$\{x \in S : \ x \leq 0\}$, respectively. The previous discussion yields

$$\mathbb{E}\left[X^{\pm}\right] = \sum_{x \in S_{\pm}} (\pm x) \cdot \mathbb{P}\left[X = x\right],$$

whence

$$\begin{aligned} \mathbb{E}\left[X\right] &= \mathbb{E}\left[X^{+}\right] - \mathbb{E}\left[X^{-}\right] \\ &= \sum_{x \in S: \ x \geq 0} x \cdot \mathbb{P}\left[X = x\right] - \sum_{x \in S: \ x \leq 0} (-x) \cdot \mathbb{P}\left[X = x\right] \\ &= \sum_{x \in S} x \cdot \mathbb{P}\left[X = x\right] \qquad (11.5) \end{aligned}$$

where the last step is justified under the condition (11.4). ■

11.4 Exercises

Ex. 11.1 Use the alternate definition (10.24) for the expectation of non-negative rvs to establish Properties **A** and **B**.

Ex. 11.2 Regarding Property **E**, explain why $\mathbb{E}\left[X + Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$ may fail to hold for rvs X and Y whose expectations $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[Y\right]$ are well defined but possibly infinite.

Ex. 11.3 Generalizing Property **E**: Show that $\mathbb{E}\left[X + Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$ still holds for rvs X and Y for which $\mathbb{E}\left[X^{+}\right] + \mathbb{E}\left[Y^{+}\right] = \infty$ and $\mathbb{E}\left[X^{-}\right] + \mathbb{E}\left[Y^{-}\right] < \infty$ (resp. $\mathbb{E}\left[X^{-}\right] + \mathbb{E}\left[Y^{-}\right] = \infty$ and $\mathbb{E}\left[X^{+}\right] + \mathbb{E}\left[Y^{+}\right] < \infty$). What is the value of $\mathbb{E}\left[X + Y\right]$?

Ex. 11.4 Consider a simple rv X with $X = \sum_{i \in I} a_i \mathbf{1}\left[A_i\right]$ for some finite $\mathcal{F}$-partition $\{A_i, \ i \in I\}$ and a collection $\{a_i, \ i \in I\}$ of scalars in $\mathbb{R}$. Show that the condition $X = 0$ a.s. implies $\mathbb{P}\left[A_i\right] = 0$ whenever $a_i \neq 0$ [**HINT:** Make use of the set $\Omega_0 \equiv [X = 0]$].

Ex. 11.5 Compute the expectation

$$\mathbb{E}\left[\frac{1}{1 + Y^{+}}\right]$$

when the rv $Y : \Omega \rightarrow \mathbb{R}$ is

a. a binomial rv $\mathrm{Bin}(n;p)$ with $n = 1, 2, \ldots$ and $0 < p < 1$,
b. a Poisson rv $\mathrm{Poi}(\lambda)$ with $\lambda > 0$,
c. a geometric rv $\mathrm{Geo}(p)$ with $0 < p < 1$,

In each case explain why the expectation $\mathbb{E}\left[\frac{1}{1+Y^+}\right]$ always exists.

Chapter 12

Mathematical expectations (II): Advanced properties

12.1 Independence and expectations

The next fact is used in many calculations. It highlights the usefulness of independence when evaluating the expectation of expressions formed through products of independent rvs.

Proposition 12.1.1 *Consider two independent rvs $X, Y : \Omega \to \mathbb{R}$. It holds*

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{12.1}$$

if either (i) the rvs are a.s. non-negative or (ii) both expectations $\mathbb{E}\left[X\right]$ and $\mathbb{E}\left[Y\right]$ exist and are finite.

Proof. Assume first that both rvs X and Y are simple rvs, say

$$X = \sum_{i\in I} a_i \mathbf{1}\left[A_i\right] \quad \text{and} \quad Y = \sum_{j\in J} b_j \mathbf{1}\left[B_j\right]$$

with a finite $\mathcal{F}$-partition $\{A_i,\ i \in I\}$ with associated collection $\{a_i,\ i \in I\}$ of scalars in $\mathbb{R}$, and a finite $\mathcal{F}$-partition $\{B_j,\ j \in J\}$ with associated collection $\{b_j,\ j \in J\}$ of scalars in $\mathbb{R}$. There is no loss of generality in assuming that the scalars $\{a_i,\ i \in I\}$ (resp. $\{b_j,\ j \in J\}$) are distinct so that $[X = a_i] = A_i$ for each i in I, and $[Y = b_j] = B_j$ for each j in J. The rvs X and Y being independent, it follows that

$$\mathbb{P}\left[A_i \cap B_j\right] = \mathbb{P}\left[A_i\right]\mathbb{P}\left[B_j\right], \quad i \in I, j \in J$$

since the events $[X = a_i]$ and $[Y = b_j]$ are independent.

Noting that

$$XY = \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbf{1}[A_i] \mathbf{1}[B_j] = \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbf{1}[A_i \cap B_j],$$

we conclude that

$$\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}\left[\sum_{i \in I} \sum_{j \in J} a_i b_j \mathbf{1}[A_i \cap B_j]\right] \\ &= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{E}[\mathbf{1}[A_i \cap B_j]] \\ &= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{P}[A_i \cap B_j] \\ &= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{P}[A_i] \mathbb{P}[B_j] \\ &= \left(\sum_{i \in I} a_i \mathbb{P}[A_i]\right) \left(\sum_{j \in J} b_j \mathbb{P}[B_j]\right) \\ &= \mathbb{E}[X] \mathbb{E}[Y]. \end{aligned}$$

Next we assume that both rvs X and Y are non-negative, so that XY is also a non-negative rv. Let $\{X_n,\ n = 1, 2, \ldots\}$ and $\{Y_n,\ n = 1, 2, \ldots\}$ denote the monotone non-negative staircase approximations of X and Y identified in Lemma 10.3.1. Note that the rvs $\{X_n Y_n,\ n = 1, 2, \ldots\}$ form a monotone sequence of staircase approximations for the rv XY since

$$0 \leq X_n Y_n \leq X_{n+1} Y_n \leq X_{n+1} Y_{n+1}, \quad n = 1, 2, \ldots$$

by the non-negativity of the rvs involved, and by the monotone nature of each sequence. For each $n = 1, 2, \ldots$, the rvs X_n and Y_n are independent rvs since $X_n = g_n(X)$ and $Y_n = g_n(Y)$ with Borel mapping $g_n : \mathbb{R} \to \mathbb{R}$ defined at (10.7) – See the construction in the proof of Lemma 10.3.1. Obviously, $\lim_{n \to \infty} X_n Y_n = (\lim_{n \to \infty} X_n) (\lim_{n \to \infty} Y_n) = XY$, whence

$$\begin{aligned} \mathbb{E}[XY] &= \lim_{n \to \infty} \mathbb{E}[X_n Y_n] \quad \text{[By the definition of } \mathbb{E}[XY]\text{]} \\ &= \lim_{n \to \infty} (\mathbb{E}[X_n] \mathbb{E}[Y_n]) \quad \text{[By independence and the first part of the proof]} \\ &= \left(\lim_{n \to \infty} \mathbb{E}[X_n]\right) \left(\lim_{n \to \infty} \mathbb{E}[Y_n]\right) \\ &= \mathbb{E}[X] \mathbb{E}[Y]. \quad \text{[By the definition of } \mathbb{E}[X] \text{ and } \mathbb{E}[Y]\text{]} \end{aligned}$$

It follows from this proof that $\mathbb{E}[XY]$ is finite if and only if both expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are finite.

For the general case, start with the decompositions $X = X^+ - X^-$ and $Y = Y^+ - Y^-$, and note that

$$\begin{aligned} XY &= \left(X^+ - X^-\right)\left(Y^+ - Y^-\right) \\ &= X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-. \end{aligned} \tag{12.2}$$

The $\mathbb{R}^2_+$-valued rvs (X^+, X^-) and (Y^+, Y^-) are independent, a fact inherited from the independence of the rvs X and Y. If the expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are both finite, then the expectation $\mathbb{E}[X^\pm]$ and $\mathbb{E}[Y^\pm]$ are all finite, whence by the earlier part of the proof (for non-negative rvs) the expectations $\mathbb{E}[X^+Y^+]$, $\mathbb{E}[X^+Y^-]$, $\mathbb{E}[X^-Y^+]$ and $\mathbb{E}[X^-Y^-]$ are all finite and given by $\mathbb{E}[X^+]\mathbb{E}[Y^+]$, $\mathbb{E}[X^+]\mathbb{E}[Y^-]$, $\mathbb{E}[X^-]\mathbb{E}[Y^+]$ and $\mathbb{E}[X^-]\mathbb{E}[Y^-]$, respectively. Thus, by Property **E** we get

$$\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}\left[X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-\right] \\ &= \mathbb{E}\left[X^+Y^+\right] - \mathbb{E}\left[X^+Y^-\right] - \mathbb{E}\left[X^-Y^+\right] + \mathbb{E}\left[X^-Y^-\right] \\ &= \mathbb{E}\left[X^+\right]\mathbb{E}\left[Y^+\right] - \mathbb{E}\left[X^+\right]\mathbb{E}\left[Y^-\right] \\ &\quad - \mathbb{E}\left[X^-\right]\mathbb{E}\left[Y^+\right] + \mathbb{E}\left[X^-\right]\mathbb{E}\left[Y^-\right] \\ &= \left(\mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right]\right)\left(\mathbb{E}\left[Y^+\right] - \mathbb{E}\left[Y^-\right]\right) \\ &= \mathbb{E}[X]\,\mathbb{E}[Y] \end{aligned} \tag{12.3}$$

as announced. ■

Proposition 12.1.1 has the following often used consequence.

Lemma 12.1.1 *Consider the mutually independent rvs $X_1 : \Omega \to \mathbb{R}^{p_1}$, ..., $X_k : \Omega \to \mathbb{R}^{p_k}$. With Borel mappings $g_1 : \mathbb{R}^{p_1} \to \mathbb{R}$, ..., $g_k : \mathbb{R}^{p_k} \to \mathbb{R}$, define the rvs*

$$Y_\ell = g_\ell(X_\ell), \quad \ell = 1, \ldots, k.$$

The $\mathbb{R}$-valued rvs $Y_1, \ldots, Y_k$ are mutually independent, and

$$\mathbb{E}\left[\prod_{\ell=1}^k Y_\ell\right] = \prod_{\ell=1}^k \mathbb{E}[Y_\ell]$$

whenever $\mathbb{E}[|Y_\ell|] < \infty$ for all $\ell = 1, \ldots, k$.

A seemingly more involved version of the impact of independence on the evaluation of expectations is given next: In the setting of Lemma 12.1.1, partition the index set $\{1, \ldots, k\}$ into r subsets, say $I_1, \ldots, I_r$ with $I_s \cap I_t = \emptyset$ for distinct $s, t = 1, \ldots, r$ and $\cup_{s=1}^r I_s = \{1, \ldots, k\}$. For each $s = 1, \ldots, r$, set $q_s = \sum_{\ell \in I_s} p_\ell$.

Lemma 12.1.2 *Consider the mutually independent rvs $X_1 : \Omega \to \mathbb{R}^{p_1}$, ..., $X_k : \Omega \to \mathbb{R}^{p_k}$. With Borel mappings $h_1 : \mathbb{R}^{q_1} \to \mathbb{R}$, ..., $h_r : \mathbb{R}^{q_r} \to \mathbb{R}^{q_r}$, define the rvs*

$$Z_s = h_s((X_\ell,\ \ell \in I_s)), \quad s = 1, \ldots, r.$$

The $\mathbb{R}$-valued rvs $Z_1, \ldots, Z_s$ are mutually independent, and

$$\mathbb{E}\left[\prod_{s=1}^r Y_s\right] = \prod_{s=1}^s \mathbb{E}\left[Z_s\right]$$

whenever $\mathbb{E}\left[|Z_s\right] < \infty$ for all $s = 1, \ldots, r$.

The proof of Lemma 12.1.1 and Lemma 12.1.2 is left as an exercise [Exercise 12.1]

12.2 Convergence results for expectations and interchange

In this section we are interested in conditions that allow the interchange of the expectation and limit operations. To set the stage consider a collection $\{X, Y, Z, X_n,\ n = 1, 2, \ldots\}$ of $\mathbb{R}$-valued rvs which are all defined on the *same* probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Monotone Convergence Theorem We begin with the situation when the rvs $\{X_n,\ n = 1, 2, \ldots\}$ are monotone; the non-decreasing and non-increasing cases are both discussed.

Theorem 12.2.1 *(i) Assume that $X \leq X_n \leq X_{n+1}$ for all $n = 1, 2, \ldots$. If the expectation $\mathbb{E}\left[X\right]$ exists with $-\infty < \mathbb{E}\left[X\right]$, then we have*

(12.4) $$\lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right]$$

monotonically.

(ii) Assume that $X_{n+1} \leq X_n \leq Y$ for all $n = 1, 2, \ldots$. If the expectation $\mathbb{E}\left[Y\right]$ exists with $\mathbb{E}\left[Y\right] < +\infty$, then we have

(12.5) $$\lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right]$$

monotonically.

Under the assumptions of Theorem 12.2.1 in both settings, the limit $\lim_{n\to\infty} X_n$ exists pointwise (and is a (possibly extended) rv), the expectations $\mathbb{E}\left[X_n\right]$ exist for all $n = 1, 2, \ldots$ and the limit $\lim_{n\to\infty} \mathbb{E}\left[X_n\right]$ also exists by monotonicity. It is plain that Claims (i) and (ii) are equivalent Exercise 12.2]. The integrability conditions on either X or Y can not be dropped as the following counterexample shows:

Counterexample 12.2.1 Consider a discrete rv Z with support $S_Z = \mathbb{N}$ with pmf $\boldsymbol{p}_Z = (p_Z(z),\ z = 0, 1, \ldots)$ given by

$$p_Z(z) = \frac{C}{1+z^2}, \quad z = 0, 1, \ldots$$

for some $C > 0$. This rv Z is defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, say the probability triple used in the proof of Lemma 8.1.1. First we note that

$$\begin{aligned} \mathbb{E}\left[(Z-n)^+\right] &= \sum_{z=0}^{\infty} (z-n)^+ p_Z(z) \\ &= \sum_{t=n+1}^{\infty} t p_Z(n+t) \\ &= C \sum_{t=n+1}^{\infty} \frac{t}{1+(t+n)^2} = \infty. \qquad (12.6) \end{aligned}$$

For Claim (i), define $X_n \equiv -(Z-n)^+$ for $n = 1, 2, \ldots$. The rvs $\{X_n,\ n = 1, 2, \ldots\}$ form an increasing sequence with $\lim_{n\to\infty} X_n = 0$ so that $\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = 0$. On the other hand, we have $\mathbb{E}\left[X_n\right] = \infty$ for each $n = 1, 2, \ldots$ and the interchange (12.4) fails! The condition that there exists a rv X such that $X \leq X_n$ for all $n = 1, 2, \ldots$ and $\mathbb{E}\left[X\right] \mid$ exists with $-\infty < \mathbb{E}\left[X\right]$ does not hold here since we automatically would have $\mathbb{E}\left[X\right] = -\infty$.

Similarly, for Claim (ii), define $X_n \equiv (Z-n)^+$ for $n = 1, 2, \ldots$. The rvs $\{X_n,\ n = 1, 2, \ldots\}$ form a decreasing sequence with $\lim_{n\to\infty} X_n = 0$ so that $\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = 0$. On the other hand, we have $\mathbb{E}\left[X_n\right] = -\infty$ for each $n = 1, 2, \ldots$ and the interchange (12.5) fails! There cannot be a rv Y such that $Y \leq X_n$ for all $n = 1, 2, \ldots$ and $\mathbb{E}\left[Y\right] \mid$ exists with $-\infty < \mathbb{E}\left[Y\right]$ for it would necessarily satisfy $\mathbb{E}\left[Y\right] = \infty$. ■

An important consequence of the Monotone Convergence Theorem is its use on series with non-negative terms: Let $\{X_n,\ n = 1, 2, \ldots\}$ denote a sequence of

$\mathbb{R}_+$-valued rvs. It follows from the Monotone Convergence Theorem that

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}\left[X_n\right]. \tag{12.7}$$

This is because, with

$$S_n = \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots$$

non-negativity implies $0 \leq S_n \leq S_{n+1}$ for all $n = 1, 2, \ldots$, whence

$$\lim_{n\to\infty} \mathbb{E}\left[S_n\right] = \mathbb{E}\left[\lim_{n\to\infty} S_n\right]$$

by (12.4) (with $X = 0$ here). By linearity, we have $\mathbb{E}\left[S_n\right] = \sum_{k=1}^{n} \mathbb{E}\left[X_k\right]$ for each $n = 1, 2, \ldots$, so that $\lim_{n\to\infty} \mathbb{E}\left[S_n\right] = \sum_{n=1}^{\infty} \mathbb{E}\left[X_n\right]$, while $\lim_{n\to\infty} S_n = \sum_{n=1}^{\infty} X_n$ – Both limiting statements are valid by the monotonocity implied by the non-negativity of the summands. The interchange (12.7) holds.

Fatou's Lemma Fatou's Lemma given next deals with situations when the limit either does not exist or is not known (yet) to exist.

Theorem 12.2.2 *(i) If $X \leq X_n$ for all $n = 1, 2, \ldots$ and $\mathbb{E}\left[X\right]$ exists with $-\infty < \mathbb{E}\left[X\right]$, we have*

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[X_n\right]. \tag{12.8}$$

(ii) If $X_n \leq Y$ for all $n = 1, 2, \ldots$ and $\mathbb{E}\left[Y\right]$ exists with $\mathbb{E}\left[Y\right] < \infty$, we have

$$\limsup_{n\to\infty} \mathbb{E}\left[X_n\right] \leq \mathbb{E}\left[\limsup_{n\to\infty} X_n\right]. \tag{12.9}$$

As with the Monotone Convergence Theorem, under the assumptions of Theorem 12.2.2 in both settings, the expectations $\mathbb{E}\left[X_n\right]$ exist for all $n = 1, 2, \ldots$. The proof of Fatou's Lemma is an easy consequence of the Monotone Convergence Theorem.

Proof. We establish only Claim as it is easy to check that Claim (i) and Claim (ii) are in fact equivalent [Exercise 12.3].

With the sequence $\{X_n,\ n = 1, 2, \ldots\}$ we associate the sequence $\{\underline{X}_n,\ n = 1, 2, \ldots\}$ given by

$$\underline{X}_n = \inf_{m\geq n} X_m, \quad n = 1, 2, \ldots \tag{12.10}$$

This sequence of rvs is monotonically increasing with $X \leq \underline{X}_n$ for all $n = 1, 2, \ldots$. As we have assumed that $\mathbb{E}[X]$ exists with $-\infty < \mathbb{E}[X]$, we can now apply the Monotone Convergence Theorem, namely Part (i) of Theorem 12.2.2, to the sequence $\{\underline{X}_n,\ n = 1, 2, \ldots\}$. This yields

$$\lim_{n\to\infty} \mathbb{E}[\underline{X}_n] = \mathbb{E}\left[\lim_{n\to\infty} \underline{X}_n\right]. \tag{12.11}$$

Obviously $\underline{X}_n \leq X_n$ for all $n = 1, 2, \ldots$, hence $\mathbb{E}[\underline{X}_n] \leq \mathbb{E}[X_n]$ for all $n = 1, 2, \ldots$. As a result,

$$\liminf_{n\to\infty} \mathbb{E}[\underline{X}_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]. \tag{12.12}$$

Combining (12.11) and (12.12) readily implies (12.8) because $\lim_{n\to\infty} \underline{X}_n = \liminf_{n\to\infty} X_n$, and $\liminf_{n\to\infty} \mathbb{E}[\underline{X}_n] = \lim_{n\to\infty} \mathbb{E}[\underline{X}_n]$ as these limits both exist. ■

The following example shows that the bounding conditions cannot be eliminated.

Counterexample 12.2.2 Take $\Omega = (0, 1)$ and $\mathcal{F} = \mathcal{B}((0, 1))$ with $\mathbb{P}$ being Lebesgue measure λ. The rvs $\{X_n,\ n = 1, 2, \ldots\}$ are given by

$$X_n(\omega) = \begin{cases} 0 & \text{if } \omega \notin [\frac{1}{n}, \frac{2}{n}] \\ \\ -n & \text{if } \omega \in [\frac{1}{n}, \frac{2}{n}] \end{cases}, \qquad \begin{array}{c} \omega \in \Omega \\ n = 2, 3, \ldots \end{array}$$

Obviously, $\mathbb{E}[X_n] = n^{-1}(-n) = -1$ for all $n = 1, 2, \ldots$, so that $\liminf_{n\to\infty} \mathbb{E}[X_n] = -1$, while $\liminf_{n\to\infty} X_n = 0$ so that $\mathbb{E}[\liminf_{n\to\infty} X_n] = 0$. ■

An interesting consequence of Fatou's Lemma is obtained by combining both parts.

Corollary 12.2.1 *Consider a sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ such that $X \leq X_n \leq Y$ for all $n = 1, 2, \ldots$. Assume that $\mathbb{E}[X]$ exists with $-\infty < \mathbb{E}[X]$ and that $\mathbb{E}[Y]$ exists with $\mathbb{E}[Y] < \infty$. If $\lim_{n\to\infty} X_n$ exists, then the interchange*

$$\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}[X_n] \tag{12.13}$$

holds

The next two results are illustrations of this fact.

Bounded Convergence Theorem The Bounded Convergence Theorem shows that the interchange always holds when the rvs $\{X_n,\ n = 1, 2, \ldots\}$ form a bounded sequence.

Theorem 12.2.3 *Assume there exists a rv* $X : \Omega \to \mathbb{R}$ *such that* $\lim_{n\to\infty} X_n = X$*. If there exists* $M > 0$ *such that* $|X_n| \leq M$*, for each* $n = 1, 2, \ldots$*, then* $\mathbb{E}[X]$ *exists and is finite with*

$$\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]. \tag{12.14}$$

Dominated Convergence Theorem The Dominated Convergence Theorem generalizes the Bounded Convergence Theorem by requiring only that the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ can be uniformly bounded by a positive rv whose expectation is finite.

Theorem 12.2.4 *Assume there exists a rv* $X : \Omega \to \mathbb{R}$ *such that* $\lim_{n\to\infty} X_n = X$*. If there exists a rv* $Y : \Omega \to \mathbb{R}_+$ *with* $\mathbb{E}[Y] < \infty$ *such that* $|X_n| < Y$ *for each* $n = 1, 2, \ldots$*, then* $\mathbb{E}[X]$ *exists and is finite with*

$$\mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]. \tag{12.15}$$

We close this section with a discussion of counterexamples to Theorem 12.2.3 and Theorem 12.2.4

Counterexample 12.2.3 Take $\Omega = (0, 1)$ and $\mathcal{F} = \mathcal{B}((0, 1))$ with $\mathbb{P}$ being Lebesgue measure λ. The rvs $\{X_n,\ n = 1, 2, \ldots\}$ are given by

$$X_n(\omega) = \begin{cases} 0 & \text{if } 0 < \omega < 1 - a_n \\ \\ b_n & \text{if } 1 - a_n \leq \omega < 1 \end{cases}, \quad \begin{array}{c} \omega \in \Omega \\ n = 1, 2, \ldots \end{array}$$

where $0 < a_n < 1$ and $b_n \neq 0$. For each $n = 1, 2, \ldots$, it is plain that $\mathbb{E}[X_n] = a_n b_n$. If $\lim_{n\to\infty} a_n = 0$, then $\lim_{n\to\infty} X_n = 0$. However, it is possible to drive $\lim_{n\to\infty} \mathbb{E}[X_n]$ to any value $c \neq 0$ by suitably selecting a_n and b_n: If we select $b_n = c a_n^{-1}$, then $\lim_{n\to\infty} \mathbb{E}[X_n] = c$, and with $b_n = \pm a_n^{-2}$, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \pm\infty$. ■

12.3 Change of variable formula for discrete rvs

We are in the setting of Section 8.4: Consider a discrete rv $X : \Omega \to \mathbb{R}^p$ with support $S_X \subseteq \mathbb{R}^p$ and pmf $\boldsymbol{p}_X = (p_X(x),\ x \in S_X)$. For any Borel mapping $g : \mathbb{R}^p \to \mathbb{R}$, we introduced the rv $Y : \Omega \to \mathbb{R}$ defined at (**??**) by composing the rv $X : \Omega \to \mathbb{R}^p$ with g, namely $Y = g(X)$.

According to Fact 8.4.1, the rv $Y : \Omega \to \mathbb{R}$ is a discrete rv with support $S_Y \equiv \{g(x) : \ x \in S_X\}$ and pmf $\boldsymbol{p}_Y = (p_Y(Y),\ y \in S_Y)$ determined through (8.11), namely

$$p_Y(y) \sum_{x \in S_X :\ g(x)=y} p_X(x), \quad y \in S_Y. \tag{12.16}$$

We seek to evaluate the expectation $\mathbb{E}\left[Y\right]$ or equivalently, $\mathbb{E}\left[g(X)\right]$. By Proposition 11.3.1 we have

$$\mathbb{E}\left[Y\right] = \sum_{y \in S_Y} y \cdot p_Y(y) \tag{12.17}$$

if either the rv Y is a.s. non-negative (with $S_Y \subseteq \mathbb{R}_+$) or if the absolute summability condition $\sum_{y \in S_Y} |y| p_Y(y) < \infty$ holds.

The difficulty with this approach is that it requires the availability of the pmf of the rv Y before one can even attempt to evaluate the sum (12.17). In many instances it is preferable to use a different computational strategy that we now explore and which requires only that the pmf of the rv X be available.

The point of departure is still the expression (12.17) but this time we use the expression (12.16) for the pmf of the rv Y: More precisely,

$$\begin{aligned}
\mathbb{E}\left[g(X)\right] &= \sum_{y \in S_Y} y \cdot p_Y(y) \\
&= \sum_{y \in S_Y} y \cdot \left(\sum_{x \in S_X :\ g(x)=y} p_X(x) \right) \\
&= \sum_{y \in S_Y} y \cdot \left(\sum_{x \in S_X} \mathbf{1}\left[g(x) = y\right] p_X(x) \right) \\
&= \sum_{x \in S_X} \left(\sum_{y \in S_Y} y \cdot \mathbf{1}\left[g(x) = y\right] \right) p_X(x) \\
&= \sum_{x \in S_X} \left(\sum_{y \in S_Y} g(x) \mathbf{1}\left[g(x) = y\right] \right) p_X(x)
\end{aligned}$$

$$= \sum_{x \in S_X} \left(\sum_{y \in S_Y} \mathbf{1}\left[g(x) = y\right] \right) g(x) p_X(x). \tag{12.18}$$

12.4 Change of variable formula

In view of the definition we have developed it is natural to write

$$\mathbb{E}\left[X\right] = \int_\Omega X(\omega) d\mathbb{P}(\omega)$$

whenever $\mathbb{E}\left[X\right]$ exists as this notation mimics the expression used for simple rvs. However, as shown in Section 10.5 this quantity depends only on the probability distribution $F_X : \mathbb{R} \to [0, 1]$.

Recall that any rv $X : \Omega \to \mathbb{R}^p$ naturally induces a probability triple on its range, namely $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)$ where $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \to [0, 1]$ is the probability measure defined by

$$\mathbb{P}_X\left[B\right] = \mathbb{P}\left[X \in B\right], \quad B \in \mathcal{B}(\mathbb{R}^p).$$

In fact, the identity mapping $\mathrm{Id} : \mathbb{R}^p \to \mathbb{R}^p : x \to x$ defines a rv $\mathbb{R}^p \to \mathbb{R}^p$ whose probability distribution (under $\mathbb{P}_X$) coincides with the probability distribution of X (under $\mathbb{P}$) since

$$\mathbb{P}_X\left[\mathrm{Id} \in B\right] = \mathbb{P}_X\left[B\right] = \mathbb{P}\left[X \in B\right], \quad B \in \mathcal{B}(\mathbb{R}^p).$$

Obviously, say with $p = 1$, the expectation of X computed under $\mathbb{P}$ has to coincide with that of the rv Id computed under $\mathbb{P}_X$ with the understanding that if one exists (resp. and is finite) so it is for the other, leading us to write

$$\mathbb{E}\left[X\right] = \int_{\mathbb{R}} x d\mathbb{P}_X(x).$$

Finally, by Carathéodory's Theorem that F_X and $\mathbb{P}_X$ contain the same probabilistic information concerning the rv X, we shall often adopt the notation

$$\mathbb{E}\left[X\right] = \int_{\mathbb{R}} x dF_X(x).$$

Proposition 12.4.1 *Consider an $\mathbb{R}^p$-valued rv $X : \Omega \to \mathbb{R}^p$. With Borel mapping $g : \mathbb{R}^p \to \mathbb{R}$, it holds that*

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}^p} g(x) dF_X(x) \tag{12.19}$$

with the understanding that if one of the quantities is well defined, so is the other and their values coincide.

Proof. If $g : \mathbb{R}^p \to \mathbb{R}$ is of the form

$$g(x) = \mathbf{1}\left[x \in B\right], \quad x \in \mathbb{R}^p$$

for some Borel set B in $\mathcal{B}(\mathbb{R}^p)$, then

$$\mathbb{E}\left[g(X)\right] = \mathbb{P}\left[X \in B\right] = \mathbb{P}_X[B] = \mathbb{E}_X[g(\cdot)] = \int_{\mathbb{R}^p} g(x) dF_X(x)$$

Assume now that $g : \mathbb{R}^p \to \mathbb{R}$ is simple in the sense that

$$g(x) = \sum_{i \in I} g_i \mathbf{1}\left[x \in B_i\right], \quad x \in \mathbb{R}^p$$

Then,

$$\begin{aligned} \mathbb{E}\left[g(X)\right] &= \mathbb{E}\left[\sum_{i \in I} g_i \mathbf{1}\left[X \in B_i\right]\right] \\ &= \sum_{i \in I} g_i \mathbb{E}\left[\mathbf{1}\left[X \in B_i\right]\right] \\ &= \sum_{i \in I} g_i \mathbb{P}\left[X \in B_i\right] \\ &= \sum_{i \in I} g_i \int_{\mathbb{R}^p} \mathbf{1}\left[B_i\right](x) dF_X(x) \\ &= \int_{\mathbb{R}^p} g(x) dF_X(x) \qquad (12.20) \end{aligned}$$

If $g : \mathbb{R}^p \to \mathbb{R}_+$, then we generate the sequence of simple mappings $\{g_n,\ n = 1, 2, \ldots\}$ where for each $n = 1, 2, \ldots$, the Borel mapping $g_n : \mathbb{R}^p \to \mathbb{R}$ is given by

$$g_n(x) = \sum_{m=0}^{n-1} \sum_{k=0}^{2^n - 1} \frac{k}{2^n} \mathbf{1}\left[\frac{k}{2^n} < x \leq \frac{k+1}{2^n}\right], \quad x \in \mathbb{R}^p$$

We already have

$$\mathbb{E}\left[g_n(X)\right] = \int_{\mathbb{R}^p} g_n(x) dF_X(x), \quad n = 1, 2, \ldots$$

and the conclusion

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}^p} g(x) dF_X(x),$$

follows by the Monotone Convergence Theorem (under $\mathbb{P}$ and $\mathbb{P}_X$).

In the general case $g : \mathbb{R}^p \to \mathbb{R}$, write

$$g(x) = g(x)^+ - g(x)^-, \quad x \in \mathbb{R}^p$$

and by linearity, we get

$$\mathbb{E}\left[g(X)\right] = \mathbb{E}\left[g(X)^+\right] - \mathbb{E}\left[g(X)^-\right]$$

■

12.5 Riemann-Stieltjes vs. Lebesgue integration

12.6 Exercises

Ex. 12.1 Prove Lemma 12.1.1 and Lemma 12.1.2

Ex. 12.2 Show that Claim (i) and Claim (ii) of the Monotone Convergence Theorem 12.2.1 are equivalent – This is already apparent in the Counterexample 12.2.1.

Ex. 12.3 Show that Claim (i) and Claim (ii) of Fatou's Lemma [Theorem 12.2.2] are equivalent.

Ex. 12.4 Let $W_1, \ldots, W_n$ denote n mutually independent Walsh rvs with same parameter p (in $(0,1)$) all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, i.e., for each $k = 1, \ldots, n$, we have

$$\mathbb{P}\left[W_k = w\right] = \begin{cases} p & \text{if } w = 1 \\ \\ 1-p & \text{if } w = -1. \end{cases}$$

For each $k = 1, \ldots, n$, write $W_k^\star$ for the product of the k rvs $W_1, \ldots, W_k$, i.e., $W_k^\star \equiv \prod_{\ell=1}^n W_\ell$.

a. For each $k = 1, 2, \ldots, n$, explain why the rv $W_k^\star$ is a Walsh rv.

b. For each $k = 1, 2, \ldots, n$, let $p_k^\star$ denote the parameter of the rv Walsh $W_k^\star$. Find a recursive relationship between $p_{k+1}^\star$ and $p_k^\star$. Can you find an explicit expression for $p_1^\star, \ldots, p_n^\star$, say by iterating this recursion?

c. An elegant way to find $p_1^\star, \ldots, p_n^\star$ is as follows: For each $k = 1, \ldots, n$, compute $\mathbb{E}\left[W_k^\star\right]$ and use Part **a**.

Ex. 12.5 Compute the first two moments $\mathbb{E}[X]$ and $\mathbb{E}\left[X^2\right]$ (and the variance $\mathrm{Var}[X]$) when the discrete rv $X : \Omega \to \mathbb{R}$ is

a. a Binomial rv $\mathrm{Bin}(n;p)$ with $n = 1, 2, \ldots$ and $0 < p < 1$.

b. a Poisson rv $\mathrm{Poi}(\lambda)$ with $\lambda > 0$.

c. a Geometric rv $\mathrm{Geo}(p)$ with $0 < p < 1$.

In each case explain why the expectations always exist.

Ex. 12.6 Using the change of variable formula, compute the expectation

$$\mathbb{E}\left[\frac{1}{1+Y^+}\right]$$

when the discrete rv $Y : \Omega \to \mathbb{R}$ is

a. a Binomial rv $\mathrm{Bin}(n;p)$ with $n = 1, 2, \ldots$ and $0 < p < 1$.

b. a Poisson rv $\mathrm{Poi}(\lambda)$ with $\lambda > 0$.

c. a Geometric rv $\mathrm{Geo}(p)$ with $0 < p < 1$.

In each case explain why the expectation $\mathbb{E}\left[\frac{1}{1+Y^+}\right]$ always exists.

Ex. 12.7 A non-empty subset of $\{1, \ldots, n\}$ (for some finite n) is selected uniformly at random.

a. Propose a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ for this random experiment.

Consider the discrete rvs $X, Y : \Omega \to \mathbb{N}_0$ given by

$$X(\omega) = \max\{k : \ k \in \omega\} \quad \text{and} \quad Y(\omega) = \min\{\ell : \ \ell \in \omega\}, \quad \omega \in \Omega.$$

b. Find the joint pmf $\boldsymbol{p}_{X,Y}$ of the discrete rv $(X, Y) : \Omega \to \mathbb{N}_0 \times \mathbb{N}_0$ – Specifiy its support $S_{X,Y}$.

c. Find the pmf $\boldsymbol{p}_X$ of the rv X – Specifiy its support S_X. Evaluate $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

d. Find the pmf $\boldsymbol{p}_Y$ of the rv Y – Specifiy its support S_Y. Evaluate $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$.

e. Show that the rvs X and $n + 1 - Y$ are equidistributed. **f.** Find the pmf $\boldsymbol{p}_{X-Y}$ of the discrete rv $X - Y$ – Specifiy its support S_{X-Y}. Evaluate $\mathbb{E}[X - Y]$, $\mathrm{Cov}[X, Y]$ and $\mathrm{Var}[X - Y]$.

Chapter 13

Moments and inequalities

All rvs are defined as Borel measurable mappings $\Omega \to \mathbb{R}$ on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and all probability distributions are computed under $\mathbb{P}$.

13.1 Moments

Consider the rv $X : \Omega \to \mathbb{R}$. With $p = 1, 2, \ldots$, we define the p^{th} moment m_p of X by

$$m_p \equiv \mathbb{E}\left[X^p\right] \tag{13.1}$$

provided the expectation exists.

For any $p \geq 0$ the *absolute* p^{th} moment of X is given by

$$\mu_p \equiv \mathbb{E}\left[|X|^p\right]. \tag{13.2}$$

This quantity is always well defined, and may possibly be infinite. Note that (13.2) in general cannot be defined for *non*-integer $p > 0$ (unless $X \geq 0$ a.s.)

Definition 13.1.1

When $p = 1$ we refer to m_1 as the first moment of X. When $p = 2$, m_2 always exists but may be infinite. We say that the rv X is a *second-order* rv if the second moment is *finite*, namely if $\mathbb{E}\left[|X|^2\right] < \infty$.

When the first moment of X exists and is finite, the definitions (13.1) and (13.2) lead naturally to the *centered* expectations given by

$$m_p^\star \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^p\right], \quad p = 1, 2, \ldots \tag{13.3}$$

and

$$\mu_p^\star \equiv \mathbb{E}\left[|X - \mathbb{E}[X]|^p\right], \quad p \geq 0 \tag{13.4}$$

provided these expectations exists.

Fact 13.1.1 *With $1 \leq p < q$ (not necessarily integers), we have*

$$\mathbb{E}\left[|X|^p\right] \leq 1 + \mathbb{E}\left[|X|^q\right].$$

In particular, the finiteness of $\mathbb{E}\left[|X|^q\right]$ implies that of $\mathbb{E}\left[|X|^p\right]$.

Proof. With $u \geq 0$ the inequality $u^p \leq 1 + u^q$ holds whenever $1 \leq p < q$ (not necessarily integers), and the result follows by the monotonicity of integration. ■

13.2 Variance and covariance

If the rv $X : \Omega \to \mathbb{R}$ is a second-order rv, then $\mathbb{E}\left[|X|\right] < \infty$ by virtue of Fact 13.1.1, and both $\mathbb{E}[X]$ and $\mathbb{E}\left[|X|\right]$ exist and are finite. The centered moment (13.3) for $p = 2$ occupies an important place in Statistics and the Data Sciences where it referred to as the variance.

Definition 13.2.1

The *variance* $\mathrm{Var}[X]$ of the second-order rv X is given by

$$\mathrm{Var}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{13.5}$$

and is well defined and finite. It is also customary to refer to the square-root of this quantity as the *standard deviation* $\sigma(X)$ of the second-order rvs X , namely

$$\sigma(X) \equiv \sqrt{\mathrm{Var}[X]}.$$

Noting that $(X - \mathbb{E}[X])^2 = X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2$, we conclude that

$$\begin{aligned} \mathrm{Var}[X] &= \mathbb{E}\left[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \qquad (13.6) \end{aligned}$$

Since $(X - \mathbb{E}[X])^2 \geq 0$ it follows that $\text{Var}[X] \geq 0$, and the inequality $(\mathbb{E}[X])^2 \leq \mathbb{E}\left[X^2\right]$ therefore holds between first and second moments. It should also be noted that $\text{Var}[X] = 0$ is equivalent to $X - \mathbb{E}[X] = 0$ a.s. In other words, a second-order rv$X$ has zero variance if and only if it is degenerate with $X = \mathbb{E}[X]$ a.s.

Definition 13.2.2

Let $X, Y : \Omega \to \mathbb{R}$ be a pair of second-order rvs. The *covariance* $\text{Cov}[X, Y]$ between the rvs X and Y is defined by

$$\text{Cov}[X, Y] \equiv \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]. \tag{13.7}$$

The quantity (13.7) is well defined by virtue of the fact $(|a| - |b|)^2 = a^2 + b^2 - 2|a| \cdot |b| \geq 0$ for arbitrary scalars a and b in $\mathbb{R}$, so that $|a| \cdot |b| \leq \frac{1}{2}(a^2 + b^2)$. As was done for the variance, we can write

$$(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y]) = XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y].$$

Taking expectations we get

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{13.8}$$

as an alternate definition for the covariance between the rvs X and Y. Note that $\text{Cov}[X, X] = \text{Var}[X]$.

We close this section with a classical definition.

Definition 13.2.3

Let $X, Y : \Omega \to \mathbb{R}$ be a pair of non-degenerate second-order rvs., i.e.. $\text{Var}[X] > 0$ and $\text{Var}[Y] > 0$. The *coefficient of correlation* $\rho(X; Y)$ between the rvs X and Y is defined by

$$\rho(X; Y) \equiv \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]} \cdot \sqrt{\text{Var}[Y]}} = \frac{\text{Cov}[X, Y]}{\sigma(X) \cdot \sigma(Y)}. \tag{13.9}$$

We stress that the probability distribution of the rv X (resp. Y) determines $\text{Var}[X]$ and $\sigma(X)$ (resp. $\text{Var}[Y]$ and $\sigma(Y)$), while $\text{Cov}[X, Y]$ and $\rho(X; Y)$ are determined by the *joint* distribution of the pair of rvs X and Y.

13.3 Sums of rvs

Consider a collection of rvs $X_1, \ldots, X_n : \Omega \to \mathbb{R}$. We already know that if $\mathbb{E}[|X_i|] < \infty$ for each $i = 1, \ldots, n$, then the expectation of the sum rv $X_1 + \ldots + X_n$ exists and is finite with

$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n].$$

It is certainly natural to wonder what would be the analog of this fact for the variance. This is discussed next.

Lemma 13.3.1 *If the rvs $X_1, \ldots, X_n$ are second-order rvs, then*

$$\text{(13.10)}\ \mathrm{Var}[X_1 + \ldots + X_n] = \sum_{k=1}^{n} \mathrm{Var}[X_k] + \sum_{k=1}^{n} \sum_{\ell=1,\, \neq k}^{n} \mathrm{Cov}[X_k, X_\ell].$$

Proof. We start by noting that the rv $X_1 + \ldots + X_n$ is also a second-order rv since

$$(X_1 + \ldots + X_n)^2 \leq n\left(|X_1|^2 + \ldots + |X_n|^2\right)$$

by the convexity of the mapping $t \to t^2$ on $\mathbb{R}$. Noting that

$$X_1 + \ldots + X_n - \mathbb{E}[X_1 + \ldots + X_n] = \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]),$$

elementary calculations give

$$\begin{aligned} & (X_1 + \ldots + X_n - \mathbb{E}[X_1 + \ldots + X_n])^2 \\ = & \sum_{k=1}^{n} \sum_{\ell=1}^{n} (X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell]) \\ = & \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k])^2 \\ & + \sum_{k=1}^{n} \sum_{\ell=1,\, \ell\neq k}^{n} (X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell]). \end{aligned}$$

Taking expectations on both sides of this last relation, we conclude that

$$\begin{aligned} \mathrm{Var}[X_1 + \ldots + X_n] &= \sum_{k=1}^{n} \mathbb{E}\left[(X_k - \mathbb{E}[X_k])^2\right] \\ & \quad + \sum_{k=1}^{n} \sum_{\ell=1,\, \ell\neq k}^{n} \mathbb{E}[(X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell])] \end{aligned}$$

and the desired conclusion (13.10) follows. ■

13.4 Uncorrelated rvs

We begin with a definition.

Definition 13.4.1

The second-order rvs X and Y are said to be *uncorrelated* if $\text{Cov}\,[X, Y] = 0$. It is customary to say that the rvs X and Y are *positively correlated* (resp. *negatively correlated*) if $\text{Cov}\,[X, Y] > 0$ (resp. $\text{Cov}\,[X, Y] < 0$).

Second-order rvs which are pairwise independent are necessarily uncorrelated.

Fact 13.4.1 *If two second-order rvs X and Y are independent, they are necessarily uncorrelated.*

Proof. The rvs X and Y being independent, the centered rvs $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ are also independent, whence

$$\mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])] = \mathbb{E}[X - \mathbb{E}[X]] \cdot \mathbb{E}[Y - \mathbb{E}[Y]] = 0$$

by virtue of Lemma **??** ■

However, the converse is not true even when the rvs X and Y are second-order rvs as the following counterexample shows.

Example 13.4.1 With the rv U uniformly distributed on $[0, 1]$, consider the bounded (hence second-order) rvs X and Y given by $X = \cos(2\pi U)$ and $Y = \sin(2\pi U)$, so that $X \cdot Y = \frac{1}{2}\sin(4\pi U)$. Note that

$$\mathbb{E}[\sin 2k\pi U] = \int_0^1 \sin(2k\pi u)du = -\frac{\cos(2k\pi) - \cos(0)}{2k\pi} = 0, \quad k = 1, 2$$

(since $\cos(2\ell\pi) = 1$ for all $\ell = 0, 1, 2, \ldots$) whence $\mathbb{E}[Y] = 0$ and $\mathbb{E}[X \cdot Y] = 0$, leading to $\text{Cov}\,[X, Y] = 0$. However, the rvs X and Y are not independent as can be seen from the fact that $X^2 + Y^2 = 1$: Knowledge of $\cos(2\pi U)$ determines $\sin(2\pi U)$ (up to a sign) with $Y = \pm\sqrt{1 - X^2}$. See Exercise 13.10 for another take on this counterexample. ■

The next result shows the usefulness of being uncorrelated in calculations dealing with the variance of sums of rvs. As a direct consequence of (13.10) we get the following often used fact.

Fact 13.4.2 *If the rvs* $X_1, \ldots, X_n$ *are* pairwise uncorrelated*, i.e.,*

$$\mathrm{Cov}\left[X_k, X_\ell\right] = 0, \quad \begin{array}{c} k \neq \ell \\ k, \ell = 1, \ldots, n \end{array}$$

then

$$\mathrm{Var}\left[X_1 + \ldots + X_n\right] = \sum_{k=1}^{n} \mathrm{Var}\left[X_k\right]. \tag{13.11}$$

In other words, the variance of a sum of uncorrelated rvs is indeed the sum of their individual variances.

13.5 The Cauchy-Schwarz inequality

We now present in this and the next two sections several important inequalities concerning expectations

Theorem 13.5.1 *(Cauchy-Schwarz inequality) For any pair of second-order rvs* $X, Y : \Omega \to \mathbb{R}$ *we have*

$$|\mathbb{E}\left[X \cdot Y\right]| \leq \sqrt{\mathbb{E}\left[|X|^2\right]} \cdot \sqrt{\mathbb{E}\left[|Y|^2\right]} \tag{13.12}$$

with equality if and only if there exist constants a *and* b *in* $\mathbb{R}$ *not simultaneously zero (i.e.,* $a^2 + b^2 > 0$*) such that* $aX + bY = 0$ *a.s.*

The moment $\mathbb{E}\left[X \cdot Y\right]$ is well defined and finite as pointed out in a comment following (13.7).

Proof. The inequality (13.12) trivially holds when either $\mathbb{E}\left[|X|^2\right] = 0$ or $\mathbb{E}\left[|Y|^2\right] = 0$ since then $X = 0$ a.s. or $Y = 0$ a.s., resulting in $XY = 0$ a.s. Therefore from now on we assume that $\mathbb{E}\left[|X|^2\right] > 0$ and $\mathbb{E}\left[|Y|^2\right] > 0$. With this in mind, consider the quadratic form $Q : \mathbb{R} \to \mathbb{R}_+$ given by

$$Q(\lambda) \equiv \mathbb{E}\left[(X + \lambda Y)^2\right], \quad \lambda \in \mathbb{R}.$$

Note that $Q(\lambda)$ is well defined and finite (since $(a+b)^2 \leq 2(a^2 + b^2)$ for all $a, b \geq 0$).

Obviously,

$$Q(\lambda) \quad = \quad \mathbb{E}\left[X^2\right] + 2\lambda\mathbb{E}\left[XY\right] + \lambda^2\mathbb{E}\left[Y^2\right], \quad \lambda \in \mathbb{R}. \tag{13.13}$$

The roots of this quadratic form are determined by the sign of the discriminant

$$\Delta = (2\mathbb{E}\left[XY\right])^2 - 4\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] = 4\left(\mathbb{E}\left[XY\right]^2 - \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]\right).$$

By its very definition, $Q(\lambda) \geq 0$ for all λ in $\mathbb{R}$, hence the quadratic equation $Q(\lambda) = 0$ on $\mathbb{R}$ cannot have two *real* distinct roots, say $\lambda_1 < \lambda_2$, as this would imply $Q(\lambda) < 0$ in the interval (λ_1, λ_2) under the condition $\mathbb{E}\left[Y^2\right] > 0$. In other words, it is not possible for $\Delta > 0$ to occur. Note that the alternative, namely $\Delta \leq 0$, is equivalent to the Cauchy-Schwarz inequality.

If (13.12) holds as an equality, then we necessarily have $\Delta = 0$, in which case there exists a unique $\lambda^\star$ in $\mathbb{R}$ such that $Q(\lambda^\star) = 0$. As this is equivalent to $\mathbb{E}\left[(X + \lambda^\star Y)^2\right] = 0$, we conclude that $X + \lambda^\star Y = 0$ a.s., hence $aX + bY = 0$ a.s. with $a = 1$ and $b = \lambda^\star$ – Obviously, $a^2 + b^2 = 1 + (\lambda^\star)^2 > 0$.

Conversely, assume that there exist constants a and b in $\mathbb{R}$ not simultaneously zero such that $aX + bY = 0$ a.s. For instance, assuming $a \neq 0$ for the sake of concreteness, we have $X + a^{-1}bY = 0$ a.s. and $Q(a^{-1}b) = 0$. Thus $a^{-1}b$ is a real root of the quadratic form, in fact its only real root. As this requires $\Delta = 0$ we get equality in the Cauchy-Schwarz inequality. The case where $b \neq 0$ is handled similarly, and details are left to the interested reader. ■

The inequality (13.12) yields a little more: Indeed, as we apply the Cauchy-Schwarz inequality to the second-order rvs $|X|$ and $|Y|$ we get

$$\mathbb{E}\left[|X| \cdot |Y|\right] \leq \sqrt{\mathbb{E}\left[|X|^2\right]} \cdot \sqrt{\mathbb{E}\left[|Y|^2\right]},$$

whence

$$|\mathbb{E}\left[X \cdot Y\right]| \leq \mathbb{E}\left[|X| \cdot |Y|\right] \leq \sqrt{\mathbb{E}\left[|X|^2\right]} \cdot \sqrt{\mathbb{E}\left[|Y|^2\right]}. \tag{13.14}$$

The Cauchy-Schwarz inequality yields the following interesting consequence concerning the range of correlation coefficients.

