LECTURE NOTES\textsuperscript{1}
ENEE 620
RANDOM PROCESSES IN COMMUNICATION
AND CONTROL

Armand M. Makowski \textsuperscript{2}

\textsuperscript{1}\textcopyright 2011 – 2017 by Armand M. Makowski
\textsuperscript{2}Department of Electrical and Computer Engineering, and Institute for Systems Research, University of Maryland, College Park, MD 20742. E-mail: armand@isr.umd.edu. Phone: (301) 405-6844
Part I

PROBABILITY MODELS
Chapter 1

Modeling random experiments

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 fluctuation of the price of a commodity on some stock exchange, etc.

In these notes we use a widely accepted approach to modeling random experiments 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 $\Omega$ lists all (elementary) outcomes (sometimes also known as samples) generated by the experiment $\mathcal{E}$, and is known as the sample space for the experiment.

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

- The “likelihood” of occurrence of events is assigned only to 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 $\sigma$-fields

With $S$ an arbitrary set, let $\mathcal{P}(S)$ denote the collection of all subsets of $S$ (including the empty set) – We often refer to $\mathcal{P}(S)$ as the power set of $S$ (sometimes also...
denoted $2^S$). Also, let $S$ denote a non-empty collection of subsets of $S$, so that $S \subseteq \mathcal{P}(S)$.

**Definition 1.1.1** The collection $S$ is said to be a field (also known as an algebra) on $S$ if

(F1) $\emptyset \in S$

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

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

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

(F3b) Closed under intersection: If $E \in S$ and $F \in S$, then $E \cap F \in S$

(F3c) Closed under differences: If $E \in S$ and $F \in S$, then $E - F \in S$, $F - E \in \mathcal{F}$ and $E \Delta F \in S$

Note that (F3) implies the seemingly more general statement

(F4) Closed under finite union: If $E_1, \ldots, E_n \in S$, then $\bigcup_{i=1}^{n} E_i \in S$

while (F3b) implies the seemingly more general statement

(F4b) Closed under finite intersection: If $E_1, \ldots, E_n \in S$, then $\bigcap_{i=1}^{n} E_i \in S$

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

**Definition 1.1.2** The non-empty collection of $S$ of subsets of $S$ is a $\sigma$-field (also known as a $\sigma$-algebra) on $S$ if

(F1) $\emptyset \in S$

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

(F3) Closed under countable union: With $I$ a countable index set, if $E_i \in S$ for each $i \in I$, then $\bigcup_{i \in I} E_i \in S$

It is plain that any $\sigma$-field is always a field. Any set $S$ always contains two $\sigma$-fields, namely the trivial $\sigma$-field $\{\emptyset, S\}$ and the full $\sigma$-field $\mathcal{P}(S)$. 
1.2 Measures and probability measures

When $S$ is an arbitrary set and $\mathcal{S}$ is a $\sigma$-field 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 measures defined on $\mathcal{S}$.

We begin with a preliminary definition.

**Definition 1.2.1** Consider an arbitrary non-empty set $S$ equipped with a field $\mathcal{S}$. A set function $\mu : \mu \to [0, \infty]$ is a finitely additive measure on $S$ if it is (finitely) additive in the sense that

$$
\mu [E \cup F] = \mu [E] + \mu [F], \quad \text{for } E, F \in \mathcal{S}, \quad E \cap F = \emptyset
$$

When $\mu [S] = 1$, we refer to such a finitely additive measure as a finitely additive probability measure.

It is immediate that for any finitely additive measure $\mu : \mu \to [0, \infty]$, we have

$$
\mu [\bigcup_{i \in I} E_i] = \sum_{i \in I} \mu [E_i]
$$

for any finite collection $\{E_i, i \in I\}$ of pairwise disjoint sets in $S$, namely

$$
E_i \cap E_j = \emptyset, \quad i \neq j, \quad i, j \in I.
$$

In order to deal with situations where the sample space is uncountable, we extend the definition of finitely additive measure in very much the same way that we extended the definition of field into a $\sigma$-field by allowing countable unions.

**Definition 1.2.2** Consider an arbitrary non-empty set $S$ equipped with a $\sigma$-field $\mathcal{S}$. A $\sigma$-additive measure $\mu$ on $S$ (or on $(S, \mathcal{S})$) is a set mapping $\mu : \mathcal{F} \to [0, \infty]$ which satisfies the following properties:

(M1) $\mu [\emptyset] = 0$

(M2) $\sigma$-additivity: With $I$ a countable index set,

$$
\mu [\bigcup_{i \in I} E_i] = \sum_{i \in I} \mu [E_i]
$$

for any collection $\{E_i, i \in I\}$ of pairwise disjoint sets in $S$, namely

$$
E_i \cap E_j = \emptyset, \quad i \neq j, \quad i, j \in I.
$$
Some easy remarks: A $\sigma$-additive measure is often referred simply as a measure. A measure is then said to be finite if $\mu(S)$ for every $S$ in $\mathcal{S}$, in which case (M1) is automatically satisfied as a consequence of (M2) since by additivity $\mu(\emptyset) = \mu(\emptyset) + \mu(\emptyset)$ on the strength of the obvious relation $\emptyset = \emptyset \cup \emptyset$.

Specializing this last definition we obtain the notion of a probability measure.

**Definition 1.2.3** Consider an arbitrary non-empty set $\Omega$ equipped with a $\sigma$-field $\mathcal{F}$. A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a measure such that $\mathbb{P}[\Omega] = 1$.

In other words, a set mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure on $(S, \mathcal{S})$ if it satisfies the following properties:

(P1) $\mathbb{P}[\Omega] = 1$

(P2) $\sigma$-additivity: With $I$ a countable index set,

$$\mathbb{P}[\bigcup_{i \in I} E_i] = \sum_{i \in I} \mathbb{P}[E_i]$$

for any collection $\{E_i, i \in I\}$ of pairwise disjoint sets in $\mathcal{S}$, namely

$$E_i \cap E_j = \emptyset, \quad i \neq j, \quad i, j \in I.$$

Again we emphasize that $\mathbb{P}[\Omega] = 1$ implies $\mathbb{P}[\emptyset] = 0$ under (P2) since $\Omega = \Omega \cup \emptyset$.

**1.3 Probability models – Definition and elementary facts**

As likelihood assignments are implemented through probability measures, we are now ready to introduce the basic model that will be used in the study of random phenomena.

**Definition 1.3.1** A probability model for the random experiment $\mathcal{E}$ is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega$ is the same space, $\mathcal{F}$ is the $\sigma$-field of events and $\mathbb{P}$ is a probability measure on $\mathcal{F}$.

Sometimes we shall refer to $(\Omega, \mathcal{F}, \mathbb{P})$ as a probability space. An event $E$ in $\mathcal{F}$ such that $\mathbb{P}[E] = 1$ is often called a certain event. Here are simple, yet useful, consequences of the definitions (F1)-(F5) and (P1)-(P2); proofs are elementary and left to the interested reader as exercises. Throughout $(\Omega, \mathcal{F}, \mathbb{P})$ is a given probability triple.
1.3. **PROBABILITY MODELS – DEFINITION AND ELEMENTARY FACTS**

**Elementary facts**  With events $E$ and $F$ in $\mathcal{F}$, we have

- Complementarity:  
  $$ \mathbb{P}[E^c] = 1 - \mathbb{P}[E] $$

- Generalizing additivity:  
  $$ \mathbb{P}[E \cup F] = \mathbb{P}[E] + \mathbb{P}[F] - \mathbb{P}[E \cap F] $$
  so that  
  $$ \mathbb{P}[E \cup F] \leq \mathbb{P}[E] + \mathbb{P}[F] $$

- Monotonicity (I):  
  $$ \mathbb{P}[F] = \mathbb{P}[F - E] + \mathbb{P}[E], \quad E \subseteq F $$

- Monotonicity (II):  
  $$ \mathbb{P}[E] \leq \mathbb{P}[F], \quad E \subseteq F $$

**Bounds**  With countable index set $I$, let $\{E_i, i \in I\}$ denote a countable collection of events in $\mathcal{F}$. The following elementary bounds can be established by induction:

- Boole’s inequality (also known as union bound) is commonly used in Information Theory and theoretical Computer Science, and states that  
  $$ \mathbb{P}[\bigcup_{i \in I} E_i] \leq \sum_{i \in I} \mathbb{P}[E_i] $$

- Bonferroni’s inequality gives a lower bound to $\mathbb{P}[\bigcup_{i \in I} E_i]$: With finite index set $I$, it holds that  
  $$ \mathbb{P}[\bigcup_{i \in I} E_i] \geq \sum_{i \in I} \mathbb{P}[E_i] - \sum_{i,j \in I: i < j} \mathbb{P}[E_i \cap E_j] $$

Both bounds are shown by induction on the size $|I|$ of the collection. The union bound is first established when $I$ is finite; the countably infinite case then follows by an application of Lemma 1.5.1.

In many applications a major question is concerned with generating the probability measure $\mathbb{P}$ that captures the salient features of the experiment $\mathcal{E}$ under consideration once its sample space $\Omega$ has been identified. In particular, this requires that the $\sigma$-field $\mathcal{F}$ of events be judiciously chosen. There are several ways to do so, and we discuss one approach in the next section.
1.4 Discrete probability models

A case of particular importance arises when $\Omega$ is countable, in which case it is customary to take $\mathcal{F} = \mathcal{P}(\Omega)$. In that setting, specifying $\mathbb{P}$ on $(\Omega, \mathcal{P}(\Omega))$ is equivalent to specifying
\[ \{\mathbb{P}[\{\omega\}], \omega \in \Omega\}. \]

Indeed, if $\mathbb{P}$ has been specified on $(\Omega, \mathcal{P}(\Omega))$, then obviously the values (1.1) are known. Conversely, assume that the values $\{\mathbb{P}[\{\omega\}], \omega \in \Omega\}$ are known, then the $\sigma$-additivity of $\mathbb{P}$ and the obvious
\[ F = \cup_{\omega \in F} \{\omega\}, \quad F \in \mathcal{P}(\Omega) \]
relations imply
\[ \mathbb{P}[F] = \sum_{\omega \in F} \mathbb{P}[\{\omega\}], \quad F \in \mathcal{P}(\Omega) \]
and $\mathbb{P}$ is specified on the whole $\sigma$-field $\mathcal{P}(\Omega)$.

Put differently, any probability measure $\mathbb{P}$ on $\sigma$-field $\mathcal{P}(\Omega)$ can be uniquely represented by a collection $\{p(\omega), \omega \in \Omega\}$ satisfying
\[ 0 \leq p(\omega) \leq 1, \quad \omega \in \Omega \quad \text{and} \quad \sum_{\omega \in \Omega} p(\omega) = 1, \]
through the identification
\[ \mathbb{P}[\{\omega\}] = p(\omega), \quad \omega \in \Omega. \]

A collection $\{p(\omega), \omega \in \Omega\}$ satisfying (1.2) is often called a probability mass function (pdf) on $\Omega$.

**Uniform probability assignments** Let $\Omega$ be an arbitrary set to be used as the sample space of a probabilistic experiment $\mathcal{E}$ where outcomes are equally likely to occur – Thus, according to $\mathcal{E}$ an element of $\Omega$ is selected at random as the saying goes, or more accurately, uniformly. A natural question is how to define the corresponding probability measure $\mathbb{P}$, hereafter referred to as the uniform probability measure. We consider several cases.

(i) The set $\Omega$ is a discrete set with a finite number of elements, say $\Omega = \{\omega_1, \ldots, \omega_N\}$ for some finite $N$. The uniform probability measure on such a finite set $\Omega$ assigns the same probability of occurrence to any outcome. Thus, $p(\omega_1) = \ldots = p(\omega_N) = p$, so that
\[ \mathbb{P}[F] = \sum_{\omega \in F} p(\omega) = |F|p, \quad F \in \mathcal{P}(\Omega) \]
whence $p = \frac{1}{|\Omega|}$ upon taking $F = \Omega$. Finally we get

$$P[F] = \frac{|F|}{|\Omega|}, \quad F \in \mathcal{P}(\Omega).$$

(ii) The set $\Omega$ is discrete with $|\Omega| = \infty$, say $\Omega = \{\omega_n, n = 1, 2, \ldots\}$: We should still take $p(\omega)$ to be independent of $\omega$, say $p(\omega) = p$ for all $\omega \in \Omega$ where $p$ is this common value. It still follows that

$$P[F] = |F|p, \quad F \in \mathcal{P}(\Omega), \quad |F| < \infty.$$  

Therefore, we get

$$|F|p \leq 1, \quad F \in \mathcal{P}(\Omega), \quad |F| < \infty$$

and this implies $p = 0$ (because we can select a sequence \{F_n, n = 1, 2, \ldots\} of subsets of $\Omega$ such that $|F_n| = n$ for all $n = 1, 2, \ldots$ Just take $F_n = \{\omega_1, \ldots, \omega_n\}$). A contradiction immediately arises since $\mathbb{P}[\Omega] = \sum_{\omega \in \Omega} p = 0$ by $\sigma$-additivity, and yet $\mathbb{P}[\Omega] = 1$. In other words, it is not possible to have a uniform probability measure on a discrete set with $|\Omega| = \infty$.

What happens when $\Omega$ is uncountable? In Chapter 2 we shall see that for the purpose of defining probability measures on non-countable sets $\Omega$, in general it is not possible to take $F = \mathcal{P}(\Omega)$. This is due to the fact that the $\sigma$-additivity of $\mathbb{P}$ imposes too many constraints, forcing a reduction of $\mathcal{P}(\Omega)$. In particular, in the non-countable case, it is not possible to assign a likelihood of occurrence (through a probability measure satisfying the axioms (P1)-(P2)) to every subset of $\Omega$! The difficulties involved will be 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 ??.

For the remainder of this Chapter, we assume given some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ which is held fixed during the discussion.

### 1.5 Continuity properties of probability measures

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

The impact of monotonicity

Lemma 1.5.1 If the sequence is monotone increasing in the sense that
\[ E_n \subseteq E_{n+1}, \quad n = 1, 2, \ldots \]
then \( \lim_{n \to \infty} P[E_n] = P[\bigcup_{n=1}^{\infty} E_n] \).

Proof. Note the relation
\[ \bigcup_{n=1}^{\infty} E_n = \bigcup_{m=1}^{\infty} F_m \]
where
\[ F_m = E_m - E_{m-1}, \quad m = 1, 2, \ldots \]
(under the convention \( E_0 = \emptyset \)). The events \( \{F_m, m = 1, 2, \ldots\} \) being pairwise disjoint, we get
\[
P[\bigcup_{n=1}^{\infty} E_n] = \sum_{m=1}^{\infty} P[F_m] \tag{1.3}
\]
[By the \( \sigma \)-additivity of \( P \)\(]
\[
= \sum_{m=1}^{\infty} (P[E_m] - P[E_{m-1}])
\]
\[
= \lim_{m \to \infty} \left( \sum_{k=1}^{m} (P[E_k] - P[E_{k-1}]) \right)
\]
(1.3)
\[
= \lim_{m \to \infty} (P[E_m] - P[E_0]) = \lim_{m \to \infty} P[E_m].
\]

This result can be interpreted as a continuity result for \( P \) in the following sense: If we define \( \lim_{n \to \infty} E_n \equiv \bigcup_{n=1}^{\infty} E_n \), then Lemma 1.5.1 states that \( \lim_{n \to \infty} P[E_n] = P[\lim_{n \to \infty} E_n] \).

Lemma 1.5.2 If the sequence is monotone decreasing in the sense that
\[ E_{n+1} \subseteq E_n, \quad n = 1, 2, \ldots \]
then \( \lim_{n \to \infty} P[E_n] = P[\bigcap_{n=1}^{\infty} E_n] \).

This result can also be recast as a continuity result for \( P \): This time, if we define \( \lim_{n \to \infty} E_n \equiv \bigcap_{n=1}^{\infty} E_n \), then \( \lim_{n \to \infty} P[E_n] = P[\lim_{n \to \infty} E_n] \) by virtue of Lemma 1.5.2. The proof of this result is similar to the one given for the case of
1.5. CONTINUITY PROPERTIES OF PROBABILITY MEASURES

a monotone increasing sequence of events. 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 close by stressing that in Lemma 1.5.1 and Lemma 1.5.2 the existence of the limit \( \lim_{n \to \infty} P[E_n] \) is trivially guaranteed by monotonicity. The added information provided by these results is an identification of the limit as the probability of the well-defined events \( \bigcup_{n=1}^{\infty} E_n \) and \( \bigcap_{n=1}^{\infty} E_n \), respectively.

Limsup and liminf, and limits 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

\[
\limsup_{n \to \infty} E_n \equiv \bigcap_{n=1}^{\infty} \left( \bigcup_{m=n}^{\infty} E_m \right) = \bigcap_{n=1}^{\infty} \bar{E}_n
\]

with

\[
\bar{E}_n = \bigcup_{m=n}^{\infty} E_m, \quad n = 1, 2, \ldots
\]

Similarly,

\[
\liminf_{n \to \infty} E_n \equiv \bigcup_{n=1}^{\infty} \left( \bigcap_{m=n}^{\infty} E_m \right) = \bigcup_{n=1}^{\infty} E_n
\]

with

\[
E_n = \bigcap_{m=n}^{\infty} E_m, \quad n = 1, 2, \ldots
\]

The events \( \limsup_{n \to \infty} E_n \) and \( \liminf_{n \to \infty} E_n \) always exist, and as expected, we refer to them as the limit sup and limit inf of the collection \( \{E_n, n = 1, 2, \ldots\} \), respectively. We have the mnemonic notation

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

and

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

**Definition 1.5.1** The collection \( \{E_n, n = 1, 2, \ldots\} \) of events is said to converge if the condition

\[
\limsup_{n \to \infty} E_n = \liminf_{n \to \infty} E_n
\]

holds, in which case we write

\[
\lim_{n \to \infty} E_n \equiv \limsup_{n \to \infty} E_n = \liminf_{n \to \infty} E_n.
\]
With the definition of set continuity given above we have the following continuity property for probability measures.

**Lemma 1.5.3** If the collection \( \{E_n, n = 1, 2, \ldots\} \) of events in \( \mathcal{F} \) converges, then

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

This result requires no monotonicity assumption on the collection \( \{E_n, n = 1, 2, \ldots\} \), only the convergence (1.4).

**Proof.** Obviously, for each \( n = 1, 2, \ldots \) the inclusion \( E_n \subseteq \bar{E}_n \) holds and we have the inclusions \( \bar{E}_{n+1} \subseteq E_n \) and \( \bar{E}_n \subseteq E_{n+1} \). By the continuity of \( P \) on monotone sequences discussed earlier it follows that

\[
P\left[ \limsup_{n \to \infty} E_n \right] = \lim_{n \to \infty} P[\bar{E}_n] \tag{1.5}
\]

and

\[
P\left[ \liminf_{n \to \infty} E_n \right] = \lim_{n \to \infty} P[E_n] \tag{1.6}
\]

By the remarks (1.5) and (1.6), the convergence of the collection \( \{E_n, n = 1, 2, \ldots\} \) implies both \( P[\lim_{n \to \infty} E_n] = \lim_{n \to \infty} P[\bar{E}_n] \) and \( P[\lim_{n \to \infty} E_n] = \lim_{n \to \infty} P[E_n] \). On the other hand, we have

\[
P[E_n] \leq P[E_n] \leq P[\bar{E}_n], \quad n = 1, 2, \ldots
\]

from the obvious \( E_n \subseteq E_n \subseteq \bar{E}_n \). A standard sandwich argument yields the desired result as we let \( n \) go to infinity in this chain of inequalities.

\( \square \)

### 1.6 Independence

Consider a collection \( \{E_i, i \in I\} \) of events in \( \mathcal{F} \) where \( I \) is an arbitrary index set.

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

\[
P[E_i \cap E_j] = P[E_i] P[E_j], \quad i \neq j
\]

all hold. When \( I \) is finite, this is a set of \( \frac{|I|(|I|-1)}{2} \) conditions.
1.6. INDEPENDENCE

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

$$ P[\cap_{j \in J} E_j] = \prod_{j \in J} P[E_j], \quad J \subset I, |J| > 0. $$

This represents $2^{|I|} - (|I| + 1)$ non-trivial conditions.

- Mutual independence (with $I$ arbitrary): 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, namely

$$ P[\cap_{j \in J} E_j] = \prod_{j \in J} P[E_j], \quad J \subseteq I, 0 < |J| < \infty. $$

Set-theoretic operations preserve independence in the following sense.

**Theorem 1.6.1** 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) The events $\{E^*_i, i \in I\}$ are mutually independent where for each $i \in I$, $E^*_i$ is either $E_i$ or its complement $E^c_i$.

(iii) 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 Part (ii) is subsumed by Part (iii); the proof of this last fact is rather cumbersome and will not be given here. With (i) in mind we should not expect pairwise independence to imply mutual independence; this is already apparent from the following simple counterexample.

**Example 1.6.1** Consider the experiment where an item is selected uniformly from a set comprising four distinct objects labelled $a, b, c, d$. Thus, $\Omega = \{a, b, c, d\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $P$ is fully determined by the assignment $p(a) = p(b) = p(c) = p(d) = \frac{1}{4}$. Now define, the events $A = \{a, b\}$, $B = \{b, c\}$ and $C = \{a, c\}$ so that $P[A] = P[B] = P[C] = \frac{1}{2}$. By inspection it is plain that $P[A \cap B] = P[B \cap C] = P[A \cap C] = \frac{1}{4}$, and the events $A$, $B$ and $C$ are indeed pairwise independent. However, $A \cap B \cap C = \emptyset$, whence $P[A \cap B \cap C] = 0 \neq \frac{1}{8}$ and the events $A$, $B$ and $C$ are not mutually independent.
1.7 Borel-Cantelli Lemmas

The Borel-Cantelli lemmas given next constitute an example of a zero-one law. Recall that if \( \{A_n, \ n = 1, 2, \ldots \} \) is a collection of events in \( \mathcal{F} \), then

\[
\limsup_{n \to \infty} A_n = [A_n \ i.o.] = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.
\]

**Lemma 1.7.1** If \( \{A_n, \ n = 1, 2, \ldots \} \) is a collection of events in \( \mathcal{F} \) such that

\[
\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty,
\]

then it is always the case that \( \mathbb{P}[A_n \ i.o.] = 0 \).

**Proof.** Obviously,

\[
\mathbb{P}[A_n \ i.o.] = \mathbb{P}\left[\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\right] = \lim_{n \to \infty} \mathbb{P}\left[\bigcup_{m=n}^{\infty} A_m\right] \quad [\text{By monotonicity in } n \text{ and Lemma 1.5.2}]
\]

\[
= \lim_{n \to \infty} \left( \lim_{k \to \infty} \mathbb{P}\left[\bigcup_{m=n}^{k} A_m\right]\right) \quad [\text{By monotonicity in } k \text{ and Lemma 1.5.1}]
\]

\[
\leq \lim_{n \to \infty} \left( \lim_{k \to \infty} \sum_{m=n}^{k} \mathbb{P}[A_m]\right) \quad [\text{By the union bound on } \mathbb{P}\left[\bigcup_{m=n}^{k} A_m\right]]
\]

\[
\leq \lim_{n \to \infty} \left( \sum_{m=n}^{\infty} \mathbb{P}[A_m]\right).
\]

The result follows since

\[
\lim_{n \to \infty} \left( \sum_{m=n}^{\infty} \mathbb{P}[A_m]\right) = 0
\]

under the convergence condition \( \sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty \). ■

It is natural to wonder what happens to the conclusion of Lemma 1.7.2 when the condition \( \sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty \) holds instead. If we add independence, then the following result holds.
1.7. BOREL-CANTELLI LEMMAS

**Lemma 1.7.2** When the events \( \{A_n, \ n = 1, 2, \ldots \} \) in \( \mathcal{F} \) are mutually independent, then \( \mathbb{P}[A_n \text{ i.o.}] = 1 \) under the condition

\[
\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty.
\]

**Proof.** Our point of departure is the observation that

\[
[A_n \text{ i.o.}]^c = \bigcup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m^c.
\]

By arguments similar to the one given in the proof of Lemma 1.7.1

\[
1 - \mathbb{P}[A_n \text{ i.o.}] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m^c\right] = \lim_{n \to \infty} \mathbb{P}\left[\cap_{m=n}^{\infty} A_m^c\right] \quad \text{[By monotonicity in } n \text{ and Lemma 1.5.1]}
\]

\[
= \lim_{n \to \infty} \left(\lim_{k \to \infty} \mathbb{P}\left[\cap_{m=n}^{k} A_m^c\right]\right) \quad \text{[By monotonicity in } k \text{ and Lemma 1.5.2].}
\]

For each \( n = 1, 2, \ldots \) and \( k = n, n + 1, \ldots \), we see that

\[
\mathbb{P}\left[\cap_{m=n}^{k} A_m^c\right] = \prod_{m=n}^{k} \mathbb{P}[A_m^c] \quad \text{[By independence]}
\]

\[
= \prod_{m=n}^{k} (1 - \mathbb{P}[A_m])
\]

\[
\leq \prod_{m=n}^{k} e^{-\mathbb{P}[A_m]} \quad \text{[Because } 1 - x \leq e^{-x}, \ x \geq 0]\]

\[
(1.7)
\]

Thus, \( \lim_{k \to \infty} \mathbb{P}\left[\cap_{m=n}^{k} A_m^c\right] = 0 \) since \( \sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty \), and the desired conclusion follows.

Without the assumption of independence the conclusion of Lemma 1.7.2 may not hold as the following example shows.

**Example 1.7.1**
1.8 Conditional probabilities

Conditional probabilities are what is needed when there independence fails to hold.

Basic definitions  We begin with a classical definition:

Definition 1.8.1  Consider an event $B$ in $\mathcal{F}$ such that $\mathbb{P}[B] > 0$. We define the conditional probability of the event $A$ in $\mathcal{F}$ given $B$ to be the ratio

$$
\mathbb{P}[A|B] \equiv \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.
$$

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

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

is always true regardless of $\mathbb{P}[B] > 0$ or not: When $\mathbb{P}[B] > 0$ this is clear from (1.8), while if $\mathbb{P}[B] = 0$, then $\mathbb{P}[A \cap B] = 0$ and $\mathbb{P}[A|B] \mathbb{P}[B] = 0$, irrespective of the arbitrary value selected for $\mathbb{P}[A|B]$.

When $\mathbb{P}[B] > 0$ we can define the mapping $Q_B : \mathcal{F} \to [0,1]$ by

$$
Q_B(A) \equiv \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}, \quad A \in \mathcal{F}.
$$

It is easy to show that $Q_B : \mathcal{F} \to \mathbb{R}_+$ is a probability measure on $\mathcal{F}$. Incidentally it is this fact that is often invoked to justify that $\mathbb{P}[\cdot |B]$ be selected as a probability measure on $\mathcal{F}$ when $\mathbb{P}[B] = 0$.

Pairwise independence can be easily characterized in terms of conditional probabilities

Fact 1.8.1  Let $A$ and $B$ be two events in $\mathcal{F}$. Under the condition $\mathbb{P}[B] > 0$, the events $A$ and $B$ are pairwise independent if and only if

$$
\mathbb{P}[A|B] = \mathbb{P}[A].
$$

In other words, the events $A$ and $B$ are pairwise independent if the conditional probability of $A$ given $B$ coincides with its (unconditional) probability $\mathbb{P}[A]$. This is a simple consequence of (1.9) and of the definition of pairwise independence.
1.8. CONDITIONAL PROBABILITIES

Three easy consequences  With \( I \) a countable index set, let \( \{ B_i, \ i \in I \} \) be events in \( \mathcal{F} \) that form an \( \mathcal{F} \)-measurable partition of \( \Omega \) in the sense that
\[
B_i \cap B_j = \emptyset, \quad \forall i, j \in I \quad \text{and} \quad \bigcup_{i \in I} B_i = \Omega.
\]

The law of total probabilities  For each \( A \) in \( \mathcal{F} \), the obvious decomposition
\[
A = \bigcup_{i \in I} (A \cap B_i)
\]
yields
\[
P[A] = \sum_{i \in I} P[A \cap B_i] = \sum_{i \in I} P[A|B_i] P[B_i], \quad A \in \mathcal{F}.
\]
(1.10)

Put differently,
\[
P[A] = \sum_{i \in I} Q_{B_i}(A) P[B_i], \quad A \in \mathcal{F}.
\]

Bayes’ rule – From prior probabilities to posterior probabilities  Consider any event \( A \) in \( \mathcal{F} \) such that \( P[A] > 0 \). For each \( k \) in \( I \), we have
\[
P[B_k|A] = \frac{P[B_k \cap A]}{P[A]} = \frac{P[A|B_k] P[B_k]}{P[A]} = \frac{P[A|B_k] P[B_k]}{\sum_{i \in I} P[A|B_i] P[B_i]} \quad [\text{By the Law of Total Probability}]
\]
(1.11)

This last relation, which gives the posterior probability \( P[B_k|A] \) in terms of the prior probabilities \( \{ P[A|B_i], \ i \in I \} \), is a celebrated relation known as Bayes’s rule.

Modeling sequential decision making  If \( I \) is a finite set, say \( I = \{1, \ldots, n\} \), we have
\[
P[A_1 \cap \ldots \cap A_n] = \prod_{i=2}^{n} P[A_i|A_1 \cap \ldots \cap A_{i-1}] \cdot P[A_1].
\]

This can shown by induction on \( n \).
1.9 Exercises

Ex. 1.1 Consider events \( \{A_n, n = 1, 2, \ldots \} \) defined on some probability triple \((\Omega, \mathcal{F}, \mathbb{P})\). Assume that these events are monotone decreasing in the sense that \( A_{n+1} \subseteq A_n \) for each \( n = 1, 2, \ldots \). Show that the events cannot be pairwise independent if \( 0 < \mathbb{P}[A_n] < 1 \) for all \( n = 1, 2, \ldots \).

Ex. 1.2 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 \( \Omega \); and (iii) the probability assignment \( \mathbb{P} \) is uniform in the sense that \( \mathbb{P}[A] = \frac{|A|}{p} \) for every subset \( A \) of \( \Omega \). Consider now two independent events \( A \) and \( B \), neither of which is empty. What can you say concerning these sets?

Ex. 1.3 A fair coin is rolled \( n \) times under “identical and independent conditions.”

a. Construct a natural probability model \((\Omega, \mathcal{F}, \mathbb{P})\) for this random experiment (Hint: Recall \( \Omega = \{0, 1\}^n \) with the usual identification. What is then \( \mathcal{F} \) and \( \mathbb{P} \)?).

b. With distinct \( i, j = 1, \ldots, n \), define the event \( A_{ij} \) 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 \( \{A_{ij}, 1 \leq i < j \leq n\} \) are pairwise independent but not mutually independent!

Ex. 1.4 Let \( \Omega \) be a countably infinite set, say \( \Omega = \mathbb{N} \). Define the collection \( \mathcal{F} \) of subsets of \( \Omega \) to be \( \mathcal{F} = \{F \subseteq \Omega : \text{Either } |F| < \infty \text{ or } |F^c| < \infty \} \).

a. Show that \( \mathcal{F} \) is an algebra on \( \Omega \). Is it a \( \sigma \)-algebra on \( \Omega \)? Explain.

b. Define the mapping \( \mathbb{P} : \mathcal{F} \to \mathbb{R}_+ \) by

\[
\mathbb{P}[E] = \begin{cases} 
0 & \text{if } |E| < \infty \\
1 & \text{if } |E^c| < \infty.
\end{cases}
\]

Show that \( \mathbb{P} \) is finitely additive. Is \( \mathbb{P} \) also \( \sigma \)-additive on \( \mathcal{F} \). Prove or give a counterexample!

Ex. 1.5 Let \( \Omega \) be an arbitrary non-empty set. Let \( \mathcal{G} \) denote the collection of all its singletons, namely \( \mathcal{G} = \{\{\omega\}, \omega \in \Omega\} \).

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

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

c. If \( \Omega \) is an uncountably infinite set, what is \( \sigma(\mathcal{G}) \)?
Ex. 1.6 Let \( \Omega \) be a set that is countably infinite set, say \( \Omega = \mathbb{N} \). Define the collection \( \mathcal{F} \) of subsets of \( \Omega \) to be

\[
\mathcal{F} = \left\{ E \subseteq \Omega : \begin{array}{l}
E \text{ is finite} \\
E^c \text{ is finite}
\end{array} \right\}
\]

a. Show that \( \mathcal{F} \) is an algebra on \( \Omega \).
b. Is \( \mathcal{F} \) a \( \sigma \)-algebra on \( \Omega \)? Prove or give a counterexample!
Define the mapping \( \mathbb{P} : \mathcal{F} \to \mathbb{R}_+ \) by

\[
\mathbb{P}[E] = \begin{cases} 
0 & \text{if } E \text{ is finite} \\
1 & \text{if } E^c \text{ is finite}.
\end{cases}
\]

c. Is the set function \( \mathbb{P} : \mathcal{F} \to \mathbb{R}_+ \) a probability measure on \( \mathcal{F} \)? Prove or give a counterexample!

Ex. 1.7 Let \( \Omega \) be a set that is uncountably infinite set, say \( \Omega = \mathbb{R} \). Define the collection \( \mathcal{F} \) of subsets of \( \Omega \) to be

\[
\mathcal{F} = \left\{ E \subseteq \Omega : \begin{array}{l}
E \text{ is countable} \\
E^c \text{ is countable}
\end{array} \right\}
\]

where here countable means either finite or countably infinite.

a. Show that \( \mathcal{F} \) is an algebra on \( \Omega \).
b. Is \( \mathcal{F} \) a \( \sigma \)-algebra on \( \Omega \)? Prove or give a counterexample!
Define the mapping \( \mathbb{P} : \mathcal{F} \to \mathbb{R}_+ \) by

\[
\mathbb{P}[E] = \begin{cases} 
0 & \text{if } E \text{ is countable} \\
1 & \text{if } E^c \text{ is countable}.
\end{cases}
\]

c. Is the set function \( \mathbb{P} : \mathcal{F} \to \mathbb{R}_+ \) a probability measure on \( \mathcal{F} \)? Prove or give a counterexample!

Ex. 1.8 Consider a sequence of events \( \{ A_n, \ n = 1, 2, \ldots \} \) on some probability triple \((\Omega, \mathcal{F}, \mathbb{P})\) such that \( \mathbb{P}[A_n] = 1 \) for all \( n = 1, 2, \ldots \)

a. Show that

\[
\mathbb{P}\left[ \bigcap_{j \in J} A_j \right] = 1 \quad \text{and} \quad \mathbb{P}\left[ \bigcup_{j \in J} A_j \right] = 1, \quad J \subseteq \{1, 2, \ldots\}, 1 \leq |J| < \infty
\]

b. What can you say concerning the value of \( \mathbb{P}[A_n \ i.o.] \)? Can you obtain your answer by making use of the Borel-Cantelli Lemmas?
Ex. 1.9 Let \((\Omega, \mathcal{F}, \mathbb{P})\) denote a probability triple, and let \(B\) an event in \(\mathcal{F}\) with \(0 < \mathbb{P}[B] < 1\) (in order to avoid trivial situations of limited interest). Define the collection \(\mathcal{A}_B\) of events in \(\mathcal{F}\) by

\[
\mathcal{A}_B = \{ A \in \mathcal{F} : \mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B] \}.
\]

a. Show that both \(\Omega\) and the empty set \(\emptyset\) belong to \(\mathcal{A}_B\).

b. Show that \(\mathcal{A}_B\) is closed under complementarity, i.e., if \(A\) is an element of \(\mathcal{A}_B\), then so is its complement \(A^c\).

c. Is the family \(\mathcal{A}_B\) a \(\sigma\)-field on \(\Omega\)? Prove or give a counterexample!

Ex. 1.10 The following problem arises in the context of random key pre-distribution for wireless sensor networks. There are \(n\) sensors, labelled \(1, \ldots, n\), and a pool of \(P\) cryptographic keys, labelled \(1, \ldots, P\). Before deployment the manufacturer inserts \(K < P\) keys (drawn form the pool of keys) into the memory module of each sensor – Here \(K\) is a fixed positive integer, the same for all nodes. For each \(i = 1, \ldots, n\), let \(\Gamma_i\) denote the set of keys assigned to node \(i\); the set \(\Gamma_i\) is also known as the key ring of node \(i\).

Throughout the sets \(\Gamma_1, \ldots, \Gamma_n\) are selected randomly and independently across sensor nodes. More precisely, the rvs \(\Gamma_1, \ldots, \Gamma_n\) are i.i.d. rvs, each of which is uniformly distributed over the set of all subsets of size \(K\) drawn from \(\{1, \ldots, P\}\).

Thus, for each \(i = 1, \ldots, n\) we have

\[
\mathbb{P}[\Gamma_i = S] = \frac{1}{\binom{P}{K}}, \quad S \subseteq \{1, \ldots, P\}, \quad |S| = K.
\]

With distinct \(i, j = 1, \ldots, n\), compute the probabilities

\[
\mathbb{P}[\Gamma_i \cap R = \emptyset], \quad R \subseteq \{1, \ldots, P\}
\]

and

\[
\mathbb{P}[\Gamma_i \cap \Gamma_j = \emptyset].
\]

Ex. 1.11 A freight train, which is being assembled at a local train yard, is scheduled to serve three different destinations. The train comprises \(K\) freight cars with \(K_i\) cars destined for destination \(D_i\) \((i = 1, 2, 3)\) – Obviously \(K_1 + K_2 + K_3 = K\). The cars are randomly scattered along the train.

a. Construct a probability model \((\Omega, \mathcal{F}, \mathbb{P})\) for this situation under the assumption that the \(K_i\) cars destined for destination \(D_i\) are distinguishable, e.g., the cars for destination \(D_i\) are labelled \((D_i, 1), \ldots, (D_i, K_i)\) \((i = 1, 2, 3)\).

b. Use this probability model to compute the probability that all the cars for each destination are side-by-side.
Ex. 1.12 Consider two urns, say $U_1$ and $U_2$, each of which contains $R$ red balls and $B$ blue balls. A ball is drawn at random from urn $U_1$, and put in urn $U_2$. Then, after stirring the balls in urn $U_2$, a ball is drawn at random from urn $U_2$.

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

b. Compute the probability that the ball drawn from urn $U_2$ is red?
CHAPTER 1. MODELING RANDOM EXPERIMENTS
Chapter 2

Events and measurability

As already mentioned in Chapter 1, determining the appropriate $\sigma$-field $F$ of events is technically more delicate when $\Omega$ is countably infinite. This is in large part due to the measure-theoretic constraints imposed by the $\sigma$-additivity of probability measures to be defined on $F$. These issues are briefly explored in the present chapter.

2.1 Generating $\sigma$-fields

In order to formalize these ideas, we introduce several definitions. Throughout this section let $S$ denote an arbitrary set.

**Fact 2.1.1** If $\{S_i, i \in I\}$ is a non-empty family of $\sigma$-fields on $S$, then the collection of subsets of $S$ defined by

$$\{E \in \mathcal{P}(S) : E \in S_i, i \in I\}$$

is also a $\sigma$-field on $S$ denoted indifferently either $\bigwedge_{i \in I} S_i$ or $\cap (S_i, i \in I)$. We refer to it as the intersection of the $\sigma$-fields $\{S_i, i \in I\}$.

The proof that $\bigwedge_{i \in I} F_i$ is a $\sigma$-field on $S$ is left as an exercise.

**Definition 2.1.1** If $\mathcal{G}$ is a collection of subsets of $S$, let $\sigma(\mathcal{G})$ denote the smallest $\sigma$-field on $S$ that contains $\mathcal{G}$.

This definition is well posed: The collection of $\sigma$-fields on $S$ that contain $\mathcal{G}$ is given by

$$\left\{S \subseteq \mathcal{P}(S) : \begin{array}{l}
S \text{ is a } \sigma\text{-field on } S \\
\text{and} \\
\mathcal{G} \text{ is contained in } S
\end{array}\right\}.$$
It is not empty because $\mathcal{P}(S)$ is a $\sigma$-field that always contains $\mathcal{G}$. It is therefore easy to see

\[ \sigma(\mathcal{G}) = \bigwedge \left\{ \mathcal{S} \subseteq \mathcal{P}(S) : \mathcal{S} \text{ is a } \sigma\text{-field on } S \text{ and } \mathcal{G} \text{ is contained in } \mathcal{S} \right\}. \tag{2.1} \]

This argument also shows that any other $\sigma$-field $\mathcal{S}$ on $S$ that contains $\mathcal{G}$ must necessarily contain the $\sigma$-field (2.1).

**Definition 2.1.2** Let $\mathcal{G}$ and $\mathcal{S}$ be two collections of subsets of $S$ with $\mathcal{G} \subseteq \mathcal{S}$. If $\mathcal{S}$ is a $\sigma$-field on $S$, we say that $\mathcal{G}$ generates the $\sigma$-field $\mathcal{S}$, or equivalently, that $\mathcal{G}$ is a generating family (or a generator) for $\mathcal{S}$, if $\mathcal{S} = \sigma(\mathcal{G})$.

The following fact is elementary.

**Fact 2.1.2** 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 $\sigma(\mathcal{G}_1) \subseteq \sigma(\mathcal{G}_2)$.

Fact 2.1.2 is often used with $\mathcal{G}_2$ already being a $\sigma$-field, in which case $\sigma(\mathcal{G}_1) \subseteq \mathcal{G}_2$.

### 2.2 Mappings and measurability

In this section, let $S_a$ and $S_b$ be arbitrary sets (possibly identical). Let $\mathcal{H}$ be a collection of subsets of $S_b$ (so $\mathcal{H} \subseteq \mathcal{P}(S_b)$). For any mapping $g : S_a \rightarrow S_b$, recall that

\[ g^{-1}(H_b) = \{ s_a \in S_a : g(s_a) \in H_b \}, \quad H_b \in \mathcal{P}(S_b). \]

Thus, with $\mathcal{H}$ denoting a collection of subsets of $S_b$, it is natural to write

\[ g^{-1}(\mathcal{H}) = \{ g^{-1}(H_b) : H_b \in \mathcal{H} \}. \]

**Lemma 2.2.1** If $\mathcal{H}$ is a $\sigma$-field on $S_b$, then the collection $g^{-1}(\mathcal{H})$ of subsets of $S_a$ is a $\sigma$-field on $S_a$. More generally, for any collection $\mathcal{H}$ of subsets of $S_b$, we have

\[ g^{-1}(\sigma(\mathcal{H})) = \sigma(g^{-1}(\mathcal{H})). \tag{2.2} \]
2.2. MAPPINGS AND MEASURABILITY

Proof. It is easy to check that the collection $g^{-1}(H)$ is a $\sigma$-field on $S_a$ when $H$ is a $\sigma$-field on $S_b$; it is left as an exercise. We now turn to establishing (2.3): The inclusion $\sigma(g^{-1}(H)) \subseteq g^{-1}(\sigma(H))$ is straightforward since $g^{-1}(\sigma(H))$ is a $\sigma$-field which contains $g^{-1}(H)$. To establish the reverse inclusion $g^{-1}(\sigma(H)) \subseteq \sigma(g^{-1}(H))$, consider the collection

$$H_g \equiv \{ H_b \in \mathcal{P}(S_b) : g^{-1}(H_b) \in \sigma(g^{-1}(H)) \}.$$ 

It is easy to check that $H_g$ is a $\sigma$-field on $S_b$ and that it contains $H$. Therefore, $H_g$ contains $\sigma(H)$, and the conclusion $g^{-1}(\sigma(H)) \subseteq \sigma(g^{-1}(H))$ follows. 

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

Definition 2.2.1 When the sets $S_a$ and $S_b$ are equipped with $\sigma$-fields $S_a$ and $S_b$, respectively, the mapping $g : S_a \to S_b$ is said to be $(S_a, S_b)$-measurable if the conditions

$$g^{-1}(F_b) \in S_a, \quad F_b \in S_b$$

all hold. The conditions (2.3) can be rewritten more compactly as

$$g^{-1}(S_b) \subseteq S_a.$$ 

When coupled with Lemma 2.2.1 this definition leads to the following fact which is operationally useful as we shall see shortly in the next section.

Lemma 2.2.2 If the $\sigma$-field $S_b$ is generated by the collection $H$, i.e., $S_b = \sigma(H)$, then the mapping $g : S_a \to S_b$ is $(S_a, S_b)$-measurable if and only if the conditions

$$g^{-1}(F_b) \in S_a, \quad F_b \in H$$

all hold. The conditions (2.5) are equivalent to

$$g^{-1}(H) \subseteq S_a.$$
Proof. The condition (2.4) obviously implies (2.6) under the condition $S_b = \sigma (\mathcal{H})$. Conversely, assume that the mapping $g : S_a \to S_b$ satisfies (2.6): We have

\[ g^{-1}(S_b) = g^{-1}(\sigma (\mathcal{H})) \quad [\text{Because } S_b = \sigma (\mathcal{H})] \]
\[ = \sigma (g^{-1}(\mathcal{H})) \quad [\text{By Lemma 2.2.1}] \]
\[ \subseteq S_a \quad [\text{By (2.6)}] \]

(2.7)

as we use the fact that $S_a$ is a itself a $\sigma$-field.

Consider now a situation where the three pairs $(S_a, S_b)$, $(S_b, S_c)$ and $(S_c, S_c)$ are measurable spaces. With the 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

\[ (h \circ g) (s_a) = h(g(s_a)), \quad s_a \in S_a. \]

Fact 2.2.1 If the mapping $g : S_a \to S_b$ is $(S_a, S_b)$-measurable and if the mapping $h : S_b \to S_c$ is $(S_b, S_c)$-measurable, then the composition mapping $h \circ g : S_a \to S_c$ is itself $(S_a, S_c)$-measurable.

Proof. The conclusion follows from the elementary fact

\[ (h \circ g)^{-1} (F_c) = g^{-1}(h^{-1}(F_c)), \quad F_c \in \mathcal{P}(S_c) \]

coupled to the definition of $(S_b, S_c)$-measurability and $(S_a, S_b)$-measurability. Details are left to the interested reader.

2.3 Borel mappings

We specialize some of the definitions of Section 2.2 to the situation when the domain is a measurable space $(S, S)$ and the range space is $\mathbb{R}^p$ for some $p$. Thus, writing $(S, S)$ for $(S_a, S_a)$ and $S_b = \mathcal{B}(\mathbb{R}^p)$, it is understood (unless specified otherwise) that we always take $S_b = \mathcal{B}(\mathbb{R}^p)$.

Definition 2.3.1 A mapping $g : S \to \mathbb{R}^p$ is said to be a Borel mapping if it is an $(S, \mathcal{B}(\mathbb{R}^p))$-measurable mapping in the sense of Definition 2.3, in which case the conditions

\[ g^{-1}(B) \in S, \quad B \in \mathcal{B}(\mathbb{R}^p) \]

all hold.
2.3. BOREL MAPPINGS

With \((S_c, S_c) = (\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))\) Fact 2.2.1 takes the following form.

**Fact 2.3.1** 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 2.2.2 leads to the following fact which is crucial for understanding the importance of probability distributions.

**Lemma 2.3.1** Let \(G\) denote a collection of subsets of \(\mathbb{R}^p\) which generates the Borel \(\sigma\)-field \(\mathcal{B}(\mathbb{R}^p)\), i.e.,
\[
\mathcal{B}(\mathbb{R}^p) = \sigma(\mathcal{G}).
\] (2.9)

The mapping \(g : S \to \mathbb{R}^p\) is a Borel mapping if and only if the weaker set of conditions
\[
g^{-1}(E) \in \mathcal{S}, \quad E \in \mathcal{G}
\] (2.10)

holds.

**Some important generators** With Lemma 2.3.1 in mind, we note that there are many generators known for the Borel \(\sigma\)-field \(\mathcal{B}(\mathbb{R}^p)\). For instance, we have (??) with \(G = \mathcal{R}_{\text{open}}(\mathbb{R}^p)\) where \(\mathcal{R}_{\text{open}}(\mathbb{R}^p)\) is the collection of all finite open rectangles, i.e.,
\[
\mathcal{R}_{\text{open}}(\mathbb{R}^p) \equiv \left\{ I_1 \times \ldots \times I_p, \quad I_k \in \mathcal{I}(\mathbb{R}) \right\}_{k=1,\ldots,p}
\]
where
\[
\mathcal{I}(\mathbb{R}) = \{ (a, b) : a, b \in \mathbb{R} \}.
\]

This can be established with the help of the next fact which is a high-dimensional analog of Fact 3.3.1 encountered in one dimension.

**Fact 2.3.2** For any open set \(U\) in \(\mathbb{R}^p\) there exists a countable family of open rectangles \(\{R_i, \ i \in I\}\) in \(\mathcal{R}_{\text{open}}(\mathbb{R}^p)\) with countable \(I\) such that \(U = \bigcup_{i \in I} R_i\).

We also have (??) with \(G = \mathcal{R}_{\text{SW}}(\mathbb{R}^p)\) where \(\mathcal{R}_{\text{SW}}(\mathbb{R}^p)\) is the collection of all closed Southwest rectangles, i.e.,
\[
\mathcal{R}_{\text{SW}}(\mathbb{R}^p) \equiv \left\{ I_1 \times \ldots \times I_p, \quad I_k = (-\infty, a_k], \quad a_k \in \mathbb{R} \right\}_{k=1,\ldots,p}
\]
2.4 From \( p \) dimensions to one dimension

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 2.4.1** 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.** Apply Lemma 2.3.1 with \( G = \mathcal{R}_{SW}(\mathbb{R}^p) \): The mapping \( g : S \to \mathbb{R}^p \) is then a Borel mapping if and only if the conditions

\[
\left\{ s \in S : g(s) \in \prod_{k=1}^{p} (-\infty, a_k] \right\} \in \mathcal{S}, \quad (a_1, \ldots, a_p) \in \mathbb{R}^p
\]

all hold.

For each \( (a_1, \ldots, a_p) \) in \( \mathbb{R}^p \) easy set-theoretic considerations show that

\[
\left\{ s \in S : g(s) \in \prod_{k=1}^{p} (-\infty, a_k] \right\} = \left\{ s \in S : g_k(s) \leq a_k, \quad k = 1, \ldots, p \right\} = \bigcap_{k=1}^{p} \left\{ s \in S : g_k(s) \leq a_k \right\}.
\]

\[(2.12)\]

If the mappings \( g_1, \ldots, g_p : S \to \mathbb{R} \) are all Borel mappings, then we have

\[
\left\{ s \in S : g_k(s) \leq a_k \right\} \in \mathcal{S}, \quad a_k \in \mathbb{R}, \quad k = 1, \ldots, p
\]

upon invoking Lemma 2.3.1 for \( p = 1 \) as we recall that

\[
\mathcal{B}(\mathbb{R}) = \sigma((-\infty, a], \ a \in \mathbb{R}).
\]

\[(2.14)\]

It then follows that (2.11) holds as needed since \( \mathcal{S} \) being a \( \sigma \)-field on \( S \), it is closed under finite intersections.

Conversely, assume that (2.11) holds for all \( (a_1, \ldots, a_p) \) in \( \mathbb{R}^p \). Pick \( \ell = 1, \ldots, p \) and \( a \) in \( \mathbb{R} \). For each \( n = 1, 2, \ldots \), define the element \( (b_{n,1}, \ldots, b_{n,p}) \) in \( \mathbb{R}^p \) given by

\[
b_{n,k} = \begin{cases} 
  a & \text{if } k = \ell \\
  n & \text{if } k \neq \ell.
\end{cases}
\]
Combining (2.11) and (2.12) (with \((a_1, \ldots, a_n)\) replaced by \((b_{n,1}, \ldots, b_{n,p})\)) we get

\[
\cap_{k=1}^{p} \left\{ s \in S : g_k(s) \leq b_{n,k} \right\} = G_{\ell}(a; n) \cap \left\{ s \in S : g_{\ell}(s) \leq a \right\} \in \mathcal{S}
\]

where we have set

\[
G_{\ell}(a; n) \equiv \cap_{k=1, k\neq \ell}^{p} \left\{ s \in S : g_k(s) \leq n \right\}.
\]

The sets \(\{G_{\ell}(a; n), n = 1, 2, \ldots\}\) are monotonically increasing with \(\cup_{n=1}^{\infty} G_{\ell}(a; n) = \mathcal{S}\). It then follows that the sets

\[
\left\{ \cap_{k=1}^{p} \left\{ s \in S : g_k(s) \leq b_k \right\}, \ n = 1, 2, \ldots \right\}
\]

are also monotonically increasing with

\[
\cup_{n=1}^{\infty} \left( \cap_{k=1}^{p} \left\{ s \in S : g_k(s) \leq b_k \right\} \right) = \left\{ s \in S : g_{\ell}(s) \leq a \right\} \cap \mathcal{S} = \left\{ s \in S : g_{\ell}(s) \leq a \right\}.
\]

A \(\sigma\)-field being closed under countable union, we can invoke (2.15) to conclude that the set \(\{s \in S : g_{\ell}(s) \leq a\}\) indeed belongs to \(\mathcal{S}\), and the mapping \(g_{\ell} : S \to \mathbb{R}\) is a Borel mapping by virtue of Lemma 2.3.1 for \(p = 1\) (with (2.14)).

Most (if not all) mappings \(\mathbb{R}^p \to \mathbb{R}^q\) encountered in applications are Borel mappings. In particular, any \textit{continuous} mapping \(\mathbb{R}^p \to \mathbb{R}^q\) can be shown to be a Borel mapping!

## 2.5 Cartesian products

We start with some standard definitions: Let \(S_a\) and \(S_b\) be two arbitrary sets (possibly identical). Recall that the \textit{Cartesian product} of \(S_a\) and \(S_b\), denoted \(S_a \times S_b\), is the set of ordered pairs defined by

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

If \(\mathcal{H}_a\) and \(\mathcal{H}_b\) are collections of subsets of \(S_a\) and \(S_b\), respectively, it is natural to set

\[
\mathcal{H}_a \times \mathcal{H}_b \equiv \{F_a \times F_b : F_a \in \mathcal{H}_a, F_b \in \mathcal{H}_b\}.
\]

A set \(F_a \times F_b\) in \(\mathcal{H}_a \times \mathcal{H}_b\) is sometimes called a \textit{rectangle} with sides \(F_a\) in \(S_a\) and \(F_b\) in \(S_b\).
In general the collection $\mathcal{H}_a \times \mathcal{H}_b$ is not a $\sigma$-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 $\sigma$-field: For instance, if we seek the complement of $F_a \times F_b$ in $S_a \times S_b$, it is given by the union of the two sets $F_a^c \times S_b$ and $S_a \times F_b^c$ (where $F_a^c$ and $F_b^c$ are the complements of $F_a$ and $F_b$ in $S_a$ and $S_b$, respectively). Obviously, $(F_a^c \times S_b) \cup (S_a \times F_b^c)$ is not a rectangle, hence is not an element of $\mathcal{H}_a \times \mathcal{H}_b$.

The next result helps put measurability on Cartesian products.

**Lemma 2.5.1** 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

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

**Proof.** As the inclusion $\mathcal{H}_a \times \mathcal{H}_b \subseteq \sigma (\mathcal{H}_a) \times \sigma (\mathcal{H}_b)$ obviously holds, we get the inclusion

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

To establish the reverse inclusion

$$\sigma (\sigma (\mathcal{H}_a) \times \sigma (\mathcal{H}_b)) \subseteq \sigma (\mathcal{H}_a \times \mathcal{H}_b),$$

we proceed as follows: Define the collections

$$\mathcal{H}_a^* \equiv \{ F_a \subseteq S_a : F_a \times S_b \in \sigma (\mathcal{H}_a \times \mathcal{H}_b) \}$$

and

$$\mathcal{H}_b^* \equiv \{ F_b \subseteq S_b : S_a \times F_b \in \sigma (\mathcal{H}_a \times \mathcal{H}_b) \}.$$

It is a simple matter to check that $\mathcal{H}_a^*$ and $\mathcal{H}_b^*$ are $\sigma$-fields on $S_a$ and $S_b$, respectively.

Pick an arbitrary subset $E$ of $S_a \times S_b$ that belongs to $\mathcal{H}_a^* \times \mathcal{H}_b^*$. Thus, $E = F_a \times F_b$ with $F_a$ in $\mathcal{H}_a^*$ and $F_b$ in $\mathcal{H}_b^*$. However, by definition $F_a \times S_b$ and $S_a \times F_b$ both belong to the $\sigma$-field $\sigma (\mathcal{H}_a \times \mathcal{H}_b)$, hence their intersection also belongs to $\sigma (\mathcal{H}_a \times \mathcal{H}_b)$. Noticing that

$$E = F_a \times F_b = (F_a \times S_b) \cap (S_a \times F_b),$$

we conclude that $E$ is also an element $\sigma (\mathcal{H}_a \times \mathcal{H}_b)$. Put differently, we have just shown that

$$\mathcal{H}_a^* \times \mathcal{H}_b^* \subseteq \sigma (\mathcal{H}_a \times \mathcal{H}_b).$$
2.6. EXTENDED BOREL MAPPINGS AND LIMITS

Obviously \( \mathcal{H}_a \subseteq \mathcal{H}_a^* \), hence \( \sigma(\mathcal{H}_a) \subseteq \mathcal{H}_a^* \) since \( \mathcal{H}_a^* \) is \( \sigma \)-field on \( S_a \). We similarly have \( \sigma(\mathcal{H}_b) \subseteq \mathcal{H}_b^* \). It follows from (2.19) that

\[
\sigma(\mathcal{H}_a) \times \sigma(\mathcal{H}_b) \subseteq \sigma(\mathcal{H}_a \times \mathcal{H}_b).
\]

and the conclusion (2.17) is now immediate.

As an example of application of Lemma 2.5.1 we offer the following useful fact

**Fact 2.5.1** For every positive integers \( p \) and \( q \) we have

\[
B(R^{p+q}) = \sigma(B(R^p) \times B(R^q)).
\]

**Proof.** This is a simple consequence of Lemma 2.5.1 once we note that for each positive integer \( s \) we have \( B(R^s) = \sigma(\mathcal{R}_{\text{open}}(R^s)) \) as pointed out in Section 2.3, and that

\[
\mathcal{R}_{\text{open}}(R^{p+q}) = \mathcal{R}_{\text{open}}(R^p) \times \mathcal{R}_{\text{open}}(R^q).
\]

Indeed,

\[
B(R^{p+q}) = \sigma(\mathcal{R}_{\text{open}}(R^{p+q})) = \sigma(\mathcal{R}_{\text{open}}(R^p) \times \mathcal{R}_{\text{open}}(R^q)) \quad \text{(2.22)}
\]

and the desired result follows.

2.6 Extended Borel mappings and limits

Sometimes the notion of Borel mappings as we defined it so far fails to cover important situations that arise in applications when the values \( \pm \infty \) naturally occur.

We begin by introducing the extended real line \( \mathbb{R} = [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\} \). The extended Borel \( \sigma \)-field \( B(\mathbb{R}) \) on \( \mathbb{R} \) is simply defined as

\[
B(\mathbb{R}) \equiv \sigma(B(\mathbb{R}), \{-\infty\}, \{\infty\}).
\]
Definition 2.6.1 Consider a measurable space \( (S, S) \). A mapping \( g : S \to [-\infty, \infty] \) is said to be an (extended) Borel mapping if

\[
g^{-1}(B) \in S, \quad B \in \mathcal{B}(\mathbb{R}).
\]

It is easy to check that this definition is equivalent to the conditions

\[
\{ s \in S : g(s) \in (-\infty, x] \} \in S, \quad x \in \mathbb{R}
\]

and

\[
S_{\pm\infty} \equiv \{ s \in S : g(s) = \pm\infty \} \in S
\]

holding.

Consider a sequence of extended Borel mappings \( \{g_n, n = 1, 2, \ldots\} \) which are all defined on the same measurable space \( (S, S) \).

Lemma 2.6.1 The following mappings \( S \to [-\infty, \infty] \) derived from the sequence \( \{g_n, n = 1, 2, \ldots\} \) are all Borel mappings in the extended sense:

(i) The supremum mapping \( S \to [-\infty, \infty] \) defined by

\[
s \mapsto \sup_{m \geq 1} g_m(s), \quad s \in S.
\]

(ii) The infimum mapping \( S \to [-\infty, \infty] \) defined by

\[
s \mapsto \inf_{m \geq 1} g_m(s), \quad s \in S.
\]

(iii) The limsup mapping \( S \to [-\infty, \infty] \) defined by

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

(iv) The liminf mapping \( S \to [-\infty, \infty] \) defined by

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

Proof. For each \( a \) in \( \mathbb{R} \), we note that

\[
\left\{ s \in S : \sup_{m \geq 1} g_m(s) \leq a \right\} = \cap_{m \geq 1} \{ s \in S : g_m(s) \leq a \} \in S
\]

since for each \( m = 1, 2, \ldots \), the mapping \( g_m : S \to \mathbb{R} \) is a Borel mapping with

\[
\{ s \in S : g_m(s) \leq a \} \in S.
\]
The Borel measurability of the supremum mapping follows from Lemma 2.3.1 as we note that the intervals \{-\infty, a], \ a \in \mathbb{R}\} generate the \(\sigma\)-field \(B(\mathbb{R})\).

The Borel measurability of the infimum mapping is an immediate consequence of that of the supremum mapping (applied to the mappings \(s \rightarrow -g_m(s)\)) upon noting that
\[
\inf_{m \geq 1} g_m(s) = -\sup_{m \geq 1} (-g_m(s)), \quad s \in S.
\]

The Borel measurability of the limsup and liminf mappings are now straightforward; the details of the proof are left to the interested reader.

It follows that
\[
S^* \equiv \left\{ s \in S : \lim_{n \to \infty} \inf g_n(s) = \lim_{n \to \infty} \sup g_n(s) \right\} \in \mathcal{S}
\]
and on \(S^*\), it holds that \(\lim_{n \to \infty} g_n(s)\) exists (possibly as an element in \([-\infty, \infty]\)), and is the common value assumed by \(\lim \inf_{n \to \infty} g_n\) and \(\lim \sup_{n \to \infty} g_n\).

2.7 Exercises
Chapter 3

Building probability models

3.1 Extensions of probability measures

In constructing a probability model \((\Omega, \mathcal{F}, \mathbb{P})\) for a random experiment \(\mathcal{E}\) we are often faced with the following situation: After identifying the sample space \(\Omega\), structural properties of \(\mathcal{E}\) suggest natural likelihood assignments for certain events in a collection of events which we denote \(\mathcal{G}\) – Let \(\text{Lik}(G)\) denote the suggested likelihood of event \(G\) in \(\mathcal{G}\). Obviously both \(\emptyset\) and \(\Omega\) would be included in \(\mathcal{G}\), and we would set \(\text{Lik}(\emptyset) = 0\) and \(\text{Lik}(\Omega) = 1\). Two points should be clear:

(i) We expect the desired \(\sigma\)-field on \(\Omega\) to contain \(\mathcal{G}\) – In fact, in the name of minimality it would be natural to require \(\mathcal{F} \equiv \sigma(\mathcal{G})\). After all, \(\sigma(\mathcal{G})\) is the smallest \(\sigma\)-field where a probability measure \(\mathbb{P}\) could be defined that is compatible with the likelihood assignments \(\{\text{Lik}(G), \, G \in \mathcal{G}\}\).

(ii) The probability assignments \(\{\text{Lik}(G), \, G \in \mathcal{G}\}\) being the values taken by the probability measure to be defined on the events in \(\mathcal{G}\) (and then ultimately, on \(\sigma(\mathcal{G})\)), it stands to reason that we should define \(\mathbb{P}\) on \(\mathcal{G}\) by

\[
\mathbb{P}[G] \equiv \text{Lik}(G), \quad G \in \mathcal{G}.
\]

However, such a definition should be compatible with the axioms (F1)-(F2) satisfied by probability measures. In particular, with \(I\) a countable index set, if the sets \(\{G_i, i \in I\}\) are pairwise disjoint with \(G_i \in \mathcal{G}\) for each \(i \in I\), then one of two possibilities can occur:

(a) If \(\bigcup_{i \in I} G_i\) is an element of \(\mathcal{G}\) as well, then the relation

\[
\text{Lik}(\bigcup_{i \in I} G_i) = \sum_{i \in I} \text{Lik}(G_i)
\]
must hold since (3.1) implies

\[ P \left( \bigcup_{i \in I} G_i \right) = \sum_{i \in I} P[G_i] \tag{3.3} \]

as required by the $\sigma$-additivity of probability measures. Similarly, if both $G$ and $G^c$ are in $\mathcal{G}$, the relation

\[ \text{Lik}(G^c) = 1 - \text{Lik}(G), \tag{3.4} \]

should hold to reflect the usual complementarity property of probability measures.

(b) On the other hand, if $\bigcup_{i \in I} G_i$ is not an element of $\mathcal{G}$, then it is natural to define the likelihood of such an event by setting

\[ P \left( \bigcup_{i \in I} G_i \right) \equiv \sum_{i \in I} \text{Lik}(G_i). \tag{3.5} \]

Similarly, if $G$ is an element of $\mathcal{G}$ but its complement $G^c$ is not, then we should set

\[ P[G^c] \equiv 1 - \text{Lik}(G). \tag{3.6} \]

In principle this provides a constructive approach to recursively building a probability measure $P$ on $\sigma(\mathcal{G})$ by taking further unions and complements of sets to which a likelihood value has been assigned under the probability measures $P$ as it is being constructed. With this in mind it is natural to wonder whether ultimately a probability measure $P$ can be constructed on $\sigma(\mathcal{G})$ which is consistent with the likelihood assignments $\{\text{Lik}(G), \ G \in \mathcal{G}\}$. A moment of reflection should convince the reader that conditions are needed on $\mathcal{G}$ for this to occur.

The following extension result is due to Carathéodory.

**Theorem 3.1.1**  Let $S$ be a field on the set $S$. Every $\sigma$-finite measure $\mu$ on $S$ admits a unique extension $\mu_{\text{Ext}} : \sigma(S) \to [0, \infty]$ to the $\sigma$-field $\sigma(S)$ generated by $S$.

In the context of Probability Theory Carathéodory’s Theorem is often applied to probability measures (which are always $\sigma$-finite).

**Theorem 3.1.2**  Let $\mathcal{G}$ be a field on the set $\Omega$. Every probability measure $P : \mathcal{G} \to [0, 1]$ admits a unique extension $P_{\text{Ext}} : \sigma(\mathcal{G}) \to [0, 1]$ to the $\sigma$-field $\sigma(\mathcal{G})$ generated by $\mathcal{G}$. 
3.2 Example 1: Infinite coin tosses

The experiment $E$ consists in repeating a coin toss under “identical and independent conditions” with a fair coin (so that the likelihood of occurrence of Head is the same as that of Tail). It is convenient to take the sample space $\Omega$ to be $\{0, 1\}^{\infty}$, i.e.,

$$\Omega = \{ \omega = (\omega_1, \omega_2, \ldots) : \omega_k \in \{0, 1\}, \ k = 1, 2, \ldots \}$$

with the understanding that $\omega_k = 1$ (resp. $\omega_k = 0$) if the $k^{th}$ toss yields Head (resp. Tail). Note that $\Omega$ has the same cardinality as the unit interval $[0, 1]$ (hence is uncountable). How should we construct $\mathcal{F}$ (and $\mathbb{P}$)?

It is natural to require that for any $n = 1, 2, \ldots$, any collection of outcomes determined by the first $n$ tosses should be an event in $\mathcal{F}$ – After all one should expect that the model we are seeking to construct would also contain a model for each of the finite toss experiments. In particular, with any given binary sequence $(b_1, \ldots, b_n)$ of length $n$, consider

$$F_n(b_1, \ldots, b_n) \equiv \left\{ \omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_k = b_k \text{ \ for } k = 1, \ldots, n \right\}. \tag{3.7}$$

It is plain that $\mathcal{F}$ must at least contain these events which are determined by a finite number of coin tosses, namely

$$F_n(b_1, \ldots, b_n) \in \mathcal{F}. \tag{3.8}$$

Fairness (which is essentially a uniformity condition) requires that $\mathbb{P}[F_n(b_1, \ldots, b_n)]$ should not depend on $(b_1, \ldots, b_n)$. As there are $2^n$ distinct sets of the form (3.7) and the decomposition

$$\bigcup_{(b_1, \ldots, b_n) \in \{0, 1\}^n} F_n(b_1, \ldots, b_n) = \Omega,$$

holds, we readily conclude that

$$\mathbb{P}[F_n(b_1, \ldots, b_n)] = 2^{-n} \tag{3.9}$$

since the collection

$$\mathcal{G} \equiv \left\{ F_n(b_1, \ldots, b_n), \ b_1, \ldots, b_n \in \{0, 1\} \right\}$$

forms a collection of non-overlapping events.

It is therefore natural to take $\mathcal{F} = \sigma(\mathcal{G})$ where the generator $\mathcal{G}$ is the collection $\mathcal{G}$. Although the $\sigma$-field $\mathcal{F}$ so defined is very large, it does not coincide with $\mathcal{P}(\Omega)$.
It does however contain interesting events that do not depend on a given finite number of tosses. For instance, consider the event \( F \) given by
\[
F = \left\{ \omega = (\omega_1, \omega_2, \ldots) \in \Omega : \text{A even number of tosses is needed before observing the first Head} \right\} = \bigcup_{k=1}^{\infty} E_{2k}
\]
where for each \( k = 1, \ldots \) we have defined \( E_k = F_k(0, \ldots, 0, 1) \). The event \( F \) is clearly an element of \( \mathcal{F} \).

Measure Theory tells us that there exists a unique probability measure \( \mathbb{P} \) on \( \mathcal{F} \) so that (3.9) holds for all \( n = 1, 2, \ldots \). In particular, it is easy to check that
\[
\mathbb{P}[F] = \mathbb{P}\left[ \bigcup_{k=1}^{\infty} E_{2k} \right] = \sum_{k=1}^{\infty} \mathbb{P}[E_{2k}] \quad \text{[By the required } \sigma \text{-additivity]}
\]
\[
= \sum_{k=1}^{\infty} 2^{-2k} = \frac{2^{-2}}{1 - 2^{-2}} = \frac{1}{3}.
\]
(3.10)

3.3 Example 2: Selecting a point at random in the interval \([0, 1]\)

Consider the random experiment where a point is selected at random in the finite interval \([0, 1]\). As it is appropriate to take \( \Omega = [0, 1] \), we now seek to turn the non-countable interval \([0, 1]\) into a measurable space on which likelihood of occurrence can be defined through a probability measure. Intuitively, following the general approach outlined in Section 3.1 we could proceed as follows to define \( \mathcal{F} \) and \( \mathbb{P} \) (denoted here \( \lambda \) for Lebesgue measure).

We begin with a well-known fact of topology on \( \mathbb{R} \): A subset \( U \) of \( \mathbb{R} \) is said to be open if for every \( x \) in \( U \), there exists an interval \( I_x \) of the form \((a_x, b_x)\) such that \( x \in I_x \) and \( I_x \subseteq U \). A set \( F \) is said to be closed if its complement \( F^c \) is open. It is elementary to check that intervals of the form \((a, b)\) (with \( a < b \) in \( \mathbb{R} \cup \{\pm \infty\} \)) are indeed open.

**Fact 3.3.1** 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
\[
U = \bigcup_{i \in I} J_i \quad \text{with} \quad J_k \cap J_\ell = \emptyset, \quad k \neq \ell, \quad k, \ell \in I.
\]
(3.11)
3.3. EXAMPLE 2: SELECTING A POINT AT RANDOM IN THE INTERVAL $[0, 1]$

To define the appropriate $\sigma$-field $\mathcal{F}$ and the probability measure $\lambda$ on it, it is natural to proceed as follows:

(i) First it is natural to require that *singletons* be events should be elements of the $\sigma$-field $\mathcal{F}$. Indeed, the model should allow one to answer questions such as “what is the probability that $\frac{2}{3}$ was selected?” As in Section 1.4, the assumption of uniform selection again would be recast as requiring $\lambda(\{\omega\})$ to be *independent* of $\omega$, and the arguments developed there also imply

$$\lambda(\{\omega\}) = 0, \quad \omega \in \Omega$$

upon using the fact that $\Omega$ contains finite subsets of unbounded cardinality. It immediately follows that $\lambda(E) = 0$ for every countable subset $E$ of $\Omega$. Furthermore, $\lambda(F) = 1$ for every subset $F$ of $\Omega$ whose complement $E^c$ (in $\Omega$) is countable.

(ii) Thus, in order to extend the definition of $\lambda$ to non-countable sets, it appears that additional constraints associated with uniform selection need to be leveraged. For instance, if a point is selected uniformly at random in $(0, 1)$, it is natural to assume that the likelihood of selecting a point in an interval $[a, b] \subseteq \Omega$ should depend only on the *size* of the interval, and not on its *location*, say

$$\lambda([a, b]) = b - a.$$

(iii) Note that (3.12) is compatible with (3.13): We get back $\lambda(\{a\}) = 0$ for all $a$ in $\Omega$ by taking $a = b$ in (3.13). It follows that

$$\lambda((a, b)) = b - a$$

upon noting that $[a, b] = \{a\} \cup (a, b) \cup \{b\}$.

(iv) The union of countable collections of open intervals should be in the $\sigma$-field $\mathcal{F}$. Therefore, by Fact 3.3.1, we see that $\mathcal{F}$ should include every open set $U \subseteq [0, 1]$ with $\sigma$-additivity requiring

$$\lambda(U) = \sum_{i \in J} \lambda(J_i)$$

where the notation and assumptions are the ones used in Fact 3.3.1.

(v) A set $F$ of $[0, 1]$ being closed if and only $F^c$ is open, we conclude that every closed set $F \subseteq (0, 1)$ must also also belong to $\mathcal{F}$ with $\lambda(F) = 1 - \lambda(F^c)$.

(vi) Clearly, any countable union (resp. intersection) of open subsets should be in $\mathcal{F}$. It follows that any countable union (resp. intersection) of closed subsets should be also in $\mathcal{F}$.

This process can be repeated countably many times, but would not generate a $\sigma$-field on $[0, 1]$. Still it naturally leads to the following definition.
CHAPTER 3. BUILDING PROBABILITY MODELS

**Definition 3.3.1** Let $\mathcal{O}([0, 1])$ denote the collection of all sets contained in $[0, 1]$. The $\sigma$-field $\sigma(\mathcal{O}([0, 1]))$ is called the Borel $\sigma$-field on $[0, 1]$ and is denoted by $\mathcal{B}([0, 1])$. An element of $\mathcal{B}([0, 1])$ is known as a Borel subset of $[0, 1]$.

The discussion leading to this definition suggests that we take $\mathcal{F} = \mathcal{B}([0, 1])$ as the natural $\sigma$-field over which to define $\lambda$. It should be noted that $\mathcal{B}([0, 1]) \neq \mathcal{P}([0, 1])$.

Upon making use of Fact 3.3.1 we can readily check that the Borel $\sigma$-field on $[0, 1]$ can also be characterized as the smallest $\sigma$-field generated by the open intervals in $[0, 1]$.

**Fact 3.3.2** We have

$$\mathcal{B}([0, 1]) = \sigma(\mathcal{I}([0, 1]))$$

where $\mathcal{I}([0, 1])$ denotes the collection of all open intervals contained in $[0, 1]$, with

$$\mathcal{I}([0, 1]) \equiv \{(a, b) : 0 \leq a < b \leq 1\}.$$

**Definition 3.3.2** For each $p = 1, 2, \ldots$, let $\mathcal{O}(\mathbb{R}^p)$ denote the collection of all open sets contained in $\mathbb{R}^p$. The $\sigma$-field $\sigma(\mathcal{O}(\mathbb{R}^p))$ is called the Borel $\sigma$-field on $\mathbb{R}^p$ and is denoted by $\mathcal{B}(\mathbb{R}^p)$. Thus,

$$\mathcal{B}(\mathbb{R}^p) \equiv \sigma(\mathcal{O}(\mathbb{R}^p)).$$

The general definition of a Borel $\sigma$-field uses the collection of open sets as a generator for in higher-dimensions there are no intervals!

### 3.4 Product probability spaces

The following situation often arises in applications: A number of $k$ distinct random experiments, say $\mathcal{E}_1, \ldots, \mathcal{E}_k$, are being individually studied. Thus, for each $\ell = 1, \ldots, k$, let $(\Omega_\ell, \mathcal{F}_\ell, \mathbb{P}_\ell)$ be the probability triple constructed to model the random experiment $\mathcal{E}_\ell$. At this point, without any further information it is not possible to formulate statements concerning the likelihood that events attached to different experiments simultaneously occur. This is so because no single probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ has yet been specified to provide likelihood assignments for the occurrence of events associated with more than one experiment.

There is however a situation that is often encountered: There might an intuitive feeling that these experiments are somewhat “unrelated” in the sense that the
3.5. EXERCISES

outcomes of different experiments are not affected by each other, suggesting independence. To make sense of this modeling assumption we proceed as follows:

First we construct the sample space $\Omega$ as the Cartesian product of the individual sample spaces. Thus,

$$\Omega \equiv \Omega_1 \times \ldots \times \Omega_k.$$ 

We then construct the product $\sigma$-field $\mathcal{F}$ on $\Omega$ given by

$$\mathcal{F} \equiv \sigma (\mathcal{F}_1 \times \ldots \times \mathcal{F}_k) = \sigma \left\{ F_1 \times \ldots \times F_k, \quad F_\ell \in \mathcal{F}_\ell, \quad \ell = 1, \ldots, k \right\}.$$ 

The product probability measure $\mathbb{P}$ is defined on $\mathcal{F}$ as follows: For any rectangle

$$R = F_1 \times \ldots \times F_k, \quad F_\ell \in \mathcal{F}_\ell$$

we set

$$(3.15) \quad \mathbb{P}[R] \equiv \prod_{\ell=1}^{k} \mathbb{P}_\ell[F_\ell].$$

So far, $\mathbb{P}$ is defined only on $\mathcal{F}_1 \times \ldots \times \mathcal{F}_p$. However, Measure Theory guarantees that there exists a unique probability measure on the $\sigma$-field $\mathcal{F}$ such that (3.15) holds. We are now in a position to make statements about the likelihood of occurrence of events that involve multiple experiments. In particular, we have the following important modeling fact which captures the original idea that these experiments were indeed probabilistically unrelated to each other: Consider the events in $\mathcal{F}$ given by

$$E_1 = A_1 \times \Omega_2 \times \ldots \times \Omega_k$$
$$E_2 = \Omega_1 \times A_2 \times \ldots \times \Omega_k$$
$$\vdots$$
$$E_p = \Omega_1 \times \Omega_2 \times \ldots \times A_k.$$ 

Under $\mathbb{P}$, it is easy to check that these events are mutually independent with

$$\mathbb{P}[E_\ell] = \mathbb{P}_\ell[A_\ell], \quad \ell = 1, \ldots, k.$$ 

3.5 Exercises
Chapter 4

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 interest is not so much 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) which we now discuss.

4.1 Random variables

Definition 4.1.1 Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, a mapping $X : \Omega \to \mathbb{R}^p$ is a random variable (rv) if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \quad B \in \mathcal{B}(\mathbb{R}^p).$$

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 2.3.1 with $S = \Omega$ and $\mathcal{S} = \mathcal{F}$. We shall often write $[X \in B]$ in lieu of $X^{-1}(B)$ and $\mathbb{P}[X \in B]$ for $\mathbb{P}[[X \in B]]$.

As discussed in Section 2.4, the mapping $X : \Omega \to \mathbb{R}^p$ is a rv if and only

$$\{\omega \in \Omega : X_k(\omega) \leq a_k, \; k = 1, \ldots, p\} \in \mathcal{F}, \quad (a_1, \ldots, a_p) \in \mathbb{R}^p$$

where it is understood that

$$X(\omega) = (X_1(\omega), \ldots, X_p(\omega)), \quad \omega \in \Omega.$$

This representation induces component mappings $X_1, \ldots, X_p : \Omega \to \mathbb{R}$ defined in an obvious manner.
The condition (4.2) can also be rewritten as

\[
\bigcap_{k=1}^p [X_k \leq a_k] \in \mathcal{F}, \quad (a_1, \ldots, a_p) \in \mathbb{R}^p.
\]  

Here as well, we conclude to the following noteworthy fact which translates Lemma 2.4.1 in the context of rvs.

**Fact 4.1.1** The mapping \( X : \Omega \to \mathbb{R}^p \) is 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.

### 4.2 Probability distribution functions

Consider an \( \mathbb{R}^p \)-valued rv \( X : \Omega \to \mathbb{R}^p \).

**Definition 4.2.1** The probability distribution (function) of the rv \( X \) is the mapping \( F_X : \mathbb{R}^p \to [0, 1] \) defined by

\[
F_X(x) \equiv \mathbb{P}[X \in (-\infty, x_1] \times \ldots \times (-\infty, x_p)] = \mathbb{P}[X_1 \leq x_1, \ldots, X_p \leq x_p], \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.
\]  

with the notation \( X = (X_1, \ldots, X_p) \).

It turns out that there is as much probabilistic information in the probability distribution \( F_X : \mathbb{R}^p \to [0, 1] \) of the rv \( X \) as in \( \{\mathbb{P}[X \in B], B \in \mathcal{B}(\mathbb{R}^p)\} \)

In fact, knowledge of \( F_X : \mathbb{R}^p \to \mathbb{R} \) allows a unique reconstruction of \( \{\mathbb{P}[X \in B], B \in \mathcal{B}(\mathbb{R}^p)\} \).

This is a consequence of Carathéodory’s Theorem.

The set function \( \mathbb{P}_X : \mathcal{B}(\mathbb{R}^p) \to [0, 1] \) defined by

\[
\mathbb{P}_X[B] \equiv \mathbb{P}[X \in B], \quad B \in \mathcal{B}(\mathbb{R}^p)
\]

is a probability measure on \( \mathcal{B}(\mathbb{R}^p) \), thereby suggesting the following 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 \( \omega \) in \( \mathcal{E} \) can only be known if the experiment \( \mathcal{E} \) is
realized, then outcome \( X(\omega) \) of the experiment \( E_X \) will be known only after \( \omega \) has been observed and the numerical value \( X(\omega) \) evaluated.

It is therefore natural to think of the triple \((\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)\) as the probability model associated with random experiment \( E_X \). If there is interest only in this associated experiment (and not in the underlying experiment \( E \)), we need only focus on the triple \((\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mathbb{P}_X)\) in view of the fact that the probability measure \( \mathbb{P}_X \) carries all the probabilistic information related to it. Therefore, 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. In particular, 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 \( F_X \).

**Marginalization** The following situation arises in many settings: Consider rvs \( X_1 : \Omega \rightarrow \mathbb{R}^{p_1}, \ldots, X_k : \Omega \rightarrow \mathbb{R}^{p_k} \). We can alternatively view this collection of rvs as a single rv \( X : \Omega \rightarrow \mathbb{R}^p \) given by

\[
X = (X_1, \ldots, X_k)
\]

with \( p = p_1 + \ldots + p_k \). As usual we have

\[
\mathbb{P}[X \in B] = \mathbb{P}[X_1 \in B_1, \ldots, X_k \in B_k], \quad B_\ell \in \mathcal{B}(\mathbb{R}^{p_\ell}) \]

with

\[
B = B_1 \times \ldots \times B_k.
\]

In particular, by taking \( B_\ell = (-\infty, x_\ell] \) with arbitrary \( x_\ell \) in \( \mathbb{R}^{p_\ell} \) for each \( \ell = 1, \ldots, k \), we conclude that the probability distribution function of the rv \( X : \Omega \rightarrow \mathbb{R}^p \) (or equivalently, the joint probability distribution function of the rvs \( X_1, \ldots, X_k \)) is given by

\[
F_{(X_1, \ldots, X_k)}(x_1, \ldots, x_k) = \mathbb{P}[X_1 \leq x_1, \ldots, X_k \leq x_k], \quad x_\ell \in \mathbb{R}^{p_\ell}, \quad \ell = 1, \ldots, k.
\]

(4.5)

Now, pick any subset \( J \subseteq \{1, \ldots, k\} \) with \( 1 \leq |J| \). It is elementary to check that

\[
F_{(X_\ell, \ell \in J)}(x_\ell, \ell \in J) = \mathbb{P}[X_\ell \leq x_\ell, \ell \in J] = \lim_{x_\ell \rightarrow \infty, \ell \in J} \mathbb{P}[X_1 \leq x_1, \ldots, X_k \leq x_k]
\]

(4.6)

\[
= \lim_{x_\ell \rightarrow \infty, \ell \in J} F_{(X_1, \ldots, X_k)}(x_1, \ldots, x_k), \quad x_\ell \in \mathbb{R}^{p_\ell}, \quad \ell \in J
\]
as we have used the facts

\[
\lim_{x_\ell \to \infty, \ell \in J^c} \bigcap_{\ell \in J^c} [X_\ell \leq x_\ell] = \Omega
\]

and

\[
[X_1 \leq x_1, \ldots, X_k \leq x_k] = (\bigcap_{\ell \in J} [X_\ell \leq x_\ell]) \cap (\bigcap_{\ell \in J^c} [X_\ell \leq x_\ell]).
\]

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 \(x_\ell = \infty\) in \(F_{(X_1, \ldots, X_k)}\) for each \(\ell\) in \(J^c\).