Fact 13.5.1 *Let $X, Y : \Omega \to \mathbb{R}$ be a pair of non-degenerate second-order rvs., i.e..* $\mathrm{Var}\left[X\right] > 0$ *and* $\mathrm{Var}\left[Y\right] > 0$*. The coefficient of correlation $\rho(X;Y)$ between the rvs X and Y satisfies*

$$|\rho(X;Y)| \leq 1 \tag{13.15}$$

with equality if and only if there exist constants a and b in $\mathbb{R}$ not simultaneously zero (i.e., $a^2 + b^2 > 0$) such that $a\left(X - \mathbb{E}\left[X\right]\right) + b\left(Y - \mathbb{E}\left[Y\right]\right) = 0$ a.s.

Proof. This fact is a straightforward consequence of Theorem 13.5.1 applied to the rvs $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$. ■

13.6 The Hölder inequality and its consequences

The Cauchy-Schwarz inequality admits the following generalization known as Hölder's inequality

Theorem 13.6.1 *(Hölder's inequality) Consider a pair of rvs $X, Y : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|^p] < \infty$ and $\mathbb{E}[|Y|^q] < \infty$ for $p, q > 1$. Whenever*

$$\frac{1}{p} + \frac{1}{q} = 1, \tag{13.16}$$

it holds that

$$\mathbb{E}[|X| \cdot |Y|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \cdot (\mathbb{E}[|Y|^q])^{\frac{1}{q}}. \tag{13.17}$$

Definition 13.6.1

Pairs of integers p and q such that (13.16) holds are said to form a *conjugate* pair. Hölder's inequality reduces to the Cauchy-Schwarz inequality when $p = q = 2$.

Under the conditions $\mathbb{E}[|X|^p] < \infty$ and $\mathbb{E}[|Y|^q] < \infty$, the inequality (13.17) necessarily implies $\mathbb{E}[|X| \cdot |Y|] < \infty$, hence $\mathbb{E}[XY]$ exists as a finite quantity satisfying

$$|\mathbb{E}[X.Y]| \leq \mathbb{E}[|X| \cdot |Y|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \cdot (\mathbb{E}[|Y|^q])^{\frac{1}{q}}$$

by the usual arguments. It should also be noted that (13.17) automatically holds if either $\mathbb{E}[|X|^p] = \infty$ or $\mathbb{E}[|Y|^q] = \infty$.

Proof. The inequality (13.17) trivially holds when either $\mathbb{E}[|X|^p] = 0$ or $\mathbb{E}[|Y|^q] = 0$ since then $X = 0$ a.s. or $Y = 0$ a.s., resulting in $XY = 0$ a.s.

From now on assume that $\mathbb{E}[|X|^p] > 0$ and $\mathbb{E}[|Y|^q] > 0$, and consider the rvs $X_p^\star$ and $Y_q^\star$ defined by

$$X_p^\star \equiv \frac{|X|^p}{\mathbb{E}[|X|^p]} \quad \text{and} \quad Y_q^\star \equiv \frac{|Y|^q}{\mathbb{E}[|Y|^q]}.$$

Obviously, we have $\mathbb{E}\left[X_p^\star\right] = 1$ and $\mathbb{E}\left[Y_q^\star\right] = 1$.

For each $x \geq 0$ and $y \geq 0$, the inequalities

$$x^\lambda \cdot y^{1-\lambda} \leq \lambda x + (1-\lambda) y, \quad 0 < \lambda < 1 \tag{13.18}$$

hold since equivalent to $\lambda \log x + (1-\lambda)\log y \leq \log\left(\lambda x + (1-\lambda)y\right)$ (which holds by virtue of the concavity of the function $t \to \log t$ on $(0,\infty)$).

Fix λ in $(0,1)$. Using (13.18) with $x = X_p^\star$ and $y = Y_q^\star$ we conclude that

$$\left(X_p^\star\right)^\lambda \cdot \left(Y_q^\star\right)^{1-\lambda} \leq \lambda X_p^\star + (1-\lambda) Y_q^\star,$$

whence

$$\begin{aligned} \mathbb{E}\left[\left(X_p^\star\right)^\lambda \cdot \left(Y_q^\star\right)^{1-\lambda}\right] &\leq \lambda \mathbb{E}\left[X_p^\star\right] + (1-\lambda)\mathbb{E}\left[Y_q^\star\right] \\ &= \lambda + (1-\lambda) = 1. \end{aligned} \tag{13.19}$$

With $\lambda = p^{-1}$ (so that $1 - \lambda = q^{-1}$) the integrand in (13.19) can be rewritten as

$$\left(X_p^\star\right)^\lambda \cdot \left(Y_q^\star\right)^{1-\lambda} = \frac{|X|}{\left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}}} \cdot \frac{|Y|}{\left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}}}$$

and (13.19) becomes

$$\mathbb{E}\left[\frac{|X|}{\left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}}} \cdot \frac{|Y|}{\left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}}}\right] \leq 1.$$

This completes the proof of (13.16) ■

Minkowski's inequality The following result is a consequence of Hölder's inequality and gives an important fact concerning the geometry of the set of rvs whose p^{th} moment is finite.

Theorem 13.6.2 *(Minkowski's inequality) For rvs $X, Y : \Omega \to \mathbb{R}$ such that $\mathbb{E}\left[|X|^p\right] < \infty$ and $\mathbb{E}\left[|Y|^p\right] < \infty$ for some $p \geq 1$, we have the inequality*

$$\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{p}} \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[|Y|^p\right]\right)^{\frac{1}{p}}. \tag{13.20}$$

Proof. The identity $|x+y|^p \leq 2^{p-1}\left(|x|^p + |y|^p\right)$ is valid for all $x, y \geq 0$ by the convexity of the mapping $t \to |t|^p$ on $\mathbb{R}$. Therefore, $\mathbb{E}\left[|X+Y|^p\right] < \infty$ if both $\mathbb{E}\left[|X|^p\right]$ and $\mathbb{E}\left[|Y|^p\right]$ are finite.

The case $p = 1$ being immediate by the last identity (which then reduces to the triangular inequality), we assume from now on that $p > 1$. Moreover, as the result automatically holds if $X + Y = 0$ a.s., we need only consider the case when $\mathbb{E}\left[|X+Y|^p\right] > 0$. Under these conditions, we begin by writing

$$\begin{aligned} |X+Y|^p &= |X+Y| \cdot |X+Y|^{p-1} \\ &\leq |X| \cdot |X+Y|^{p-1} + |Y| \cdot |X+Y|^{p-1}, \end{aligned}$$

whence

$$\begin{aligned} &\mathbb{E}\left[|X+Y|^p\right] \\ &\quad\leq \mathbb{E}\left[|X| \cdot |X+Y|^{p-1}\right] + \mathbb{E}\left[|Y| \cdot |X+Y|^{p-1}\right]. \end{aligned} \tag{13.21}$$

Choose $q > 1$ conjugate to p so that $q = \frac{p}{p-1}$ by (13.16), and note that

$$\mathbb{E}\left[\left(|X+Y|^{p-1}\right)^q\right] = \mathbb{E}\left[|X+Y|^p\right] < \infty.$$

Applying Hölder's inequality to the rvs $|X|$ (with $\mathbb{E}\left[|X|^p\right] < \infty$) and $|X+Y|^{p-1}$ (with $\mathbb{E}\left[\left(|X+Y|^{p-1}\right)^q\right] < \infty$) we conclude that

$$\begin{aligned} \mathbb{E}\left[|X| \cdot |X+Y|^{p-1}\right] &\leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \cdot \left(\mathbb{E}\left[\left(|X+Y|^{p-1}\right)^q\right]\right)^{\frac{1}{q}} \\ &= \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \cdot \left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{q}}. \end{aligned} \tag{13.22}$$

Similarly, we have

$$\mathbb{E}\left[|Y| \cdot |X+Y|^{p-1}\right] \leq \left(\mathbb{E}\left[|Y|^p\right]\right)^{\frac{1}{p}} \cdot \left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{q}}. \tag{13.23}$$

Combining (13.21), (13.22) and (13.23) we conclude that

$$\mathbb{E}\left[|X+Y|^p\right] \leq \left(\left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[|Y|^p\right]\right)^{\frac{1}{p}}\right) \cdot \left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{q}}.$$

Upon dividing both sides by $\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{q}}$, we obtain

$$\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{1-\frac{1}{q}} \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[|Y|^p\right]\right)^{\frac{1}{p}},$$

and the proof of (13.20) is now complete since $1 - \frac{1}{q} = \frac{1}{p}$. ■

13.7 Jensen's inequality and its consequences

Several useful bounds involving moments of rvs are a byproduct of convexity; in its general form this is expressed through *Jensen's inequality* which is now discussed.

Recall the definition of convexity: A mapping $g : \mathbb{R} \to (-\infty, +\infty]$ is *convex* if the conditions

$$g((1-\lambda)x_0 + \lambda x_1) \leq (1-\lambda)g(x_0) + \lambda g(x_1), \qquad \begin{array}{c} \lambda \in [0,1] \\ x_0, x_1 \in (-\infty, +\infty] \end{array}$$

hold. The *effective domain* $\mathrm{Dom}(g)$ of g is the subset of $\mathbb{R}$ given by

$$\mathrm{Dom}(g) = \{x \in \mathbb{R} : \ g(x) \in \mathbb{R}\} .$$

Theorem 13.7.1 *(Jensen's inequality) Consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$. For any convex mapping $g : \mathbb{R} \to (-\infty, +\infty]$, whenever $\mathbb{E}[X]$ belongs to $\mathrm{Dom}(g)$, it holds that*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)] \tag{13.24}$$

if $\mathbb{E}[g(X)^-] < \infty$.

The assumption $\mathbb{E}[|X|] < \infty$ ensures that $\mathbb{E}[X]$ is well defined and finite. As will become apparent in the forthcoming proof, $\mathbb{E}[g(X)]$ is *not* well defined if $\mathbb{E}[g(X)^-] = \infty$ for then it is necessarily the case that $\mathbb{E}[g(X)^+] = \infty$. However, under the condition $\mathbb{E}[g(X)^-] < \infty$, the expectation $\mathbb{E}[g(X)]$ is well defined, though possibly infinite; see an example below.

Proof. If $\mathbb{E}[X]$ belongs to $\mathrm{Dom}(g)$, then the sub-differential of g is well defined at that point and non-empty. Therefore, for any v in $\partial g(\mathbb{E}[X])$ it holds that

$$g(\mathbb{E}[X]) + v\,(x - \mathbb{E}[X]) \leq g(x), \quad x \in \mathbb{R}.$$

Evaluating this last inequality at $x = X$ and using the decomposition $g(X) = g(X)^+ - g(X)^-$ we conclude that

$$g(\mathbb{E}[X]) + v\,(X - \mathbb{E}[X]) + g(X)^- \leq g(X)^+. \tag{13.25}$$

Taking expectations we note the following: Under the assumption $\mathbb{E}[|X|] < \infty$, the expectation of the rv $g(\mathbb{E}[X]) + v\,(X - \mathbb{E}[X]) + g(X)^-$ is well defined and given by

$$g(\mathbb{E}[X]) + v\mathbb{E}[X - \mathbb{E}[X]] + \mathbb{E}\left[g(X)^-\right] = g(\mathbb{E}[X]) + \mathbb{E}\left[g(X)^-\right].$$

Thus, (13.25) now gives

$$g(\mathbb{E}[X]) + \mathbb{E}\left[g(X)^-\right] \leq \mathbb{E}\left[g(X)^+\right]; \tag{13.26}$$

this holds even if $\mathbb{E}[g(X)^+] = \infty$ and $\mathbb{E}[g(X)^-] = \infty$. However, if $\mathbb{E}[g(X)^-]$ is finite, then $\mathbb{E}[g(X)]$ is well defined in the usual manner as $\mathbb{E}[g(X)] = \mathbb{E}[g(X)^+] - \mathbb{E}[g(X)^-]$, and the desired conclusion (13.24) follows from (13.26). ■

Back to the variance We have shown earlier that if X is a second-order rv, then $(\mathbb{E}[X])^2 \leq \mathbb{E}\left[X^2\right]$. Note that this is also a simple consequence of Jensen's inequality applied to the mapping $g : \mathbb{R} \to \mathbb{R} : t \to t^2$ – Here $\mathbb{E}[g(X)^-] = 0$ and obviously there is no guarantee that $\mathbb{E}\left[X^2\right] < \infty$ simply because $\mathbb{E}[|X|]$ is finite as discussed above. In fact, it is trivially the case that $(\mathbb{E}[X])^2 \leq \mathbb{E}\left[X^2\right]$ when $\mathbb{E}\left[X^2\right] = \infty$ as soon as $\mathbb{E}[X]$ exists, finite or not.

Lyapounov's inequality In Section 13.1, with $1 \leq p < q$ we noted that $\mathbb{E}[|X|^p] < \infty$ holds whenever $\mathbb{E}[|X|^q] < \infty$, thereby suggesting a possible monotonicity for $\mathbb{E}[|X|^p]$ as a function of p. Lyapounov's inequality given next is an easy consequence of Jensen's inequality, and provides a more precise version of this suggestion.

Lemma 13.7.1 *(Lyapounov's inequality) For any rv $X : \Omega \to \mathbb{R}$, the monotonicity property*

$$(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq (\mathbb{E}[|X|^q])^{\frac{1}{q}} \quad 1 \leq p < q \tag{13.27}$$

holds.

Proof. Pick p and q such that $1 \leq p < q$, and introduce $\lambda > 1$ such that $q = \lambda p$. The mapping $x \to |x|^\lambda$ being convex on $\mathbb{R}$, we conclude with the help of Jensen's inequality that

$$\mathbb{E}[|X|^q] = \mathbb{E}\left[(|X|^p)^\lambda\right] \geq (\mathbb{E}[|X|^p])^\lambda .$$

Exponentiating both sides of this last inequality to power q^{-1} yields (13.27) since $\lambda q^{-1} = qp^{-1}q^{-1} = p^{-1}$. ■

13.8 On the way to normed spaces of rvs

For any rv $X : \Omega \to \mathbb{R}$ we write

$$\|X\|_p \equiv \left(|X|^p\right)^{\frac{1}{p}}, \quad p \geq 1. \tag{13.28}$$

With this notation, Lyapounov's inequality (13.27) states that the mapping $[1, \infty) \to [0, \infty] : p \to \|X\|_p$ is non-decreasing.

A semi-norm Fix $p \geq 1$: Let $\mathcal{L}_p$ denote the set of all rvs $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}\left[|X|^p\right] < \infty$. Equivalently, $\mathcal{L}_p$ can be defined as

$$\mathcal{L}_p \equiv \left\{ \text{Rv } X : \Omega \to \mathbb{R} : \ \|X\|_p < \infty \right\}.$$

Obviously, for arbitrary rvs X and Y in $\mathcal{L}_p$, it holds that

$$\|t \cdot X\|_p = |t| \cdot \|X\|_p, \quad t \in \mathbb{R} \tag{13.29}$$

while Minkowsk's inequality gives

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p. \tag{13.30}$$

The properties (13.29) and (13.30) are known as *(positive) homogeneity* and the *triangle inequality*, respectively; they turn the mapping $X \to \|X\|_p$ into a *semi-norm* on $\mathcal{L}_p$. While it is always the case that $\|X\|_p \geq 0$, the condition $\|X\|_p = 0$ only implies $X = 0$ a.s. (and not $X = 0$ on Ω), and the mapping $X \to \|X\|_p$ is therefore *not* a norm on $\mathcal{L}_p$ (as this would require $X = 0$ on Ω).

Constructing a norm There is a simple way to turn this semi-norm into a *bona fide* norm. Note that a.s. equality (under $\mathbb{P}$) defines an *equivalence relation*, denoted $\sim_{\mathbb{P}}$, on all rvs $\Omega \to \mathbb{R}$ defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$: For rvs $X, Y : \Omega \to \mathbb{R}$ defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, we write

$$X \sim_{\mathbb{P}} Y \quad \text{if and only if} \quad \mathbb{P}\left[X \neq Y\right] = 0.$$

With rv $X : \Omega \to \mathbb{R}$, the equivalence class $[X]_{\mathbb{P}}$ of X under $\sim_{\mathbb{P}}$ is given by

$$[X]_{\mathbb{P}} \equiv \left\{ \text{Rv } Y : \Omega \to \mathbb{R} : \ Y \sim_{\mathbb{P}} X \right\}.$$

Note that $[X]_{\mathbb{P}} = [Y]_{\mathbb{P}}$ whenever $X \sim_{\mathbb{P}} Y$, so that $\|Y\|_p = \|X\|_p$ for every rv Y in $[X]_{\mathbb{P}}$. It follows that $[X]_{\mathbb{P}}$ is a subset of $\mathcal{L}_p$ whenever X belongs to $\mathcal{L}_p$. The definition

$$L_p \equiv \left\{ [X]_{\mathbb{P}} : \ X \in \mathcal{L}_p \right\}$$

is therefore well posed, and set

$$\|[X]_{\mathbb{P}}\|_p \equiv \|Y\|_p, \quad [X]_{\mathbb{P}} \in L_p$$

where Y is any element in $\mathcal{L}_p$ belonging to $[X]_{\mathbb{P}}$. It is easy to check that this definition is also well posed (and independent of the selection of Y in $[X]_{\mathbb{P}}$), and that it defines a semi-norm on L_p. Moreover, $\|[X]_{\mathbb{P}}\|_p = 0$ is equivalent to $\|[X]_{\mathbb{P}} = [0]_{\mathbb{P}}$, hence it is a norm on L_p (since $[0]_{\mathbb{P}}$ is the zero element in L_p).

13.9 In the limit

Pick a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}\left[|X|^q\right] < \infty$ for all $q \geq 1$. The monotonicity property (13.27) ensures that the convergence $\lim_{q\to\infty} \|X\|_q$ takes place monotonically with the limit identified as

$$\lim_{q\to\infty} \|X\|_q = \sup_{q\geq 1} \|X\|_q. \tag{13.31}$$

The question naturally arises as to whether this limiting value can be given a more operational (and therefore more useful) form.

To that end we introduce the quantity

$$\|X\|_\infty \equiv \inf\left\{a > 0 : \mathbb{P}\left[|X| > a\right] = 0\right\} \tag{13.32}$$

with the understanding that $\|X\|_\infty = \infty$ if the set $\{a > 0 : \mathbb{P}\left[|X| > a\right] = 0\}$ is empty. This definition is well posed as an element of $[0, \infty]$ upon noting that $\mathbb{P}\left[|X| > a\right]$ decreases with increasing a.

The reader with some knowledge of Measure Theory will identify $\|X\|_\infty$ as the $\mathbb{P}$-Essential Supremum of the rv $|X|$. The main result concerning this quantity is given next.

Proposition 13.9.1 *For any rv $X : \Omega \to \mathbb{R}$ it holds that*

$$\sup_{q\geq 1} \|X\|_q = \|X\|_\infty. \tag{13.33}$$

Proof. Assume first that $\|X\|_\infty = 0$: Then, for all $a > 0$ we have $\mathbb{P}\left[|X| > a\right] = 0$, or equivalently $\mathbb{P}\left[|X| \leq a\right] = 1$. Obviously $X = 0$ a.s., and $\|X\|_q = 0$ for all $q \geq 1$, so that (13.33) automatically holds.

For the remainder of the proof we assume $\|X\|_\infty > 0$. Fix $q > 1$ and pick $a > 0$ arbitrary. The easy decomposition

$$\mathbb{E}\left[|X|^q\right] = \mathbb{E}\left[|X|^q \mathbf{1}\left[|X| \le a\right]\right] + \mathbb{E}\left[|X|^q \mathbf{1}\left[|X| > a\right]\right] \tag{13.34}$$

implies the bound

$$a^q \cdot \mathbb{E}\left[\mathbf{1}\left[|X| > a\right]\right] \le \mathbb{E}\left[|X|^q\right].$$

It then follows that

$$a \cdot \left(\mathbb{P}\left[|X| > a\right]\right)^{\frac{1}{q}} \le \left(\mathbb{E}\left[|X|^q\right]\right)^{\frac{1}{q}}, \quad q > 1. \tag{13.35}$$

If $\|X\|_\infty = \infty$, then the set $\{a > 0 : \ \mathbb{P}\left[|X| > a\right] = 0\}$ being empty, we have $\mathbb{P}\left[|X| > a\right] > 0$ for all $a > 0$, in which case $\lim_{q\to\infty}\left(\mathbb{P}\left[|X| > a\right]\right)^{\frac{1}{q}} = 1$. Letting q go to infinity in (13.35) we conclude that $a \le \sup_{q\ge 1}\|X\|_q$ for each $a > 0$. Therefore, $\sup_{q\ge 1}\|X\|_q = \infty$ and we obtain (13.33).

If $0 < \|X\|_\infty < \infty$, then on the range $0 < a < \|X\|_\infty$, we have $\mathbb{P}\left[|X| > a\right] > 0$ and again it follows that $\lim_{q\to\infty}\left(\mathbb{P}\left[|X| > a\right]\right)^{\frac{1}{q}} = 1$. Letting q go to infinity in (13.35) we conclude that $a \le \sup_{q\ge 1}\|X\|_q$ whenever $0 < a < \|X\|_\infty$, whence $\|X\|_\infty \le \sup_{q\ge 1}\|X\|_q$. This shows that (13.33) holds.

Next, still under the condition $\|X\|_\infty < \infty$, for $a > \|X\|_\infty$ it holds that $\mathbb{P}\left[|X| > a\right] = 0$, or equivalently $\mathbb{E}\left[\mathbf{1}\left[|X| > a\right]\right] = 0$ so that $\mathbf{1}\left[|X| > a\right] = 0$ a.s., whence $|X|^q \mathbf{1}\left[|X| > a\right] = 0$ a.s. It then follows from (13.34) that

$$\mathbb{E}\left[|X|^q\right] = \mathbb{E}\left[|X|^q \mathbf{1}\left[|X| \le a\right]\right] \le a^q \cdot \mathbb{P}\left[|X| \le a\right] = a^q, \qquad \begin{array}{c} \|X\|_\infty < a \\ q > 1 \end{array}$$

and we obtain the bounds $\left(\mathbb{E}\left[|X|^q\right]\right)^{\frac{1}{q}} \le a$ on the range $\|X\|_\infty < a$. It follows that $\sup_{q\ge 1}\|X\|_q \le a$ whenever $\|X\|_\infty < a$, whence $\sup_{q\ge 1}\|X\|_q \le \|X\|_\infty$. Combining with the earlier conclusion we get the equality (13.33) under the condition $0 < \|X\|_\infty < \infty$. ■

13.10 Hölder's inequality when $p = 1$

The attentive reader may have noticed that Hölder's inequality was given only for conjugate pairs p and q such that $p, q > 1$, and it is therefore natural to wonder what happens when $p = 1$: Formally the defining relation (13.16) suggests that the conjugate q of $p = 1$ should be taken to be $q = \infty$.

Theorem 13.10.1 *(Hölder's inequality when $p = 1$) For rvs $X, Y : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$ and $\|Y\|_\infty < \infty$, it holds that*

$$\mathbb{E}[|X| \cdot |Y|] \le \mathbb{E}[|X|] \cdot \|Y\|_\infty. \tag{13.36}$$

Proof. First note that whenever $\|Y\|_\infty$ is finite, then $\|Y\|_q < \infty$ for all $q \ge 1$ by virtue of Proposition 13.9.1. Next we pick $n = 1, 2, \ldots$ and fix $p > 1$: It is plain that $\| \min(n, |X|) \|_p < \infty$. Applying Hölder's inequality for $p > 1$ and its conjugate $q = \frac{p}{p-1}$ we get

$$\mathbb{E}[| \min(n, |X|) \cdot Y|] \le \mathbb{E}[| \min(n, |X|) |^p]^{\frac{1}{p}} \cdot \mathbb{E}[|Y|^q]^{\frac{1}{q}} .$$

Now let p go down to 1, say $p \downarrow 1$ (so that $q \uparrow \infty$), in this last inequality: Bounded convergence yields

$$\lim_{p \downarrow 1} \mathbb{E}[| \min(n, |X|) |^p] = \mathbb{E}[\min(n, |X|)] ,$$

whence $\lim_{p \downarrow 1} \mathbb{E}[| \min(n, |X|) |^p]^{\frac{1}{p}} = \mathbb{E}[\min(n, |X|)]$. On the other hand, we get $\lim_{q \uparrow \infty} \mathbb{E}[|Y|^q]^{\frac{1}{q}} = \|Y\|_\infty$ by Proposition 13.9.1. Collecting these facts we conclude that

$$\mathbb{E}[| \min(n, |X|) \cdot Y|] \le \mathbb{E}[\min(n, |X|)] \cdot \|Y\|_\infty, \quad n = 1, 2, \ldots$$

Let n go to infinity in this last inequality. By monotone convergence we get both

$$\lim_{n \to \infty} \mathbb{E}[| \min(n, |X|) \cdot Y|] = \mathbb{E}\left[\lim_{n \to \infty} \min(n, |X|) \cdot |Y|\right] = \mathbb{E}[|X| \cdot |Y|]$$

and $\lim_{n \to \infty} \mathbb{E}[\min(n, |X|)] = \mathbb{E}[|X|]$. The conclusion $\mathbb{E}[|X| \cdot |Y|] \le \mathbb{E}[|X|] \cdot \|Y\|_\infty$ follows, and the proof of (13.36) is now complete. ■

13.11 Exercises

All rvs are defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 13.1 (Optimality of the first moment) The following fact lies at the root of the popular Minimum Mean Square Estimation (MMSE) procedures: For a second-order rv $X : \Omega \to \mathbb{R}$, show that

$$\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right] \leq \mathbb{E}\left[|X - a|^2\right], \quad a \in \mathbb{R}.$$

In other words, the first moment $\mathbb{E}[X]$ solves the minimization problem

$$\text{Minimize } \left\{\mathbb{E}\left[|X - a|^2\right], \, a \in \mathbb{R}\right\}.$$

Ex. 13.2 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$ such that either $\mathbb{E}[X] = \mathbb{E}[Y]$ or $\mathbb{E}[X] = -\mathbb{E}[Y]$. Give conditions for the rvs $X - Y$ and $X + Y$ to be uncorrelated.

Ex. 13.3 Let N be a discrete rv with support contained in $\mathbb{N}_0$ (i.e., $\mathbb{P}[N \in \mathbb{N}_0] = 1$) with a finite second moment, i.e., $\mathbb{E}\left[N^2\right] < \infty$. Also let $\{X_n, \ n = 1, 2, \ldots\}$ denote a collection of second-order rvs. Assume the rvs $\{N, X_n, \ n = 1, 2, \ldots\}$ to be mutually independent.

a. Compute the first moment $\mathbb{E}\left[\sum_{n=1}^{N} X_n\right]$.

b. Compute the variance $\text{Var}\left[\sum_{n=1}^{N} X_n\right]$.

c. Specialize the results of Parts **a** and **b** when the rvs $\{X_n, \ n = 1, 2, \ldots\}$ have identical mean and variance, namely $\mu \equiv \mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots$ and $\sigma^2 \equiv \text{Var}[X_1] = \text{Var}[X_2] = \ldots$

Ex. 13.4 Let N be a discrete rv with support contained in $\mathbb{N}_0$ (i.e., $\mathbb{P}[N \in \mathbb{N}_0] = 1$) with a finite second moment, i.e., $\mathbb{E}\left[N^2\right] < \infty$. Also let $\{X_n, \ n = 1, 2, \ldots\}$ denote a collection of second-order rvs. Assume the rvs $\{N, X_n, \ n = 1, 2, \ldots\}$ to be mutually independent.

a. Compute the first moment

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right].$$

b. Compute the variance

$$\text{Var}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right].$$

c. Specialize the results of Parts **a** and **b** when the rvs $\{X_n, \ n = 1, 2, \ldots\}$ have identical mean and variance, namely $\mu \equiv \mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots$ and $\sigma^2 \equiv \text{Var}[X_1] = \text{Var}[X_2] = \ldots$

Ex. 13.5 We start with a collection $U_1, U_2, \ldots, U_n$ of n rvs, each uniformly distributed over the interval $(0,1)$. Also available is a rv P with the property that $\mathbb{P}[0 < P \leq 1] = 1$. Assume that the $n+1$ rvs $P, U_1, \ldots, U_n$ are mutually independent rvs, and that the rv P is a discrete rv with $\mathbb{P}[P \in S] = 1$ for some countable subset $S \subseteq (0,1]$.[1]

a. For $i, j = 1, 2, \ldots, n$, compute $\mathbb{E}[\mathbf{1}[U_i \leq P]]$ and $\text{Cov}[\mathbf{1}[U_i \leq P], \mathbf{1}[U_j \leq P]]$ (in terms of momnents of P)

b. For $i, j = 1, 2, \ldots, n$, are the rvs $\mathbf{1}[U_i \leq P]$ and $\mathbf{1}[U_j \leq P]$ uncorrelated?

c. Under what conditions are the rvs $\mathbf{1}[U_1 \leq P], \ldots, \mathbf{1}[U_n \leq P]$ uncorrelated? pairwise independent? mutually independent?

Ex. 13.6 The setting is that of Exercise 13.5. Under the assumptions there, we are interested in the rv X defined by

$$X \equiv \sum_{i=1}^{n} \mathbf{1}[U_i \leq P].$$

a. Compute $\mathbb{E}[X]$ in terms of $\mathbb{E}[P]$.

b. How many moments of P are needed to compute $\text{Var}[X]$?

c. When S contains at least two elements, are the rvs $\mathbf{1}[U_1 \leq P], \ldots, \mathbf{1}[U_n \leq P]$ (i) mutually independent (ii) pairwise uncorrelated ?

d. Compute the probabilities

$$\mathbb{P}[X = k], \quad k = 0, 1, \ldots, n.$$

How many moments of P are needed?

Ex. 13.7 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$ such that $X + Y = a$ a.s. for some constant a.

a. Show that it is always the case that $\text{Cov}[X, Y] \leq 0$.

b. Under what conditions are the rvs X and Y negatively correlated?

Ex. 13.8 With $0 < p < 1$, let $X(p), Y(p) : \Omega \to \mathbb{R}$ be a pair of independent Bernoulli rvs with $\mathbb{P}[X(p) = 1] = 1 - \mathbb{P}[X(p) = 0] = p$ and $\mathbb{P}[Y(p) = 1] = 1 - \mathbb{P}[Y(p) = 0] = p$.

a. Compute the covariance $\text{Cov}[|X(p) - Y(p)|, X(p) + Y(p)]$ between the rvs $|X(p) - Y(p)|$ and $X(p) + Y(p)$ as a function of p.

[1] The results of Exercises 13.5 and 13.6 also hold for any rv P, discrete or not, satisfying $0 < P \leq 1$ a.s. However the general case is best handled through the use of conditional expectations discussed in Chapter **??**.

b. Find all the values of p in $(0,1)$ such that rvs $|X(p)-Y(p)|$ and $X(p)+Y(p)$ are uncorrelated rvs.

c. Find all the values of p in $(0,1)$ such that rvs $|X(p)-Y(p)|$ and $X(p)+Y(p)$ are independent rvs. [HINT: For the values of p in Part **b**, compute the joint probability $\mathbb{P}\left[|X(p)-Y(p)|=0, X(p)+Y(p)=0\right]$ and compare it to the product $\mathbb{P}\left[|X(p)-Y(p)|=0\right]\mathbb{P}\left[X(p)+Y(p)=0\right]$.]

Ex. 13.9 Consider four second-order rvs $X_1, X_2, Y_1, Y_2 : \Omega \to \mathbb{R}$. If the two-dimensonal rvs $X : \Omega \to \mathbb{R}^2$ and $Y : \Omega \to \mathbb{R}^2$ are independent where $X = (X_1, X_2)$ and (Y_1, Y_2), show that

$$\text{Cov}\left[X_1+Y_1, X_2+Y_2\right] = \text{Cov}\left[X_1, X_2\right] + \text{Cov}\left[Y_1, Y_2\right].$$

Ex. 13.10 We start with the rv $U : \Omega \to \mathbb{R}$ which has a symmetric distribution, i.e., $U =_{st} -U$ (where as usual $=_{st}$ refers to equality in distribution). Given are two Borel mappings $f, g : \mathbb{R} \to \mathbb{R}$ with the following properties: $f(-x) = -f(x)$ and $g(-x) = g(x)$ for all x in $\mathbb{R}$. Assume that $\mathbb{E}\left[|f(U)|^2\right] < \infty$ and $\mathbb{E}\left[|g(U)|^2\right] < \infty$,

a. Show that $\text{Cov}\left[f(U), g(U)\right] = 0$, i.e., the rvs $X = f(U)$ and $Y = g(U)$ are always uncorrelated.

b. Consider now the case when $f(x) = \sin(x)$ and $g(x) = \cos(x)$ for all x in $\mathbb{R}$ (in which case $|f(x)|^2 + |g(x)|^2 = 1$). In the spirit of Exercise 13.7, show that $\text{Cov}\left[|f(U)|^2, |g(U)|^2\right] < 0$ with the implication that the rvs $|X|^2$ and $|Y|^2$ are not independent, and *a fortiori* X and Y cannot be independent!

Ex. 13.11 Let A and B be events in $\mathcal{F}$, and let $X_A = \mathbf{1}[A]$ and $X_B = \mathbf{1}[B]$ denote their indicator functions.

a. Compute the covariance $\text{Cov}\left[X_A, X_B\right]$

b. Show that rvs X_A and X_B are uncorrelated if and only if the events A and B are independent.

c. Show that the rvs X_A and X_B are independent if and only if the rvs X_A and X_B are uncorrelated.

Ex. 13.12 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$ which are independent with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Define the rvs $U, V : \Omega \to \mathbb{R}$ by $U \equiv \min(X, Y)$ and $V \equiv \max(X, Y)$.

a. Compute $\mathbb{E}[UV]$.

b. Are the rvs U and V second-order rvs?.

c. Are the rvs U and V uncorrelated? Are they independent?

Ex. 13.13 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$. We assume that (i) the rvs X and Y are independent rvs and that (ii) each has a symmetric distribution, i.e., $X =_{st} -X$ and $Y =_{st} -Y$. Define the rvs $U, V : \Omega \to \mathbb{R}$ by $U \equiv \min(X, Y)$ and $V \equiv \max(X, Y)$.

a. Show that $V =_{st} -U$.

b. Compute $\text{Cov}\,[U, V]$

c. Under what additional conditions are the rvs U and V uncorrelated?

Ex. 13.14 Consider three second-order rvs $X, Y, Z : \Omega \to \mathbb{R}$ such that (i) $X + Y + Z = 1$ a.s. and (ii) $\text{Var}\,[X] \leq \text{Var}\,[Y] \leq \text{Var}\,[Z]$.

a. Show that the rvs X and Z (resp. Y and Z) are negatively correlated in the sense that $\text{Cov}\,[X, Z] \leq 0$ (resp. $\text{Cov}\,[Y, Z] \leq 0$).

b. Show that $\text{Cov}\,[X, Y] \geq 0$ (i.e., the rvs X and Y are positively correlated) if and only if

$$\text{Var}\,[X] + \text{Var}\,[Y] \leq \text{Var}\,[Z]\,.$$

c. Show that it is always the case that $|\text{Cov}\,[X, Z]| \leq |\text{Cov}\,[Y, Z]|$.

Ex. 13.15 Consider a rv $X : \Omega \to \mathbb{R}$.

a. Show that there always exists a scalar M in $\mathbb{R}$ such that

$$\mathbb{P}\,[X \leq M] \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}\,[X \geq M] \geq \frac{1}{2}.$$

Such a scalar is called a *median* for the probability distribution function of the rv X. Is such scalar unique?

b. Let $F_X : \mathbb{R} \to [0, 1]$ denotes the probability distribution function of the rv X. If $F_X : \mathbb{R} \to [0, 1]$ is a *strictly increasing* and *continuous* function, show that there is only one median and it is characterized by

$$F_X(t) = \frac{1}{2}, \quad t \in \mathbb{R}$$

c. If $\mathbb{E}\,[|X|] < \infty$, then show that

$$\mathbb{E}\,[|X - M|] \leq \mathbb{E}\,[|X - a|]\,, \quad a \in \mathbb{R}$$

for every median M of X.

Ex. 13.16 The following (important) inequality happens to be true: With $f, g : \mathbb{R} \to \mathbb{R}$ monotone non-decreasing mappings,[2] it holds that

$$\mathbb{E}\,[f(X)] \cdot \mathbb{E}\,[g(X)] \leq \mathbb{E}\,[f(X) \cdot g(X)] \tag{13.37}$$

[2] This means $f(x) \leq f(y)$ and $g(x) \leq g(y)$ whenever $x < y$ in $\mathbb{R}$

whenever the expectations are well defined, e.g., when f and g take non-negative values.

a. Try to give a proof of this fact when X is a discrete rv with support on some countable set $S \subseteq \mathbb{R}$ with pdf $\boldsymbol{p}_X = (\mathbb{P}[X = x], \ x \in S)$. So you will need to show that

$$\left(\sum_{x \in S} f(x) p_X(x)\right) \cdot \left(\sum_{x \in S} g(x) p_X(x)\right) \leq \sum_{x \in S} f(x) g(x) p_X(x).$$

It is feasible but not a pleasant exercise. Just think about it for a few minutes! Here is now a *probabilistic* proof in a few lines, said proof illustrating the power of probabilistic thinking!

b. Explain why it is always the case that

$$\Delta(y,z) \equiv (f(y) - f(z)) \cdot (g(y) - g(z)) \geq 0, \quad y, z \in \mathbb{R}.$$

c. Let Y and Z be two *independent* rvs $\Omega \to \mathbb{R}$, each with the same probability distribution as X. What is the sign of $\mathbb{E}[\Delta(Y,Z)]$?

d. Use Part **c** to conclude that (13.37) holds!

Ex. 13.17 Consider a rv $X : \Omega \to \mathbb{R}$ defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The quantity $\|X\|_\infty$ is completely determined by the probability distribution of the rv $|X|$ (through the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$). More precisely: For $k = 1, 2$, the rv $X_k : \Omega_k \to \mathbb{R}$ is a rv defined on the probability triple $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$. If the distribution of $|X_1|$ under $\mathbb{P}_1$ coincides with that of $|X_2|$ under $\mathbb{P}_2$, then the quantities $\|X_1\|_\infty$ (computed under $\mathbb{P}_1$) and $\|X_2\|_\infty$ (computed under $\mathbb{P}_2$) coincide.

a. Show that $\|X\|_\infty \leq \sup\{|X(\omega)| : \omega \in \Omega\}$.

b. Give an example when this inequality is strict with both quantities $\|X\|_\infty$ and $\sup\{|X(\omega)| : \omega \in \Omega\}$ finite

c. Give an example when $\|X\|_\infty$ is finite but $\sup\{|X(\omega)| : \omega \in \Omega\} = \infty$.

Ex. 13.18 For any rv $X : \Omega \to \mathbb{R}$, it was noted that $\mathbb{E}[|X|^p] < \infty$ for all p in the interval $[1, q]$ as soon as $\mathbb{E}[|X|^q] < \infty$ for some $q \geq 1$, an observation which translates into the nested inclusions $\mathcal{L}_q \subseteq \mathcal{L}_p \subseteq \mathcal{L}_1$ when $1 \leq p < q$. It is natural to wonder as to what is the set $\cap_{q \geq 1} \mathcal{L}_q$. In view of Proposition 13.9.1 it might be tempted to conclude that $\cap_{q \geq 1} \mathcal{L}_q = \mathcal{L}_\infty$ where

$$\mathcal{L}_\infty \equiv \{\text{Rv } X : \Omega \to \mathbb{R} : \|X\|_\infty < \infty\}.$$

Show by a conterexample that this guess is incorrect and that $\mathcal{L}_\infty$ is a strict subset of $\cap_{q \geq 1} \mathcal{L}_q$ – In other words, it is possible for $\mathbb{E}[|X|^q] < \infty$ for every $q \geq 1$ and yet $\|X\|_\infty = \infty$. [HINT: Consider a Gaussian rv or an exponetial rv.]

Ex. 13.19 Show that (13.32) defines a semi-norm on the linear space of all rvs defined on $(\Omega, \mathcal{F}, \mathbb{P})$, namely for arbitrary rvs $X, Y : \Omega \to \mathbb{R}$, it holds that $\|tX\|_\infty = |t|\|X\|_\infty$ for each t in $\mathbb{R}$ and $\|X + Y\|_\infty \leq \|X\|_\infty + \|Y\|_\infty$.

a. Prove these facts by a limiting argument based on (13.31) and Proposition 13.9.1

b. Prove these facts by a direct argument based on the definition (13.32).

Chapter 14

Bounding probabilities

All rvs are defined as Borel measurable mappings $\Omega \to \mathbb{R}$ defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and all probability distributions are computed under $\mathbb{P}$.

14.1 Markov's inequality and consequences

In this section we present a number of inequalities that prove useful in many contexts. They are all implied by the very simple observation embedded in Markov's inequality.

Theorem 14.1.1 *(Markov's inequality) For any rv $X : \Omega \to \mathbb{R}$ with $X \geq 0$ a.s., it holds that*

$$\mathbb{P}\left[X \geq t\right] \leq \frac{1}{t} \cdot \mathbb{E}\left[X\right], \quad t > 0. \tag{14.1}$$

Markov's inequality can be quite poor for some values of $t > 0$: Indeed the bound will be useless whenever $\mathbb{E}\left[X\right] > t$ (as will occur for small t) since yielding a right hand side greater than unity! However, even for large values of t the bound may not capture the tail behavior of $\mathbb{P}\left[X > t\right]$. Here is an example: If X is an exponential rv with unit parameter, then $\mathbb{P}\left[X > t\right] = e^{-t}$ for all $t > 0$ while $\mathbb{E}\left[X\right] = 1$. Clearly e^{-t} decays much faster than t^{-1}.

Proof. As the bound trivially holds if $\mathbb{E}\left[X\right] = \infty$, it suffices to consider the case when $\mathbb{E}\left[X\right]$ is finite: Fix $t > 0$, and note that

$$\begin{aligned} X &= X \cdot \mathbf{1}\left[X < t\right] + X \cdot \mathbf{1}\left[X \geq t\right] \\ &\geq X \cdot \mathbf{1}\left[X \geq t\right] \quad a.s. \\ &\geq t \cdot \mathbf{1}\left[X \geq t\right] \quad a.s. \end{aligned}$$

Taking expectations in this last inequality, we find that $t \cdot \mathbb{E}\left[\mathbf{1}\left[X \geq t\right]\right] \leq \mathbb{E}\left[X\right]$, and the conclusion (14.1) follows. ■

Markov's inequality gives rise to several useful inequalities.

The Bienaymé-Tchebychev inequality Consider a second-order rv X. Applying Markov's inequality to the rv $(X - \mathbb{E}\left[X\right])^2$ we get

$$\mathbb{P}\left[(X - \mathbb{E}\left[X\right])^2 \geq t^2\right] \leq \frac{1}{t^2} \cdot \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right], \quad t > 0.$$

This is often written in the equivalent form

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq t\right] \leq \frac{1}{t^2} \cdot \mathrm{Var}\left[X\right], \quad t > 0 \tag{14.2}$$

and is known as the *Bienaymé-Tchebychev* inequality.

A particularly useful form arises when using (14.2) with $t = \lambda\sigma(X)$ for some $\lambda > 0$, in which case we get

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq \lambda\sigma(X)\right] \leq \frac{1}{\lambda^2}, \quad \lambda > 0. \tag{14.3}$$

This easy fact is often used in statistical studies to assert that the probability that an observed rv X deviates from its mean $\mathbb{E}\left[X\right]$ by at least 3 standard deviations is less than $1/9$.

Chernoff bounds As another application of Markov's inequality consider the following observation: Fix t in $\mathbb{R}$. With $\theta > 0$, we note that $X \geq t$ if and only if $e^{\theta X} \geq e^{\theta t}$, hence

$$\mathbb{P}\left[X \geq t\right] = \mathbb{P}\left[e^{\theta X} \geq e^{\theta t}\right].$$

Applying Markov's inequality to the rv $e^{\theta X}$ yields

$$\mathbb{P}\left[e^{\theta X} \geq e^{\theta t}\right] \leq e^{-\theta t}\mathbb{E}\left[e^{\theta X}\right].$$

Collecting these facts we conclude that

$$\mathbb{P}\left[X > t\right] \leq e^{-\theta t}\mathbb{E}\left[e^{\theta X}\right]. \tag{14.4}$$

Such a bound is known as a *Chernoff* bound. Note that $\mathbb{E}\left[e^{\theta X}\right]$ always exists, although possibly infinite. However, only when $\mathbb{E}\left[e^{\theta X}\right] < \infty$ with $\theta > 0$ will the bound just derived be useful.

Thus far, $\theta > 0$ is a free parameter, hence we can seek to optimize the bound (14.4) by selecting $\theta > 0$ that achieves the best possible upper bound. This yields

$$\mathbb{P}\left[X \geq t\right] \leq \inf\left(e^{-\theta t}\mathbb{E}\left[e^{\theta X}\right] : \begin{array}{c} \theta > 0, \\ \mathbb{E}\left[e^{\theta X}\right] < \infty \end{array} \right). \tag{14.5}$$

A general Markov inequality The idea behind the Bienaymé-Tchebychev inequality and the Chernoff bound can be further generalized as follows: Consider a *non-decreasing* function $g : \mathbb{R} \to \mathbb{R}_+$ such that $\mathbb{E}\left[g(X)\right]$ is finite. Fix t in $\mathbb{R}$. Under these conditions, we note that $X \geq t$ implies $g(X) \geq g(t)$, whence $\mathbb{P}\left[X \geq t\right] \leq \mathbb{P}\left[g(X) \geq g(t)\right]$, and the basic Markov inequality (applied to $g(X)$) yields

$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}\left[g(X)\right]}{g(t)}$$

whenever $g(t) > 0$.

14.2 Concentration inequalities

Hoeffding's inequality

Bernstein's inequality

Bennett's inequality

Azuma's inequality

14.3 Exercises

All rvs are defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 14.1 Is it possible for a non-negative rv X to satisfy Markov's inequality (14.1) as an equality for *all* $t \geq EX$, namely

$$\mathbb{P}\left[X \geq t\right] = \frac{t}{\mathbb{E}\left[X\right]}, \quad t \geq \mathbb{E}\left[X\right].$$

Explain!

Ex. 14.2 (14.1) still holds.

Ex. 14.3 If the rv X satisfies $\mathbb{E}\left[2^X\right] = 4$, show that $\mathbb{P}\left[X \geq 3\right] \leq \frac{1}{2}$.

Ex. 14.4 Consider a collection $\{X_n,\ n = 1, 2, \ldots\}$ of second-order $\mathbb{R}$-valued rvs such that $\mathbb{E}\left[|X_n|^2\right] = c > 0$ for all $n = 1, 2, \ldots$.

a. Show that $\mathbb{P}\left[X_n > n \text{ i.o.}\right] = 0$.

b. Still assuming that $\mathbb{E}\left[|X_n|^2\right] = a_n < \infty$ for all $n = 1, 2, \ldots$, find a strictly weaker condition on the second moments, hence on the sequence $\{a_n,\ n = 1, 2, \ldots\}$, that ensures $\mathbb{P}\left[X_n > n \text{ i.o.}\right] = 0$.

c. Assume now that there exists $\nu > 0$ such that $\mathbb{E}\left[|X_n|^{1+\nu}\right] = a_n < \infty$ for all $n = 1, 2, \ldots$, find a condition on the sequence $\{a_n,\ n = 1, 2, \ldots\}$ that ensures $\mathbb{P}\left[X_n > n \text{ i.o.}\right] = 0$.

Ex. 14.5 For any rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{P}\left[X \in \mathbb{N}\right] = 1$, i.e., $X \in \mathbb{N}$ a.s., show that the inequality $\mathbb{P}\left[X > 0\right] \leq \mathbb{E}\left[X\right]$ always holds. This simple observation is the basis for the method of first moment often used in the theory of random graphs and in Combinatorics.

Ex. 14.6 For any second-order rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{P}\left[X \geq 0\right] = 1$, i.e., $X \geq 0$ a.s., show that

$$\frac{\left(\mathbb{E}\left[X\right]\right)^2}{\mathbb{E}\left[X^2\right]} \leq \mathbb{P}\left[X > 0\right]$$

provided $\mathbb{E}\left[X^2\right] > 0$ [**HINT:** Note that $X = X \cdot \mathbf{1}\left[X > 0\right]$ a.s. and apply the Cauchy-Schwartz inequality]. This inequality is the starting point for the method of second moment often used in the theory of random graphs and in Combinatorics where it is applied to integer-valued count rvs in the form

$$\mathbb{P}\left[X = 0\right] \leq 1 - \frac{\left(\mathbb{E}\left[X\right]\right)^2}{\mathbb{E}\left[X^2\right]}.$$

Ex. 14.7 With Hólder's inequality generalizing the Cauchy-Schwartz inequality, Exercise 14.6 suggests the following inequality:

a. Consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{P}\left[X \geq 0\right] = 1$, i.e., $X \geq 0$ a.s. with $\mathbb{E}\left[|X|^p\right] < \infty$ for some $p > 1$. With q the conjugate of p, show that

$$\left(\frac{\mathbb{E}\left[X\right]}{\left(\mathbb{E}\left[X^p\right]\right)^{\frac{1}{p}}}\right)^q \leq \mathbb{P}\left[X > 0\right]$$

provided $\mathbb{E}\left[|X|^p\right] > 0$ [**HINT:** Note that $X = X \cdot \mathbf{1}\left[X > 0\right]$ a.s. and apply Hólder's inequality].

b. Apply the result of Part **a** when X is an exponential rv with unit parameter and p is an integer, and explore how the bounds improve as p increases.

Chapter 15

Conditional expectations: The case of partitions

We now turn to the important notion of conditioning in its various forms. In this chapter the focus will be on the notion of conditional expectations with respect to a partition. The general notion of conditional expectation is then developed in Chapter 16.

Throughout we assume given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, and all the rvs are defined on it.