### 4.3 Properties of probability distribution functions

#### The case \(p = 1\)

It is easy to see that the following properties hold when \(p = 1\).

**Proposition 4.3.1** Given a rv \(X : \Omega \to \mathbb{R}\) with probability distribution function \(F_X : \mathbb{R} \to [0, 1]\), the following properties hold:

(i) Monotonicity:

\[
F_X(x) \leq F_X(y), \quad x < y, \quad x, y \in \mathbb{R}
\]

(ii) Right-continuity:

\[
\lim_{y \downarrow x} F_X(y) = F_X(x), \quad x \in \mathbb{R}
\]

(iii) Existence of a left limit:

\[
\lim_{y \uparrow x} F_X(y) = F_X(x-), \quad x \in \mathbb{R}
\]

with \(\Pr[X = x] = F_X(y) - F_X(x-), \quad x \in \mathbb{R}\)

(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 \(\Pr\) one we note that with \(x\) and \(y\) in \(\mathbb{R}\), we have \([X \leq x] \subseteq [X \leq y]\) as soon as \(x < y\). Indeed, we have

\[
\Pr[X \leq y] = \Pr[X \leq x] + \Pr[x < X \leq y]
\]

or equivalently,

\[
F_X(y) - F_X(x) = \Pr[x < X \leq y].
\]
(ii) Pick \( x \in \mathbb{R} \), and let \( \{y_n, n = 1, 2, \ldots \} \) denote a decreasing sequence in \( \mathbb{R} \) such that \( x \leq y_n \) for each \( n = 1, 2, \ldots \). 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 set sequence with

\[
\cap_{n=1}^{\infty} [x < X \leq y_n] = \emptyset
\]

and the conclusion \( \lim_{n \to \infty} \mathbb{P}[x < X \leq y_n] = 0 \) follows, whence \( \lim_{n \to \infty} F_X(y_n) = F_X(x) \).

(iii) Similarly, pick \( x \in \mathbb{R} \), and let \( \{y_n, n = 1, 2, \ldots \} \) denote an increasing sequence in \( \mathbb{R} \) such that \( y_n \leq x \) for each \( n = 1, 2, \ldots \). 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 an increasing set sequence with

\[
\cup_{n=1}^{\infty} [y < X \leq x_n] = [X = x]
\]

This time we get \( \lim_{n \to \infty} \mathbb{P}[y_n < X \leq x] = \mathbb{P}[X = x] \), and the limit \( \mathbb{P}[X = x] \) being independent of the sequence, the desired result follows.

(iv) Finally,

The case \( p \geq 1 \) As the quantity \( \mathbb{P}[x_k < X_k \leq y_k] \) 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

\[
\mathbb{P}[a < X_1 \leq b, \alpha < X_2 \leq \beta]
\]

\[
= \mathbb{P}[X_1 \leq b, \alpha < X_2 \leq \beta] - \mathbb{P}[X_1 \leq a, \alpha < X_2 \leq \beta]
\]

\[
= (\mathbb{P}[X_1 \leq b, X_2 \leq \beta] - \mathbb{P}[X_1 \leq b, X_2 \leq \alpha])
\]

\[
- (\mathbb{P}[X_1 \leq a, X_2 \leq \beta] - \mathbb{P}[X_1 \leq a, X_2 \leq \alpha])
\]

\[
= (F_{(X_1,X_2)}(b, \beta) - F_{(X_1,X_2)}(b, \alpha))
\]

\[
- (F_{(X_1,X_2)}(a, \beta) - F_{(X_1,X_2)}(a, \alpha)), \quad a < b
\]

\[
\alpha < \beta
\]

(4.7)

**Proposition 4.3.2** 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:
4.4 Probability distribution functions \((p = 1)\)

Turning these properties into a definition we introduce the concept of a probability distribution (function).

**Definition 4.4.1** A probability distribution (function) on \(\mathbb{R}\) is any mapping \(F : \mathbb{R} \rightarrow [0, 1]\) such that

- **Monotonicity:** \(F(x) \leq F(y), \ x, y \in \mathbb{R}\)
- **Right-continuity:** \(\lim_{y \downarrow x} F(y) = F(x), \ x \in \mathbb{R}\)
- **Existence of left limits:** \(\lim_{y \uparrow x} F(y) = F(x-) \ x \in \mathbb{R}\)
- **Behavior at infinity:** Monotonically, we have
  \[
  \lim_{x \rightarrow -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} F(x) = 1
  \]
4.5. PROBABILITY DISTRIBUTION FUNCTIONS ($P \geq 1$)

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 4.4.1. The converse allows us to equate a probability distribution function with a rv understood as a Borel mapping defined on some probability triple.

**Lemma 4.4.1** For any probability distribution function $F : \mathbb{R} \to [0, 1]$, there exists a probability triple $(\Omega^*, F^*, \mathbb{P}^*)$ and a rv $X^* : \Omega^* \to \mathbb{R}$ defined on it such that

$$\mathbb{P}^* [X^* \leq x] = F(x), \quad x \in \mathbb{R}.$$ 

This is the basis of Monte-Carlo simulation. There exists a multi-dimensional analog to this fact to be discussed later on.

**Proof.** Take $\Omega^* = [0, 1], F^* = \mathcal{B}([0, 1])$ and $\mathbb{P}^* = \lambda$. Define the rv $X^* : \Omega^* \to \mathbb{R}$ by setting

$$X^*(\omega^*) = F^-(\omega^*), \quad \omega^* \in [0, 1]$$

where $F^- : [0, 1] \to [-\infty, \infty]$ is the **generalized inverse** of $F$ given by

$$F^-(u) = \inf \{x \in \mathbb{R} : u \leq F(x)\}, \quad 0 \leq u \leq 1$$

with the understanding that $F^-(u) = \infty$ if the defining set is empty, i.e., $F(x) < u$ for all $x$ in $\mathbb{R}$.

4.5 Probability distribution functions ($p \geq 1$)

4.6 Discrete distributions

**Definition 4.6.1** A rv $X : \Omega \to \mathbb{R}^p$ is a **discrete** rv if there exists a countable subset $S \subseteq \mathbb{R}^p$ such that

$$\mathbb{P} [X \in S] = 1.$$ 

We refer to the countable $S$ entering this definition as the **support** of the discrete rv $X$. Under this definition, the basic relation

$$\mathbb{P} [X \in B] = \sum_{x \in S \cap B} \mathbb{P} [X = x], \quad B \in \mathcal{B}(\mathbb{R}^p)$$

holds. It is often more convenient to characterize the distributional properties of the rv $X$ through the **probability mass function** (pmf) $p_X \equiv (p_X(x), x \in S)$ of the rv $X$ given by

$$p_X(x) = \mathbb{P} [X = x], \quad x \in S.$$
Note that
\[ 0 \leq p_X(x) \leq 1, \quad x \in S \quad \text{and} \quad \sum_{x \in S} p_X(x) = 1. \]

This leads to the following definition:

**Definition 4.6.2** With \( S \) a countable subset of \( \mathbb{R}^p \), a pmf with support on \( S \) is any collection \( p = (p(x), x \in S) \) such that
\[ 0 \leq p(x) \leq 1, \quad x \in S \quad \text{and} \quad \sum_{x \in S} p(x) = 1. \]

The analog of Lemma 4.4.1 for pmfs takes the following form.

**Lemma 4.6.1** With \( S \) a countable subset of \( \mathbb{R}^p \), for any pmf \( p = (p(x), x \in S) \) with support on \( S \), there exists a probability triple \((\Omega^*, \mathcal{F}^*, \mathbb{P}^*)\) and a rv \( X^* : \Omega^* \rightarrow \mathbb{R} \) defined on it such that
\[ \mathbb{P}^*[X^* = x] = p(x), \quad x \in S. \]

**Proof.** Take \( \Omega^* = S, \mathcal{F}^* = \mathcal{P}(S) \) and define the probability measure \( \mathbb{P}^* \) on \( \mathcal{P}(S) \) by setting
\[ \mathbb{P}^*[E] = \sum_{\omega^* \in E} p(\omega^*), \quad E \in \mathcal{P}(S). \]

The rv \( X^* : \Omega^* \rightarrow \mathbb{R}^p \) defined by
\[ X^*(\omega^*) = \omega^*, \quad \omega^* \in \Omega^* \]
is clearly a discrete rv with support \( S \). Furthermore,
\[ \mathbb{P}^*[X^* = x] = p(x), \quad x \in S \]
since \( [X^* = x] = \{x\} \), and \( p \) is indeed the pdf of \( X^* \) under \( \mathbb{P}^* \).

Well-known examples of discrete rvs (and of their distributions) include

**Bernoulli** \( \text{Ber}(p) \) \((0 \leq p \leq 1)\): Here \( S = \{0, 1\} \) with
\[
(4.8) \quad p(1) = p \quad \text{and} \quad p(0) = 1 - p.
\]
4.7 Absolutely continuous distributions

**Binomial** $\text{Bin}(n; p) \ (n = 1, 2, \ldots \text{ and } 0 \leq p \leq 1)$: Here $S = \{0, 1, \ldots, n\}$ with

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

(4.9)

**Poisson** $\text{Poi}(\lambda) \ (\lambda > 0)$: Here $S = \mathbb{N}$ with

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots$$

(4.10)

**Geometric** $\text{Geo}(p) \ (0 \leq p \leq 1)$: Here $S = \mathbb{N}$ with

$$p(x) = p(1-p)^x, \quad x = 0, 1, \ldots$$

(4.11)

Sometimes, with $S = \mathbb{N}_0$ and

$$p(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

(4.12)

### 4.7 Absolutely continuous distributions

A rv $X : \Omega \to \mathbb{R}^p$ is an (absolutely) continuous rv if there exists a Borel mapping $f_X : \mathbb{R}^p \to \mathbb{R}_+$ such that

$$\mathbb{P}[X_i \leq x_i, \ i = 1, \ldots, p] = \int_{-\infty}^x f_X(\xi) d\xi, \quad x = (x_1, \ldots, x_p) \in \mathbb{R}^p.$$  

(4.13)

The mapping $f_X : \mathbb{R}^p \to \mathbb{R}_+$ is known as the probability density function of the (absolutely) continuous rv $X$. In general the integral (4.13) is understood as a Lesbegue integral. However, in most applications the integrand in (4.13) is a well behaved function and the corresponding integral can be performed through Riemann integration.

Well-known examples of continuous rvs (and of their distributions) are given next: Throughout $f : \mathbb{R} \to \mathbb{R}_+$ denotes the probability distribution function while $F : \mathbb{R} \to [0, 1]$ is the corresponding probability distribution function.

**Uniform** $U(a, b) \ (a < b \text{ in } \mathbb{R})$: Its density is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

(4.14)
and

\[ F(x) = \begin{cases} 
0 & \text{if } x \leq a \\
\frac{x-a}{b-a} & \text{if } a \leq x \leq b \\
1 & \text{if } b \leq x
\end{cases} \]  
(4.15)

**Exponential** \( \text{Exp}(\lambda) (\lambda > 0) \):

\[ f(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\lambda e^{-\lambda x} & \text{if } x \geq 0
\end{cases} \]  
(4.16)

and

\[ F(x) = 1 - e^{-\lambda x^+}, \quad x \in \mathbb{R} \]  
(4.17)

**Gaussian** \( N(m, \sigma^2) (m \text{ in } \mathbb{R} \text{ and } \sigma > 0) \):

\[ f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in \mathbb{R} \]  
(4.18)

There is no closed form expression for the integral

\[ F(x) = \int_{-\infty}^{x} f(t) dt, \quad x \in \mathbb{R} \]  
(4.19)

We refer to the case \( m = 0 \) and \( \sigma^2 = 1 \) as the *standard Gaussian* distribution; sometimes it is also known as the *normal* distribution. With these parameter values we follow customs and change the notation to write

\[ \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R} \]  
(4.20)

and

\[ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad x \in \mathbb{R} \]  
(4.21)

**Cauchy** \( C(m, a) (m, a \text{ in } \mathbb{R}) \):

\[ f(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R} \]  
(4.22)

and

\[ F(x) = \, \, x \in \mathbb{R} \]  
(4.23)
4.8. FUNCTIONS OF RVS

\textbf{Pareto} \( \text{Par}(a, \nu) \) \((a > 0 \text{ and } \nu > 0)\):

\begin{equation}
  f(x) = \begin{cases} 
  0 & \text{if } x < 0 \\
  \frac{\nu}{a} \left(\frac{a}{a+x}\right)^{\nu+1} & \text{if } x \geq 0 
  \end{cases}
\end{equation}

(4.24)

and

\begin{equation}
  F(x) = \begin{cases} 
  0 & \text{if } x < 0 \\
  1 - \left(\frac{a}{a+x}\right)^{\nu} & \text{if } x \geq 0
  \end{cases}
\end{equation}

(4.25)

4.8 Functions of rvs

Consider a rv \( X : \Omega \to \mathbb{R}^p \). For any Borel mapping \( g : \mathbb{R}^p \to \mathbb{R}^q \), define the mapping \( Y : \Omega \to \mathbb{R}^q \) by composing the rv \( X : \Omega \to \mathbb{R}^p \) with \( g \), namely

\[ Y(\omega) = g(X(\omega)), \quad \omega \in \Omega. \]

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{equation}
  \mathbb{P}[Y \in B] = \mathbb{P}[g(X) \in B] = \mathbb{P}[X \in g^{-1}(B)], \quad B \in \mathbb{R}^q
\end{equation}

(4.26)

\textbf{Discrete rvs} \quad \text{Assume that the rv } X : \Omega \to \mathbb{R}^p \text{ is a rv with support } S_X \subseteq \mathbb{R}^p \text{ so that } \mathbb{P}[X \in S_X] = 1. \text{ The set } S_Y \equiv \{g(x) : x \in S_X\} \text{ is also countable subset of } \mathbb{R}^q, \text{ and it is plain that } \mathbb{P}[Y \in S_Y] = 1. \text{ Thus, } Y \text{ is a discrete rv with support } S_Y, \text{ and its pmf is easily determined: Indeed,}

\begin{equation}
  \mathbb{P}[Y \in B] = \mathbb{P}[g(X) \in B] = \sum_{x \in S_X} \mathbb{P}[X = x, g(X) \in B] = \sum_{x \in S_X : g(x) \in B} \mathbb{P}[X = x], \quad B \in \mathbb{R}^q
\end{equation}

(4.27)

Therefore,

\[ \mathbb{P}[Y = y] = \sum_{x \in S_X : g(x) = y} \mathbb{P}[X = x], \quad y \in S_Y. \]
CHAPTER 4. RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Continuous rvs If the rv $X : \Omega \to \mathbb{R}^p$ is a continuous rv with probability density function $f_X : \mathbb{R}^p \to \mathbb{R}^+$, then

$$P\{Y \in B\} = P\{X \in g^{-1}(B)\} = \int_{g^{-1}(B)} f_X(t) dt, \quad t \in \mathbb{R}^q$$

(4.28)

In general, despite this last expression, it is not true that the rv $Y$ is itself a continuous rv. A trivial case arises when using the constant mapping $g : \mathbb{R}^p \to \mathbb{R}^q : x \to b$ for some given value $b$, in which case $Y$ is a discrete rv with support reduced to the singleton $\{b\}$.

4.9 Independence of rvs Consider a collection of rvs $\{X_i, i \in I\}$ which are all defined on some probability triple $(\Omega, \mathcal{F}, P)$. 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 4.9.1 With $I$ finite, the rvs $\{X_i, i \in I\}$ are mutually independent if for each selection of $B_i$ 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 1.6, we see that the rvs $\{X_i, i \in I\}$ are mutually independent according to Definition 4.9.1 if the conditions

$$P\{X_j \in B_j, j \in J\} = \prod_{j \in J} P\{X_j \in B_j\}, \quad B_j \in \mathcal{B}(\mathbb{R}^{p_j})$$

(4.29)

$$1 \leq |J| \leq |I|$$

all hold. It is now easy to see that the rvs $\{X_i, i \in I\}$ are mutually independent if only the smaller set of conditions

$$P\{X_i \in B_i, i \in I\} = \prod_{i \in I} P\{X_i \in B_i\}, \quad B_i \in \mathcal{B}(\mathbb{R}^{p_i})$$

(4.30)

$$i \in I$$

hold. Indeed, while (4.29) implies (4.30), it is easy to see that (4.30) implies (4.29) – Just take $B_j = \mathbb{R}^{p_j}$ for $j$ in $J^c$!
Definition 4.9.2 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.

With \( k \) some fixed integer, in what follows consider a collection of rvs \( X_1, \ldots, X_k \) 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 an \( \mathbb{R}^{p_i} \)-valued rv for some positive integer \( p_i \).

Lemma 4.9.1 The rvs \( \{X_1, \ldots, X_k\} \) are mutually independent if and only if

\[
F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i), \quad x_i \in \mathbb{R}^{p_i}
\]

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 rv \((X_1, \ldots, X_k)\) with \( p = p_1 + \ldots + p_k \).

Note that the mapping \( \Omega \to \mathbb{R}^p : \omega \to (X_1(\omega), \ldots, X_n(\omega)) \) is indeed an \( \mathbb{R}^p \)-valued rv by virtue of Fact 4.1.1.

4.10 Taking limits

Consider the sequence of \( \mathbb{R} \)-valued rvs \( \{X_n, \ n = 1, 2, \ldots\} \) which are all defined on the same probability triple \((\Omega, \mathcal{F}, \mathbb{P})\). The following mappings \( \Omega \to [\infty, \infty] \) are rvs in the extended sense:

The supremum mapping \( \Omega \to [\infty, \infty] \) defined by

\[
\omega \to \sup_{n \geq 1} X_n(\omega), \quad \omega \in \Omega
\]

The infimum mapping \( \Omega \to [\infty, \infty] \) defined by

\[
\omega \to \inf_{n \geq 1} X_n(\omega), \quad \omega \in \Omega.
\]

The limsup mapping \( \Omega \to [\infty, \infty] \) defined by

\[
\omega \to \lim_{n \to \infty} \sup X_n(\omega), \quad \omega \in \Omega.
\]

The liminf mapping \( \Omega \to [\infty, \infty] \) defined by

\[
\omega \to \lim_{n \to \infty} \inf X_n(\omega), \quad \omega \in \Omega.
\]
It follows that
\[ \Omega^* \equiv \left[ \liminf_{n \to \infty} X_n = \limsup_{n \to \infty} X_n \right] \in \mathcal{F} \]
and on \( \Omega^* \), \( \lim_{n \to \infty} X_n \) 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 \( \mathbb{P}[\Omega^*] = 1 \) it is customary to say that the sequence \( \{X_n, n = 1, 2, \ldots\} \) converges almost surely (a.s.) (under \( \mathbb{P} \)), written
\[ \lim_{n \to \infty} X_n \mathbb{P}-\text{a.s.} \]
In that case, for any rv \( X : \Omega \to \mathbb{R} \) such that
\[ X(\omega) = \lim_{n \to \infty} X_n(\omega), \quad \omega \in \Omega^* \]
we shall write
\[ \lim_{n \to \infty} X_n = X \mathbb{P}-\text{a.s.} \]
Such a rv \( X \) always exists when \( \mathbb{P}[\Omega^*] = 1 \) but is not unique. Existence is immediate since we can always take
\[ X(\omega) \equiv \begin{cases} 
\liminf_{n \to \infty} X_n(\omega) = \limsup_{n \to \infty} X_n(\omega) & \text{if } \omega \in \Omega^* \\
Z(\omega) & \text{if } \omega \notin \Omega^* 
\end{cases} \]
where \( Z : \Omega \to \mathbb{R} \) is some arbitrary rv, and non-uniqueness is obvious.

### 4.11 Exercises

**Ex. 4.1** Consider a pair of rvs \( X, Y : \Omega \to \mathbb{R} \). By direct arguments show that the following mappings \( \Omega \to \mathbb{R} \) are rvs:

a. \( Z = X + Y \)

b. \( Z = \alpha X + \beta Y \) with \( \alpha, \beta \) arbitrary in \( \mathbb{R} \).

c. \( V = \max(X, Y) \) and \( W = \min(X, Y) \)

**Ex. 4.2** With \( m \) in \( \mathbb{R} \) and \( \sigma^2 > 0 \), show the following facts:

a. If \( X \sim \mathcal{N}(m, \sigma^2) \), show that \( \sigma^{-1}(X - m) \sim \mathcal{N}(0, 1) \).

b. Conversely, if \( Y \sim \mathcal{N}(0, 1) \), then \( m + \sigma Y \sim \mathcal{N}(m, \sigma^2) \)
Ex. 4.3 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 \( 0 < p < 1 \)).

a. Develop a probability model \((\Omega, \mathcal{F}, \mathbb{P})\) for this experiment.

b. Define
\[
X(\omega) = \begin{cases} 
\text{The number of tosses before} & \\
\text{the first Head appears in the sample } \omega & \\
\end{cases}, \quad \omega \in \Omega.
\]

Explain why the mapping \( X : \Omega \rightarrow \mathbb{R} \) so defined is indeed a rv. Is it a discrete rv?

c. Find the pdf of this rv, i.e., \( \{ \mathbb{P}[X = m], m = 1, 2, \ldots \} \).

d. On the probability triple constructed in Part a, is it possible to define a rv \( Y : \Omega \rightarrow \mathbb{R} \) which is not a discrete rv?

Ex. 4.4 Consider a mapping \( F : \mathbb{R} \rightarrow \mathbb{R}_+ \) 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 \( F^{-} : \mathbb{R}_+ \rightarrow [−\infty, \infty] \) given by
\[
F^{-}(u) = \inf \{ x \in \mathbb{R} : u \leq F(x) \}, \quad u \geq 0
\]

with \( F^{-}(u) = \infty \) if the set \( \{ x \in \mathbb{R} : u \leq F(x) \} \) is empty.

Assume that \( F \) is a probability distribution function \( F : \mathbb{R} \rightarrow [0, 1] \).

a. What is the value of \( F^{-}(u) \) 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 \( F^{-} : \mathbb{R}_+ \rightarrow [−\infty, \infty] \). Compute \( F^{-}(F(x)) \) 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 \( F^{-}(F(x)) \) for all \( x \) in \( \mathbb{R} \).

Ex. 4.5 Let \( F_1, \ldots, F_n \) denote probability distribution functions \( \mathbb{R} \rightarrow [0, 1] \). Determine which of the following mappings \( G : \mathbb{R} \rightarrow \mathbb{R} \) 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

$$G(x) = \sum_{i=1}^{n} \alpha_i F_i(x), \quad x \in \mathbb{R}$$

b. The product

$$G(x) = F_1(x) \ldots F_n(x), \quad x \in \mathbb{R}$$

c. The product

$$G(x) = F_1(x-)F_1(x), \quad x \in \mathbb{R}$$

d. With $0 < u < 1$,

$$G(x) = 1 - u F_1(x), \quad x \in \mathbb{R}$$

Ex. 4.6 Define the binary rvs

$$\chi_{ij} = 1 [\Gamma_i \cap \Gamma_j \neq \emptyset], \quad i \neq j, \quad i, j = 1, \ldots, n.$$ 

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.

a. Compute the probabilities

$$\mathbb{P} [\Gamma_i \cap \Gamma_j = \emptyset], \quad i \neq j, \quad i, j = 1, \ldots, n.$$ 

b. Do the rvs $\{\chi_{1j}, \ j = 2, \ldots, n\}$ form a collection of mutually independent rvs?

c. Are the rvs $\chi_{12}, \chi_{23}$ and $\chi_{31}$ mutually independent?

Ex. 4.7 For $k = 1, 2$, consider the rv $X_k : \Omega_k \rightarrow \mathbb{R}^p$ 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 $g : \mathbb{R}^p \rightarrow \mathbb{R}^q$, 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.

Ex. 4.8 A rv $X : \Omega \rightarrow \mathbb{R}$ 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., $X =_{st} -X$.

a. Give necessary and sufficient conditions on $F_X : \mathbb{R} \rightarrow [0, 1]$ for the rv $X$ to have a symmetric probability distribution.

b. Specialize your answer to the case when $X$ is a discrete rv with support $S \subseteq \mathbb{R}^p$ and pmf $p = (p(x), \ x \in S)$.  

4.11. EXERCISES

Ex. 4.9 Assume the rv $X : \Omega \to \mathbb{R}$ 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. 4.10 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 $X \sim \mathcal{N}(0, 1)$ but the result holds more generally and requires very little computation. Again the power of probabilistic thinking at work!

Ex. 4.11 Let $X, Y : \Omega \to \mathbb{R}$ be a pair of independent rvs. Assume that the rvs are exponentially distributed in the sense that

$$ \mathbb{P}[X \leq t] = \mathbb{P}[Y \leq t] = 1 - e^{-\lambda t^+}, \quad t \in \mathbb{R} $$

for some $\lambda > 0$ with $t^+ = \max(t, 0)$ for each $t$ in $\mathbb{R}$. Using direct integration arguments\(^1\)

a. Compute $\mathbb{P}[X \leq Y]$ and $\mathbb{P}[X = Y]$.

b. Compute

$$ \mathbb{P}[Z \leq z], \quad z \in \mathbb{R} $$

where we have set

$$ Z = \begin{cases} X & \text{if } X + Y > 0 \\ X + Y & \text{if } X + Y \leq 0. \end{cases} $$

Ex. 4.12 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. Identify its parameter. Generalize to $K$ mutually independent Poisson rvs $X_1, \ldots, X_K$ with

$$ X_k \overset{d}{=} \text{Poi}(\lambda_k), \quad \lambda_k > 0, \quad k = 1, \ldots, K. $$

Carefully explain your reasoning.

\(^1\)[HINT: Recall that the $\mathbb{R}^2$-valued rv $(X, Y)$ is of continuous type [WHY?] and its probability density function is given by $\ldots$]
Ex. 4.13 Let $N$ be a Poisson rv, and let $\{B_n, n = 1, 2, \ldots\}$ be a collection of Bernoulli rvs with

$$
\mathbb{P}[B_n = 1] = 1 - \mathbb{P}[B_n = 0] = p, \quad n = 1, 2, \ldots,
0 < p < 1.
$$

If the rvs $\{N, B_n, n = 1, 2 \ldots\}$ are mutually independent, show that 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 with $X$ and $Y$ Poisson rvs with parameters $\lambda p$ and $\lambda(1 - p)$, respectively. Can you use this result to provide an alternative solution to Exercise? Explain! Again a case of probabilistic reasoning at work!
Chapter 5

Mathematical expectations

The probability distribution function of a rv \( X : \Omega \to \mathbb{R} \) is a complicated object – for all intents 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 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 given in the context of Measure Theory. This is developed in the next section.

5.1 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 \( X : \Omega \to \mathbb{R} \) 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 \( \mathbb{E}[X] \), when it exists, as the expectation of \( X \). This definition for the expectation operator is guided by the following requirements.

A. The expectation \( \mathbb{E}[X] \) 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 \( F_{X'} : \mathbb{R} \to \mathbb{R} \) (under \( \mathbb{P}' \)), then the distributional equality \( F_X = F_{X'} \) implies \( \mathbb{E}[X] = \mathbb{E}'[X'] \). Put differently, the existence of \( \mathbb{E}[X] \) and its value computable once \( F_X \) on the basis of \( F_X \) alone.

The definition of the quantity \( \mathbb{E}[X] \) does not depend on the type of distribution of the rv \( X \), say discrete or absolute continuous, but does coincide with the usual
elementary definitions given in elementary courses in Probability Theory.

**B. 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 $X = 1[A]$, then

$$E[X] = E[1[A]] = \mathbb{P}[A], \quad B \in \mathcal{F}.$$  

**C. Expectations for non-negative rvs** The expectation of non-negative rvs is always well defined (although it could be infinite) with

$$E[X] \geq 0 \quad \text{if} \quad X \geq 0.$$  

**D. 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

$$E[aX + bY] = aE[X] + bE[Y]$$

holds whenever the expression $aE[X] + bE[Y]$ is well defined. In particular this will happen when $E[X]$ and $E[Y]$ are both finite. When either $E[X]$ or $E[Y]$ is infinite, this requirement may put conditions on the sign of $a$ and $b$ for the right-handside to be well defined: No meaning is attributed to $\infty - \infty$. Throughout we will use the following algebraic conventions:

$$0 \cdot (\pm \infty) = \pm \infty$$

and

$$c \pm \infty = c \pm \infty, \quad c \in \mathbb{R}.$$  

A definition of $E[X]$ that meets the requirements (A)-(D) 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 in terms of simple rvs (to be defined next)
- **Step 3:** For arbitrary rvs through a decomposition in positive and negative parts.

We shall see that the expectation operation so constructed has a couple of useful by-products:
5.2. SIMPLE RVS

Monotonicity The operation $X \rightarrow \mathbb{E}[X]$ is monotone in the following sense: If two rvs $X, Y : \Omega \rightarrow \mathbb{R}$ are such that $X \leq Y$, then $Y - X \geq 0$ with both $EX$ and $E[Y]$ are well defined and finite, then $E[Y - X] = E[Y] - E[X]$ by linearity (D) and $E[Y - X] \geq 0$ by (C). As a result, $E[X] \leq E[Y]$. It turns out that a somewhat stronger result holds.

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 $\omega$ in an event $\Omega_*$ with $\mathbb{P}[\Omega_*] = 1$. 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 interchange of limit and expectation

$$
\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X]
$$

does hold!

5.2 Simple rvs

First a couple of definitions and some terminology.

**Definition 5.2.1** With $I$ an index set, an $\mathcal{F}$-partition of $\Omega$ is a non-empty collection $\{A_i, i \in I\}$ of events in $\mathcal{F}$ such that

$$
A_i \cap A_j = \emptyset, \quad i \neq j \quad \text{and} \quad \bigcup_{i \in I} A_i = \Omega.
$$

Such an $\mathcal{F}$-partition is said to be finite (resp. countable) if the index set $I$ is finite (resp. countable).

**Definition 5.2.2** A rv $X : \Omega \rightarrow \mathbb{R}$ is said to be simple rv if it is of the form

$$
X = \sum_{i \in I} a_i 1[A_i]
$$

(5.1)

for some finite $\mathcal{F}$-partition $\{A_i, i \in I\}$ and a collection $\{a_i, i \in I\}$ of scalars in $\mathbb{R}$.

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. The representation (5.2.2) of a simple rv is not necessarily unique. However, in many arguments it is customary to assume that the values $\{a_k, k \in I\}$ are distinct scalars and
that the events \( \{ A_k, \ k \in I \} \) forming the \( \mathcal{F} \)-partition are all non-empty, in which case \( \{ X(\omega), \ \omega \in \Omega \} = \{ a_k, \ k \in I \} \) and
\[
A_k = [X = a_k], \quad k \in I.
\]

We refer to this representation as the generic representation of the simple rv.

Here are some easy facts concerning simple rvs.

\textbf{Fact 5.2.1} If \( X, Y : \Omega \to \mathbb{R} \) are simple rvs, then the rvs \( X + Y \) and \( cX \) (with scalar \( c \)) are also simple rvs.

We close with a useful definition.

\textbf{Definition 5.2.3} The sequence of rvs \( \{ X_n, n = 1, 2, \ldots \} \) is called a staircase approximation 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 monotone increasing in the sense that for every \( \omega \) in \( \Omega \), the sequence \( \{ X_n(\omega), n = 1, 2, \ldots \} \) is monotone increasing with
\[
X_n(\omega) \leq X_{n+1}(\omega) \leq X(\omega), \quad \omega \in \Omega, \quad n = 1, 2, \ldots
\]

(ii) The sequence converges pointwise with
\[
\lim_{n \to \infty} X_n(\omega) = X(\omega), \quad \omega \in \Omega.
\]

\section{5.3 Approximating with simple rvs}

For the purpose of defining expectations the key observation concerning simple rvs is contained in the following lemma.

\textbf{Lemma 5.3.1} 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 existence of the limit at (??) is ensured by the monotonicity (??)

**Proof.** For each \( n = 1, 2, \ldots \), consider the mapping \( X_n : \Omega \to \mathbb{R}_+ \) given by

\[
X_n = \begin{cases} 
  k2^{-n} & \text{if } k2^{-n} \leq X < (k+1)2^{-n}, \\
  0 & \text{otherwise}
\end{cases}
\]

(5.4)

It is easy to 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

\[
A_{n,k} = \begin{cases} 
  [k2^{-n} \leq X < (k+1)2^{-n}] & \text{if } k = 0, 1, \ldots, 4^n - 1 \\
  [X \geq 2^n] & \text{if } k = 4^n
\end{cases}
\]

(5.5)

with associated values \( \{ a_{n,k}, k = 0, 1, \ldots, 4^n - 1, 4^n \} \) given by

\[
a_{n,k} = \begin{cases} 
  k2^{-n} & \text{if } k = 0, 1, \ldots, 4^n - 1 \\
  0 & \text{if } k = 4^n.
\end{cases}
\]

(5.5)

That the partition (5.5) is an \( \mathcal{F} \)-partition is a consequence 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 for each \( n = 1, 2, \ldots \), set

\[
k_n(x) \equiv \lfloor x2^n \rfloor \quad \text{and} \quad x_n \equiv k_n(x)2^{-n}.
\]

It is a simple matter to check that \( 2k_n(x) \leq k_{n+1}(x) \) so that \( x_n \leq x_{n+1} \leq x \) with \( x_n \leq x < x_n + 2^{-n} \). The sequence \( \{ x_n, n = 1, 2, \ldots \} \) is therefore monotone increasing with \( \lim_{n \to \infty} x_n = x \). Obviously \( x_n \geq 0 \) for all \( n = 1, 2, \ldots \) if \( x \geq 0 \).

Note that for each \( n = 1, 2, \ldots \), the rv \( X_n \) can be defined as

\[
X_n = \begin{cases} 
  \lfloor X2^n \rfloor 2^{-n} & \text{if } X < 2^n \\
  0 & \text{if } X \geq 2^n.
\end{cases}
\]

(5.6)

The sequence \( \{ X_n, n1,2,\ldots \} \) whose existence is announced in Lemma 5.3.1 is not unique.
5.4 Defining the expectation of a rv

We are now ready to define the expectation of a rv $X : \Omega \rightarrow \mathbb{R}$. This will be done according to the three steps announced earlier.

**Simple rvs** Assume the rv $X : \Omega \rightarrow \mathbb{R}$ to be a simple rv of the form

$$X = \sum_{i \in I} a_i 1_{[A]}$$

(5.7)

for some finite $\mathcal{F}$-partition $\{A_i, i \in I\}$ with associated collection $\{a_i, i \in I\}$ of scalars in $\mathbb{R}$. For such a rv we define its expectation $\mathbb{E}[X]$ by

$$\mathbb{E}[X] \equiv \sum_{i \in I} a_i \mathbb{P}[A_i].$$

(5.8)

The definition (5.8) does not depend on the particular representation used for the simple rv $X$, and is therefore well posed. This a consequence of the following fact.

**Lemma 5.4.1** Consider a simple rv $X : \Omega \rightarrow \mathbb{R}$ of the form

$$X = \sum_{i \in I} a_i 1_{[A]}$$

(5.9)

for some finite $\mathcal{F}$-partition $\{A_i, i \in I\}$ with associated collection $\{a_i, i \in I\}$ of scalars in $\mathbb{R}$. If it holds that

$$X = \sum_{j \in J} b_j 1_{[B]}$$

(5.10)

for some finite $\mathcal{F}$-partition $\{B_j, j \in J\}$ with associated collection $\{b_j, j \in J\}$ of scalars in $\mathbb{R}$, then we necessarily have

$$\sum_{i \in I} a_i \mathbb{P}[A_i] = \sum_{j \in J} b_j \mathbb{P}[B_j].$$

**Proof.** Indeed, for $i$ in $I$ and $j$ in $J$, $X = a_i$ on $A_i$ and $X = b_j$ on $B_j$, hence $a_i = b_j$ on $A_i \cap B_j$. Since $\{B_j, j \in J\}$ is an finite $\mathcal{F}$-partition, we get

Linearity and monotonicity hold on the class of simple rvs.
Lemma 5.4.2 Assume the rvs $X, Y : \Omega \rightarrow \mathbb{R}$ to be simple rvs. It holds that

$$E[aX + bY] = aE[X] + bE[Y], \quad a, b \in \mathbb{R}.$$  

Furthermore, if $X \leq Y$, then $E[X] \leq E[Y]$.

Given the requirements stated in Section 5.1 the definition (5.8) is the only definition possible: Indeed, we must have

$$E[X] = E \left[ \sum_{i \in I} a_i 1[A_i] \right] = \sum_{i \in I} a_i E[1[A_i]] \quad \text{[By linearity (D)]}$$

(5.11)

$$= \sum_{i \in I} a_i P[A_i] \quad \text{[By (B)]}$$

Non-negative rvs If the rv $X : \Omega \rightarrow \mathbb{R}$ is non-negative, then let $\{X_n, n = 1, 2, \ldots\}$ denote a collection of simple non-negative rvs associated with $X$ whose existence was established in Lemma 5.3.1: We define $E[X]$ by the limiting process

$$E[X] \equiv \lim_{n \rightarrow \infty} E[X_n].$$

(5.12)

Note that $E[X]$ always exists as an element in $[0, +\infty]$ due to the fact that the sequence $\{E[X_n], n = 1, 2, \ldots\}$ is increasing in $\mathbb{R}_+$ by Lemma 5.4.2 (and the monotonicity of the sequence $\{X_n, n = 1, 2, \ldots\}$). Moreover, it can be shown that the limit is independent of the approximating sequence of simple rvs being used:

Lemma 5.4.3 If $\{X_{1,n}, n = 1, 2, \ldots\}$ and $\{X_{2,n}, n = 1, 2, \ldots\}$ are two sequences of simple $\mathbb{R}_+$-valued rvs which monotonically approximate $X$ from below, i.e., for each $k = 1, 2, X_{k,n} \leq X_{k,n+1} \leq X$ for $n = 1, 2, \ldots$ with $\lim_{n \rightarrow \infty} X_{k,n} = X$ pointwise, then

$$\lim_{n \rightarrow \infty} E[X_{1,n}] = \lim_{n \rightarrow \infty} E[X_{2,n}],$$

and their common value is $E[X]$.

The general case Setting $X^+ = \max(0, X)$ and $X^- = \max(0, -X)$, we note the decomposition

$$X = X^+ - X^-$$

(5.13)
as well as the identity
\[ |X| = X^+ + X^- .\]

We define
\[ \mathbb{E}[X] \equiv \mathbb{E}[X^+] - \mathbb{E}[X^-] \tag{5.13} \]
with the understanding that at least one of the terms \( \mathbb{E}[X^+] \) and \( \mathbb{E}[X^-] \) is finite. 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 \( \mathbb{E}[X] = \infty \). 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 \( \mathbb{E}[|X|] = \infty \).

**Definition 5.4.1** *The expectation \( \mathbb{E}[X] \) of the rv \( X \) is said to exist if*
\[ \min(\mathbb{E}[X^+], \mathbb{E}[X^-]) < \infty. \]

*It will be finite if the stronger condition \( \mathbb{E}[X^+] + \mathbb{E}[X^-] < \infty \) holds.*

### 5.5 TBD

In view of the definition we have developed it is natural to write
\[ \mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \]
whenever \( \mathbb{E}[X] \) exists as this notation mimics the expression used for simple rvs. However, a careful look at the construction we have used also shows that this quantity depends only on the probability distribution \( F_X : \mathbb{R} \to [0, 1] \). There are several ways to convince oneself of the validity of this statement: For instance, if \( X \) is a simple rv in its generic form, then
\[ \mathbb{E}[X] = \sum_{i \in I} a_i \mathbb{P}[X = a_i] = \sum_{i \in I} a_i (F_X(a_i) - F_X(a_i^-)) \]
with the usual notation.

Recall that any rv \( X : \Omega \to \mathbb{R}^p \) naturally induces a probability triple on its range, namely \( (\mathbb{R}^p, B(\mathbb{R}^p), \mathbb{P}_X) \) where \( P_X : B(\mathbb{R}^p) \to [0, 1] \) is the probability measure defined by
\[ \mathbb{P}_X[B] = \mathbb{P}[X \in B], \quad B \in B(\mathbb{R}^p). \]
5.6. BASIC PROPERTIES (I)

In fact, the identity mapping \( \text{Id} : \mathbb{R}^p \to \mathbb{R}^p : x \to x \) 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

\[
\mathbb{P}_X [\text{Id} \in B] = \mathbb{P}_X [B] = \mathbb{P} [X \in B], \quad B \in \mathcal{B}(\mathbb{R}^p).
\]

Obviously, say with \( p = 1 \), the expectation of \( X \) computed under \( \mathbb{P} \) has to coincide with that of the rv \( \text{Id} \) 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

\[
\mathbb{E}[X] = \int_{\mathbb{R}} x \mathbb{P}_X(x).
\]

Finally, recalling 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

\[
\mathbb{E}[X] = \int_{\mathbb{R}} x \mathbb{F}_X(x).
\]

5.6 Basic properties (I)

Throughout this section, we are given rvs \( X, Y : \Omega \to \mathbb{R} \).