15.1 Conditional distributions and their expectations

We first return to the definitions given in Section 2.5: Let D be an event in $\mathcal{F}$ – We shall often refer to D as the c*onditioning* event. With $\mathbb{P}[D] > 0$, the conditional probability measure $\mathbb{Q}_D : \mathcal{F} \to [0,1]$ is a well-defined probability measure given by

$$\mathbb{Q}_D(E) = \frac{\mathbb{P}[E \cap D]}{\mathbb{P}[D]}, \quad E \in \mathcal{F}.$$

When $\mathbb{P}[D] = 0$, it is convenient to take $\mathbb{Q}_D : \mathcal{F} \to [0,1]$ to be an arbitrary probability measure on $(\Omega, \mathcal{F})$ – This issue will be revisited at some later time.

Assume $\mathbb{P}[D] > 0$. It is now possible to define the *conditional expectation* of the rv X *given* D, denoted $\mathbb{E}[X|D]$: It is simply the expectation of the rv X evaluated under the conditional probability measure $\mathbb{Q}_D$ defined on $(\Omega, \mathcal{F})$. The requirement $\mathbb{E}[\mathbf{1}[E]|D] = \mathbb{Q}_D[A]$ has to be satisfied for any event E in $\mathcal{F}$ when constructing the mathematical expectation operator associated with $\mathbb{Q}_D$. These conditions imply that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left[E\right]|D\right] &= \mathbb{Q}_D\left[E\right] \\ &= \frac{\mathbb{P}\left[E\cap D\right]}{\mathbb{P}\left[D\right]} \\ &= \frac{\mathbb{E}\left[\mathbf{1}\left[D\right]\mathbf{1}\left[E\right]\right]}{\mathbb{P}\left[D\right]}, \quad E\in\mathcal{F}. \end{aligned} \tag{15.1}$$

This observation leads readily to the following characterization.

Lemma 15.1.1 *Let D be an event in $\mathcal{F}$ such that $\mathbb{P}\left[D\right]>0$. For any rv $X:\Omega\to\mathbb{R}$, the conditional expectation of X given D exists (resp. exists and is finite) if the expectation $\mathbb{E}\left[X\right]$ exists (resp. exists and is finite), in which case the relation*

$$\mathbb{E}\left[X|D\right]=\frac{\mathbb{E}\left[\mathbf{1}\left[D\right]X\right]}{\mathbb{P}\left[D\right]} \tag{15.2}$$

holds.

Proof. The proof is carried out through the usual three step process: It holds for indicator rvs by virtue of (15.1), thus for simple rvs by linearity of expectation. Non-negative rvs are handled via a staircase approximation argument with simple rvs, and the case of arbitrary rvs uses the standard decomposition into the positive and negative parts. Details are left to the interested reader. ■

When $\mathbb{P}\left[D\right]=0$, it is convenient to take $\mathbb{Q}_D:\mathcal{F}\to[0,1]$ to be an arbitrary probability measure on $(\Omega,\mathcal{F})$, say even $\mathbb{P}$, for the sake of concreteness; such a choice will allow us to make sense of $\mathbb{E}\left[X|D\right]$ as an expectation of the rv X under the selected probability measure. However, regardless of the choice made for $\mathbb{Q}_D$ it is always the case that

$$\mathbb{E}\left[\mathbf{1}\left[D\right]X\right]=\mathbb{E}\left[X|D\right]\cdot\mathbb{P}\left[D\right] \tag{15.3}$$

as this is true for indicator rvs. More generally it can also be seen as follows: By Property **F** of expectation we have $\mathbb{E}\left[\mathbf{1}\left[D\right]X\right]=0$ if $\mathbb{P}\left[D\right]=0$ since then $\mathbf{1}\left[D\right]X=0$ a.s. – See Section 11.2.

15.2 Conditioning with respect to a partition

The next step on our way towards a general notion of conditional expectation passes through the notion of *conditional expectation given a countable partition*.

Given is a *countable* $\mathcal{F}$-partition $\{D_i,\ i \in I\}$ of Ω; see Definition 10.2.1 where I is now a countable index set (instead of a finite index set as was required in the definition of simple rvs). The condition $\cup_{i\in I} D_i = \Omega$ implies

$$\sum_{i\in I} \mathbf{1}[D_i] = 1. \tag{15.4}$$

The events in the partition are assumed to be non-empty although it is possible to have $\mathbb{P}[D_i] = 0$ for some i in I (but not for all indices since $\mathbb{P}[\Omega] = 1$). If the event D in $\mathcal{F}$ is of the form

$$D = \cup_{j\in J} D_j$$

for some $J \subseteq I$, then the decomposition

$$\sum_{j\in J} \mathbf{1}[D_j] = \mathbf{1}[D] \tag{15.5}$$

generalizes (15.4) and will be used on several occasions.

Definition 15.2.1

Consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[X]$ exists. The conditional expectation of X given the countable $\mathcal{F}$-partition $\{D_i,\ i \in I\}$ is the extended rv $\Omega \to [-\infty, \infty]$ defined by

$$\mathbb{E}[X|D_i,\ i \in I] \equiv \sum_{i\in I} \mathbb{E}[X|D_i]\, \mathbf{1}[D_i] \tag{15.6}$$

where for each i in I, $\mathbb{E}[X|D_i]$ is the expectation of X under the conditional probability distribution of X given D_i.

Definition 15.2.1 is well posed as a result of Lemma 15.1.1. In particular, the rv $\mathbb{E}[X|D_i,\ i \in I]$ is an $\mathbb{R}$-valued rv as soon as $\mathbb{E}[X]$ is finite. We stress that $\mathbb{E}[X|D_i,\ i \in I]$ is a *random variable* and not merely a constant – See (15.7) below. This rv *compactly* encodes *all* the conditional expectations $\{\mathbb{E}[X|D_i],\ i \in I\}$, hence the notation $\mathbb{E}[X|D_i,\ i \in I]$. The advantage of collecting all these conditional expectations into a *single* mathematical construct will become apparent when carrying out computations.

15.3 Elementary properties

We now present some elementary properties of the conditional expectation of X given the countable $\mathcal{F}$-partition $\{D_i,\ i \in I\}$. The key observation, derived from

(15.6), is that for every i in I, the equality

$$\mathbb{E}\left[X|D_i,\ i \in I\right] = \mathbb{E}\left[X|D_i\right] \quad \text{on } D_i \tag{15.7}$$

holds. It is therefore not surprising that the elementary properties given below do hold; in fact Properties **A**–**D** are immediate consequences of similar properties for expectations which were discussed in Section 11.1 and Section 11.2. The proofs are elementary and therefore omitted.

A. Mutiplying by a constant

For any $X : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$, and any scalar c in $\mathbb{R}$, we have

$$\mathbb{E}\left[c \cdot X|D_i,\ i \in I\right] = c \cdot \mathbb{E}\left[X|D_i,\ i \in I\right].$$

B. Addition

For any rvs $X, Y : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$ and $\mathbb{E}\left[|Y|\right] < \infty$, we have

$$\mathbb{E}\left[X + Y|D_i,\ i \in I\right] = \mathbb{E}\left[X|D_i,\ i \in I\right] + \mathbb{E}\left[Y|D_i,\ i \in I\right].$$

C. Monotonicity

Consider rvs $X, Y : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$ and $\mathbb{E}\left[|Y|\right] < \infty$. Whenever $X \leq Y$ a.s., we have

$$\mathbb{E}\left[X|D_i,\ i \in I\right] \leq \mathbb{E}\left[Y|D_i,\ i \in I\right].$$

D. Taking absolute values

For any rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$, we have

$$\left|\mathbb{E}\left[X|D_i,\ i \in I\right]\right| \leq \mathbb{E}\left[|X|\,|D_i,\ i \in I\right].$$

The forthcoming sections develop additional properties of conditioning with respect to a partition. Although proofs are provided for the sake of completeness, they are somewhat tedious and can be safely omitted in a first reading. Moreover, the notion of conditioning with respect to a partition is generalized in Chapter 16 where the corresponding properties are shown with typically shorter proofs which better highlight the structure of conditioning.

In the discussion we shall often make use of the following facts inherited from the definition: It holds that

$$\mathbb{E}\left[X|D_i,\ i\in I\right] = \sum_{i\in I}\mathbb{E}\left[X|D_i\right]\cdot\mathbf{1}\left[D_i\right]$$

where

$$\mathbb{E}\left[X|D_i\right] = \frac{\mathbb{E}\left[\mathbf{1}\left[D_i\right]\cdot X\right]}{\mathbb{P}\left[D_i\right]}, \qquad \begin{array}{c} i\in I \\ \mathbb{P}\left[D_i\right]>0. \end{array} \tag{15.8}$$

However, it is always the case that

$$\mathbb{E}\left[\mathbf{1}\left[D_i\right]\cdot X\right] = \mathbb{E}\left[X|D_i\right]\cdot\mathbb{P}\left[D_i\right], \quad i\in I. \tag{15.9}$$

15.4 The localization lemma

Lemma 15.4.1 *Consider a rv $X:\Omega\to\mathbb{R}$ with $\mathbb{E}\left[|X|\right]<\infty$. Let the rv $Z:\Omega\to\mathbb{R}$ be of the form*

$$Z = \sum_{j\in J} c_j\mathbf{1}\left[D_j\right] \tag{15.10}$$

for some countable subset $J\subseteq I$ and scalars $\{c_j,\ j\in J\}$. Whenever $\mathbb{E}\left[|ZX|\right]<\infty$, we have

$$\mathbb{E}\left[ZX|D_i,\ i\in I\right] = Z\cdot\mathbb{E}\left[X|D_i,\ i\in I\right] \quad \text{a.s.} \tag{15.11}$$

In other words, when Z is a rv of the form (15.10) it acts as a *constant* in the conditioning process and can therefore be taken out of the conditional expectation.

Proof. The proof proceeds by considering three separate cases, namely when the index J is a singleton, has a finite size and is countably infinite, respectively.

J is a singleton We start with $Z=\mathbf{1}\left[D_j\right]$ for some j in I. Note that

$$\begin{aligned} \mathbb{E}\left[ZX|D_i,\ i\in I\right] &= \mathbb{E}\left[\mathbf{1}\left[D_j\right]\cdot X|D_i,\ i\in I\right] \\ &= \sum_{i\in I}\mathbb{E}\left[\mathbf{1}\left[D_j\right]\cdot X|D_i\right]\cdot\mathbf{1}\left[D_i\right] \\ &= \sum_{i\in I:\ \mathbb{P}[D_i]>0}\mathbb{E}\left[\mathbf{1}\left[D_j\right]\cdot X|D_i\right]\cdot\mathbf{1}\left[D_i\right] \quad \text{a.s.} \end{aligned} \tag{15.12}$$

with

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left[D_j\right]X|D_i\right] &= \frac{\mathbb{E}\left[\mathbf{1}\left[D_i\right]\mathbf{1}\left[D_j\right]\cdot X\right]}{\mathbb{P}\left[D_i\right]} \\
&= \frac{\mathbb{E}\left[\mathbf{1}\left[D_i\cap D_j\right]\cdot X\right]}{\mathbb{P}\left[D_i\right]} \\
&= \delta(i,j)\cdot\frac{\mathbb{E}\left[\mathbf{1}\left[D_i\right]\cdot X\right]}{\mathbb{P}\left[D_i\right]} \\
&= \delta(i,j)\cdot\mathbb{E}\left[X|D_i\right], \qquad \begin{array}{c} i\in I \\ \mathbb{P}\left[D_i\right]>0. \end{array}
\end{aligned}$$

Combining this information in (15.12) with the fact that $\delta(i,j)\mathbf{1}\left[D_i\right]=\mathbf{1}\left[D_i\right]\mathbf{1}\left[D_j\right]$ for all i in I, we conclude that

$$\begin{aligned}
\mathbb{E}\left[ZX|D_i,\ i\in I\right] &= \sum_{i\in I:\ \mathbb{P}[D_i]>0}\delta(i,j)\cdot\mathbb{E}\left[X|D_i\right]\cdot\mathbf{1}\left[D_i\right] \\
&= \sum_{i\in I}\delta(i,j)\cdot\mathbb{E}\left[X|D_i\right]\cdot\mathbf{1}\left[D_i\right] \quad \text{a.s.} \\
&= \sum_{I\in I}\mathbb{E}\left[X|D_i\right]\mathbf{1}\left[D_i\right]\cdot\mathbf{1}\left[D_j\right] \\
&= \mathbf{1}\left[D_j\right]\cdot\mathbb{E}\left[X|D_i,\ i\in I\right] \qquad (15.13)
\end{aligned}$$

This establishes (15.11) when $Z=\mathbf{1}\left[D_j\right]$ for some j in I. ■

Finite J Next we consider the case when Z is of the form (15.10) for some *finite* index set $J\subseteq I$. By linearity we get

$$\begin{aligned}
\mathbb{E}\left[ZX|D_i,\ i\in I\right] &= \mathbb{E}\left[\left(\sum_{j\in J}c_j\mathbf{1}\left[D_j\right]\right)\cdot X\middle|D_i,\ i\in I\right] \\
&= \sum_{j\in J}c_j\mathbb{E}\left[\mathbf{1}\left[D_j\right]\cdot X\middle|D_i,\ i\in I\right] \\
&= \sum_{j\in J}c_j\mathbf{1}\left[D_j\right]\cdot\mathbb{E}\left[X|D_i,\ i\in I\right] \quad \text{a.s.}
\end{aligned}$$

by the first part of the proof. Noting that

$$\begin{aligned}
\sum_{j\in J}c_j\mathbf{1}\left[D_j\right]\cdot\mathbb{E}\left[X|D_i,\ i\in I\right] &= \left(\sum_{j\in J}c_j\mathbf{1}\left[D_j\right]\right)\cdot\mathbb{E}\left[X|D_i,\ i\in I\right] \\
&= Z\cdot\mathbb{E}\left[X|D_i,\ i\in I\right],
\end{aligned}$$

we obtain (15.11) when Z is a rv of the form (15.10) with J finite. ■

Countably infinite J Finally we turn to the case when Z is of the form (15.10) for some *countably* infinite index set $J \subseteq I$. There is no loss of generality in assuming that $J = \mathbb{N}_0$. For each $n = 1, 2, \ldots$, write $J_n = \{1, \ldots, n\}$ and introduce the rv Z_n given by

$$Z_n \equiv \sum_{j \in J_n} c_j \mathbf{1}[D_j] = \sum_{j=1}^{n} c_j \mathbf{1}[D_j].$$

This rv Z_n is of the form (15.10) with finite index set J_n. Direct inspection reveals that $|Z_n X| \leq |ZX|$ with $|Z_n X| = \sum_{j=1}^{n} |c_j \cdot X| \cdot \mathbf{1}[D_j]$ and $|ZX| = \sum_{j=1}^{\infty} |c_j \cdot X| \cdot \mathbf{1}[D_j]$. Under the assumed integrability condition $\mathbb{E}[|ZX|] < \infty$, the convergence $\lim_{n\to\infty} Z_n = Z$ yields

$$\lim_{n \to \infty} \mathbb{E}[|Z_n X|] = \mathbb{E}[|ZX|]$$

by the Dominated Convergence Theorem.

Fix $n = 1, 2, \ldots$. By an earlier part of the proof we have

$$\mathbb{E}[Z_n \cdot X | D_i, \ i \in I] = Z_n \cdot \mathbb{E}[X | D_i, \ i \in I] \quad \text{a.s.}$$

Let n go to infinity in this last relation. It is plain that

$$\begin{aligned} \lim_{n \to \infty} \mathbb{E}[Z_n \cdot X | D_i, \ i \in I] &= \lim_{n \to \infty} (Z_n \cdot \mathbb{E}[X | D_i, \ i \in I]) \quad \text{a.s.} \\ &= Z \cdot \mathbb{E}[X | D_i, \ i \in I]. \end{aligned} \tag{15.14}$$

We next show that

$$\lim_{n \to \infty} \mathbb{E}[Z_n \cdot X | D_i, \ i \in I] = \mathbb{E}[Z \cdot X | D_i, \ i \in I], \tag{15.15}$$

in which case combining (15.14) and (15.15) we obtain (15.11) when Z is a rv of the form (15.10) with J countably infinite.

To establish (15.15), for each $n = 1, 2, \ldots$ we note that

$$\begin{aligned} \mathbb{E}[Z_n \cdot X | D_i, \ i \in I] &= \sum_{i \in I} \mathbb{E}[Z_n \cdot X | D_i] \cdot \mathbf{1}[D_i] \\ &= \sum_{i \in I: \ \mathbb{P}[D_i] > 0} \mathbb{E}[Z_n \cdot X | D_i] \cdot \mathbf{1}[D_i] \quad \text{a.s.} \end{aligned} \tag{15.16}$$

with

$$\mathbb{E}\left[Z_n \cdot X | D_i\right] \quad = \quad \frac{\mathbb{E}\left[Z_n \cdot X \cdot \mathbf{1}\left[D_i\right]\right]}{\mathbb{P}\left[D_i\right]}, \qquad \begin{array}{c} i \in I \\ \mathbb{P}\left[D_i\right] > 0. \end{array}$$

Letting n go to infinity and invoking again the Dominated Convergence Theorem we get $\lim_{n\to\infty} \mathbb{E}\left[Z_n \cdot X \cdot \mathbf{1}\left[D_i\right]\right] = \mathbb{E}\left[Z \cdot X \cdot \mathbf{1}\left[D_i\right]\right]$, and the conclusion

$$\lim_{n\to\infty} \mathbb{E}\left[Z_n \cdot X | D_i\right] = \mathbb{E}\left[Z \cdot X | D_i\right], \qquad \begin{array}{c} i \in I \\ \mathbb{P}\left[D_i\right] > 0 \end{array}$$

follows.

Finally let n go to infinity in (15.16): Using the fact that the collection $\{D_i,\ i \in I\}$ is a partition, we obtain

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}\left[Z_n \cdot X | D_i,\ i \in I\right] &= \sum_{i \in I:\ \mathbb{P}[D_i]>0} \left(\lim_{n\to\infty} \mathbb{E}\left[Z_n \cdot X | D_i\right] \right) \cdot \mathbf{1}\left[D_i\right] \quad \text{a.s.} \\ &= \sum_{i \in I:\ \mathbb{P}[D_i]>0} \mathbb{E}\left[Z \cdot X | D_i\right] \cdot \mathbf{1}\left[D_i\right] \\ &= \mathbb{E}\left[Z \cdot X | D_i,\ i \in I\right] \quad \text{a.s.} \end{aligned}$$

and this immediately yields the desired result (15.15). ■

15.5 Taking the expectation of conditional expectations

The evaluation of the expectation of a conditional expectation with respect to a partition is discussed next. Below we give a direct proof of this result but it can also be viewed as a special case of iterated conditioning discussed in Section 15.6.

Lemma 15.5.1 *For any rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}\left[|X|\right] < \infty$, the rv $\mathbb{E}\left[X | D_i,\ i \in I\right]$ has a finite expectation with*

$$\mathbb{E}\left[\mathbb{E}\left[X | D_i,\ i \in I\right]\right] = \mathbb{E}\left[X\right]. \tag{15.17}$$

Proof. The proof proceeds along the usual steps.

Non-negative rvs If $X \geq 0$, then $\mathbb{E}\left[X|D_i\right] \geq 0$ for all i in I. Apply the Monotone Convergence Theorem to the series $\sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbf{1}\left[D_i\right]$ with non-negative terms and observe that

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}\left[X|D_i,\ i \in I\right]\right] &= \mathbb{E}\left[\sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbf{1}\left[D_i\right]\right] \\ &= \sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbb{E}\left[\mathbf{1}\left[D_i\right]\right] \\ &= \sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbb{P}\left[D_i\right] \\ &= \sum_{i \in I} \mathbb{E}\left[\mathbf{1}\left[D_i\right] \cdot X\right] \end{aligned} \tag{15.18}$$

as we make use of (15.8).

Using the Monotone Convergence Theorem once again we obtain

$$\begin{aligned} \sum_{i \in I} \mathbb{E}\left[\mathbf{1}\left[D_i\right] \cdot X\right] &= \mathbb{E}\left[\sum_{i \in I} \mathbf{1}\left[D_i\right] \cdot X\right] \\ &= \mathbb{E}\left[\left(\sum_{i \in I} \mathbf{1}\left[D_i\right]\right) \cdot X\right] = \mathbb{E}\left[X\right] \end{aligned} \tag{15.19}$$

by virtue of (15.4). Combining (15.18) and (15.19) yields (15.17) for non-negative rvs. It also follows that when $X \geq 0$, then the non-negative rv $\mathbb{E}\left[X|D_i,\ i \in I\right]$ has a finite expectation.

Arbitrary rvs For the general case, write $X = X^+ - X^-$. The finiteness assumption $\mathbb{E}\left[|X|\right] < \infty$ is equivalent to $\mathbb{E}\left[X^\pm\right] < \infty$, and the rv $\mathbb{E}\left[X^\pm|D_i,\ i \in I\right]$ therefore has a finite expectation with

$$\mathbb{E}\left[\mathbb{E}\left[X^\pm|D_i,\ i \in I\right]\right] = \mathbb{E}\left[X^\pm\right] \tag{15.20}$$

by the first part of the proof. Direct inspection gives

$$\begin{aligned} \mathbb{E}\left[X|D_i,\ i \in I\right] &= \sum_{i \in I} \left(\mathbb{E}\left[X^+|D_i\right] - \mathbb{E}\left[X^-|D_i\right]\right) \cdot \mathbf{1}\left[D_i\right] \\ &= \mathbb{E}\left[X^+|D_i,\ i \in I\right] - \mathbb{E}\left[X^-|D_i,\ i \in I\right] \end{aligned}$$

by linearity. The desired conclusion follows upon noting that

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}\left[X|D_i,\ i \in I\right]\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[X^+|D_i,\ i \in I\right] - \mathbb{E}\left[X^-|D_i,\ i \in I\right]\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^+|D_i,\ i \in I\right]\right] - \mathbb{E}\left[\mathbb{E}\left[X^-|D_i,\ i \in I\right]\right] \\ &= \mathbb{E}\left[X\right] \end{aligned}$$

with the help of (15.20). ■

15.6 Iterated conditioning

The properties discussed in this section are the ones that make conditioning such a powerful tool when carrying out calculations: Consider the situation where two countable $\mathcal{F}$-partitions $\{D_i,\ i \in I\}$ and $\{D'_j,\ j \in I'\}$ are available where I and I' are countable index sets.

Definition 15.6.1

The countable $\mathcal{F}$-partition $\{D'_j,\ j \in J'\}$ is said to be *finer* than the countable $\mathcal{F}$-partition $\{D_i,\ i \in I\}$ if for every i in I there exists a subset $J'(i) \subseteq I'$ such that

$$D_i = \cup_{j \in J'(i)} D'_j. \qquad (15.21)$$

Equivalently, the partition $\{D_i,\ i \in I\}$ is said to be *coarser* than the partition $\{D'_j,\ j \in I'\}$.

It is plain that the index sets $\{J'(i),\ i \in I\}$ must be disjoint since both collections $\{D_i,\ j \in I\}$ and $\{D'_j,\ j \in J'\}$ are $\mathcal{F}$-partitions of Ω. The following properties of *iterated conditioning* can be shown by direct arguments.

Lemma 15.6.1 *Assume the countable $\mathcal{F}$-partition $\{D'_j,\ j \in I'\}$ to be finer than the countable $\mathcal{F}$-partition $\{D_i,\ i \in I\}$. For any rv $X : \Omega \rightarrow \mathbb{R}$ such that $\mathbb{E}\left[|X|\right] < \infty$, it holds that*

$$\mathbb{E}\left[\mathbb{E}\left[X|D_i,\ i \in I\right]|D'_j,\ j \in I'\right] = \mathbb{E}\left[X|D_i,\ i \in I\right] \quad \text{a.s.} \qquad (15.22)$$

and

$$\mathbb{E}\left[\mathbb{E}\left[X|D'_j,\ j \in I'\right]|D_i,\ i \in I\right] = \mathbb{E}\left[X|D_i,\ i \in I\right] \quad \text{a.s.} \qquad (15.23)$$

Proof.

A proof of (15.22) In the notation of Definition 15.6.1 we note that

$$\mathbf{1}\left[D_i\right] = \sum_{j \in J'(i)} \mathbf{1}\left[D'_j\right], \quad i \in I$$

so that

$$\begin{aligned}\mathbb{E}\left[X|D_i,\, i \in I\right] &= \sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbf{1}\left[D_i\right] \\ &= \sum_{i \in I} \mathbb{E}\left[X|D_i\right] \cdot \left(\sum_{j \in J'(i)} \mathbf{1}\left[D'_j\right] \right) \\ &= \sum_{j \in I'} c'_j \cdot \mathbf{1}\left[D'_j\right] \qquad (15.24)\end{aligned}$$

with

$$c'_j \equiv \sum_{i \in I:\, j \in J'(i)} \mathbb{E}\left[X|D_i\right], \quad j \in I'.$$

Thus, the rv $\mathbb{E}\left[X|D_i,\, i \in I\right]$ is of the form (15.10) with respect to the finer partition $\{D',\, j \in I'\}$, and Lemma 15.4.1 (applied with $Z = \mathbb{E}\left[X|D_i,\, i \in I\right]$ and $X = 1$) yields the desired result (15.22) provided

$$\mathbb{E}\left[\left|\sum_{j \in I'} c'_j \cdot \mathbf{1}\left[D'_j\right]\right|\right] < \infty. \qquad (15.25)$$

This inetgrability condition is automatically satisfied by virtue of (15.24) and the integrability condition $\mathbb{E}\left[|X|\right] < \infty$ (which automatically implies $\mathbb{E}\left[|\mathbb{E}\left[X|D_i,\, i \in I\right]|\right] < \infty$. ■

A proof of (15.23) We start with the case of non-negative rvs. So assume first that $X \geq 0$. Applying the definition we start with

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left[X|D'_j,\, j \in I'\right]|D_i,\, i \in I\right] \\ &\quad = \sum_{i \in I} \mathbb{E}\left[\mathbb{E}\left[X|D'_j,\, j \in I'\right]|D_i\right] \cdot \mathbf{1}\left[D_i\right] \qquad (15.26)\end{aligned}$$

where the rv $\mathbb{E}\left[X|D'_j,\, j \in I'\right]$ is given by

$$\mathbb{E}\left[X|D'_j,\, j \in I'\right] = \sum_{j \in I'} \mathbb{E}\left[X|D'_j\right] \cdot \mathbf{1}\left[D'_j\right]. \qquad (15.27)$$

Fix i in I with $\mathbb{P}\left[D_i\right] > 0$, and consider the corresponding term in (15.26): Again applying the definition we find

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[X|D'_j,\ j\in I'\right]|D_i\right] &= \mathbb{E}\left[\sum_{j\in I'}\mathbb{E}\left[X|D'_j\right]\cdot\mathbf{1}\left[D'_j\right]|D_i\right] \\
&= \sum_{j\in I'}\mathbb{E}\left[\mathbb{E}\left[X|D'_j\right]\cdot\mathbf{1}\left[D'_j\right]|D_i\right] \qquad (15.28)\\
&= \sum_{j\in I'}\mathbb{E}\left[X|D'_j\right]\cdot\mathbb{E}\left[\mathbf{1}\left[D'_j\right]|D_i\right] \qquad (15.29)
\end{aligned}$$

with

$$\mathbb{E}\left[\mathbf{1}\left[D'_j\right]|D_i\right] = \frac{\mathbb{E}\left[\mathbf{1}\left[D'_j\right]\cdot\mathbf{1}\left[D_i\right]\right]}{\mathbb{P}\left[D_i\right]} = \frac{\mathbb{P}\left[D'_j\cap D_i\right]}{\mathbb{P}\left[D_i\right]}, \quad j\in I'.$$

The equality (15.28) is a consequence of the Monotone Convergence Theorem applied to a series with non-negative terms. But D_i being of the form $D_i = \cup_{j\in J'(i)}D'_j$ for some subset $J'(i)\subseteq I'$, it follows that

$$\mathbb{P}\left[D'_j\cap D_i\right] = \begin{cases}\mathbb{P}\left[D'_j\right] & \text{if } j\in J'(i)\\ 0 & \text{if } j\notin J'(i).\end{cases} \qquad (15.30)$$

As a result, we conclude from (15.29) and (15.30) that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[X|D'_j,\ j\in I'\right]|D_i\right]\\
&= \sum_{j\in I'}\mathbb{E}\left[X|D'_j\right]\cdot\frac{\mathbb{P}\left[D'_j\right]}{\mathbb{P}\left[D_i\right]}\cdot\mathbf{1}\left[j\in J'(i)\right]\\
&= \frac{1}{\mathbb{P}\left[D_i\right]}\left(\sum_{j\in J'(i)}\mathbb{E}\left[X|D'_j\right]\cdot\mathbb{P}\left[D'_j\right]\right)\\
&= \frac{1}{\mathbb{P}\left[D_i\right]}\left(\sum_{j\in J'(i)}\mathbb{E}\left[X\cdot\mathbf{1}\left[D'_j\right]\right]\right)\\
&= \frac{1}{\mathbb{P}\left[D_i\right]}\left(\mathbb{E}\left[X\cdot\left(\sum_{j\in J'(i)}\mathbf{1}\left[D'_j\right]\right)\right]\right) \qquad (15.31)\\
&= \frac{\mathbb{E}\left[X\cdot\mathbf{1}\left[D_i\right]\right]}{\mathbb{P}\left[D_i\right]}\\
&= \mathbb{E}\left[X|D_i\right]
\end{aligned}$$

where the equality (15.31) follows by an application of the Monotone Convergence Theorem applied to a series with non-negative terms. This completes the proof of (15.23) when X is a non-negative rv.

The case of an arbitrary rv X is treated in the usual manner: With $X = X^+ - X^-$, the first part of the proof for non-negative rvs yields

$$\mathbb{E}\left[\mathbb{E}\left[X^\pm | D'_j,\ j \in I'\right] | D_i,\ i \in I\right] = \mathbb{E}\left[X^\pm | D_i,\ i \in I\right] \tag{15.32}$$

while we have

$$\mathbb{E}\left[X | D'_j,\ j \in I'\right] = \mathbb{E}\left[X^+ | D'_j,\ j \in I'\right] - \mathbb{E}\left[X^- | D'_j,\ j \in I'\right].$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[X | D'_j,\ j \in I'\right] | D_i,\ i \in I\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[X^+ | D'_j,\ j \in I'\right] - \mathbb{E}\left[X^- | D'_j,\ j \in I'\right] | D_i,\ i \in I\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[X^+ | D'_j,\ j \in I'\right] | D_i,\ i \in I\right] - \mathbb{E}\left[\mathbb{E}\left[X^- | D'_j,\ j \in I'\right] | D_i,\ i \in I\right] \\
&= \mathbb{E}\left[X^+ | D_i,\ i \in I\right] - \mathbb{E}\left[X^- | D_i,\ i \in I\right] \\
&= \mathbb{E}\left[X | D_i,\ i \in I\right]
\end{aligned}$$

by the usual arguments. This completes the proof of (15.23) when X is an arbitrary rv. ■

15.7 Conditioning and independence

The next result explores the interplay between conditioning and independence.

Lemma 15.7.1 *Consider a rv $X : \Omega \to \mathbb{R}$ which is independent of the the $\mathcal{F}$-partition $\{D_i,\ i \in I\}$. If $\mathbb{E}\left[|X|\right] < \infty$, then*

$$\mathbb{E}\left[X | D_i,\ i \in I\right] = \mathbb{E}\left[X\right] \quad \textit{a.s.} \tag{15.33}$$

Here, the independence of the rv X from the $\mathcal{F}$-partition $\{D_i,\ i \in I\}$ means that for each i in I, the rvs $\mathbf{1}\left[D_i\right]$ and X are independent rvs. Under the assumption $\mathbb{E}\left[|X|\right] < \infty$ this independence condition implies

$$\mathbb{E}\left[X \cdot \mathbf{1}\left[D_i\right]\right] = \mathbb{P}\left[D_i\right] \cdot \mathbb{E}\left[X\right], \quad i \in I. \tag{15.34}$$

Proof. Using (15.8) with (15.34) yields

$$\mathbb{E}\left[X|D_i\right] = \mathbb{E}\left[X\right], \qquad \begin{array}{c} i \in I \\ \mathbb{P}\left[D_i\right] > 0. \end{array}$$

It then follows that

$$\begin{aligned} \mathbb{E}\left[X|D_i,\ i \in I\right] &= \sum_{i\in I} \mathbb{E}\left[X|D_i\right] \cdot \mathbf{1}\left[D_i\right] \\ &= \sum_{i\in I:\ \mathbb{P}[D_i]>0} \mathbb{E}\left[X|D_i\right] \cdot \mathbf{1}\left[D_i\right] \quad \text{a.s.} \\ &= \left(\sum_{i\in I:\ \mathbb{P}[D_i]} \cdot\mathbf{1}\left[D_i\right] \right) \cdot \mathbb{E}\left[X\right] \\ &= \mathbb{E}\left[X\right] \quad \text{a.s.} \end{aligned}$$

since $\sum_{i\in I:\ \mathbb{P}[D_i]>0} \mathbf{1}\left[D_i\right] = \sum_{i\in I} \mathbf{1}\left[D_i\right] = 1$ a.s. ■

15.8 Characterizing the conditional expectation with respect to a partition

The following characterization foreshadows forthcoming developments in the general case discussed in Section 16.2.

Lemma 15.8.1 *For any rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$, it holds that*

$$\mathbb{E}\left[\mathbf{1}\left[D\right] \mathbb{E}\left[X|D_i,\ i \in I\right]\right] = \mathbb{E}\left[\mathbf{1}\left[D\right] X\right] \tag{15.35}$$

for any event D in $\mathcal{F}$ of the form

$$D = \cup_{j\in J} D_j \tag{15.36}$$

for some index set $J \subseteq I$.

The conditions (15.35)-(15.36) collectively characterize the rv $\mathbb{E}\left[X|D_i,\ i \in I\right]$ given in Definition 15.2.1 as essentially the *only* rv Z of the form

$$Z = \sum_{i\in I} c_i \mathbf{1}\left[D_i\right] \tag{15.37}$$

with scalars $\{c_i,\ i \in I\}$. This is a consequence of the next result.

Lemma 15.8.2 *Consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$, and let $Z_1, Z_2 : \Omega \to \mathbb{R}$ be rvs of the form (15.37) such that both $\mathbb{E}[Z_1]$ and $\mathbb{E}[Z_2]$ exist. If both rvs satisfy the conditions*

$$\mathbb{E}\left[\mathbf{1}\left[D\right] Z_k\right] = \mathbb{E}\left[\mathbf{1}\left[D\right] X\right], \quad \begin{array}{c} D \in \mathcal{D} \\ k = 1, 2, \end{array} \tag{15.38}$$

then they have finite expectations with $\mathbb{E}[|Z_1|] < \infty$ and $\mathbb{E}[|Z_2|] < \infty$, and $Z_1 = Z_2$ a.s.

As this result is subsumed by Claim (ii) of Theorem 16.2.1, its proof will not be provided; see also Section 16.7 to map conditioning with respect to a partition into the general notion of conditioning.

15.9 A proof of Lemma 15.8.1

Throughout fix D in $\mathcal{F}$ of the form $D = \cup_{j \in J} D_j$ for some $J \subseteq I$.

Non-negative rvs We first assume that the rv X is non-negative, say $X : \Omega \to \mathbb{R}_+$, in which case we have $\mathbb{E}[X|D_i] \geq 0$ for each i in I. Fix D in $\mathcal{F}$ of the form $D = \cup_{j \in J} D_j$ for some $J \subseteq I$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left[D\right] X\right] &= \mathbb{E}\left[\left(\sum_{j \in J} \mathbf{1}\left[D_j\right]\right) X\right] \\
&= \sum_{j \in J} \mathbb{E}\left[\mathbf{1}\left[D_j\right] X\right] && (15.39) \\
&= \sum_{j \in J:\ \mathbb{P}[D_j] > 0} \mathbb{E}\left[X | D_j\right] \cdot \mathbb{P}\left[D_j\right] \\
&= \sum_{j \in J:\ \mathbb{P}[D_j] > 0} \mathbb{E}\left[\mathbf{1}\left[D_j\right] \cdot \mathbb{E}\left[X | D_j\right]\right] \\
&= \sum_{j \in J:\ \mathbb{P}[D_j] > 0} \mathbb{E}\left[\mathbf{1}\left[D_j\right] \cdot \mathbb{E}\left[X | D_i,\ i \in I\right]\right] \\
&= \mathbb{E}\left[\sum_{j \in J:\ \mathbb{P}[D_j] > 0} \mathbf{1}\left[D_j\right] \cdot \mathbb{E}\left[X | D_i,\ i \in I\right]\right] && (15.40) \\
&= \mathbb{E}\left[\left(\sum_{j \in J:\ \mathbb{P}[D_j] > 0} \mathbf{1}\left[D_j\right]\right) \cdot \mathbb{E}\left[X | D_i,\ i \in I\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\sum_{j\in J}\mathbf{1}\left[D_j\right]\right)\cdot\mathbb{E}\left[X|D_i,\ i\in I\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\mathbb{E}\left[X|D_i,\ i\in I\right]\right]
\end{aligned}
$$

as desired – Both (15.39) and (15.40) are validated by monotone convergence if J is countably infinite. In the equality before last we have used the fact that

$$\sum_{j\in J:\ \mathbb{P}[D_j]>0}\mathbf{1}\left[D_j\right]=\sum_{j\in J}\mathbf{1}\left[D_j\right]=\mathbf{1}\left[D\right]\quad\text{a.s.}$$

Arbitrary rvs To obtain the result in the arbitrary case, use the usual decomposition $X = X^+ - X^-$ with $\mathbb{E}\left[X^\pm\right]$ finite since $\mathbb{E}\left[|X|\right] < \infty$. Therefore, starting with the definition we find that $\mathbb{E}\left[X^\pm|D_i,\ i\in I\right]$ is finite with

$$\mathbb{E}\left[X^\pm|D_i,\ i\in I\right]=\sum_{i\in I}\mathbb{E}\left[X^\pm|D_i\right]\cdot\mathbf{1}\left[D_i\right]$$

It is now plain that

$$
\begin{aligned}
&\mathbb{E}\left[X^+|D_i,\ i\in I\right]-\mathbb{E}\left[X^-|D_i,\ i\in I\right] \\
&\quad= \sum_{i\in I}\mathbb{E}\left[X^+|D_i\right]\cdot\mathbf{1}\left[D_i\right]-\sum_{i\in I}\mathbb{E}\left[X^-|D_i\right]\cdot\mathbf{1}\left[D_i\right] \\
&\quad= \sum_{i\in I}\left(\mathbb{E}\left[X^+|D_i\right]-\mathbb{E}\left[X^-|D_i\right]\right)\cdot\mathbf{1}\left[D_i\right] \\
&\quad= \sum_{i\in I}\mathbb{E}\left[X|D_i\right]\cdot\mathbf{1}\left[D_i\right] \\
&\quad= \mathbb{E}\left[X|D_i,\ i\in I\right] \qquad (15.41)
\end{aligned}
$$

as we make use of the fact $\mathbb{E}\left[X|D_i\right]=\mathbb{E}\left[X^+|D_i\right]-\mathbb{E}\left[X^-|D_i\right]$ for each i in I.

The first part of the proof yields

$$\mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\mathbb{E}\left[X^\pm|D_i,\ i\in I\right]\right]=\mathbb{E}\left[\mathbf{1}\left[D\right]\cdot X^\pm\right]. \qquad (15.42)$$

Noting that $\mathbb{E}\left[X^\pm|D_i,\ i\in I\right]\geq 0$ a.s., we conclude from (15.42) (with $D=\Omega$) that the rv $\mathbb{E}\left[X^\pm|D_i,\ i\in I\right]$ has finite expectation and so does $\mathbb{E}\left[X|D_i,\ i\in I\right]$ by virtue of (15.41). Using (15.41) again we conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\mathbb{E}\left[X|D_i,\ i\in I\right]\right] \\
&\quad= \mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\left(\mathbb{E}\left[X^+|D_i,\ i\in I\right]-\mathbb{E}\left[X^-|D_i,\ i\in I\right]\right)\right] \\
&\quad= \mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\mathbb{E}\left[X^+|D_i,\ i\in I\right]\right]-\mathbb{E}\left[\mathbf{1}\left[D\right]\cdot\mathbb{E}\left[X^-|D_i,\ i\in I\right]\right] \\
&\quad= \mathbb{E}\left[\mathbf{1}\left[D\right]\cdot X^+\right]-\mathbb{E}\left[\mathbf{1}\left[D\right]\cdot X^-\right] \qquad (15.43) \\
&\quad= \mathbb{E}\left[\mathbf{1}\left[D\right]X\right]
\end{aligned}
$$

where (15.43) follows from the validity of the result for non-negative rvs. This completes the proof of Lemma 15.8.1. ■

15.10 Exercises

Ex. 15.1 Consider a rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$.

a. Compute $\mathbb{E}[X|D]$ when $D = \Omega$.

b. Compute the conditional expectation of X given the $\mathcal{F}$-partition $\{\Omega\}$.

Ex. 15.2 Let X denote a geometric rv with parameter $0 < a < 1$, namely

$$\mathbb{P}[X = k] = (1-a)a^k, \quad k = 0, 1, \ldots$$

a. Compute the conditional probabilities

$$\mathbb{P}[X = k + \ell | X \geq k], \quad k, \ell = 0, 1, \ldots$$

b. (Converse) Consider now a discrete rv Y with support $\{0, 1, \ldots\}$ and pmf $(p_r, \ r = 0, 1, \ldots)$. Define

$$q_{\ell|k} := \mathbb{P}[Y = k + \ell | Y \geq k], \quad k, \ell = 0, 1, \ldots$$

For each $k = 0, 1, \ldots$, $(q_{\ell|k}, \ \ell = 0, 1, \ldots)$ can be viewed as the pmf for a discrete rv with support $\{0, 1, \ldots\}$. Determine *all* the pmfs $(p_r, \ r = 0, 1, \ldots)$ with the property that

$$q_{\ell|k} = p_\ell, \quad k, \ell = 0, 1, \ldots$$

simultaneously!

Ex. 15.3 With a in $(0, 1)$, consider a collection of mutually independent Bernoulli rvs $\{B_k, \ k = 1, 2, \ldots\}$ with

$$\mathbb{P}[X_k = 1] = 1 - \mathbb{P}[X_k = 0] = a, \quad k = 1, 2, \ldots$$

For each $n = 1, 2, \ldots$, define the partial sums $S_n \equiv B_1 + \ldots + B_n$.

a. For each $n = 1, 2, \ldots$ compute the conditional probabilities

$$\mathbb{P}[B_k = b | S_n = s], \quad \begin{array}{c} b = 0, 1 \\ s = 0, 1, \ldots, n \\ k = 1, \ldots, n \end{array}$$

Is the result surprising?

b. For each $n = 1, 2, \ldots$ compute the conditional probabilities

$$\mathbb{P}\left[B_k = b_k, B_\ell = b_\ell | S_n = s\right], \qquad \begin{array}{c} b_k, b_\ell \in \{0,1\} \\ s = 0, 1, \ldots, n \\ k \neq \ell, \ k, \ell = 1, \ldots, n \end{array}$$

Are the rvs B_k and B_ℓ conditionally independent given that $S_n = s$?

Ex. 15.4 Consider a rv $X : \Omega \to \mathbb{R}$ of the discrete type with $\mathbb{P}\left[X \in S\right] = 1$ where $S \equiv \{a_i, \ i \in I\}$ for some countable index set I. Let B an event such that $\mathbb{P}\left[B\right] > 0$ – Obviously both X and B are defined on the same sample space Ω with B in $\mathcal{F}$. Define the function $F(\cdot|B) : \mathbb{R} \to [0,1]$ by

$$F(x|B) \equiv \mathbb{P}\left[X \leq x|B\right], \quad x \in \mathbb{R}.$$

If this probability distribution function were to be of the discrete type, show that its atoms are also atoms for the discrete rv X, i.e., if $\mathbb{P}\left[X = a|B\right] > 0$ for some a in $\mathbb{R}$, then $\mathbb{P}\left[X = a\right] > 0$.

Ex. 15.5 If the discrete rv $X : \Omega \to \mathbb{R}$ has pmf given by

$$\mathbb{P}\left[X = 1\right] = \mathbb{P}\left[X = 0\right] = \frac{1}{2},$$

define the rv $Y \equiv 1 + (-1)^X$ Show that the atoms of the conditional distribution of Y given $X = 1$ form a strict subset of the set of atoms of Y.

Ex. 15.6 Let the second-order rv N be a discrete rv whose support is contained in $\mathbb{N}_0$ (i.e., $\mathbb{P}\left[N \in \mathbb{N}_0\right] = 1$), and let $\{X_n, \ n = 1, 2, \ldots\}$ denote a collection of second-order rvs. Assume the rvs $\{N, X_n, \ n = 1, 2, \ldots\}$ to be mutually independent.

a. Using pre-conditioning arguments compute the expectation

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right].$$

b. Using pre-conditioning arguments compute the variance

$$\text{Var}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right].$$

Chapter 16

Conditional expectations: The general case

We are now ready to define the general notion of conditional expectation. Rather than conditioning with respect to a single event (as in Section 15.1) or even with respect to a family of events forming a partition (as in Section 15.2), it turns out that the appropriate setting is that of conditioning with respect to a σ-field.

The rvs introduced next are all defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

16.1 Sub-σ-fields generated by rvs

The following terminology will be useful throughout.

Definition 16.1.1

A collection $\mathcal{D}$ of subsets of Ω is called a sub-σ-field of $\mathcal{F}$ if $\mathcal{D}$ is a σ-field on Ω such that $\mathcal{D} \subseteq \mathcal{F}$.

We begin with a simple definition.

Definition 16.1.2

Let $\mathcal{D}$ be a sub-σ-field of $\mathcal{F}$. An rv $\mathbb{R}^p$-valued rv $X : \Omega \to \mathbb{R}^p$ is said to be $\mathcal{D}$-measurable if

$$[X \in B] \in \mathcal{D}, \quad B \in \mathcal{B}(\mathbb{R}^p)$$

(and not merely in $\mathcal{F}$ as required in the definition of X as a rv).

These definitions are often used when the σ-field $\mathcal{D}$ is itself generated by some rv $Y : \Omega \to \mathbb{R}^q$ in the following sense: This σ-field, denoted $\sigma(Y)$, is defined by

$$\sigma(Y) \equiv \left\{ Y^{-1}(C) : \ C \in \mathcal{B}(\mathbb{R}^q) \right\}$$

as expected. We have the following important operational characterization of $\sigma(Y)$-measurability.

Lemma 16.1.1 *Assume the σ-field $\mathcal{D}$ is generated by the rv $Y : \Omega \to \mathbb{R}^q$, so that $\mathcal{D} = \sigma(Y)$:*

(i) For any Borel mapping $g : \mathbb{R}^q \to \mathbb{R}$, the rv $Z = g(Y)$ is $\mathcal{D}$-measurable.

(ii) Conversely, any $\mathcal{D}$-measurable rv $Z : \Omega \to \mathbb{R}$ can be written in the form $Z = g(Y)$ for some Borel mapping $g : \mathbb{R}^q \to \mathbb{R}$.

A proof is available in Section 16.10.

16.2 The general definition of conditional expectations

We now present a general definition for the conditional expectation given an arbitrary σ-field.

Theorem 16.2.1 *Let $\mathcal{D}$ be a sub-σ-field of $\mathcal{F}$, and consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$.*

(i) (Existence) There exists a $\mathcal{D}$-measurable rv $Z : \Omega \to \mathbb{R}$ with $\mathbb{E}[|Z|] < \infty$ such that

$$\mathbb{E}[\mathbf{1}[D]\, Z] = \mathbb{E}[\mathbf{1}[D]\, X], \quad D \in \mathcal{D}. \tag{16.1}$$

(ii) (Uniqueness) Let $Z_1, Z_2 : \Omega \to \mathbb{R}$ be $\mathcal{D}$-measurable rvs such that $\mathbb{E}[Z_1]$ and $\mathbb{E}[Z_2]$ exist. If they both satisfy (16.1), namely

$$\mathbb{E}[\mathbf{1}[D]\, Z_k] = \mathbb{E}[\mathbf{1}[D]\, X], \quad \begin{matrix} k = 1,2 \\ D \in \mathcal{D} \end{matrix} \tag{16.2}$$

then $\mathbb{E}[|Z_1|] < \infty$ and $\mathbb{E}[|Z_2|] < \infty$, and $Z_1 = Z_2$ a.s.

Proof. Claim (i): We consider first the case when $X \geq 0$: Introduce the set function $\mathbb{Q} : \mathcal{D} \to [0, +\infty)$ given by

$$\mathbb{Q}[D] \equiv \mathbb{E}[\mathbf{1}[D]\, X], \quad D \in \mathcal{D}. \tag{16.3}$$

The following three facts hold for the set function $\mathbb{Q} : \mathcal{D} \to [0, +\infty)$:

$\mathbb{Q}$ **is a measure** Obviously, we have $\mathbb{Q}[\emptyset] = 0$. Next, consider a countable collection $\{E_i,\ i \in I\}$ of disjoint events in $\mathcal{F}$, so that $\mathbf{1}[\cup_{i\in I} E_i] = \sum_{i\in I} \mathbf{1}[E_i]$. We have

$$\begin{aligned}
\mathbb{Q}[\cup_{i\in I} E_i] &= \mathbb{E}[\mathbf{1}[\cup_{i\in I} E_i] \cdot X] \\
&= \mathbb{E}\left[\left(\sum_{i\in I} \mathbf{1}[E_i]\right) \cdot X\right] \\
&= \mathbb{E}\left[\sum_{i\in I} (\mathbf{1}[E_i] \cdot X)\right] \\
&= \sum_{i\in I} \mathbb{E}[\mathbf{1}[E_i] \cdot X] \\
&= \sum_{i\in I} \mathbb{Q}[E_i]
\end{aligned}$$

where the equality before last is validated by the Monotone Convergence Theorem applied to a series with non-negative random summands. ■

Absolute continuity For any event event D in $\mathcal{D}$ we have the bounds

$$\mathbb{E}[\mathbf{1}[D] \min(X, n)] \leq n\mathbb{P}[D], \quad n = 1, 2, \ldots.$$

If $\mathbb{P}[D] = 0$, then $\mathbb{E}[\mathbf{1}[D] \min(X, n)] = 0$ for all $n = 1, 2, \ldots$, and it follows that $\lim_{n\to\infty} \mathbb{E}[\mathbf{1}[D] \min(X, n)] = 0$. Using the Monotone Convergence Theorem and the fact that $\lim_{n\to\infty} \mathbf{1}[D] \min(X, n) = \mathbf{1}[D] X$ we conclude that $\mathbb{E}[\mathbf{1}[D] X] = 0$. Thus, whenever $\mathbb{P}[D] = 0$ for an event D in $\mathcal{D}$, we have $\mathbb{Q}[D] = 0$, and $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{Q}$ on $\mathcal{D}$. ■

Finiteness Both the measure $\mathbb{Q} : \mathcal{D} \to [0, \infty)$ and the underlying probability measure $\mathbb{P}$ are finite measures with $\mathbb{Q}[\Omega] = \mathbb{E}[X]$ and $\mathbb{P}[\Omega] = 1$. ■

The celebrated Radon-Nikodym Theorem can now be applied to the pair of measures $\mathbb{Q}$ and $\mathbb{P}$ (restricted to $\mathcal{D}$): It implies the existence of a $\mathcal{D}$-measurable rv $L : \Omega \to [0, +\infty)$ such that

$$\mathbb{Q}[D] = \mathbb{E}[\mathbf{1}[D] L], \quad D \in \mathcal{D}. \tag{16.4}$$

Combining (16.3) and (16.5) we conclude that

$$\mathbb{E}\left[\mathbf{1}\left[D\right]L\right] = \mathbb{E}\left[\mathbf{1}\left[D\right]X\right], \quad D \in \mathcal{D}. \tag{16.5}$$

Using $D = \Omega$ in (16.5) we see that $\mathbb{E}[L]$ is finite since $\mathbb{E}[X]$ is finite. Obviously, the $\mathcal{D}$-measurable rv L is a candidate for the $\mathcal{D}$-measurable rv Z which satisfies (16.1).

To handle the case of an arbitrary rv $X : |\Omega \to \mathbb{R}$, we proceed in the usual manner: Write $X = X^+ - X^-$ and use the earlier part of the proof. There exist $\mathcal{D}$-measurable rvs $L^\pm : \Omega \to [0, +\infty)$ such that

$$\mathbb{E}\left[\mathbf{1}\left[D\right]L^\pm\right] = \mathbb{E}\left[\mathbf{1}\left[D\right]X^\pm\right], \quad D \in \mathcal{D}. \tag{16.6}$$

with $\mathbb{E}[L^\pm] < \infty$ by the first part of the proof. It is now immediate from (16.6) that

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}\left[D\right]\left(L^+ - L^-\right)\right] &= \mathbb{E}\left[\mathbf{1}\left[D\right]X^+\right] - \mathbb{E}\left[\mathbf{1}\left[D\right]X^-\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[D\right]X\right], \quad D \in \mathcal{D}. \end{aligned} \tag{16.7}$$

This time the $\mathcal{D}$-measurable rv $L^+ - L^-$ is a candidate for the desired $\mathcal{D}$-measurable rv Z which satisfies (16.1).

Claim (ii): The condition $\mathbb{E}[|X|] < \infty$ guarantees that $\mathbb{E}[\mathbf{1}[D]X]$ is finite for *any* event D in $\mathcal{D}$. Therefore, for each $k = 1, 2$, introducing the $\mathcal{D}$-measurable events $D_k^+ = [Z_k \geq 0]$ and $D_k^- = [Z_k \leq 0]$ in (16.2), we conclude that $\mathbb{E}\left[\mathbf{1}\left[D_k^\pm\right]Z_k\right]$ is finite. Since $Z_k^+ = \mathbf{1}\left[D_k^+\right]Z_k$ while $Z_k^- = -\mathbf{1}\left[D_k^-\right]Z_k$, it immediately follows that the rv $Z_k^\pm$ has a finite expectation, and so does the rv Z_k.

By linearity the condition (16.2) now implies

$$\mathbb{E}\left[\mathbf{1}\left[D\right]Z\right] = 0, \quad D \in \mathcal{D} \tag{16.8}$$

with $\mathcal{D}$-measurable rv $Z = Z_1 - Z_2$. Using (16.8) with $\mathcal{D}$-measurable events $D_+ = [Z \geq 0]$ and $D_- = [Z < 0]$, we readily conclude that $\mathbf{1}[D_+]Z = Z^+ = 0$ a.s., and $\mathbf{1}[D_-]Z = -Z^- = 0$ a.s., whence $Z = Z^+ - Z^- = 0$ a.s. ■

The $\mathcal{D}$-measurable rvs with finite expectation satisfying (16.1) form an equivalence class; any one of its representatives will be denoted by $\mathbb{E}[X|\mathcal{D}]$.

16.3 Elementary properties

We now list several basic properties of conditional expectations. They can all be derived through the characterization and uniqueness of Theorem 16.2.1.

A. Mutiplying by a constant

For any $X : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$, and any c in $\mathbb{R}$, we have

$$\mathbb{E}[c \cdot X|\mathcal{D}] = c \cdot \mathbb{E}[X|\mathcal{D}] \quad \text{a.s.}$$

Indeed, for any event D in $\mathcal{D}$,

$$\begin{aligned} \mathbb{E}[\mathbf{1}[D]\,\mathbb{E}[c \cdot X|\mathcal{D}]] &= \mathbb{E}[\mathbf{1}[D]\,c \cdot X] \\ &= c\ddot{\mathbb{E}}[\mathbf{1}[D]\,X] \\ &= c \cdot \mathbb{E}[\mathbf{1}[D]\,\mathbb{E}[X|\mathcal{D}]] \\ &= \mathbb{E}[\mathbf{1}[D]\,c \cdot \mathbb{E}[X|\mathcal{D}]] \end{aligned}$$

and the conclusion follows by uniqueness as we note that the rv $c \cdot \mathbb{E}[X|\mathcal{D}]$ is $\mathcal{D}$-measurable. ■

B. Addition

For any rvs $X, Y : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$, we have

$$\mathbb{E}[X + Y|\mathcal{D}] = \mathbb{E}[X|\mathcal{D}] + \mathbb{E}[Y|\mathcal{D}] \quad \text{a.s.}$$

For any event D in $\mathcal{D}$, both rvs $\mathbb{E}[X|\mathcal{D}]$ and $\mathbb{E}[Y|\mathcal{D}]$ are integrable. Next note that

$$\begin{aligned} \mathbb{E}[\mathbf{1}[D]\,\mathbb{E}[X+Y|\mathcal{D}]] &= \mathbb{E}[\mathbf{1}[D]\,(X+Y)] \\ &= \mathbb{E}[\mathbf{1}[D]\,X] + \mathbb{E}[\mathbf{1}[D]\,Y] \\ &= \mathbb{E}[\mathbf{1}[D]\,\mathbb{E}[X|\mathcal{D}]] + \mathbb{E}[\mathbf{1}[D]\,\mathbb{E}[Y|\mathcal{D}]] \\ &= \mathbb{E}[\mathbf{1}[D]\,(\mathbb{E}[X|\mathcal{D}] + \mathbb{E}[Y|\mathcal{D}])] \end{aligned} \tag{16.9}$$

and the conclusion follows by uniqueness since the rv $\mathbb{E}[X|\mathcal{D}] + \mathbb{E}[Y|\mathcal{D}]$ is $\mathcal{D}$-measurable. ■

C. Monotonicity

Consider rvs $X, Y : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$. Whenever $X \leq Y$ a.s., we have

$$\mathbb{E}[X|\mathcal{D}] \leq \mathbb{E}[Y|\mathcal{D}] \quad \text{a.s.}$$

For any D in $\mathcal{D}$, the inequality $X \leq Y$ a.s. implies $\mathbb{E}\left[\mathbf{1}\left[D\right]X\right] \leq \mathbb{E}\left[\mathbf{1}\left[D\right]Y\right]$ and using (16.1) we get

$$\mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \mathbb{E}\left[X|\mathcal{D}\right]\right] \leq \mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \mathbb{E}\left[Y|\mathcal{D}\right]\right].$$

and using linearity we obtain

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \mathbb{E}\left[Y|\mathcal{D}\right]\right] - \mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \mathbb{E}\left[X|\mathcal{D}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \left(\mathbb{E}\left[Y|\mathcal{D}\right] - \mathbb{E}\left[X|\mathcal{D}\right]\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[D\right] \cdot \mathbb{E}\left[Y - X|\mathcal{D}\right]\right]. \end{aligned}$$

Using the $\mathcal{D}$-measurable event $D^- \equiv \left[\mathbb{E}\left[Y-X|\mathcal{D}\right] \leq 0\right]$ in this last expression we get

$$0 \leq \mathbb{E}\left[\mathbf{1}\left[D^-\right] \cdot \mathbb{E}\left[Y - X|\mathcal{D}\right]\right] \leq 0$$

since $\mathbf{1}\left[D^-\right] \cdot \mathbb{E}\left[Y-X|\mathcal{D}\right] \leq 0$. Therefore, we have $\mathbb{E}\left[\mathbf{1}\left[D^-\right] \cdot \mathbb{E}\left[Y-X|\mathcal{D}\right]\right] = 0$ and the conclusion $\mathbf{1}\left[D^-\right] \cdot \mathbb{E}\left[Y-X|\mathcal{D}\right] = 0$ a.s. follows.

■

D. Taking absolute values

If $\mathbb{E}\left[|X|\right] < \infty$, then $|\mathbb{E}\left[X|\mathcal{D}\right]| \leq \mathbb{E}\left[|X||\mathcal{D}\right]$ a.s.

The result is a simple consequence of Property **C** as we note that $-|X| \leq X \leq |X|$. ■

16.4 The localization lemma

Lemma 16.4.1 *For any rvs $X, Z : \Omega \to \mathbb{R}$ with $\mathbb{E}\left[|X|\right] < \infty$ and $\mathbb{E}\left[|ZX|\right] < \infty$, we have*

$$\mathbb{E}\left[ZX|\mathcal{D}\right] = Z\mathbb{E}\left[X|\mathcal{D}\right] \quad \text{a.s.} \tag{16.10}$$

whenever the rv Z is $\mathcal{D}$-measurable.

In other words, when Z is a $\mathcal{D}$-measurable rv it acts as a constant in the conditioning process with respect to the partition, and can therefore be taken out of the conditional expectation.

Proof.

Simple rvs Let D denote an event in $\mathcal{D}$. If the $\mathcal{D}$-measurable rv Z is of the form $Z = \mathbf{1}[D']$ for some event D' in $\mathcal{D}$, then by repeated application of Theorem 16.2.1 we get

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}[D] \cdot \mathbb{E}\left[\mathbf{1}\left[D'\right] \cdot X|\mathcal{D}\right]\right] &= \mathbb{E}\left[\mathbf{1}[D] \cdot \left(\mathbf{1}\left[D'\right] \cdot X\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[D \cap D'\right] \cdot X\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[D \cap D'\right] \cdot \mathbb{E}[X|\mathcal{D}]\right] \\ &= \mathbb{E}\left[\mathbf{1}[D] \cdot \left(\mathbf{1}\left[D'\right] \cdot \mathbb{E}[X|\mathcal{D}]\right)\right] \qquad (16.11) \end{aligned}$$

since the event $D \cap D'$ is in the σ-field $\mathcal{D}$ given that both are in it. It follows that (16.10) holds for a $\mathcal{D}$-measurable rv Z of the form $Z = \mathbf{1}[D']$ with D' in $\mathcal{D}$.

If is now straightforward by linearity that (16.10) also holds for any simple $\mathcal{D}$-measurable rv Z, i.e., a rv Z of the form $Z = \sum_{i \in I} c_i \mathbf{1}[D_i]$ for a finite $\mathcal{D}$-partition $\{D_i,\ i \in I\}$ and scalars $\{c_i,\ i \in I\}$.

Non-negative rvs Now assume that the rv Z is non-negative. Lemma 10.3.1 guarantees the existence of a staircase approximation $\{Z_n,\ n = 1, 2, \ldots\}$ of Z made of simple non-negative $\mathcal{D}$-measurable rvs $\Omega \to \mathbb{R}_+$. For each $n = 1, 2, \ldots$, we have $|Z_n X| \leq |ZX|$, hence the integrability condition $\mathbb{E}[|ZX|] < \infty$ insures $\mathbb{E}[|Z_n X|] < \infty$ and we conclude to

$$\mathbb{E}[Z_n \cdot X|\mathcal{D}] = Z_n \cdot \mathbb{E}[X|\mathcal{D}] \quad \text{a.s.}$$

by the first part of the proof.

Arbitrary rvs ■

16.5 Taking expectations and iterated conditioning

The first result is a simple consequence of the characterization (16.1) with $D = \Omega$.

Lemma 16.5.1 *For any rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$, the rv $\mathbb{E}[X|\mathcal{D}]$ has a finite expectation with*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{D}]] = \mathbb{E}[X]. \qquad (16.12)$$

The next result forms the basis of the operational usefulness of conditioning through *pre-conditioning*.

Lemma 16.5.2 *Let $\mathcal{D}$ and $\mathcal{D}'$ be two sub-σ-fields of $\mathcal{F}$ with $\mathcal{D} \subseteq \mathcal{D}'$. For any rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$, we have*

$$\mathbb{E}\left[\mathbb{E}[X|\mathcal{D}]\,|\mathcal{D}'\right] = \mathbb{E}[X|\mathcal{D}] \quad \text{a.s.} \tag{16.13}$$

and

$$\mathbb{E}\left[\mathbb{E}\left[X|\mathcal{D}'\right]|\mathcal{D}\right] = \mathbb{E}[X|\mathcal{D}] \quad \text{a.s.} \tag{16.14}$$

Proof. Obviously, the rv $\mathbb{E}[X|\mathcal{D}]$ is $\mathcal{D}$-measurable, hence $\mathcal{D}'$-measurable. Using the localization property of conditional expectation given in Lemma 16.4.1 we get (16.13).

Pick D in $\mathcal{D}$, and note that

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}[D]\,\mathbb{E}\left[\mathbb{E}\left[X|\mathcal{D}'\right]|\mathcal{D}\right]\right] &= \mathbb{E}\left[\mathbf{1}[D]\,\mathbb{E}\left[X|\mathcal{D}'\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}[D]\,X\right] \\ &= \mathbb{E}\left[\mathbf{1}[D]\,\mathbb{E}[X|\mathcal{D}]\right]. \end{aligned} \tag{16.15}$$

The first equality used (16.1) when taking the conditional expectation of the rv $\mathbb{E}[X|\mathcal{D}']$ with respect to the sub-σ-field $\mathcal{D}$, while the second equality uses (16.1) when taking the conditional expectation of the rv X with respect to the sub-σ-field $\mathcal{D}'$ – Recall that the $\mathcal{D}$-measurable event D is also $\mathcal{D}'$-measurable. The final equality uses (16.1) when taking the conditional expectation of the rv X with respect to the sub-σ-field $\mathcal{D}$. The desired conclusion then follows by uniqueness. ■

16.6 Conditional expectations and independence

Lemma 16.6.1 *Consider a rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$. If the rv X is independent of the σ-field $\mathcal{D}$, then*

$$\mathbb{E}[X|\mathcal{D}] = \mathbb{E}[X] \quad \text{a.s.} \tag{16.16}$$

Here, the independence of the rv X from the σ-field $\mathcal{D}$ means that for each D in $\mathcal{D}$, the rvs X and $\mathbf{1}[D]$ are independent.

Proof. By independence we note that

$$\mathbb{E}\left[\mathbf{1}\left[D\right]X\right]=\mathbb{P}\left[D\right]\mathbb{E}\left[X\right]=\mathbb{E}\left[\mathbf{1}\left[D\right]\mathbb{E}\left[X\right]\right], \quad D\in\mathcal{D}$$

and the conclusion follows by uniqueness since the defining condition (16.1) holds for the constant rv $\mathbb{E}\left[X\right]$ (which is of course $\mathcal{D}$-measurable). ■

Lemma 16.6.1 is often used when the conditioning σ-field $\mathbb{D}$ is generated by some rv $Y:\Omega\rightarrow\mathbb{R}^q$, thus $\mathcal{D}=\sigma(Y)$, and the rvs X and Y are independent (in which case the σ-field $\mathcal{D}$ and the rv X are independent in the sense used earlier).

Consider a Borel mapping $h:\mathbb{R}^p\times\mathbb{R}^q\rightarrow\mathbb{R}$, and rvs $X:\Omega\rightarrow\mathbb{R}^p$ and $Y:\Omega\rightarrow\mathbb{R}^q$ such that $\mathbb{E}\left[|h(X,Y)|\right]<\infty$. Define the mapping $\widehat{h}:\mathbb{R}^q\rightarrow\mathbb{R}$ given by

$$\widehat{h}(y)=\mathbb{E}\left[h(X,y)\right], \quad y\in\mathbb{R}^q.$$

This definition is always well posed, and produces a Borel mapping $\mathbb{R}^q\rightarrow\mathbb{R}$.

Lemma 16.6.2 *If the rv X is independent of the σ-field $\mathcal{D}$ and the rv Y is $\mathcal{D}$-measurable, then*

$$\mathbb{E}\left[h(X,Y)|\mathcal{D}\right]=\widehat{h}(Y) \quad \textit{a.s.}$$

Proof. The proof proceeds according to the usual pattern.

Case I Consider first the case where the Borel mapping $h:\mathbb{R}^p\times\mathbb{R}^q\rightarrow\mathbb{R}$ is of the form

$$h(x,y)=\mathbf{1}\left[y\in C\right]g(x), \quad \begin{array}{c}x\in\mathbb{R}^p\\ y\in\mathbb{R}^q\end{array}$$

with Borel mapping $g:\mathbb{R}^p\rightarrow\mathbb{R}$ such that $\mathbb{E}\left[|g(X)|\right]<\infty$ and Borel subset C in $\mathcal{B}\left(\mathbb{R}^q\right)$. For every D in $\mathcal{D}$, the event $D\cap\left[Y\in C\right]$ belongs to $\mathcal{D}$ under the foregoing assumptions. It follows that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left[D\right]\mathbb{E}\left[h(X,Y)|\mathcal{D}\right]\right] &= \mathbb{E}\left[\mathbf{1}\left[D\right]h(X,Y)\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[D\right]\mathbf{1}\left[Y\in C\right]g(X)\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[D\cap\left[Y\in C\right]\right]g(X)\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[D\cap\left[Y\in C\right]\right]\mathbb{E}\left[g(X)|\mathcal{D}\right]\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[D\cap\left[Y\in C\right]\right]\mathbb{E}\left[g(X)\right]\right]\\ &= \mathbb{E}\left[\mathbf{1}\left[D\right]\mathbf{1}\left[Y\in C\right]\mathbb{E}\left[g(X)\right]\right] \qquad (16.17)\end{aligned}$$

as we made use of the fact that the rv X is independent of the σ-field $\mathcal{D}$. By the usual uniqueness argument, we conclude that

$$\mathbb{E}\left[h(X,Y)|\mathcal{D}\right] = \mathbf{1}\left[Y \in C\right]\mathbb{E}\left[g(X)\right] = \widehat{h}(Y) \quad \text{a.s.}$$

upon noting that here

$$\widehat{h}(y) = \mathbb{E}\left[h(X,y)\right] = \mathbb{E}\left[\mathbf{1}\left[y \in C\right]g(X)\right] = \mathbf{1}\left[y \in C\right]\mathbb{E}\left[g(X)\right], \quad y \in \mathbb{R}^q.$$

Case II The result immediately follows for any Borel mapping $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the form

$$h(x,y) = \sum_{i\in I} \mathbf{1}\left[y \in C_i\right] g_i(x), \qquad \begin{array}{c} x \in \mathbb{R}^p \\ y \in \mathbb{R}^q \end{array}$$

with I a finite index set, and for each i in I, Borel mapping $g_i : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\left[|g_i(X)|\right] < \infty$ and Borel subset C_i in $\mathcal{B}\left(\mathbb{R}^q\right)$. By additivity, we get

$$\begin{aligned} \mathbb{E}\left[h(X,Y)|\mathcal{D}\right] &= \mathbb{E}\left[\sum_{i\in I} \mathbf{1}\left[Y \in C_i\right] g_i(X)|\mathcal{D}\right] \quad \text{a.s.} \\ &= \sum_{i\in I} \mathbb{E}\left[\mathbf{1}\left[Y \in C_i\right] g_i(X)|\mathcal{D}\right] \quad \text{a.s.} \\ &= \sum_{i\in I} \mathbf{1}\left[Y \in C_i\right] \mathbb{E}\left[g_i(X)\right] \quad \text{a.s.} \\ &= \widehat{h}(Y) \quad \text{a.s.} \qquad (16.18) \end{aligned}$$

as we note that

$$\begin{aligned} \widehat{h}(y) &= \mathbb{E}\left[h(X,y)\right] \\ &= \mathbb{E}\left[\sum_{i\in I} \mathbf{1}\left[y \in C_i\right] g_i(X)\right] \\ &= \sum_{i\in I} \mathbf{1}\left[y \in C_i\right] \mathbb{E}\left[g_i(X)\right], \quad y \in \mathbb{R}^q. \qquad (16.19) \end{aligned}$$

Case III ■

16.7 The σ-field generated by a countable partition

At this point the reader may wonder as to what is the connection between the conditioning with respect to a partition discussed in Chapter 15 and the notion of conditioning with respect to a σ-field defined in Section 16.2. The easiest way to understand how the latter indeed subsumes the former is to associate a sub-σ-field with the countable partition.

To that end, let $\{D_i,\ i \in I\}$ be a countable $\mathcal{F}$-partition of Ω, and consider the sub-σ-field $\mathcal{D} = \sigma\left(D_i,\ i \in I\right)$ generated by the partition $\{D_i,\ i \in I\}$. It is easy to check that every element D of $\mathcal{D}$ is of the form

$$D = \cup_{j \in J} D_j \tag{16.20}$$

for some countable subset $J \subseteq I$ (possibly empty if $D = \emptyset$ or $J = I$ if $D = \Omega$).

Fact 16.7.1 *Consider an $\mathcal{D}$-measurable rv $X : \Omega \to \mathbb{R}^p$ where $\mathcal{D} = \sigma\left(D_i,\ i \in I\right)$. For each i in I, the rv X is constant on the event D_i, and the values $\{X(\omega),\ \omega \in \Omega\}$ achieved by X form a countable set of points in $\mathbb{R}^p$.*

Proof. For each x in $\mathbb{R}^p$, the $\mathcal{D}$-measurability of X implies that the event $[X = x]$ is an element in $\mathcal{D}$. The result follows since any element D of $\mathcal{D}$ is necessarily of the form (16.20) for some countable subset $J \subseteq I$. ■

In particular, the rv $\mathbb{E}\left[X|D_i,\ i \in I\right]$ is an $\mathbb{R}$-valued rv as soon as $\mathbb{E}\left[X\right]$ is finite; as an extended rv, it is $\mathcal{D}$-measurable rv in the sense that

$$\left[\mathbb{E}\left[X|D_i,\ i \in I\right] \in C\right] \in \mathcal{D}, \quad C \in \mathcal{B}\left(\mathbb{R}\right).$$

This is a consequence of the observation that for each j in I, we have $\mathbb{E}\left[X|D_i,\ i \in I\right] = \mathbb{E}\left[X|D_j\right]$ on the event D_j.

In view of Lemma 15.8.2, the rv $\mathbb{E}\left[X|D_i,\ i \in I\right]$ belongs to an *a.s. equivalence class* of $\mathcal{D}$-measurable rvs which all satisfy (15.38). It is just a *representative* of this equivalence class. Following usage we shall refer to this equivalence class as the *conditional expectation* of the rv X *given* the σ-field $\mathcal{D}$, here $\sigma\left(D_i,\ i \in I\right)$, and we denote any of its representative by $\mathbb{E}\left[X|\mathcal{D}\right]$.

16.8 Countable partitions and discrete rvs

We briefly discuss how $\mathcal{F}$-partitions are induced by discrete rvs, and how this ultimately relates to conditional expectations with respect to such rvs: Consider a discrete rv $Y : \Omega \to \mathbb{R}^q$. By definition there exists a countable subset $S \subseteq \mathbb{R}^p$ such that $\mathbb{P}[Y \in S] = 1$. For ease of notation, with I countable we shall use the representation $S = \{y_i,\ i \in I\}$ where the elements are *distinct* and each of the events $\{[Y = y_i],\ i \in I\}$ is *non*-empty. So far we can only assert that the event

$$\Omega_Y \equiv \cup_{i \in I} [Y = y_i]$$

has probability one, or equivalently, that the complement Ω_Y^c has zero probability. Nothing precludes the set of values

$$\{Y(\omega),\ \omega \notin \Omega_Y\}$$

to form an *uncountable* set. Only when that set is empty, will the collection $\{[Y = y_i],\ i \in I\}$ be an $\mathcal{F}$-partition of Ω.

To remedy this difficulty, pick an element b *not* in S and define the discrete rv $Y_b : \Omega \to \mathbb{R}^q$ by

$$Y_b(\omega) \equiv \begin{cases} Y(\omega) & \text{if } \omega \in \Omega_Y \\ & \\ b & \text{if } \omega \notin \Omega_Y. \end{cases}$$

The collection $\{[Y = b],\ [Y = y_i],\ i \in I\}$ is now an $\mathcal{F}$-partition of Ω. The following facts are easy consequences from the following observation

$$\mathbb{P}[Y \neq Y_b] \leq \mathbb{P}[\Omega_Y^c] = 0.$$

(i) The rvs Y and Y_b have the same probability distribution under $\mathbb{P}$. If $X : \Omega \to \mathbb{R}^p$ is another rv, the pairs (X, Y) and (X, Y_b) have the same probability distribution under $\mathbb{P}$.

(ii) Consider a Borel mapping $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ such that $\mathbb{E}[|h(X, Y)|] < \infty$. With

$$S_b = \{y_i, \in I;\ b\} = S \cup \{b\},$$

and $\mathcal{D}_b = \sigma([Y = b],\ [Y = y_i],\ i \in I)$, we note that

$$\begin{aligned} & \mathbb{E}[h(X, Y)|\mathcal{D}_b] \\ & \quad = \ \mathbb{E}[h(X, Y_b)|\mathcal{D}_b] \\ & \quad = \ \sum_{y \in S_b} \mathbb{E}[h(X, Y_b)|Y_b = y]\, \mathbf{1}[Y_b = y] \\ & \quad = \ \sum_{y \in S} \mathbb{E}[h(X, Y_b)|Y_b = y]\, \mathbf{1}[Y_b = y] + \mathbb{E}[h(X, Y_b)|Y_b = b]\, \mathbf{1}[Y_b = b] \\ & \quad = \ \sum_{y \in S} \frac{\mathbb{E}[\mathbf{1}[Y_b = y]\, h(X, Y_b)]}{\mathbb{P}[Y_b = y]} \mathbf{1}[Y_b = y] + \mathbb{E}[h(X, Y_b)|Y_b = b]\, \mathbf{1}[Y_b = b] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{y\in S} \frac{\mathbb{E}\left[\mathbf{1}\left[Y_b = y\right] h(X,y)\right]}{\mathbb{P}\left[Y_b = y\right]} \mathbf{1}\left[Y_b = y\right] + \mathbb{E}\left[h(X,Y_b)|Y_b = b\right] \mathbf{1}\left[Y_b = b\right] \\
&= \sum_{y\in S} \frac{\mathbb{E}\left[\mathbf{1}\left[Y = y\right] h(X,y)\right]}{\mathbb{P}\left[Y = y\right]} \mathbf{1}\left[Y = y\right] + \mathbb{E}\left[h(X,Y_b)|Y_b = b\right] \mathbf{1}\left[Y_b = b\right]
\end{aligned}
$$

It follows that

$$
\mathbb{E}\left[h(X,Y_b)|\mathcal{D}_b\right] = \sum_{y\in S} \frac{\mathbb{E}\left[\mathbf{1}\left[Y = y\right] h(X,y)\right]}{\mathbb{P}\left[Y = y\right]} \mathbf{1}\left[Y = y\right] \quad \mathbb{P}\text{-a.s.}
$$

(iii) In light of this last calculation, with Borel mapping $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ such that $\mathbb{E}\left[|h(X,Y)|\right] < \infty$, for distinct values $b \neq c$ in $\mathbb{R}^q$, we have

$$
\mathbb{E}\left[h(X,Y_b)|\mathcal{D}_b\right] = \mathbb{E}\left[h(X,Y_c)|\mathcal{D}_c\right] \quad \mathbb{P}\text{-}a.s.
$$

where we use the notation $\mathcal{D}_b = \sigma([Y = b],\ [Y = y_i],\ i \in I)$ and $\mathcal{D}_c = \sigma([Y = c],\ [Y = y_i],\ i \in I)$.

In other words, although the two conditional expectation rvs are *not* necessarily identical (as mappings $\Omega \to \mathbb{R}$), they are equal to each other except on a set of zero probability measure (under $\mathbb{P}$). As this notion defines an equivalence relation on rvs, we write $\mathbb{E}\left[h(X,Y)|Y\right]$ (or sometimes $\mathbb{E}\left[h(X,Y)|\sigma(Y)\right]$) to denote *any* representative in the equivalence class.

(iv) One standard representative in that class of $\mathbb{P}$-equivalent rvs is given by

$$
\sum_{y\in S} \mathbb{E}\left[h(X,Y)|Y = y\right] \mathbf{1}\left[Y = y\right] \tag{16.21}
$$

Note that all the terms in (16.21) are well defined in terms of Y! It is *convenient* to use this expression when *representing* the conditional expectation of $h(X,Y)$ given Y.

(v) Next, observe that

$$
\begin{aligned}
&\sum_{y\in S} \mathbb{E}\left[h(X,Y)|Y = y\right] \mathbf{1}\left[Y = y\right] \\
&\qquad = \sum_{y\in S} \frac{\mathbb{E}\left[\mathbf{1}\left[Y = y\right] h(X,Y)\right]}{\mathbb{P}\left[Y = y\right]} \mathbf{1}\left[Y = y\right] \\
&\qquad = \sum_{y\in S} \frac{\mathbb{E}\left[\mathbf{1}\left[Y = y\right] h(X,y)\right]}{\mathbb{P}\left[Y = y\right]} \mathbf{1}\left[Y = y\right] \\
&\qquad = \sum_{y\in S} \mathbb{E}\left[h(X,y)|Y = y\right] \mathbf{1}\left[Y = y\right] \qquad (16.22)
\end{aligned}
$$

This last expression suggests introducing the mapping $\widehat{h} : \mathbb{R}^q \to \mathbb{R}$ given by

$$\widehat{h}(y) = \begin{cases} \mathbb{E}\left[h(X,y)|Y=y\right] & \text{if } y \in S \\ \\ h^\star(y) & \text{if } y \notin S \end{cases}$$

where $h^\star : \mathbb{R}^q \to \mathbb{R}$ is an arbitrary Borel mapping such that $\mathbb{E}\left[|h^\star(Y)|\right] < \infty$. This definition is always well posed, and produces a Borel mapping $\mathbb{R}^q \to \mathbb{R}$.

With this notation we conclude that

$$\begin{aligned} &\sum_{y\in S} \mathbb{E}\left[h(X,Y)|Y=y\right] \mathbf{1}\left[Y=y\right] \\ &\quad= \sum_{y\in S} \mathbb{E}\left[h(X,y)|Y=y\right] \mathbf{1}\left[Y=y\right] \\ &\quad= \sum_{y\in S} \widehat{h}(y)\mathbf{1}\left[Y=y\right] \\ &\quad= \sum_{y\in S} \widehat{h}(Y)\mathbf{1}\left[Y=y\right] \\ &\quad= \widehat{h}(Y)\left(\sum_{y\in S} \mathbf{1}\left[Y=y\right]\right) \\ &\quad= \widehat{h}(Y) \quad \mathbb{P}\text{-a.s.} \end{aligned} \tag{16.23}$$

since

$$\sum_{y\in S} \mathbf{1}\left[Y=y\right] = \mathbf{1}\left[Y\in S\right] = 1 \quad \mathbb{P}\text{-a.s.}$$

(vi) Symbolically, this last discussion can be summarized as follows:

$$\begin{aligned} \mathbb{E}\left[h(X,Y)|Y\right] &= \left(\mathbb{E}\left[h(X,Y)|Y=y\right]\right)_{y=Y} \\ &= \left(\mathbb{E}\left[h(X,y)|Y=y\right]\right)_{y=Y} \quad \mathbb{P}\text{-a.s.} \end{aligned} \tag{16.24}$$

16.9 The absolutely continuous case

Consider rvs $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$. If the rv Y is absolutely continuous, then

$$\mathbb{P}\left[Y=y\right] = 0, \quad y \in \mathbb{R}^q$$

since

$$\int_{\{y\}} f_Y(\eta)d\eta = 0.$$

As a result, for each y in $\mathbb{R}^q$ we *cannot* define the conditional probabilities

$$\mathbb{P}\left[X \in B | Y = y\right] = \frac{\mathbb{P}\left[X \in B, Y = y\right]}{\mathbb{P}\left[Y = y\right]}, \quad B \in \mathcal{B}(\mathbb{R}^p).$$

With y in $\mathbb{R}^q$, the ball centered at y with radius $\varepsilon > 0$ is denoted by $B_\varepsilon(y)$, i.e.,

$$B_\varepsilon(y) \equiv \left\{\eta \in \mathbb{R}^q : \ \|\eta - y\| \leq \varepsilon\right\}.$$

Pick y in $\mathbb{R}^q$ such that $f_Y(y) > 0$ and assume there exists $\varepsilon_0 > 0$ such that

$$\mathbb{P}\left[Y \in B_\varepsilon(y)\right] > 0, \quad 0 < \varepsilon \leq \varepsilon_0.$$

The basic idea is as follows: Pick B in $\mathcal{B}(\mathbb{R}^p)$. Whatever definition is given to the conditional probability $\mathbb{P}\left[X \in B | Y = y\right]$, it is reasonable to expect that it should be compatible with the limiting value $\lim_{\varepsilon \downarrow 0} \mathbb{P}\left[X \in B | Y \in B_\varepsilon(y)\right]$ if it exists.

With this in mind we note that

$$\begin{aligned}
\mathbb{P}\left[X \in B | Y \in B_\varepsilon(y)\right] &= \frac{\mathbb{P}\left[[X \in B] \cap B_\varepsilon(y)\right]}{\mathbb{P}\left[B_\varepsilon(y)\right]} \\
&= \frac{\int_{B \times B_\varepsilon(y)} f_{XY}(\xi, \eta) d\xi d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta} \\
&= \frac{\int_B \left(\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta\right) d\xi}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta} \\
&= \int_B \left(\frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta}\right) d\xi \qquad (16.25)
\end{aligned}$$

Note that

$$\lim_{\varepsilon \downarrow 0} \int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta = 0, \quad \xi \in \mathbb{R}^p$$

and

$$\lim_{\varepsilon \downarrow 0} \int_{B_\varepsilon(y)} f_Y(\eta) d\eta = 0.$$

However, in many cases of interest in applications, we find that these limits have the same rate of convergence so that the limit

$$\lim_{\varepsilon \downarrow 0} \frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta}$$

in fact exists. This is analogous to the situation handled by L'Hospital's rule when the indeterminate form $\frac{0}{0}$ arises. Indeed note that under broad conditions it holds

$$\lim_{\varepsilon \downarrow 0} \frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\lambda(B_\varepsilon(y))} = f_{XY}(\xi, y), \quad \xi \in \mathbb{R}^p$$

and

$$\lim_{\varepsilon \downarrow 0} \frac{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta}{\lambda(B_\varepsilon(y))} = f_Y(y).$$

where $\lambda(B_\varepsilon(y))$ denotes the Lebesgue measure of the ball $B_\varepsilon(y)$. It now follows that

$$\begin{aligned} \lim_{\varepsilon \downarrow 0} \frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta} &= \lim_{\varepsilon \downarrow 0} \frac{\frac{\int_{B_\varepsilon(y)} f_{XY}(\xi,\eta) d\eta}{\lambda(B_\varepsilon(y))}}{\frac{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta}{\lambda(B_\varepsilon(y))}} \\ &= \frac{f_{XY}(\xi, y)}{f_Y(y)}, \quad \xi \in \mathbb{R}^p. \end{aligned} \tag{16.26}$$

This suggests

$$\begin{aligned} \lim_{\varepsilon \downarrow 0} \mathbb{P}\left[X \in B | Y \in B_\varepsilon(y)\right] &= \lim_{\varepsilon \downarrow 0} \int_B \left(\frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta} \right) d\xi \\ &= \int_B \lim_{\varepsilon \downarrow 0} \left(\frac{\int_{B_\varepsilon(y)} f_{XY}(\xi, \eta) d\eta}{\int_{B_\varepsilon(y)} f_Y(\eta) d\eta} \right) d\xi \\ &= \int_B \frac{f_{XY}(\xi, y)}{f_Y(y)} d\xi \end{aligned} \tag{16.27}$$

under the assumption that the interchange of limit and integration is permissible.

With y in $\mathbb{R}^q$, define the mapping $f_{X|Y}(\cdot|y) : \mathbb{R}^p \to \mathbb{R}_+$ by

$$f_{X|Y}(x|y) \equiv \begin{cases} \frac{f_{XY}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ & \\ g(x) & \text{if } f_Y(y) = 0 \end{cases}$$

where the Borel mapping $g : \mathbb{R}^p \to \mathbb{R}_+$ is a probability density function, hence satisfies

$$\int_{\mathbb{R}^p} g(x) dx = 1.$$

Computing conditional expectations (I) Consider a Borel mapping $u : \mathbb{R}^p \to \mathbb{R}$ such that that $\mathbb{E}\left[|u(X)|\right] < \infty$, and pick a Borel set C in $\mathcal{B}(\mathbb{R}^q)$. Note that

$$\mathbb{P}\left[[Y \in C] \cap [f_Y(Y) = 0]\right] = 0$$

since

$$\mathbb{P}\left[f_Y(Y) = 0\right] = \int_{\{\eta \in \mathbb{R}^q:\ f_Y(\eta)=0\}} f_Y(\eta) d\eta = 0.$$

With

$$C_Y^+ \equiv \left\{\eta \in \mathbb{R}^q :\ f_Y(\eta) > 0\right\},$$

this becomes

$$\mathbb{P}\left[Y \notin C_Y^+\right] = \mathbb{P}\left[f_Y(Y) = 0\right] = 0.$$

We find

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left[Y \in C\right] u(X)\right] &= \mathbb{E}\left[\mathbf{1}\left[Y \in C, f_Y(Y) > 0\right] u(X)\right] \\
&= \int_{\mathbb{R}^p \times (C \cap C_Y^+)} u(\xi) f_{XY}(\xi, \eta) d\xi d\eta \\
&= \int_{C \cap C_Y^+} \left(\int_{\mathbb{R}^p} u(\xi) f_{XY}(\xi, \eta) d\xi \right) d\eta
\end{aligned}$$

by Fubini's Theorem.

If $f_Y(\eta) > 0$, then

$$\begin{aligned}
\int_{\mathbb{R}^p} u(\xi) f_{XY}(\xi, \eta) d\xi &= \int_{\mathbb{R}^p} u(\xi) f_{X|Y}(\xi|\eta) f_Y(\eta) d\xi \\
&= \left(\int_{\mathbb{R}^p} u(\xi) f_{X|Y}(\xi|\eta) d\xi \right) f_Y(\eta) \\
&= \widehat{u}(\eta) f_Y(\eta) \qquad (16.28)
\end{aligned}$$

as we define $\widehat{u} : \mathbb{R}^q \to \mathbb{R}$ given by

$$\widehat{u}(y) = \int_{\mathbb{R}^p} u(\xi) f_{X|Y}(\xi|y) d\xi, \quad y \in \mathbb{R}^q.$$

It can be shown that the mapping $\widehat{u} : \mathbb{R}^q \to \mathbb{R}$ is well defined and Borel.

It follows that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left[Y \in C\right] u(X)\right] &= \int_{C \cap C_Y^+} \widehat{u}(\eta) f_Y(\eta) d\eta \\
&= \int_C \widehat{u}(\eta) f_Y(\eta) d\eta \\
&= \mathbb{E}\left[\mathbf{1}\left[Y \in C\right] \widehat{u}(Y)\right]. \qquad (16.29)
\end{aligned}$$

Recalling that $\sigma(Y) = \{Y \in C,\ C \in \mathcal{B}(\mathbb{R}^q)\}$, we conclude that

$$\mathbb{E}\left[u(X)|\sigma(Y)\right] = \widehat{u}(Y) \quad \mathbb{P}\text{-a.s.}$$

Computing conditional expectations (II) In a similar way, consider a Borel mapping $v : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ such that $\mathbb{E}\left[|v(X,Y)|\right] < \infty$, Then,

$$\mathbb{E}\left[\mathbf{1}\left[Y \in C\right] v(X,Y)\right] \quad = \quad \mathbb{E}\left[\mathbf{1}\left[Y \in C\right] \widehat{v}(Y)\right]. \tag{16.30}$$

where we define $\widehat{v} : \mathbb{R}^q \to \mathbb{R}$ given by

$$\widehat{v}(y) = \int_{\mathbb{R}^p} v(\xi, y) f_{X|Y}(\xi|y) d\xi, \quad y \in \mathbb{R}^q.$$

It can be shown that the mapping $\widehat{v} \to \mathbb{R}$ is well defined and Borel. Here as well we have

$$\mathbb{E}\left[v(X,Y)|\sigma(Y)\right] = \widehat{v}(Y) \quad \mathbb{P}\text{-a.s.}$$

16.10 A proof of Lemma 16.1.1

Claim (i): The conclusion is immediate from the fact that

$$\begin{aligned}[Z \in B] &= [g(Y) \in B] \\ &= [Y \in g^{-1}(B)] \in \mathcal{D}, \quad B \in \mathcal{B}(\mathbb{R}).\end{aligned}$$

since Y is $\mathcal{D}$-measurable and $g^{-1}(B)$ belongs to $\mathcal{B}(\mathbb{R}^p)$ by the Borel measurability of g.

Claim (ii): Conversely, assume that the rv $Z : \Omega \to \mathbb{R}$ is $\mathcal{D}$-measurable. The proof proceeds in three standard steps:

Simple rvs First assume that $Z = \mathbf{1}\left[D\right]$ for some D in $\sigma(Y)$, in which case $D = [Y \in C]$ for some C in $\mathcal{B}(\mathbb{R}^q)$. It is now plain that $Z = g_C(Y)$ with Borel mapping $g_C : \mathbb{R}^q \to \mathbb{R}$ given by

$$g_C(y) = \begin{cases} 0 & \text{if } y \notin C \\ \\ 0 & \text{if } y \in C. \end{cases} \tag{16.31}$$

The desired conclusion is readily seen to hold for simple $\mathcal{D}$-measurable rvs of the form

$$Z = \sum_{i \in I} a_i \mathbf{1}\left[D_i\right]$$

where I is a finite index, $\{D_i,\ i \in I\}$ form a $\mathcal{D}$-partition of Ω and $\{a_i,\ i \in I\}$ are the associated scalars. Indeed, for each i in I, we have $D_i = [Y \in C_i]$ for some C_i in $\mathcal{B}(\mathbb{R}^p)$, so that $Z = g(Y)$ with Borel mapping $g : \mathbb{R}^q \to \mathbb{R}$ is given by

$$g(y) = \sum_{i \in I} a_i g_{C_i}(y), \quad y \in \mathbb{R}^q$$

where the mapping $g_{C_i} : \mathbb{R}^q \to \mathbb{R}$ is associated with C_i through (16.31).

Non-negative rvs For any non-negative $\mathcal{D}$-measurable rv $Z : \Omega \to \mathbb{R}_+$, we introduce the usual monotone increasing sequence of simple rvs $\{Z_n,\ n = 1, 2, \ldots\}$ given by

$$Z_n = \sum_{m=0}^{n-1} \sum_{k=0}^{2^n-1} \frac{k}{2^n} \mathbf{1}\left[\frac{k}{2^n} < Z \leq \frac{k+1}{2^n}\right], \quad n = 1, 2, \ldots$$

with $\lim_{n\to\infty} Z_n = Z$. Obviously, the simple rvs $\{Z_n,\ n = 1, 2, \ldots\}$ are all $\mathcal{D}$-measurable, hence by the last part of the proof, for each $n = 1, 2, \ldots$, there exists a Borel mapping $g_n : \mathbb{R}^q \to \mathbb{R}$ such that

$$Z_n = g_n(Y), \quad n = 1, 2, \ldots$$

with the point wise convergence implying

$$Z(\omega) = \lim_{n\to\infty} Z_n(\omega) = \lim_{n\to\infty} g_n(Y(\omega)), \quad \omega \in \Omega.$$

Now define the subset $L \subseteq \mathbb{R}^q$ by $L \equiv \{y \in \mathbb{R}^q :\ \lim_{n\to\infty} g_n(y) \text{ exists in } \mathbb{R}\}$. The set L being a Borel subset of $\mathbb{R}^q$, it readily follows that the mapping $g : \mathbb{R}^q \to \mathbb{R}$ given by

$$g(y) \equiv \begin{cases} \lim_{n\to\infty} g_n(y) & \text{if } y \in L \\ \\ 0 & \text{if } y \notin L \end{cases}$$

is a Borel mapping. By construction it is plain that $Z = g(Y)$ since $Y(\omega)$ lies in L for each ω in Ω.

The general case The case of an arbitrary $\mathcal{D}$-measurable rv $Z : \Omega \to \mathbb{R}$ is handled in the usual manner: Just write $Z = Z^+ - Z^-$, and apply the last conclusion to each of the rvs Z^+ and Z^-. In particular, there exist Borel mappings $g_+ : \mathbb{R}^q \to \mathbb{R}_+$ and $g_- : \mathbb{R}^q \to \mathbb{R}_+$ such that $Z^+ = g_+(Y)$ and $Z^- = g_-(Y)$. The desired Borel mapping $g : \mathbb{R}^q \to \mathbb{R}$ is then simply given by

$$g(y) = g_+(y) - g_-(y), \quad y \in \mathbb{R}^q.$$

Note that it is not necessarily the case that $g_\pm(y) = \max(0, \pm g(y))$. ■

16.11 Exercises

Ex. 16.1 Consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$.

a. Compute $\mathbb{E}[X|\mathcal{T}]$ where $\mathcal{T}$ denotes the trivial σ-field on Ω.

b. Compute $\mathbb{E}[X|\mathcal{F}]$.

Ex. 16.2 With $\mathcal{D}$ a sub-σ-field of $\mathcal{F}$, let $X : \Omega \to \mathbb{R}^p$ denote a $\mathcal{D}$-measurable rv. List all the rvs $\Omega \to \mathbb{R}^p$ which are $\mathcal{D}$-measurable when $\mathcal{D}$ is the trivial σ-field $\mathcal{D} = \{\emptyset, \Omega\}$.

Ex. 16.3 Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two sub-σ-fields of $\mathcal{F}$ such that $\mathcal{D}_1 \subseteq \mathcal{D}_2$ (so $\mathcal{D}_1$ is a sub-σ-field of $\mathcal{D}_2$).

a. Show that a rv $X : \Omega \to \mathbb{R}^p$ which is $\mathcal{D}_1$-measurable is also $\mathcal{D}_2$-measurable.

b. Consider now a rv $X : \Omega \to \mathbb{R}^p$ which is $\mathcal{D}_2$-measurable. Is it automatically $\mathcal{D}_1$-measurable? Either prove or give a counterexample.

Ex. 16.4 With $\Omega = \mathbb{N}$, consider the σ-fields $\mathcal{F}$ and $\mathcal{D}$ on Ω defined by $\mathcal{F} \equiv \sigma(\{n\},\ n = 0, 1, \ldots)$ and $\mathcal{D} \equiv \sigma(\{2n, 2n+1\},\ n = 0, 1, \ldots)$.

a. Show that $\mathcal{D}$ is a strict sub-σ-field of $\mathcal{F}$ by giving an event E in $\mathcal{F}$ which is not in $\mathcal{D}$.

b. Give a rv $X : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F})$ which is not $\mathcal{D}$-measurable.

c. Give a rv $X : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F})$ which **is** $\mathcal{D}$-measurable.

Ex. 16.5 We start with a collection $\{U_1, U_2, \ldots, U_n\}$ of n rvs, each uniformly distributed over the interval $(0, 1)$, and let P denote a rv with the property that $\mathbb{P}[0 < P \le 1] = 1$. Moreover assume that the $n+1$ rvs $P, U_1, \ldots, U_n$ are mutually independent rvs. Under these assumptions we are interested in the rv X defined by

$$X \equiv \sum_{i=1}^{n} \mathbf{1}[U_i \le P].$$

Using pre-conditioning arguments to answer the following questions:

a. Compute $\mathbb{E}[X]$ in terms of $\mathbb{E}[P]$.

b. How many moments of P do you need to know in order to compute $\mathrm{Var}[X]$?

c. Are the rvs $\mathbf{1}[U_1 \le P], \ldots, \mathbf{1}[U_n \le P]$ (i) mutually independent (ii) pairwise uncorrelated when S contains at least two elements?

d. Compute the probabilities

$$\mathbb{P}\left[X = k\right], \quad k = 0, 1, \ldots, n.$$

How many moments of P are needed?

Ex. 16.6 The rvs $X, X_1, \ldots, X_n$, all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, are i.i.d. rvs with $\mathbb{E}\left[|X|\right] < \infty$.

a. Compute

$$\mathbb{E}\left[X_i | X_1 + \ldots + X_n\right]$$

for each $i = 1, \ldots, n$. The answer does not depend on the (common) probability distrubition function of $X_1, \ldots, X_n$! [HINT: Does the probability distribution of the pair $(X_i, X_1 + \ldots + X_n)$ depend on i?]

b. When $1 \leq k < n$, compute

$$\mathbb{E}\left[X_1 + \ldots + X_k | X_1 + \ldots + X_n\right]$$

c. When $1 \leq k < n$, compute

$$\mathbb{E}\left[X_1 + \ldots + X_n | X_1 + \ldots + X_k\right]$$

Ex. 16.7 The rvs $X, X_1, \ldots, X_n, Y, Y_1, \ldots, Y_n$, all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. We assume the following: (i) The rvs $X, X_1, \ldots, X_n, Y, Y_1, \ldots, Y_n$ are mutually independent; (ii) The rvs $X, X_1, \ldots, X_n$ are i.i.d. rvs with $\mathbb{E}\left[|X|\right] < \infty$; and (iii) The rvs $Y, Y_1, \ldots, Y_n$ are i.i.d. rvs with $\mathbb{E}\left[|Y|\right] < \infty$ – The rvs X and Y do not necessarily have the same probability distribution.