A. Multiplying by a constant

If \( \mathbb{E}[X] \) exists, then for each \( c \) in \( \mathbb{R} \), \( \mathbb{E}[cX] \) also exists and \( \mathbb{E}[cX] = c \mathbb{E}[X] \).

It is clearly true for simple rvs – See the first part of Lemma 5.4.2 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 5.3.1, so that \( \mathbb{E}[X] = \lim_{n \to \infty} \mathbb{E}[X_n] \).
Then \( \{cX_n, \ n = 1, 2, \ldots\} \) are simple non-negative rvs associated with \( cX \), hence \( \mathbb{E}[cX_n] = c \mathbb{E}[X_n] \) for all \( n = 1, 2, \ldots \).

If \( c \geq 0 \) the sequence is also monotone increasing with \( \lim_{n \to \infty} cX_n = cX \) pointwise and we get

\[
\mathbb{E}[cX] = \lim_{n \to \infty} \mathbb{E}[cX_n] = \lim_{n \to \infty} c \mathbb{E}[X_n] = c \lim_{n \to \infty} \mathbb{E}[X_n] = c \mathbb{E}[X].
\]

(5.14)

If \( c < 0 \), then \((cX)^+ = 0\) and \((cX)^- = -cX = |c|X\). As a result, \( \mathbb{E}[(cX)^+] = 0 \) and \( \mathbb{E}[(cX)^-] = \mathbb{E}[c|X|] = |c|\mathbb{E}[X] \) by the earlier part of the proof and we conclude that \( \mathbb{E}[X] = -\mathbb{E}[(cX)^-] = -|c|\mathbb{E}[X] = c \mathbb{E}[X] \) as desired.
CHAPTER 5. MATHEMATICAL EXPECTATIONS

For the general case, first consider the case $c > 0$. Note that $(cX)^+ = cX^+$ and $(cX)^- = cX^-$, whence $E[cX]$ is well defined as soon as $E[X]$ is well defined with $E[cX] = E[(cX)^+] - E[(cX)^-] = cE[X^+] - cE[X^-] = cE[X]$. The case $c < 0$ is handled mutatis mutandis.

**B. Monotonicity** If $X \leq Y$, then $E[X] \leq E[Y]$ with the understanding that (i) if $-\infty < E[X]$, then $-\infty < E[Y]$ and $E[X] \leq E[Y]$, or (ii) if $E[Y] < \infty$, then $E[X] < \infty$ and $E[X] \leq E[Y]$

**C. Taking absolute values** If $E[X]$ exists, then $|E[X]| \leq E[|X|]$

Note that $-|X| \leq X \leq |X|$ and apply Property B twice.

**D. Localization** If $E[X]$ exists, then $E[X1_A]$ exists for any event $A$ in $F$. If $E[X]$ is finite, then $E[X1_A]$ is finite

For any $A$ in $F$, introduce the rv $X_A = X1_A$. We have $0 \leq X_A^+ = X^+1_A$ by direct inspection so that $X_A^+ \leq X^+$. Obviously $E[X_A^+] \leq E[X^+]$ by Property B, whence $\min(E[X_A^+], E[X_A^-]) \leq \min(E[X^+], E[X^-])$ and $E[X_A^+] + E[X_A^-] \leq E[X^+] + E[X^-]$. The conclusions is now straightforward from Definition 5.4.1 as $\min(E[X^+], E[X^-]) < \infty$ (resp. $E[X^+] + E[X^-] < \infty$) implies $\min(E[X_A^+], E[X_A^-]) < \infty$ (resp. $E[X_A^+] + E[X_A^-] < \infty$).

**E. Adding rvs** We have $E[X + Y] = E[X] + E[Y]$ if (i) the rvs $X$ and $Y$ are non-negative or (ii) if $E[X]$ and $E[Y]$ are both finite

---

5.7 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 sample $\omega$ in $\Omega$. We shall that property $P$ holds almost surely (under $P$) 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 $P$-a.s. when we wish to emphasize the fact that relevant probabilities are evaluated under $P$. For instance, for rvs $X, Y : \Omega \to \mathbb{R}$, we write $X = Y$ a.s. (resp. $X \leq Y$ a.s.) to express the fact that $P [X = Y] = 1$ (resp. $P [X \leq Y] = 1$).

**F. If $X = 0$ a.s., then $E [X] = 0$**

First assume the rv $X$ to be simple with $X = \sum_{i \in I} a_i 1 [A_i]$ in the usual notation. The condition $X = 0$ a.s. implies $P [A_i] = 0$ whenever $a_i \neq 0$, whence $E [X] = \sum_{i \in I} a_i P [A_i] = 0$.

If $X \geq 0$, then the approximating sequence $\{X_n, n = 1, 2, \ldots\}$ of Lemma 5.3 satisfies $0 \leq X_n \leq X$ for all $n = 1, 2, \ldots$, and the constraint $X = 0$ a.s. implies $X_n = 0$ a.s. for all $n = 1, 2, \ldots$, hence $E [X_n] = 0$ by the first part of the proof. Therefore, $E [X] = \lim_{n \to \infty} E [X_n] = 0$ by invoking the definition of $E [X]$.

For arbitrary rv $X$, note that $X \pm = 0$ a.s. if $X = 0$ a.s., whence $E [X \pm] = 0$ and $E [X] = E [X^+] - E [X^-] = 0$. ■

**G. Almost sure equality**  If $X = Y$ a.s. with $E [|X|] < \infty$, then $E [|Y|] < \infty$ and $E [X] = E [Y]$

Write

$$E \equiv \{ \omega \in \Omega : X(\omega) = Y(\omega) \}.$$  

Recall that $X = X 1 [E] + X 1 [E^c]$ and $Y = Y 1 [E] + Y 1 [E^c]$, so that

$$E [X] = E [X 1 [E] + X 1 [E^c]]$$

$$= E [X 1 [E]] + E [X 1 [E^c]] \quad [\text{By Property D and Property E}]$$

$$= E [Y 1 [E]] + E [X 1 [E^c]]$$

$$= E [Y 1 [E]] \quad [\text{By Property F}]$$

(5.15)

CHAPTER 5. MATHEMATICAL EXPECTATIONS

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 $\mathbb{P} [A] = 0$.

For each $n = 1, 2, \ldots$, define the rv $X_n \equiv X1_{E_n}$. 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 inequality $0 \leq \frac{1}{n} 1_{[E_n]} \leq X_n$ then yields

$$0 \leq \frac{1}{n} \mathbb{P} [E_n] \leq \mathbb{E} [X_n] = 0$$

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 Lemma 1.5.1.

I. Assume $\mathbb{E} [\lvert X \rvert] < \infty$ and $\mathbb{E} [\lvert Y \rvert] < \infty$. If $\mathbb{E} [X1_A] \leq \mathbb{E} [Y1_A]$ for all $A$ in $\mathcal{F}$, then $X \leq Y$ a.s.

By Property D, the condition $\mathbb{E} [\lvert X \rvert] < \infty$ (resp. $\mathbb{E} [\lvert Y \rvert] < \infty$) implies the finiteness of $\mathbb{E} [X1_A]$ (resp. $\mathbb{E} [Y1_A]$) for all $A$ in $\mathcal{F}$. Define the event $B$ by

$$B \equiv \{ \omega \in \Omega : Y(\omega) < X(\omega) \}.$$ 

It is plain that $Y1_B \leq X1_B$ whence $\mathbb{E} [Y1_B] \leq \mathbb{E} [X1_B]$ by Property B, while $\mathbb{E} [X1_B] \leq \mathbb{E} [Y1_B]$ by assumption, and the conclusion $\mathbb{E} [X1_B] = \mathbb{E} [Y1_B]$ follows, or equivalently, $\mathbb{E} [(X - Y)1_B] = 0$. But $(X - Y)1_B \geq 0$ and Property H yields $(X - Y)1_B = 0$ a.s.

With $A = [(X - Y)1_B = 0]$, pick $\omega$ in $A$: If $\omega$ also lies in $B$, then $(X(\omega) - Y(\omega))1_B(\omega) = 0$, hence $0 = X(\omega) - Y(\omega) > 0$ by the definition of $B$, and a contradiction occurs. We conclude that $A \cap B = \emptyset$ or equivalently that $A \subseteq B^c$, whence $\mathbb{P} [A] \leq \mathbb{P} [B^c]$ with $\mathbb{P} [A] = 1$. In fine, $\mathbb{P} [B^c] = 1$ and $X \leq Y$ a.s.

J. Extended rvs For any extended rv $X : \Omega \to [\neg \infty, \infty]$, the condition $\mathbb{E} [\lvert X \rvert] < \infty$ implies $\lvert X \rvert < \infty$ a.s.

With $A \equiv \{ \omega \in \Omega : \lvert X(\omega) \rvert = \infty \}$, assume that $\mathbb{P} [A] > 0$. Then, by Property D we have $\mathbb{E} [\lvert X1_A \rvert] \leq \mathbb{E} [\lvert X \rvert]$. But $\mathbb{E} [\lvert X1_A \rvert] \geq \infty \mathbb{P} [A]$ while
5.8 Independence and expectations

The next fact is used in many calculations.

**Proposition 5.8.1** Consider two independent rvs $X, Y: \Omega \to \mathbb{R}$. It holds

\begin{equation}
\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]
\end{equation}

if either (i) the rvs are a.s. non-negative or (ii) both expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist and are finite.

**Proof.** Assume first that both rvs $X$ and $Y$ are simple rvs, say

$X = \sum_{i \in I} a_i 1[A_i] \quad \text{and} \quad Y = \sum_{j \in J} b_j 1[B_j]$

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 $\mathbb{R}$. We shall assume 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

$\mathbb{P}[A_i \cap B_j] = \mathbb{P}[A_i] \mathbb{P}[B_j], \quad i \in I, j \in J$

since the events $[X = a_i]$ and $[Y = b_j]$ are independent.

Noting that

$XY = \sum_{i \in I} \sum_{j \in J} a_i b_j 1[A_i] 1[B_j] = \sum_{i \in I} \sum_{j \in J} a_i b_j 1[A_i \cap B_j],$

we conclude that

$\mathbb{E}[XY] = \mathbb{E}\left[ \sum_{i \in I} \sum_{j \in J} a_i b_j 1[A_i \cap B_j] \right]$

$= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{E}[1[A_i \cap B_j]]$
CHAPTER 5. MATHEMATICAL EXPECTATIONS

\[= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{P}[A_i \cap B_j]\]
\[= \sum_{i \in I} \sum_{j \in J} a_i b_j \mathbb{P}[A_i] \mathbb{P}[B_j]\]
\[= \left( \sum_{i \in I} a_i \mathbb{P}[A_i] \right) \left( \sum_{j \in J} b_j \mathbb{P}[B_j] \right)\]
\[= \mathbb{E}[X] \mathbb{E}[Y].\]  
(5.17)

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 5.3.1. Note that the rvs \(\{X_nY_n, n = 1, 2, \ldots\}\) form a monotone sequence of staircase approximations for the rv \(XY\) since

\[0 \leq X_nY_n \leq X_{n+1}Y_n \leq X_{n+1}Y_{n+1}, \quad n = 1, 2, \ldots\]

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 \(Y_n = h_n(Y)\) for some Borel mappings \(g_n, h_n : \mathbb{R} \rightarrow \mathbb{R}^+\). See the construction in the proof of Lemma 5.3.1. Obviously, \(\lim_{n \rightarrow \infty} X_nY_n = (\lim_{n \rightarrow \infty} X_n) (\lim_{n \rightarrow \infty} Y_n) = XY\), whence

\[\mathbb{E}[XY] = \lim_{n \rightarrow \infty} \mathbb{E}[X_nY_n] \quad [\text{By the definition of } \mathbb{E}[XY]]\]
\[= \lim_{n \rightarrow \infty} \mathbb{E}[X_n] \mathbb{E}[Y_n] \quad [\text{By independence}]\]
\[= \left( \lim_{n \rightarrow \infty} \mathbb{E}[X_n] \right) \left( \lim_{n \rightarrow \infty} \mathbb{E}[Y_n] \right) \quad [\text{By the definition of } \mathbb{E}[X] \text{ and } \mathbb{E}[Y]]\]
\[= \mathbb{E}[X] \mathbb{E}[Y].\]  
(5.18)

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, we use the decompositions \(X = X^+ - X^-\) and \(Y = Y^+ - Y^-\) so that

\[XY = (X^+ - X^-) (Y^+ - Y^-)\]
\[= X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-.\]  
(5.19)

Note that the \(\mathbb{R}_+^2\)-valued rvs \((X^+, X^-)\) and \((Y^+, Y^-)\) are independent, a fact inherited from the independence of the rvs. 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) we have that the expectations $E[X+Y]$, $E[X+Y^-]$, $E[X^-Y^+]$ and $E[X^-Y^-]$ are all finite and given by $E[X+]E[Y^+]$, $E[X+]E[Y^-]$, $E[X^-]E[Y^+]$ and $E[X^-]E[Y^-]$, respectively. Thus, by Properties A and E we get

$$
E[XY] = E[X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-] \\
= E[X^+Y^+] - E[X^+Y^-] - E[X^-Y^+] + E[X^-Y^-] \\
- E[X^-]E[Y^+] + E[X^-]E[Y^-] \\
= (E[X^+] - E[X^-]) (E[Y^+] - E[Y^-]) \\
= E[X]E[Y]
$$

(5.20)

as announced.

Proposition 5.8.1 has the following often used consequence.

**Lemma 5.8.1** The rvs $X_1 : \Omega \to \mathbb{R}^{p_1}$, ..., $X_k : \Omega \to \mathbb{R}^{p_k}$ are assumed to be mutually independent. 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

$$E[\prod_{\ell=1}^k Y_\ell] = \prod_{\ell=1}^k E[Y_\ell]$$

whenever $E[|Y_\ell|] < \infty$ for all $\ell = 1, \ldots, k$.

### 5.9 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 $P$ 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 5.9.1** Take $\Omega = [0, 1]$, $\mathcal{F} = B([0, 1])$ and with $a < b$ in $[0, 1]$, define the probability measure $P$ on $\mathcal{F}$ by setting

$$P[E] = \frac{|E \cap \{a, b\}|}{2}, \quad E \in \mathcal{F}.$$
CHAPTER 5. MATHEMATICAL EXPECTATIONS

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 elementary treatment of Probability Theory.

**Proposition 5.9.1** If $X : \Omega \to \mathbb{R}$ is a discrete rv with (countable) support $S$, then

$$E[X] = \sum_{x \in S} x \cdot \mathbb{P}[X = x]$$

(5.21)

if either the rv $X$ is a.s. non-negative (with $S \subseteq \mathbb{R}_+$) or if the absolute summability condition

$$\sum_{x \in S} |x| \cdot \mathbb{P}[X = x] < \infty$$

(5.22)

holds.

**Proof.** Assume first that the set $S$ contains only finitely many elements. The rv $X^* : \Omega \to \mathbb{R}$ defined by

$$X^* \equiv \sum_{x \in S} x \cdot 1[X = x] + 0 \cdot 1[X \notin S]$$

is a simple rv with expectation given by

$$E[X^*] = \sum_{x \in S} x \cdot \mathbb{P}[X = x].$$

Note that $X = X^*$ on $[X \in S]$, hence $X = X^*$ a.s.. Using Property G, we conclude that $E[|X|] < \infty$ since $E[|X^*|] < \infty$ and we conclude $E[X] = E[X^*]$, hence

$$E[X] = \sum_{x \in S} x \cdot \mathbb{P}[X = x].$$

Assume now that $S$ has countably many elements. If $S \subseteq \mathbb{R}_+$, then $X = X1[X \in S]$ a.s. with $X1[X \in S] \geq 0$. Introduce the simple non-negative rvs

$$\{X_n, n = 1, 2, \ldots\}$$

given by

$$X_n \equiv \sum_{\ell=1}^n x_\ell \cdot 1[X = x_\ell], \quad n = 1, 2, \ldots$$
where \( \{x_\ell, \ell = 1, 2, \ldots\} \) is a labelling of \( S \). Note that the sequence \( \{X_n, n = 1, 2, \ldots\} \) is monotone with \( \lim_{n \to \infty} X_n = X1 [X \in S] \). Thus,

\[
\mathbb{E} [X1 [X \in S]] = \lim_{n \to \infty} \mathbb{E} [X_n] \quad \text{[By the definition of } \mathbb{E} [X1 [X \in S]]] \\
= \lim_{n \to \infty} \left( \sum_{\ell=1}^{n} x_\ell \cdot \mathbb{P} [X = x_\ell] \right) \quad \text{[By the first part of the proof]} \\
= \sum_{\ell=1}^{\infty} x_\ell \cdot \mathbb{P} [X = x_\ell] \quad \text{[By monotonicity since } S \subseteq \mathbb{R}_+ \text{]} \\
= \sum_{x \in S} x \cdot \mathbb{P} [X = x].
\]

(5.23)

Using Property G, if \( \mathbb{E} [X1 [X \in S]] < \infty \), then \( \mathbb{E} [|X|] < \infty \) and we conclude

\[
\mathbb{E} [X] = \mathbb{E} [X1 [X \in S]] = \sum_{x \in S} x \cdot \mathbb{P} [X = x].
\]

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. By the previous discussion, we have

\[
\mathbb{E} [X^\pm] = \sum_{x \in S \pm} (\pm x) \cdot \mathbb{P} [X = x],
\]

whence

\[
\mathbb{E} [X] = \mathbb{E} [X^+] - \mathbb{E} [X^-] = \sum_{x \in S : x \geq 0} x \cdot \mathbb{P} [X = x] - \sum_{x \in S : x \leq 0} (-x) \cdot \mathbb{P} [X = x].
\]

(5.24)

5.10 Lebesgue-Stieltjes integration

5.11 Convergence results for expectations

In this section we are considering a sequence \( \{X, Y, Z, X_n, n = 1, 2, \ldots\} \) of \( \mathbb{R} \)-valued rvs which are all defined on some probability triple \( (\Omega, \mathcal{F}, \mathbb{P}) \). We shall be
interested in conditions that allow the interchange of expectation and limit operations.

We begin with the Monotone Convergence Theorem which deals with monotone sequences.

**Theorem 5.11.1** With \( Y \leq X_n \leq X_{n+1} \) for all \( n = 1, 2, \ldots \) where \( E[Y] > -\infty \), we have

\[
\lim_{n \to \infty} E[X_n] = E\left[ \lim_{n \to \infty} X_n \right]
\]

monotonically. With \( X \geq X_n \geq X_{n+1} \) for all \( n = 1, 2, \ldots \) where \( E[X] < \infty \), we have

\[
\lim_{n \to \infty} E[X_n] = E\left[ \lim_{n \to \infty} X_n \right]
\]

monotonically.

Under the assumptions of Theorem 5.11.1 the limit \( \lim_{n \to \infty} X_n \) exists pointwise, while the limit \( \lim_{n \to \infty} E[X_n] \) also exists by monotonicity.

An important consequence of the Monotone Convergence Theorem is as follows: Let \( \{X_n, n = 1, 2, \ldots\} \) denote a sequence of \( \mathbb{R}_+ \)-valued rvs. It follows from the Monotone Convergence Theorem that

\[
E\left[ \sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} E[X_n]
\]

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

\[
\lim_{n \to \infty} E[S_n] = E\left[ \lim_{n \to \infty} S_n \right]
\]

by (5.25). By linearity, we have

\[
E[S_n] = \sum_{k=1}^{n} E[X_k]
\]

so that \( \lim_{n \to \infty} E[S_n] = \sum_{n=1}^{\infty} E[X_n] \), while \( \lim_{n \to \infty} S_n = \sum_{n=1}^{\infty} X_n \).

Fatou’s Lemma is given next and deals with situations when the limit may not exist or is not known (yet) to exist.
5.11. CONVERGENCE RESULTS FOR EXPECTATIONS

Theorem 5.11.2 With $X_n \geq Y$ for all $n = 1, 2, \ldots$ where $\mathbb{E}[Y] > -\infty$, we have

$$\mathbb{E}\left[\lim\inf_{n \to \infty} X_n\right] \leq \lim\inf_{n \to \infty} \mathbb{E}[X_n].$$

With $X_n \leq Y$ for all $n = 1, 2, \ldots$ where $\mathbb{E}[Y] < \infty$, we have

$$\lim\sup_{n \to \infty} \mathbb{E}[X_n] \leq \mathbb{E}\left[\lim\sup_{n \to \infty} X_n\right].$$

The following example shows that the bounding conditions cannot be eliminated.

Example 5.11.1 Take $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(I)$ with $\mathbb{P}$ being Lebesgue measure $\lambda$.

The rvs $\{X_n, n = 1, 2, \ldots\}$ are given by

$$X_n(\omega) = \begin{cases} 0 & \text{if } \omega \notin \left[\frac{1}{n}, \frac{2}{n}\right], \\ -n & \text{if } \omega \in \left[\frac{1}{n}, \frac{2}{n}\right], \end{cases} \quad n = 2, 3, \ldots$$

Obviously, $\mathbb{E}[X_n] = n^{-1}(-n) = -1$ for all $n = 1, 2, \ldots$, so that $\lim\inf_{n \to \infty} \mathbb{E}[X_n] = -1$, while $\lim\inf_{n \to \infty} X_n = 0$ so that $\mathbb{E}\left[\lim\inf_{n \to \infty} X_n\right] = 0$.

The Bounded Convergence Theorem shows that the interchange always holds when the rvs $\{X_n, n = 1, 2, \ldots\}$ form a bounded sequence.

Theorem 5.11.3 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

$$\mathbb{E}\left[\lim_{n \to \infty} X_n\right] = \lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

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 5.11.4 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 $|X| < Y$ for each $n = 1, 2, \ldots$, then $\mathbb{E}[X]$ exists and is finite with

$$\mathbb{E}\left[\lim_{n \to \infty} X_n\right] = \lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$
5.12 Change of variable formula

**Proposition 5.12.1** 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

$$E[g(X)] = \int_{\mathbb{R}^p} g(x) dF_X(x) \quad (5.31)$$

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

$$g(x) = 1 \begin{bmatrix} x \in B \end{bmatrix}, \quad x \in \mathbb{R}^p$$

for some Borel set $B$ in $\mathcal{B}(\mathbb{R}^p)$, then

$$E[g(X)] = P[X \in B] = P_X[B] = E_X[g(\cdot)] = \int_{\mathbb{R}^p} g(x) dF_X(x)$$

Assume now that $g : \mathbb{R}^p \to \mathbb{R}$ is simple in the sense that

$$g(x) = \sum_{i \in I} g_i 1 \begin{bmatrix} x \in B_i \end{bmatrix}, \quad x \in \mathbb{R}^p$$

Then,

$$E[g(X)] = E \left[ \sum_{i \in I} g_i 1 \begin{bmatrix} X \in B_i \end{bmatrix} \right] = \sum_{i \in I} g_i E \left[ 1 \begin{bmatrix} X \in B_i \end{bmatrix} \right] = \sum_{i \in I} g_i P_X \left[ X \in B_i \right] = \sum_{i \in I} g_i \int_{\mathbb{R}^p} 1 \begin{bmatrix} B_i \end{bmatrix}(x) dF_X(x) = \int_{\mathbb{R}^p} g(x) dF_X(x) \quad (5.32)$$

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

$$g_n(x) = \sum_{m=0}^{n-1} \sum_{k=0}^{2^n-1} \frac{k}{2^n} 1 \begin{bmatrix} \frac{k}{2^n} < x \leq \frac{k+1}{2^n} \end{bmatrix}, \quad x \in \mathbb{R}^p$$
We already have
\[ E[g_n(X)] = \int_{\mathbb{R}} g_n(x) dF_X(x), \quad n = 1, 2, \ldots \]
and the conclusion
\[ E[g(X)] = \int_{\mathbb{R}} g(x) dF_X(x), \]
follows by the Monotone Convergence Theorem (under \( P \) and \( P_X \)).

In the general case \( g : \mathbb{R}^P \rightarrow \mathbb{R} \), write
\[ g(x) = g(x)^+ - g(x)^-, \quad x \in \mathbb{R}^P \]
and by linearity, we get
\[ E[g(X)] = E[g(X)^+] - E[g(X)^-] \]

5.13 Exercises

Ex. 5.1 Let \( X : \Omega \rightarrow \mathbb{R} \) be a rv with finite expectation, i.e., \( E[|X|] < \infty \).
   
a. If \( X \geq 0 \), show that \( \lim_{n \to \infty} nP[X \geq n] = 0 \) (so that \( \lim_{n \to \infty} nP[X > n] = 0 \) as well). [HINT: If \( X \geq 0 \), recall that the value \( E[X] \) 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. 5.2 Compute the expectation
\[ E\left[\frac{1}{1+Y^+}\right] \]
when the rv \( Y : \Omega \rightarrow \mathbb{R} \) is
   
a. a binomial rv \( \text{Bin}(n; p) \) with \( n = 1, 2, \ldots \) and \( 0 < p < 1 \),
   
b. a Poisson rv \( \text{Poi}(\lambda) \) with \( \lambda > 0 \),
   
c. a geometric rv \( \text{Geo}(p) \) with \( 0 < p < 1 \).

In each case explain why the expectation \( E\left[\frac{1}{1+Y^+}\right] \) always exists.
Ex. 5.3 Assume the rv $X : \Omega \to \mathbb{R}^p$ defined on the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ to be a symmetric rv.
   a. Can $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ assume different values?
   b. Show that $\mathbb{E}[X]$ is well defined, it finite with $\mathbb{E}[X] = 0$.
   c. Give an example of a symmetric rv $X$ for which $\mathbb{E}[X]$ is not well defined

Ex. 5.4 We start with a rv $\xi : \Omega \to \mathbb{R}$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.
Next we introduce the rvs $X, Y, Z : \Omega \to \mathbb{R}$ given by $X \equiv \sin(\xi), Y \equiv \xi^1 + \xi^2$ 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 $\xi$. In each case justify your answer!
   In what follows, assume $\xi$ to be a symmetric rv under $\mathbb{P}$ in the sense that the rvs $\xi$ and $-\xi$ have the same probability distribution under $\mathbb{P}$.
   b. Evaluate $\mathbb{E}[X]$.
   c. Evaluate $\mathbb{E}[Y]$.
   d. Give an example that shows that $\mathbb{E}[Z]$ may not always exist. Give an additional condition on $\xi$ to ensure that the expectation $\mathbb{E}[Z]$ can be evaluated and find its value.

Ex. 5.5 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. 5.6 It is November 1, 1920 in Paris and on that autumn evening a party of $n$ friends enter the restaurant “Le Languedoc” to celebrate the birthday of Antoine Lepoutre, a fellow none of them knew. Upon arrival they all give their hats to the coat check. Unfortunately, the attendant is quite absent minded, and the hats end up being thoroughly mixed up by the time the guests depart at the end of the evening. As a result, the hats are randomly handed out to these $n$ customers – What else is there to do?
   a. Construct a probability model $(\Omega, \mathcal{F}, \mathbb{P})$ for this unfortunate situation in order to answer the following questions in Parts b-c:
   For each $k = 1, \ldots, n$, let $X_k$ denote a binary rv defined to $X_k = 1$ (resp. $X_k = 0$) if the $k^{th}$ person gets (resp. does not get) her hat back. If $S_n$ denotes the number of diners that get their hats back, then $S_n = X_1 + \ldots + X_n$. 


b. Compute $P [X_k = 1]$ for each $k = 1, \ldots, n$ and evaluate $E [S_n]$

c. With distinct $k, \ell = 1, \ldots, n$, determine whether the rvs $X_k$ and $X_\ell$ are (i) mutually independent or (ii) uncorrelated

d. Compute the variance $\text{Var} [S_n]$

e. Show that $P [S_n \geq 11] \leq 0.01$ for $n \geq 11$.

**Ex. 5.7** Let $N$ be a discrete rv with support contained in $\mathbb{N}_0$ (i.e., $P [N \in \mathbb{N}_0] = 1$) with a finite second moment, i.e., $E [N^2] < \infty$. 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

$$E \left[ \frac{1}{N} \sum_{n=1}^{N} X_n \right].$$

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 $\mu \equiv E [X_1] = E [X_2] = \ldots$ and $\sigma^2 \equiv \text{Var}[X_1] = \text{Var}[X_2] = \ldots$

**Ex. 5.8** 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 $P [0 < P \leq 1] = 1$. 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

$$X \equiv \sum_{i=1}^{n} 1 [U_i \leq P].$$

Further assume that the rv $P$ is a discrete rv with $P [P \in S] = 1$ for some countable subset $S \subseteq (0, 1]$.


b. How many moments of $P$ are needed to compute $\text{Var} [X]$?

c. When $S$ contains at least two elements, are the rvs $1 [U_1 \leq P], \ldots, 1 [U_n \leq P] (i)$ mutually independent (ii) pairwise uncorrelated?

d. Compute the probabilities

$$P [X = k], \quad k = 0, 1, \ldots, n.$$ 

How many moments of $P$ are needed?
Chapter 6

Moments, bounds and inequalities

All rvs are defined on the same probability triple \((\Omega, \mathcal{F}, \mathbb{P})\).

6.1 Moments, variance and covariance

Consider the rv \(X: \Omega \to \mathbb{R}\). With \(r = 1, 2, \ldots\), we define the \(r^{th}\) moment \(m_r\) of \(X\) by
\[
m_r \equiv \mathbb{E}[X^r]
\]
provided the expectation exists. For any \(r \geq 0\) the absolute \(r^{th}\) moment of \(X\) is given by
\[
\mu_r \equiv \mathbb{E}[|X|^r].
\]
It is always well defined, and may possibly be infinite. When \(r = 1\) we refer to \(m_1\) as the first moment of \(X\). When \(r = 2\), \(m_2\) 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 \(\mathbb{E}[|X|^2] < \infty\).

When the first moment of \(X\) exists and is finite, then the definitions (6.1) and (6.2) lead naturally to the centered expectations given by
\[
m^*_r \equiv \mathbb{E}[(X - \mathbb{E}[X])^r]
\]
and
\[
\mu^*_r \equiv \mathbb{E}[|X - \mathbb{E}[X]|^r].
\]
provided the expectation exists.

For all \(u \geq 0\) the inequality \(u^r \leq 1 + u^s\) holds whenever \(1 \leq r < s\) (not necessarily integers). This leads through Property B to the following useful fact.
**Fact 6.1.1** With $1 \leq r < s$ (not necessarily integers), we have
\[
\mathbb{E}[|X|^r] \leq 1 + \mathbb{E}[|X|^s].
\]
In particular, the finiteness of $\mathbb{E}[|X|^s]$ implies that of $\mathbb{E}[|X|^r]$.

Thus, if $X$ is a second-order rv, then $\mathbb{E}[|X|] < \infty$ by virtue of Fact 6.1.1, and both $\mathbb{E}[X]$ and $\mathbb{E}[|X|]$ exist and are finite. The variance $\text{Var}[X]$ of the rv $X$ is then defined as
\[
\text{Var}[X] \equiv \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\]
Since $(X - \mathbb{E}[X])^2 \geq 0$ we conclude that $\text{Var}[X] \geq 0$, whence $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$.

Let $X,Y : \Omega \to \mathbb{R}$ be a pair of rvs with finite second moments, namely $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$. The covariance $\text{Cov}[X,Y]$ between the rvs $X$ and $Y$ is defined by
\[
\text{Cov}[X,Y] \equiv \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\]
This quantity is well defined when the rvs $X$ and $Y$ are second-order rvs This is a consequence of the identity $(a-b)^2 = a^2 + b^2 - 2ab \geq 0$ for arbitrary scalars $a$ and $b$ in $\mathbb{R}$, so that $|a||b| \leq \frac{1}{2}(a^2 + b^2)$. This is also a consequence of the Cauchy-Schwarz inequality (6.14).

**Definition 6.1.1** The rvs $X$ and $Y$ are said to be uncorrelated if $\text{Cov}[X,Y] = 0$.

**Fact 6.1.2** If two second-order rvs $X$ and $Y$ are independent, they are necessarily uncorrelated.

However, the converse is not true even when the rvs $X$ and $Y$ are second-order rvs as the following counterexample shows

**Example 6.1.1**

We conclude with the following useful result.

**Lemma 6.1.1** If the rvs $X_1, \ldots, X_n$ are second-order rvs, then
\[
\text{Var}[X_1 + \ldots + X_n] = \sum_{k=1}^{n} \text{Var}[X_k] + \sum_{k=1}^{n} \sum_{\ell=1, \ell \neq k}^{n} \text{Cov}[X_k, X_\ell].
\]
6.2. MARKOV’S INEQUALITY AND CONSEQUENCES

If the rvs $X_1, \ldots, X_n$ are pairwise uncorrelated, i.e.,

$$\text{Cov}[X_k, X_\ell] = 0, \quad k \neq \ell, \quad k, \ell = 1, \ldots, n$$

then the variance of a sum is indeed the sum of the individual variances as (6.5) becomes

(6.6) $$\text{Var}[X_1 + \ldots + X_n] = \sum_{k=1}^n \text{Var}[X_k].$$

**Proof.** We start by noting that

$$X_1 + \ldots + X_n - \mathbb{E}[X_1 + \ldots + X_n] = \sum_{k=1}^n (X_k - \mathbb{E}[X_k])$$

and

$$(X_1 + \ldots + X_n - \mathbb{E}[X_1 + \ldots + X_n])^2 = \sum_{k=1}^n \sum_{\ell=1}^n (X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell]).$$

Taking expectations on both sides of this last relation, we conclude that

$$\text{Var}[X_1 + \ldots + X_n]
= \mathbb{E}\left[(X_1 + \ldots + X_n - \mathbb{E}[X_1 + \ldots + X_n])^2\right]
= \sum_{k=1}^n \sum_{\ell=1}^n \mathbb{E}[(X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell])]
= \sum_{k=1}^n \mathbb{E}\left[(X_k - \mathbb{E}[X_k])^2\right]
+ \sum_{k=1}^n \sum_{\ell=1, \ell \neq k}^n \mathbb{E}[(X_k - \mathbb{E}[X_k]) \cdot (X_\ell - \mathbb{E}[X_\ell])].$$

(6.7)

and the desired conclusion (6.5) follows.

6.2 Markov’s inequality and consequences

We present here and in the next section a number of inequalities that prove useful in many contexts.
Theorem 6.2.1 (Markov’s inequality) For any rv $X : \Omega \to \mathbb{R}$ with $\mathbb{E}[|X|] < \infty$, we have
\begin{equation}
\mathbb{P}[|X| \geq t] \leq \frac{1}{t} \cdot \mathbb{E}[|X|], \quad t > 0
\end{equation}

Proof. Fix $t > 0$, and note that
\[
|X| = |X| \cdot 1[|X| < t] + |X| \cdot 1[|X| \geq t] \\
\geq |X| \cdot 1[|X| \geq t] \\
\geq t \cdot 1[|X| \geq t].
\]
Taking expectations in this inequality, we find
\[
t \cdot \mathbb{E}[1[|X| \geq t]] \leq \mathbb{E}[|X|]
\]
and the conclusion follows.

This basic inequality gives rise to several useful inequalities.

Bienayme-Tchebychev inequality For instance, consider a second-order rv $X$. Applying Markov’s inequality to the rv $(X - \mathbb{E}[X])^2$ we get
\begin{equation}
\mathbb{P}[(X - \mathbb{E}[X])^2 \geq t^2] \leq \frac{1}{t^2} \cdot \mathbb{E}[(X - \mathbb{E}[X])^2], \quad t > 0.
\end{equation}
This is often written in the equivalent form
\begin{equation}
\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{1}{t^2} \cdot \text{Var}[X], \quad t > 0
\end{equation}
and is known as the Bienayme-Tchebychev inequality.

Chernoff bounds Fix $t$ in $\mathbb{R}$. With $\theta > 0$, we note that
\[
\mathbb{P}[X > t] = \mathbb{P}[e^{\theta X} > e^{\theta t}].
\]
Applying Markov’s inequality to the rv $e^{\theta X}$ we get
\[
\mathbb{P}[e^{\theta X} > e^{\theta t}] \leq e^{-\theta t} \mathbb{E}
\end{equation}
6.3. **INEQUALITIES**

Such a bound is known as a Chernoff bound. Of course only when $\theta > 0$ is such that $\mathbb{E} [e^{\theta X}] < \infty$ will the bound just derived be useful.

Up to this point $\theta > 0$ is a free parameter, so that we can optimize the previous bound by properly selecting $\theta > 0$. This yields

$$\mathbb{P} [X > t] \leq \inf \left( e^{-\theta t} \mathbb{E} [e^{\theta X}] : \theta > 0, \mathbb{E} [e^{\theta X}] < \infty \right).$$

**A general Markov inequality** Consider a strictly monotone increasing function $g : \mathbb{R} \to \mathbb{R}_+$ such that $\mathbb{E} [g(X)]$ is finite. Fix $t$ in $\mathbb{R}$. Note that $X > t$ if and only if $g(X) > g(t)$. It follows then that

$$\mathbb{P} [X > t] = \mathbb{P} [g(X) > g(t)] \leq \frac{\mathbb{E} [g(X)]}{g(t)}$$

by the basic Markov inequality.

### 6.3 Inequalities

In this section we present a number of useful bounds on the moments of rvs. With $I$ an interval in $\mathbb{R}$, a mapping $g : I \to \mathbb{R}$ is convex if the conditions

$$g((1 - \lambda)x_0 + \lambda x_1) \leq (1 - \lambda)g(x_0) + \lambda g(x_1), \quad x_0, x_1 \in I, \quad \lambda \in [0, 1]$$

hold.

**Theorem 6.3.1 (Jensen’s inequality)** For any rv $X : \Omega \to \mathbb{R}$, we have

$$g(\mathbb{E} [X]) \leq \mathbb{E} [g(X)]$$

for any convex mapping $g : \mathbb{R} \to \mathbb{R}$.

**Theorem 6.3.2 (Cauchy-Schwarz inequality)** For any pair of second-order rvs $X, Y : \Omega \to \mathbb{R}$ we have

$$\mathbb{E} [|X||Y|] \leq \sqrt{\mathbb{E} [|X|^2]} \cdot \sqrt{\mathbb{E} [|Y|^2]}$$

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

$$aX + bY = 0 \quad a.s.$$
Proof. Define the quadratic form $Q : \mathbb{R} \to \mathbb{R}_+$ given by
\[
Q(\lambda) = \mathbb{E}[(X + \lambda Y)^2], \quad \lambda \in \mathbb{R}.
\]
Obviously,
\[
Q(\lambda) = \mathbb{E}[X^2] + 2\lambda \mathbb{E}[XY] + \lambda^2 \mathbb{E}[Y^2], \quad \lambda \in \mathbb{R}.
\]
(6.15)

The roots of this quadratic form are determined by the sign of the discriminant
\[
\Delta = (2 \mathbb{E}[XY])^2 - 4 \mathbb{E}[Y^2] = 4 \left( \mathbb{E}[XY]^2 - \mathbb{E}[Y^2] \right)
\]

However, by its very definition, $Q(\lambda) \geq 0$ for all $\lambda$ in $\mathbb{R}$. Therefore, the equation $Q(\lambda) = 0$ on $\mathbb{R}$ cannot have two distinct roots, say $\lambda_1 < \lambda_2$, as this would imply $Q(\lambda) < 0$ in the interval $(\lambda_1, \lambda_2)$. In other words, it is not possible for $\Delta > 0$ to occur. Note that $\Delta \geq 0$ is equivalent to the Cauchy-Schwarz inequality.

If (6.14) holds as an equality, then we necessarily have $\Delta = 0$, and there exists a unique $\lambda^*$ in $\mathbb{R}$ such that $Q(\lambda^*) = 0$, or equivalently, $\mathbb{E}[(X + \lambda^* Y)^2] = 0$. This implies $(X + \lambda^* Y)^2 = 0$ a.s., hence $aX + bY = 0$ a.s. with $a = 1$ and $b = \lambda^*$.

Conversely, assume there exist constants $a$ and $b$ in $\mathbb{R}$ not simultaneously zero such that $aX + bY = 0$ a.s. For instance assuming $a \neq 0$, then $X + a^{-1} b Y = 0$ 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 Cauchy-Schwarz inequality admits the following generalization known as H"older’s inequality

Theorem 6.3.3 (H"older’s inequality) Consider rvs $X, Y : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|^r] < \infty$ and $\mathbb{E}[|Y|^s] < \infty$ for $r, s > 1$. Whenever
\[
\frac{1}{r} + \frac{1}{s} = 1,
\]
(6.16)

it holds that
\[
\mathbb{E}[|X||Y|] \leq \mathbb{E}[|X|^r]^{\frac{1}{r}} \mathbb{E}[|Y|^s]^{\frac{1}{s}}.
\]
(6.17)
Pairs of integers $r$ and $s$ such that (6.16) holds are said to form a *conjugate* pair. Hölder’s inequality reduces to the Cauchy-Schwarz inequality when $r = s = 2$.

**Theorem 6.3.4** (*Minkowski’s inequality*) For rvs $X, Y : \Omega \to \mathbb{R}$ such that $\mathbb{E} [|X|^r] < \infty$ and $\mathbb{E} [|Y|^r] < \infty$ for some $r \geq 1$, we have the inequality

$$
\mathbb{E} [|X + Y|^r] \leq \mathbb{E} [|X|^r]^{\frac{1}{r}} + \mathbb{E} [|Y|^r]^{\frac{1}{r}}.
$$

(6.18)

**Proof.**

The following consequence of Hölder’s inequality is known as Lyapounov’s inequality.

**Theorem 6.3.5** (*Lyapounov’s inequality*) For any rv $X : \Omega \to \mathbb{R}$, the mapping $g : \mathbb{R}_+ \to [0, \infty]$ defined by

$$
g(r) = \log (\mathbb{E} [|X|^r]), \quad r \geq 0.
$$

(6.19)

is convex.

### 6.4 Exercises

**Ex. 6.1** If the rv $X : \Omega \to \mathbb{R}$ is bounded in the sense that $\mathbb{P} [|X| \leq M] = 1$ for some $M > 0$, show that the expectation of the rv $X$ always exists and is finite with $|\mathbb{E} [X]| \leq M$.

**Ex. 6.2** 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?

c. If $\mathbb{E} [|X|^r] < \infty$ for some $r > 0$, show that $\lim_{n \to \infty} n^{-r} \mathbb{P} [|X| > n] = 0$. 