By using basic properties of conditional expectations, compute

$$\mathbb{E}\left[X_1 Y_1 + \ldots + X_n Y_n \middle| \begin{array}{c} X_1 + \ldots + X_n \\ Y_1 + \ldots + Y_n \end{array}\right]$$

[HINT: Use iterated conditioning and compute

$$\mathbb{E}\left[X_1 Y_1 + \ldots + X_n Y_n \middle| \begin{array}{c} X_1, \ldots, X_n \\ Y_1 + \ldots + Y_n \end{array}\right].$$

Ex. 16.8 Consider the $\mathbb{R}$-valued rvs $U, U_1, \ldots, U_n$ which are all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The rvs $U, U_1, \ldots, U_n$ are assumed to be i.i.d. rvs, each of which is uniformly distributed on the interval $(0, 1)$. Now define the $\mathbb{R}$-valued rv X by

$$X = \sum_{k=1}^{n} \mathbf{1}\left[U_k \leq U\right]$$

a. With this definition as starting point, use *direct* probabilistic arguments to show that

$$\mathbb{P}\left[X = k\right] = \frac{1}{n+1}, \quad k = 0, \ldots, n.$$

b. Using conditioning arguments compute the conditional probability

$$\mathbb{P}\left[X = k|U\right], \quad k = 0, \ldots, n.$$

c. Use Parts **a** and **b** to evaluate the integrals

$$I_n(k) \equiv \int_0^1 t^k (1-t)^{n-k} dt, \quad k = 0, \ldots, n$$

Chapter 17

Probability distributions and their transforms

A number of developments concerning rvs and their probability distribution functions are sometimes best handled through transforms associated with them. There are a number of such transforms with varying ranges of applications. Here we focus mainly on the notion of characteristic function.

17.1 Definitions

All rvs are defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. For any element v in $\mathbb{R}^p$ (viewed as a column vector), we write v^t for its transpose, so that $v^t u$ is simply the scalar product $\sum_{i=1}^p u_i v_i$ between the two (column) vectors u and v. We begin with a basic definition.

Definition 17.1.1

With any rv $X : \Omega \to \mathbb{R}^p$, we associate its characteristic function $\Phi_X : \mathbb{R}^p \to \mathbb{C}$ given by

$$\Phi_X(\theta) \equiv \mathbb{E}\left[e^{i\theta^t X}\right], \quad \theta \in \mathbb{R}^p. \tag{17.1}$$

Characteristic functions are always well defined regardless of the type of probability distribution function for the rv X: Indeed the definition (17.1) is well posed since for each θ in $\mathbb{R}^p$, the rvs $\Omega \to \mathbb{R} : \omega \to \cos\left(\theta^t X(\omega)\right)$ and $\Omega \to \mathbb{R} : \omega \to$

$\sin\left(\theta^t X(\omega)\right)$ are both bounded. As a result, their expected values $\mathbb{E}\left[\cos\left(\theta^t X\right)\right]$ and $\mathbb{E}\left[\sin\left(\theta^t X\right)\right]$ are well defined and finite with

$$\left|\mathbb{E}\left[\cos\left(\theta^t X\right)\right]\right| \leq 1 \quad \text{and} \quad \left|\mathbb{E}\left[\sin\left(\theta^t X\right)\right]\right| \leq 1.$$

This fact allows us to make sense of (17.1) by linearity through the relations

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t X}\right] &= \mathbb{E}\left[\cos\left(\theta^t X\right) + i\sin\left(\theta^t X\right)\right] \\ &= \mathbb{E}\left[\cos\left(\theta^t X\right)\right] + i\mathbb{E}\left[\sin\left(\theta^t X\right)\right], \quad \theta \in \mathbb{R}^p. \end{aligned} \tag{17.2}$$

Characteristic functions are akin to Fourier transforms. For instance, if the rv X admits a probability density function $f_X : \mathbb{R}^p \to \mathbb{R}_+$, then

$$\Phi_X(\theta) = \int_{\mathbb{R}^p} e^{i\theta^t x} f_X(x) dx, \quad \theta \in \mathbb{R}^p.$$

If the rv X is a discrete rv with support $S \subseteq \mathbb{R}^p$, then

$$\Phi_X(\theta) = \sum_{x \in S} e^{i\theta^t x} \mathbb{P}\left[X = x\right], \quad \theta \in \mathbb{R}^p.$$

Obviously, the characteristic function Φ_X of the rv X is determined by its probability distribution function $F_X : \mathbb{R}^p \to [0,1]$. In fact we could rewrite (17.1) as

$$\Phi_X(\theta) = \int_{\mathbb{R}^p} e^{i\theta^t x} dF_X(x), \quad \theta \in \mathbb{R}^p. \tag{17.3}$$

This suggests writing Φ_X as Φ_{F_X}, and leads to the following definition.

Definition 17.1.2

With any probability distribution function $F : \mathbb{R}^p \to [0,1]$, we associate its characteristic function $\Phi_F : \mathbb{R}^p \to \mathbb{C}$ defined by

$$\Phi_F(\theta) \equiv \int_{\mathbb{R}^p} e^{i\theta^t x} dF(x), \quad \theta \in \mathbb{R}^p. \tag{17.4}$$

17.2 An inversion formula and uniqueness

The next result provides an *inversion formula* when $p = 1$ [**?**, Thm. 6.2.1, p. 153]. This result is of theoretical importance, and establish a one-to-one correspondence between a probability distribution function and its characteristic function. We give it here without proof. A more general version is also available; see [, Thm. , p.].

Theorem 17.2.1 *Consider a probability distribution function $F : \mathbb{R} \to [0,1]$, and let $\Phi_F : \mathbb{R} \to \mathbb{C}$ be its characteristic function. For $a < b$ in $\mathbb{R}$, it holds that*

$$\begin{aligned} & F(b-) - F(a) + \frac{F(a) - F(a-)}{2} + \frac{F(b) - F(b-)}{2} \\ = & \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \cdot \Phi_F(t) dt \end{aligned} \tag{17.5}$$

with the integrand being defined by continuity at $t = 0$.

Thus, when the probability distribution function $F : \mathbb{R} \to [0,1]$ is a continuous function, then (17.5) yields

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \cdot \Phi_F(t) dt \tag{17.6}$$

In the language of rvs, Theorem 17.2.1 can be reformulated as follows: For any rv $X : \Omega \to \mathbb{R}$ with characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$, it holds that

$$\begin{aligned} & \mathbb{P}\left[a < X < b\right] + \frac{\mathbb{P}\left[X = a\right]}{2} + \frac{\mathbb{P}\left[X = b\right]}{2} \\ = & \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \cdot \Phi_X(t) dt \end{aligned} \tag{17.7}$$

with arbitrary $a < b$ in $\mathbb{R}$.

The usefulness of the notion of characteristic function comes in part from the following uniqueness result which an easy byproduct of the inversion formula.

Theorem 17.2.2 *Let $F, G : \mathbb{R}^p \to [0,1]$ be two probability distribution functions on $\mathbb{R}^p$. If their characteristic functions coincide, namely*

$$\Phi_F(\theta) = \Phi_G(\theta), \quad \theta \in \mathbb{R}^p,$$

then the two probability distribution functions coincide, namely

$$F(x) = G(x), \quad x \in \mathbb{R}^p.$$

Thus, $\Phi_F = \Phi_G$ implies $F = G$. In other words, if a function $\mathbb{R}^p \to \mathbb{C}$ is known to be the characteristic function of some probability distribution function, there is no other probability distribution function that can generate this characteristic function. In the language of rvs, Theorem 17.2.2 states that if two rvs X and

Y (possibly defined on different probability triples) taking values in $\mathbb{R}^p$ have the same characteristic function, say $\Phi_X = \Phi_Y$, then their probability distributions must coincide, namely $F_X = F_Y$.

Sometimes a function $\Phi : \mathbb{R} \to \mathbb{C}$ arises in the discussion, and it is imperative to know whether it is the characteristic function of some rv. The terminology given next should facilitate the discussion of this issue presented in Sections 17.3 and 17.4.

Definition 17.2.1

A function $\Phi : \mathbb{R} \to \mathbb{C}$ is said to be a characteristic function if there exists a probability distribution $F : \mathbb{R}^p \to [0,1]$ such that

$$\Phi(\theta) = \int_{\mathbb{R}^p} e^{i\theta^t x} dF(x), \quad \theta \in \mathbb{R}^p \tag{17.8}$$

in which case $\Phi = \Phi_F$ by Theorem 17.2.2.

Alternatively, a function $\Phi : \mathbb{R} \to \mathbb{C}$ is said to be a characteristic function if there exists a rv $X : \Omega :\to \mathbb{R}^p$ such that

$$\Phi(\theta) = \mathbb{E}\left[e^{i\theta^t X}\right] = \Phi_X(\theta), \quad \theta \in \mathbb{R}^p. \tag{17.9}$$

17.3 Basic properties

Not every function $\mathbb{R}^p \to \mathbb{C}$ is a characteristic function. That much is clear from the basic properties derived in Theorem 17.3.1 given next.

Theorem 17.3.1 *Consider a rv $X : \Omega \to \mathbb{R}^p$ with characteristic function $\Phi_X : \mathbb{R}^p \to \mathbb{C}$ given by (17.1). It satisfies the following properties:*

(i) Boundedness: We have

$$|\Phi_X(\theta)| \leq \Phi_X(0) = 1 \quad \theta \in \mathbb{R}^p. \tag{17.10}$$

(ii) Uniform continuity on $\mathbb{R}^p$: We have

$$\lim_{\delta \to 0} \sup \left(|\Phi_X(\theta + \delta) - \Phi_X(\theta)| \, , \, \theta \in \mathbb{R}^p \right) = 0. \tag{17.11}$$

(iii) Positive semi-definiteness: For every $n = 1, 2, \ldots$, we have

$$\sum_{k=1}^{n} \sum_{\ell=1}^{n} \Phi_X(\theta_k - \theta_\ell) z_k z_\ell^\star \geq 0 \tag{17.12}$$

with arbitrary $z_1, \ldots, z_n$ in $\mathbb{C}$ and arbitrary $\theta_1, \ldots, \theta_n$ in $\mathbb{R}^p$.

(iv) Hermitian symmetry: We have

$$\Phi_X(-\theta) = \Phi_X(\theta)^\star, \quad \theta \in \mathbb{R}^p. \tag{17.13}$$

Much of the discussion makes use of the elementary relation

$$e^{i\theta x} - 1 = \int_0^x i\theta e^{i\theta s} ds, \quad x, \theta \in \mathbb{R} \tag{17.14}$$

so that the bounds

$$\left| e^{i\theta x} - 1 \right| \leq \int_0^x \left| i\theta e^{i\theta s} \right| ds \leq |\theta| x \tag{17.15}$$

hold.[1]

Proof. (i) It is plain that $\Phi_X(0) = 1$. Next,

$$|\Phi_X(\theta)| \leq \mathbb{E}\left[\left| e^{i\theta^t X} \right| \right] = 1, \quad \theta \in \mathbb{R}^p.$$

(ii) Fix θ and δ in $\mathbb{R}^p$. Since

$$e^{i(\theta+\delta)^t X} - e^{i\theta^t X} = e^{i\theta^t X} \left(e^{i\delta^t X} - 1 \right),$$

it follows that

$$\begin{aligned} |\Phi_X(\theta+\delta) - \Phi_X(\theta)| &= \left| \mathbb{E}\left[e^{i(\theta+\delta)^t X} \right] - \mathbb{E}\left[e^{i\theta^t X} \right] \right| \\ &= \left| \mathbb{E}\left[\left(e^{i\delta^t X} - 1 \right) e^{i\theta^t X} \right] \right| \\ &\leq \mathbb{E}\left[\left| \left(e^{i\delta^t X} - 1 \right) e^{i\theta^t X} \right| \right] \\ &= \mathbb{E}\left[\left| e^{i\delta^t X} - 1 \right| \right], \end{aligned}$$

so that

$$\sup\left(|\Phi_X(\theta+\delta) - \Phi_X(\theta)|, \ \theta \in \mathbb{R}^p \right) \leq \mathbb{E}\left[\left| e^{i\delta^t X} - 1 \right| \right]. \tag{17.16}$$

Uniform continuity follows if we can show that

$$\lim_{\delta \to 0} \mathbb{E}\left[\left| e^{i\delta^t X} - 1 \right| \right] = 1.$$

[1] With ab in $\mathbb{R}$, we have

$$|a + ib| = \sqrt{a^2 + b^2} \leq |a| + |b|.$$

This last statement is a simple consequence of the Bounded Convergence Theorem. as we note that $\lim_{\delta\to 0}\cos\left(\delta^t X\right) = 1$ and $\lim_{\delta\to 0}\sin\left(\delta^t X\right) = 0$.

(iii) Fix $n = 1, 2, \ldots$ and pick arbitrary $z_1, \ldots, z_n$ in $\mathbb{C}$: It is plain that

$$\begin{aligned}
&\sum_{k=1}^n \sum_{\ell=1}^n \Phi_X(\theta_k - \theta_\ell) z_k z_\ell^\star \\
&= \sum_{k=1}^n \sum_{\ell=1}^n \mathbb{E}\left[e^{j(\theta_k-\theta_\ell)^t X}\right] z_k z_\ell^\star \\
&= \mathbb{E}\left[\sum_{k=1}^n \sum_{\ell=1}^n e^{j(\theta_k-\theta_\ell)^t X} z_k z_\ell^\star\right] \\
&= \mathbb{E}\left[\sum_{k=1}^n \sum_{\ell=1}^n e^{j\theta_k^t X} e^{-\theta_\ell^t X} z_k z_\ell^\star\right] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^n e^{j\theta_k^t X} z_k\right)\left(\sum_{\ell=1}^n e^{j\theta_\ell^t X} z_\ell\right)^\star\right] \\
&= \mathbb{E}\left[\left|\sum_{k=1}^n e^{j\theta_k X} z_k\right|^2\right] \geq 0. \qquad (17.17)
\end{aligned}$$

(iv) Fix θ in $\mathbb{R}^p$. We note that

$$\begin{aligned}
\Phi_X(-\theta) &= \mathbb{E}\left[e^{-\theta^t X}\right] \\
&= \mathbb{E}\left[\cos\left(-\theta^t X\right)\right] + i\mathbb{E}\left[\sin\left(-\theta^t X\right)\right] \\
&= \mathbb{E}\left[\cos\left(\theta^t X\right)\right] - i\mathbb{E}\left[\sin\left(\theta^t X\right)\right] \\
&= \left(\mathbb{E}\left[\cos\left(\theta^t X\right)\right] + i\mathbb{E}\left[\sin\left(\theta^t X\right)\right]\right)^\star \\
&= \Phi_X(\theta)^\star \qquad (17.18)
\end{aligned}$$

as desired. ■

17.4 Bochner's Theorem

Interestingly enough the first three properties given in Theorem 17.3.1 turn out to be sufficient. This is a consequence of a deep result of Harmonic Analysis, known as the Bochner-Herglotz Theorem [**?**, Thm. 6.5.2, p. 179].

Theorem 17.4.1 *A function $\Phi : \mathbb{R}^p \to \mathbb{C}$ is a characteristic function if it is (i) bounded with $|\Phi(\theta)| \leq \Phi(0) = 1$ for all θ in $\mathbb{R}^p$; (ii) uniformly continuous on $\mathbb{R}^p$; and (iii) positive semi-definite.*

The property of positive semi-definiteness already implies the boundedness property (i). It also implies uniform continuity if $\Phi : \mathbb{R} \to \mathbb{C}$ is continuous at $\theta = 0$ [**?**, Thm. 6.5.1, p. 178]. This gives rise to the following sharp characterization.

Theorem 17.4.2 *A function $\Phi : \mathbb{R}^p \to \mathbb{C}$ is a characteristic function if and only if it is positive semi-definite and continuous at $\theta = 0$ with $\Phi(0) = 1$.*

17.5 Examples

Bernoulli rvs

$$\Phi_X(\theta) = pe^{i\theta} + (1-p), \quad \theta \in \mathbb{R} \tag{17.19}$$

Binomial rvs

$$\begin{aligned} \Phi_X(\theta) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{ik\theta} \\ &= \sum_{k=0}^{n} \binom{n}{k} \left(e^{i\theta} p\right)^k (1-p)^{n-k} \\ &= \left(1 - p + pe^{i\theta}\right), \quad \theta \in \mathbb{R} \end{aligned} \tag{17.20}$$

Poisson rvs

$$\begin{aligned} \Phi_X(\theta) &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{ik\theta} \\ &= \left(\sum_{k=0}^{\infty} \frac{(\lambda e^{i\theta})^k}{k!} \right) e^{-\lambda} \\ &= e^{-\lambda} e^{\lambda e^{i\theta}} \\ &= e^{-\lambda(1-e^{i\theta})}, \quad \theta \in \mathbb{R} \end{aligned} \tag{17.21}$$

Geometric rvs

$$\begin{aligned}\Phi_X(\theta) &= \sum_{k=0}^{\infty} p(1-p)^k e^{ik\theta} \\ &= \sum_{k=0}^{\infty} p\left((1-p)^k e^{i\theta}\right)^k \\ &= \frac{p}{1-(1-p)e^{i\theta}}, \quad \theta \in \mathbb{R} \qquad (17.22)\end{aligned}$$

Exponential rvs

$$\begin{aligned}\Phi_X(\theta) &= \int_0^{\infty} \lambda e^{-\lambda x} e^{i\theta x} dx \\ &= \lambda \int_0^{\infty} e^{(i\theta-\lambda)x} dx \\ &= \frac{\lambda}{i\theta-\lambda} \int_0^{\infty} (i\theta-\lambda) e^{(i\theta-\lambda)x} dx \\ &= \frac{\lambda}{i\theta-\lambda} \cdot \left[e^{(i\theta-\lambda)x} \right]_0^{\infty} \\ &= \frac{\lambda}{\lambda-i\theta}, \quad \theta \in \mathbb{R} \qquad (17.23)\end{aligned}$$

as we note that

$$\lim_{x\to\infty} e^{(i\theta-\lambda)x} = 0.$$

17.6 Independence via characteristic functions

The setting is as follows: Consider a collection of rvs $X_1, \ldots, X_k$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. For each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ has characteristic function $\Phi_{X_\ell} : \mathbb{R}^{p_\ell} \to \mathbb{C}$. We concatenate the rvs $X_1, \ldots, X_k$ into the rv $X : \Omega \to \mathbb{R}^p$ given by

$$X \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$$

where $p = p_1 + \ldots + p_k$. We denote the characteristic function of the rv X by $\Phi_X : \mathbb{R}^p \to \mathbb{C}$. We have the following useful characterization of independence in terms of characteristic functions.

Theorem 17.6.1 *If the rvs $X_1, \ldots, X_k$ are mutually independent, then*

$$\Phi_X(\theta) = \prod_{\ell=1}^k \Phi_{X_\ell}(\theta_\ell), \qquad \begin{array}{c} \theta_\ell \in \mathbb{R}^{p_\ell} \\ \ell = 1, \ldots, k \end{array} \tag{17.24}$$

with

$$\theta \equiv \left(\begin{array}{c} \theta_1 \\ \vdots \\ \theta_k \end{array} \right)$$

Conversely, if (17.24) holds on $\mathbb{R}^p$, then the rvs $X_1, \ldots, X_k$ are mutually independent.

Proof. Fix θ in $\mathbb{R}^p$. Noting that

$$\theta^t X = \sum_{\ell=1}^k \theta_\ell^t X_\ell,$$

we get

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t X}\right] &= \mathbb{E}\left[e^{i\sum_{\ell=1}^k \theta_\ell^t X_\ell}\right] \\ &= \mathbb{E}\left[\prod_{\ell=1}^k e^{i\theta_\ell^t X_\ell}\right] \\ &= \prod_{\ell=1}^k \mathbb{E}\left[e^{i\theta_\ell^t X_\ell}\right] \end{aligned} \tag{17.25}$$

by independence. The relation (17.24) follows.

Conversely, if (17.24) holds on $\mathbb{R}^p$, then

■

When $p_1 = \ldots = p_k = p$, consider the rv $S : \Omega \to \mathbb{R}^p$ given by

$$S = X_1 + \ldots + X_k.$$

Theorem 17.6.2 *If the rvs $X_1, \ldots, X_k$ are mutually independent, then*

$$\Phi_S(\theta) = \prod_{\ell=1}^k \Phi_{X_\ell}(\theta), \quad \theta \in \mathbb{R}^p. \tag{17.26}$$

Proof. Fix θ in $\mathbb{R}^p$. This time noting that $\theta^t S = \sum_{\ell=1}^k \theta^t X_\ell$, we get

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t S}\right] &= \mathbb{E}\left[e^{i\sum_{\ell=1}^k \theta^t X_\ell}\right] \\ &= \mathbb{E}\left[\prod_{\ell=1}^k e^{i\theta^t X_\ell}\right] \\ &= \prod_{\ell=1}^k \mathbb{E}\left[e^{i\theta^t X_\ell}\right] \qquad (17.27) \end{aligned}$$

by independence. ■

A case of particular interest arises when the rvs $X, X_1, \ldots, X_k$ are i.i..d. rvs. In that case, Theorem 17.6.2 yields

$$\Phi_S(\theta) = \Phi_X(\theta)^k, \quad \theta \in \mathbb{R}^p. \qquad (17.28)$$

17.7 Easy analytical facts

We consider the case $p = 1$. We begin with a simple fact that will prove useful in a number of places.

Theorem 17.7.1 *Fix x and θ in $\mathbb{R}$. For each $k = 1, 2, \ldots$, the expansion*

$$e^{i\theta x} = \sum_{\ell=0}^k \frac{1}{\ell!}\left(i\theta x\right)^\ell + R_k(x;\theta) \qquad (17.29)$$

holds with the remainder term given by

$$R_k(x;\theta) = (i\theta)^k \int_0^x \frac{(x-t)^{k-1}}{(k-1)!}\left(e^{i\theta t} - 1\right) dt. \qquad (17.30)$$

Proof. The proof proceed by induction: Throughout θ and x in $\mathbb{R}$ are scalars held fixed.

Basis step For $k = 1$, we use (17.14) to get

$$\begin{aligned} e^{i\theta x} - 1 &= \int_0^x i\theta e^{i\theta t} dt \\ &= \int_0^x i\theta \left(e^{i\theta t} - 1\right) dt + \int_0^x i\theta dt \\ &= i\theta x + i\theta \int_0^x \left(e^{i\theta t} - 1\right) dt \\ &= i\theta x + R_1(x;\theta) \qquad (17.31) \end{aligned}$$

by direct inspection.

Induction step Now assume that (17.30)-(17.30) holds for some $k = 1, 2, \ldots$. It is plain that

$$\begin{aligned} &\int_0^x \frac{(x-t)^{k-1}}{(k-1)!} \left(e^{i\theta t} - 1\right) dt \\ &= \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} \left(\int_0^t i\theta e^{i\theta s} ds\right) dt \\ &= \int_0^x \left(\int_0^t \frac{(x-t)^{k-1}}{(k-1)!} i\theta e^{i\theta s} ds\right) dt \\ &= \int_0^x \left(\int_s^x \frac{(x-t)^{k-1}}{(k-1)!} i\theta e^{i\theta s} dt\right) ds \\ &= \int_0^x \left(\int_s^x \frac{(x-t)^{k-1}}{(k-1)!} dt\right) i\theta e^{i\theta s} ds \\ &= \int_0^x i\theta \frac{(x-s)^k}{k!} e^{i\theta s} ds \qquad (17.32) \end{aligned}$$

since

$$\int_s^x \frac{(x-t)^{k-1}}{(k-1)!} dt = \left[-\frac{(x-t)^k}{k!}\right]_s^x = \frac{(x-s)^k}{k!}, \quad 0 \leq s \leq x.$$

Therefore, we have

$$\begin{aligned} R_k(x;\theta) &= (i\theta)^k \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} \left(e^{i\theta t} - 1\right) dt \\ &= (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} e^{i\theta s} ds \end{aligned}$$

$$
\begin{aligned}
&= (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} \left(e^{i\theta s} - 1 \right) ds + (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} ds \\
&= R_{k+1}(x;\theta) + (i\theta)^{k+1} \frac{x^{k+1}}{(k+1)!} \qquad (17.33)
\end{aligned}
$$

and the proof of the induction step is now completed. ■

17.8 Characteristic functions and moments

Since the probability distribution function of the rv X can be recovered from its characteristic function, it is not unreasonable to expect that there might be simple ways to recover moments whenever they exist and are finite. This is explored below.

Consider a rv $X : \Omega \to \mathbb{R}$ with characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ given by (17.1). Fix θ in $\mathbb{R}$. It follows from Theorem 17.7.1 that

$$
e^{i\theta X} - \sum_{\ell=0}^{k} \frac{1}{\ell!} (i\theta X)^\ell = R_k(X;\theta) \qquad (17.34)
$$

Therefore, if the rv X has a finite moment of order k for some $k = 1, 2, \ldots$, the expectation

$$
\mathbb{E}\left[R_k(X;\theta)\right]
$$

exists and is well defined since all the moments of X of order $\ell = 1, 2, \ldots, k$ exist and are finite. Thus, the relationship

$$
\mathbb{E}\left[e^{i\theta X}\right] = \sum_{\ell=0}^{k} \frac{1}{\ell!} (i\theta)^\ell \, \mathbb{E}\left[X^\ell\right] + \mathbb{E}\left[R_k(X;\theta)\right] \qquad (17.35)
$$

does hold. This suggests the following result.

Theorem 17.8.1 *Consider a rv $X : \Omega \to \mathbb{R}$ with characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ given by (17.1). If $\mathbb{E}\left[|X|^n\right] < \infty$ for some $n = 1, 2, \ldots$, then for each $k = 1, 2, \ldots, n$, the characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ is everywhere k^{th} differentiable with*

$$
\frac{d^k}{d\theta^k} \Phi_X(\theta) = \mathbb{E}\left[(iX)^k \, e^{i\theta X}\right], \quad \theta \in \mathbb{R}. \qquad (17.36)
$$

In particular,

$$
\left. \frac{d^k}{d\theta^k} \Phi_X(\theta) \right|_{\theta=0} = i^k \mathbb{E}\left[X^k\right]. \qquad (17.37)
$$

Proof. If $k = 1$. Fix θ in $\mathbb{R}$ and for each $h \neq 0$ note that

$$\begin{aligned} \Phi_X(\theta + h) - \Phi_X(\theta) &= \mathbb{E}\left[e^{i\theta X}\left(e^{ihX} - 1\right)\right] \\ &= \mathbb{E}\left[e^{i\theta X}\int_0^X ihe^{iht}dt\right] \quad (17.38) \end{aligned}$$

so that

$$\frac{1}{h}\left(\Phi_X(\theta + h) - \Phi_X(\theta)\right) = \mathbb{E}\left[e^{i\theta X}\int_0^X ie^{iht}dt\right].$$

The bound

$$\left|e^{i\theta X}\int_0^X ie^{iht}dt\right| = \left|e^{i\theta X}\right|\left|\int_0^X ie^{iht}dt\right| \leq |X| \quad (17.39)$$

holds *uniformly* in $h \neq 0$, whence

$$\lim_{h\to 0}\left(e^{i\theta X}\int_0^X ie^{iht}dt\right) = (iX)\,e^{i\theta X}$$

by the Bounded Convergence Theorem. We now conclude that

$$\begin{aligned} \lim_{h\to 0}\frac{1}{h}\left(\Phi_X(\theta + h) - \Phi_X(\theta)\right) &= \lim_{h\to 0}\mathbb{E}\left[e^{i\theta X}\int_0^X ie^{iht}dt\right]. \\ &= \mathbb{E}\left[\lim_{h\to 0}\left(e^{i\theta X}\int_0^X ie^{iht}dt\right)\right] \\ &= \mathbb{E}\left[(iX)\,e^{i\theta X}\right] \quad (17.40) \end{aligned}$$

by the Dominated Convergence Theorem and the conclusion (**??**) holds for $k = 1$.

If $k \geq 2$, we proceed by induction: The basis step was just established. To establish the induction step, assume that for each $\ell = 1, \ldots, k-1$, the characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ is everywhere ℓ^{th} differentiable with

$$\frac{d^\ell}{d\theta^\ell}\Phi_X(\theta) = \mathbb{E}\left[(iX)^\ell\,e^{i\theta X}\right], \quad \theta \in \mathbb{R}. \quad (17.41)$$

Under the assumption $\mathbb{E}\left[|X|^k\right] < \infty$, we shall now show that the characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ is everywhere $(\ell + 1)^{rst}$ differentiable with

$$\frac{d^{\ell+1}}{d\theta^{\ell+1}}\Phi_X(\theta) = \mathbb{E}\left[(iX)^{\ell+1}\,e^{i\theta X}\right], \quad \theta \in \mathbb{R}. \quad (17.42)$$

Indeed, for every $h \neq 0$, we have

$$\begin{aligned}\frac{d^\ell}{d\theta^\ell}\Phi_X(\theta+h) - \frac{d^\ell}{d\theta^\ell}\Phi_X(\theta) &= \mathbb{E}\left[(iX)^\ell \left(e^{i(\theta+h)X} - e^{i\theta X}\right)\right] \\ &= \mathbb{E}\left[(iX)^\ell e^{i\theta X}\left(e^{ihX} - 1\right)\right] \\ &= \mathbb{E}\left[(iX)^\ell e^{i\theta X}\int_0^X ihe^{iht}dt\right]\end{aligned}$$

so that

$$\frac{1}{h}\left(\frac{d^\ell}{d\theta^\ell}\Phi_X(\theta+h) - \frac{d^\ell}{d\theta^\ell}\Phi_X(\theta)\right) = \mathbb{E}\left[(iX)^\ell e^{i\theta X}\int_0^X ie^{iht}dt\right]$$

Again we see that

$$\left|(iX)^\ell e^{i\theta X}\int_0^X ie^{iht}dt\right| \leq |X|^{\ell+1}$$

uniformly in $h \neq 0$ with $\mathbb{E}\left[|X|^{\ell+1}\right] < \infty$ by assumption. Invoking the Dominated Convergence Theorem we conclude that

$$\begin{aligned}&\lim_{h\to 0}\frac{1}{h}\left(\frac{d^\ell}{d\theta^\ell}\Phi_X(\theta+h) - \frac{d^\ell}{d\theta^\ell}\Phi_X(\theta)\right) \\ &= \lim_{h\to 0}\mathbb{E}\left[(iX)^\ell e^{i\theta X}\int_0^X ie^{iht}dt\right] \\ &= \mathbb{E}\left[(iX)^\ell e^{i\theta X}\lim_{h\to 0}\int_0^X ie^{iht}dt\right] \\ &= \mathbb{E}\left[(iX)^{\ell+1} e^{i\theta X}\right], \qquad (17.43)\end{aligned}$$

and this establishes (17.42) holds. This concludes the induction step as we have now shown that (17.41) holds for $\ell = 1, \ldots, k$. ■

17.9 Moment generating functions

Definition 17.9.1

With any rv $X : \Omega \to \mathbb{R}^p$, we associate its *moment generating function* (MGF) $M_X : \mathbb{R}^p \to \mathbb{R}$ given by

$$M_X(\theta) \equiv \mathbb{E}\left[e^{\theta^t X}\right], \quad \theta \in \mathbb{R}^p. \qquad (17.44)$$

While the moment generating function of any rv X is always well defined – After all $e^{\theta^t X} \geq 0$ for all $\theta \in \mathbb{R}^p$, it may not be finite. In fact it is not too difficult to find examples for which $M_X(\theta) = \infty$ for all θ in $\mathbb{R}^p$ except $\theta = 0_p$ (in which case $M_X(\theta) = 1$. This limits the use of moment generating functions

17.10 Laplace transforms

Definition 17.10.1

With any rv $X : \Omega \to \mathbb{R}^p_+$, we associate its *moment generating function* (MGF) $M_X : \mathbb{R}^p \to \mathbb{R}$ given by

$$L_X(s) \equiv \mathbb{E}\left[e^{-\theta^t X}\right], \quad s \in \mathbb{R}^p. \tag{17.45}$$

17.11 Probability generating functions

Definition 17.11.1

With any rv $X : \Omega \to \mathbb{N}$, we associate its *probability generating function* (PGF) $G_X : \mathbb{R} \to \mathbb{R}$ given by

$$G_X(z) \equiv \mathbb{E}\left[z^X\right], \quad z \in \mathbb{R}. \tag{17.46}$$

17.12 Exercises

Ex. 17.1 If the rv $X : \Omega \to \mathbb{R}^p$ is symmetric, then its characteristic function Φ_X is real-valued, i.e., $\Phi_X(\theta)$ is an element of $\mathbb{R}$ for every θ in $\mathbb{R}^p$.

Ex. 17.2 With $a > 0$ and $\nu \geq 0$, consider the function $\Phi_{a,\nu} : \mathbb{R} \to \mathbb{R}$ given by

$$\Phi_a(\theta) \equiv e^{-a|\theta|^{1+\nu}}, \quad \theta \in \mathbb{R}.$$

Determine whether the function $\Phi_{a,\nu} : \mathbb{R} \to \mathbb{R}$ is the characteristic function associated with a probability distribution $F_{a,\nu} : \mathbb{R} \to [0,1]$?

Ex. 17.3

Ex. 17.4

Chapter 18

Gaussian random variables

This chapter is devoted to a brief discussion of the class of Gaussian rvs. In particular, for easy reference we have collected various facts and properties to be used repeatedly, here and in many applications.

18.1 Scalar Gaussian rvs

Definition 18.1.1

With scalars μ (in $\mathbb{R}$) and $\sigma \geq 0$, a rv $X : \Omega \to \mathbb{R}$ (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$) is a Gaussian (or normally distributed) rv with parameters μ and σ^2 if *either* $\sigma = 0$ and X is a degenerate rv with $X = \mu$ a.s., *or* $\sigma > 0$ and the probability distribution of X is of the form

$$\mathbb{P}\left[X \leq x\right] = \int_{-\infty}^{x} f_{\mu,\sigma^2}(t)dt, \quad x \in \mathbb{R}$$

where

$$f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \quad t \in \mathbb{R}.$$

We leave it as a simple exercise to check that

$$(18.1) \qquad \int_{\mathbb{R}} t f_{\mu,\sigma^2}(t)dt = \mu \quad \text{and} \quad \int_{\mathbb{R}} t f_{\mu,\sigma^2}(t)dt = \mu^2 + \sigma^2$$

so that $\mathbb{E}\left[X\right] = \mu$ and $\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$, hence $\text{Var}[X] = \sigma^2$. This shows that the parameters μ and σ^2 are the mean and variance, respectively, of the rv X. In fact, a Gaussian rv is completely characterized by its first and second moments – As a result it is customary to refer to the rv X in Definition 18.1.1 as a Gaussian rv with mean μ and variance σ^2, written $X \sim \mathrm{N}(\mu, \sigma^2)$.

The characteristic function of Gaussian rvs takes a very simple form.

Fact 18.1.1 *If the rv $X : \Omega \to \mathbb{R}$ (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$) is a Gaussian rv with mean μ (in $\mathbb{R}$) and variance $\sigma^2 \geq 0$. its characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ is given by*

$$\Phi_X(\theta) = \mathbb{E}\left[e^{i\theta X}\right] = e^{i\theta\mu - \frac{\sigma^2}{2}\cdot\theta^2}, \quad \theta \in \mathbb{R}. \tag{18.2}$$

This fact is established in Section 18.13. and allows us to give a definition which is equivalent to Definition 18.1.1 and which covers both cases.

Definition 18.1.2

A rv $X : \Omega \to \mathbb{R}$ (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$) is a Gaussian rv with mean μ (in $\mathbb{R}$) and variance $\sigma^2 \geq 0$ if its characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ is given by

$$\Phi_X(\theta) = e^{i\theta\mu - \frac{\sigma^2}{2}\cdot\theta^2}, \quad \theta \in \mathbb{R}. \tag{18.3}$$

The relations (18.1) can also be established by differentiating the expression (18.3) and using Theorem 17.8.1. It is a simple matter to check that if X is normally distributed with mean μ and variance σ^2, then for scalars a and b, the rv $aX + b$ is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular, with $\sigma > 0$, the rv $\sigma^{-1}(X - \mu)$ is a Gaussian rv with mean zero and unit variance.

18.2 The standard Gaussian rv

The Gaussian rv with mean zero ($\mu = 0$) and unit variance ($\sigma^2 = 1$) is known as the *standard* Gaussian rv, and occupies a very special place among Gaussian rvs. Throughout, we denote by U a Gaussian rv with zero mean and unit variance (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$). Its probability distribution function is given by

$$\mathbb{P}\left[U \leq x\right] = \Phi(x) \equiv \int_{-\infty}^{x} \phi(t)dt, \quad x \in \mathbb{R} \tag{18.4}$$

with probability density function $\phi : \mathbb{R} \to \mathbb{R}_+$ given by

$$\phi(t) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad t \in \mathbb{R}. \tag{18.5}$$

As should be clear from earlier comments, for any Gaussian rv X with mean μ and variance σ^2, it holds that $X =_{st} \mu + \sigma U$, so that

$$\begin{aligned} \mathbb{P}[X \leq x] &= \mathbb{P}\left[\sigma^{-1}(X-\mu) \leq \sigma^{-1}(x-\mu)\right] \\ &= \mathbb{P}\left[U \leq \sigma^{-1}(x-\mu)\right] \\ &= \Phi(\sigma^{-1}(x-\mu)), \quad x \in \mathbb{R}. \end{aligned}$$

The evaluation of probabilities involving Gaussian rvs thus reduces to the evaluation of related probabilities for the standard Gaussian rv. It also follows readily by differentiation of (18.6) that

$$f_{\mu,\sigma^2}(x) = \sigma^{-1}\phi(\sigma^{-1}(x-\mu)), \quad x \in \mathbb{R}$$

as expected.

The standard Gaussian rv U is a symmetric rv: Indeed, for each x in $\mathbb{R}$, the symmetry of the probability density function $\phi : \mathbb{R} \to \mathbb{R}_+$ readily implies $\mathbb{P}[U \leq -x] = \mathbb{P}[U > x]$, so that $\Phi(-x) = 1 - \Phi(x)$, and Φ is therefore fully determined by the complementary probability distribution function of U on $[0, \infty)$, namely

$$Q(x) \equiv 1 - \Phi(x) = \mathbb{P}[U > x], \quad x \geq 0. \tag{18.6}$$

The evaluation of the so-called Q-function is given in Section 18.10 together with some of its properties (which are often used in Communication Theory).

In Section 18.12 we evaluate the moments of the standard zero-mean unit Gaussian rv U.

Fact 18.2.1 *If U is a standard zero-mean unit variance Gaussian rv, then its moments are all finite and given by*

$$m_k \equiv \mathbb{E}\left[U^k\right] = \begin{cases} 0 & \text{if } k = 2\ell+1 \text{ with } \ell = 0, 1, \ldots \\ \frac{(2\ell)!}{2^\ell \ell!} & \text{if } k = 2\ell \text{ with } \ell = 1, 2 \ldots \end{cases} \tag{18.7}$$

18.3 A little Linear Algebra

Before introducing the notion of a multi-dimensional Gaussian rv, we present some standard facts from Linear Algebra that are needed in developing the appropriate definition. Throughout p is a positive integer, and unless specified otherwise, elements of $\mathbb{R}^p$ are understood as column vectors. If u is an element in $\mathbb{R}^p$, then its k^{th} component is denoted by u_k for $k = 1, \ldots, p$, and $u = (u_1, \ldots, u_p)^t$ with the superscript ^{t} denoting transposition.

Definition 18.3.1

A square $p \times p$ matrix R is said to be

(i) symmetric if $R^t = R$, namely

$$R_{k\ell} = R_{\ell k}, \quad k, \ell = 1, \ldots, p.$$

(ii) positive semi-definite if

$$u^t R u \geq 0, \quad u \in \mathbb{R}^p$$

(iii) positive definite if it is positive semi-definite and the condition $u^t R u = 0$ implies $u = (0, \ldots, 0)^t$.

The facts given next concern the eigenvalues and eigenvectors of symmetric matrices, and are well known:

Theorem 18.3.1 *Let R denote a symmetric $p \times p$ matrix. It has p eigenvalues, not necessarily distinct, all of which are real, say $\lambda_1, \ldots, \lambda_p$. Moreover, there exists vectors $u_1, \ldots, u_p$ in $\mathbb{R}^p$ with the following properties:*

(i) The vectors $u_1, \ldots, u_p$ form an orthonormal family in the sense that

$$u_k^t u_\ell = \delta(k, \ell), \quad k, \ell = 1, \ldots, p.$$

(ii) For each $k = 1, \ldots, p$, the vector u_k is an eigenvector for the eigenvalue λ_k in that

$$R u_k = \lambda_k u_k.$$

(iii) If in addition, the matrix R is positive semi-definite, then $\lambda_k \geq 0$ for each $k = 1, \ldots, p$.

The following calculations are standard: It is customary to introduce the $p \times p$ matrix T formed by taking its columns to be the eigenvectors $u_1, \ldots, u_p$, namely

$$T \equiv \begin{pmatrix} u_1 & u_2 & \ldots & u_p \end{pmatrix}.$$

The transpose T^t of T is given by

$$T^t = \begin{pmatrix} u_1^t \\ u_2^t \\ \vdots \\ u_p^t \end{pmatrix}.$$

From Theorem 18.3.1 we conclude that

$$RT = \begin{pmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_p u_p \end{pmatrix}$$

and

$$\begin{aligned} T^t RT &= \begin{pmatrix} u_1^t \\ u_2^t \\ \vdots \\ u_p^t \end{pmatrix} \begin{pmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_p u_p \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 u_1^t u_1 & \lambda_2 u_1^t u_2 & \dots & \lambda_p u_1^t u_p \\ \lambda_1 u_2^t u_1 & \lambda_2 u_2^t u_2 & \dots & \lambda_p u_2^t u_p \\ \vdots & & & \\ \lambda_1 u_p^t u_1 & \lambda_2 u_p^t u_2 & \dots & \lambda_p u_1^p u_p \end{pmatrix} \\ &= \mathrm{Diag}(\lambda_1, \lambda_2, \dots, \lambda_p) \qquad (18.8) \end{aligned}$$

where $\mathrm{Diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ is the diagonal matrix whose diagonal elements are $\lambda_1, \lambda_2, \dots, \lambda_p$. A similar line of reasoning also shows that

$$T^t T = \begin{pmatrix} u_1^t \\ u_2^t \\ \vdots \\ u_p^t \end{pmatrix} \begin{pmatrix} u_1 & u_2 & \dots & u_p \end{pmatrix} = I_p$$

where I_p denotes the p-dimensional unite matrix. By the uniqueness of the inverse of a matrix, we conclude that T is invertible with $T^{-1} = T^t$. Since $TT^{-1} = T^{-1}T = I_p$ it follows that

$$TT^t = T^t T = I_p.$$

The relation $T^t RT = \mathrm{Diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ yields

$$R = T\left(T^t RT\right)T^t = T\left(\mathrm{Diag}(\lambda_1, \lambda_2, \dots, \lambda_p)\right)T^t.$$

If in addition to being a symmetric matrix, R was also positive semi-definite, then its eigenvalues are now non-negative and we can write

$$\begin{aligned} R &= \left(T\mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p})\right) \cdot \left(\mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p})\right) T^t \\ &= \left(T\mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p})\right) \cdot \left(T\mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_p})\right)^t. \end{aligned}$$

The $p \times p$ matrix

$$B \equiv T\mathrm{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2} \dots, \sqrt{\lambda_p})$$

has the property that $R = BB^t$, and is known as the *square root* of the positive semi-definite symmetric matrix R.

18.4 Gaussian random vectors

There are several equivalent ways to define multi-dimensional Gausssian rvs. Throughout, let μ denote a vector in $\mathbb{R}^p$ and let Σ be a $p \times p$ symmetric and positive semi-definite matrix, thus $\Sigma^t = \Sigma$ and $\theta^t \Sigma \theta \geq 0$ for all θ in $\mathbb{R}^p$.

A definition via characteristic functions The most convenient definition is given in terms of characteristic functions.

Definition 18.4.1

An $\mathbb{R}^p$-valued rv X (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$) is said to be a Gaussian rv (or a p-dimensional Gaussian random vector) with mean vector μ and covariance matrix Σ if its characteristic function is given by

$$\mathbb{E}\left[e^{i\theta^t X}\right] = e^{i\theta^t \mu - \frac{1}{2}\theta^t \Sigma \theta}, \quad \theta \in \mathbb{R}^p. \tag{18.9}$$

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

For the right-handside of (18.9) to be a characteristic function we must have

$$\left|\mathbb{E}\left[e^{i\theta^t X}\right]\right| = e^{-\frac{1}{2}\theta^t \Sigma \theta} \leq 1, \quad \theta \in \mathbb{R}^p.$$

This implies $\theta^t \Sigma \theta \geq 0$ for each θ in $\mathbb{R}^p$, making it necessary for the $p \times p$ matrix Σ to be a positive semi-definite matrix.

Next, fix θ in $\mathbb{R}^p$ and use (18.9) with $a\theta$ where a ranges in $\mathbb{R}$. It follows that

$$\mathbb{E}\left[e^{ia\theta^t X}\right] = e^{ia\theta^t \mu - \frac{a^2}{2}\theta^t \Sigma \theta}, \quad a \in \mathbb{R} \tag{18.10}$$

and by virtue of Definition 18.1.2 we conclude that the *scalar* rv $\theta^t X$ is a Gaussian rv with mean $\theta^t \mu$ and variance $\theta^t \Sigma \theta$. But $\theta^t \mu = \mathbb{E}\left[\theta^t X\right] = \theta^t \mathbb{E}\left[X\right]$ and $\theta^t \Sigma \theta = \mathrm{Var}\left[\theta^t X\right] = \theta^t \mathrm{Cov}\left[X\right] \theta$. As these equalities hold for *all* θ in $\mathbb{R}^p$ we conclude that $\mu = \mathbb{E}\left[X\right]$ and $\Sigma = \mathrm{Cov}\left[X\right]$. In other words, the parameters μ and Σ indeed have the interpretation of mean and covariance for the rv X. The latter conclusion also shows that the matrix Σ appearing in (18.9) is necessarily symmetric.

A constructive definition We now present another definition of multi-dimensional Gaussian rvs. This definition is not give in terms of characteristic functions, but instead uses only the existence of standard (one-dimensional) Gaussian rvs.

Definition 18.4.2

An $\mathbb{R}^p$-valued rv X (defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$) is said to be a Gaussian rv (or a p-dimensional Gaussian random vector) if for some positive integer d, there exists an element b in $\mathbb{R}^p$, a $p \times d$ matrix B and i.i.d. standard Gaussian rvs $U_1, \dots, U_d$ (defined on $(\Omega, \mathcal{F}, \mathbb{P})$) such that

$$X =_{st} b + B \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}. \tag{18.11}$$

By linearity of expectations it is plain from (18.11) that

$$\mathbb{E}[X] = \mathbb{E}\left[b + B \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}\right] = b + B \begin{pmatrix} \mathbb{E}[U_1] \\ \vdots \\ \mathbb{E}[U_d] \end{pmatrix} = b$$

and

$$\begin{aligned} \mathbb{E}\left[(X-b)(X-b)^t\right] &= \mathbb{E}\left[B \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix} \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}^t B^t\right] \\ &= B\mathbb{E}\left[\begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix} \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}^t\right] B^t \\ &= BI_dB^t \\ &= BB^t. \end{aligned} \tag{18.12}$$

In short we have shown that if the rv $X : \Omega \to \mathbb{R}^p$ is Gaussian according to Definition 18.4.2, then

$$\mathbb{E}[X] = b \quad \text{and} \quad \text{Cov}[X] = BB^t.$$

18.5 Equivalence of the two definitions

We now discuss the equivalence of these two definitions.

Definition 18.4.2 implies Definition 18.4.1 Next, pick θ in $\mathbb{R}^p$. We note that

$$\theta^t (X-b) =_{st} \theta^t B \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix} = (B^t\theta)^t \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix} = \sum_{k=1}^d \left(B^t\theta\right)_k U_k$$

where for each $k = 1, \ldots, d$, $\left(B^t\theta\right)_k$ denotes the k^{th} component of the vector $B^t\theta$ in $\mathbb{R}^d$. It follows that

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t(X-b)}\right] &= \mathbb{E}\left[e^{i\sum_{k=1}^d \left(B^t\theta\right)_k U_k}\right] \\ &= \mathbb{E}\left[\prod_{k=1}^d e^{i\left(B^t\theta\right)_k U_k}\right] \\ &= \prod_{k=1}^d \mathbb{E}\left[e^{i\left(B^t\theta\right)_k U_k}\right] \end{aligned}$$

by the mutual independence of the rvs $U_1, \ldots, U_d$ with

$$\mathbb{E}\left[e^{i\left(B^t\theta\right)_k U_k}\right] = e^{-\frac{1}{2}|\left(B^t\theta\right)_k|^2}$$

upon using the fact that $U_k \sim \mathrm{N}(0,1)$ for each $k = 1, \ldots, d$. Collecting terms we conclude that

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t(X-b)}\right] &= \prod_{k=1}^d e^{-\frac{1}{2}|\left(B^t\theta\right)_k|^2} \\ &= e^{-\frac{1}{2}\sum_{k=1}^d |\left(B^t\theta\right)_k|^2} \\ &= e^{-\frac{1}{2}\theta^t BB^t\theta} \qquad (18.13) \end{aligned}$$

as we note that

$$\sum_{k=1}^d | \left(B^t\theta\right)_k |^2 = \theta^t BB^t\theta.$$

We conclude that

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t X}\right] &= e^{i\theta^t b}\mathbb{E}\left[e^{i\theta^t(X-b)}\right] \\ &= e^{i\theta^t b} e^{-\frac{1}{2}\theta^t BB^t\theta}, \quad \theta \in \mathbb{R}^p \qquad (18.14) \end{aligned}$$

and $X \sim \mathrm{N}(b, BB^t)$ according to Definition 18.4.1, On the basis of this discussion the reader might wonder whether *any* $p \times p$ matrix Σ which is both positive semi-definite and symmetric can be realized in this manner, namely as $\Sigma = BB^t$ for some $d \times p$ matrix B. The answer is obviously in view of the discussion of Section 18.3: There we showed that for any $p \times p$ matrix Σ which is symmetric and positive semi-definite, there always exists a $p \times p$ matrix B such that $\Sigma = BB^t$ – Its square root! This also shows that although the pair (d, B) may not be unique, there is always one with smallest dimension, namely $d = p$ in which case B is taken to be the square-root of the target covariance Σ. ■

Definition 18.4.1 implies Definition 18.4.2 Consider a rv $X : \Omega \to \mathbb{R}^p$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ which is a Gaussian rv with mean vector μ and covariance matrix Σ according to Definition Definition 18.4.1. The matrix Σ being symmetric and positive semi-definite, there exists a $p \times p$ matrix B such that $\Sigma = BB^t$. Consider the rv $X^\star : \Omega \to \mathbb{R}^p$ given by

$$X^\star \equiv \mu + B \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix}$$

where $U_1, \ldots, U_p$ are i.i.d. standard Gaussian rvs. As shown earlier in this section, $\mathbb{E}[X^\star] = \mu$ and $\text{Cov}[X^\star] = BB^t = \Sigma$, while

$$\mathbb{E}\left[e^{i\theta^t X^\star}\right] = e^{i\theta^t\mu - \frac{1}{2}\theta^t\Sigma\theta}, \quad \theta \in \mathbb{R}^p$$

Therefore, $X^\star$ and X have identical characteristic functions, hence they have the same probability distribution functions and we can write $X =_{st} X^\star$, just another way to say that X is Gaussian according to Definition 18.4.2. ■

18.6 Existence of a density

In general, an $\mathbb{R}^p$-valued Gaussian rv as defined above may not admit a density function: To see why, consider a Gaussian rv $X : \Omega \to \mathbb{R}^p$ with mean vector μ and covariance matrix Σ. The *kernel* $\text{Ker}(\Sigma)$ of its covariance matrix Σ, also known as its *null space*, is the linear subspace of $\mathbb{R}^p$ given by

$$\text{Ker}(\Sigma) \equiv \{x \in \mathbb{R}^p : \Sigma x = 0_p\}.$$

Observe that $\theta^t\Sigma\theta = 0$ if and only if θ belongs to $\mathrm{Ker}(\Sigma)$, in which case (18.9) yields

$$\mathbb{E}\left[e^{i\theta^t(X-\mu)}\right] = 1$$

and we conclude that

$$\theta^t(X-\mu) = 0 \quad \text{a.s.}$$

In other words, with probability one, the rv $X-\mu$ is orthogonal to the linear space $\mathrm{Ker}(\Sigma)$.