Ex. 6.3 Let $X : \Omega \to \mathbb{R}$ be a discrete rv such that $\mathbb{P}[X \in \mathbb{N}] = 1$. Using the fact that $X = \sum_{n=0}^\infty 1 [X > n]$ 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. 6.4 For any two second-order rvs $X, Y : \Omega \to \mathbb{R}$, the rvs $X-Y$ and $X+Y$ are uncorrelated.

Ex. 6.5 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 $\text{Cov} [\vert X(p) - Y(p)\vert, X(p) + Y(p)]$ of the pair of rvs $\vert X(p) - Y(p)\vert$ and $X(p) + Y(p)$ as a function of $p$.

b. Find all the values of $p$ in $(0, 1)$ such that rvs $\vert X(p) - Y(p)\vert$ and $X(p) + Y(p)$ are uncorrelated rvs.

c. Find all the values of $p$ in $(0, 1)$ such that rvs $\vert X(p) - Y(p)\vert$ and $X(p) + Y(p)$ are independent rvs.

Ex. 6.6 Consider four second-order rvs $X_1, X_2, Y_1, Y_2 : \Omega \to \mathbb{R}$. Assume the rvs $X : \Omega \to \mathbb{R}^2$ and $Y : \Omega \to \mathbb{R}^2$ to be independent where $X = (X_1, X_2)$ and $(Y_1, Y_2)$, show that
\[
\text{Cov} [X_1 + Y_1, X_2 + Y_2] = \text{Cov} [X_1, X_2] + \text{Cov} [Y_1, Y_2].
\]

Ex. 6.7 Let $X, Y : \Omega \to \mathbb{R}$ denote two second-order rvs.

a. If the rvs $X$ and $Y$ are independent, they are uncorrelated.

b. The converse is not true in general: Give a counterexample showing that the rvs $X$ and $Y$ are uncorrelated but not independent.

Ex. 6.8 Let $A$ and $B$ be events in $\mathcal{F}$, and let $X_A = 1[A]$ and $X_B = 1[B]$ denote their indicator functions.

a. Compute the covariance $\text{Cov} [X_A, X_B]$

b. Show that the events $A$ and $B$ are independent if and only if the rvs $X_A$ and $X_B$ are uncorrelated.

c. Show that the rvs $X_A$ and $X_B$ are uncorrelated if and only if the rvs $X_A$ and $X_B$ are independent.
6.4. EXERCISES

Ex. 6.9 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$ defined on some probability triple $(\Omega, \mathcal{F}, P)$. These two rvs are assumed to be independent rvs with $E[X] = E[Y] = 0$. Define the rvs $U, V : \Omega \to \mathbb{R}$ by $U \equiv \min(X, Y)$ and $V \equiv \max(X, Y)$.

a. Compute $E[U \cdot V]$.

b. Are the rvs $U$ and $V$ second-order rvs? Explain your answer.

c. Are the rvs $U$ and $V$ uncorrelated? Are they independent?

Ex. 6.10 Consider two second-order rvs $X, Y : \Omega \to \mathbb{R}$ defined on some probability triple $(\Omega, \mathcal{F}, P)$. We assume that (i) these rvs $X$ and $Y$ are independent rvs and that (ii) each of the rvs $X$ and $Y$ has a symmetric distribution, i.e., $X =_{st} -X$ and $Y =_{st} -Y$ (where as usual $=_{st}$ refers to equality in distribution). Define the rvs $U, V : \Omega \to \mathbb{R}$ by $U \equiv \min(X, Y)$ and $V \equiv \max(X, Y)$.

a. Show that $V =_{st} -U$.


c. Under what additional conditions are the rvs $U$ and $V$ uncorrelated?

Ex. 6.11 The following (important) inequality happens to be true: With $f, g : \mathbb{R} \to \mathbb{R}$ monotone non-decreasing mappings,\footnote{This means $f(x) \leq f(y)$ and $g(x) \leq g(y)$ whenever $x < y$ in $\mathbb{R}$} it holds that

$$E[f(X)] \cdot E[g(X)] \leq E[f(X) \cdot g(X)]$$

(6.20)

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 $p_X = (P[X = x], x \in S)$. 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 $E[\Delta(Y, Z)]$?

d. Use Part c to conclude that (6.20) holds!
Chapter 7

Conditioning and conditional expectations

We now turn to the important notion of conditioning in its various forms. Throughout we assume given a probability triple \((\Omega, \mathcal{F}, \mathbb{P})\). The following terminology will be useful in a number of places.

**Definition 7.0.1** A collection \(\mathcal{D}\) of subsets of \(\Omega\) is called a sub-\(\sigma\)-field of \(\mathcal{F}\) if \(\mathcal{D}\) is a \(\sigma\)-field on \(\Omega\) such that \(\mathcal{D} \subseteq \mathcal{F}\).

### 7.1 Sub-\(\sigma\)-fields generated by rvs

We begin with a simple definition.

**Definition 7.1.1** Let \(\mathcal{D}\) be a sub-\(\sigma\)-field of \(\mathcal{F}\). An \(\mathbb{R}^p\)-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).

This definition is often used when the \(\sigma\)-field \(\mathcal{D}\) is itself generated by some rv \(Y : \Omega \to \mathbb{R}^q\); this \(\sigma\)-field, denoted \(\sigma(Y)\), is defined by

\[
\sigma(Y) \equiv \{Y^{-1}(C) : C \in \mathcal{B}(\mathbb{R}^q)\}
\]

as expected.

We have the following important operational characterization of \(\sigma(Y)\)-measurability.
Lemma 7.1.1 Assume the \( \sigma \)-field \( \mathcal{D} \) to be 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 \( X = g(Y) \) is \( \mathcal{D} \)-measurable.

(ii) Conversely, any \( \mathcal{D} \)-measurable rv \( X : \Omega \to \mathbb{R} \) can be written in the form \( X = g(Y) \) for some Borel mapping \( g : \mathbb{R}^q \to \mathbb{R} \).

Proof. (i) The conclusion is immediate from the fact that

\[
[X \in B] = [g(Y) \in B] = [Y \in g^{-1}(B)] \in \mathcal{D}, \quad B \in \mathcal{B}(\mathbb{R}).
\]

(7.1)

since \( Y \) is \( \mathcal{D} \)-measurable and \( g^{-1}(B) \) belongs to \( \mathcal{B}(\mathbb{R}^p) \) by the Borel measurability of \( g \).

(ii) Conversely, assume that the rv \( X : \Omega \to \mathbb{R} \) is \( \mathcal{D} \)-measurable. The proof proceeds in three standard steps:

Simple rvs First assume that \( X = 1[D] \) 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 \( X = g_C(Y) \) with Borel mapping \( g_C : \mathbb{R}^q \to \mathbb{R} \) given by

\[
g_C(y) = \begin{cases} 
0 & \text{if } y \notin C \\
0 & \text{if } y \in C.
\end{cases}
\]

(7.2)

The desired conclusion is readily seen to hold for simple \( \mathcal{D} \)-measurable rvs of the form

\[X = \sum_{i \in I} a_i 1[D_i]\]

where \( I \) is a finite index, \( \{D_i, \ i \in I\} \) form a \( \mathcal{D} \)-partition of \( \Omega \) 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 \( X = g(Y) \) 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 (7.2).
### Non-negative rvs

For any non-negative $\mathcal{D}$-measurable rv $X : \Omega \to \mathbb{R}_+$, we introduce the usual monotone increasing sequence of simple rvs $\{X_n, n = 1, 2, \ldots\}$ given by

$$X_n = \sum_{m=0}^{n-1} \sum_{k=0}^{2^n-1} \frac{k}{2^n} \mathbf{1} \left[ \frac{k}{2^n} < X \leq \frac{k+1}{2^n} \right], \quad n = 1, 2, \ldots$$

with $\lim_{n \to \infty} X_n = X$. Obviously, the simple rvs $\{X_n, n = 1, 2, \ldots\}$ 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

$$X_n = g_n(Y), \quad n = 1, 2, \ldots$$

with the pointwise convergence implying

$$X(\omega) = \lim_{n \to \infty} X_n(\omega) = \lim_{n \to \infty} g_n(Y(\omega)), \quad \omega \in \Omega.$$

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 is easy to check 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 $\mathbb{R}^q \to \mathbb{R}$. By construction it is plain that $X = g(Y)$ since $Y(\omega)$ lies in $L$ for each $\omega$ in $\Omega$.

### The general case

The case of an arbitrary $\mathcal{D}$-measurable rv $X : \Omega \to \mathbb{R}$ is handled in the usual manner: Just write $X = X^+ - X^-$, and apply the last conclusion to each of the rvs $X^+$ and $X^-$. In particular, there exist Borel mappings $g_+ : \mathbb{R}^q \to \mathbb{R}_+$ and $g_- : \mathbb{R}^q \to \mathbb{R}_+$ such that $X^+ = g_+(Y)$ and $X^- = g_-(Y)$.

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 $g_\pm(y) = \max(0, \pm g(y))$. 

\[ \square \]
CHAPTER 7. CONDITIONING AND CONDITIONAL EXPECTATIONS

7.2 Conditional distributions and their expectations

We first return to the definitions given in Section 1.8: Let $D$ be an event in $\mathcal{F}$. With $\mathbb{P}[D] > 0$, the conditional probability measure $Q_D : \mathcal{F} \to [0, 1]$ was defined by

$$Q_D(A) = \frac{\mathbb{P}[A \cap D]}{\mathbb{P}[D]}, \quad A \in \mathcal{F}.$$  

When $\mathbb{P}[D] = 0$, it is convenient to take $Q_D : \mathcal{F} \to [0, 1]$ to be an arbitrary probability measure on $(\Omega, \mathcal{F})$ – This issue will be revisited later in this chapter.

It is now possible to define the conditional expectation of $X$ given $D$, denoted $\mathbb{E}[X | D]$: It is simply the expectation of the rv $X$ evaluated under the conditional probability measure $Q_D$ defined on $(\Omega, \mathcal{F})$. In particular, we note that

$$\mathbb{E}[1[A] | D] = Q_D[A] = \frac{\mathbb{P}[A \cap D]}{\mathbb{P}[D]} = \frac{\mathbb{E}[1[D]1[A]]}{\mathbb{P}[D]}, \quad A \in \mathcal{F}. \quad (7.3)$$

This observation leads readily to the following characterization.

**Lemma 7.2.1** Let $D$ be an event in $\mathcal{F}$ such that $\mathbb{P}[D] > 0$. For any rv $X : \Omega \to \mathbb{R}$, the conditional expectation of $X$ given $D$ exists (resp. exists and is finite) if the expectation $\mathbb{E}[X]$ exists (resp. exists and is finite), in which case the relation

$$\mathbb{E}[X | D] = \frac{\mathbb{E}[1[D]X]}{\mathbb{P}[D]} \quad (7.4)$$

holds.

The proof of this result is carried out through the usual three step process: It holds for indicator rvs by virtue of (7.3), thus for simple rvs by linearity of expectation. Non-negative rvs are handled via a staircase approximation argument with simple rvs, and arbitrary rvs by means of the standard decomposition.

When $\mathbb{P}[D] = 0$, it is will be convenient to take $Q_D : \mathcal{F} \to [0, 1]$ to be an arbitrary probability measure on $(\Omega, \mathcal{F})$, say $\mathbb{P}$ for the sake of concreteness. However, regardless of the choice we will always have

$$\mathbb{E}[1[D]X] = \mathbb{E}[X | D] \cdot \mathbb{P}[D] \quad (7.5)$$

as this the case for indicator rvs.
7.3 Conditioning with respect to a partition

In the next step we condition with respect to a partition: Thus, consider an $\mathcal{F}$-partition $\{D_i, i \in I\}$ where $I$ is a countable index set, i.e.,

$$D_i \cap D_j = \emptyset, \quad i \neq j \quad \text{and} \quad \cup_{i \in I} D_i = \Omega.$$ 

It is plain that

$$\sum_{i \in I} 1[D_i] = 1. \quad (7.6)$$

Throughout the events in the partition are assumed to be non-empty although it is possible to have $P[D_i] = 0$ for some $i$ in $I$ (but not for all indices $i$ in $I$ since $P[\Omega] = 1$).

We introduce the sub-$\sigma$-field $\mathcal{D} = \sigma(D_i, i \in I)$ induced by the partition $\{D_i, i \in I\}$: It is easy to check that every element $D$ of $\mathcal{D}$ is of the form

$$D = \cup_{j \in J} D_j$$

for some countable subset $J \subseteq I$ (possibly empty if $D = \emptyset$ or $J = I$ if $D = \Omega$).

The decomposition

$$\sum_{j \in J} 1[D_j] = 1[D] \quad (7.8)$$

generalizes (7.6) and will be used on several occasions.

**Fact 7.3.1** Consider an $\mathcal{D}$-measurable rv $X : \Omega \to \mathbb{R}^p$ where $\mathcal{D} = \sigma(D_i, i \in I)$. 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 $[X = x]$ is an element in $\mathcal{D}$. The result immediately follows since any element $D$ of $\mathcal{D}$ is necessarily of the form (7.7) for some countable subset $J \subseteq I$. \[ \square \]

**Definition 7.3.1** Let $\{D_i, i \in I\}$ denote a $\mathcal{F}$-partition with countable index set $I$. Consider a rv $X : \Omega \to \mathbb{R}$ such that $E[X]$ exists. The conditional expectation of $X$ given the partition $\{D_i, i \in I\}$ is the extended rv $\Omega \to [-\infty, \infty]$ defined by

$$E[X|D_i, i \in I] \equiv \sum_{i \in I} E[X|D_i] 1[D_i]$$

where for each $i$ in $I$, $E[X|D_i]$ is the expectation of $X$ under the conditional probability distribution of $X$ given $D_i$. 

This definition is well posed as a result of Lemma 7.2.1. In particular, the rv
\( E[X_D], i \in I \) is an \( \mathbb{R} \)-valued rv as soon as \( E[X] \) is finite; as an extended rv, it is \( D \)-measurable rv in the sense that
\[
[E[X_D], i \in I] \in C \quad C \in \mathcal{B}(\mathbb{R}).
\]
Indeed we have
\[
[E[X|D_D], i \in I] \in C = \bigcup_{i \in I}([E[X|B_D], i \in I] \in C \cap D_D)
\]
\[
= \bigcup_{i \in I}([E[X|B_D] \in C \cap D_D].
\]

The following characterization should be pointed out, and foreshadows forthcoming developments.

**Lemma 7.3.1** For any rv \( X : \Omega \to \mathbb{R} \) with \( E[|X|] < \infty \), it holds that
\[
(7.9) \quad E[1_D E[X|D_D], i \in I]] = E[1_D X], \quad D \in D.
\]

**Proof.** It suffices to assume that the rv \( X \) is non-negative, say \( X : \Omega \to \mathbb{R}_+ \). Obviously \( E[X|D_D] \geq 0 \) for each \( i \) in \( I \). Fix \( D \) in \( D \), say \( D = \bigcup_{j \in J} D_j \) for some \( J \subseteq I \). Thus,
\[
E[1_D X] = E\left[\left( \sum_{j \in J} 1[D_j] \right) X \right]
\]
\[
= \sum_{j \in J} E[1[D_j]X] \quad \text{[By Monotone Convergence]}
\]
\[
= \sum_{j \in J: \mathbb{P}[D_j] > 0} E[X|D_j] \cdot \mathbb{P}[D_j]
\]
\[
= \sum_{j \in J: \mathbb{P}[D_j] > 0} E[E[X|D_j] 1[D_j]]
\]
\[
= \sum_{j \in J: \mathbb{P}[D_j] > 0} E[E[X|D_i, i \in I] 1[D_j]]
\]
\[
= E\left[\sum_{j \in J: \mathbb{P}[D_j] > 0} E[X|D_i, i \in I] 1[D_j] \right] \quad \text{[By Monotone Convergence]}
\]
7.3. CONDITIONING WITH RESPECT TO A PARTITION

\[ = \mathbb{E} \left( \left( \sum_{j \in J : \mathbb{P}(D_j) > 0} 1[D_j] \right) \mathbb{E}[X | D_i, i \in I] \right) \]

\[ = \mathbb{E} \left( \sum_{j \in J} 1[D_j] \right) \mathbb{E}[X | D_i, i \in I] \]

\[ = \mathbb{E}[1[D] \mathbb{E}[X | D_i, i \in I]] \]

as desired. In the equality before last we have used the fact that

\[ \sum_{j \in J : \mathbb{P}(D_j) > 0} 1[D_j] = \sum_{j \in J} 1[D_j] \quad \text{a.s.} \]

The arbitrary case is obtained by using the usual decomposition \( X = X^+ - X^- \) with \( \mathbb{E}[X^\pm] \) finite: Starting with the definition, we get

\[ \mathbb{E}[X^\pm | D_i, i \in I] = \sum_{i \in I} \mathbb{E}[X^\pm | D_i] \]

so that

\[ \mathbb{E}[X^+ | D_i, i \in I] - \mathbb{E}[X^- | D_i, i \in I] \]

\[ = \sum_{i \in I} \mathbb{E}[X^+ | D_i] 1[D_i] - \sum_{i \in I} \mathbb{E}[X^- | D_i] 1[D_i] \]

\[ = \sum_{i \in I} \mathbb{E}[X^+ - X^- | D_i] 1[D_i] \]

\[ = \sum_{i \in I} \mathbb{E}[X | D_i] 1[D_i] \]

(7.11)

The first part of the proof yields

\[ \mathbb{E}[1[D] \mathbb{E}[X^\pm | D_i, i \in I]] = \mathbb{E}[1[D] X^\pm], \quad D \in \mathcal{D}. \]

Noting that \( \mathbb{E}[X^\pm | D_i, i \in I] \geq 0 \) a.s., we conclude from (7.12) (by taking \( D = \Omega \)) that the rv \( \mathbb{E}[X^\pm | D_i, i \in I] \) has a finite expectation and so does \( \mathbb{E}[X | D_i, i \in I] \) by virtue of (7.11). Therefore,

\[ \mathbb{E}[1[D] \mathbb{E}[X | D_i, i \in I]] \]

\[ = \mathbb{E}[1[D] \mathbb{E}[X^+ | D_i, i \in I] - \mathbb{E}[X^- | D_i, i \in I]] \]

\[ = \mathbb{E}[1[D] \mathbb{E}[X^+ | D_i, i \in I]] - \mathbb{E}[1[D] \mathbb{E}[X^- | D_i, i \in I]] \]

\[ = \mathbb{E}[1[D] X^+] - \mathbb{E}[1[D] X^-] \]

(7.13)

\[ = \mathbb{E}[1[D] X]. \]
The rv $\mathbb{E}[X|D_i, i \in I]$ given in Definition 7.3.1 is essentially the only $\mathcal{D}$-measurable rv which satisfies the conditions (7.9).

**Lemma 7.3.2** Consider a rv $X : \Omega \rightarrow \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$, and let $Z_1, Z_2 : \Omega \rightarrow \mathbb{R}$ be $\mathcal{D}$-measurable rvs such that both $\mathbb{E}[Z_1]$ and $\mathbb{E}[Z_2]$ exist. If both satisfy (7.9), namely

$$
\mathbb{E}[1[D]Z_k] = \mathbb{E}[1[D]X], \quad D \in \mathcal{D}, \quad k = 1, 2,
$$

then $\mathbb{E}[|Z_1|] < \infty$ and $\mathbb{E}[|Z_2|] < \infty$, and $Z_1 = Z_2$ a.s.

**Proof.** Fix $k = 1, 2$. By using the $\mathcal{D}$-measurable events $D_k^+ = [Z_k \geq 0]$ and $D_k^- = [Z_k \leq 0]$ in (7.14), we conclude that $\mathbb{E}[1[D]Z_k]$ is finite by virtue of $\mathbb{E}[|X|] < \infty$. Thus, the rv $Z_k^\pm$ has a finite expectation, and so does the rv $Z_k$. The condition (7.14) can now be rewritten as

$$
\mathbb{E}[1[D]Z] = 0, \quad D \in \mathcal{D}
$$

with $\mathcal{D}$-measurable rv $Z = Z_1 - Z_2$. Using (7.15) with $\mathcal{D}$-measurable events $D_+ = [Z > 0]$ and $D_- = [Z < 0]$, we readily conclude that $1[D_\pm]Z = Z^\pm = 0$ a.s., whence $Z = Z^+ - Z^- = 0$ a.s. ■

In view of Lemma 7.3.2, the rv $\mathbb{E}[X|D_i, i \in I]$ belongs to an a.s. equivalence class of $\mathcal{D}$-measurable rvs which all satisfy (7.14). 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 $\sigma$-field $\mathcal{D}$, and we denote any of its representative by $\mathbb{E}[X|\mathcal{D}]$.

### 7.4 Conditioning with respect to a partition – Miscellaneous properties

The form of $\mathbb{E}[X|\mathcal{D}]$ suggests that conditional expectations inherit most, if not all, the properties of (unconditional) expectations. There is however a set of properties that are unique to conditional expectations; all of them can be established by taking advantage of Definition 7.3.1. We leave the details as exercises to the interested reader.
7.4. CONDITIONING WITH RESPECT TO A PARTITION – MISCELLANEOUS PROPERTIES

The first result gives support to the notion that when conditioning with respect to the $\sigma$-field $\mathcal{D}$, any $\mathcal{D}$-measurable rv is essentially fixed (i.e., can be considered to be a constant) in that conditional integration.

Lemma 7.4.1 Consider a rv $X : \Omega \to \mathbb{R}$ such that $E[|X|] < \infty$. If the rv $Z : \Omega \to \mathbb{R}$ is a $\mathcal{D}$-measurable rv with $E[|ZX|] < \infty$, then

$$E[ZX|\mathcal{D}] = ZE[X|\mathcal{D}] \text{ a.s.}$$

Next, we consider two $\mathcal{F}$-partitions $\{D_i, i \in I\}$ and $\{D'_j, j \in J\}$ where $I$ and $J$ are countable index sets.

Definition 7.4.1 The partition $\{D'_j, j \in J\}$ is said to be finer than the partition $\{D_i, i \in I\}$ if for every $i$ in $I$ there exists a subset $J(i) \subseteq J$ such that

$$D_i = \bigcup_{j \in J(i)} D'_j.$$  

Equivalently, the partition $\{D_i, i \in I\}$ is said to be coarser than the partition $\{D'_j, j \in J\}$.

In that case, $\mathcal{D}$ is a sub-$\sigma$-field of $\mathcal{D}'$, i.e., $\mathcal{D} \subseteq \mathcal{D}'$, where $\mathcal{D} \equiv \sigma(D_i, i \in I)$ as before and $\mathcal{D}' \equiv \sigma(D'_j, j \in J)$. The following properties of iterated conditioning can be shown by direct arguments.

Lemma 7.4.2 Assume the $\mathcal{F}$-partition $\{D'_j, j \in J\}$ to be finer than the $\mathcal{F}$-partition $\{D_i, i \in I\}$. For any rv $X : \Omega \to \mathbb{R}$ such that $E[|X|] < \infty$, it holds that

$$E[E[X|\mathcal{D}']|\mathcal{D}] = E[X|\mathcal{D}] \text{ a.s.}$$

and

$$E[E[X|\mathcal{D}]|\mathcal{D}'] = E[X|\mathcal{D}] \text{ a.s.}$$

The next lemma explores the interplay between conditioning and independence.

Lemma 7.4.3 Consider a rv $X : \Omega \to \mathbb{R}$ which is independent of the $\sigma$-field $\mathcal{D}$. If $E[|X|] < \infty$, then

$$E[X|\mathcal{D}] = E[X] \text{ a.s.}$$
Here, the independence of the rv $X$ from the $\sigma$-field $D$ means that for each $D$ in $D$, the rvs $X$ and $1_D$ are independent.

**Proof.** By definition we have
\[
\mathbb{E}[X|D] = \sum_{i \in I} \mathbb{E}[X|D_i] \mathbb{1}[D_i] \quad \text{a.s.}
\]
\[
= \sum_{i \in I} \frac{\mathbb{E}[1_D X]}{\mathbb{P}[D_i]} \mathbb{1}[D_i] \quad \text{a.s.}
\]
\[
= \sum_{i \in I: \mathbb{P}[D_i] > 0} \frac{\mathbb{E}[1_D] \mathbb{E}[X]}{\mathbb{P}[D_i]} \mathbb{1}[D_i] \quad \text{[By independence]}
\]
\[
= \sum_{i \in I: \mathbb{P}[D_i] > 0} \mathbb{E}[X] \mathbb{1}[D_i]
\]
\[
= \left( \sum_{i \in I: \mathbb{P}[D_i] > 0} \mathbb{1}[D_i] \right) \mathbb{E}[X] \quad \text{(7.16)}
\]
\[
= \mathbb{E}[X] \quad \text{a.s.}
\]
since $\sum_{i \in I: \mathbb{P}[D_i] > 0} 1_D = 1$ a.s.

\[
(7.16)
\]

### 7.5 The general definition of conditional expectations

We now present a general definition for the conditional expectation given an arbitrary $\sigma$-field.

**Theorem 7.5.1** Let $D$ be a sub-$\sigma$-field of $\mathcal{F}$, and consider a rv $X : \Omega \to \mathbb{R}$ such that $\mathbb{E}[|X|] < \infty$.

(i) (Existence) There exists a $D$-measurable rv $Z : \Omega \to \mathbb{R}$ with $\mathbb{E}[|Z|] < \infty$ such that
\[
\mathbb{E}[1_D Z] = \mathbb{E}[1_D X], \quad D \in D.
\]

(ii) (Uniqueness) If the $D$-measurable rvs $Z_1, Z_2 : \Omega \to \mathbb{R}$ with $\mathbb{E}[|Z_1|] < \infty$ and $\mathbb{E}[|Z_2|] < \infty$ both satisfy (7.17), namely
\[
\mathbb{E}[1_D Z_k] = \mathbb{E}[1_D X], \quad k = 1, 2, \quad D \in D
\]
then $Z_1 = Z_2$ a.s.
7.5. THE GENERAL DEFINITION OF CONDITIONAL EXPECTATIONS

It is easy to see

**Proof.** Existence is a consequence of the Radon-Nikodym Theorem. The \( D \)-measurable rvs with finite expectation satisfying (7.17) form an equivalence class; any one of its representatives will be denoted by \( \mathbb{E}[X|D] \).

We are now listing several basic properties of conditional expectations.

**A. Multiplying by a constant**  
*For any \( X : \Omega \to \mathbb{R} \) with \( \mathbb{E}[|X|] < \infty \), and any \( \alpha \) in \( \mathbb{R} \), we have*  
\[
\mathbb{E}[\alpha X|D] = \alpha \mathbb{E}[X|D] \quad \text{a.s.}
\]

Indeed, for any event \( D \) in \( \mathcal{D} \),
\[
\mathbb{E}[1[D] \mathbb{E}[\alpha X|D]] = \mathbb{E}[1[D] \alpha X] = \alpha \mathbb{E}[1[D] X] = \alpha \mathbb{E}[1[D] \mathbb{E}[X|D]] = \mathbb{E}[1[D] \alpha \mathbb{E}[X|D]]
\]

(7.19)

and the conclusion follows by uniqueness since the rv \( \alpha \mathbb{E}[X|D] \) is \( D \)-measurable.

**B. 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*  
\[
\mathbb{E}[X|D] \leq \mathbb{E}[Y|D] \quad \text{a.s.}
\]

**C. Taking absolute values**  
*If \( \mathbb{E}[|X|] < \infty \), then \( |\mathbb{E}[X|D]| \leq \mathbb{E}[|X||D] \) a.s.*  
The result is a simple consequence of B as we note that \( -|X| \leq X \leq |X| \).

**D. Localization**  
*For any rvs \( X, Z : \Omega \to \mathbb{R} \) with \( \mathbb{E}[|X|] < \infty \) and \( \mathbb{E}[|ZX|] < \infty \), we have*  
\[
\mathbb{E}[ZX|D] = Z \mathbb{E}[X|D] \quad \text{a.s.}
\]

whenever the rv \( Z \) is \( D \)-measurable.
E. Addition  For any rvs $X, Y : \Omega \to \mathbb{R}$ with $E[|X|] < \infty$ and $E[|Y|] < \infty$, we have

$$E[X + Y|D] = E[X|D] + E[Y|D] \quad a.s.$$  

For any event $D$ in $\mathcal{D}$,

$$E[1[D]E[X + Y|D]] = E[1[D](X + Y)]$$


$$= E[1[D]E[X|D]] + E[1[D]E[Y|D]]$$

(7.20)

$$= E[1[D](E[X|D] + E[Y|D])]$$

and the conclusion follows by uniqueness since the rv $E[X|D] + E[Y|D]$ is $\mathcal{D}$-measurable.

7.6 Iterated conditioning

Lemma 7.6.1 For any rv $X : \Omega \to \mathbb{R}$ such that $E[|X|] < \infty$, the rv $E[X|\mathcal{D}]$ has a finite expectation with

(7.21)

$$E[E[X|\mathcal{D}]] = E[X].$$

Proof. Use (7.17) with $D = \Omega$.

Lemma 7.6.2 Let $\mathcal{D}$ and $\mathcal{D}'$ be two sub-$\sigma$-fields of $\mathcal{F}$ with $\mathcal{D} \subseteq \mathcal{D}'$. For any rv $X : \Omega \to \mathbb{R}$ with $E[|X|] < \infty$, we have

(7.22)

$$E[E[X|\mathcal{D}]|\mathcal{D}'] = E[X|\mathcal{D}] \quad a.s.$$ and

(7.23)

$$E[E[X|\mathcal{D}']|\mathcal{D}] = E[X|\mathcal{D}] \quad a.s.$$  

Proof. Obviously, the rv $E[X|\mathcal{D}]$ is $\mathcal{D}$-measurable, hence $\mathcal{D}'$-measurable. Applying Property D we get (7.22).
Pick $D$ in $\mathcal{D}$. We get

$$E \left[ 1 \left[D\right] E \left[X \mid D' \right] \mid D \right] = E \left[ 1 \left[D\right] \right] E \left[X \mid D' \right] \quad \text{[By (7.17) for $E \left[X \mid D' \right]$ and $D$]}
$$

$$= E \left[ 1 \left[D\right] X \right] \quad \text{[By (7.17) for $X$ and $D'$]}
$$

(7.24)

$$= E \left[ 1 \left[D\right] E \left[X \mid D \right] \right] \quad \text{[By (7.17) for $X$ and $D$]}
$$

The conclusion follows by uniqueness.

7.7 Conditional expectations and independence

**Lemma 7.7.1** Consider a rv $X : \Omega \to \mathbb{R}$ with $E \left[|X|\right] < \infty$. If the rv $X$ is independent of the $\sigma$-field $\mathcal{D}$, then

$$E \left[X \mid \mathcal{D} \right] = E \left[X \right] \quad \text{a.s.}
$$

(7.25)

Here, the independence of the rv $X$ from the $\sigma$-field $\mathcal{D}$ means that for each $D$ in $\mathcal{D}$, the rvs $X$ and $1 \left[D\right]$ are independent.

**Proof.** By independence we note that

$$E \left[ 1 \left[D\right] X \right] = P \left[D\right] E \left[X \right] = E \left[ 1 \left[D\right] E \left[X \right] \right], \quad D \in \mathcal{D}
$$

and the conclusion follows by uniqueness since the defining condition (7.17) holds for the constant rv $E \left[X \right]$ (which is of course $\mathcal{D}$-measurable).

Consider a Borel mapping $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$, and rvs $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$ such that $E \left[|h(X,Y)|\right] < \infty$. Define the mapping $\hat{h} : \mathbb{R}^q \to \mathbb{R}$ given by

$$\hat{h}(y) = E \left[h(X,y) \right], \quad y \in \mathbb{R}^q.
$$

This definition is always well posed, and produces a Borel mapping $\mathbb{R}^q \to \mathbb{R}$.

**Lemma 7.7.2** If the rv $X$ is independent of the $\sigma$-field $\mathcal{D}$ and the rv $Y$ is $\mathcal{D}$-measurable, then

$$E \left[h(X,Y) \mid \mathcal{D} \right] = \hat{h}(Y) \quad \text{a.s.}
$$

**Proof.** The proof proceeds according to the usual pattern.
CHAPTER 7. CONDITIONING AND CONDITIONAL EXPECTATIONS

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

\[
h(x, y) = 1 \{ y \in C \} g(x), \quad x \in \mathbb{R}^p, \quad y \in \mathbb{R}^q
\]

with Borel mapping \( g : \mathbb{R}^p \rightarrow \mathbb{R} \) such that \( \mathbb{E}[|g(X)|] < \infty \) and Borel subset \( C \) in \( \mathcal{B}(\mathbb{R}^q) \). For every \( D \) in \( \mathcal{D} \), the event \( D \cap \{ Y \in C \} \) belongs to \( \mathcal{D} \) under the foregoing assumptions. It follows that

\[
\mathbb{E}[1_D \mathbb{E}[h(X,Y)|D]] = \mathbb{E}[1_D h(X,Y)]
\]

and by uniqueness, we conclude that

\[
\mathbb{E}[h(X,Y)|D] = 1 \{ Y \in C \} \mathbb{E}[g(X)] = \hat{h}(Y) \quad a.s.
\]

upon noting that here

\[
\hat{h}(y) = \mathbb{E}[h(X,y)] = \mathbb{E}[1 \{ y \in C \} g(X)] = 1 \{ y \in C \} \mathbb{E}[g(X)], \quad y \in \mathbb{R}^q.
\]

Case II  The result immediately follows for any Borel mapping \( h : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R} \) of the form

\[
h(x, y) = \sum_{i \in I} 1 \{ y \in C_i \} g_i(x), \quad x \in \mathbb{R}^p, \quad y \in \mathbb{R}^q
\]

with \( I \) a finite index set, and for each \( i \) in \( I \), Borel mapping \( g_i : \mathbb{R}^p \rightarrow \mathbb{R} \) such that \( \mathbb{E}[|g_i(X)|] < \infty \) and Borel subset \( C_i \) in \( \mathcal{B}(\mathbb{R}^q) \). By additivity, we get

\[
\mathbb{E}[h(X,Y)|D] = \mathbb{E}\left[ \sum_{i \in I} 1 \{ Y \in C_i \} g_i(X)|D\right] \quad a.s.
\]

\[
= \sum_{i \in I} \mathbb{E}[1 \{ Y \in C_i \} g_i(X)|D] \quad a.s.
\]

\[
= \sum_{i \in I} 1 \{ Y \in C_i \} \mathbb{E}[g_i(X)] \quad a.s.
\]

\[
(7.27) \quad = \hat{h}(Y) \quad a.s.
\]
as we note that

\[ \hat{h}(y) = \mathbb{E}[h(X, y)] = \mathbb{E} \left[ \sum_{i \in I} 1 \{ y \in C_i \} g_i(X) \right] = \sum_{i \in I} 1 \{ y \in C_i \} \mathbb{E}[g_i(X)], \quad y \in \mathbb{R}^q. \] (7.28)

Case III

### 7.8 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

\[ \Omega_Y \equiv \bigcup_{i \in I} \{Y = y_i\} \]

has probability one, or equivalently, that the complement \( \Omega^c_Y \) 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 \( \Omega \).

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 \( \Omega \). The following facts are easy consequences from the following observation

\[ \mathbb{P}[Y \neq Y_b] \leq \mathbb{P}[\Omega^c_Y] = 0. \]
(i) The rvs $Y$ and $Y_b$ have the same probability distribution under $P$. If $X : \Omega \to \mathbb{R}^p$ is another rv, the pairs $(X, Y)$ and $(X, Y_b)$ have the same probability distribution under $P$.

(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 $D_b = \sigma([Y = b], [Y = y_i], i \in I)$, we note that
\[ \mathbb{E}[h(X, Y)|D_b] = \sum_{y \in S_b} \mathbb{E}[h(X, Y_b) | Y = y] 1 [Y_b = y] \]
\[ = \sum_{y \in S} \frac{\mathbb{E}[1 [Y_b = y] h(X, Y_b)]}{\mathbb{P}[Y_b = y]} 1 [Y_b = y] + \mathbb{E}[h(X, Y_b) | Y_b = b] 1 [Y_b = b] \]
\[ = \sum_{y \in S} \frac{\mathbb{E}[1 [Y_b = y] h(X, y)]}{\mathbb{P}[Y = y]} 1 [Y_b = y] + \mathbb{E}[h(X, Y_b) | Y_b = b] 1 [Y_b = b] \]
\[ = \sum_{y \in S} \frac{\mathbb{E}[1 [Y = y] h(X, y)]}{\mathbb{P}[Y = y]} 1 [Y = y] + \mathbb{E}[h(X, Y_b) | Y_b = b] 1 [Y_b = b] \]

It follows that
\[ \mathbb{E}[h(X, Y_b)|D_b] = \sum_{y \in S} \frac{\mathbb{E}[1 [Y = y] h(X, y)]}{\mathbb{P}[Y = y]} 1 [Y = y] \quad \mathbb{P}\text{-a.s.} \]

(iii) In light of this last calculation, with Borel mapping $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ such that $\mathbb{E}[|h(X, Y)|] < \infty$, for distinct values $b \neq c$ in $\mathbb{R}^q$, we have
\[ \mathbb{E}[h(X, Y_b)|D_b] = \mathbb{E}[h(X, Y_c)|D_c] \quad \mathbb{P}\text{-a.s.} \]
where we use the notation $D_b = \sigma([Y = b], [Y = y_i], i \in I)$ and $D_c = \sigma([Y = c], [Y = y_i], i \in I)$.

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 $P$). As this notion defines an equivalence relation on rvs, we write $\mathbb{E}[h(X, Y)|Y]$ (or sometimes $\mathbb{E}[h(X, Y)|\sigma(Y)]$) to denote any representative in the equivalence class.
7.8. PARTITIONS AND DISCRETE RVS

(iv) One standard representative in that class of $\mathbb{P}$-equivalent rvs is given by

$$\sum_{y \in S} \mathbb{E} [h(X, Y) | Y = y] \mathbbm{1} [Y = y]$$

(7.29)

Note that all the terms in (7.29) 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

$$\sum_{y \in S} \mathbb{E} [h(X, Y) | Y = y] \mathbbm{1} [Y = y] = \sum_{y \in S} \frac{\mathbb{E} \left[ \mathbbm{1} [Y = y] h(X, Y) \right]}{\mathbb{P} [Y = y]} \mathbbm{1} [Y = y]$$

$$= \sum_{y \in S} \frac{\mathbb{E} \left[ \mathbbm{1} [Y = y] h(X, y) \right]}{\mathbb{P} [Y = y]} \mathbbm{1} [Y = y]$$

(7.30)

This last expression suggests introducing the mapping $\hat{h} : \mathbb{R}^q \to \mathbb{R}$ given by

$$\hat{h}(y) = \begin{cases} \mathbb{E} [h(X, y) | Y = y] & \text{if } y \in S \\ h^*(y) & \text{if } y \notin S \end{cases}$$

where $h^* : \mathbb{R}^q \to \mathbb{R}$ is an arbitrary Borel mapping such that $\mathbb{E} [|h^*(Y)|] < \infty$.

This definition is always well posed, and produces a Borel mapping $\mathbb{R}^q \to \mathbb{R}$.

With this notation we conclude that

$$\sum_{y \in S} \mathbb{E} [h(X, Y) | Y = y] \mathbbm{1} [Y = y]$$

$$= \sum_{y \in S} \mathbb{E} [h(X, y) | Y = y] \mathbbm{1} [Y = y]$$

$$= \sum_{y \in S} \hat{h}(y) \mathbbm{1} [Y = y]$$

$$= \sum_{y \in S} \hat{h}(Y) \mathbbm{1} [Y = y]$$

$$= \hat{h}(Y) \left( \sum_{y \in S} \mathbbm{1} [Y = y] \right)$$

(7.31)
since
\[ \sum_{y \in S} 1 \left[ Y = y \right] = 1 \left[ Y \in S \right] = 1 \text{ } \mathbb{P}\text{-a.s.} \]

(vi) Symbolically, this last discussion can be summarized as follows:
\[
\mathbb{E} \left[ h(X, Y) \mid Y \right] = \left( \mathbb{E} \left[ h(X, Y) \mid Y = y \right] \right)_{y=Y} \\
= \left( \mathbb{E} \left[ h(X, y) \mid Y = y \right] \right)_{y=Y} \text{ } \mathbb{P}\text{-a.s.} \tag{7.32}
\]

7.9 The absolutely continuous case

Consider rvs \( X : \Omega \rightarrow \mathbb{R}^p \) and \( Y : \Omega \rightarrow \mathbb{R}^q \). If the rv \( Y \) is absolutely continuous, then
\[
\mathbb{P} \left[ Y = y \right] = 0, \text{ } y \in \mathbb{R}^q
\]
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
\[
\mathbb{P} \left[ X \in B \mid Y = y \right] = \frac{\mathbb{P} \left[ X \in B, Y = y \right]}{\mathbb{P} \left[ Y = y \right]}, \text{ } B \in \mathcal{B}(\mathbb{R}^p).
\]

With \( y \) in \( \mathbb{R}^q \), the ball centered at \( y \) with radius \( \varepsilon > 0 \) is denoted by \( B_{\varepsilon}(y) \), i.e.,
\[
B_{\varepsilon}(y) \equiv \{ \eta \in \mathbb{R}^q : \| \eta - y \| \leq \varepsilon \}.
\]

Pick \( y \) in \( \mathbb{R}^q \) such that \( f_Y(y) > 0 \) and assume there exists \( \varepsilon_0 > 0 \) such that
\[
\mathbb{P} \left[ Y \in B_{\varepsilon}(y) \right] > 0, \text{ } 0 < \varepsilon \leq \varepsilon_0.
\]

The basic idea is as follows: Pick \( B \) in \( \mathcal{B}(\mathbb{R}^p) \). Whatever definition is given to the conditional probability \( \mathbb{P} \left[ X \in B \mid Y = y \right] \), it is reasonable to expect that it should be compatible with the limiting value \( \lim_{\varepsilon \downarrow 0} \mathbb{P} \left[ X \in B \mid Y \in B_{\varepsilon}(y) \right] \) if it exists.