To proceed, assume that the covariance matrix Σ is not trivial (in that it has some non-zero entries) for otherwise $X = \mu$ a.s. In the non-trivial case, there are now two possibilities depending on whether the $p \times p$ matrix Σ is positive definite or not. Note that the positive definiteness of Σ, i.e., $\theta^t\Sigma\theta = 0$ necessarily implies $\theta = \mathbf{0}_d$, is equivalent to the condition $\mathrm{Ker}(\Sigma) = 0_p$.

If the $p\times p$ matrix Σ is not positive definite, namely only positive semi-definite, then the mass of the rv $X-\mu$ is concentrated on the orthogonal space $\mathrm{Ker}(\Sigma)^\perp$ of $\mathrm{Ker}(\Sigma)$, and the distribution of X has its support on the linear manifold $\mu + \mathrm{Ker}(\Sigma)^\perp$ and must be singular with respect to Lebesgue measure – The probability distribution function of the gausssian rv X does not admit a probability density function.

On the other hand, if the $p\times p$ matrix Σ is positive definite, then the matrix Σ is invertible, $\det(\Sigma) \neq 0$ and the Gaussian rv X with mean vector μ and covariance matrix Σ admits a probability density function $f : \mathbb{R}^p \to \mathbb{R}_+$ given by

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^p.$$

18.7 Linear transformations

The following result is very useful in many contexts, and shows that linear transformations preserve the Gaussian character:

Lemma 18.7.1 *let ν be an element of $\mathbb{R}^q$ and let A be an $q \times p$ matrix. Then, for any Gaussian rv $\mathbb{R}^p$-valued rv X with mean vector μ and covariance matrix Σ, the $\mathbb{R}^q$-valued rv Y given by*

$$Y = \nu + AX$$

is also a Gaussian rv with mean vector $\nu + A\mu$ and covariance matrix $A\Sigma A^t$.

Proof. First, by linearity we note that

$$\mathbb{E}[Y] = \mathbb{E}[\nu + AX] = \nu + A\mu$$

so that

$$\begin{aligned}\mathrm{Cov}[Y] &= \mathbb{E}\left[A(X-\mu)\left(A(X-\mu)\right)^t\right] \\ &= A\mathbb{E}\left[(X-\mu)(X-\mu)^t\right]A^t \\ &= A\Sigma A^t. \qquad (18.15)\end{aligned}$$

Consequently, the $\mathbb{R}^q$-valued rv Y has mean vector $\nu + A\mu$ and covariance matrix $A\Sigma A^t$.

Pick α arbitrary in $\mathbb{R}^q$. We have

$$\begin{aligned}\mathbb{E}\left[e^{i\alpha^t Y}\right] &= \mathbb{E}\left[e^{i\alpha^t(\nu+AX)}\right] \\ &= e^{i\alpha^t\nu}\mathbb{E}\left[e^{i\alpha^t AX}\right] \\ &= e^{i\alpha^t\nu}\mathbb{E}\left[e^{i\left(A^t\alpha\right)^t X}\right] \\ &= e^{i\alpha^t\nu}e^{-\frac{1}{2}\left(A^t\alpha\right)^t\Sigma\left(A^t\alpha\right)} \\ &= e^{i\alpha^t\nu}e^{-\frac{1}{2}\alpha^t A\Sigma A^t\alpha} \qquad (18.16)\end{aligned}$$

as required. ■

This result can also be established through the evaluation of the characteristic function of the rv Y. As an immediate consequence of Lemma 18.7.1 we get the following fact whose proof is left as an easy exercise.

Corollary 18.7.1 *Consider a Gaussian rv $\mathbb{R}^p$-valued rv X with mean vector μ and covariance matrix Σ. For any subset I of $\{1,\ldots,d\}$ with $|I| = q \leq d$, the $\mathbb{R}^q$-valued rv X_I given by $X_I = (X_i,\ i \in I)^t$ is a Gaussian rv with mean vector $(\mu_i,\ i \in I)^t$ and covariance matrix $(\Sigma_{ij},\ i,j \in I)$.*

18.8 Independence of Gaussian rvs

Characterizing the mutual independence of Gaussian rvs turns out to be quite straightforward: Consider the second-order rvs $X_1,\ldots,X_k$, all defined on the same probability triple $(\Omega,\mathcal{F},\mathbb{P})$, where for each $\ell = 1,\ldots,k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ has mean vector μ_ℓ and covariance matrix Σ_ℓ. With $p = p_1 + \ldots + p_r$, let X denote the $\mathbb{R}^p$-valued rv obtained by concatenating $X_1,\ldots,X_k$, namely

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}. \qquad (18.17)$$

Its mean vector μ is simply

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \tag{18.18}$$

while its covariance matrix Σ can be written in block form as

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} & \dots & \Sigma_{1,k} \\ \Sigma_{2,1} & \Sigma_2 & \dots & \Sigma_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{k,1} & \Sigma_{k,2} & \dots & \Sigma_k \end{pmatrix} \tag{18.19}$$

with the notation

$$\Sigma_{i,j} \equiv \mathrm{Cov}[X_i, X_j] \quad i, j = 1, \dots, k.$$

Lemma 18.8.1 *With the notation above, assume the rv $X : \Omega \to \mathbb{R}^p$ to be a Gaussian rv with mean vector μ and covariance matrix Σ. Then, for each $\ell = 1, \dots, k$, the rv X_ℓ is a Gaussian rv with mean vector μ_ℓ and covariance matrix Σ_ℓ. Moreover, the rvs $X_1, \dots, X_k$ are mutually independent Gaussian rvs if and only they are uncorrelated, i.e.,*

$$\Sigma_{i,j} = \delta(i,j)\Sigma_j, \quad i, j = 1, \dots, k. \tag{18.20}$$

The first part of Lemma 18.8.1 is a simple rewrite of Corollary 18.7.1. Sometimes we refer to the fact that the rv X is Gaussian by saying that the rvs $X_1, \dots, X_r$ are *jointly* Gaussian. A converse to Lemma 18.8.1 is available:

Lemma 18.8.2 *Assume that for each $\ell = 1, \dots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a Gaussian rv with mean vector μ_ℓ and covariance matrix Σ_ℓ. If the rvs $X_1, \dots, X_k$ are mutually independent, then the rv $X : \Omega \to \mathbb{R}^p$ is a Gaussian rv with mean vector μ and covariance matrix Σ as given by (18.19) with (18.20).*

It might be tempting to conclude that the Gaussian character of *each* of the rvs $X_1, \dots, X_k$ *alone* suffices to imply the Gaussian character of the combined rv X. However, it can be shown through simple counterexamples that this is not so. In other words, the joint Gaussian character of X does not follow merely from that of its components $X_1, \dots, X_k$ *without* further assumptions. A counterexample is given in Exercise 18.13.

18.9 Conditional distributions

Consider the following situation: The rv $Z : \Omega \to \mathbb{R}^{p+q}$ is defined on some probability triple and is of the form

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}$$

with component rvs $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$.

Lemma 18.9.1 *There always exists a $p \times q$ matrix $A^\star$ such that the rvs $V = X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])$ and Y are uncorelated. This matrix is any solution of the matrix equation*

$$\text{Cov}[X, Y] = A\text{Cov}[Y], \quad p \times q \textit{ matrix } A. \tag{18.21}$$

When $\text{Cov}[Y]$ *is invertible, then the matrix A is unique and is given by $A^\star = \text{Cov}[X, Y] \text{Cov}[Y]^{-1}$.*

Proof. For any $p \times q$ matrix A, define the rv $V_A : \Omega \to \mathbb{R}^p$ by

$$V_A \equiv X - \mathbb{E}[X] - A(Y - \mathbb{E}[Y]).$$

Note that

$$\begin{aligned} \text{Cov}[V_A, Y] &= \text{Cov}[X - \mathbb{E}[X] - A(Y - \mathbb{E}[Y]), Y] \\ &= \text{Cov}[X - \mathbb{E}[X], Y] - \text{Cov}[A(Y - \mathbb{E}[Y]), Y] \\ &= \text{Cov}[X, Y] - A\text{Cov}[Y]. \end{aligned} \tag{18.22}$$

The condition that the rvs V_A and Y are uncorrelated reads $\text{Cov}[V_A, Y] = O_{p\times q}$, or equivalently, (18.24). If $\text{Cov}[Y]$ is invertible, then clearly there is only one solution to this matrix equation (in A), and it is given by $A^\star = \text{Cov}[X, Y] \text{Cov}[Y]^{-1}$. ■

If $\text{Cov}[Y]$ is not invertible, then

Some important consequences can be derived from Lemma 18.9.1, and are given next.

Lemma 18.9.2 *Assume the rv $Z : \Omega \to \mathbb{R}^{p+q}$ to be a Gaussian rv. With the notation of Lemma 18.9.1, the rvs $V = X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])$ and Y are independent, each of which is Gaussian.*

Proof. Now consider the rv $W : \Omega \to \mathbb{R}^{p+q}$ given by

$$W \equiv \begin{pmatrix} X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y]) \\ Y \end{pmatrix}.$$

We can rewrite W in a more compact form as

$$W = B \begin{pmatrix} X \\ Y \end{pmatrix} - b = BZ - b$$

where the $(p+q) \times (p+q)$ matrix B and the element b of $\mathbb{R}^{p+q}$ are given by

$$B \equiv \begin{pmatrix} I_p & -A^\star \\ O_{q\times p} & I_q \end{pmatrix} \quad \text{and} \quad b \equiv \begin{pmatrix} \mathbb{E}[X] - A^\star \mathbb{E}[Y] \\ 0_q \end{pmatrix}$$

The Gaussian character of the rv Z implies that the rv $W : \Omega \to \mathbb{R}^{p+q}$ is also Gaussian by Lemma 18.7.1. Therefore, the rvs $V = X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])$ and Y being uncorrelated, we can use Lemma 18.8.1 to conclude that they are independent! ■

We shall use these facts and basic properties of conditional expectations to evaluate the conditional expectation of the rv X given Y.

Lemma 18.9.3 *Assume the rv $Z : \Omega \to \mathbb{R}^{p+q}$ to be a Gaussian rv. It holds that*

$$\mathbb{E}[X|Y] = \mathbb{E}[X] + A^\star(Y - \mathbb{E}[X|Y]) \quad a.s. \tag{18.23}$$

where the $p \times q$ matrix $A^\star$ is any solution of the matrix equation (18.24).

Proof. First, by the independence established in Lemma 18.9.2 we have

$$\begin{aligned} &\mathbb{E}[X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])|Y] \\ &= \mathbb{E}[X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])] = (0, \ldots, 0)^t \quad \text{a.s.} \end{aligned}$$

On the other hand, by linearity of conditional expectations we get

$$\begin{aligned} &\mathbb{E}[X - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y])|Y] \\ &= \mathbb{E}[X|Y] - \mathbb{E}[X] - A^\star(Y - \mathbb{E}[Y]) \quad \text{a.s.} \end{aligned}$$

Combining these two evaluations we conclude to (18.23). ■

Finally, we are in a position to identify the conditional distribution of the rv X given Y.

Proposition 18.9.1 *Assume the rv $Z : \Omega \to \mathbb{R}^{p+q}$ to be a Gaussian rv. It holds that*

$$\begin{aligned} & \mathbb{E}\left[e^{i\theta^t X}\middle|Y\right] \\ = \quad & e^{i\theta^t \mathbb{E}[X|Y]} \cdot e^{-\frac{1}{2}\theta^t(\text{Cov}[X]-A^\star \text{Cov}[Y,X])\theta}, \quad \theta \in \mathbb{R}^p \end{aligned} \tag{18.24}$$

The conditional distribution of the rv X given Y is therefore also Gaussian with (conditional) mean $\mathbb{E}\left[X|Y\right]$ and covariance matrix $\text{Cov}\left[X\right] - A^\star \text{Cov}\left[Y, X\right]$.

Proof. Fix θ in $\mathbb{R}^p$. Noting that

$$X = V + \mathbb{E}\left[X\right] + A^\star(Y - \mathbb{E}\left[Y\right]) = V + \mathbb{E}\left[X|Y\right]$$

where $A^\star$ is as before, we get

$$\begin{aligned} \mathbb{E}\left[e^{i\theta^t X}|Y\right] &= \mathbb{E}\left[e^{i\theta^t V} \cdot e^{i\theta^t \mathbb{E}[X|Y]}|Y\right] \\ &= \mathbb{E}\left[e^{i\theta^t V}|Y\right] \cdot e^{i\theta^t \mathbb{E}[X|Y]} \\ &= \mathbb{E}\left[e^{i\theta^t V}\right] \cdot e^{i\theta^t \mathbb{E}[X|Y]} \quad \text{a.s.} \end{aligned} \tag{18.25}$$

where in the last step we used the fact that the rv V is independent of the rv Y, a fact established in Lemma 18.9.3.

But, the rv V is a Gaussian rv $\Omega \to \mathbb{R}^p$ as shown in Lemma 18.9.3; its characteristic function is therefore determined by its mean and its covariance: First, we see that

$$\mathbb{E}\left[V\right] = \mathbb{E}\left[X - \mathbb{E}\left[X\right] - A^\star(Y - \mathbb{E}\left[Y\right])\right] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

while

$$\begin{aligned} \text{Cov}\left[V\right] &= \text{Cov}\left[X - \mathbb{E}\left[X\right] - A^\star(Y - \mathbb{E}\left[Y\right])\right] \\ &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right] - A^\star(Y - \mathbb{E}\left[Y\right]))\,(X - \mathbb{E}\left[X\right] - A^\star(Y - \mathbb{E}\left[Y\right]))^t\right] \\ &= \text{Cov}\left[X\right] - \text{Cov}\left[X, Y\right](A^\star)^t - A^\star\text{Cov}\left[Y, X\right] + A^\star\text{Cov}\left[Y\right](A^\star)^t \\ &= \text{Cov}\left[X\right] - A^\star\text{Cov}\left[Y, X\right] \end{aligned} \tag{18.26}$$

as we note that

$$A^\star\text{Cov}\left[Y\right](A^\star)^t = \text{Cov}\left[X, Y\right](A^\star)^t.$$

This is a simple consequence of the equation (18.24) satisfied by $A^\star$.

Thus,

$$\mathbb{E}\left[e^{i\theta^t V}\right] = e^{-\frac{1}{2}\theta^t(\mathrm{Cov}[X]-A^\star \mathrm{Cov}[Y,X])\theta},$$

and combining with (18.25) we get (18.24). ■

A case of particular interest arises when $\mathrm{Cov}\,[Y]$ is invertible in which case $A^\star = \mathrm{Cov}\,[X,Y]\,\mathrm{Cov}\,[Y]^{-1}$, whence

$$\mathrm{Cov}\,[V] = \mathrm{Cov}\,[X] - \mathrm{Cov}\,[X,Y]\,\mathrm{Cov}\,[Y]^{-1}\,\mathrm{Cov}\,[Y,X]$$

18.10 Evaluating $Q(x)$

The complementary distribution function (18.6) repeatedly enters the computation of various probabilities of error. Given its importance, we need to develop good approximations to $Q(x)$ over the entire range $x \geq 0$.

The error function In the literature on digital communications, probabilities of error are often expressed in terms of the so-called *error function* $\mathrm{Erf} : \mathbb{R}_+ \to \mathbb{R}$ and of its complement $\mathrm{Erfc} : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt, \quad x \geq 0 \tag{18.27}$$

and

$$\mathrm{Erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt, \quad x \geq 0. \tag{18.28}$$

A simple change of variables ($t = \frac{u}{\sqrt{2}}$) in these integrals leads to the relationships

$$\mathrm{Erf}(x) = 2\left(\Phi(x\sqrt{2}) - \frac{1}{2}\right) \quad \text{and} \quad \mathrm{Erfc}(x) = 2Q(x\sqrt{2}),$$

so that

$$\mathrm{Erf}(x) = 1 - \mathrm{Erfc}(x), \quad x \geq 0.$$

Conversely, we also have

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\right) \quad \text{and} \quad Q(x) = \frac{1}{2}\mathrm{Erfc}\left(\frac{x}{\sqrt{2}}\right).$$

Thus, knowledge of any one of the quantities Φ, Q, Erf or Erfc is equivalent to that of the other three quantities. Although the last two quantities do not have a probabilistic interpretation, evaluating Erf is computationally more efficient. Indeed, $\mathrm{Erf}(x)$ is an integral of a positive function over the *finite* interval $[0, x]$ (and not over an infinite interval as in the other cases).

Chernoff bounds To approximate $Q(x)$ we begin with a crude bound which takes advantage of (**??**): Fix $x > 0$. For each $\theta > 0$, the usual Chernoff bound argument gives

$$\begin{aligned} \mathbb{P}\left[U > x\right] &\leq \mathbb{E}\left[e^{\theta U}\right] e^{-\theta x} \\ &= e^{-\theta x + \frac{\theta^2}{2}} \\ &= e^{-\frac{x^2}{2}} e^{\frac{(\theta - x)^2}{2}} \qquad (18.29) \end{aligned}$$

where in the last equality we made use of a completion-of-square argument. The best lower bound

$$Q(x) \leq e^{-\frac{x^2}{2}}, \quad x \geq 0 \qquad (18.30)$$

is achieved upon selecting $\theta = x$ in (18.29). The bound (18.30) is referred to as a Chernoff bound; it is not very accurate for small $x > 0$ since $\lim_{x\to 0} Q(x) = \frac{1}{2}$ while $\lim_{x \to 0} e^{-\frac{x^2}{2}} = 1$.

Approximating $Q(x)$ ($x \to \infty$) The Chernoff bound shows that $Q(x)$ decays to zero for large x at least as fast as $e^{-\frac{x^2}{2}}$. However, sometimes more precise information is needed regarding the rate of decay of $Q(x)$. This issue is addressed as follows:

For each $x \geq 0$, a straigthforward change of variable yields

$$\begin{aligned} Q(x) &= \int_x^\infty \phi(t) dt \\ &= \int_0^\infty \phi(x+t) dt \\ &= \phi(x) \int_0^\infty e^{-xt} e^{-\frac{t^2}{2}} dt. \qquad (18.31) \end{aligned}$$

With the Taylor series expansion of $e^{-\frac{t^2}{2}}$ in mind, approximations for $Q(x)$ of increased accuracy thus suggest themselves by simply approximating the second exponential factor (namely e^{-xt}) in the integral at (18.31) by terms of the form

$$\sum_{k=0}^n \frac{(-1)^k}{2^k k!} t^{2k}, \quad n = 0, 1, \ldots \qquad (18.32)$$

To formulate the resulting approximation contained in Proposition 18.10.1 given next, we set

$$Q_n(x) = \phi(x) \int_0^\infty \left(\sum_{k=0}^n \frac{(-1)^k}{2^k k!} t^{2k} \right) e^{-xt} dt, \quad x \geq 0$$

for each $n = 0, 1, \ldots$.

Proposition 18.10.1 *Fix $n = 0, 1, \ldots$. For each $x > 0$ it holds that*

$$Q_{2n+1}(x) \leq Q(x) \leq Q_{2n}(x), \tag{18.33}$$

with

$$\mid Q(x) - Q_n(x) \mid \leq \frac{(2n)!}{2^n n!} x^{-(2n+1)} \phi(x). \tag{18.34}$$

where

$$Q_n(x) = \phi(x) \sum_{k=0}^{n} \frac{(-1)^k (2k)!}{2^k k!} x^{-(2k+1)}. \tag{18.35}$$

A proof of Proposition 18.10.1 can be found in Section **??**. Upon specializing (18.33) to $n = 0$ we get

$$\frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}} \left(1 - \frac{1}{x^2}\right) \leq Q(x) \leq \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}, \quad x > 0 \tag{18.36}$$

and the asymptotics

$$Q(x) \sim \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}} \quad (x \to \infty) \tag{18.37}$$

follow. Note that the lower bound in (18.36) is meaningful only when $x \geq 1$.

18.11 Rvs derived from Gaussian rvs

Rayleigh rvs A rv X is said to be a *Rayleigh* rv with parameter σ ($\sigma > 0$) if

$$X =_{st} \sqrt{Y^2 + Z^2} \tag{18.38}$$

with Y and Z independent zero mean Gaussian rvs with variance σ^2. It is easy to check that

$$\mathbb{P}[X > x] = e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0 \tag{18.39}$$

with corresponding density function

$$\frac{d}{dx}\mathbb{P}[X \leq x] = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0. \tag{18.40}$$

It is also well known that the rv Θ given by

$$\Theta := \arctan\left(\frac{Z}{Y}\right) \tag{18.41}$$

is uniformly distributed over $[0, 2\pi)$ and independent of the Rayleigh rv X, i.e.,

$$\mathbb{P}\left[X \leq x, \Theta \leq \theta\right] = \frac{\theta}{2\pi}\left(1 - e^{-\frac{x^2}{2\sigma^2}}\right), \quad \theta \in [0, 2\pi),\ x \geq 0. \tag{18.42}$$

Rice rvs A rv X is said to be a *Rice* rv with parameters α (in $\mathbb{R}$) and σ ($\sigma > 0$) if

$$X =_{st} \sqrt{(\alpha + Y)^2 + Z^2} \tag{18.43}$$

with Y and Z independent zero mean Gaussian rvs with variance σ^2. It is easy to check that X admits a probability density function given by

$$\frac{d}{dx}\mathbb{P}\left[X \leq x\right] = \frac{x}{\sigma^2} e^{-\frac{x^2+\alpha^2}{2\sigma^2}} \cdot I_0\left(\frac{\alpha x}{\sigma^2}\right), \quad x \geq 0. \tag{18.44}$$

Here,

$$I_0(x) := \frac{1}{2\pi}\int_0^{2\pi} e^{x\cos t} dt, \quad x \in \mathbb{R} \tag{18.45}$$

is the modified Bessel function of the first kind of order zero.

Chi-square rvs For each $n = 1, 2, \ldots$, the Chi-square rv with n degrees of freedom is the rv defined by

$$\chi_n^2 =_{st} U_1^2 + \ldots + U_n^2$$

where $U_1, \ldots, U_n$ are n i.i.d. standard Gaussian rvs.

18.12 Evaluating the moments of the standard Gaussian distribution

In this section we evaluate the moments of the standard zero-mean unite variance Gaussian rv U. Recall that its probability density function $\phi : \mathbb{R} \to \mathbb{R}_+$ is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

As before we write

$$m_k \equiv \mathbb{E}\left[U^k\right] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x^k e^{-\frac{x^2}{2}} dx, \quad k = 0, 1, \ldots \tag{18.46}$$

and note by symmetry that $m_{2\ell+1} = 0$ for every $\ell = 0, 1, \ldots$. Therefore we need only focus on the moments of even order $k = 2\ell$ with $\ell = 1, 2, \ldots$.

To that end, fix $\ell = 0, 1, \ldots$. Standard arguments using integration by parts yield

$$\begin{aligned}
\int_{\mathbb{R}} x^{2(\ell+1)} e^{-\frac{x^2}{2}} dx &= 2\int_0^\infty x^{2(\ell+1)} e^{-\frac{x^2}{2}} dx \\
&= 2\int_0^\infty x^{2\ell+1} \cdot \left(x e^{-\frac{x^2}{2}} \right) dx \\
&= 2\int_0^\infty x^{2\ell+1} \cdot \frac{d}{dx} \left(-e^{-\frac{x^2}{2}} \right) dx \\
&= 2\left(\left[-x^{2\ell+1} e^{-\frac{x^2}{2}} \right]_0^\infty + \int_0^\infty (2\ell+1) x^{2\ell} e^{-\frac{x^2}{2}} dx \right) \\
&= 2(2\ell+1) \int_0^\infty x^{2\ell} e^{-\frac{x^2}{2}} dx. \qquad (18.47)
\end{aligned}$$

by virtue of the fact that $\lim_{x\to\infty} x^{2\ell} e^{-\frac{x^2}{2}} = 0$. In other words, multiplying both sides by $\sqrt{2\pi}$, we conclude that

$$m_{2(\ell+1)} = (2\ell+1) m_{2\ell}, \quad \ell = 0, 1, \ldots$$

Iterating we easily get

$$\begin{aligned}
m_{2\ell} &= (2\ell-1) m_{2(\ell-1)} \\
&= (2\ell-1)(2\ell-3) m_{2(\ell-2)} \\
&\vdots \\
&= (2\ell-1)(2\ell-3)(2\ell-5) \cdot \ldots \cdot 5 \cdot 3 \cdot 1 \cdot m_0. \qquad (18.48)
\end{aligned}$$

Obviously $m_0 = 1$, and the conclusion

$$m_{2\ell} = \frac{(2\ell)!}{(2\ell)(2(\ell-1))(2(\ell-2)) \cdots (2\cdot 3)(2\cdot 2)(2\cdot 1)} = \frac{(2\ell)!}{2^\ell \ell!}, \quad \ell = 0, 1, \ldots$$

follows. The expressions (18.7) are now established. ■

18.13 Evaluating the characteristic function of Gaussian rvs

As we seek to establish the expression (18.2) for the characteristic function of Gaussian rvs, we need only consider the case of zero-mean unit variance standard Gaussian rvs, i.e., $\mu = 0$ and $\sigma^2 = 1$.

We seek to evaluate

$$\Phi_U(\theta) = \mathbb{E}\left[e^{i\theta U}\right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\theta x} e^{-\frac{1}{2}x^2} dx, \quad \theta \in \mathbb{R}.$$

Fix θ in $\mathbb{R}$. Our starting point is the Taylor series expansion

$$e^{i\theta x} = \sum_{k=0}^{\infty} \frac{(i\theta x)^k}{k!}, \quad x \in \mathbb{R}.$$

Assuming a valid interchange of integration and summation (to be justified below), we get

$$\begin{aligned} \int_{\mathbb{R}} e^{i\theta x} e^{-\frac{1}{2}x^2} dx &= \int_{\mathbb{R}} \left(\sum_{k=0}^{\infty} \frac{(i\theta x)^k}{k!} \right) e^{-\frac{1}{2}x^2} dx \\ &= \sum_{k=0}^{\infty} \int_{\mathbb{R}} \frac{(i\theta x)^k}{k!} e^{-\frac{1}{2}x^2} dx \\ &= \sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} \int_{\mathbb{R}} x^k e^{-\frac{1}{2}x^2} dx \\ &= \sqrt{2\pi} \left(\sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} \cdot m_k \right) \qquad (18.49) \end{aligned}$$

with the notation (18.46).

Using the expressions (18.7) we conclude that

$$\begin{aligned} \int_{\mathbb{R}} e^{i\theta x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx &= \sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} \cdot m_k \\ &= \sum_{\ell=0}^{\infty} \frac{(i\theta)^{2\ell}}{(2\ell)!} \cdot \frac{(2\ell)!}{2^\ell \ell!} \\ &= \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \frac{(-\theta)^2}{\ell!} \qquad (18.50) \end{aligned}$$

and the desired conclusion

$$\mathbb{E}\left[e^{i\theta U}\right] = e^{-\frac{\theta^2}{2}}, \quad \theta \in \mathbb{R}$$

follows. The general case is now immediate once we recall that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then the rv $\frac{X-\mu}{\sigma}$ is a standard zero-mean unit variance Gaussian rv. ■

18.14 Exercises

Ex. 18.1 Establish the relations (18.1) through direct integration.

Ex. 18.2 Let U denote a zero-mean unit variance Gaussian rv. With the help of Theorem 17.8.1 use the moment information contained in the characteristic function (18.2) to evaluate all the moments $\mathbb{E}\left[U^p\right]$ $(p = 1, \ldots)$. You may want to compare your results with those obtained in Section

Ex. 18.3 Find all the moments $\mathbb{E}\left[U^p\right]$ $(p = 1, \ldots)$ where X is a χ^2_n-rv with n degrees of freedom.

Ex. 18.4 Derive the relationships between the quantities Φ, Q, Erf or Erfc which are given in Section 18.10.

Ex. 18.5 Given the covariance matrix Σ, explain why the representation (**??**)–(**??**) may not be unique. Give a counterexample.

Ex. 18.6 Give a proof for Lemma 18.8.1 and of Lemma 18.8.2.

Ex. 18.7 Construct an $\mathbb{R}^2$-valued rv $X = (X_1, X_2)$ such that the $\mathbb{R}$-valued rvs X_1 and X_2 are each Gaussian but the $\mathbb{R}^2$-valued rv X is not (jointly) Gaussian.

Ex. 18.8 Derive the probability distribution function (18.39) of a Rayleigh rv with parameter σ $(\sigma > 0)$.

Ex. 18.9 Show by direct arguments that if X is a Rayleigh distribution with parameter σ, then X^2 is exponentially distributed with parameter $(2\sigma^2)^{-1}$ **[HINT:** Compute $\mathbb{E}\left[e^{-\theta X^2}\right]$ for a Rayleigh rv X for $\theta \geq 0$.]

Ex. 18.10 Derive the probability distribution function (18.44) of a Rice rv with parameters α (in $\mathbb{R}$) and σ $(\sigma > 0)$.

Ex. 18.11 Write a program to evaluate $Q_n(x)$.

Ex. 18.12 Let $X_1, \ldots, X_n$ be i.i.d. Gaussian rvs with zero mean and unit variance and write $S_n = X_1 + \ldots + X_n$. For each $a > 0$ show that

$$\mathbb{P}\left[S_n > na\right] \sim \frac{e^{-\frac{na^2}{2}}}{a\sqrt{2\pi n}} \quad (n \to \infty). \tag{18.51}$$

This asymptotic is known as the Bahadur-Rao correction to the large deviations asymptotics of S_n.

Ex. 18.13 Consider three rvs mutually independent rvs Y, Z and U defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that (i) The rv $U : \Omega \to \mathbb{R}$ is a Bernoulli rv with

$$\mathbb{P}[U = 1] = p = 1 - \mathbb{P}[U = 0]$$

for some p in $(0, 1)$; (ii) The rvs $Y, Z : \Omega \to \mathbb{R}^2$ are two-dimensional zero-mean Gaussian rvs with covariance matrices R_a and R_b, respectively, given by

$$R_\star = \left(\begin{array}{cc} 1 & \rho_\star \\ \rho_\star & 1 \end{array} \right), \quad \star = a, b$$

with $\rho_a \neq \rho_b$. The conditions $|\rho_a| \leq 1$ and $|\rho_b| \leq 1$ are assumed in order to ensure that the matrices R_a and R_b are legitimate covariance matrices.

a. Compute the characteristic function $\Phi_X : \mathbb{R} \to \mathbb{C}$ of the rv $X : \Omega \to \mathbb{R}^2$ given by

$$X = UY + (1 - U)Z.$$

b. If $X = (X_1, X_2)$, show that the component rvs X_1 and X_2 are each standard Gaussian rv.

c. Explain why the rv X is *not* a Gaussian rv. .

Ex. 18.14 The following arises in classical Statistics: Let $X_1, \ldots, X_n$ denote n i.i.d. Gaussian rvs, each with mean μ and variance $\sigma^2 > 0$. Define the rvs $\bar{X}$ and $Z_1, \ldots, Z_n$ by

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \quad \text{and} \quad Z_k = X_k - \bar{X}, \quad k = 1, 2, \ldots, n.$$

a. Compute the joint characteristic function of the $n + 1$ rvs $Z_1, \ldots, Z_n$ and $\bar{X}$.

b. Use Part **a** to establish the independence of the rvs $\bar{X}$ and S^2 where

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2.$$

Ex. 18.15 The rvs $X_1, \ldots, X_n$ are jointly Gaussian, e.g., with $\boldsymbol{X} = (X_1, \ldots, X_n)'$, namely $X \sim \mathrm{N}(\mu, R)$ for some vector μ in $\mathbb{R}^n$ and $n \times n$ covariance matrix R. With a and b elements in $\mathbb{R}^n$, define the $\mathbb{R}$-valued rvs A and B by

$$A \equiv a^t X = \sum_{k=1}^n a_k X_k \quad \text{and} \quad B \equiv b^t X = \sum_{k=1}^n b_k X_k.$$

a. Compute the characteristic function of the $\mathbb{R}^2$-valued rv $(A, B)'$, namely

$$\varphi(s,t) = \mathbb{E}\left[e^{i(sA+tB)}\right], \quad s,t \in \mathbb{R}.$$

Carefully explain your calculations!

b. With the help of your answer in Part **a** derive a necessary and sufficient condition on the parameters μ, a, $\boldsymbol{b}$ and R for the rvs A and B to be independent. Carefully explain your calculations!

c. What form does this condition take when the rvs $X_1, \ldots, X_n$ are i.i.d. Gaussian rvs, say $X \sim \mathrm{N}(\mu, \sigma^2 I_n)$ with $\sigma^2 > 0$?

Ex. 18.16 Consider the bivariate Gaussian rv $(X, Y)'$ with probability density function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(2x^2+y^2+2xy-22x-14y+65\right)}, \quad (x,y) \in \mathbb{R}^2.$$

Evaluate the quantities $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathrm{Var}[X]$, $\mathrm{Var}[Y]$ and $\mathrm{Cov}[X,Y]$.

Ex. 18.17 Let $\xi, \eta : \Omega \to \mathbb{R}$ be independent rvs, each of which is distributed according to a standard Gaussian distribution. Define the rv $(\xi^\star, \eta^\star) : \Omega \to \mathbb{R}^2$ given by

$$\begin{pmatrix} \xi^\star \\ \eta^\star \end{pmatrix} = \begin{cases} \begin{pmatrix} \xi \\ |\eta| \end{pmatrix} & \text{if } \xi \geq 0 \\ \\ \begin{pmatrix} \xi \\ -|\eta| \end{pmatrix} & \text{if } \xi < 0. \end{cases}$$

Show that rvs $\xi^\star$ and $\eta^\star$ are standard Gaussian rvs but that the rv $(\xi^\star, \eta^\star) : \Omega \to \mathbb{R}^2$ is *not* Gaussian. Contrast with the statement: The rv $(\xi, \eta) : \Omega \to \mathbb{R}^2$ is a jointly Gaussian rv $\mathrm{N}(\mathbf{0}_2, \boldsymbol{I}_2)$ with $\mathbf{0}_2 = (0,0)'$ and $\boldsymbol{I}_2$ is the identity on $\mathbb{R}^2$. What explain the difference?

Chapter 19

Convergence of random variables

We now turn to developing a convergence theory for sequences of rvs. We assume that all the rvs are defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{X_n, n = 1, 2, \ldots\}$ denote the sequence of rvs $\Omega \to \mathbb{R}^p$ whose limiting behavior is being investigated, and let $X : \Omega \to \mathbb{R}^p$ be a possible limit. Most of the discussion will be given for the case $p = 1$, as the case general case $p \geq 1$ can easily be inferred from the one-dimensional case; see Section 19.6 for comments and pointers.

We stress that the four modes of convergence to be introduced shortly are *compatible* with the usual convergence on $\mathbb{R}$ in the following sense: If the sequence $\{X_n, n = 1, 2, \ldots\}$ comprises degenerate rvs, say for each $n = 1, 2, \ldots$ we have $X_n = a_n$ a.s. for some scalar a_n, then the convergence of the sequence $\{X_n, n = 1, 2, \ldots\}$ in any one of the four sense is *equivalent* to the usual convergence of the deterministic sequence $\{a_n,\ n = 1, 2, \ldots\}$.

Basic notions of convergence in $\mathbb{R}$ are reviewed in Appendix 22.

19.1 Almost sure convergence

Almost sure convergence is the mode of convergence that is easiest to understand as it mimics most closely usual convergence.

Definition 19.1.1

The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges *almost surely (a.s.)* to the rv $X : \Omega \to \mathbb{R}$ if $\mathbb{P}[C] = 1$ where C is the event

$$C \equiv \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega) \text{ in } \mathbb{R}\}. \tag{19.1}$$

We shall write $\lim_{n\to\infty} X_n = X$ a.s.

Sometimes the qualifier “almost sure(ly)” is replaced by the qualifier “with probability one” (often abbreviated as w.p. 1), in which case we write $\lim_{n\to\infty} X_n = X$ w.p. 1. It is easy to see that the convergence set C is indeed an event in $\mathcal{F}$ since

$$C = \cap_{k=1}^{\infty} \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} \left[|X_m - X| \leq \frac{1}{k} \right].$$

The following notation will prove convenient in what follows: With $\varepsilon > 0$, for each $n = 1, 2, \ldots$, we define the events

$$A_n(\varepsilon) \equiv [|X_n - X| \leq \varepsilon]$$

and

$$\begin{aligned} B_n(\varepsilon) &\equiv \cap_{m\geq n} A_m(\varepsilon) \\ &= [|X_m - X| \leq \varepsilon,\ m = n, n+1, \ldots]. \end{aligned} \tag{19.2}$$

Theorem 19.1.1 *The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges a.s. to the rv X if and only if*

$$\mathbb{P}\left[B_\infty(\varepsilon)\right] = 1, \quad \varepsilon > 0 \tag{19.3}$$

with

$$B_\infty(\varepsilon) = \cup_{n=1}^{\infty} B_n(\varepsilon). \tag{19.4}$$

Proof. With this notation, the characterization of C given earlier can now be expressed in the more compact form

$$C = \cap_{k=1}^{\infty} B_\infty(k^{-1}).$$

Note also that $B_\infty(\varepsilon') \subseteq B_\infty(\varepsilon)$ whenever $0 < \varepsilon' < \varepsilon$. Hence, by the continuity from above of $\mathbb{P}$ under monotone decreasing sequences we get

$$\mathbb{P}\left[C\right] = \lim_{k\to\infty} \mathbb{P}\left[B_\infty(k^{-1})\right]. \tag{19.5}$$

As this last convergence is monotonically decreasing as k increases, we conclude that $\mathbb{P}\left[C\right] = 1$ if and only if

$$\mathbb{P}\left[B_\infty(k^{-1})\right] = 1, \quad k = 1, 2, \ldots.$$

The conclusion (19.3) follows since for every $\varepsilon > 0$ there exists a positive integer k such that $(k+1)^{-1} \leq \varepsilon < k^{-1}$ with $B_\infty((k+1)^{-1}) \subseteq B_\infty(\varepsilon) \subseteq B_\infty(k^{-1})$. ■

This simple observation paves the way for the following simple criterion for a.s. convergence.

Theorem 19.1.2 *The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges a.s. to the rv X if for every $\varepsilon > 0$, it holds that*

$$\sum_{n=1}^{\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] < \infty. \tag{19.6}$$

Proof. Pick $\varepsilon > 0$. Note that $B_\infty(\varepsilon) = \liminf_{n\to\infty} A_n(\varepsilon)$, or equivalently, $B_\infty(\varepsilon)^c = \limsup_{n\to\infty} A_n(\varepsilon)^c$. The first Borel-Cantelli Lemma (Lemma 3.3.1) now yields $\mathbb{P}\left[B_\infty(\varepsilon)^c\right] = 0$ provided

$$\sum_{n=1}^{\infty} \mathbb{P}\left[A_n(\varepsilon)^c\right] < \infty.$$

This statement is equivalent to $\mathbb{P}\left[B_\infty(\varepsilon)\right] = 1$ provided (19.6) holds, and the proof is completed by invoking Theorem 19.1.1. ■

The condition (19.6) is sufficient, but *not* necessary, to ensure a.s. convergence. However, it occurs sufficiently often that it has been given a name.

Definition 19.1.2

The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ is said to be *completely convergent* to the rv X if for every $\varepsilon > 0$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] < \infty. \tag{19.7}$$

Theorem 19.1.2 states that complete convergence implies a.s. convergence. That complete convergence is only a sufficient condition for a.s. convergence, and not a necessary condition for it, is confirmed by the next example.

Counterexample 19.1.1 A.s. convergence does not imply complete convergence Take Ω = $[0, 1]$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P}$ is Lebesgue measure λ. Define the rvs $\{X_n,\ n = 1, 2, \ldots\}$ to be

$$X_n = \begin{cases} 0 & \text{if } 0 \le \omega \le 1 - \frac{1}{n} \\ 1 & \text{if } 1 - \frac{1}{n} < \omega \le 1 \end{cases}$$

for every $n = 1, 2, \ldots$. Fix ω in $[0, 1)$. It is plain that $\lim_{n\to\infty} X_n(\omega) = 0$, and the sequence $\{X_n,\ n = 1, 2, \ldots\}$ converges a.s. to the rv $X \equiv 0$. However, for every ε in $(0, 1)$, we get

$$\mathbb{P}\left[|X_n| > \varepsilon\right] = \frac{1}{n}, \quad n = 1, 2, \ldots$$

whence (19.7) fails since $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ by the divergence of the harmonic series. ■

19.2 Convergence in probability

The next mode of convergence is closely related to almost sure convergence, but less demanding, hence more likely to hold.

Definition 19.2.1

The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges *in probability* to the rv X if for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] = 0. \tag{19.8}$$

We shall write $X_n \xrightarrow{P}_n X$.

As expected s.s. convergence is a stronger notion of convergence than convergence in probability.

Theorem 19.2.1 *Almost sure convergence implies convergence in probability: If the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges a.s. to the rv X, then it also converges in probability to the rv X.*

Proof. Pick $\varepsilon > 0$ arbitrary. We have $B_n(\varepsilon) \subseteq A_n(\varepsilon)$ for each $n = 1, 2, \ldots$, whence

$$\mathbb{P}\left[B_n(\varepsilon)\right] \leq \mathbb{P}\left[A_n(\varepsilon)\right], \quad n = 1, 2, \ldots$$

The sets $\{B_n(\varepsilon),\ n = 1, 2, \ldots\}$ being non-decreasing, we readily conclude that $\lim_{n\to\infty} \mathbb{P}\left[B_n(\varepsilon)\right] = \mathbb{P}\left[B_\infty(\varepsilon)\right]$ with $B_\infty(\varepsilon)$ defined at (19.4). It is now plain that

$$\mathbb{P}\left[B_\infty(\varepsilon)\right] = \lim_{n\to\infty} \mathbb{P}\left[B_n(\varepsilon)\right] \leq \liminf_{n\to\infty} \mathbb{P}\left[A_n(\varepsilon)\right].$$

By Theorem 19.1.1 the a.s. convergence of the sequence $\{X_n,\ n = 1, 2, \ldots\}$ implies $\mathbb{P}[B_\infty(\varepsilon)] = 1$, and this immediately implies $\liminf_{n\to\infty} \mathbb{P}[A_n(\varepsilon)] = 1$. Thus, $\lim_{n\to\infty} \mathbb{P}[A_n(\varepsilon)] = 1$, and the sequence $\{X_n,\ n = 1, 2, \ldots\}$ converges in probability. ■

Here is an example of a sequence which converges in probability but not almost surely:

Counterexample 19.2.1 Convergence in probability does not imply a.s. convergence Take Ω = $[0, 1]$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P}$ is Lebesgue measure λ. Define the rvs $\{X_n,\ n = 1, 2, \ldots\}$ as follows: For each $n = 1, 2, \ldots$, there exists a unique integer $k = 0, 1, \ldots$ such that $2^k \le n < 2^{k+1}$ so that $n = 2^k + m$ for some unique $m = 0, \ldots, 2^k - 1$. Define

$$X_n = \begin{cases} 1 & \text{if } \omega \in I_n \\ \\ 0 & \text{if } \omega \notin I_n \end{cases}$$

where $I_n = (m2^{-k}, (m + 1)2^{-k})$.

The set Ω_b of boundary points

$$\Omega_b = \left\{ m2^{-k},\ m = 0, \ldots, 2^k,\ k = 0, 1, \ldots \right\}$$

is countable, hence $\mathbb{P}[\Omega_b] = 0$. With ω not in Ω_b we note that $X_n(\omega) = 0$ and $X_n(\omega) = 1$ infinitely often, so that $\liminf_{n\to\infty} X_n(\omega) = 0 < \limsup_{n\to\infty} X_n(\omega) = 1$. The sequence $\{X_n,\ n = 1, 2, \ldots\}$ therefore does not converge a.s.. However, with $X = 0$, we have $\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$ for every $\varepsilon > 0$ since

$$\mathbb{P}[|X_n - X| > \varepsilon] = \begin{cases} \mathbb{P}[I_n] & \text{if } 0 < \varepsilon < 1 \\ \\ 0 & \text{if } 1 \ge \varepsilon. \end{cases}$$

The sequence $\{X_n,\ n = 1, 2, \ldots\}$ indeed converges in probability. ■

Yet, despite this counterexample which shows that a.s. convergence is strictly stronger than convergence in probability, there is a partial converse in the following sense.

Theorem 19.2.2 *Convergence in probability implies almost sure convergence but only along a subsequence: If the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in* *probability to the rv X, then there exists a (deterministic) subsequence $\mathbb{N}_0 \to \mathbb{N}_0$ with*

$$n_k < n_{k+1}, \quad k = 1, 2, \ldots$$

such that the subsequence of rvs $\{X_{n_k},\ k = 1, 2, \ldots\}$ converges almost surely to X.

The constraint on the subsequence $\{n_k,\ k = 1, 2, \ldots\}$ implies $\lim_{k\to\infty} n_k = \infty$. We stress that this subsequence is *independent* of the sample ω (in the appropriate certain event) for which the a.s. convergence of $\{X_{n_k}(\omega),\ k = 1, 2, \ldots\}$ is established.

Proof. The assumed convergence in probability of the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ to the rv X amounts to

$$\lim_{n\to\infty} \mathbb{P}\left[|X - X_n| > \varepsilon\right] = 0, \quad \varepsilon > 0.$$

Fix $\varepsilon > 0$: For every $\delta > 0$ there exists a positive integer $n^\star(\varepsilon, \delta)$ such that

$$\mathbb{P}\left[|X - X_n| > \varepsilon\right] \leq \delta, \quad n \geq n^\star(\varepsilon, \delta).$$

We now use this observation (with $\varepsilon = k^{-1}$ and $\delta = 2^{-k}$) as follows: For each $k = 1, 2, \ldots$, there exists a positive integer n_k such that

$$\mathbb{P}\left[|X - X_n| > k^{-1}\right] \leq 2^{-k}, \quad n \geq n_k.$$

It is always possible to recursively select n_k as any positive integer satisfying

$$\max\left(n^\star(\varepsilon, \delta), n_{k-1}\right) < n_k$$

with the convention $n_0 = 0$. This construction guarantees $n_k < n_{k+1}$ for all $k = 1, 2, \ldots$.

Pick $\varepsilon > 0$ and introduce the integer $k(\varepsilon) = \lfloor \varepsilon^{-1} \rfloor$. With the quantities just introduced we have

$$\begin{aligned}
&\sum_{k=1}^{\infty} \mathbb{P}\left[|X_{n_k} - X| > \varepsilon\right] \\
&\quad = \sum_{k=1,2,\ldots:\ k^{-1} > \varepsilon} \mathbb{P}\left[|X_{n_k} - X| > \varepsilon\right] + \sum_{k=1,2,\ldots:\ k^{-1} \leq \varepsilon} \mathbb{P}\left[|X_{n_k} - X| > \varepsilon\right] \\
&\quad \leq k(\varepsilon) + \sum_{k=k(\varepsilon)}^{\infty} \mathbb{P}\left[|X_{n_k} - X| > k^{-1}\right] \\
&\quad \leq k(\varepsilon) + \sum_{k=k(\varepsilon)}^{\infty} 2^{-k}.
\end{aligned}$$

As the conclusion $\sum_{k=1}^{\infty} \mathbb{P}\left[|X_{n_k} - X| > \varepsilon\right] < \infty$ follows, the a.s. convergence of the sequence of rvs $\{X_{n_k},\ k = 1, 2, \ldots\}$ is now a consequence of Theorem 19.1.2.
■

19.3 Convergence in the r^{th} mean

Definition 19.3.1

With $r \geq 1$, the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges to the rv X in the r^{th} mean if the rvs $\{X_n,\ n = 1, 2, \ldots\}$ satisfy

$$\mathbb{E}\left[|X_n|^r\right] < \infty, \quad n = 1, 2, \ldots \tag{19.9}$$

and

$$\lim_{n \to \infty} \mathbb{E}\left[|X_n - X|^r\right] = 0. \tag{19.10}$$

We shall write $X_n \xrightarrow{L^r}_n X$.