With this in mind we note that
\[
\mathbb{P} \left[ X \in B \mid Y \in B_{\varepsilon}(y) \right] = \frac{\mathbb{P} \left[ X \in B \cap B_{\varepsilon}(y) \right]}{\mathbb{P} \left[ B_{\varepsilon}(y) \right]} \cdot \frac{\int_{B \times B_{\varepsilon}(y)} f_{XY}(\xi, \eta) d\xi d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) d\eta}
\]
7.9. THE ABSOLUTELY CONTINUOUS CASE

\[
\begin{align*}
\int_B \left( \frac{\int_{B_{\varepsilon}(y)} f_{XY}(\xi, \eta) \, d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) \, d\eta} \right) \, d\xi \\
= \int_B \left( \frac{\int_{B_{\varepsilon}(y)} f_{XY}(\xi, \eta) \, d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) \, d\eta} \right) \, d\xi
\end{align*}
\]

(7.33)

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

\[
\lim_{\varepsilon \downarrow 0} \frac{\int_{B_{\varepsilon}(y)} f_{XY}(\xi, \eta) \, d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) \, d\eta} = \frac{\int_{B_{\varepsilon}(y)} f_{XY}(\xi, \eta) \, d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) \, d\eta}
\]

(7.34)

This suggests

\[
\lim_{\varepsilon \downarrow 0} \mathbb{P} \left[ X \in B \mid Y \in B_{\varepsilon}(y) \right] = \lim_{\varepsilon \downarrow 0} \int_B \left( \frac{\int_{B_{\varepsilon}(y)} f_{XY}(\xi, \eta) \, d\eta}{\int_{B_{\varepsilon}(y)} f_Y(\eta) \, d\eta} \right) \, d\xi
\]
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

\[
f_{X|Y}(x|y) = \begin{cases} 
\frac{f_{XY}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\
g(x) & \text{if } f_Y(y) = 0
\end{cases}
\]

where the Borel mapping \( g : \mathbb{R}^p \to \mathbb{R}_+ \) is a probability density, 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 \( B(\mathbb{R}^q) \). 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

\[
C_Y^+ \equiv \{ \eta \in \mathbb{R}^q : f_Y(\eta) > 0 \},
\]

this becomes

\[
\mathbb{P} \left[ Y \notin C_Y^+ \right] = \mathbb{P} \left[ f_Y(Y) = 0 \right] = 0.
\]

We find

\[
\mathbb{E} \left[ 1 \{ Y \in C \} u(X) \right] = \mathbb{E} \left[ 1 \{ Y \in C, f_Y(Y) > 0 \} u(X) \right] = \int_{\mathbb{R}^p \times (C \cap C_Y^+)} u(\xi) f_{XY}(\xi, \eta) d\xi d\eta = \int_{C \cap C_Y^+} \left( \int_{\mathbb{R}^p} u(\xi) f_{XY}(\xi, \eta) d\xi \right) d\eta
\]

by Fubini’s Theorem.
7.9. THE ABSOLUTELY CONTINUOUS CASE

If \( f_Y(\eta) > 0 \), then

\[
\int_{\mathbb{R}^p} u(\xi) f_{XY}(\xi, \eta) d\xi = \int_{\mathbb{R}^p} u(\xi) f_{X|Y}(\xi|\eta) f_Y(\eta) d\xi
\]

\[
= \left( \int_{\mathbb{R}^p} u(\xi) f_{X|Y}(\xi|\eta) d\xi \right) f_Y(\eta)
\]

(7.36)

\[
= \widehat{u}(\eta) f_Y(\eta)
\]

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

\[
\mathbb{E} \left[ 1 \left[ Y \in C \right] u(X) \right] = \int_{C \cap C_Y^*} \widehat{u}(\eta) f_Y(\eta) d\eta
\]

\[
= \int_C \widehat{u}(\eta) f_Y(\eta) d\eta
\]

(7.37)

\[
= \mathbb{E} \left[ 1 \left[ Y \in C \right] \widehat{u}(Y) \right] .
\]

Recalling that \( \sigma(Y) = \{ Y \in C, \ C \in \mathcal{B}(\mathbb{R}^q) \} \), we conclude that

\[
\mathbb{E} [u(X)|\sigma(Y)] = \widehat{u}(Y) \quad \mathbb{P}\text{-a.s.}
\]

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 \( \mathbb{E}[|v(X, Y)|] < \infty \), Then,

(7.38)

\[
\mathbb{E} \left[ 1 \left[ Y \in C \right] v(X, Y) \right] = \mathbb{E} \left[ 1 \left[ Y \in C \right] \widehat{v}(Y) \right] .
\]

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

\[
\mathbb{E} [v(X, Y)|\sigma(Y)] = \widehat{v}(Y) \quad \mathbb{P}\text{-a.s.}
\]
7.10 Exercises

Ex. 7.1 With $\mathcal{D}$ a sub-$\sigma$-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 $\sigma$-field $\mathcal{D} = \{\emptyset, \Omega\}$.

Ex. 7.2 Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two sub-$\sigma$-fields of $\mathcal{F}$ such that $\mathcal{D}_1 \subseteq \mathcal{D}_2$ (so $\mathcal{D}_1$ is a sub-$\sigma$-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. 7.3 Let $X$ denote a geometric rv with parameter $0 < a < 1$, namely

$$P[X = k] = (1 - a)a^k, \quad k = 0, 1, \ldots$$

a. Compute the conditional probabilities

$$P[X = k + \ell | X \geq k], \quad k, \ell = 0, 1, \ldots$$

b. (Converse) Consider now a discrete rv $Y$ with support $\{0, 1, \ldots\}$ and pmf $(p_r, \ r = 0, 1, \ldots)$. Define

$$q_{\ell|k} := P[Y = k + \ell | Y \geq k], \quad k, \ell = 0, 1, \ldots$$

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. 7.4 With $a$ in $(0, 1)$, consider a collection of mutually independent Bernoulli rvs $\{B_k, \ k = 1, 2, \ldots\}$ with

$$P[X_k = 1] = 1 - P[X_k = 0] = a, \quad k = 1, 2, \ldots$$

For each $n = 1, 2, \ldots$, define the partial sums $S_n = B_1 + \ldots + B_n$.

a. For each $n = 1, 2, \ldots$ compute the conditional probabilities

$$P[B_k = b | S_n = s], \quad b = 0, 1 \quad s = 0, 1, \ldots, n \quad k = 1, \ldots, n$$
Is the result surprising?

b. For each \( n = 1, 2, \ldots \) compute the conditional probabilities

\[
\mathbb{P} [B_k = b_k, B_\ell = b_\ell | S_n = s], \quad b_k, b_\ell \in \{0, 1\}, \quad s = 0, 1, \ldots, n \\
k \neq \ell, k, \ell = 1, \ldots, n
\]

Are the rvs \( B_k \) and \( B_\ell \) conditionally independent given that \( S_n = s \)?

**Ex. 7.5** Consider a rv \( X : \Omega \to \mathbb{R} \) of the discrete type with \( \mathbb{P} [X \in S] = 1 \) where \( S \equiv \{a_i, \ i \in I\} \) for some countable index set \( I \). Let \( B \) an event such that \( \mathbb{P} [B] > 0 \) – Obviously both \( X \) and \( B \) are defined on the same sample space \( \Omega \) with \( B \) in \( \mathcal{F} \). Define the function \( F(\cdot | B) : \mathbb{R} \to [0, 1] \) by

\[
F(x|B) \equiv \mathbb{P} [X \leq x | B], \quad x \in \mathbb{R}.
\]

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 \( \mathbb{P} [X = a | B] > 0 \) for some \( a \) in \( \mathbb{R} \), then \( \mathbb{P} [X = a] > 0 \).

**Ex. 7.6** If the discrete rv \( X : \Omega \to \mathbb{R} \) has pmf given by

\[
\mathbb{P} [X = 1] = \mathbb{P} [X = 0] = \frac{1}{2},
\]

define the rv \( Y \equiv Y := 1 + (-1)^X \) 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. 7.7** Let \( N \) be a discrete rv whose support is contained in \( \mathbb{N}_0 \) (i.e., \( \mathbb{P} [N \in \mathbb{N}_0] = 1 \)) with a finite second moment, i.e., \( \mathbb{E} [N^2] < \infty \). 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. 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 7. CONDITIONING AND CONDITIONAL EXPECTATIONS

**Ex. 7.8** 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 \( \mathbb{P} [0 < P \leq 1] = 1 \). 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

\[
X \equiv \sum_{i=1}^{n} 1 [U_i \leq P].
\]

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 \( \text{Var} [X] \)?

c. Are the rvs \( 1 [U_1 \leq P], \ldots, 1 [U_n \leq P] \) (i) mutually independent (ii) pairwise uncorrelated when \( S \) contains at least two elements?

d. Compute the probabilities

\[
\mathbb{P} [X = k], \quad k = 0, 1, \ldots, n.
\]

How many moments of \( P \) are needed?

**Ex. 7.9** 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 \( \mathbb{E} [|X|] < \infty \).

a. Compute

\[
\mathbb{E} [X_i | X_1 + \ldots + X_n]
\]

for each \( i = 1, \ldots, n \). The answer does not depend on the (common) probability distribution 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

\[
\mathbb{E} [X_1 + \ldots + X_k | X_1 + \ldots + X_n]
\]

c. When \( 1 \leq k < n \), compute

\[
\mathbb{E} [X_1 + \ldots + X_n | X_1 + \ldots + X_k]
\]

**Ex. 7.10** 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 \( \mathbb{E} [|X|] < \infty \); and (iii) The rvs \( Y, Y_1, \ldots, Y_n \) are i.i.d. rvs with \( \mathbb{E} [|Y|] < \infty \) – The rvs \( X \) and \( Y \) do not necessarily have the same probability distribution.

By using basic properties of conditional expectations, compute

\[
\mathbb{E} \left[ X_1 Y_1 + \ldots + X_n Y_n \bigg| X_1 + \ldots + X_n, Y_1 + \ldots + Y_n \right]
\]
7.10. EXERCISES

[HINT: Use iterated conditioning and compute]

\[
\mathbb{E} \left[ X_1 Y_1 + \ldots + X_n Y_n \mid X_1, \ldots, X_n, Y_1 + \ldots + Y_n \right].
\]

**Ex. 7.11** 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

\[
X = \sum_{k=1}^{n} 1 \{ U_k \leq U \}
\]

**a.** With this definition as starting point, use direct probabilistic arguments to show that

\[
\mathbb{P} [ X = k ] = \frac{1}{n+1}, \quad k = 0, \ldots, n.
\]

**b.** Using conditioning arguments compute the conditional probability

\[
\mathbb{P} [ X = k \mid U ], \quad k = 0, \ldots, n.
\]

**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 8

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.

8.1 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 8.1.1** With any rv $X : \Omega \rightarrow \mathbb{R}^p$, we associate its characteristic function $\Phi_X : \mathbb{R}^p \rightarrow \mathbb{C}$ given by

$$ \Phi_X(\theta) \equiv \mathbb{E} \left[ e^{i \theta^t X} \right], \quad \theta \in \mathbb{R}^p. \quad (8.1) $$

Characteristic functions are always well defined regardless of the type of probability distribution function for the rv $X$: Indeed the definition (8.1) is well posed since for each $\theta$ in $\mathbb{R}^p$, the rvs $\Omega \rightarrow \mathbb{R} : \omega \rightarrow \cos (\theta^t X(\omega))$ and $\Omega \rightarrow \mathbb{R} : \omega \rightarrow \sin (\theta^t X(\omega))$ are both bounded. As a result, their expected values $\mathbb{E} \left[ \cos (\theta^t X) \right]$ and $\mathbb{E} \left[ \sin (\theta^t X) \right]$ are well defined and finite with

$$ |\mathbb{E} \left[ \cos (\theta^t X) \right]| \leq 1 \quad \text{and} \quad |\mathbb{E} \left[ \sin (\theta^t X) \right]| \leq 1. $$
This fact allows us to make sense of (8.1) by linearity through the relations
\[
\mathbb{E} \left[ e^{i\theta t X} \right] = \mathbb{E} \left[ \cos (\theta t X) + i \sin (\theta t X) \right]
\]
\begin{align*}
&= \mathbb{E} \left[ \cos (\theta t X) \right] + i \mathbb{E} \left[ \sin (\theta t X) \right], \quad \theta \in \mathbb{R}^p. \\
\end{align*}

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
\[
\Phi_X(\theta) = \sum_{x \in S} e^{i\theta t x} \mathbb{P} [X = x], \quad \theta \in \mathbb{R}^p.
\]

Obviously, the characteristic function \( \Phi_X \) of the rv \( X \) is determined by its probability distribution function \( F_X : \mathbb{R} \to [0, 1] \). In fact we could rewrite (8.1) as
\begin{equation}
\Phi_X(\theta) = \int_{\mathbb{R}^p} e^{i\theta t x} dF_X(x), \quad \theta \in \mathbb{R}^p.
\end{equation}

This suggests writing \( \Phi_X \) as \( \Phi_{F_X} \), and leads to the following definition.

**Definition 8.1.2** 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
\begin{equation}
\Phi_F(\theta) \equiv \int_{\mathbb{R}^p} e^{i\theta t x} dF(x), \quad \theta \in \mathbb{R}^p.
\end{equation}

### 8.2 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 8.2.1** 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
\[
F(b-) - F(a) + \frac{F(a) - F(a-)}{2} + \frac{F(b) - F(b-)}{2}
\]
\begin{equation}
= \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \cdot \Phi_F(t) \, dt
\end{equation}

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 (8.5) yields

\begin{equation}
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{8.6}
\end{equation}

In the language of rvs, Theorem 8.2.1 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{equation}
\mathbb{P}[a < X < b] + \mathbb{P}[X = a] + \mathbb{P}[X = b] = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \cdot \Phi_X(t) dt \tag{8.7}
\end{equation}

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 8.2.2** 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 8.2.2 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 8.3 and 8.4.

**Definition 8.2.1** 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

\begin{equation}
\Phi(\theta) = \int_{\mathbb{R}^p} e^{i\theta^T x} dF(x), \quad \theta \in \mathbb{R}^p \tag{8.8}
\end{equation}

in which case $\Phi = \Phi_F$ by Theorem 8.2.2.
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

\[
\Phi(\theta) = \mathbb{E} \left[ e^{i\theta^\top X} \right] = \Phi_X(\theta), \quad \theta \in \mathbb{R}^p.
\]

### 8.3 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 8.4 given next.

**Theorem 8.3.1** Consider a rv \( X : \Omega :\to \mathbb{R}^p \) with characteristic function \( \Phi_X : \mathbb{R}^p \to \mathbb{C} \) given by (8.1). It satisfies the following properties:

(i) **Boundedness:** We have

\[
|\Phi_X(\theta)| \leq \Phi_X(0) = 1, \quad \theta \in \mathbb{R}^p.
\]  

(ii) **Uniform continuity on** \( \mathbb{R}^p \): We have

\[
\lim_{\delta \to 0} \sup_{\theta} (|\Phi_X(\theta + \delta) - \Phi_X(\theta)|, \theta \in \mathbb{R}^p) = 0.
\]

(iii) **Positive semi-definiteness:** For every \( n = 1, 2, \ldots \), we have

\[
\sum_{k=1}^{n} \sum_{\ell=1}^{n} \Phi_X(\theta_k - \theta_\ell)z_kz_\ell^* \geq 0
\]

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

\[
\Phi_X(-\theta) = \Phi_X(\theta)^*, \quad \theta \in \mathbb{R}^p.
\]

Much of the discussion makes use of the elementary relation

\[
e^{i\theta x} - 1 = \int_0^x i\theta e^{i\theta s} \, ds, \quad x, \theta \in \mathbb{R}
\]

so that the bounds

\[
|e^{i\theta x} - 1| \leq \int_0^x |i\theta e^{i\theta s}| \, ds \leq |\theta|x
\]
hold.\footnote{With $ab$ in $\mathbb{R}$, we have $|a + ib| = \sqrt{a^2 + b^2} \leq |a| + |b|$.}

**Proof.** (i) It is plain that $\Phi_X(0) = 1$. Next,  
\[ |\Phi_X(\theta)| \leq \mathbb{E} \left[ |e^{i \theta^T X}| \right] = 1, \quad \theta \in \mathbb{R}^p. \]

(ii) Fix $\theta$ and $\delta$ 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  
\[ |\Phi_X(\theta + \delta) - \Phi_X(\theta)| = \mathbb{E} \left[ |e^{i(\theta + \delta)^T X} - e^{i \theta^T X}| \right] \leq \mathbb{E} \left[ \left| \left( e^{i \delta^T X} - 1 \right) e^{i \theta^T X} \right| \right] = \mathbb{E} \left[ \left| e^{i \delta^T X} - 1 \right| \right], \]

so that

\[ (8.16) \quad \sup_{\theta \in \mathbb{R}^p} (|\Phi_X(\theta + \delta) - \Phi_X(\theta)|) \leq \mathbb{E} \left[ \left| e^{i \delta^T X} - 1 \right| \right]. \]

Uniform continuity follows if we can show that  
\[ \lim_{\delta \to 0} \mathbb{E} \left[ \left| e^{i \delta^T X} - 1 \right| \right] = 1. \]

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  
\[ \sum_{k=1}^{n} \sum_{\ell=1}^{n} \Phi_X(\theta_k - \theta_\ell) z_k z_\ell^* \]
\[ = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \mathbb{E} \left[ e^{i(\theta_k - \theta_\ell)^T X} \right] z_k z_\ell^* \]
\[ = \mathbb{E} \left[ \sum_{k=1}^{n} \sum_{\ell=1}^{n} e^{i(\theta_k - \theta_\ell)^T X} z_k z_\ell^* \right] \]
\[ \sum_{k=1}^{n} \sum_{\ell=1}^{n} e^{j\theta_k X} e^{-\theta_\ell X} z_k z_\ell^* \]
\[ = \mathbb{E} \left[ \left( \sum_{k=1}^{n} e^{j\theta_k X} z_k \right) \left( \sum_{\ell=1}^{n} e^{j\theta_\ell X} z_\ell \right)^* \right] \]
\[ = \mathbb{E} \left[ \left| \sum_{k=1}^{n} e^{j\theta_k X} z_k \right|^2 \right] \geq 0. \]

(8.17)

(iv) Fix \( \theta \) in \( \mathbb{R}^p \). We note that
\[
\Phi_X(\theta) = \mathbb{E} \left[ e^{-\theta^t X} \right]
\]
\[
= \mathbb{E} \left[ \cos(-\theta^t X) \right] + i \mathbb{E} \left[ \sin(-\theta^t X) \right]
\]
\[
= \mathbb{E} \left[ \cos(\theta^t X) \right] - i \mathbb{E} \left[ \sin(\theta^t X) \right]
\]
\[
= (\mathbb{E} \left[ \cos(\theta^t X) \right] + i \mathbb{E} \left[ \sin(\theta^t X) \right])^*
\]
\[ = \Phi_X(\theta)^* \]

(8.18)
as desired.

### 8.4 Bochner’s Theorem

Interestingly enough the first three properties given in Theorem 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 8.4.1** 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 \( \theta \) 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 8.4.2** 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 \( \Phi(0) = 1 \).
8.5 Examples

Bernoulli rvs

\( \Phi_X(\theta) = pe^{i\theta} + (1 - p), \quad \theta \in \mathbb{R} \)  
(8.19)

Binomial rvs

\[
\Phi_X(\theta) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} e^{ik\theta} \\
= \sum_{k=0}^{n} \binom{n}{k} (e^{i\theta} p)^k (1 - p)^{n-k} \\
= \sum_{k=0}^{n} \binom{n}{k} \left( e^{i\theta} p \right)^k (1 - p)^{n-k} \\
= \left( 1 - p + pe^{i\theta} \right), \quad \theta \in \mathbb{R} \]  
(8.20)

Poisson rvs

\[
\Phi_X(\theta) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{ik\theta} \\
= \left( \sum_{k=0}^{\infty} \frac{(\lambda e^{i\theta})^k}{k!} \right) e^{-\lambda} \\
= e^{-\lambda} e^{\lambda e^{i\theta}} \\
= e^{-\lambda(1-e^{i\theta})}, \quad \theta \in \mathbb{R} \]  
(8.21)

Geometric rvs

\[
\Phi_X(\theta) = \sum_{k=0}^{\infty} p(1 - p)^k e^{ik\theta} \\
= \sum_{k=0}^{\infty} p \left( (1 - p)^k e^{i\theta} \right)^k \\
= \frac{p}{1 - (1 - p)e^{i\theta}}, \quad \theta \in \mathbb{R} \]  
(8.22)

Exponential rvs

\[
\Phi_X(\theta) = \int_{0}^{\infty} \lambda e^{-\lambda x} e^{i\theta x} dx 
\]
\[ \lambda \int_{0}^{\infty} e^{(i\theta - \lambda)x} \, dx = \frac{\lambda}{i\theta - \lambda} \int_{0}^{\infty} (i\theta - \lambda) e^{(i\theta - \lambda)x} \, dx \]
\[ = \frac{\lambda}{i\theta - \lambda} \cdot \left[ e^{(i\theta - \lambda)x} \right]_{0}^{\infty} \]
\[ = \frac{\lambda}{\lambda - i\theta}, \quad \theta \in \mathbb{R} \]  
(8.23)

as we note that
\[ \lim_{x \to \infty} e^{(i\theta - \lambda)x} = 0. \]

### 8.6 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 8.6.1

If the rvs \( X_1, \ldots, X_k \) are mutually independent, then

\[ \Phi_X(\theta) = \prod_{\ell=1}^{k} \Phi_{X_\ell}(\theta_\ell), \quad \theta_\ell \in \mathbb{R}^{p_\ell}, \quad \ell = 1, \ldots, k \]  
(8.24)

with

\[ \theta \equiv \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} \]

Conversely, if (8.24) holds on \( \mathbb{R}^{p} \), then the rvs \( X_1, \ldots, X_k \) are mutually independent.
**Proof.** Fix $\theta$ in $\mathbb{R}^p$. Noting that

$$\theta^t X = \sum_{\ell=1}^{k} \theta_{\ell}^t X_{\ell},$$

we get

$$E \left[ e^{i\theta^t X} \right] = E \left[ e^{i \sum_{\ell=1}^{k} \theta_{\ell}^t X_{\ell}} \right] = E \left[ \prod_{\ell=1}^{k} e^{i\theta_{\ell}^t X_{\ell}} \right] = \prod_{\ell=1}^{k} E \left[ e^{i\theta_{\ell}^t X_{\ell}} \right] \quad (8.25)$$

by independence. The relation (8.24) follows.

Conversely, if (8.24) holds on $\mathbb{R}^p$, then

$$\Phi_S(\theta) = \prod_{\ell=1}^{k} \Phi_{X_{\ell}}(\theta), \quad \theta \in \mathbb{R}^p.$$

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 8.6.2** If the rvs $X_1, \ldots, X_k$ are mutually independent, then

$$\Phi_S(\theta) = \prod_{\ell=1}^{k} \Phi_{X_{\ell}}(\theta), \quad \theta \in \mathbb{R}^p. \quad (8.26)$$

**Proof.** Fix $\theta$ in $\mathbb{R}^p$. This time noting that $\theta^t S = \sum_{\ell=1}^{k} \theta_{\ell}^t X_{\ell}$, we get

$$E \left[ e^{i\theta^t S} \right] = E \left[ e^{i \sum_{\ell=1}^{k} \theta_{\ell}^t X_{\ell}} \right] = E \left[ \prod_{\ell=1}^{k} e^{i\theta_{\ell}^t X_{\ell}} \right] = \prod_{\ell=1}^{k} E \left[ e^{i\theta_{\ell}^t X_{\ell}} \right] \quad (8.27)$$
CHAPTER 8. PROBABILITY DISTRIBUTIONS AND THEIR TRANSFORMS

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 8.6.2 yields

$$\Phi_S(\theta) = \Phi_X(\theta)^k, \quad \theta \in \mathbb{R}^p. \quad (8.28)$$

8.7 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 8.7.1** Fix $x$ and $\theta$ in $\mathbb{R}$. For each $k = 1, 2, \ldots$, the expansion

$$e^{i\theta x} = \sum_{\ell=0}^k \frac{1}{\ell!} (i\theta x)^\ell + R_k(x; \theta) \quad (8.29)$$

holds with the remainder term given by

$$R_k(x; \theta) = (i\theta)^k \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} \left( e^{i\theta t} - 1 \right) dt. \quad (8.30)$$

**Proof.** The proof proceed by induction: Throughout $\theta$ and $x$ in $\mathbb{R}$ are scalars held fixed.

**Basis step** For $k = 1$, we use (8.14) to get

$$e^{i\theta x} - 1 = \int_0^x i\theta e^{i\theta t} dt$$

$$= \int_0^x i\theta \left( e^{i\theta t} - 1 \right) dt + \int_0^x i\theta dt$$

$$= i\theta x + i\theta \int_0^x \left( e^{i\theta t} - 1 \right) dt$$

$$= i\theta x + R_1(x; \theta) \quad (8.31)$$

by direct inspection.
Induction step. Now assume that (8.30)-(8.30) holds for some \( k = 1, 2, \ldots \). It is plain that

\[
\int_0^x (x-t)^{k-1} \frac{e^{i\theta t} - 1}{(k-1)!} dt
\]

\[
= \int_0^x (x-t)^{k-1} \left( \int_0^t i\theta e^{i\theta s} ds \right) dt
\]

\[
= \int_0^x \left( \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} i\theta e^{i\theta s} ds \right) dt
\]

\[
= \int_0^x \left( \int_s^x \frac{(x-t)^{k-1}}{(k-1)!} i\theta e^{i\theta s} ds \right) dt
\]

\[
= \int_0^x \frac{i\theta (x-s)^k}{k!} e^{i\theta s} ds
\]

(8.32)

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

\[
R_k(x; \theta) = (i\theta)^k \int_0^x \frac{(x-t)^{k-1}}{(k-1)!} \left( e^{i\theta t} - 1 \right) dt
\]

\[
= (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} e^{i\theta s} ds
\]

\[
= (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} \left( e^{i\theta s} - 1 \right) ds + (i\theta)^{k+1} \int_0^x \frac{(x-s)^k}{k!} ds
\]

(8.33)

and the proof of the induction step is now completed.

\[\blacksquare\]

8.8 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 (8.1). Fix \( \theta \) in \( \mathbb{R} \). It follows from Theorem 8.7.1 that

\[
e^{i\theta X} - \sum_{\ell=0}^{k} \frac{1}{\ell!} (i\theta X)^\ell = R_k(X; \theta)
\]

Therefore, if the rv \( X \) has a finite moment of order \( k \) for some \( k = 1, 2, \ldots \), the expectation

\[
\mathbb{E}[R_k(X; \theta)]
\]

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

\[
\mathbb{E}[e^{i\theta X}] = \sum_{\ell=0}^{k} \frac{1}{\ell!} (i\theta)^\ell \mathbb{E}[X^\ell] + \mathbb{E}[R_k(X; \theta)]
\]

does hold. This suggests the following result.

**Theorem 8.8.1** Consider a rv \( X : \Omega \to \mathbb{R} \) with characteristic function \( \Phi_X : \mathbb{R} \to \mathbb{C} \) given by (8.1). If \( \mathbb{E}[|X|^n] < \infty \) 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

\[
\frac{d^k}{d\theta^k} \Phi_X(\theta) = \mathbb{E}
\left[
(iX)^k e^{i\theta X}
\right], \quad \theta \in \mathbb{R}.
\]

**Proof.** If \( k = 1 \). Fix \( \theta \) in \( \mathbb{R} \) and for each \( h \neq 0 \) note that

\[
\Phi_X(\theta + h) - \Phi_X(\theta) = \mathbb{E}\left[e^{i\theta X} \left(e^{ihX} - 1\right)\right]
\]

\[
= \mathbb{E}\left[e^{i\theta X} \int_0^X ihe^{iht} dt\right]
\]

so that

\[
\frac{1}{h} (\Phi_X(\theta + h) - \Phi_X(\theta)) = \mathbb{E}\left[e^{i\theta X} \int_0^X ie^{iht} dt\right].
\]

The bound

\[
\left|\int_0^X ie^{iht} dt\right| = \left|\int_0^X ie^{iht} dt\right| \leq |X|
\]
holds uniformly in \( h \neq 0 \), whence
\[
\lim_{h \to 0} \left( e^{i\theta X} \int_0^X i e^{iht} dt \right) = (iX) e^{i\theta X}
\]
by the Bounded Convergence Theorem. We now conclude that
\[
\lim_{h \to 0} \frac{1}{h} \left( \Phi_X(\theta + h) - \Phi_X(\theta) \right) = \lim_{h \to 0} \mathbb{E} \left[ e^{i\theta X} \int_0^X i e^{iht} dt \right] = \mathbb{E} \left[ e^{i\theta X} \int_0^X i e^{iht} dt \right] = \mathbb{E} \left[ (iX) e^{i\theta X} \right]
\]
by the Dominated Convergence Theorem and the conclusion (8.36) 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 \( \ell^{th} \) differentiable with
\[
\frac{d^\ell}{d\theta^\ell} \Phi_X(\theta) = \mathbb{E} \left[ (iX)^\ell e^{i\theta X} \right], \quad \theta \in \mathbb{R}.
\]
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 \( (\ell + 1)^{st} \) differentiable with
\[
\frac{d^{\ell+1}}{d\theta^{\ell+1}} \Phi_X(\theta) = \mathbb{E} \left[ (iX)^{\ell+1} e^{i\theta X} \right], \quad \theta \in \mathbb{R}.
\]
Indeed, for every \( h \neq 0 \), we have
\[
\frac{d^\ell}{d\theta^\ell} \Phi_X(\theta + h) - \frac{d^\ell}{d\theta^\ell} \Phi_X(\theta) = \mathbb{E} \left[ (iX)^\ell \left( e^{i(\theta+h)X} - e^{i\theta X} \right) \right] = \mathbb{E} \left[ (iX)^\ell e^{i\theta X} (e^{ihX} - 1) \right] = \mathbb{E} \left[ (iX)^\ell e^{i\theta X} \int_0^X ihe^{iht} dt \right]
\]
so that
\[
\frac{1}{h} \left( \frac{d^\ell}{d\theta^\ell} \Phi_X(\theta + h) - \frac{d^\ell}{d\theta^\ell} \Phi_X(\theta) \right) = \mathbb{E} \left[ (iX)^\ell e^{i\theta X} \int_0^X i e^{iht} dt \right]
\]
Again we see that
\[
\left| (iX)^\ell e^{i\theta X} \int_0^X i e^{iht} dt \right| \leq |X|^{\ell+1}
\]
uniformly in \( h \neq 0 \) with \( \mathbb{E} \left[ |X|^{\ell+1} \right] < \infty \) by assumption. Invoking the Dominated
Convergence Theorem we conclude that

\[
\lim_{h \to 0} \frac{1}{h} \left( \frac{d^\ell}{d\theta^\ell} \Phi_X(\theta + h) - \frac{d^\ell}{d\theta^\ell} \Phi_X(\theta) \right) = \lim_{h \to 0} \mathbb{E} \left[ (iX)^\ell e^{i\theta X} \int_0^X i e^{iht} dt \right] = \mathbb{E} \left[ (iX)^\ell e^{i\theta X} \lim_{h \to 0} \int_0^X i e^{iht} dt \right]
\]

(8.42)

and this establishes (8.41) holds. This concludes the induction step as we have now shown that (8.40) holds for \( \ell = 1, \ldots, k \).

\[ \square \]

### 8.9 Exercises

**Ex. 8.1** If the rv \( X : \Omega \to \mathbb{R}^p \) is symmetric, then its characteristic function \( \Phi_X \) is
real-valued, i.e., \( \Phi_X(t) \) is an element of \( \mathbb{R} \) for every \( t \) in \( \mathbb{R}^p \).

**Ex. 8.2**

**Ex. 8.3**

**Ex. 8.4**
Chapter 9

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.

9.1 Scalar Gaussian rvs

Definition 9.1.1 With scalars $\mu$ (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 said to be a Gaussian (or normally distributed) rv with parameters $\mu$ and $\sigma^2$ 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

$$
\mathbb{P}[X \leq x] = \int_{-\infty}^{x} f_{\mu,\sigma^2}(t) \, dt, \quad x \in \mathbb{R}
$$

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

(9.1) \quad \int_{\mathbb{R}} tf_{\mu,\sigma^2}(t) \, dt = \mu \quad \text{and} \quad \int_{\mathbb{R}} t^2 f_{\mu,\sigma^2}(t) \, dt = \mu^2 + \sigma^2

so that $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \mu^2 + \sigma^2$, hence $\text{Var}[X] = \sigma^2$. This shows that the parameters $\mu$ and $\sigma^2$ are the mean and variance, respectively, of the rv $X$. As a result it is customary to refer to the rv $X$ in Definition 9.1.1 as a Gaussian rv with mean $\mu$ and variance $\sigma^2$, written $X \sim N(\mu, \sigma^2)$.
CHAPTER 9. GAUSSIAN RANDOM VARIABLES

If \( X \sim N(\mu, \sigma^2) \) it can be shown that

\[
\mathbb{E} \left[ e^{i\theta X} \right] = e^{i\theta \mu - \frac{\sigma^2}{2} \theta^2}, \quad \theta \in \mathbb{R}.
\]

This fact is established in Section 9.12, and allows us to give the following definition which is equivalent to Definition 9.1.1 and which covers both cases.

**Definition 9.1.2** A rv \( X : \Omega \to \mathbb{R} \) (defined on some probability triple \((\Omega, \mathcal{F}, \mathbb{P})\)) is said to be a Gaussian rv with mean \( \mu \) (in \( \mathbb{R} \)) and variance \( \sigma^2 > 0 \) if its characteristic function is given by

\[
\mathbb{E} \left[ e^{i\theta X} \right] = e^{i\theta \mu - \frac{\sigma^2}{2} \theta^2}, \quad \theta \in \mathbb{R}.
\]

The relations (9.1) can also be established by differentiating the expression (9.3) and using Theorem 8.8.1. It is a simple matter to check that if \( X \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \), 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.

### 9.2 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

\[
\mathbb{P} \left[ U \leq x \right] = \Phi(x) \equiv \int_{-\infty}^{x} \phi(t)dt, \quad x \in \mathbb{R}
\]

with probability density function \( \phi : \mathbb{R} \to \mathbb{R}_+ \) given by

\[
\phi(t) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \quad t \in \mathbb{R}.
\]

As should be clear from earlier comments, for any Gaussian rv \( X \) with mean \( \mu \) and variance \( \sigma^2 \), it holds that \( X =_{st} \mu + \sigma U \), so that

\[
\mathbb{P} \left[ X \leq x \right] = \mathbb{P} \left[ \sigma^{-1}(X - \mu) \leq \sigma^{-1}(x - \mu) \right] \\
= \mathbb{P} \left[ U \leq \sigma^{-1}(x - \mu) \right] \\
= \Phi(\sigma^{-1}(x - \mu)), \quad x \in \mathbb{R}.
\]
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 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 \( \Phi \) is therefore fully determined by the complementary probability distribution function of \( U \) on \([0, \infty)\), namely

\[ Q(x) \equiv 1 - \Phi(x) = \mathbb{P}[U > x], \quad x \geq 0. \]

(9.6)

The evaluation of the so-called \( Q \)-function is given in Section 9.10 to gather with some of its properties.

\[ 9.3 \quad \text{A little Linear Algebra} \]

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 9.3.1** 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

\[ u^t Ru \geq 0, \quad u \in \mathbb{R}^p \]

(iii) positive definite if it is positive semi-definite and the condition \( u^t Ru = 0 \) implies \( u = (0, \ldots, 0)^t \).

The facts given next concern the eigenvalues and eigenvectors of symmetric matrices, and are well known:
Theorem 9.3.1  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 $\lambda_k$ in that

$$Ru_k = \lambda_k u_k.$$ 

(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

$$T \equiv \left( \begin{array}{cccc} u_1 & u_2 & \cdots & u_p \end{array} \right).$$

The transpose $T^t$ of $T$ is given by

$$T^t = \left( \begin{array}{c} u_1^t \\ u_2^t \\ \vdots \\ u_p^t \end{array} \right).$$

From Theorem 9.3.1 we conclude that

$$RT = \left( \begin{array}{cccc} \lambda_1 u_1 & \lambda_2 u_2 & \cdots & \lambda_p u_p \end{array} \right)$$

and

$$T^t RT = \left( \begin{array}{c} u_1^t \\ u_2^t \\ \vdots \\ u_p^t \end{array} \right) \left( \begin{array}{cccc} \lambda_1 u_1 & \lambda_2 u_2 & \cdots & \lambda_p u_p \\ \lambda_1 u_1^t u_1 & \lambda_2 u_1^t u_2 & \cdots & \lambda_p u_1^t u_p \\ \lambda_1 u_2^t u_1 & \lambda_2 u_2^t u_2 & \cdots & \lambda_p u_2^t u_p \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 u_p^t u_1 & \lambda_2 u_p^t u_2 & \cdots & \lambda_p u_p^t u_p \end{array} \right)$$

$$= \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

(9.7)
9.4 Gaussian random vectors

where \( \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \) is the diagonal matrix whose diagonal elements are \( \lambda_1, \lambda_2, \ldots, \lambda_p \). A similar line of reasoning also shows that

\[
T^t T = \begin{pmatrix}
u_1^t \\
u_2^t \\
\vdots \\
u_p^t 
\end{pmatrix}
\begin{pmatrix}
u_1 & \nu_2 & \cdots & \nu_p 
\end{pmatrix} = I_p
\]

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 \( T T^{-1} = T^{-1} T = I_p \) it follows that

\[
T T^t = T^t T = I_p.
\]

The relation \( T^t R T = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \) yields

\[
R = T (T^t R T) T^t = T (\text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)) T^t.
\]

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

\[
R = \left(T \text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})\right) \cdot \left(\text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})\right) T^t
\]

\[
= \left(T \text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})\right) \cdot \left(T \text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})\right)^t.
\]

The \( p \times p \) matrix

\[
B \equiv T \text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})
\]

has the property that \( R = BB^t \), and is known as the square root of the positive semi-definite symmetric matrix \( R \).

9.4 Gaussian random vectors

There are several equivalent ways to define multi-dimensional Gaussian rvs. Throughout, let \( \mu \) denote a vector in \( \mathbb{R}^p \) and let \( \Sigma \) be a \( p \times p \) symmetric and positive semi-definite matrix, thus \( \Sigma^t = \Sigma \) and \( \theta^t \Sigma \theta \geq 0 \) for all \( \theta \) in \( \mathbb{R}^p \).

A definition via characteristic functions The most convenient definition is given in terms of characteristic functions.
CHAPTER 9. GAUSSIAN RANDOM VARIABLES

Definition 9.4.1 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 $\mu$ and covariance matrix $\Sigma$ if its characteristic function is given by

$$
E \left[ e^{i\theta^t X} \right] = e^{i\theta^t \mu - \frac{1}{2} \theta^t \Sigma \theta}, \quad \theta \in \mathbb{R}^p.
$$

(9.8)

We shall write $X \sim N(\mu, \Sigma)$.

For the right-hand side of (9.8) to be a characteristic function we must have

$$
\left| E \left[ e^{i\theta^t X} \right] \right| = e^{-\frac{1}{2} \theta^t \Sigma \theta} \leq 1, \quad \theta \in \mathbb{R}^p.
$$

This implies $\theta^t \Sigma \theta \geq 0$ for each $\theta$ in $\mathbb{R}^p$, making it necessary for the $p \times p$ matrix $\Sigma$ to be a positive semi-definite matrix.

Next, fix $\theta$ in $\mathbb{R}^p$ and use (9.8) with $a\theta$ where $a$ ranges in $\mathbb{R}$. It follows that

$$
E \left[ e^{ia\theta^t X} \right] = e^{ia\theta^t \mu - \frac{a^2}{2} \theta^t \Sigma \theta}, \quad a \in \mathbb{R}
$$

(9.9)

and by virtue of Definition 9.1.2 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 $\theta^t \mu = E \left[ \theta^t X \right] = \theta^t E \left[ X \right]$ and $\theta^t \Sigma \theta = \text{Var} \left[ \theta^t X \right] = \theta^t \text{Cov} \left[ X \right] \theta$. As these equalities hold for all $\theta$ in $\mathbb{R}^p$ we conclude that $\mu = E \left[ X \right]$ and $\Sigma = \text{Cov} \left[ X \right]$. In other words, the parameters $\mu$ and $\Sigma$ indeed have the interpretation of mean and covariance for the rv $X$. The latter conclusion also shows that the matrix $\Sigma$ appearing in (9.8) 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 9.4.2 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 $p \times p$ matrix $B$ and i.i.d. standard Gaussian rvs $U_1, \ldots, U_d$ (defined on $(\Omega, \mathcal{F}, \mathbb{P})$) such that

$$
X =_{st} b + B \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}.
$$

(9.10)
9.5. EQUIVALENCE OF THE TWO DEFINITIONS

By linearity of expectations it is plain from (9.10) that

\[ E[X] = E \left[ b + B \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right) \right] = b + B \left( \begin{array}{c} E[U_1] \\ \vdots \\ E[U_d] \end{array} \right) = b \]

and

\[ E \left[ (X - b)(X - b)^t \right] = E \left[ B \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right) \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right)^t \right] \\
= B E \left[ \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right) \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right)^t \right] B^t \\
= B I_d B^t \\
= B B^t. \quad (9.11) \]

In short we have shown that if the rv \( X : \Omega \to \mathbb{R}^p \) is Gaussian according to Definition 9.4.2, then

\[ E[X] = b \quad \text{and} \quad \text{Cov}[X] = BB^t. \]

9.5 Equivalence of the two definitions

We now discuss the equivalence of these two definitions.