The case $r = 2$ is often used in applications where it is referred as *mean-square* convergence. The case $r = 1$ also occurs with some regularity, and is referred as *mean* convergence. It follows from (19.10) that $\mathbb{E}\left[|X_n - X|^r\right] < \infty$ for all n sufficiently large, whence the rv X necessarily has a finite moment of order r by virtue of Minkowski's inequality under (19.9).

Convergence in the r^{th} mean becomes more stringent as r increases.

Theorem 19.3.1 *With $1 \leq s < r$, convergence in the r^{th} mean implies convergence in the s^{th}mean: If the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in the r^{th} mean to the rv X, then the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ also converges in the s^{th} mean to the rv X.*

Proof. This is a simple consequence of Lyapounov's inequality

$$\mathbb{E}\left[|X_n - X|^s\right]^{\frac{1}{s}} \leq \mathbb{E}\left[|X_n - X|^r\right]^{\frac{1}{r}}, \quad n = 1, 2, \ldots$$

■

Next, we relate r^{th} mean convergence to convergence in probability.

Theorem 19.3.2 *Convergence in the r^{th} mean implies convergence in probability: If the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in r^{th} mean to the rv X for some $r \geq 1$, then it also converges in probability to the rv X.*

Proof. Pick $\varepsilon > 0$ arbitrary. Markov's inequality yields

$$\begin{aligned} \mathbb{P}\left[|X_n - X| > \varepsilon\right] &= \mathbb{P}\left[|X_n - X|^r > \varepsilon^r\right] \\ &\leq \frac{\mathbb{E}\left[|X_n - X|^r\right]}{\varepsilon^r}, \quad n = 1, 2, \ldots \end{aligned} \tag{19.11}$$

and $\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] = 0$ as soon as $\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0$. ■

The converse is more delicate as the next example already illustrates; see also Section 19.5.

Counterexample 19.3.1 Consider a collection of rvs $\{X_n,\ n = 1, 2, \ldots\}$ such that

$$X_n = \begin{cases} 0 & \text{with probability } 1 - n^{-\alpha} \\ & \\ n^\beta & \text{with probability } n^{-\alpha} \end{cases}$$

for each $n = 1, 2, \ldots$ where $\alpha > 0$ and $\beta > 0$. Thus,

$$\mathbb{P}\left[|X_n| > \varepsilon\right] = n^{-\alpha}, \quad n = 1, 2, \ldots$$

as soon as $0 < \varepsilon \leq 1$ so that $X_n \xrightarrow{P}_n 0$.

On the other hand, with $r \geq 1$, elementary calculations show that

$$\mathbb{E}\left[|X_n|^r\right] = 0\left(1 - n^{-\alpha}\right) + n^{r\beta} n^{-\alpha} = n^{r\beta - \alpha}, \quad n = 1, 2, \ldots$$

so that

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n|^r\right] = \begin{cases} 0 & \text{if } r\beta < \alpha \\ 1 & \text{if } r\beta = \alpha \\ \infty & \text{if } r\beta > \alpha. \end{cases}$$

It is now plain that $X_n \xrightarrow{L^r}_n 0$ when $r\beta < \alpha$ but no such conclusion can be reached when $r\beta \geq \alpha$. ■

We close this section with a simple observation, based on Theorem 19.1.2, which allows us to determine a.s. convergence in the presence of convergence in the r^{th} mean.

Theorem 19.3.3 *If the sequence of rvs $\{X_n,\ n = 1,2,\ldots\}$ converges in r^{th} mean to the rv X for some $r \geq 1$, then it also converges almost surely to the rv X whenever the condition*

$$\sum_{n=1}^{\infty} \mathbb{E}\left[|X_n - X|^r\right] < \infty \tag{19.12}$$

holds.

The convergence condition (19.10) is automatically satisfied under the summability condition (19.12).

Proof. As already shown at (19.11), Markov's inequality leads to

$$\mathbb{P}\left[|X_n - X| > \varepsilon\right] \leq \frac{\mathbb{E}\left[|X_n - X|^r\right]}{\varepsilon^r}, \quad n = 1, 2, \ldots$$

for every $\varepsilon > 0$. The assumed condition (19.12) yields complete the convergence condition

$$\sum_{n=1}^{\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] \leq \frac{1}{\varepsilon^r} \sum_{n=1}^{\infty} \mathbb{E}\left[|X_n - X|^r\right] < \infty$$

and the conclusion is immediate by Theorem 19.1.2. ■

19.4 Convergence in distribution

For any rv $X : \Omega \to \mathbb{R}$, recall the properties satisfied by its probability distribution function $F_X : \mathbb{R} \to [0,1]$ – See Section 7.4 for details: (i) It is non-decreasing; (ii) It has left-limit and is right-continuous at every point; and (iii) The limits $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$ hold.

Let $\mathcal{C}(F_X)$ denote the set of points in $\mathbb{R}$ where $F_X : \mathbb{R} \to [0,1]$ is continuous; see Definition 7.4.1. Recall from Lemma 7.4.1 that $\mathcal{C}(F_X)^c$ is a countable subset of $\mathbb{R}$.

Definition 19.4.1

The sequence of rvs $\{X_n,\ n = 1,2,\ldots\}$ converges in distribution to the rv X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in \mathcal{C}(F_X). \tag{19.13}$$

We shall write $X_n \Longrightarrow_n X$ or $X_n \xrightarrow{\mathcal{L}}_n X$. Some authors refer to this mode of convergence as convergence in law or as weak convergence.

As this mode of convergence involves *only* the probability distribution functions of the rvs involved, it is sometimes convenient to define this notion without any reference to the rvs (viewed as mappings):

Definition 19.4.2

The sequence of probability distribution functions $\{F_n,\ n = 1, 2, \ldots\}$ converges in distribution to the probability distribution function F if

$$\lim_{n\to\infty} F_n(x) = F(x), \quad x \in \mathcal{C}(F) \tag{19.14}$$

where $\mathcal{C}(F)$ denotes the set of continuity of the probability distribution function F; see Definition 7.4.1. This time we write $F_n \Longrightarrow_n F$ or $F_n \xrightarrow{\mathcal{L}}_n F$.

At this point the reader may wonder as to why the definition of distribution convergence requires the convergence (19.13) only on the set of points of continuity of the limit. This is best seen on the following example.

Example 19.4.1 The importance of discontinuity points Consider the two sequences of rvs $\{X_n,\ n = 1, 2, \ldots\}$ and $\{Y_n,\ n = 1, 2, \ldots\}$ given by

$$X_n = -\frac{1}{n} \quad \text{and} \quad Y_n = \frac{1}{n}, \quad n = 1, 2, \ldots$$

defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Both sequences converge as *deterministic* sequences with $\lim_{n\to\infty} X_n(\omega) = 0$ and $\lim_{n\to\infty} Y_n(\omega) = 0$ for every ω in Ω. Yet it is easy to check that

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \\ 1 & \text{if } x \geq 0 \end{cases} \quad \text{and} \quad \lim_{n\to\infty} F_{Y_n}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \\ 1 & \text{if } x > 0. \end{cases}$$

It should be noted that $\lim_{n\to\infty} F_{X_n}$ is a probability distribution function while $\lim_{n\to\infty} F_{Y_n}$ is not – This second limit is a left-continuous function with right limits, and fails to be right-continuous at $x = 0$. It might be natural to state that $X_n \Longrightarrow_n X$ but not that $Y_n \Longrightarrow_n Y$ even though the sequence $\{Y_n,\ n = 1, \ldots\}$ converges pointwise. Compatibility. ■

The next result relates convergence in distribution to convergence in probability – The latter always implies the former!

Theorem 19.4.1 *Convergence in probability implies convergence in distribution: If the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in probability to the rv X, then it also converges in distribution to the rv X.*

Proof. Fix $n = 1, 2, \ldots$ and pick x in $\mathbb{R}$. With $\varepsilon > 0$, we note that

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}\left[X_n \leq x\right] \\ &= \mathbb{P}\left[X_n \leq x, X \leq x + \varepsilon\right] + \mathbb{P}\left[X_n \leq x, x + \varepsilon < X\right] \\ &\leq \mathbb{P}\left[X \leq x + \varepsilon\right] + \mathbb{P}\left[|X_n - X| > \varepsilon\right] \\ &= F_X(x + \varepsilon) + \mathbb{P}\left[|X_n - X| > \varepsilon\right]. \end{aligned}$$

In a similar way, we find

$$\begin{aligned} F_X(x - \varepsilon) &= \mathbb{P}\left[X \leq x - \varepsilon\right] \\ &= \mathbb{P}\left[X \leq x - \varepsilon, X_n \leq x\right] + \mathbb{P}\left[X \leq x - \varepsilon, x < X_n\right] \\ &\leq \mathbb{P}\left[X_n \leq x\right] + \mathbb{P}\left[|X_n - X| > \varepsilon\right] \\ &= F_{X_n}(x) + \mathbb{P}\left[|X_n - X| > \varepsilon\right]. \end{aligned}$$

Let n go to infinity in these inequalities. Under the assumed convergence in probability, we find $\limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x + \varepsilon)$ and $F_X(x - \varepsilon) \leq \liminf_{n\to\infty} F_{X_n}(x)$. Picking x to be a point of continuity for F_X, we obtain

$$\begin{aligned} \limsup_{n\to\infty} F_{X_n}(x) &= \lim_{\varepsilon\downarrow 0}\left(\limsup_{n\to\infty} F_{X_n}(x)\right) \\ &\leq \lim_{\varepsilon\downarrow 0} F_X(x + \varepsilon) \\ &= F_X(x) \end{aligned}$$

and

$$\begin{aligned} F_X(x) &= \lim_{\varepsilon\downarrow 0} F_X(x - \varepsilon) \\ &\leq \lim_{\varepsilon\downarrow 0}\left(\liminf_{n\to\infty} F_{X_n}(x)\right) \\ &= \liminf_{n\to\infty} F_{X_n}(x) \end{aligned}$$

whence $\liminf_{n\to\infty} F_{X_n}(x) = \limsup_{n\to\infty} F_{X_n}(x) = F_X(x)$. It follows that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in \mathcal{C}(F_X)$$

and the desired convergence in distribution takes place. ■

Although weak convergence is weaker than convergence in probability, there is one situation where they are equivalent. With c a scalar in $\mathbb{R}$, we refer to any rv X such that $X = c$ a.s. as the degenerate rv $X = c$.

Theorem 19.4.2 *With c a scalar in $\mathbb{R}$, the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in probability to the degenerate rv $X = c$ if and only if the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in distribution to the degenerate rv $X = c$.*

Proof. In view of Theorem 19.4.1 we need only show that if the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in distribution to the degenerate rv $X = c$, then the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in probability to the degenerate rv $X = c$.

Fix $\varepsilon > 0$. For every $n = 1, 2, \ldots$, we observe that

$$\begin{aligned}\mathbb{P}\left[|X_n - X| \le \varepsilon\right] &= \mathbb{P}\left[c - \varepsilon \le X_n \le c + \varepsilon\right] \\ &= \mathbb{P}\left[X_n \le c + \varepsilon\right] - \mathbb{P}\left[X_n < c - \varepsilon\right] \\ &= F_{X_n}(c + \varepsilon) - F_{X_n}((c - \varepsilon)-) \qquad (19.15)\end{aligned}$$

so that

$$\begin{aligned}\mathbb{P}\left[|X_n - X| > \varepsilon\right] &= 1 - F_{X_n}(c + \varepsilon) + F_{X_n}((c - \varepsilon)-) \\ &\le 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon).\end{aligned}$$

Recall that $F_X(x) = 0$ (resp. $F_X(x) = 1$) if $x < c$ (resp. $c \le x$) so that the only point of discontinuity of F_X is located at $x = c$. Thus, under the assumed convergence in distribution of the sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$, we have $\lim_{n\to\infty} F_{X_n}(c+\varepsilon) = 1$ and $\lim_{n\to\infty} F_{X_n}(c-\varepsilon) = 0$, and the desired conclusion $\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] = 0$ follows. ■

19.5 Uniform integrability

If a rv X has a finite first moment, we know that

$$\lim_{B\to\infty} \mathbb{E}\left[\mathbf{1}\left[|X| > B\right] \cdot |X|\right] = 0. \qquad (19.16)$$

This is a simple consequence of the Dominated Convergence Theorem since $\mathbf{1}\left[|X| > B\right] \cdot |X| \le |X|$ for all $B > 0$. Thus, for every $\varepsilon > 0$, there exists $B^\star(\varepsilon) > 0$ such that

$$\mathbb{E}\left[\mathbf{1}\left[|X| > B\right] \cdot |X|\right] \le \varepsilon, \quad B \ge B^\star(\varepsilon). \qquad (19.17)$$

As we consider a collection of rvs $\{X_n,\ n = 1, 2, \ldots\}$ with finite first moments, we can certainly assert the following: For each $n = 1, 2, \ldots$ and every $\varepsilon > 0$, there exists $B^\star(\varepsilon; n) > 0$ such that

$$\mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] \leq \varepsilon, \quad B \geq B^\star(\varepsilon; n). \tag{19.18}$$

This is a direct consequence of (19.17). However, sometimes it is required that this condition holds *uniformly* with respect to $n = 1, 2, \ldots$ in that $B^\star(\varepsilon; n)$ can be selected *independently* of n. This leads to the following stronger notion of integrability for a *sequence* of rvs, rather than for a *single* rv.

Definition 19.5.1

The collection of rvs $\{X_n,\ n = 1, 2, \ldots\}$ is *uniformly integrable* if

$$\lim_{B \to \infty} \left(\sup_{n=1,2,\ldots} \mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] \right) = 0. \tag{19.19}$$

In other words, for every $\varepsilon > 0$, there exists $B^\star(\varepsilon) > 0$ such that

$$\sup_{n=1,2,\ldots} \mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] \leq \varepsilon, \quad B \geq B^\star(\varepsilon). \tag{19.20}$$

The uniform integrability of the rvs $\{X_n,\ n = 1, 2, \ldots\}$ readily implies [Exercise 19.10] the boundedness condition

$$\sup_{n=1,2,\ldots} \mathbb{E}\left[|X_n|\right] < \infty. \tag{19.21}$$

While this condition is not sufficient to imply uniform integrability, a slight strengthening of it will.

Lemma 19.5.1 *The collection of rvs $\{X_n,\ n = 1, 2, \ldots\}$ is uniformly integrable if there exists $r > 0$ such that*

$$\sup_{n=1,2,\ldots} \mathbb{E}\left[|X_n|^{1+r}\right] < \infty. \tag{19.22}$$

Proof. Fix $n = 1, 2, \ldots$ and $B > 0$. Applying Hölder's inequality to the rv $|X_n|$

and $\mathbf{1}\left[|X_n| > B\right]$ (with conjugate exponents $p = \frac{r+1}{r}$ and $q = 1 + r$), we get

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] &= \left(\mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right]\right]\right)^{\frac{r}{r+1}} \cdot \left(\mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{1}{1+r}} \\
&= \left(\mathbb{P}\left[|X_n| > B\right]\right)^{\frac{r}{r+1}} \cdot \left(\mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{1}{1+r}} \\
&\leq \left(\frac{1}{B} \cdot \mathbb{E}\left[|X_n|\right]\right)^{\frac{r}{r+1}} \cdot \left(\mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{1}{1+r}} .
\end{aligned}$$

with the help of Markov's inequality in the last step.

Using the standard inequality $\mathbb{E}\left[|X_n|\right] \leq 1 + \mathbb{E}\left[|X_n|^{1+r}\right]$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] \\
&\quad \leq \ B^{-\frac{r}{r+1}} \cdot \left(1 + \mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{r}{r+1}} \cdot \left(\mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{1}{1+r}} \\
&\quad \leq \ C \cdot B^{-\frac{r}{r+1}}
\end{aligned}$$

with finite constant C given by

$$C \equiv \sup_{n=1,2,\ldots} \left(\left(1 + \mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{r}{r+1}} \cdot \left(\mathbb{E}\left[|X_n|^{1+r}\right]\right)^{\frac{1}{1+r}}\right).$$

In other words, the uniform bound

$$\sup_{n=1,2,\ldots} \ \mathbb{E}\left[\mathbf{1}\left[|X_n| > B\right] \cdot |X_n|\right] \leq C B^{-\frac{r}{r+1}}, \quad B > 0$$

holds, and the uniform integrability condition (19.19) follows. ■

Interest in this notion arises from the need to have an easy characterization of situations where interchange between limits and expectation can take place. This is captured by the next result.

Theorem 19.5.1 *Consider a collection of rvs* $\{X, X_n, \ n = 1, 2, \ldots\}$ *such that* $\lim_{n\to\infty} X_n = X$ *a.s. (resp.* $X_n \xrightarrow{P}_n X$, $X_n \Longrightarrow_n X$*). If the collection of rvs* $\{X_n, \ n = 1, 2, \ldots\}$ *is uniformly integrable, then* $\mathbb{E}\left[|X|\right] < \infty$ *and*

$$\lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \mathbb{E}\left[X\right]. \tag{19.23}$$

19.6 Convergence in higher dimensions

The discussion so far has been in the context of $\mathbb{R}$-valued rvs. We now outline the corresponding theory for $\mathbb{R}^p$-valued rvs with $p \geq 1$. The first observation is that the three first modes of convergence, namely a.s. convergence, convergence in probability and convergence in the r^{th} mean are "metric" notions in the following sense: The rvs $\{X_n,\ n = 1, 2, \ldots\}$

- converge a.s. to the rv X if
$$\lim_{n\to\infty} |X_n - X| = 0 \quad a.s.$$
- converge in probability to the rv X if
$$\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon\right] = 0, \quad \varepsilon > 0$$
- converge in the r^{th} mean (for some $r \geq 1$) to the rv X if
$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0.$$

They are all expressed in terms of the *distance* $|X_n - X|$ of X_n to X.

In $\mathbb{R}^p$ there are a number of ways to define the distance between two vectors. Here we limit ourselves to metrics that are induced by norms, so that distance is measured by
$$d(x, y) = \|x - y\|, \quad x, y \in \mathbb{R}^p$$
where $\|\cdot\| : \mathbb{R}^p \to \mathbb{R}_+$ is a norm. Therefore, a natural to define the modes of convergence for $\mathbb{R}^p$-valued rvs as follows:

Consider any norm $\|\cdot\| : \mathbb{R}^p \to \mathbb{R}_+$. The $\mathbb{R}^p$-valued rvs $\{X_n,\ n = 1, 2, \ldots\}$

- converge a.s. to the rv X if
$$\lim_{n\to\infty} \|X_n - X\| = 0 \quad a.s.$$
- converge in probability to the rv X if
$$\lim_{n\to\infty} \mathbb{P}\left[\|X_n - X\| > \varepsilon\right] = 0, \quad \varepsilon > 0$$
- converge in the r^{th} mean (for some $r \geq 1$) to the rv X if
$$\lim_{n\to\infty} \mathbb{E}\left[\|X_n - X\|^r\right] = 0.$$

Note that all norms on $\mathbb{R}^p$ are equivalent in the following sense: If $\|\cdot\|_a : \mathbb{R}^p \to \mathbb{R}_+$ and $\|\cdot\|_b : \mathbb{R}^p \to \mathbb{R}_+$ are two different norms, then there exists constants $c_{a|b} > 0$ and $C_{a|b} >$ such that

$$c_{a|b}\|x\|_a \leq \|x\|_b \leq C_{a|b}\|x\|_a, \quad x \in \mathbb{R}^p.$$

Norms often used in applications include

- The Euclidean norm (or L$_1$-norm):

$$\|x\|_2 = \sqrt{\sum_{k=1}^{p} |x_k|^2}, \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

- The L$_1$-norm:

$$\|x\|_1 = \sum_{k=1}^{p} |x_k|, \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

- The Manhattan norm

$$\|x\|_\infty = \max(|x_k|,\ k = 1, \ldots, p), \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$

However when it comes to convergence in distribution matters are quite different because this notion does not rely on a notion of proximity in the range of the rvs under consideration. Furthermore, probability distribution functions on $\mathbb{R}^p$ are more cumbersome to characterize. So instead of using the definition given in Section 19.4 we instead rely on the equivalence given in Theorem 20.4

19.7 Exercises

Unless explicitly stated otherwise, all rvs are defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Ex. 19.1 (The impact of dependencies on almost sure convergence) Consider a collection of i.i.d. rvs $\{U, U_n,\ n = 1, 2, \ldots\}$, all which are uniformly distributed on the interval $[0, 1]$.

a. Define the rvs $\{V_n,\ n = 1, 2, \ldots\}$ by

$$V_n \equiv \max\left(U_1, \ldots, U_n\right), \quad n = 1, 2, \ldots$$

Show that the sequence $\{V_n,\ n = 1, 2, \ldots\}$ converges a.s. and identify its limit.

b. The rvs $\{W_n,\ n=1,2,\ldots\}$ are defined by

$$W_n \equiv \max\left(U_1^\star,\ldots,U_n^\star\right), \quad n=1,2,\ldots$$

where

$$U_n^\star \equiv \begin{cases} U & \text{if } n=2p+1 \text{ with } p=0,1,\ldots \\ \\ 1-U & \text{if } n=2p \text{ with } p=1,\ldots \end{cases}$$

Show that the sequence $\{W_n,\ n=1,2,\ldots\}$ converges a.s. and identify its limit.

Both sequences $\{U_n,\ n=1,2,\ldots\}$ and $\{U_n^\star,\ n=1,2,\ldots\}$ comprise identically distributed rvs but have very different dependency structures!

Ex. 19.2 Consider the i.i.d discrete rvs $\{W_k,\ k=1,2,\ldots\}$ with finite support $S=\{-1,1\}$ and common pmf given by

$$\mathbb{P}\left[W_k=1\right]=\alpha \quad \text{and} \quad \mathbb{P}\left[W_k=-1\right]=1-\alpha, \quad k=1,2,\ldots$$

for some $0<\alpha<1$. Define the rvs $\{W_k^\star,\ k=1,2,\ldots\}$ to be

$$W_k^\star \equiv \prod_{\ell=1}^{k} W_\ell, \quad k=1,2,\ldots$$

a. Show that the rvs $\{W_k^\star,\ k=1,2,\ldots\}$ converge in distribution to a rv $W_\infty^\star$. Identify the probability distribution function of the limiting rv $W_\infty^\star$.

b. Is it the case that the rvs $\{W_k^\star,\ k=1,2,\ldots\}$ converge in probability? Prove or disprove!

Ex. 19.3 The i.i.d. $\mathbb{R}$-valued rvs $\{X, X_n,\ n=1,2,\ldots\}$ have a common probability distribution given by

$$\mathbb{P}\left[X\le x\right]=1-e^{-x^+}, \quad x\in\mathbb{R}.$$

For all $n=1,2,\ldots$, consider the rv Y_n given by

$$Y_n \equiv \prod_{k=1}^{n}\left(X_k\cdot\mathbf{1}\left[X_k\ge 0\right]\right).$$

a. For each $n=1,2,\ldots$, show that $\mathbb{E}\left[Y_n\right]=1$ while $\mathbb{E}\left[\sqrt{Y_n}\right]=\left(\frac{\sqrt{\pi}}{2}\right)^n$.

b. Use Part **a** to identify a range for $\theta>0$ where the convergence

$$\lim_{n\to\infty}\mathbb{P}\left[Y_n>\theta^n\right]=0$$

takes place.

c. Use Part **a** to show that the sequence $\{Y_n,\ n = 1, 2, \ldots\}$ converges in probability (and identify the limiting rv) but that it does not converge in the r^{th} mean with $r = 1$.

Ex. 19.4 Consider a sequence of i.i.d. $\mathbb{R}$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ with $\mathbb{E}\left[|X|\right] < \infty$.

a. Assume that the sequence $\{X_n,\ n = 1, 2, \ldots\}$ converges almost surely, i.e., there exists an $\mathbb{R}$-valued rv X_∞ defined on the same probability triple such that $\lim_{n\to\infty} X_n = X_\infty$ a.s.. Characterize the (probability distribution of the) limiting rv X_∞. What does it imply regarding the probability distribution of the rv X.

b. Assume next that the sequence $\{X_n,\ n = 1, 2, \ldots\}$ converges in probability (and not necessarily almost surely), i.e., there exists an $\mathbb{R}$-valued rv X'_∞ defined on the same probability triple such that $X_n \xrightarrow{P}_n X'_\infty$. Characterize the (probability distribution of the) limiting rv X'_∞. What does it imply regarding the probability distribution of the rv X.

Ex. 19.5 Consider a sequence of i.i.d. $\mathbb{R}$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ which are all defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

a. Does the sequence $\{X_n,\ n = 1, 2, \ldots\}$ converge in distribution, and in the affirmative, identify the limit.

b. Determine whether the sequence $\{\frac{X_n}{n},\ n = 1, 2, \ldots\}$ converges in probability, and in the affirmative, identify the limit.

c. Give a condition to ensure that the sequence $\{\frac{X_n}{n},\ n = 1, 2, \ldots\}$ converges almost surely and identify the limit.

Ex. 19.6 Consider a collection of i.i.d. $\mathbb{R}_+$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ such that $\mathbb{P}\left[X > 0\right] = 1$ and $\mathbb{E}\left[|X|^2\right] < \infty$. The rvs $\{T_n,\ n = 0, 1, \ldots\}$ are defined by

$$T_0 \equiv 0,\ T_n \equiv \sum_{k=1}^{n} X_k \cdot \mathbf{1}\left[X_k > 0\right], \quad n = 1, 2, \ldots$$

and set

$$D_{n+1} \equiv \sqrt{T_{n+1}} - \sqrt{T_n}, \quad n = 0, 1, \ldots$$

a. Show that the sequence $\{D_n,\ n = 1, 2, \ldots\}$ converges almost surely and identify the limit.

b. Does the sequence $\{D_n,\ n = 1, 2, \ldots\}$ converge in mean-square? In the affirmative identify the limit. [**HINT:** There are a number of ways to solve Part **b**. In particular it might be useful to note the following facts: (i) For each $n = 0, 1, \ldots$, we have $D_{n+1} =_{st} \sqrt{T_n + X} - \sqrt{T_n}$ and $D_{n+2} =_{st} \sqrt{T_{n+1} + X} - \sqrt{T_{n+1}}$ with $T_n \leq T_{n+1}$; and (ii) the mapping $\mathbb{R}_+ \to \mathbb{R}_+ : x \to \sqrt{x}$ is a concave function.]

Ex. 19.7 (Generalizing Exercise 19.6) Consider a mapping $g : \mathbb{R}_+ \to \mathbb{R}_+$ which is concave and strictly increasing. In the framework of Exercise 19.6, set

$$D^g_{n+1} \equiv g(T_{n+1}) - g(T_n), \quad n = 0, 1, \ldots$$

Ex. 19.8 Let p be a fixed parameter in $(0, 1)$. Consider a family of Binomial rvs $\{X_n,\ n = 1, 2, \ldots\}$ defined on the same probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ with

$$\mathbb{P}\left[X_n = k\right] = \binom{n}{k} p^k (1-p)^{n-k}, \quad \begin{array}{c} k = 0, 1, \ldots, n \\ n = 1, 2, \ldots \end{array}$$

Define the sequence of rvs $\{Z_n,\ n = 1, 2, \ldots\}$ given by

$$Z_n \equiv \min\left(X_n, n - X_n\right), \quad n = 1, 2, \ldots$$

a. Does the sequence $\{\frac{Z_n}{n},\ n = 1, 2, \ldots\}$ converge in probability? Carefully explain your answer and identify the limiting rv (if appropriate).

b. Does the sequence $\{\frac{Z_n}{n},\ n = 1, 2, \ldots\}$ converge in distribution? Carefully explain your answer and identify the limiting rv (if appropriate).

c. Show that the limit

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{Z_n}{n}\right]$$

exists and find its value [**HINT:** Uniform integrability]

d. Is it possible to construct a probability triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ and rvs $\{Z^\star_n,\ n = 1, 2, \ldots\}$ defined on the triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ such that (i) for each $n = 1, 2, \ldots$, the probability distribution of the rv $Z^\star_n$ under $\mathbb{P}^\star$ coincides with that of the rv Z_n under $\mathbb{P}$, and (ii) the sequence $\{\frac{Z^\star_n}{n},\ n = 1, 2, \ldots\}$ is a.s. convergent under $\mathbb{P}^\star$?

Ex. 19.9 Consider the triangular array of rvs $\{X_{n,k},\ k = 1, \ldots, n;\ n = 1, 2, \ldots\}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. For each $n = 1, 2, \ldots$, we assume that the rvs $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. rvs with

$$\mathbb{P}\left[X_{n,k} = -\sqrt{n}\right] = \mathbb{P}\left[X_{n,k} = \sqrt{n}\right] = \frac{1}{2n}, \quad k = 1, \ldots, n$$

and

$$\mathbb{P}\left[X_{n,k} = 0\right] = 1 - \frac{1}{n}, \quad k = 1, \ldots, n.$$

We write

$$S_n = \sum_{k=1}^{n} X_{n,k}, \quad n = 1, 2, \ldots.$$

a. For each $n = 1, 2, \ldots$, compute $\mathbb{E}\left[X_{n,k}\right]$ and $\mathrm{Var}\left[X_{n,k}\right]$ for all $k = 1, \ldots, n$.

6.b. Does the sequence $\{\frac{S_n}{n},\ n = 1, 2, \ldots\}$ converge in probability? In the event it does, identify the limiting rv.

Exercise 20.1 explores a related CLT-like result.

Ex. 19.10 Show that the uniform integrability condition (19.20) implies the boundedness (19.21) of the moments.

Chapter 20

From convergence in distribution to weak convergence

This chapter develops some useful tools to establish convergence in distribution. We start by recalling the definition of convergence in distribution as given in Definition 19.4.1

Definition 20.0.1

The sequence of rvs $\{X_n,\ n = 1, 2, \ldots\}$ converges in distribution to the rv X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in \mathcal{C}(F_X). \tag{20.1}$$

We shall write $X_n \Longrightarrow_n X$ or $X_n \xrightarrow{\mathcal{L}}_n X$. Some authors refer to this mode of convergence as convergence in law or as weak convergence.

20.1 Weak convergence via characteristic functions

Weak convergence of a sequence of rvs can be characterized through the limiting behavior of the corresponding sequence of characteristic functions.

Theorem 20.1.1 *The sequence of rvs $\{X_n, n = 1, 2, \ldots\}$ converges distribution to the rv X if and only if*

$$\lim_{n\to\infty} \Phi_{X_n}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}.$$

This result suggests the following strategy: Consider the limit

$$\Phi(\theta) = \lim_{n\to\infty} \Phi_{X_n}(\theta), \quad \theta \in \mathbb{R} \tag{20.2}$$

and identify the rv X whose characteristic function coincides with $\Phi : \mathbb{R} \to \mathbb{C}$. However, a word of caution is in order as the limit (20.2) may not necessarily define the characteristic function of a rv as can be seen from the following example.

Example 20.1.1 The limit of characteristic functions is not always a characteristic function For each $n = 1, 2, \ldots$, the rv X_n is the uniform rv on the interval $(-n, n)$. Easy calculations show that

$$\Phi_{X_n}(\theta) = \int_{-n}^{n} \frac{e^{i\theta x}}{2n} dx = \begin{cases} \frac{\sin(n\theta)}{n} & \text{if } \theta \neq 0 \\ 1 & \text{if } \theta = 0, \end{cases} \tag{20.3}$$

so that

$$\Phi(\theta) = \lim_{n\to\infty} \Phi_{X_n}(\theta) = \begin{cases} 0 & \text{if } \theta \neq 0 \\ 1 & \text{if } \theta = 0. \end{cases}$$

Obviously, there are no rv X whose characteristic function coincides with the limit. ■

This difficulty can be remedied with the help of the next result by simply checking *continuity* at $\theta = 0$ for the limit (20.2). This is a consequence of the Bochner-Herglotz Theorem.

Theorem 20.1.2 *Consider a sequence of rvs $\{X_n, n = 1, 2, \ldots\}$ such that the limits*

$$\Phi(\theta) = \lim_{n\to\infty} \Phi_{X_n}(\theta), \quad \theta \in \mathbb{R}$$

all exist. If $\Phi : \mathbb{R} \to \mathbb{C}$ is continuous at $\theta = 0$, then it is the characteristic function of some rv X, and $X_n \Longrightarrow_n X$.

Proof. For each $n = 1, 2, \ldots$, the function $\Phi_{X_n} : \mathbb{R} \to \mathbb{C}$ is a characteristic function. Therefore, by Theorem 17.4.1 it is (i) bounded with $|\Phi_{X_n}(\theta)| \leq \Phi_{X_n}(0) = 1$ for all θ in $\mathbb{R}$; (ii) uniformly continuous on $\mathbb{R}$; and (iii) positive semi-definite. Properties (i) and (iii) are clearly inherited by the limit $\Phi : \mathbb{R} \to \mathbb{C}$. Therefore, by Theorem 17.4.2 the assumed continuity of Φ implies that it is a characteristic function, i.e., there exists a rv X such that $\Phi = \Phi_X$. Invoking Theorem 20.1.1 we conclude that $X_n \Longrightarrow_n X$. ■

20.2 Weak convergence via the Skorokhod representation

Consider a collection $\{F, F_n,\ n = 1, 2, \ldots\}$ of probability distribution functions on $\mathbb{R}$. The Skorokhod representation discussed in Section 7.6 provides a natural connection between convergence in distribution and a.s. convergence. This is developed in the next result.

Theorem 20.2.1 *If the sequence of probability distribution functions $\{F_n,\ n = 1, 2, \ldots\}$ converges weakly to F, then there exists a probability triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ and a collection of $\mathbb{R}$-valued rvs $\{X^\star, X^\star_n,\ n = 1, 2, \ldots\}$ all defined on $\Omega^\star$ with the following properties:*

(i) We have

$$F_n(x) = \mathbb{P}^\star[X^\star_n \leq x] \quad \begin{array}{l} x \in \mathbb{R} \\ n = 1, 2, \ldots \end{array} \tag{20.4}$$

and

$$F(x) = \mathbb{P}^\star[X^\star \leq x], \quad x \in \mathbb{R}. \tag{20.5}$$

(ii) The rvs $\{X^\star_n,\ n = 1, 2, \ldots\}$ converges a.s. to $X^\star$ (under $\mathbb{P}^\star$), i.e.,

$$\mathbb{P}^\star\left[\left\{\omega^\star \in \Omega^\star : \lim_{n\to\infty} X^\star_n(\omega^\star) = X^\star(\omega^\star)\right\}\right] = 1.$$

Proof. The existence follows from the Skorokhod representation described in Lemma 7.6.1 with $\Omega^\star = [0, 1]$, $\mathcal{F}^\star = \mathcal{B}([0, 1])$ and $\mathbb{P}$ taken to be Lebesgue measure λ. The rvs $\{X^\star, X^\star_n,\ n = 1, 2, \ldots\}$ are taken to be

$$X^\star_n \equiv F^\leftarrow_n(\omega^\star), \quad \begin{array}{l} \omega^\star \in [0, 1] \\ n = 1, 2, \ldots \end{array}$$

and

$$X^\star \equiv F^\leftarrow(\omega^\star), \quad \omega^\star \in [0, 1].$$

It is easy to check that $\lim_{n\to\infty} X^\star_n(\omega^\star) = X^\star(\omega^\star)$ for every $\omega^\star$ in $\Omega^\star$ as a result of the weak convergence condition $\lim_{n\to n} F_n(x) = F(x)$ for every x in $\mathcal{C}(F)$. ■

The a.s. convergence (under $\mathbb{P}^\star$) of the sequence of rvs $\{X^\star_n,\ n = 1, 2, \ldots\}$ in now way implies the a.s. convergence of any other collection of rvs $\{X_n,\ n = 1, 2, \ldots\}$ defined on some other probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ such that the probability distribution function of the rv X_n under $\mathbb{P}$ coincides with F_{X_n} fort all $n = 1, 2, \ldots$.

20.3 Functional characterization of convergence in distribution

The following equivalent characterizations of distributional convergence have many use.

Theorem 20.3.1 *Consider the $\mathbb{R}$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The following three statements are equivalent:*
(i) The rvs $\{X_n,\ n = 1, 2, \ldots\}$ converge in distribution to the rv X, i.e.,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in \mathcal{C}(F_X).$$

(ii) For every bounded continuous mapping $g : \mathbb{R} \to \mathbb{R}$, it holds that

$$\lim_{n\to\infty} \mathbb{E}\left[g(X_n)\right] = \mathbb{E}\left[g(X)\right]. \tag{20.6}$$

(iii) The characteristic functions converge in the sense that

$$\lim_{n\to\infty} \Phi_{X_n}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}. \tag{20.7}$$

Proof. It follows from Theorem 20.2.1 that (i) implies the validity of (ii): Indeed, with the notation used in that result, consider the probability triple $(\Omega^\star, \mathcal{F}^\star, \mathbb{P}^\star)$ and the $\mathbb{R}$-valued rvs $\{X^\star, X_n^\star,\ n = 1, 2, \ldots\}$ all defined on $\Omega^\star$ such that

$$\mathbb{P}\left[X \leq x\right] = \mathbb{P}^\star[X^\star \leq x], \quad x \in \mathbb{R} \tag{20.8}$$

and

$$\mathbb{P}\left[X_n \leq x\right] = \mathbb{P}^\star[X_n^\star \leq x] \quad \begin{array}{l} x \in \mathbb{R} \\ n = 1, 2, \ldots \end{array} \tag{20.9}$$

with

$$\mathbb{P}^\star\left[\omega^\star \in \Omega^\star : \lim_{n\to\infty} X_n^\star(\omega^\star) = X^\star(\omega^\star)\right] = 1.$$

Consider a mapping $g : \mathbb{R} \to \mathbb{R}$ which is continuous and bounded - Set

$$B_g \equiv \sup_{x\in\mathbb{R}} |g(x)| < \infty.$$

We obviously have $\mathbb{E}\left[g(X)\right] = \mathbb{E}^\star\left[g^\star(X^\star)\right]$ and

$$\mathbb{E}\left[g(X_n)\right] = \mathbb{E}^\star\left[g^\star(X_n^\star)\right], \quad n = 1, 2, \ldots$$

By the continuity of g we get

$$\lim_{n\to\infty} g(X_n^\star) = g(X^\star) \quad \mathbb{P}^\star\text{-a.s.}$$

with

$$|g(X_n^\star(\omega^\star))| \leq B_g, \quad \begin{matrix} \omega^\star \in \Omega^\star \\ n = 1, 2, \ldots \end{matrix}$$

Invoking the Dominated Convergence Theorem we readily conclude that

$$\lim_{n\to\infty} \mathbb{E}^\star\left[g^\star(X_n^\star)\right] = \mathbb{E}^\star\left[g^\star(X^\star)\right].$$

This completes the proof of the validity of (ii). The proof that (ii) implies (i) is rather technical and is omitted; see [] for details.

The equivalence of (i) and (iii) is just Theorem 20.1.1. Note that (iii) is a simple consequence of (ii) since for every θ in $\mathbb{R}$ the mappings $x \to \cos(\theta x)$ and $x \to \sin(\theta x)$ are bounded and continuous on $\mathbb{R}$. ■

An immediate consequence of Theorem 20.4 is the following continuity result for weak convergence.

Theorem 20.3.2 *Consider the $\mathbb{R}$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. If the rvs $\{X_n,\ n = 1, 2, \ldots\}$ converge in distribution to the rv X, then the $\mathbb{R}$-valued rvs $\{h(X_n),\ n = 1, 2, \ldots\}$ converge in distribution to the rv $h(X)$ for any continuous mapping $h : \mathbb{R} \to \mathbb{R}$, namely*

$$h(X_n) \Longrightarrow_n h(X).$$

Proof. The proof follows by a simple application of Theorem 20.4: Pick a bounded continuous mapping $g : \mathbb{R} \to \mathbb{R}$. Given the continuous mapping $h : \mathbb{R} \to \mathbb{R}$, we note that the mapping $g \circ h : \mathbb{R} \to \mathbb{R}$ given by

$$g \circ h(x) = g(h(x)), \quad x \in \mathbb{R}$$

is also a bounded continuous mapping $\mathbb{R} \to \mathbb{R}$. Therefore, by Part (ii) of Theorem 20.4 we conclude from the assumed convergence $X_n \Longrightarrow_n X$ that

$$\lim_{n\to\infty} \mathbb{E}\left[g \circ h(X_n)\right] = \mathbb{E}\left[g \circ h(X)\right].$$

or equivalently,

$$\lim_{n\to\infty} \mathbb{E}\left[g(h(X_n))\right] = \mathbb{E}\left[g(h(X))\right].$$

Invoking one more time Part (ii) of Theorem 20.4 we now conclude that $h(X_n) \Longrightarrow_n h(X)$ as desired. ■

20.4 Weak convergence of discrete rvs

In this section we consider a collection of *discrete* rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ with

$$\mathbb{P}\left[X \in S\right] = \mathbb{P}\left[X_n \in S\right] = 1, \quad n = 1, 2, \ldots$$

where $S = \{a_i, i \in I\}$ is a countable subset of $\mathbb{Z}$.

Theorem 20.4.1 *The sequence of discrete rvs $X_n \Longrightarrow_n X$ converges weakly to the rv X if and only if*

$$\lim_{n\to\infty} \mathbb{P}\left[X_n = a_i\right] = \mathbb{P}\left[X = a_i\right], \quad i \in I.$$

Proof. Assume first that $X_n \Longrightarrow_n X$. Let a be a point of discontnuity for F_X. By assumption a is an element of $\mathbb{Z}$, and therefore ε can be selected in $(0, 1)$ so that both $a \pm \varepsilon$ are not in $\mathbb{Z}$ and are points of continuity for F_X. It follows that

$$(20.10) \qquad \lim_{n\to\infty} \mathbb{P}\left[X_n \le a \pm \varepsilon\right] = \mathbb{P}\left[X \le a \pm \varepsilon\right].$$

Note however that

$$(20.11) \quad \mathbb{P}\left[X_n \le a - \varepsilon\right] = \mathbb{P}\left[X_n \le a + \varepsilon\right] + \mathbb{P}\left[X_n = a\right], \quad n = 1, 2, \ldots$$

and

$$(20.12) \qquad \mathbb{P}\left[X \le a - \varepsilon\right] = \mathbb{P}\left[X \le a + \varepsilon\right] + \mathbb{P}\left[X = a\right].$$

since the probability distribution functions are piecewise constant with jumps only at points in $\mathbb{Z}$.

Let n go to infinity in (20.11). It is plain from (20.10) that $\lim_{n\to\infty} \mathbb{P}\left[X_n = a\right]$ exists and is given by

$$\lim_{n\to\infty} \mathbb{P}\left[X_n = a\right] = \mathbb{P}\left[X \le a + \varepsilon\right] - \mathbb{P}\left[X \le a - \varepsilon\right] = \mathbb{P}\left[X = a\right]$$

as we make use of (20.12). ■

Conversely, assume that

$$\lim_{n\to\infty} \mathbb{P}\left[X_n = a\right] = \mathbb{P}\left[X = a\right], \quad a \notin \mathcal{C}(F_X). \tag{20.13}$$

With any Borel subset B in $\mathbb{R}$, we shall show that

$$\lim_{n\to\infty} \mathbb{P}\left[X_n \in B\right] = \mathbb{P}\left[X \in B\right]. \tag{20.14}$$

This will immediately imply $X_n \Longrightarrow_n X$ upon specializing B to sets of the form $B = (-\infty, x]$ with x in $\mathcal{C}(F_X)$.

To establish (20.14), fix $n = 1, 2, \ldots$ and pick an arbitrary positive integer A: We see that

$$\begin{aligned} &\mathbb{P}\left[X_n \in B\right] \\ &\quad = \mathbb{P}\left[|X_n| \leq A, X_n \in B\right] + \mathbb{P}\left[|X_n| > A, X_n \in B\right] \\ &\quad = \sum\nolimits_{a\in\mathbb{Z}\cap B: |a| \leq A} \mathbb{P}\left[X_n = a\right] + \mathbb{P}\left[|X_n| > A, X_n \in B\right] \end{aligned} \tag{20.15}$$

while

$$\begin{aligned} \mathbb{P}\left[X \in B\right] &= \mathbb{P}\left[|X| \leq A, X \in B\right] + \mathbb{P}\left[|X| > A, X \in B\right] \\ &= \sum\nolimits_{a\in\mathbb{Z}\cap B: |a| \leq A} \mathbb{P}\left[X = a\right] + \mathbb{P}\left[|X| > A, X \in B\right]. \end{aligned}$$

Substracting we conclude that

$$\begin{aligned} &\left|\mathbb{P}\left[X_n \in B\right] - \mathbb{P}\left[X \in B\right]\right| \\ &\quad \leq \sum\nolimits_{a\in\mathbb{Z}\cap B: |a| \leq A} \left|\mathbb{P}\left[X_n = a\right] - \mathbb{P}\left[X = a\right]\right| + \mathbb{P}\left[|X_n| > A\right] + \mathbb{P}\left[|X| > A\right]. \end{aligned}$$

Let n go to infinity in this last inequality: Using (20.13) we get

$$\lim_{n\to\infty} \sum\nolimits_{a\in\mathbb{Z}\cap B: |a| \leq A} \left|\mathbb{P}\left[X_n = a\right] - \mathbb{P}\left[X = a\right]\right| = 0$$

since this sum has at most $2A + 1$ terms, while

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left[|X_n| > A\right] &= \lim_{n\to\infty} \left(1 - \mathbb{P}\left[|X_n| > A\right]\right) \\ &= 1 - \mathbb{P}\left[|X| \leq A\right] = \mathbb{P}\left[|X| \leq A\right] \end{aligned} \tag{20.16}$$

by a similar argument.

Collecting these facts we conclude that

$$\limsup_{n\to\infty} \left(\left| \mathbb{P}\left[X_n \in B\right] - \mathbb{P}\left[X \in B\right] \right| \right) \leq 2\mathbb{P}\left[|X| > A\right].$$

Now, letting A go to infinity in this last inequality, we note that

$$\limsup_{n\to\infty} \left| \mathbb{P}\left[X_n \in B\right] - \mathbb{P}\left[X \in B\right] \right| = 0$$

since the left handside does not depend on A. The desired conclusion (20.14) immediately follows. ■

In the more restrictive setting where $S \subseteq \mathbb{N}$, probability generating functions can be defined, and the following analog of Theorem 20.1.1 holds.

Theorem 20.4.2 *The sequence of* $\mathbb{N}$*-valued rvs* $\{X_n, n = 1, 2, \ldots\}$ *converges weakly to the rv* X *if and only if*

$$\lim_{n\to\infty} G_{X_n}(z) = G_X(z), \quad |z| \leq 1.$$

The sequence of $\mathbb{R}^p$-valued rvs $\{X_n,\ n = 1, 2, \ldots\}$ *converges in distribution* to the $\mathbb{R}^p$-valued rv X if for every bounded continuous mapping $g : \mathbb{R}^p \to \mathbb{R}$, it holds that

$$\lim_{n\to\infty} \mathbb{E}\left[g(X_n)\right] = \mathbb{E}\left[g(X)\right]. \tag{20.17}$$

Here as well we shall write $X_n \Longrightarrow_n X$ or $X_n \xrightarrow{\mathcal{L}}_n X$. Some authors also refer to this mode of convergence as *convergence in law* or as *weak convergence*.

Theorem 20.4 has the following multi-dimensional analog.

Theorem 20.4.3 *Consider the* $\mathbb{R}^p$*-valued rvs* $\{X, X_n,\ n = 1, 2, \ldots\}$ *defined on some probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*. Then, the rvs* $\{X_n,\ n = 1, 2, \ldots\}$ *converge in distribution to the rv* X *if and only if*

$$\lim_{n\to\infty} \Phi_{X_n}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}. \tag{20.18}$$

This amounts to

$$\lim_{n\to\infty} \mathbb{E}\left[e^{i\theta' X_n}\right] = \mathbb{E}\left[e^{i\theta' X}\right], \quad \theta \in \mathbb{R}.$$

In the same way that Theorem implied Theorem 20.3.2, we readily see that Theorem 20.4.3has the following important consequence.

Theorem 20.4.4 *Consider the $\mathbb{R}^p$-valued rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. If the rvs $\{X_n,\ n = 1, 2, \ldots\}$ converge in distribution to the rv X, then the $\mathbb{R}^q$-valued rvs $\{h(X_n),\ n = 1, 2, \ldots\}$ converge in distribution to the $\mathbb{R}^q$-valued rv $h(X)$ for any continuous mapping $h : \mathbb{R}^p \to \mathbb{R}^q$, namely*

$$h(X_n) \Longrightarrow_n h(X).$$

20.5 Convergence of Gaussian rvs

Gaussian rvs have a very compact characterization in terms of their characteristic functions. This can be used to show that the class of Gaussian distributions is stable under weak convergence in the following sense.

Lemma 20.5.1 *Let $\{X_k,\ k = 1, 2, \ldots\}$ denote a collection of $\mathbb{R}^p$-valued Gaussian rvs. For each $k = 1, 2, \ldots$, let μ_k and Σ_k denotes the mean vector and covariance matrix of the rv X_k. The rvs $\{X_k,\ k = 1, \ldots\}$ converge in distribution (in law) if and only there exist an element μ in $\mathbb{R}^p$ and a $p \times p$ matrix Σ such that*

$$\lim_{k\to\infty} \mu_k = \mu \quad \text{and} \quad \lim_{k\to\infty} \Sigma_k = \Sigma, \tag{20.19}$$

in which case, $X_k \Longrightarrow_k X$ where X is an $\mathbb{R}^d$-valued Gaussian rv $\mathrm{N}(\mu, \Sigma)$ with mean vector μ and covariance matrix Σ.

Proof. Assume first that the conditions (20.19) hold. Using the fact that

$$\Phi_{X_k}(\theta) = e^{i\theta^t \mu_k - \frac{1}{2}\theta^t \Sigma_k \theta}, \qquad \begin{array}{c} \theta \in \mathbb{R}^k \\ k = 1, 2, \ldots \end{array}$$

we note that

$$\lim_{k\to\infty} \Phi_{X_k}(\theta) = e^{i\theta^t \mu - \frac{1}{2}\theta^t \Sigma \theta}, \quad \theta \in \mathbb{R}$$

The second half of condition (20.19) ensures that the matrix Σ is symmetric and non-negative definite, hence a covariance matrix. Therefore, $\lim_{k\to\infty} \Phi_{X_k}$ is the characteristic function of a Gaussian rv X with $X \sim \mathrm{N}(\mu, \Sigma)$. Applying Theorem 20.1.1 we conclude that $X_k \Longrightarrow_k X$ where $X \sim \mathrm{N}(\mu, \Sigma)$.

Conversely, assume $X_k \Longrightarrow_k X$. Applying Theorem 20.1.1 again we conclude that

$$\lim_{k\to\infty} \Phi_{X_k}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}. \tag{20.20}$$

Using the decomposition into real and imaginary components yields

$$\Phi_{X_k}(\theta) = \cos\left(\theta^t \mu_k\right) e^{-\frac{1}{2}\theta^t \Sigma_k \theta} + i \sin\left(\theta^t \mu_k\right), \quad \begin{array}{c} \theta \in \mathbb{R}^p \\ k = 1, 2, \ldots \end{array}$$

It is plain that the convergence (20.20) is equivalent to the simultaneous validity of the two convergence statements

$$\lim_{k\to\infty} \cos\left(\theta^t \mu_k\right) e^{-\frac{1}{2}\theta^t \Sigma_k \theta} = \mathbb{E}\left[\cos\left(\theta^t X\right)\right], \quad \theta \in \mathbb{R}^p \tag{20.21}$$

and

$$\lim_{k\to\infty} \sin\left(\theta^t \mu_k\right) = \mathbb{E}\left[\sin\left(\theta^t X\right)\right], \quad \theta \in \mathbb{R}^p. \tag{20.22}$$

For any any convergent subsequence $\{\mu_{k_\ell}, \ell = 1, 2, \ldots\}$ with $\lim_{\ell\to\infty} \mu_{k_\ell} = \mu$ for some vector μ in $\mathbb{R}^p$, the convergence (20.22) yields

$$\sin\left(\theta^t \mu\right) = \lim_{\ell\to\infty} \sin\left(\theta^t \mu_{k_\ell}\right) = \mathbb{E}\left[\sin\left(\theta^t X\right)\right], \quad \theta \in \mathbb{R}^p.$$

Therefore, if $\mu^\star$ and $\mu_\star$ are accumulation points of the sequence $\{\mu_k, k = 1, 2, \ldots\}$ we must necessarily have

$$\sin\left(\theta^t \mu^\star\right) = \sin\left(\theta^t \mu_\star\right), \quad \theta \in \mathbb{R}^p$$

and the equality $\mu^\star = \mu_\star$ follows. Therefore, all accumulation points of the sequence $\{\mu_k,\ k = 1, 2, \ldots\}$ coincide and the sequence converges, say with limit μ in $\mathbb{R}^p$.

For any convergent subsequence $\{\Sigma_{k_\ell}, \ell = 1, 2, \ldots\}$ with $\lim_{\ell\to\infty} \Sigma_{k_\ell} = \Sigma$ for some $p \times p$ matrix Σ, the convergence (20.21) yields

$$\cos\left(\theta^t \mu\right) e^{-\frac{1}{2}\theta^t \Sigma \theta} = \lim_{\ell\to\infty} \cos\left(\theta^t \mu_{k_\ell}\right) e^{-\frac{1}{2}\theta^t \Sigma_{k_\ell} \theta} = \mathbb{E}\left[\cos\left(\theta^t X\right)\right], \quad \theta \in \mathbb{R}^p$$

and we conclude that

$$e^{-\frac{1}{2}\theta^t \Sigma \theta} = \frac{\mathbb{E}\left[\cos\left(\theta^t X\right)\right]}{\cos\left(\theta^t \mu\right)}, \quad \theta \in \mathbb{R}^p \quad \text{whenever } \cos\left(\theta^t \mu\right) \neq 0$$

As a result, all accumulation points of the sequence $\{\Sigma_k,\ k = 1, 2, \ldots\}$ coincide and the sequence converges, say with limit Σ, said limit being a $p \times p$ being necessarily a covariance matrix. ■

20.6 Exercises

Ex. 20.1 As in Exercise 19.9, consider the triangular array of rvs $\{X_{n,k},\ k = 1,\ldots,n;\ n = 1,2,\ldots\}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. For each $n = 1,2,\ldots$, we assume that the rvs $X_{n,1},\ldots,X_{n,n}$ are i.i.d. rvs with

$$\mathbb{P}\left[X_{n,k} = -\sqrt{n}\right] = \mathbb{P}\left[X_{n,k} = \sqrt{n}\right] = \frac{1}{2n}, \quad k = 1,\ldots,n$$

and

$$\mathbb{P}\left[X_{n,k} = 0\right] = 1 - \frac{1}{n}, \quad k = 1,\ldots,n.$$

We write

$$R_n \equiv \frac{S_n}{\sqrt{\mathrm{Var}\left[S_n\right]}} \quad \text{with} \quad S_n = \sum_{k=1}^{n} X_{n,k}, \quad n = 1,2,\ldots$$

Explore the weak convergence of the sequence $\{R_n,\ n = 1,2,\ldots\}$. In particular, identify a rv R such that $R_n \Longrightarrow_n R$ [**HINT:** What is the characteristic function of Poisson rvs?].

Ex. 20.2 Consider a sequence of $p \times p$ matrices $\{R_k,\ k = 1,2,\ldots\}$ such that for each $k = 1,2,\ldots$, the matrix R_k is a covariance matrix. Show that if $R \equiv \lim_{k\to\infty} R_k$ exists entrywise as a $p \times p$, then R is also a covariance matrix.

Chapter 21

The classical limit theorems

In this chapter we explore the classical limit theorem of Probability Theory. The setting of the next four sections is as follows: The rvs $\{X_n,\ n = 1, 2, \ldots\}$ are rvs defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. With this sequence we associate the sums

$$S_n = \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots$$

Two types of results will be discussed: The first class of results, known as Laws of Large Numbers, deal with the convergence of the sample averages

$$\bar{S}_n = \frac{1}{n} \sum_{k=1}^{n} X_k, \quad n = 1, 2, \ldots$$

The second class of results are called Central Limit Theorems and provide a rate of convergence in the Laws of Large Numbers.

21.1 Weak Laws of Large Numbers (I)

Laws of Large Numbers come in two types which are distinguished by the mode of convergence used. When convergence in probability is used, we refer to such results as Weak Laws of Large Numbers. The most basic such result is given next, and constitutes the Weak Laws of Large Numbers in its original version.

Theorem 21.1.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}\left[|X|^2\right] < \infty$. Then,*

$$\frac{S_n}{n} \xrightarrow{L^2}_n \mathbb{E}\left[X\right], \tag{21.1}$$

whence

$$\frac{S_n}{n} \xrightarrow{P}_n \mathbb{E}\left[X\right]. \tag{21.2}$$

Proof. For each $n = 1, 2, \ldots$, we have

$$\mathbb{E}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[X\right]\right|^2\right] = \mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \mathbb{E}\left[X\right]\right)\right|^2\right] = \frac{\mathrm{Var}[S_n]}{n^2}.$$

By the comments following Lemma 13.3.1 we conclude that $\mathrm{Var}[S_n] = n\mathrm{Var}[X]$ since

$$\mathrm{Cov}[X_k, X_\ell] = \delta(k;\ell)\mathrm{Var}[X], \quad k, \ell = 1, \ldots, n$$

under the enforced independence assumptions. As a result,

$$\mathbb{E}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[X\right]\right|^2\right] = \frac{\mathrm{Var}[X]}{n}$$

and the desired conclusion (21.1) is now immediate, with the convergence (21.2) following by Theorem 19.3.2. ■

21.2 Weak Laws of Large Numbers (II)

A careful inspection of the proof of Theorem 21.1.1 suggests a more general result. Assume that the rvs $\{X_n,\ n = 1, 2, \ldots\}$ are second-order rvs: For each $n = 1, 2, \ldots$, it is still the case that

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \mathbb{E}\left[X_k\right]\right)\right|^2\right] = \frac{\mathrm{Var}[S_n]}{n^2}.$$

Again making use of Lemma 13.3.1 we now obtain

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \mathbb{E}\left[X_k\right]\right)\right|^2\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[X_k] + \frac{1}{n^2}\sum_{k,\ell=1,\ k\neq\ell}\mathrm{Cov}[X_k, X_\ell].$$

Letting n go to infinity in this last relation we conclude to the next result – Here as well the passage from mean-square convergence to convergence in probability is validated by Theorem 19.3.2.

Proposition 21.2.1 *Consider a collection* $\{X_n,\ n=1,2,\ldots\}$ *of second-order rvs such that*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[X_k]=0. \tag{21.3}$$

We have

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\mathbb{E}\left[X_k\right]\right)\xrightarrow{L^2}_n 0 \tag{21.4}$$

and

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\mathbb{E}\left[X_k\right]\right)\xrightarrow{P}_n 0 \tag{21.5}$$

whenever either one of the following conditions holds:

(i) The rvs $\{X_n,\ n=1,2,\ldots\}$ *are uncorrelated*

(ii) The rvs $\{X_n,\ n=1,2,\ldots\}$ *are negatively correlated, i.e.,*

$$\mathrm{Cov}[X_k,X_\ell]\leq 0, \qquad \begin{array}{c} k\neq\ell \\ k,\ell=1,\ldots,n. \end{array}$$

(iii) The rvs $\{X_n,\ n=1,2,\ldots\}$ *satisfy the averaging condition*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{k,\ell=1,\ k\neq\ell}\mathrm{Cov}[X_k,X_\ell]=0. \tag{21.6}$$

Proof. In each case it suffices to show that

$$\lim_{n\to\infty}\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\mathbb{E}\left[X_k\right]\right)\right|^2\right]=0. \tag{21.7}$$

Case (i) is already contained in Case (ii) for which we have

$$0\leq\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\mathbb{E}\left[X_k\right]\right)\right|^2\right]\leq\frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}[X_k].$$

Letting n go to infinity in this chain of inequalities we get (21.7) by making use of the conditions (21.3).

In Case (iii) the limit (21.7) holds by virtue of (21.6) and (21.7). ■

This result is often applied when the rvs $\{X_n,\ n = 1, 2, \ldots\}$ have identical means and variances, namely there exist μ and $\sigma^2 > 0$ such that

$$\mathbb{E}\left[X_n\right] = \mu \quad \text{and} \quad \text{Var}[X_n] = \sigma^2, \quad n = 1, 2, \ldots$$

In that case, condition (21.3) is automatically satisfied and the convergence statements take the simpler form

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{L^2}_n \mu \quad \text{and} \quad \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P}_n \mu. \tag{21.8}$$

21.3 The classical Weak Law of Large Numbers (III)

As we now show, in Theorem 21.1.1 the finiteness of the second moments can be dropped while still insuring the result (21.2) under a finite first moment assumption. This is done by leveraging the equivalence between convergence in probability and weak convergence when the limit is a constant, thereby opening up the possibility to use methods based on characteristic functions.

However note that the mean-square convergence (21.1) is now obviously out of reach since none of the rvs involved may have finite second moments under the weaker first moment assumption.

Theorem 21.3.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}\left[|X|\right] < \infty$. Then, we have*

$$\frac{S_n}{n} \xrightarrow{P}_n \mathbb{E}\left[X\right]. \tag{21.9}$$

Proof. Fix $n = 1, 2, \ldots$ and θ in $\mathbb{R}$. Note that

$$\begin{aligned}
\mathbb{E}\left[e^{i\theta\left(\frac{S_n}{n} - \mathbb{E}[X]\right)}\right] &= \mathbb{E}\left[e^{i\frac{\theta}{n}\sum_{k=1}^n (X_k - \mathbb{E}[X])}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^{n} e^{i\frac{\theta}{n}(X_k - \mathbb{E}[X])}\right] \\
&= \prod_{k=1}^{n} \mathbb{E}\left[e^{i\frac{\theta}{n}(X_k - \mathbb{E}[X])}\right] \\
&= \left(\mathbb{E}\left[e^{i\frac{\theta}{n}(X - \mathbb{E}[X])}\right]\right)^n. \qquad (21.10)
\end{aligned}$$

Theorem 17.7.1 (for $k = 1$ and $x = X - \mathbb{E}[X]$) gives

$$e^{i\theta(X-\mathbb{E}[X])} = 1 + i\theta(X - \mathbb{E}[X]) + i\theta \cdot \int_0^{X-\mathbb{E}[X]} \left(e^{i\theta t} - 1\right) dt,$$

whence

$$\mathbb{E}\left[e^{i\theta(X-\mathbb{E}[X])}\right] = 1 + i\theta \cdot \mathbb{E}\left[\int_0^{X-\mathbb{E}[X]} \left(e^{i\theta t} - 1\right) dt\right] = 1 + i\theta \cdot C_1(\theta)$$

upon taking expectations with

$$C_1(\theta) \equiv \mathbb{E}\left[\int_0^{X-\mathbb{E}[X]} \left(e^{i\theta t} - 1\right) dt\right].$$

Substituting θ by $\frac{\theta}{n}$ in these relations leads to a rewriting of (21.10) as

$$\begin{aligned} \mathbb{E}\left[e^{i\theta\left(\frac{S_n}{n}-\mathbb{E}[X]\right)}\right] &= \left(\mathbb{E}\left[e^{i\frac{\theta}{n}(X-\mathbb{E}[X])}\right]\right)^n \\ &= \left(1 + \frac{i\theta}{n} \cdot C_1\left(\frac{\theta}{n}\right)\right)^n. \qquad (21.11) \end{aligned}$$

By Dominated Convergence, we conclude that $\lim_{n\to\infty} C_1\left(\frac{\theta}{n}\right) = 0$, whence

$$\lim_{n\to\infty} \left(\mathbb{E}\left[e^{i\theta\left(\frac{S_n}{n}-\mathbb{E}[X]\right)}\right]\right)^n = \lim_{n\to\infty} \left(1 + \frac{i\theta}{n} \cdot C_1\left(\frac{\theta}{n}\right)\right)^n = 1.$$

It follows that $\frac{S_n}{n} - \mathbb{E}[X] \xrightarrow{P}_n 0$, and this conclude the proof of (21.9). ■

21.4 The Strong Law of Large Numbers

Strong Laws of Large Numbers are convergence statements in the a.s. sense. The classical Strong Law of Large Numbers in its strongest form was proved by Kolmogorov.

Theorem 21.4.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}[|X|] < \infty$. Then,*

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}[X] \quad \text{a.s.} \qquad (21.12)$$

We give two proofs of this result under stronger assumptions on the moments of X. One proof assumes $\mathbb{E}\left[|X|^4\right] < \infty$ while the second proof is given under the condition $\mathbb{E}\left[|X|^2\right] < \infty$. A proof under the first moment condition $\mathbb{E}[|X|] < \infty$ is available in a number of references, see [**?**, **?**].

Proof 1 Assume $\mathbb{E}\left[|X|^4\right] < \infty$ – Note that there is no loss in generality in assuming that $\mathbb{E}[X] = 0$ as we do from now on in this proof. The basic idea of the proof is as follows: By the Monotone Convergence Theorem it is always the case that

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right] = \sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right]$$

Therefore, if we could show that

$$\sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] < \infty, \tag{21.13}$$

we immediately conclude that

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right] < \infty.$$

As a result,

$$\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4 < \infty \quad \text{a.s.}$$

and the conclusion $\lim_{n\to\infty}\frac{S_n}{n} = 0$ a.s. is now straightforward.

In order to establish (21.13) our starting point is the observation that

$$\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] = \frac{1}{n^4}\cdot\mathbb{E}\left[\left(\sum_{k=1}^{n}X_k\right)^4\right]$$

with

$$\mathbb{E}\left[\left(\sum_{k=1}^{n}X_k\right)^4\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathbb{E}\left[X_iX_jX_kX_\ell\right]. \tag{21.14}$$

Under the enforced independence assumptions it is plain (with $\mathbb{E}[X] = 0$) that $\mathbb{E}[X_iX_jX_kX_\ell] = 0$ as soon as one of the indices i, j, k, ℓ is different from all the other three, e.g., $i \notin \{j, k, \ell\}$, etc. The only cases when $\mathbb{E}[X_iX_jX_kX_\ell] \neq 0$ are as follows: (i) If $i = j = k = \ell$, then $\mathbb{E}[X_iX_jX_kX_\ell] = \mathbb{E}\left[X^4\right]$; there are n such configurations; (ii) If $\{i, j, k, \ell\}$ contains only two distinct values, say $a \neq b$ appearing as $aabb$, $abab$ and $abba$ in (21.14), then $\mathbb{E}[X_iX_jX_kX_\ell] = (\mathbb{E}\left[X^2\right])^2$; there are $3n(n-1)$ such configurations. It follows that

$$\mathbb{E}\left[\left(\sum_{k=1}^{n}X_k\right)^4\right] = n\mathbb{E}\left[X^4\right] + 3n(n-1)(\mathbb{E}\left[X^2\right])^2,$$

whence

$$\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] = \frac{1}{n^3}\cdot\mathbb{E}\left[X^4\right] + 3\frac{n-1}{n^3}\cdot(\mathbb{E}\left[X^2\right])^2 \sim \frac{3(\mathbb{E}\left[X^2\right])^2}{n^2}.$$

The conclusion (21.13) readily follows as we recall that $\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty$. This completes the proof. ■

Proof 2 Assume $\mathbb{E}\left[|X|^2\right] < \infty$ – For each $k = 1, 2, \ldots$, we note that

$$\text{Var}\left[\frac{S_{k^2}}{k^2}\right] = \frac{\text{Var}\left[X\right]}{k^2}$$

so that

$$\sum_{k=1}^{\infty}\mathbb{P}\left[\left|\frac{S_{k^2}}{k^2}\right| > \varepsilon\right] \le \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}\frac{\text{Var}\left[X\right]}{k^2} < \infty, \quad \varepsilon > 0.$$

It follows from Theorem 19.1.2 that

$$\lim_{k\to\infty}\frac{S_{k^2}}{k^2} = \mathbb{E}\left[X\right] \quad \text{a.s.} \tag{21.15}$$

Now assume that the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ are non-negative, i.e., $X \ge 0$ a.s. (in which case obviously $\mathbb{E}\left[X\right] \ge 0$). The case when the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ are non-positive, i.e., $X \le 0$ a.s., can be handed *mutatis mutandis*.

Fix $n = 1, 2, \ldots$. There exists a unique positive integer $k(n)$ such that

$$k(n)^2 \le n < (k(n)+1)^2. \tag{21.16}$$

Under the non-negativity assumption, we have $X_\ell \ge 0$ a.s. for $\ell = k(n)^2, \ldots, (k(n)+1)^2 - 1$, and the inequalities

$$S_{k(n)^2} \le S_n \le S_{(k(n)+1)^2} \quad \text{a.s.}$$

hold. It follows that

$$\frac{k(n)^2}{n}\cdot\left(\frac{S_{k(n)^2}}{k(n)^2}\right) \le \frac{S_n}{n} \le \frac{(k(n)+1)^2}{n}\cdot\left(\frac{S_{(k(n)+1)^2}}{(k(n)+1)^2}\right). \tag{21.17}$$

Using (21.16) we readily get

$$\frac{k(n)^2}{n} \le 1 < \frac{k(n)^2}{n} + 2\cdot\frac{k(n)}{\sqrt{n}}\cdot\frac{1}{\sqrt{n}} + \frac{1}{n}, \quad n = 1, 2, \ldots \tag{21.18}$$

From the first inequality in (21.18) we conclude that $\limsup_{n\to\infty} \frac{k(n)^2}{n} \leq 1$, while the second inequality leads to $1 \leq \liminf_{n\to\infty} \frac{k(n)^2}{n}$ since $\frac{k(n)}{\sqrt{n}} \leq 1$ for all $n = 1, 2, \ldots$. As a result, $\lim_{n\to\infty} \frac{k(n)^2}{n} = 1$ (whence $\lim_{n\to\infty} k(n) = \infty$ as expected). Finally let n go to infinity in (21.17): We readily get (21.12) upon combining this last conclusion with the convergence (21.15).

To complete the proof note that $\mathbb{E}\left[(X^\pm)^2\right] < \infty$ since $\mathbb{E}\left[|X|^2\right] = \mathbb{E}\left[(X^+)^2\right] + \mathbb{E}\left[(X^-)^2\right]$ (as we note that $X^+X^- = 0$). Thus, it holds that

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n X_k^\pm}{n} = \mathbb{E}\left[X^\pm\right] \quad \text{a.s.} \tag{21.19}$$

since the rvs $\{X^\pm, X_k^\pm,\ k = 1, 2, \ldots\}$ form an i.i.d. sequence of second-order rvs. The desired result (21.12) automatically follows upon noting that

$$X_n = X_n^+ - X_n^-, \quad n = 1, 2, \ldots$$

and $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. ■

21.5 The Central Limit Theorem

The Central Limit Theorem complements the Law of Large Numbers, in that it provides some indication as to the rate at which convergence takes place.

Theorem 21.5.1 *Assume the rvs $\{X, X_n,\ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E}\left[|X|^2\right] < \infty$. Then, we have*

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}[X]\right) \Longrightarrow_n \sqrt{\mathrm{Var}[X]} \cdot U \tag{21.20}$$

where U is a standard zero-mean unit-variance Gaussian rv.

Proof. Fix $n = 1, 2, \ldots$ and θ in $\mathbb{R}$. This time, as in the proof of Theorem 21.3.1 we get

$$\mathbb{E}\left[e^{i\theta\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}[X]\right)}\right] = \left(\mathbb{E}\left[e^{i\frac{\theta}{\sqrt{n}}(X - \mathbb{E}[X])}\right]\right)^n$$

under the enforced independence.

This time Theorem 17.7.1 (with $k = 2$ and $x = X - \mathbb{E}[X]$) yields

$$\begin{aligned} & e^{i\theta(X-\mathbb{E}[X])} \\ &= 1 + i\theta(X - \mathbb{E}[X]) - \frac{\theta^2}{2}(X - \mathbb{E}[X])^2 \\ & -\frac{\theta^2}{2}\int_0^{X-\mathbb{E}[X]} (X - \mathbb{E}[X] - t)\left(e^{i\theta t} - 1\right) dt, \end{aligned} \tag{21.21}$$

Taking expectations we get

$$\mathbb{E}\left[e^{i\theta(X-\mathbb{E}[X])}\right] = 1 - \frac{\theta^2}{2}\cdot \mathrm{Var}[X] - \frac{\theta^2}{2}\cdot C_2(\theta) \tag{21.22}$$

with

$$C_2(\theta) \equiv \mathbb{E}\left[\int_0^{X-\mathbb{E}[X]} (X - \mathbb{E}[X] - t)\left(e^{i\theta t} - 1\right) dt\right]. \tag{21.23}$$

Substituting θ by $\frac{\theta}{\sqrt{n}}$ in this last relation leads to

$$\mathbb{E}\left[e^{i\frac{\theta}{\sqrt{n}}(X-\mathbb{E}[X])}\right] = 1 - \frac{\theta^2}{2n}\cdot \mathrm{Var}[X] - \frac{\theta^2}{2n}\cdot C_2\left(\frac{\theta}{\sqrt{n}}\right)$$

so that

$$\mathbb{E}\left[e^{i\theta\sqrt{n}\left(\frac{S_n}{n}-\mathbb{E}[X]\right)}\right] = \left(1 - \frac{\theta^2}{2n}\cdot \mathrm{Var}[X] - \frac{\theta^2}{2n}\cdot C_2\left(\frac{\theta}{\sqrt{n}}\right)\right)^n.$$

Again, by Dominated Convergence, we obtain

$$\lim_{n\to\infty} C_2\left(\frac{\theta}{\sqrt{n}}\right) = 0$$

under the second moment condition $\mathbb{E}\left[|X|^2\right] < \infty$, whence

$$\lim_{n\to\infty} n\left(\frac{\theta^2}{2n}\cdot \mathrm{Var}[X] - \frac{\theta^2}{2n}\cdot C_2\left(\frac{\theta}{\sqrt{n}}\right)\right) = \frac{\theta^2}{2}\cdot \mathrm{Var}[X]$$

It follows that

$$\lim_{n\to\infty} \mathbb{E}\left[e^{i\theta\sqrt{n}\left(\frac{S_n}{n}-\mathbb{E}[X]\right)}\right] = e^{-\frac{\theta^2}{2}\cdot \mathrm{Var}[X]}$$

This complete the proof of (21.20). ■

21.6 The Central Limit Theorem – An application

We are still in the setting of Theorem 21.5.1. We can rephrase (21.20) as

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\left[\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[X\right]\right) \le x\right] \\ &= \mathbb{P}\left[\sqrt{\mathrm{Var}[X]}\cdot U \le x\right], \quad x \in \mathbb{R}. \qquad (21.24)\end{aligned}$$

as we recall that every point in $\mathbb{R}$ is a point of continuity for the rv U (or $\sqrt{\mathrm{Var}[X]}\cdot U$).

It follows that

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\left[\left|\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[X\right]\right)\right| \le x\right] \\ &= \mathbb{P}\left[\sqrt{\mathrm{Var}[X]}\cdot U \le x\right] - \mathbb{P}\left[\sqrt{\mathrm{Var}[X]}\cdot U \le -x\right] \\ &= \Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right) - \Phi\left(-\frac{x}{\sqrt{\mathrm{Var}[X]}}\right) \\ &= 2\Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right) - 1, \quad x \ge 0. \qquad (21.25)\end{aligned}$$

Fix $x \ge 0$ and $n = 1, 2, \ldots$: We have

$$\left|\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[X\right]\right)\right| \le x$$

if and only if

$$-x \le \sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\left[X\right]\right) \le x$$

if and only if

$$\mathbb{E}\left[X\right] \in \left[\frac{S_n}{n} - \frac{x}{\sqrt{n}}, \frac{S_n}{n} + \frac{x}{\sqrt{n}}\right].$$

Thus, if we think of

$$\widehat{X}_n = \frac{S_n}{n}, \quad n = 1, 2, \ldots$$

as an estimate of $\mathbb{E}\left[X\right]$ on the basis of the observations $X_1, \ldots, X_n$, then the SLLNs already tells us that the estimate is increasingly accurate as n gets large since

$$\lim_{n\to\infty} \widehat{X}_n = \mathbb{E}\left[X\right] \quad a.s.$$

The calculations above show via (21.25) that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left[\mathbb{E}\left[X\right]\in\left[\widehat{X}_n-\frac{x}{\sqrt{n}},\widehat{X}_n+\frac{x}{\sqrt{n}}\right]\right]\\ = \quad 2\Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right)-1,\quad x\geq 0.\end{aligned}\tag{21.26}$$

In other words, for large n, the *unknown* value $\mathbb{E}\left[X\right]$ lies in a symmetric interval centered at the estimate $\widehat{X}_n$ (obtained from the *observed* data $X_1,\ldots,X_n$) of width $\frac{2x}{\sqrt{n}}$ with a probability approximately given by

$$2\Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right)-1,$$

the accuracy of this approximation improving with increasing $n.$ With α in $(0,1)$ *given*, we can ensure that

$$\mathbb{P}\left[\mathbb{E}\left[X\right]\in\left[\widehat{X}_n-\frac{x}{\sqrt{n}},\widehat{X}_n+\frac{x}{\sqrt{n}}\right]\right]\simeq 1-\alpha$$

for large n if we select $x\geq 0$ such that

$$2\Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right)-1=1-\alpha,$$

or equivalently,

$$\Phi\left(\frac{x}{\sqrt{\mathrm{Var}[X]}}\right)=1-\frac{\alpha}{2}.$$

With λ in $(0,1)$ let z_λ denote the unique solution to the nonlinear equation

$$1-\Phi(x)=\lambda,\quad x\in\mathbb{R}.$$

Equivalently,

$$\mathbb{P}\left[U>x\right]=\lambda,\quad x\in\mathbb{R}.$$

With this notation we see that the *random* interval

$$\left[\frac{S_n}{n}-\frac{z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}[X]}}{\sqrt{n}},\frac{S_n}{n}+\frac{z_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}[X]}}{\sqrt{n}}\right]$$

is known as the *confidence interval* for estimating $\mathbb{E}\left[X\right]$ on the basis data $X_1,\ldots,X_n$ with confidence $(1-\alpha)\%$

Note that this analysis is predicated on knowing the variance $\mathrm{Var}[X]$. When this value is unknown, we replace $\mathrm{Var}[X]$ by the *sample variance* S_n^2 given by

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^n \left(X_k - \frac{1}{n}\sum_{\ell=1}^n X_\ell \right)^2, \quad n = 2, 3, \ldots$$

21.7 Poisson convergence

The setting is a follows: For each $n = 1, 2, \ldots$, let $X_1(p_n), \ldots, X_n(p_n)$ denote a collection of i.i.d. Bernoulli rvs with parameters p_n in $(0, 1)$. i.e.,

$$\mathbb{P}\left[X_{k,n}(p_n) = 1\right] = 1 - \mathbb{P}\left[X_{k,n}(p_n) = 0\right] = p_n, \quad k = 1, \ldots, n$$

Write

$$S_n = \sum_{k=1}^n X_k(p_n), \quad n = 1, 2, \ldots$$

Theorem 21.7.1 *Assume there exists $\lambda > 0$ such that*

$$\lim_{n\to\infty} np_n = \lambda. \tag{21.27}$$

Then, we have

$$S_n \Longrightarrow_n \Pi(\lambda) \tag{21.28}$$

where $\Pi(\lambda)$ denotes a Poisson rv with parameter λ.

The convergence (21.28) can be restated as

$$\lim_{n\to\infty} \mathbb{P}\left[S_n = k\right] = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, \ldots \tag{21.29}$$

We give two proofs of this important result.

Proof 1 The first proof uses the characterization of weak convergence for integer-valued rvs given in Theorem 20.4.1: Fix $n = 1, 2, \ldots$. Under the independence assumptions, the rv S_n is a binomial rv $\mathrm{Bin}(n; p_n)$. Thus, Fix $k = 0, 1, \ldots$. For every integer n such that $k \leq n$ we have

$$\begin{aligned} \mathbb{P}\left[S_n = k\right] &= \binom{n}{k} p_n^k \left(1 - p_n\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \cdot p_n^k \left(1 - p_n\right)^{n-k} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{k!}\left(\frac{p_n}{1-p_n}\right)^k \cdot \frac{n!}{(n-k)!} \cdot (1-p_n)^n \\
&= \frac{1}{k!}\left(\frac{np_n}{1-p_n}\right)^k \cdot \frac{n!}{n^k(n-k)!} \cdot (1-p_n)^n . \qquad (21.30)
\end{aligned}
$$

It is plain that

$$\lim_{n\to\infty} \frac{n!}{n^k(n-k)!} = \lim_{n\to\infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} = 1$$

while (21.27) implies

$$\lim_{n\to\infty} (1-p_n)^n = \lim_{n\to\infty} \left(1 - \frac{np_n}{n}\right)^n = e^{-\lambda}$$

and

$$\lim_{n\to\infty} \frac{p_n}{1-p_n} = \lambda$$

since $\lim_{n\to\infty} p_n = 0$. Collecting we conclude to (21.29) as we make use of Theorem 20.4.1. ■

Proof 2 This second proof relies on the characterization of weak convergence for integer-valued rvs given in terms of probability generating functions: Fix $n = 1, 2, \ldots$. For each θ in $\mathbb{R}$ we get

$$
\begin{aligned}
\mathbb{E}\left[e^{i\theta S_n}\right] &= \mathbb{E}\left[e^{i\theta \sum_{k=1}^n X_k(p_n)}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^n e^{i\theta X_k(p_n)}\right] \\
&= \prod_{k=1}^n \mathbb{E}\left[e^{i\theta X_k(p_n)}\right] \\
&= \left(1 - p_n + p_n e^{i\theta}\right)^n \\
&= \left(1 - p_n\left(1 - e^{i\theta}\right)\right)^n . \qquad (21.31)
\end{aligned}
$$

Under (21.27) we get that

$$\lim_{n\to\infty} np_n \left(1 - e^{i\theta}\right) = \lambda\left(1 - e^{i\theta}\right).$$

Thus,

$$\lim_{n\to\infty}\mathbb{E}\left[e^{i\theta S_n}\right] = e^{-\lambda(1-e^{i\theta})}, \quad \theta\in\mathbb{R}$$

and the conclusion (21.28) follows since

$$\begin{aligned}\mathbb{E}\left[e^{i\theta\Pi(\lambda)}\right] &= \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}\cdot e^{ik\theta} \\ &= \left(\sum_{k=0}^{\infty}\frac{1}{k!}\left(\lambda e^{i\theta}\right)^k\right)e^{-\lambda} = e^{-\lambda(1-e^{i\theta})}, \quad \theta\in\mathbb{R} \qquad (21.32)\end{aligned}$$

as we use Theorem 20.4.2. ■

Chapter 22

Appendix A: Limits in $\mathbb{R}$

We begin with several standard definitions. We refer to a mapping $a : \mathbb{N}_0 \to \mathbb{R}$ as a ($\mathbb{R}$-valued) sequence; sometimes we also use the notation $\{a_n,\ n = 1, 2, \ldots\}$.

Definition 22.0.1

A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ *converges* to $a^\star$ in $\mathbb{R}$ if for every $\varepsilon > 0$, there exists an integer $n^\star(\varepsilon)$ (which depends on ε) such that

$$|a_n - a^\star| \leq \varepsilon, \quad n \geq n^\star(\varepsilon). \tag{22.1}$$

We shall write $\lim_{n\to\infty} a_n = a^\star$, and refer to the scalar $a^\star$ as the *limit* of the sequence.

Sometimes it is desirable to make sense of situations where values of the sequence become either unboundedly large or unboundely negative, in which case we shall write $\lim_{n\to\infty} a_n = +\infty$ and $\lim_{n\to\infty} a_n = -\infty$, respectively. A precise definition of such occurences is as follows: We write $\lim_{n\to\infty} a_n = \infty$ to signify that for every $M > 0$, there exists a finite integer $n^\star(M)$ (which depends on M) in $\mathbb{N}_0$ such that

$$a_n > M, \quad n \geq n^\star(M). \tag{22.2}$$

It is natural to define $\lim_{n\to\infty} a_n = -\infty$ if $\lim_{n\to\infty} (-a_n) = \infty$.

If there exists $a^\star$ in $\overline{\mathbb{R}}$ such that $\lim_{n\to\infty} a_n = a^\star$, we shall simply say that the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ *converges* or *is convergent* (without any reference to its limit). Sometimes we shall also say that the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges *in* $\mathbb{R}$ to indicate that the limit $a^\star$ is an element of $\mathbb{R}$ (thus finite).

Applying the definition (22.1) requires that the limit be *known*. Often this information is not available, and yet the need remains to determine whether the sequence converges. The notion of *Cauchy sequence*, which is instrumental in that respect, is built around the following observation: If the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges to $a^\star$ in $\mathbb{R}$, then for every $\varepsilon > 0$, there exists a finite integer $n^\star(\varepsilon)$ such that (22.1) holds, and by the triangular inequality we conclude that

$$|a_n - a_m| \leq |a_n - a^\star| + |a^\star - a_m| \leq \varepsilon + \varepsilon = 2\varepsilon, \quad n, m \geq n^\star(\varepsilon).$$

This observation is turned into the following definition.

Definition 22.0.2

A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is said to be a *Cauchy sequence* if for every $\varepsilon > 0$, there exists an integer $n^\star(\varepsilon)$ such that

$$|a_n - a_m| \leq \varepsilon, \quad m, n \geq n^\star(\varepsilon). \tag{22.3}$$

As observed earlier, a convergent sequence $a : \mathbb{N}_0 \to \mathbb{R}$ *in* $\mathbb{R}$ is always a Cauchy sequence. It is a deep fact concerning the topological properties of $\mathbb{R}$ that being a Cauchy sequence is sufficient to ensure the convergence of the sequence in $\mathbb{R}$.

Theorem 22.0.1 *(Cauchy criterion) A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is convergent in $\mathbb{R}$ if and only if it is a Cauchy sequence.*

This provides a convergence criterion which does *not* require knowledge of the limit.

22.1 Two important facts

In addition to the Cauchy convergence criterion, here are two facts that are often found useful in studying convergence, namely *monotonicy* and *boundedness*.

Definition 22.1.1

A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is said to be *non-decreasing* (resp. *non-increasing*) if

$$a_n \leq a_{n+1} \quad \text{(resp. } a_{n+1} \leq a_n), \quad n = 1, 2, \ldots$$

A *monotone* sequence is a sequence that is either non-decreasing or non-increasing.

Convergence is automatically guaranteed for monotone sequences.

Theorem 22.1.1 *A monotone sequence* $a : \mathbb{N}_0 \to \mathbb{R}$ *always converges and we have* $\lim_{n\to\infty} a_n = \sup(a_n,\ n = 1, 2, \ldots)$ *(resp.* $\lim_{n\to\infty} a_n = \inf(a_n,\ n = 1, 2, \ldots)$*) if the sequence is non-decreasing (resp. non-increasing).*

A convergent sequence in $\mathbb{R}$ is always bounded in the following sense.

Definition 22.1.2

The sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is said to be *bounded* if there exists some $B > 0$ such that

$$\sup(|a_n|,\ n = 1, 2, \ldots) \leq B.$$

While a bounded sequence may not be convergent, some of the *subsequences* obtained by sampling the original sequence are convergent in $\mathbb{R}$: Consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$. A *subsequence* of the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is any sequence of the form $\mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k}$ where

$$n_k < n_{k+1}, \quad k = 1, 2, \ldots$$

The strict inequality forces $\lim_{k\to\infty} n_k = \infty$.

Theorem 22.1.2 *(Bolzano-Weierstrass) For any bounded sequence* $a : \mathbb{N}_0 \to \mathbb{R}$*, there exists a convergent subsequence* $\mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k}$ *with* $\lim_{k\to\infty} a_{n_k} = a^\star$ *for some* $a^\star$ *in* $\mathbb{R}$*.*

22.2 Accumulation points

Since not all sequences converge, it is important to understand how non-convergence occurs.

Definition 22.2.1

An *accumulation point* for the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is defined as any element $a^\star$ in $\overline{\mathbb{R}}$ such that

$$\lim_{k\to\infty} a_{n_k} = a^\star$$

for *some* subsequence $\mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k}$.

Obviously a convergent sequence $a : \mathbb{N}_0 \to \mathbb{R}$ has exactly *one* accumulation point, namely its limit. In fact, were the sequence *not* convergent, it must necessarily have distinct accumulation points (in $\overline{\mathbb{R}}$), in which case there is a smallest and a largest accumulation point. The next definition formalizes this observation.

Definition 22.2.2

Given a sequence $a : \mathbb{N}_0 \to \mathbb{R}$, the quantities

$$\overline{A} = \limsup_{n\to\infty} A_n = \inf_{n\geq 1}\left(\sup_{m\geq n} a_m\right)$$

and

$$\underline{A} = \liminf_{n\to\infty} A_n, = \sup_{n\geq 1}\left(\inf_{m\geq n} a_m\right)$$

are known as the *limsup* and *liminf* of the sequence $a : \mathbb{N}_0 \to \mathbb{R}$.

The following notation is found to be convenient when using liminf and limsup quantities: For each $n = 1, 2, \ldots$, we define the quantities

$$\overline{A}_n = \sup_{m\geq n} a_m \quad \text{and} \quad \underline{A}_n = \inf_{m\geq n} a_m \tag{22.4}$$

Note that $\underline{A}_n \leq \overline{A}_n$, and that the sequences $n \to \overline{A}_n$ and $n \to \underline{A}_n$ are non-increasing and non-decreasing, respectively. Therefore, by Theorem 22.1.2 the limits $\overline{A} = \lim_{n\to\infty} \overline{A}_n$ and $\underline{A} = \lim_{n\to\infty} \underline{A}_n$ both exist, but are possibly infinite, and we always have $\underline{A} \leq \overline{A}$.

Theorem 22.2.1 *Consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$. If it converges to $a^\star$, then $\overline{A} = \underline{A} = a^\star$. Conversely, if $\overline{A} = \underline{A} = a^\star$ for some $a^\star$ in $\overline{\mathbb{R}}$, then the sequence converges to $a^\star$.*

If $a, b : \mathbb{N}_0 \to \mathbb{R}$ are two sequences such that

$$a_n \leq b_n, \quad n = 1, 2, \ldots$$

then $\overline{A} \leq \overline{B}$ and $\underline{A} \leq \underline{B}$. The following arguments will often be made on the basis of this observation: Consider a sequence $\{p_n,\ n = 1, 2, \ldots\}$ where for each $n = 1, 2, \ldots$, p_n is the probability of some event so that

$$0 \leq p_n \leq 1, \quad n = 1, 2, \ldots \tag{22.5}$$

If we show that

$$1 \leq \liminf_{n\to\infty} p_n, \tag{22.6}$$

then we necessarily have convergence of the sequence with $\lim_{n\to\infty} p_n = 1$: Indeed, we always have $\limsup_{n\to\infty} p_n \leq 1$ as a result of (22.5), whence

$$\liminf_{n\to\infty} p_n = \limsup_{n\to\infty} p_n = 1$$

upon using (22.6). In a similar vein, if we show $\limsup_{n\to\infty} p_n = 0$, then we necessarily have convergence of the sequence with $\lim_{n\to\infty} p_n = 0$.

22.3 Cesàro convergence

With any sequence $a : \mathbb{N}_0 \to \mathbb{R}$ we associate the *Cesàro* sequence $a^c : \mathbb{N}_0 \to \mathbb{R}$ given by

$$a_n^c = \frac{1}{n}\left(a_1 + \ldots + a_n\right), \quad n = 1, 2, \ldots$$

Theorem 22.3.1 *(Cesàro convergence) If the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges to $a^\star$, then the Cesàro sequence $a^c : \mathbb{N}_0 \to \mathbb{R}$ also converges with same limit $a^\star$.*

The convergence of the sequence $\{a_n^c,\ n = 1, 2, \ldots\}$ is referred to as the Cesàro convergence of the sequence $\{a_n,\ n = 1, 2, \ldots\}$

Proof. First we assume the convergent sequence $a : \mathbb{N}_0 \to \mathbb{R}$ to have a finite limit $a^\star$ in $\mathbb{R}$. Note that

$$a_n^c - a^\star = \frac{1}{n}\sum_{k=1}^n \left(a_k - a^\star\right), \quad n = 1, 2, \ldots$$

Now, for every $\varepsilon > 0$, there exists an integer $n^\star(\varepsilon)$ such that $|a_n - a^\star| \leq \varepsilon$ whenever $n \geq n^\star(\varepsilon)$. On that range, with $B(\varepsilon) = \sum_{k=1}^{n^\star(\varepsilon)} |a_k - a^\star|$, we have

$$\begin{aligned}
|a_n^c - a^\star| &\leq \frac{1}{n}\sum_{k=1}^n |a_k - a^\star| \\
&= \frac{1}{n}\sum_{k=1}^{n^\star(\varepsilon)} |a_k - a^\star| + \frac{1}{n}\sum_{k=n^\star(\varepsilon)+1}^{n} |a_k - a^\star| \\
&\leq \frac{B(\varepsilon)}{n} + \frac{n - n^\star(\varepsilon)}{n}\cdot\varepsilon \\
&\leq \frac{B(\varepsilon)}{n} + \varepsilon \qquad (22.7)
\end{aligned}$$

Since $\lim_{n\to\infty} \frac{1}{n} = 0$, for every $\varepsilon > 0$, there exists a finite integer $n^{\star\star}(\varepsilon)$ such that

$$\frac{1}{n} < \frac{\varepsilon}{B(\varepsilon)}, \quad n \geq n^{\star\star}(\varepsilon).$$

Just take $n^{\star\star}(\varepsilon) = \lceil \frac{B(\varepsilon)}{\varepsilon} \rceil$. As a result, we have $|a_n^c - a^\star| \leq \varepsilon + \varepsilon = 2\varepsilon$ whenever $n \geq \max(n^\star(\varepsilon), n^{\star\star}(\varepsilon))$, and the proof is now complete since ε is arbitrary. We leave it as an exercise to show the result when $a^\star = \pm\infty$. ■

However, the converse is not true:

Counterexample 22.3.1 The sequence $a : \mathbb{N}_0 \to \mathbb{R}$ given by $a_n = (-1)^n$ for each $n = 1, 2, \ldots$ does not converge since $\liminf_{n\to\infty} a_n = -1$ and $\limsup_{n\to\infty} a_n = 1$. Yet $\lim_{n\to\infty} a_n^c = 0$ since

$$a_n^c = \begin{cases} 0 & \text{if } n = 2p \\ & . \\ -\frac{1}{2p-1} & \text{if } n = 2p-1 \end{cases}, \quad p = 1, 2, \ldots$$

This example nicely illustrates the smoothing effect of averaging. It might be tempting to conjecture that such averaging always produces a convergent sequence. However, this is not so as the following example shows:

Counterexample 22.3.2 Consider the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ given by

$$a_n = (-1)^k, \quad \begin{matrix} 2^{2^k} \leq n < 2^{2^{k+1}} \\ k = 0, 1, \ldots \end{matrix}$$

with $a_1 = 1$. Having two distinct accumulation points, namely ± 1, the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ does not converge. However, it is also not Cesàro convergent.

Chapter 23

Appendix A: Sums, series and summation

23.1 Series

Starting with a sequence $a : \mathbb{N}_0 \to \mathbb{R}$, we define the partial sums

$$s_n = a_1 + \ldots + a_n, \quad n = 1, 2, \ldots$$

where s_n is known as the n^{th} *partial sum*. We refer to the sequence $s : \mathbb{N}_0 \to \mathbb{R} : n \to s_n$ as the sequence of partial sums associated with the sequence $a : \mathbb{N}_0 \to \mathbb{R}$.

Definition 23.1.1

The series $\sum_{n=1}^{\infty} a_n$ is said to converge (or to be *summable*) if the sequence $s : \mathbb{N}_0 \to \mathbb{R}$ converges to some $s^\star$ in $\mathbb{R}$, in which case we write $\sum_{n=1}^{\infty} a_n$ as its limit (and refer to $s^\star$ as its *sum*).

Summability amounts to the following: For every $\varepsilon > 0$ there exists a finite integer $n^\star(\varepsilon)$ such that $|s_n - s^\star| < \varepsilon$ whenever $n \geq n^\star(\varepsilon)$. This readily implies the following fact:

Lemma 23.1.1 *For any sequence $a : \mathbb{N}_0 \to \mathbb{R}$ whose sequence of partial sums converges in $\mathbb{R}$, we have* $\lim_{n\to\infty} a_n = 0$

Proof. Since the sequence of partial sums $s : \mathbb{N}_0 \to \mathbb{R}$ converges in $\mathbb{R}$, it is a Cauchy sequence. Thus, for every $\varepsilon > 0$, there exists a finite integer $n^\star(\varepsilon)$

such that $|s_n - s_m| \leq \varepsilon$ whenever $n, m \geq n^\star(\varepsilon)$. Selecting $m = n + 1$ with $n \geq n^\star(\varepsilon)$, we get $|a_{n+1}| = |s_n - s_{n+1}| \leq \varepsilon$ whenever $n \geq n^\star(\varepsilon)$, and the conclusion $\lim_{n\to\infty} a_n = 0$ follows. ■

The following stronger form of convergence is often invoked for series.

Definition 23.1.2

The series $s : \mathbb{N}_0 \to \mathbb{R}$ associated with the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ is said to be *absolutely convergent* if the series associated with the sequence of absolute values $\mathbb{N}_0 \to \mathbb{R}_+ : n \to |a_n|$ does itself converge in $\mathbb{R}$.

A series which is absolutely convergent is also convergent in the usual sense: Indeed, note that

$$\left| \sum_{k=n+1}^{m} a_k \right| \leq \sum_{k=n+1}^{m} |a_k| , \qquad \begin{array}{c} m = n+1, \ldots \\ n = 1, 2, \ldots \end{array}$$

and apply the Cauchy convergencce criterion. However, the converse is not true as is easily seen through the example

$$a_n = \frac{(-1)^n}{n}, \quad n = 1, 2, \ldots$$

Definition 23.1.3

A series which is convergent in the usual sense but not absolutely convergent is said to be *conditionally* convergent.

When the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ assumes only non-negative values, i.e., $a_n \geq 0$ for all $n = 1, 2, \ldots$, then the corresponding the sequence $s : \mathbb{N}_0 \to \mathbb{R}_+$ of partial sums is non-decreasing, so that $\lim_{n\to\infty} s_n$ always exists, possibly infinite. Many tests exist to check the convergence of series with non-negative terms The most basic one is the Comparison Test given next.

Theorem 23.1.1 *(Comparison Test) Consider two sequences $a, b : \mathbb{N}_0 \to \mathbb{R}_+$ such that*

$$0 \leq a_n \leq b_n, \quad n = 1, 2, \ldots$$

If $\sum_{n=1}^\infty b_n$ converges in $\mathbb{R}$, then $\sum_{n=1}^\infty a_n$ also converges in $\mathbb{R}$ with

$$0 \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} b_n.$$

On the other hand, if $\sum_{n=1}^\infty a_n = \infty$, then we necessarily have $\sum_{n=1}^\infty b_n = \infty$.

Geometric series play a pivotal role in determining the convergence of series through the Comparison Test. The *geometric* series with reason $\rho > 0$ is the series associated with the sequence $a : \mathbb{N}_0 \to \mathbb{R}$ given by

$$a_n = \rho^n, \quad n = 1, 2, \ldots$$

It well known that

$$s_n = a_1 + \ldots + a_n = \begin{cases} \frac{\rho}{1-\rho}\left(1 - \rho^n\right) & \text{if } \rho \neq 1 \\ \\ \quad n & \text{if } \rho = 1. \end{cases}$$

Therefore,

$$\lim_{n\to\infty} s_n = \frac{\rho}{1-\rho} \quad \text{if } |\rho| < 1.$$

When coupled with the Comparisdion Test of Theorem 23.1.1 this observation constitutes the basis for two criteria to dteermine the absolute convergence of series, namely the criteria of Cauchy and d' Alembert, also known as the Root Test and Ratio Test, respectively.

Theorem 23.1.2 *(Ratio Test) Consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$. Assume that the limit*

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = R$$

exists (possibly infinite). Then, $\sum_{n=1}^{\infty} |a_n| < \infty$ if $R < 1$ and $\sum_{n=1}^{\infty} |a_n| = \infty$ if $1 < R$.

Theorem 23.1.3 *(Root Test) Consider a sequence $a : \mathbb{N}_0 \to \mathbb{R}$. Assume that the limit*

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = R$$

exists. Then, $\sum_{n=1}^{\infty} |a_n| < \infty$ if $R < 1$ and $\sum_{n=1}^{\infty} |a_n| = \infty$ if $1 < R$.