**Definition 9.4.2 implies Definition 9.4.1** Next, pick \( \theta \) in \( \mathbb{R}^p \). We note that

\[ \theta^t (X - b) = \theta^t B \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right) = (B^t \theta)^t \left( \begin{array}{c} U_1 \\ \vdots \\ U_d \end{array} \right) = \sum_{k=1}^d (B^t \theta)_k U_k \]

where for each \( k = 1, \ldots, p \), \( (B^t \theta)_k \) denotes the \( k^{th} \) component of the vector \( B^t \theta \) in \( \mathbb{R}^d \). It follows that

\[ E \left[ e^{i\theta^t(X-b)} \right] = E \left[ e^{i \sum_{k=1}^d (B^t \theta)_k U_k} \right] \\
= E \left[ \prod_{k=1}^d e^{i(B^t \theta)_k U_k} \right] \\
= \prod_{k=1}^d E \left[ e^{i(B^t \theta)_k U_k} \right] \quad \text{[By independence]} \]
with
\[ \mathbb{E} \left[ e^{i(B^t \theta)_k U_k} \right] = e^{-\frac{1}{2} |(B^t \theta)_k|^2} \quad \text{[Since } U_k \sim \mathcal{N}(0, 1) \text{ for each } k = 1, \ldots, d]. \]

Collecting terms we conclude that
\[
\mathbb{E} \left[ e^{i \theta^t (X-b)} \right] = \prod_{k=1}^{d} e^{-\frac{1}{2} |(B^t \theta)_k|^2}
= e^{-\frac{1}{2} \sum_{k=1}^{d} |(B^t \theta)_k|^2}
= e^{-\frac{1}{2} \theta^t BB^t \theta}
\]
(9.12)
as we note that
\[ \sum_{k=1}^{d} |(B^t \theta)_k|^2 = \theta^t BB^t \theta. \]

We conclude that
\[
\mathbb{E} \left[ e^{i \theta^t X} \right] = e^{i \theta^t b} \mathbb{E} \left[ e^{i \theta^t (X-b)} \right]
= e^{i \theta^t b} e^{-\frac{1}{2} \theta^t BB^t \theta}, \quad \theta \in \mathbb{R}^p
\]
(9.13)
and \( X \sim \mathcal{N}(b, BB^t) \) according to Definition 9.4.1.

On the basis of this discussion the reader might wonder whether any \( p \times p \) matrix \( \Sigma \) 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 9.3: There we showed that for any \( p \times p \) matrix \( \Sigma \) 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 \( \Sigma \). ■

**Definition 9.4.1 implies Definition 9.4.2** Consider a rv \( X : \Omega \rightarrow \mathbb{R}^p \) defined on some probability triple \((\Omega, \mathcal{F}, \mathbb{P})\) which is a Gaussian rv with mean vector \( \mu \) and covariance matrix \( \Sigma \) according to Definition 9.4.1. The matrix \( \Sigma \) being symmetric and positive semi-definite, there exists a \( p \times p \) matrix \( B \) such that \( \Sigma = BB^t \). Consider the rv \( X^* : \Omega \rightarrow \mathbb{R}^p \) given by
\[
X^* \equiv \mu + B \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix}
\]
9.6. EXISTENCE OF A DENSITY

where $U_1, \ldots, U_p$ are i.i.d. standard Gaussian rvs. As shown earlier in this section, $E[X^*] = \mu$ and $\text{Cov}[X^*] = BB^t = \Sigma$, while

$$E\left[e^{i\theta^t X^*}\right] = e^{i\theta^t \mu - \frac{1}{2} \theta^t \Sigma \theta}, \quad \theta \in \mathbb{R}^p$$

Therefore, $X^*$ and $X$ have identical characteristic functions, hence they have the same probability distribution functions and we can write $X =_{st} X^*$, just another way to say that $X$ is Gaussian according to Definition 9.4.2. \hfill \Box

9.6 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 $X : \Omega \rightarrow \mathbb{R}^p$ with mean vector $\mu$ and covariance matrix $\Sigma$. The kernel $\text{Ker}(\Sigma)$ of its covariance matrix $\Sigma$, also known as its null space. It is the linear subspace of $\mathbb{R}^p$ given by

$$\text{Ker}(\Sigma) := \{x \in \mathbb{R}^p : \Sigma x = (0, \ldots, o)^t\}.$$ 

Observe that $\theta^t \Sigma \theta = 0$ if and only if $\theta$ belongs to $\text{Ker}(\Sigma)$, in which case (9.8) yields

$$E\left[e^{i\theta^t (X - \mu)}\right] = 1$$

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 $\text{Ker}(\Sigma)$.

To proceed, assume that the covariance matrix $\Sigma$ 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 $\Sigma$ is positive definite or not. Note that the positive definiteness of $\Sigma$, i.e., $\theta^t \Sigma \theta = 0$ necessarily implies $\theta = 0_d$, is equivalent to the condition $\text{Ker}(\Sigma) = \{(0, \ldots, o)^t\}$.

If the $d \times p$ matrix $\Sigma$ is not positive definite, hence only positive semi-definite, then the mass of the rv $X - \mu$ is concentrated on the orthogonal space $\text{Ker}(\Sigma)^\perp$ of $\text{Ker}(\Sigma)$, whence the distribution of $X$ has its support on the linear manifold $\mu + \text{Ker}(\Sigma)^\perp$ and is singular with respect to Lebesgue measure.

On the other hand, if the $d \times p$ matrix $\Sigma$ is positive definite, then the matrix $\Sigma$ is invertible, $\det(\Sigma) \neq 0$ and the Gaussian rv $X$ with mean vector $\mu$ and covariance
matrix $\Sigma$ admits a probability density function $f : \mathbb{R}^p \rightarrow \mathbb{R}_+$ given by
\[
f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu)}, \quad x \in \mathbb{R}^p.
\]

9.7 Linear transformations

The following result is very useful in many contexts, and shows that linear transformations preserve the Gaussian character:

**Lemma 9.7.1** let $\nu$ 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 $\mu$ and covariance matrix $\Sigma$, 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 $A\Sigma A^\top$.

**Proof.** First, by linearity we note that
\[
E[Y] = E[\nu + AX] = \nu + A\mu
\]
so that
\[
\text{Cov}[Y] = E[A(X - \mu)(A(X - \mu))^\top] = A\Sigma A^\top.
\]
(9.14)

Consequently, the $\mathbb{R}^q$-valued rv $Y$ has mean vector $\nu + A\mu$ and covariance matrix $A\Sigma A^\top$.

Pick $\alpha$ arbitrary in $\mathbb{R}^q$. We have
\[
E[e^{i\alpha^\top Y}] = E[e^{i\alpha^\top (\nu + AX)}] = e^{i\alpha^\top \nu} \mathbb{E}[e^{i\alpha^\top AX}] = e^{i\alpha^\top \nu} e^{-\frac{1}{2} (A^\alpha)^\top \Sigma (A^\alpha)} = e^{i\alpha^\top \nu} e^{-\frac{1}{2} \alpha^\top A \Sigma A^\top \alpha}
\]
(9.15)
as required. \(\blacksquare\)
This result can also be established through the evaluation of the characteristic function of the rv $Y$. As an immediate consequence of Lemma 9.7.1 we get

**Corollary 9.7.1** Consider a Gaussian rv $\mathbb{R}^p$-valued rv $X$ with mean vector $\mu$ and covariance matrix $\Sigma$. 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 $(\Sigma_{ij}, \ i, j \in I)$.

### 9.8 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 $(\Omega, \mathcal{F}, P)$, where for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ has mean vector $\mu_\ell$ and covariance matrix $\Sigma_\ell$. With $p = p_1 + \ldots + p_k$, let $X$ denote the $\mathbb{R}^p$-valued rv obtained by concatenating $X_1, \ldots, X_k$, namely

\[
X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.
\]  

(9.16)

Its mean vector $\mu$ is simply

\[
\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}
\]  

(9.17)

while its covariance matrix $\Sigma$ can be written in block form as

\[
\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{1,2} & \ldots & \Sigma_{1,k} \\ \Sigma_{2,1} & \Sigma_2 & \ldots & \Sigma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{k,1} & \Sigma_{k,2} & \ldots & \Sigma_k \end{pmatrix}
\]  

(9.18)

with the notation

\[
\Sigma_{i,j} := \text{Cov}[X_i, X_j] \quad i, j = 1, \ldots, k.
\]

**Lemma 9.8.1** With the notation above, assume the rv $X : \Omega \to \mathbb{R}^p$ to be a Gaussian rv with mean vector $\mu$ and covariance matrix $\Sigma$. Then, for each $\ell = 1, \ldots, k,$
the rv $X_\ell$ is a Gaussian rv with mean vector $\mu_\ell$ and covariance matrix $\Sigma_\ell$. Moreover, the rvs $X_1, \ldots, X_k$ are mutually independent Gaussian rvs if and only they are uncorrelated, i.e.,

$$\Sigma_{i,j} = \delta(i,j)\Sigma_j, \quad i,j = 1, \ldots, k.$$  \hspace{1cm} (9.19)

The first part of Lemma 9.8.1 is a simple rewrite of Corollary 9.7.1. Sometimes we refer to the fact that the rv $X$ is Gaussian by saying that the rvs $X_1, \ldots, X_r$ are jointly Gaussian. A converse to Lemma 9.8.1 is available:

**Lemma 9.8.2** Assume that for each $\ell = 1, \ldots, k$, the rv $X_\ell : \Omega \to \mathbb{R}^{p_\ell}$ is a Gaussian rv with mean vector $\mu_\ell$ and covariance matrix $\Sigma_\ell$. If the rvs $X_1, \ldots, X_k$ are mutually independent, then the rv $X : \Omega \to \mathbb{R}^p$ is a Gaussian rv with mean vector $\mu$ and covariance matrix $\Sigma$ as given by (9.18) with (9.19).

It might be tempting to conclude that the Gaussian character of each of the rvs $X_1, \ldots, X_r$ 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 $X_1, \ldots, X_k$ without further assumptions.

**Counterexample 9.8.1**

### 9.9 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 9.9.1** There always exists a $p \times q$ matrix $A^*$ such that the rvs $V = X - \mathbb{E}[X] - A^*(Y - \mathbb{E}[Y])$ and $Y$ are uncorrelated. This matrix is any solution of the matrix equation

$$\text{Cov}[X,Y] = A \text{Cov}[Y], \quad p \times q \text{ matrix } A.$$  \hspace{1cm} (9.20)

When $\text{Cov}[Y]$ is invertible, then the matrix $A$ is unique and is given by $A^* = \text{Cov}[X,Y] \text{Cov}[Y]^{-1}$. 
Proof. For any $p \times q$ matrix $A$, define the rv $V_A : \Omega \rightarrow \mathbb{R}^p$ by

$$V_A \equiv X - \mathbb{E}[X] - A(Y - \mathbb{E}[Y]).$$

Note that

$$\text{Cov}[V_A, Y] = \text{Cov}[X - \mathbb{E}[X] - A(Y - \mathbb{E}[Y]), Y] = \text{Cov}[X - \mathbb{E}[X], Y] - \text{Cov}[A(Y - \mathbb{E}[Y]), Y]$$

(9.21)

$$= \text{Cov}[X, Y] - A \text{Cov}[Y].$$

The condition that the rvs $V_A$ and $Y$ are uncorrelated reads $\text{Cov}[V_A, Y] = O_{p \times q}$, or equivalently, (9.25). If $\text{Cov}[Y]$ is invertible, then clearly there is only one solution to this matrix equation (in $A$), and it is given by $A^* = \text{Cov}[X, Y] \text{Cov}[Y]^{-1}$.

If $\text{Cov}[Y]$ is not invertible, then

Some important consequences flow from Lemma 9.9.1

Lemma 9.9.2 Assume the rv $Z : \Omega \rightarrow \mathbb{R}^{p+q}$ to be a Gaussian rv. With the notation of Lemma 9.9.1, the rvs $V = X - \mathbb{E}[X] - A^*(Y - \mathbb{E}[Y])$ and $Y$ are independent, each of which is Gaussian.

Proof. Now consider the rv $W : \Omega \rightarrow \mathbb{R}^{p+q}$ given by

$$W \equiv \left( X - \mathbb{E}[X] - A^*(Y - \mathbb{E}[Y]) \right)_Y$$

We can rewrite it as

$$W = B \left( \begin{array}{c} X \\ Y \end{array} \right) + b = BZ + b$$

where the $(p + q) \times (p + q)$ matrix $B$ and the element $b$ of $\mathbb{R}^{p+q}$ are given by

$$B \equiv \begin{pmatrix} I_p & -A^* \\ O_{q \times p} & I_q \end{pmatrix} \quad \text{and} \quad b \equiv \begin{pmatrix} \mathbb{E}[X] - A^*\mathbb{E}[Y] \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The Gaussian character of the rv $Z$ implies that the rv $W : \Omega \rightarrow \mathbb{R}^{p+q}$ is also Gaussian by Lemma 9.7.1. Therefore, the rvs $V = X - \mathbb{E}[X] - A^*(Y - \mathbb{E}[Y])$
and \( Y \) being uncorrelated, we can use Lemma 9.8.1 to conclude that they are independent! ■

We shall use these facts and basic properties of conditional expectations to evaluate the conditional expectation of the \( \text{rv} \; X \) given \( Y \).

**Lemma 9.9.3** Assume the \( \text{rv} \; Z : \Omega \to \mathbb{R}^{p+q} \) to be a Gaussian \( \text{rv} \). It holds that

\[
E[X|Y] = E[X] + A^*(Y - E[Y]) \quad \text{a.s.}
\]

where the \( p \times q \) matrix \( A^* \) is any solution of the matrix equation 9.25.

**Proof.** First, by the independence established in Lemma 9.9.2 we have

\[
E[X - E[X] - A^*(Y - E[Y])|Y] = E[X - E[X] - A^*(Y - E[Y])] = (0, \ldots, 0)^T \quad \text{a.s.}
\]

On the other hand, by linearity of conditional expectations we get

\[
\]

Combining these two evaluations we conclude to (9.22). ■

Finally, we are in a position to identify the conditional distribution of the \( \text{rv} \; X \) given \( Y \).

**Proposition 9.9.1** Assume the \( \text{rv} \; Z : \Omega \to \mathbb{R}^{p+q} \) to be a Gaussian \( \text{rv} \). It holds that

\[
E[e^{i\theta^T X}|Y] = e^{i\theta^T E[X|Y]} \cdot e^{-\frac{1}{2} \theta^T (\text{Cov}[X] - A^* \text{Cov}[Y,X]) \theta}, \quad \theta \in \mathbb{R}^p
\]

The conditional distribution of the \( \text{rv} \; X \) given \( Y \) is therefore also Gaussian with (conditional) mean \( E[X|Y] \) and covariance matrix \( \text{Cov}[X] - A^* \text{Cov}[Y,X] \).

**Proof.** Fix \( \theta \) in \( \mathbb{R}^p \). Noting that

\[
X = V + E[X] + A^*(Y - E[Y]) = V + E[X|Y]
\]
where $A^*$ is as before, we get

$$
\mathbb{E} \left[ e^{i\theta t} X \right] = \mathbb{E} \left[ e^{i\theta t} V \cdot e^{i\theta t} \mathbb{E}[X | Y] \right] = \mathbb{E} \left[ e^{i\theta t} V \right] \cdot e^{i\theta t} \mathbb{E}[X | Y] = \mathbb{E} \left[ e^{i\theta t} V \right] \cdot e^{i\theta t} \mathbb{E}[X | Y] \text{ a.s. (9.26)}
$$

where in the last step we used the fact that the rv $V$ is independent of the rv $Y$, a fact established in Lemma 9.9.3.

But, the rv $V$ is a Gaussian rv $\Omega \to \mathbb{R}^p$ as shown in Lemma 9.9.3; its characteristic function is therefore determined by its mean and its covariance: First, we see that

$$
\mathbb{E} [V] = \mathbb{E} [X - \mathbb{E} [X] - A^* (Y - \mathbb{E} [Y])] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}
$$

while

$$
\text{Cov} [V] = \text{Cov} [X - \mathbb{E} [X] - A^* (Y - \mathbb{E} [Y])] = \mathbb{E} \left[ (X - \mathbb{E} [X] - A^* (Y - \mathbb{E} [Y])) (X - \mathbb{E} [X] - A^* (Y - \mathbb{E} [Y]))^t \right] = \text{Cov} [X] - \text{Cov} [X,Y] (A^*)^t - A^* \text{Cov} [Y,X] + A^* \text{Cov} [Y] (A^*)^t
$$

(9.27)

as we note that

$$
A^* \text{Cov} [Y] (A^*)^t = \text{Cov} [X,Y] (A^*)^t.
$$

This is a simple consequence of the equation (9.25) satisfied by $A^*$.

Thus,

$$
\mathbb{E} \left[ e^{i\theta t} V \right] = e^{-\frac{1}{2} \theta^t \text{Cov}[X] - A^* \text{Cov}[Y,X] \theta},
$$

and combining with (9.26) we get (9.25).

A case of particular interest arises when $\text{Cov} [Y]$ is invertible in which case $A^* = \text{Cov} [X,Y] \text{Cov} [Y]^{-1}$, whence

$$
\text{Cov} [V] = \text{Cov} [X] - \text{Cov} [X,Y] \text{Cov} [Y]^{-1} \text{Cov} [Y,X]
$$

9.10 Evaluating $Q(x)$

The complementary distribution function (9.6) 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 \( \text{Erf} : \mathbb{R}_+ \to \mathbb{R} \) and of its complement \( \text{Erfc} : \mathbb{R}_+ \to \mathbb{R} \) defined by

\[
\text{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad x \geq 0
\]

and

\[
\text{Erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt, \quad x \geq 0.
\]

A simple change of variables \((t = \frac{u}{\sqrt{2}})\) in these integrals leads to the relationships

\[
\text{Erf}(x) = 2 \left( \Phi(x\sqrt{2}) - \frac{1}{2} \right) \quad \text{and} \quad \text{Erfc}(x) = 2Q(x\sqrt{2}),
\]

so that

\[
\text{Erf}(x) = 1 - \text{Erfc}(x), \quad x \geq 0.
\]

Conversely, we also have

\[
\Phi(x) = \frac{1}{2} \left( 1 + \text{Erf} \left( \frac{x}{\sqrt{2}} \right) \right) \quad \text{and} \quad Q(x) = \frac{1}{2} \text{Erfc} \left( \frac{x}{\sqrt{2}} \right).
\]

Thus, knowledge of any one of the quantities \( \Phi, Q, \text{Erf} \) or \( \text{Erfc} \) is equivalent to that of the other three quantities. Although the last two quantities do not have a probabilistic interpretation, evaluating \( \text{Erf} \) is computationally more efficient. Indeed, \( \text{Erf}(x) \) 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

\[
\mathbb{P}[U > x] \leq \mathbb{E}[e^{\theta U}] e^{-\theta x} = e^{-\theta x + \frac{\theta^2 x^2}{2}} e^{\frac{(\theta - x)^2}{2}}
\]

\[
(9.30)
\]

where in the last equality we made use of a completion-of-square argument. The best lower bound

\[
Q(x) \leq e^{-\frac{x^2}{2}}, \quad x \geq 0
\]

is achieved upon selecting \( \theta = x \) in (9.30). We refer to the bound (9.31) as a Chernoff bound; it is not very accurate for small \( x > 0 \) since \( \lim_{x \to 0} Q(x) = \frac{1}{2} \) while \( \lim_{x \to 0} e^{-\frac{x^2}{2}} = 1 \).
### Evaluating $Q(x)$ ($x \to \infty$)

The Chernoff bound shows that $Q(x)$ decays to zero for large $x$ at least as fast as $e^{-x^2/2}$. 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 straightforward change of variable yields

$$Q(x) = \int_{x}^{\infty} \phi(t) dt$$

$$= \int_{0}^{\infty} \phi(x + t) dt$$

$$= \phi(x) \int_{0}^{\infty} e^{-xt} e^{-t^2/2} dt. \quad (9.32)$$

With the Taylor series expansion of $e^{-t^2/2}$ 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 (9.32) by terms of the form

$$\sum_{k=0}^{n} \frac{(-1)^k t^{2k}}{2^k k!}, \quad n = 0, 1, \ldots \quad (9.33)$$

To formulate the resulting approximation contained in Proposition 9.10.1 given next, we set

$$Q_n(x) = \phi(x) \int_{0}^{\infty} \left( \sum_{k=0}^{n} \frac{(-1)^k t^{2k}}{2^k k!} \right) e^{-xt} dt, \quad x \geq 0$$

for each $n = 0, 1, \ldots$.

**Proposition 9.10.1** Fix $n = 0, 1, \ldots$. For each $x > 0$ it holds that

$$Q_{2n+1}(x) \leq Q(x) \leq Q_{2n}(x), \quad (9.34)$$

with

$$|Q(x) - Q_n(x)| \leq \frac{(2n)!}{2^n n!} x^{-(2n+1)} \phi(x). \quad (9.35)$$

where

$$Q_n(x) = \phi(x) \sum_{k=0}^{n} \frac{(-1)^k (2k)!}{2^k k!} x^{-(2k+1)}. \quad (9.36)$$
A proof of Proposition 9.10.1 can be found in Section ??.

Upon specializing (9.34) to \( n = 0 \) we get

\[
\frac{e^{-x^2}}{x\sqrt{2\pi}} \left( 1 - \frac{1}{x^2} \right) \leq Q(x) \leq \frac{e^{-x^2}}{x\sqrt{2\pi}}, \quad x > 0
\]

and the asymptotics

\[
Q(x) \sim \frac{e^{-x^2}}{x\sqrt{2\pi}}, \quad (x \to \infty)
\]

follow. Note that the lower bound in (9.37) is meaningful only when \( x \geq 1 \).

### 9.11 Rvs derived from Gaussian rvs

**Rayleigh rvs** A rv \( X \) is said to be a Rayleigh rv with parameter \( \sigma (\sigma > 0) \) if

\[
X = \sqrt{Y^2 + Z^2}
\]

with \( Y \) and \( Z \) independent zero mean Gaussian rvs with variance \( \sigma^2 \). It is easy to check that

\[
P[X > x] = e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0
\]

with corresponding density function

\[
\frac{d}{dx} P[X \leq x] = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0.
\]

It is also well known that the rv \( \Theta \) given by

\[
\Theta := \arctan \left( \frac{Z}{Y} \right)
\]

is uniformly distributed over \([0, 2\pi)\) and independent of the Rayleigh rv \( X \), i.e.,

\[
P[X \leq x, \Theta \leq \theta] = \frac{\theta}{2\pi} \left( 1 - e^{-\frac{\theta^2}{2\sigma^2}} \right), \quad \theta \in [0, 2\pi), \quad x \geq 0.
\]

**Rice rvs** A rv \( X \) is said to be a Rice rv with parameters \( \alpha \) (in \( \mathbb{R} \)) and \( \sigma (\sigma > 0) \) if

\[
X = \sqrt{(\alpha + Y)^2 + Z^2}
\]

with \( Y \) and \( Z \) independent zero mean Gaussian rvs with variance \( \sigma^2 \). It is easy to check that \( X \) admits a probability density function given by

\[
\frac{d}{dx} P[X \leq x] = \frac{x}{\sigma^2} e^{-\frac{x^2 + \alpha^2}{2\sigma^2}} \cdot I_0 \left( \frac{\alpha x}{\sigma^2} \right), \quad x \geq 0.
\]
9.12. A PROOF OF (9.2)

Here,
\[ I_0(x) := \frac{1}{2\pi} \int_0^{2\pi} e^{x \cos t} dt, \quad x \in \mathbb{R} \] (9.46)
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
\[ \chi^2_n = \sum U_1^2 + \ldots + U_n^2 \]
where \( U_1, \ldots, U_n \) are \( n \) i.i.d. standard Gaussian rvs.

9.12 A proof of (9.2)

Assume \( \mu = 0 \) and \( \sigma^2 = 1 \). Fix \( \theta \) in \( \mathbb{R} \). We need to evaluate
\[ \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\theta x} e^{-\frac{x^2}{2}} dx \]

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
\[
\int_{\mathbb{R}} e^{i\theta x} e^{-\frac{x^2}{2}} dx = \int_{\mathbb{R}} \left( \sum_{k=0}^{\infty} \frac{(i\theta x)^k}{k!} \right) e^{-\frac{x^2}{2}} dx \\
= \sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} \int_{\mathbb{R}} x^k e^{-\frac{x^2}{2}} dx \\
= \sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} m_k \]
(9.47)

where we have set
\[ m_k = \int_{\mathbb{R}} x^k e^{-\frac{x^2}{2}} dx, \quad k = 0, 1, \ldots \]

Note that
\[ m_k = 0, \quad k = 1, 3, 5, \ldots \]
by symmetry, so that
\begin{equation}
\int_{\mathbb{R}} e^{i\theta x} e^{-x^2/2} dx = \sum_{\ell=0}^{\infty} \frac{(i\theta)2\ell}{(2\ell)!} m_{2\ell} = \sum_{\ell=0}^{\infty} \frac{(-\theta^2)^{\ell}}{(2\ell)!} m_{2\ell}.
\end{equation}

Therefore, it remains to compute \(m_{2\ell}, \ell = 0, 1, \ldots\).

To that end, fix \(\ell = 0, 1, \ldots\). By integration by parts yields
\begin{equation}
m_{2(\ell+1)} = \int_{\mathbb{R}} x^{2(\ell+1)} e^{-x^2/2} dx
= 2 \int_{0}^{\infty} x^{2(\ell+1)} e^{-x^2/2} dx
= 2 \int_{0}^{\infty} x^{2\ell+1} \left( xe^{-x^2/2} \right) dx
= 2 \int_{0}^{\infty} x^{2\ell+1} \left( -e^{-x^2/2} \right)^t dx
= 2 \left( -x^{2\ell+1} e^{-\frac{x^2}{2}} \right)_0^{\infty} + \int_{0}^{\infty} (2\ell + 1)x^{2\ell} e^{-x^2/2} dx
\end{equation}
\begin{equation}
= 2(2\ell + 1) \int_{0}^{\infty} x^{2\ell} e^{-x^2/2} dx.
\end{equation}

In other words,
\begin{equation}
m_{2(\ell+1)} = (2\ell + 1)m_{2\ell}, \quad \ell = 0, 1, \ldots
\end{equation}

Iterating we get
\begin{equation}
m_{2\ell} = (2\ell - 1)m_{2(\ell-1)}
= (2\ell - 1)(2\ell - 3)m_{2(\ell-2)}
\vdots
\end{equation}
\begin{equation}
= (2\ell - 1)(2\ell - 3)(2\ell - 5) \cdots 5 \cdot 3 \cdot 1 \cdot m_0.
\end{equation}

It follows that
\begin{equation}
m_{2\ell} = \frac{(2\ell)!}{(2\ell)(2(\ell-1))(2(\ell-2)) \cdots (2 \cdot 3)(2 \cdot 2)(2 \cdot 1)} \cdot m_0 = \frac{(2\ell)!}{(2\ell)!} \cdot m_0
\end{equation}
for each \(\ell = 1, 2, \ldots\). Collecting we conclude that
\begin{equation}
\int_{\mathbb{R}} e^{i\theta x} e^{-x^2/2} dx = \sum_{\ell=0}^{\infty} \frac{(-\theta^2)^{\ell}}{(2\ell)!} \cdot \frac{(2\ell)!}{(2\ell)!} \cdot m_0
\end{equation}
\[ \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left( -\frac{\theta^2}{2} \right)^\ell \cdot m_0 \]

(9.51)

\[ = e^{-\frac{\theta^2}{2}} \cdot m_0. \]

The desired conclusion now follows from the fact that \( m_0 = \sqrt{2\pi} \) since
\[ \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx = 1. \]

### 9.13 Exercises

#### Ex. 9.1
Establish the relations (9.1) through direct integration.

#### Ex. 9.2
Derive the relationships between the quantities \( \Phi, Q, \text{Erf} \) or \( \text{Erfc} \) which are given in Section ??.

#### Ex. 9.3
Given the covariance matrix \( \Sigma \), explain why the representation (??)–(??) may not be unique. Give a counterexample.

#### Ex. 9.4
Give a proof for Lemma 9.8.1 and of Lemma 9.8.2.

#### Ex. 9.5
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. 9.6
Derive the probability distribution function (9.40) of a Rayleigh rv with parameter \( \sigma \) (\( \sigma > 0 \)).

#### Ex. 9.7
Show by direct arguments that if \( X \) is a Rayleigh distribution with parameter \( \sigma \), 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. 9.8
Derive the probability distribution function (9.45) of a Rice rv with parameters \( \alpha \) (in \( \mathbb{R} \)) and \( \sigma \) (\( \sigma > 0 \)).

#### Ex. 9.9
Write a program to evaluate \( Q_n(x) \).

#### Ex. 9.10
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
\[ P \left[ S_n > na \right] \sim \frac{e^{-na^2}}{a\sqrt{2\pi n}} \quad (n \to \infty). \]

(9.52)

This asymptotic is known as the Bahadur-Rao correction to the large deviations asymptotics of \( S_n \).
CHAPTER 9. GAUSSIAN RANDOM VARIABLES

Ex. 9.11 Find all the moments $E[U^p]$ $(p = 1, \ldots)$ where $U$ is a zero-mean unit variance Gaussian rv.

Ex. 9.12 Find all the moments $E[U^p]$ $(p = 1, \ldots)$ where $X$ is a $\chi^2_n$-rv with $n$ degrees of freedom.

Ex. 9.13 Consider three rvs mutually independent rvs $\xi_a, \xi_b$ and $U$ defined on some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that (i) The rv $U : \Omega \rightarrow \mathbb{R}$ is a Bernoulli rv with
$$\mathbb{P}[U = 1] = p = 1 - \mathbb{P}[U = 0]$$
for some $p$ in $(0, 1)$; (ii) The rvs $\xi_a, \xi_b : \Omega \rightarrow \mathbb{R}^2$ are two-dimensional zero-mean Gaussian rvs with covariance matrices $R_a$ and $R_b$, respectively, given by
$$R_\star = \begin{pmatrix} 1 & \rho_\star \\ \rho_\star & 1 \end{pmatrix}, \quad \star = a, b$$
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 $\Phi_X : \mathbb{R} \rightarrow \mathbb{C}$ of the rv $X : \Omega \rightarrow \mathbb{R}^2$ given by
$$X = U\xi_a + (1 - U)\xi_b.$$

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. 9.14 The following arises in classical Statistics: Let $X_1, \ldots, X_n$ denote $n$ i.i.d. Gaussian rvs, each with mean $\mu$ and variance $\sigma^2 > 0$. Define the rvs $\bar{X}$ and $Z_1, \ldots, Z_n$ by
$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$
and
$$Z_k = X_k - \bar{X}, \quad k = 1, 2, \ldots, n.$$

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. 9.15 The rvs $X_1, \ldots, X_n$ are jointly Gaussian, e.g., with $X = (X_1, \ldots, X_n)'$, namely $X \sim N(\mu, R)$ for some vector $\mu$ 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

$$A \equiv a'X = \sum_{k=1}^n a_k X_k \quad \text{and} \quad B \equiv b'X = \sum_{k=1}^n b_k X_k.$$ 

a. Compute the characteristic function of the $\mathbb{R}^2$-valued rv $(A, B)'$, namely

$$\varphi(s, t) = \mathbb{E}[e^{i(sA + tB)}], \quad s, t \in \mathbb{R}.$$ 

Carefully explain your calculations!

b. With the help of your answer in Part a derive a necessary and sufficient condition on the parameters $\mu$, $a$, $b$ 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 $X \sim N(\mu, \sigma^2 I_n)$ with $\sigma^2 > 0$?

Ex. 9.16 Consider the bivariate Gaussian rv $(X, Y)'$ with probability density function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(2x^2 + y^2 + 2xy - 22x - 14y + 65)}, \quad (x, y) \in \mathbb{R}^2.$$ 

Evaluate the quantities $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\text{Var}[X]$, $\text{Var}[Y]$ and $\text{Cov}[X, Y]$.

Ex. 9.17 Let $\xi, \eta : \Omega \to \mathbb{R}$ be independent rvs, each of which is distributed according to a standard Gaussian distribution. Define the rv $(\xi^*, \eta^*) : \Omega \to \mathbb{R}^2$ given by

$$\begin{pmatrix} \xi^* \\ \eta^* \end{pmatrix} = \begin{cases} \begin{pmatrix} \xi \\ |\eta| \end{pmatrix} & \text{if } \xi \geq 0 \\ \begin{pmatrix} \xi \\ -|\eta| \end{pmatrix} & \text{if } \xi < 0. \end{cases}$$

Show that rvs $\xi^*$ and $\eta^*$ are standard Gaussian rvs but that the rv $(\xi^*, \eta^*) : \Omega \to \mathbb{R}^2$ is not Gaussian. Contrast with the statement: The rv $(\xi, \eta) : \Omega \to \mathbb{R}^2$ is a jointly Gaussian rv $N(0_2, I_2)$ with $0_2 = (0, 0)'$ and $I_2$ is the identity on $\mathbb{R}^2$. What explain the difference?
Part II

CONVERGENCE
Chapter 10

Convergence of random variables

Basic notions of convergence in $\mathbb{R}$ are reviewed in Appendix ??, and used extensively in earlier Chapters. 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 $(\Omega, \mathcal{F}, P)$. 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 10.6 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 $\{a_n, n = 1, 2, \ldots\}$.

10.1 Almost sure convergence

Almost sure convergence is the mode of convergence that is easiest to understand as it mimics most closely usual convergence.

Definition 10.1.1 The sequence of rvs $\{X_n, n = 1, 2, \ldots\}$ converges almost surely (a.s.) to the rv $X$ if $P[C] = 1$ where $C$ is the event

\begin{equation}
C = \{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\}.
\end{equation}

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 wp 1), in which case we write \( \lim_{n \to \infty} X_n = X \) wp 1. It is easy to see that the convergence set \( C \) is indeed an event in \( F \) since

\[
C = \cap_{k=1}^{\infty} \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} \left[ |X_n - X| \leq \frac{1}{k} \right].
\]

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

\[
B_n(\varepsilon) \equiv \cap_{m \geq n} A_m(\varepsilon)
= \{|X_n - X| \leq \varepsilon, \ m = n, n+1, \ldots \}.
\]  

(10.2)

**Theorem 10.1.1** The sequence of rvs \( \{X_n, \ n = 1, 2, \ldots \} \) converges a.s. to the rv \( X \) if and only if

\[
P[ B_\infty(\varepsilon) ] = 1, \ \varepsilon > 0
\]

with

\[
B_\infty(\varepsilon) = \cup_{n=1}^{\infty} B_n(\varepsilon).
\]

(10.3)

(10.4)

**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 property of \( P \) under monotone limits, we get

\[
P[C] = \lim_{k \to \infty} P[ B_\infty(k^{-1}) ].
\]

(10.5)

This last convergence being monotonically decreasing as \( k \) increases, we conclude that \( P[C] = 1 \) if and only if

\[
P[ B_\infty(k^{-1}) ] = 1, \ \ k = 1, 2, \ldots
\]

The conclusion follows since for every \( \varepsilon > 0 \) there exists a positive integer \( k \) such that \((k+1)^{-1} \leq \varepsilon \leq k^{-1}\) 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 10.1.2** 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

\[
\sum_{n=1}^{\infty} P[|X_n - X| > \varepsilon] < \infty.
\]  

**(10.6)**

**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 part of the Borel-Cantelli Lemma now yields \( P[B_{\infty}(\varepsilon)^c] = 0 \) provided

\[
\sum_{n=1}^{\infty} P[A_n(\varepsilon)^c] < \infty.
\]

This is equivalent to \( P[B_{\infty}(\varepsilon)] = 1 \) provided (10.6) holds, and the proof is completed by invoking Theorem 10.1.1.

The condition (10.6) is sufficient to ensure a.s. convergence, but not necessary. It occurs sufficiently often that it has been given a name.

**Definition 10.1.2** 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

\[
\sum_{n=1}^{\infty} P[|X_n - X| > \varepsilon] < \infty.
\]  

**(10.7)**

By Theorem 10.1.2 we see that complete convergence implies a.s. convergence. However, complete convergence is only a sufficient condition for a.s. convergence, and not a necessary condition. This is confirmed by the next example.

**Counterexample 10.1.1** A.s. convergence does not imply complete convergence. Take \( \Omega = [0, 1], \mathcal{F} = \mathcal{B}(\mathbb{R}) \) and \( \mathbb{P} \) is Lebesgue measure \( \lambda \). Define the rvs \( \{X_n, \ n = 1, 2, \ldots\} \) to be

\[
X_n = \begin{cases} 
1 & \text{if } 0 \leq \omega \leq 1 - \frac{1}{n} \\
0 & \text{if } 1 - \frac{1}{n} < \omega \leq 1
\end{cases}
\]
for every \( n = 1, 2, \ldots \). Fix \( \omega \) 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 \( \varepsilon \) in \((0, 1)\), we get
\[
P[|X_n| > \varepsilon] = \frac{1}{n}, \quad n = 1, 2, \ldots
\]
whence (10.7) fails since \( \sum_{n=1}^{\infty} \frac{1}{n} = \infty \) by the divergence of the harmonic series.

## 10.2 Convergence in probability

### Definition 10.2.1
The sequence of rvs \( \{X_n, \, n = 1, 2, \ldots\} \) converges in probability to the rv \( X \) if for every \( \varepsilon > 0 \), we have
\[
\lim_{n \to \infty} P[|X_n - X| > \varepsilon] = 0.
\]
(10.8)
We shall write \( X_n \xrightarrow{P} n X \).

Convergence in probability admits the following Cauchy criterion.

### Theorem 10.2.1 (Cauchy criterion for convergence in probability)
The sequence of rvs \( \{X_n, \, n = 1, 2, \ldots\} \) converges in probability if and only if for every \( \varepsilon > 0 \), we have
\[
\lim_{n \to \infty} \left( \sup_{m \geq n} P[|X_n - X_m| > \varepsilon] \right) = 0.
\]
(10.9)

A.s. convergence is a stronger notion of convergence than convergence in probability.

### Theorem 10.2.2
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
\[
P[B_n(\varepsilon)] \leq P[A_n(\varepsilon)], \quad n = 1, 2, \ldots
\]
The sets \( \{B_n(\varepsilon), \, n = 1, 2, \ldots\} \) being non-decreasing, we readily conclude that \( \lim_{n \to \infty} P[B_n(\varepsilon)] = P[B_\infty(\varepsilon)] \) with \( B_\infty(\varepsilon) \) defined at (10.4). It is now plain that
\[
P[B_\infty(\varepsilon)] = \lim_{n \to \infty} P[B_n(\varepsilon)] \leq \liminf_{n \to \infty} P[A_n(\varepsilon)].
\]
10.2. CONVERGENCE IN PROBABILITY

By Theorem 10.1.1 the a.s. convergence of the sequence \( \{X_n, n = 1, 2, \ldots\} \) implies \( P[B_\infty(\varepsilon)] = 1 \), and this immediately implies \( \liminf_{n \to \infty} P[A_n(\varepsilon)] = 1 \). Thus, \( \lim_{n \to \infty} P[A_n(\varepsilon)] = 1 \), and the sequence \( \{X_n, n = 1, 2, \ldots\} \) converges in probability.

Here is an example of a sequence which converges in probability but does not converge almost surely:

Counterexample 10.2.1 Convergence in probability does not imply a.s. convergence

Take \( \Omega = [0, 1] \), \( \mathcal{F} = \mathcal{B}(\mathbb{R}) \) and \( P \) is Lebesgue measure \( \lambda \). 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 \( 2^k \leq n < 2^{k+1} \) 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 \( \Omega_b \) of boundary points

\[
\Omega_b = \left\{ m2^{-k}, m = 0, \ldots, 2^k, k = 0, 1, \ldots \right\}
\]

is countable, hence \( P[\Omega_b] = 0 \). With \( \omega \) not in \( \Omega_b \) 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 \( \lim_{n \to \infty} P[|X_n - X| > \varepsilon] = 0 \) for every \( \varepsilon > 0 \) since

\[
P[|X_n - X| > \varepsilon] = \begin{cases} 
P[I_n] & \text{if } 0 < \varepsilon < 1 \\
0 & \text{if } 1 \geq \varepsilon.
\end{cases}
\]

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 10.2.3 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 $N_0 \to N_0$ 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 sequence $\{n_k, k = 1, 2, \ldots\}$ implies $\lim_{k \to \infty} n_k = \infty$.

**Proof.** The assumed convergence in probability of the sequence of rvs $\{X_n, n = 1, 2, \ldots\}$ to the rv $X$ amounts to

$$\lim_{n \to \infty} P[|X - X_n| > \varepsilon] = 0, \quad \varepsilon > 0.$$

Fix $\varepsilon > 0$: For every $\delta > 0$ there exists a positive integer $n^*(\varepsilon, \delta)$ such that

$$P[|X - X_n| > \varepsilon] \leq \delta, \quad n \geq n^*(\varepsilon, \delta).$$

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

$$P[|X - X_n| > k^{-1}] \leq 2^{-k}, \quad n \geq n_k.$$

It is always possible to select $n_k$ as any positive integer satisfying

$$\max(n^*(\varepsilon, \delta), n_{k-1}) < n_k$$

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

$$\sum_{k=1}^{\infty} P[|X_{n_k} - X| > \varepsilon]$$

$$= \sum_{k=1, 2, \ldots, k^{-1} > \varepsilon} P[|X_{n_k} - X| > \varepsilon] + \sum_{k=1, 2, \ldots, k^{-1} \leq \varepsilon} P[|X_{n_k} - X| > \varepsilon]$$

$$\leq k(\varepsilon) + \sum_{k = k(\varepsilon)}^{\infty} P[|X_{n_k} - X| > k^{-1}]$$

$$\leq k(\varepsilon) + \sum_{k = k(\varepsilon)}^{\infty} 2^{-k}$$
and the conclusion \( \sum_{k=1}^{\infty} P [|X_{n_k} - X| > \varepsilon] < \infty \) follows. The desired a.s. convergence of the sequence of rvs \( \{X_{n_k}, \ k = 1, 2, \ldots \} \) is now a consequence of Theorem 10.1.2.

\[ \]

### 10.3 Convergence in the \( r^{th} \) mean

**Definition 10.3.1** 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

\[
E [|X_n|^r] < \infty, \quad n = 1, 2, \ldots
\]

and

\[
\lim_{n \to \infty} E [|X_n - X|^r] = 0. \tag{10.11}
\]

We shall write \( X_n \xrightarrow{L^r} X \).

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 (10.11) that \( E [|X_n - X|^r] < \infty \) for all \( n \) sufficiently large, whence the rv \( X \) necessarily has a finite moment of order \( r \) by virtue of Minkowski’s inequality under (10.10).

Convergence in the \( r^{th} \) mean also admits a Cauchy criterion which is given next.

**Theorem 10.3.1** *(Cauchy criterion for \( r^{th} \) mean convergence)* With \( r \geq 1 \), the sequence of rvs \( \{X_n, \ n = 1, 2, \ldots \} \) converges in the \( r^{th} \) mean if and only if

\[
\lim_{n \to \infty} \left( \sup_{m \geq n} E [|X_n - X_m|^r] \right) = 0. \tag{10.12}
\]

Convergence in the \( r^{th} \) mean becomes more stringent as \( r \) increases.

**Theorem 10.3.2** 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

\[
E [|X_n - X|^s]^{\frac{1}{s}} \leq E [|X_n - X|^r]^{\frac{1}{r}}, \quad n = 1, 2, \ldots
\]
Next, we compare \( r \)-th mean convergence to convergence in probability.

**Theorem 10.3.3** 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

\[
P [|X_n - X| > \varepsilon] \leq \frac{E [|X_n - X|^r]}{\varepsilon^r}, \quad n = 1, 2, \ldots
\]

so that \( \lim_{n \to \infty} P [|X_n - X| > \varepsilon] = 0 \) as soon as \( \lim_{n \to \infty} E [|X_n - X|^r] = 0 \). ■

The converse is more delicate as the next example already illustrates; see also Section 10.5.

**Counterexample 10.3.1** 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,

\[
P [|X_n| > \varepsilon] = n^{-\alpha}, \quad n = 1, 2, \ldots
\]

as soon as \( 0 < \varepsilon \leq 1 \) so that \( X_n \xrightarrow{P} n0 \).

On the other hand, with \( r \geq 1 \), we find

\[
E [|X_n|^r] = 0 \left(1 - n^{-\alpha}\right) + n^{r\beta}n^{-\alpha} = n^{r\beta-\alpha}, \quad n = 1, 2, \ldots
\]

so that

\[
\lim_{n \to \infty} E [|X_n|^r] = \begin{cases} 
0 & \text{if } r\beta < \alpha \\
1 & \text{if } r\beta = \alpha \\
\infty & \text{if } r\beta > \alpha.
\end{cases}
\]

It is now plain that \( X_n \xrightarrow{L^r} 0 \) 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 10.1.2, which allows us to determine a.s. convergence in the presence of convergence in the \( r \)th mean.

**Theorem 10.3.4** 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
\[
\sum_{n=1}^{\infty} \mathbb{E}[|X_n - X|^r] < \infty
\]
holds.

**Proof.** By Markov's inequality, we have
\[
\mathbb{P}[|X_n - X| > \varepsilon] \leq \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r}, \quad n = 1, 2, \ldots
\]
for every \( \varepsilon > 0 \), whence
\[
\sum_{n=1}^{\infty} \mathbb{P}[|X_n - X| > \varepsilon] \leq \frac{1}{\varepsilon^r} \sum_{n=1}^{\infty} \mathbb{E}[|X_n - X|^r],
\]
and the conclusion is immediate by Theorem 10.1.2.

---

### 10.4 Convergence in distribution

For any rv \( X : \Omega \to \mathbb{R} \), its probability distribution function \( F_X : \mathbb{R} \to [0, 1] \) satisfies the following properties: (i) it is non-decreasing; (ii) it has left-limit and is right-continuous at every point; and (iii) \( \lim_{x \to -\infty} F_X(x) = 0 \) and \( \lim_{x \to \infty} F_X(x) = 1 \).

Let \( C(F_X) \) denote the set of points in \( \mathbb{R} \) where \( F_X : \mathbb{R} \to [0, 1] \) is continuous, i.e.,
\[
C(F_X) = \{ x \in \mathbb{R} : F_X(x-) = F_X(x) \}.
\]
The complement \( C(F_X)^c \) of \( C(F_X) \) in \( \mathbb{R} \) consists of the points where \( F_X : \mathbb{R} \to [0, 1] \) is not continuous.

**Theorem 10.4.1** For any rv \( X : \Omega \to \mathbb{R} \), its probability distribution function \( F_X : \mathbb{R} \to [0, 1] \) has the property that \( C(F_X) \) is a countable subset of \( \mathbb{R} \).
**Proof.** For each $n = 1, 2, \ldots$, let $D_n$ denote the collection of points of discontinuity in $C(F_X)^c$ whose discontinuity jump lies in the interval $(\frac{1}{n+1}, \frac{1}{n}]$, i.e.,

$$D_n \equiv \left\{ x \in C(F_X)^c : \frac{1}{n+1} < F_X(x) - F_X(x-) \leq \frac{1}{n} \right\}$$

Noting that

$$|D_n| \cdot \frac{1}{n+1} \leq \sum_{x \in D_n} (F_X(x) - F_X(x-)) \leq 1,$$

it follows that $|D_n| \leq n + 1$. The desired result is now immediate since $C(F_X)^c = \bigcup_{n=1}^{\infty} D_n$.

**Definition 10.4.1** The sequence of rvs $\{X_n, n = 1, 2, \ldots\}$ converges in distribution to the rv $X$ if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x), \quad x \in C(F_X). \tag{10.13}$$

We shall write $X_n \xrightarrow{n} X$ or $X_n \xleftarrow{\mathcal{L}} X$. 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 10.4.2** The sequence of probability distribution functions $\{F_n, n = 1, 2, \ldots\}$ converges in distribution to the probability distribution function $F$ if

$$\lim_{n \to \infty} F_n(x) = F(x), \quad x \in C(F_X). \tag{10.14}$$

We shall write $F_n \xrightarrow{n} F$ or $F_n \xleftarrow{\mathcal{L}} F$.

At this point the reader may wonder as to why the definition of distributional convergence requires the convergence (10.13) only on the set of points of continuity of the limit. This is best seen on the following example.

**Example 10.4.1** 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 \( \omega \) in \( \Omega \). Yet it is easy to check that

\[
\lim_{n \to \infty} F_{X_n}(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1 & \text{if } x \geq 0
\end{cases} \quad \text{and} \quad \lim_{n \to \infty} F_{Y_n}(x) = \begin{cases} 
0 & \text{if } x \leq 0 \\
1 & \text{if } x > 0
\end{cases}
\]

\[\square\]

**Theorem 10.4.2** 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.

**Proof.** Fix \( n = 1, 2, \ldots \) and pick \( x \) in \( \mathbb{R} \). With \( \varepsilon > 0 \), we note that

\[
F_{X_n}(x) = \mathbb{P}[X_n \leq x] = \mathbb{P}[X_n \leq x, X \leq x+\varepsilon] + \mathbb{P}[X_n \leq x, x+\varepsilon < X] 
\leq \mathbb{P}[X \leq x+\varepsilon] + \mathbb{P}[|X_n - X| > \varepsilon] 
= F_X(x+\varepsilon) + \mathbb{P}[|X_n - X| > \varepsilon].
\]

In a similar way, we find

\[
F_X(x-\varepsilon) = \mathbb{P}[X \leq x-\varepsilon] 
\leq \mathbb{P}[X \leq x-\varepsilon, X_n \leq x] + \mathbb{P}[|X_n - X| > \varepsilon] 
= F_{X_n}(x) + \mathbb{P}[|X_n - X| > \varepsilon].
\]

Let \( n \) go to infinity in these inequalities. Under the assumed convergence in probability, we find

\[
\limsup_{n \to \infty} F_{X_n}(x) \leq F_X(x + \varepsilon) \quad \text{and} \quad F_X(x - \varepsilon) \leq \liminf_{n \to \infty} F_{X_n}(x).
\]

Picking \( x \) to be a point of continuity for \( F_X \), we obtain

\[
\limsup_{n \to \infty} F_{X_n}(x) = \lim_{\varepsilon \downarrow 0} \left( \limsup_{n \to \infty} F_{X_n}(x) \right) 
\leq \lim_{\varepsilon \downarrow 0} F_X(x + \varepsilon) 
= F_X(x)
\]
CHAPTER 10. CONVERGENCE OF RANDOM VARIABLES

and

\[ F_X(x) = \lim_{\varepsilon \downarrow 0} F_X(x - \varepsilon) \]
\[ \leq \lim_{\varepsilon \downarrow 0} \left( \liminf_{n \to \infty} F_{X_n}(x) \right) \]
\[ = \liminf_{n \to \infty} F_{X_n}(x) \]

whence \( \liminf_{n \to \infty} F_{X_n}(x) = \limsup_{n \to \infty} F_{X_n}(x) = F_X(x) \). It follows that

\[ \lim_{n \to \infty} F_{X_n}(x) = F_X(x), \quad x \in \mathcal{C}(F_X). \]

Although weak convergence is weaker than convergence in probability, there is one situation where they are equivalent.

**Theorem 10.4.3** 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.** Assume that the sequence of rvs \( \{X_n, \ n = 1, 2, \ldots\} \) converges in distribution to the degenerate rv \( X = c \). Fix \( \varepsilon > 0 \). For every \( n = 1, 2, \ldots \), we observe that

\[ P[|X_n - X| \leq \varepsilon] = P[|X_n - c| \leq \varepsilon] \]
\[ = P[c - \varepsilon \leq X_n \leq c + \varepsilon] \]
\[ = P[X_n \leq c + \varepsilon] - P[X_n < c - \varepsilon] \]
\[ = F_{X_n}(c + \varepsilon) - F_{X_n}((c - \varepsilon) -) \]

(10.15)

so that

\[ P[|X_n - X| > \varepsilon] = 1 - F_{X_n}(c + \varepsilon) + F_{X_n}((c - \varepsilon) -) \]
\[ \leq 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon). \]

Recall that \( F_X(x) = 0 \) (resp. \( F_X(x) = 1 \)) if \( x < c \) (resp. \( c \leq x \)) so that the only point of discontinuity of \( F_X \) is located at \( x = c \). Thus, under the assumed convergence in distribution, we have \( \lim_{n \to \infty} F_{X_n}(c + \varepsilon) = 1 \) and \( \lim_{n \to \infty} F_{X_n}(c - \varepsilon) = 0 \), whence \( \lim_{n \to \infty} P[|X_n - X| > \varepsilon] = 0 \) as desired. \( \blacksquare \)
10.5 Uniform integrability

If a rv $X$ has a finite first moment, we know that

$$
\lim_{B \to \infty} E \left[ 1 \left[ |X| > B \right] |X| \right] = 0.
$$

(10.16)

This is a simple consequence of the Dominated Convergence Theorem (since $Y_B \leq |X|$ where $Y_B = 1 \left[ |X| > B \right] |X|$ for all $B > 0$). Thus, for every $\varepsilon > 0$, there exists $B^*(\varepsilon) > 0$ such that

$$
E \left[ 1 \left[ |X| > B \right] |X| \right] \leq \varepsilon, \quad B \geq B^*(\varepsilon).
$$

(10.17)

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 $B^*(\varepsilon; n) > 0$ such that

$$
E \left[ 1 \left[ |X_n| > B \right] |X_n| \right] \leq \varepsilon, \quad B \geq B^*(\varepsilon; n).
$$

(10.18)

This is a direct consequence of (10.17). However, sometimes it is required that this condition holds uniformly with respect to $n = 1, 2, \ldots$ in that $B^*(\varepsilon; n)$ can be selected independently of $n$. This leads to the following stronger notion of integrability for a sequence of rvs.

**Definition 10.5.1** The collection of rvs $\{X_n, \ n = 1, 2, \ldots\}$ is said to be uniformly integrable if

$$
\lim_{B \to \infty} \left( \sup_{n=1,2,\ldots} E \left[ 1 \left[ |X_n| > B \right] |X_n| \right] \right) = 0.
$$

(10.19)

In other words, for every $\varepsilon > 0$, there exists $B^*(\varepsilon) > 0$ such that

$$
\sup_{n=1,2,\ldots} E \left[ 1 \left[ |X_n| > B \right] |X_n| \right] \leq \varepsilon, \quad B \geq B^*(\varepsilon).
$$

(10.20)

The uniform integrability of the rvs $\{X_n, \ n = 1, 2, \ldots\}$ implies the boundedness condition

$$
\sup_{n=1,2,\ldots} E |X_n| < \infty.
$$

(10.21)

While this condition is not sufficient to imply uniform integrability, a slight strengthening of it will.
Lemma 10.5.1. The collection of rvs \( \{X_n, n = 1, 2, \ldots \} \) is uniformly integrable if there exists \( r > 0 \) such that

\[
\sup_{n=1,2,\ldots} E [|X_n|^{1+r}] < \infty. \tag{10.22}
\]

Proof. Fix \( n = 1, 2, \ldots \) and \( B > 0 \). By the usual monotonicity argument we have

\[
B \cdot P [|X_n| > B] \leq E [1 [|X_n| > B] |X_n|] \leq \frac{1}{B} \cdot E [|X_n|].
\]

Next, applying Hölder’s inequality to the rv \( |X_n| \) and \( 1 [|X_n| > B] \) (with conjugate exponents \( p = 1 + r \) and \( q = \frac{r+1}{r} \)), we get

\[
E [1 [|X_n| > B] |X_n|] \leq E \left[ 1 \left[ |X_n| > B \right] \right]^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}
= E \left[ 1 \left[ |X_n| > B \right] \right]^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}
= P \left[ |X_n| > B \right]^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}
\leq \left( \frac{1}{B} \cdot E [|X_n|] \right)^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}} \leq B^{-\frac{r}{r+1}} (\ldots)
\tag{10.23}
\]

with

\[
\ldots \quad = \ E [|X_n|^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}
\tag{10.24}
= \left( 1 + E \left[ |X_n|^{1+r} \right] \right)^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}.
\]

It follows that

\[
\sup_{n=1,2,\ldots} E [1 [|X_n| > B] |X_n|] \leq C B^{-\frac{r}{r+1}} \tag{10.25}
\]

with finite constant \( C \) given by

\[
C \equiv \sup_{n=1,2,\ldots} \left( 1 + E \left[ |X_n|^{1+r} \right] \right)^{\frac{r+1}{r}} \cdot E \left[ |X_n|^{1+r} \right]^{\frac{1}{1+r}}.
\]
The desired conclusion is now immediate.

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 10.5.1** Consider a collection of rvs \( \{ X, X_n, n = 1, 2, \ldots \} \) such that \( \lim_{n \to \infty} X_n = X \) a.s. (resp. \( X_n \overset{P}{\to} X, X_n \Rightarrow_n X \)) If the collection of rvs \( \{ X_n, n = 1, 2, \ldots \} \) is uniformly integrable, then \( \mathbb{E}[|X|] < \infty \) and

\[
\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X].
\]

**10.6 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
  \[
  \lim_{n \to \infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0, \quad \varepsilon > 0
  \]
- converge in the \( r \)th mean (for some \( r \geq 1 \)) to the rv \( X \) if
  \[
  \lim_{n \to \infty} \mathbb{E}[|X_n - X|^r] = 0.
  \]

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 
\[
\lim_{n \to \infty} \|X_n - X\| = 0 \text{ a.s.}
\]

• converge in probability to the rv $X$ if 
\[
\lim_{n \to \infty} P[\|X_n - X\| > \varepsilon] = 0, \quad \varepsilon > 0
\]

• converge in the $r^{th}$ mean (for some $r \geq 1$) to the rv $X$ if 
\[
\lim_{n \to \infty} E[\|X_n - X\|^r] = 0.
\]

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 $C_{a|b} > 0$ 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 $L_2$-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 10.4 we instead rely on the equivalence given in Theorem 11.4.
Chapter 11

Weak convergence

11.1 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 11.1.1 The sequence of rvs \( \{X_n, n = 1, 2, \ldots\} \) converges weakly 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

\[
(11.1) \quad \Phi(\theta) = \lim_{n \to \infty} \Phi_{X_n}(\theta), \quad \theta \in \mathbb{R}
\]

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 (11.1) may not necessarily define the characteristic function of a rv as can be seen from the following example.

Example 11.1.1 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

\[
(11.2) \quad \Phi_{X_n}(\theta) = \int_{-n}^{n} e^{i\theta x} \frac{1}{2n} \, dx = \begin{cases} 
\frac{\sin(n\theta)}{n} & \text{if } \theta \neq 0 \\
1 & \text{if } \theta = 0,
\end{cases}
\]
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. \[ \blacksquare \]

This difficulty can be remedied with the help of the next result by simply checking continuity at \( \theta = 0 \) for the limit (11.1). This is a consequence of the Bochner-Herglotz Theorem.

**Theorem 11.1.2** 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 \( X_n \Rightarrow n X \).

**Proof.** For each \( n = 1, 2, \ldots \), the function \( \Phi_{X_n} : \mathbb{R} \to \mathbb{C} \) is a characteristic function. Therefore, by Theorem 8.4.1 it is (i) bounded with \( |\Phi_{X_n}(\theta)| \leq \Phi_{X_n}(0) = 1 \) for all \( \theta \) 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 8.4.2 the assumed continuity of \( \Phi \) implies that it is a characteristic function, i.e., there exists a rv \( X \) such that \( \Phi = \Phi_X \). Invoking Theorem 11.1.1 we conclude that \( X_n \Rightarrow n X \). \[ \blacksquare \]

### 11.2 Weak convergence via the Skorokhod representation

Consider a collection \( \{F, F_n, \ n = 1, 2, \ldots\} \) of probability distribution functions on \( \mathbb{R} \).

**Theorem 11.2.1** If the sequence of probability distribution functions \( \{F_n, \ n = 1, 2, \ldots\} \) converges weakly to \( F \), then there exists a probability triple \( (\Omega^*, \mathcal{F}^*, \mathbb{P}^*) \) and a collection of \( \mathbb{R} \)-valued rvs \( \{X^*, X^*_n, \ n = 1, 2, \ldots\} \) all defined on \( \Omega^* \) with the following properties:
11.3. FUNCTIONAL CHARACTERIZATION OF CONVERGENCE IN DISTRIBUTION

(i) We have

\[ F_n(x) = \mathbb{P}[X^*_n \leq x], \quad x \in \mathbb{R}, \quad n = 1, 2, \ldots \]  

and

\[ F(x) = \mathbb{P}[X^* \leq x], \quad x \in \mathbb{R}. \]  

(ii) The rvs \( \{X^*_n, n = 1, 2, \ldots\} \) converges a.s. to \( X^* \) (under \( \mathbb{P}^* \)), i.e.,

\[ \mathbb{P}^* \left( \{\omega^* \in \Omega^*: \lim_{n \to \infty} X^*_n(\omega^*) = X^*(\omega^*)\} \right) = 1 \]

This result also holds in higher dimensions.

11.3 Functional characterization of convergence in distribution

The following equivalent characterizations of distributional convergence have many use.

**Theorem 11.3.1** 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

\[ \lim_{n \to \infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]. \]  

(iii) The characteristic functions converge in the sense that

\[ \lim_{n \to \infty} \Phi_{X_n}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}. \]

**Proof.** It follows from Theorem 11.2.1 that (i) implies the validity of (ii): Indeed, with the notation used in that result, consider the probability triple \( (\Omega^*, \mathcal{F}^*, \mathbb{P}^*) \) and the \( \mathbb{R} \)-valued rvs \( \{X^*, X^*_n, n = 1, 2, \ldots\} \) all defined on \( \Omega^* \) such that

\[ \mathbb{P}[X \leq x] = \mathbb{P}^*[X^* \leq x], \quad x \in \mathbb{R} \]
and

(11.8) \[ P[X_n \leq x] = P^*[X_n^* \leq x] \quad x \in \mathbb{R} \]

with

\[ P^* \left[ \Omega^* : \lim_{n \to \infty} X_n^*(\omega^*) = X^*(\omega^*) \right] = 1. \]

Pick a mapping \( g : \mathbb{R} \to \mathbb{R} \) which is continuous and bounded - Set

\[ B_g \equiv \sup_{x \in \mathbb{R}} |g(x)| < \infty. \]

Obviously,

\[ E[g(X)] = E^*[g^*(X^*)] \quad \text{and} \quad E[g(X_n)] = E^*[g^*(X_n^*)], \quad n = 1, 2, \ldots \]

It is plain that

\[ \lim_{n \to \infty} g(X_n^*) = g(X^*) \quad P^*-a.s. \]

by the continuity of \( g \), with

\[ |g(X_n^*(\omega^*))| \leq B_g, \quad \omega^* \in \Omega^*, \quad n = 1, 2, \ldots \]

Invoking the Dominated Convergence Theorem we readily conclude that

\[ \lim_{n \to \infty} E^*[g^*(X_n^*)] = E^*[g^*(X^*)]. \]

This completes the proof of the validity of (ii). The proof that (ii) implies (i) is omitted.

The equivalence of (i) and (iii) is just Theorem ???. Note that (iii) is a simple consequence of (ii) since for every \( \theta \) 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 is the following continuity result for weak convergence.

**Theorem 11.3.2** Consider the \( \mathbb{R} \)-valued rvs \( \{X, X_n, \ n = 1, 2, \ldots\} \) defined on some probability triple \((\Omega, \mathcal{F}, P)\). 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

\[ h(X_n) \Rightarrow_n h(X). \]
**Proof.** The proof follows by a simple application of Theorem 11.4: 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 11.4 we conclude from the assumed convergence $X_n \Rightarrow_n X$ that

$$\lim_{n \to \infty} \mathbb{E}[g \circ h(X_n)] = \mathbb{E}[g \circ h(X)].$$

or equivalently,

$$\lim_{n \to \infty} \mathbb{E}[g(h(X_n))] = \mathbb{E}[g(h(X))].$$

Invoking one more time Part (ii) of Theorem 11.4 we now conclude that $h(X_n) \Rightarrow_n h(X)$ as desired. 

---

**11.4 Weak convergence of discrete rvs**

In this section we consider a collection of discrete rvs $\{X, X_n, n = 1, 2, \ldots\}$ with

$$\mathbb{P}[X \in S] = \mathbb{P}[X_n \in S] = 1, \quad n = 1, 2, \ldots$$

where $S = \{a_i, i \in I\}$ is a countable subset of $\mathbb{Z}$.

**Theorem 11.4.1** The sequence of discrete rvs $X_n \Rightarrow_n X$ converges weakly to the rv $X$ if and only if

$$\lim_{n \to \infty} \mathbb{P}[X_n = a_i] = \mathbb{P}[X = a_i], \quad i \in I.$$

**Proof.** Assume first that $X_n \Rightarrow_n X$. Pick a a point of discontinuity for $F_X$. By assumption $a$ is an element of $\mathbb{Z}$, so that $\varepsilon > 0$ can be selected so that both $a \pm \varepsilon$ are not in $\mathbb{Z}$, whence

$$\lim_{n \to \infty} \mathbb{P}[X_n \leq a \pm \varepsilon] = \mathbb{P}[X \leq a \pm \varepsilon].$$
Note however that
\[
\mathbb{P}[X_n \leq a - \varepsilon] = \mathbb{P}[X_n \leq a + \varepsilon] + \mathbb{P}[X_n = a], \quad n = 1, 2, \ldots
\]
and
\[
\mathbb{P}[X \leq a - \varepsilon] = \mathbb{P}[X \leq a + \varepsilon] + \mathbb{P}[X = a].
\]
since the probability distribution functions are piecewise constant with jumps only at points in \(\mathbb{Z}\).

Let \(n\) go to infinity in (11.10). It is plain from (11.9) that
\[
\lim_{n \to \infty} \mathbb{P}[X_n = a] \quad \text{exists and is given by}
\]
\[
\lim_{n \to \infty} \mathbb{P}[X_n = a] = \mathbb{P}[X \leq a + \varepsilon] - \mathbb{P}[X \leq a - \varepsilon] = \mathbb{P}[X = a]
\]
where the last equality follows from (11.11).

Conversely, assume that
\[
\lim_{n \to \infty} \mathbb{P}[X_n = a] = \mathbb{P}[X = a], \quad a \notin \mathcal{C}(\mathcal{F}_X)
\]
With a Borel subset \(B\) in \(\mathbb{R}\), we shall show that
\[
\lim_{n \to \infty} \mathbb{P}[X_n \in B] = \mathbb{P}[X \in B].
\]
This will immediately imply \(X_n \Rightarrow X\) upon specializing \(B\) to sets of the form \(B = (-\infty, x]\) with \(x\) in \(\mathcal{C}(\mathcal{F}_X)\). To do so, fix \(n = 1, 2, \ldots\) and pick \(A\) an arbitrary positive integer \(A\):

We see that
\[
\mathbb{P}[X_n \in B]
= \mathbb{P}[|X_n| \leq A, X_n \in B] + \mathbb{P}[|X_n| > A, X_n \in B]
= \sum_{a \in \mathbb{Z} \cap B : |a| \leq A} \mathbb{P}[X_n = a] + \mathbb{P}[|X_n| > A, X_n \in B]
\]
while
\[
\mathbb{P}[X \in B]
= \mathbb{P}[|X| \leq A, X \in B] + \mathbb{P}[|X| > A, X \in B]
= \sum_{a \in \mathbb{Z} \cap B : |a| \leq A} \mathbb{P}[X = a] + \mathbb{P}[|X| > A, X \in B].
\]
Subtracting we conclude that
\[
\mathbb{P}[X_n \in B] - \mathbb{P}[X \in B] \leq \sum_{a \in \mathbb{Z} \cap B : |a| \leq A} \mathbb{P}[X_n = a] - \mathbb{P}[X = a] + \mathbb{P}[|X_n| > A] + \mathbb{P}[|X| > A].
\]
Let \( n \) go to infinity in this last inequality: Using (11.12) we get
\[
\lim_{n \to \infty} \sum_{a \in \mathbb{Z} \cap B: |a| \leq A} |\mathbb{P}[X_n = a] - \mathbb{P}[X = a]| = 0
\]
since this sum has at most \( 2A + 1 \) terms, while
\[
\lim_{n \to \infty} \mathbb{P}[|X_n| > A] = \lim_{n \to \infty} (1 - \mathbb{P}[|X_n| \leq A]) = 1 - \mathbb{P}[|X| \leq A] = \mathbb{P}[|X| > A]
\]
by a similar argument. Collecting these facts we obtain
\[
\limsup_{n \to \infty} |\mathbb{P}[X_n \in B] - \mathbb{P}[X \in B]| \leq 2\mathbb{P}[|X| > A].
\]
Now let \( A \) go to infinity in this last inequality and note that
\[
\lim_{A \to \infty} \left( \limsup_{n \to \infty} |\mathbb{P}[X_n \in B] - \mathbb{P}[X \in B]| \right) = 0
\]
and the desired conclusion (11.13) follows since the left handside does not depend on \( A \). 

In the more restrictive setting where \( S \subseteq \mathbb{N} \), probability generating functions can be defined, and the following analog of Theorem 11.1.1 holds.

**Theorem 11.4.2** 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
\[
\lim_{n \to \infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)].
\]
Here as well we shall write \( X_n \Rightarrow_n X \) or \( X_n \xrightarrow{L} X \). Some authors also refer to this mode of convergence as *convergence in law* or as *weak convergence*.

Theorem 11.4 has the following multi-dimensional analog.
CHAPTER 11. WEAK CONVERGENCE

Theorem 11.4.3 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

$$
\lim_{n \to \infty} \Phi_{X_n}(\theta) = \Phi_X(\theta), \quad \theta \in \mathbb{R}.
$$

(11.18)

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 11.3.2, we readily see that Theorem 11.4.3 has the following important consequence.

Theorem 11.4.4 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

$$
h(X_n) \quad \rightarrow_n h(X).
$$

11.5 Convergence and limits of Gaussian rvs

In later chapters we will need to define integrals with respect to Gaussian processes. As in the deterministic case, these stochastic integrals will be defined as limits of partial sums of the form

$$
X_n := \sum_{i=1}^{k_n} a_j^{(n)} Y_j^{(n)}, \quad n = 1, 2, \ldots
$$

(11.19)

where for each $n = 1, 2, \ldots$, the integer $k_n$ and the coefficients $a_j^{(n)}$, $j = 1, \ldots, k_n$, are non-random while the rvs $\{Y_j^{(n)}, j = 1, \ldots, k_n\}$ are jointly Gaussian rvs. Typically, as $n$ goes to infinity so does $k_n$. Note that under the foregoing assumptions for each $n = 1, 2, \ldots$, the rv $X_n$ is Gaussian with

$$
\mathbb{E} [X_n] = \sum_{i=1}^{k_n} a_j^{(n)} \mathbb{E} [Y_j^{(n)}]
$$

(11.20)
11.5. CONVERGENCE AND LIMITS OF GAUSSIAN RVS

and

\begin{equation}
\text{Var}[X_n] = \sum_{i=1}^{k_n} \sum_{j=1}^{k_n} a_i^{(n)} a_j^{(n)} \text{Cov}[Y_i^{(n)}, Y_j^{(n)}].
\end{equation}

Therefore, the study of such integrals is expected to pass through the convergence of sequence of rvs \( \{X_n, \ n = 1, 2, \ldots\} \) of the form (11.19). Such considerations lead naturally to the need for the following result [, Thm. , p.]:

**Lemma 11.5.1**  Let \( \{X_k, \ k = 1, 2, \ldots\} \) denote a collection of \( \mathbb{R}^p \)-valued Gaussian rvs. For each \( k = 1, 2, \ldots \), let \( \mu_k \) and \( \Sigma_k \) 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 \( \mu \) in \( \mathbb{R}^p \) and a \( d \times d \) matrix \( \Sigma \) such that

\begin{equation}
\lim_{k \to \infty} \mu_k = \mu \quad \text{and} \quad \lim_{k \to \infty} \Sigma_k = \Sigma.
\end{equation}

In that case,

\[ X_k \rightarrow_k X \]

where \( X \) is an \( \mathbb{R}^d \)-valued Gaussian rv with mean vector \( \mu \) and covariance matrix \( \Sigma \).

The second half of condition (11.22) ensures that the matrix \( \Sigma \) is symmetric and non-negative definite, hence a covariance matrix.

Returning to the partial sums (11.19) we see that Lemma 11.5.1 (applied with \( d = 1 \)) requires identifying the limits \( \mu = \lim_{n \to \infty} \mathbb{E}[X_n] \) and \( \sigma^2 = \lim_{n \to \infty} \text{Var}[X_n] \), in which case \( X_n \overset{d}{\rightarrow} X \) where \( X \) is an \( \mathbb{R} \)-valued Gaussian rv with mean \( \mu \) and variance \( \Sigma \). In Section ?? we discuss a situation where this can be done quite easily.
Chapter 12

The classical limit theorems

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 are known as Laws of Large Numbers; they 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 Large Numbers.

12.1 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 results is given first.

**Theorem 12.1.1** 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,

\[
(12.1) \quad \frac{S_n}{n} \xrightarrow{L^2} n \mathbb{E} [X],
\]

whence

\[
(12.2) \quad \frac{S_n}{n} \xrightarrow{P} n \mathbb{E} [X].
\]
Proof. For each \( n = 1, 2, \ldots \), we note that

\[
\mathbb{E} \left[ \left| \frac{S_n}{n} - \mathbb{E}[X] \right|^2 \right] = \mathbb{E} \left[ \left| \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X]) \right|^2 \right]
\]

(12.3)

\[
= \frac{1}{n^2} \cdot \text{Var}[S_n]
\]

with

\[
\text{Var}[S_n] = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \text{Cov}[X_k, X_\ell]
\]

\[
= \sum_{k=1}^{n} \text{Var}[X_k]
\]

(12.4)

since

\[
\text{Cov}[X_k, X_\ell] = \delta(k; \ell) \text{Var}[X], \quad k, \ell = 1, \ldots, n
\]

under the enforced independence assumptions.

As a result,

\[
\mathbb{E} \left[ \left| \frac{S_n}{n} - \mathbb{E}[X] \right|^2 \right] = \frac{n \text{Var}[X]}{n^2} = \frac{\text{Var}[X]}{n^2}
\]

and the desired conclusions follow. \( \square \)

12.2 Weak Laws of Large Numbers (II)

A careful inspection of the proof of Theorem 12.1.1 suggests a more general result. Assume that the rvs \( \{X_n, n = 1, 2, \ldots\} \) are second-order rvs. For each \( n = 1, 2, \ldots \), we note that

\[
\mathbb{E} \left[ \left| \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \right|^2 \right] = \frac{\text{Var}[S_n]}{n^2}.
\]
By computations similar to the ones used in the proof of Theorem 12.1.1, we find

\[ \text{Var}[S_n] = \text{Var}\left[ \sum_{k=1}^{n} X_k \right] \]
\[ = \sum_{k=1}^{n} \sum_{\ell=1}^{n} \text{Cov}[X_k, X_\ell] \]
\[ = \sum_{k=1}^{n} \text{Var}[X_k] + \sum_{k,\ell=1, k\neq \ell}^{n} \text{Cov}[X_k, X_\ell], \]

whence

\[ \mathbb{E} \left[ \left| \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \right|^2 \right] \]
\[ = \frac{1}{n^2} \sum_{k=1}^{n} \text{Var}[X_k] + \frac{1}{n^2} \sum_{k,\ell=1, k\neq \ell}^{n} \text{Cov}[X_k, X_\ell]. \]

**Theorem 12.2.1** Consider a collection \{\(X_n, n = 1, 2, \ldots\)\} of second-order rvs such that

\( \lim_{n \to \infty} \frac{1}{n^2} \sum_{k=1}^{n} \text{Var}[X_k] = 0. \)

We have

\( \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \xrightarrow{L^2} 0 \)
and

\( \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \xrightarrow{P} 0 \)

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.,

\[ \text{Cov}[X_k, X_\ell] \leq 0, \quad k \neq \ell, \quad k, \ell = 1, \ldots, n \]

(iii) The rvs \( \{X_n, n = 1, 2, \ldots\} \) satisfy the condition

\( \lim_{n \to \infty} \frac{1}{n^2} \sum_{k,\ell=1, k\neq \ell}^{n} \text{Cov}[X_k, X_\ell] = 0. \)
12.3 The classical Weak Law of Large Numbers (III)

As we now show, the finiteness of the second moment of $X$ can be dropped.

**Theorem 12.3.1** Assume the rvs $\{X, X_n, n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E} [|X|] < \infty$. Then, we have

\[
\frac{S_n}{n} \xrightarrow{P} \mathbb{E} [X].
\]

**Proof.** Fix $n = 1, 2, \ldots$ and $\theta$ in $\mathbb{R}$. Note that

\[
\mathbb{E} \left[ e^{i\theta \left( \frac{S_n}{n} - \mathbb{E}[X] \right)} \right] = \mathbb{E} \left[ e^{i\theta \sum_{k=1}^{n} (X_k - \mathbb{E}[X])} \right]
\]

\[
= \mathbb{E} \left[ \prod_{k=1}^{n} e^{i\frac{\theta}{n} (X_k - \mathbb{E}[X])} \right]
\]

\[
= \prod_{k=1}^{n} \mathbb{E} \left[ e^{i\frac{\theta}{n} (X_k - \mathbb{E}[X])} \right]
\]

\[
= \left( \mathbb{E} \left[ e^{i\frac{\theta}{n} (X - \mathbb{E}[X])} \right] \right)^n.
\]

(12.14)

so that

\[
\mathbb{E} \left[ e^{i\theta \left( \frac{S_n}{n} - \mathbb{E}[X] \right)} \right] = \left( \mathbb{E} \left[ e^{i\frac{\theta}{n} (X - \mathbb{E}[X])} \right] \right)^n.
\]
As pointed out in Section 8.7, using Theorem 8.7.1 (for $k = 1$ and $x = X - \mathbb{E}[X]$), we get
\[ e^{i\theta(X - \mathbb{E}[X])} = 1 + i\theta(X - \mathbb{E}[X]) + i\theta \int_0^{X - \mathbb{E}[X]} (e^{i\theta t} - 1) \, dt, \]
whence
\[ \mathbb{E} \left[ e^{i\theta(X - \mathbb{E}[X])} \right] = 1 + i\theta \mathbb{E} \left[ \int_0^{X - \mathbb{E}[X]} (e^{i\theta t} - 1) \, dt \right] \]
on taking expectations. Substituting $\theta$ by $\frac{\theta}{n}$, we obtain the relation
\[ \mathbb{E} \left[ e^{i\frac{\theta}{n}(X - \mathbb{E}[X])} \right] = 1 + \frac{i\theta}{n} \cdot C_1 \left( \frac{\theta}{n} \right) \]
where
\[ C_1(\theta) \equiv \mathbb{E} \left[ \int_0^{X - \mathbb{E}[X]} (e^{i\theta t} - 1) \, dt \right]. \]
It follows that
\[ (12.15) \quad \mathbb{E} \left[ e^{i\frac{\theta}{n}(S_n - \mathbb{E}[X])} \right] = \left( 1 + \frac{i\theta}{n} \cdot C_1 \left( \frac{\theta}{n} \right) \right)^n. \]

By Dominated Convergence, we conclude that $\lim_{n \to \infty} C_1 \left( \frac{\theta}{n} \right) = 0$, whence
\[ \lim_{n \to \infty} \left( \mathbb{E} \left[ e^{i\frac{\theta}{n}(X - \mathbb{E}[X])} \right] \right)^n = \lim_{n \to \infty} \left( 1 + \frac{i\theta}{n} \cdot C_1 \left( \frac{\theta}{n} \right) \right)^n = 1. \]
It follows that $\frac{S_n}{n} = \mathbb{E}[X] \xrightarrow{P} 0$, and this conclude the proof of (12.13).

\section{The Strong Law of Large Numbers}

Strong Laws of Large Numbers are given are convergence statements in the a.s. sense. The classical Strong Law of Large Numbers is given next.

\textbf{Theorem 12.4.1} Assume the rvs $\{X, X_n, \ n = 1, 2, \ldots\}$ to be i.i.d. rvs with $\mathbb{E} [|X|] < \infty$. Then,
\[ (12.16) \quad \lim_{n \to \infty} \frac{S_n}{n} = \mathbb{E}[X] \quad \text{a.s.} \]
We give two proofs of this result under stronger assumptions on the moments of \( X \). One proof assumes \( \mathbb{E} [ |X|^4] < \infty \) while the second proof is given under the condition \( \mathbb{E} [ |X|^2] < \infty \). A proof under the first moment condition \( \mathbb{E} [ |X|] < \infty \) is available in [1].

**Proof 1** Assume \( \mathbb{E} [ |X|^4] < \infty \) — 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

\[
\sum_{n=1}^{\infty} \mathbb{E} \left[ \left( \frac{S_n}{n} \right)^4 \right] < \infty,
\]

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 (12.17) we note 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

\[
\mathbb{E} \left[ \left( \sum_{k=1}^{n} X_k \right)^4 \right] = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \mathbb{E} [X_iX_jX_kX_\ell].
\]

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 \( i, j, k, \ell \) 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 \( \mathbb{E} [X_iX_jX_kX_\ell] = \mathbb{E} [X^4] \); there are \( n \)
12.4. THE STRONG LAW OF LARGE NUMBERS

such configurations; (ii) If \( \{i, j, k, \ell\} \) contains only two distinct values, say \( a \neq b \) appearing as \( aabb, abab \) and \( abba \) in (12.18), then \( \mathbb{E}[X_i X_j X_k X_\ell] = (\mathbb{E}[X^2])^2 \); there are \( 3n(n-1) \) such configurations. It follows that

\[
\mathbb{E} \left[ \left( \sum_{k=1}^{n} X_k \right)^4 \right] = n \mathbb{E}[X^4] + 3n(n-1)(\mathbb{E}[X^2])^2,
\]

whence

\[
\mathbb{E} \left[ \left( \frac{S_n}{n} \right)^4 \right] = \frac{1}{n^3} \mathbb{E}[X^4] + 3 \frac{n-1}{n^3} (\mathbb{E}[X^2])^2.
\]

The conclusion (12.17) readily follows, and this completes the proof.

Proof 2 Assume \( \mathbb{E}[|X|^2] < \infty \) – For each \( k = 1, 2, \ldots \), we note that

\[
\text{Var} \left[ \frac{S_k}{k^2} \right] = \frac{\text{Var}[X]}{k^2}
\]

so that

\[
\sum_{k=1}^{\infty} \mathbb{P} \left[ \left| \frac{S_k}{k^2} \right| > \varepsilon \right] \leq \frac{1}{\varepsilon^2} \sum_{k=1}^{\infty} \frac{\text{Var}[X]}{k^2} < \infty, \quad \varepsilon > 0.
\]

It follows from Theorem 10.1.2 that

(12.19) \[
\lim_{k \to \infty} \frac{S_k}{k^2} = \mathbb{E}[X] \quad \text{a.s.}
\]

Now assume that the rvs \( \{X, X_n, n = 1, 2, \ldots\} \) are non-negative, i.e., \( X \geq 0 \) a.s. (in which case obviously \( \mathbb{E}[X] \geq 0 \). The case when the rvs \( \{X, X_n, n = 1, 2, \ldots\} \) are non-positive, i.e., \( X \leq 0 \) a.s., can be handed mutatis mutandis.

Fix \( n = 1, 2, \ldots \). There exists a unique positive integer \( k(n) \) such that

(12.20) \[
k(n)^2 \leq n < (k(n)+1)^2.
\]

Under the non-negativity assumption we note the inequalities

\[
S_{k(n)^2} \leq S_n \leq S_{(k(n)+1)^2} \quad \text{a.s.}
\]

by virtue of the fact that \( X_\ell \geq 0 \) a.s. for \( \ell = k(n)^2, \ldots, (k(n)+1)^2 - 1 \). It follows that

(12.21) \[
\frac{k(n)^2}{n} \cdot \left( \frac{S_{k(n)^2}}{k(n)^2} \right) \leq \frac{S_n}{n} \leq \frac{(k(n)+1)^2}{n} \cdot \left( \frac{S_{(k(n)+1)^2}}{(k(n)+1)^2} \right).
\]
Using (12.20) we readily get
\[ \frac{k(n)^2}{n} \leq 1 < \frac{k(n)^2}{n} + 2 \cdot \frac{k(n)}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} + \frac{1}{n} \]  
(12.22)

It is now straightforward to conclude from the first inequality in (12.22) that
\[ \limsup_{n \to \infty} \frac{k(n)^2}{n} \leq 1 \]
with \( \frac{k(n)}{\sqrt{n}} \leq 1 \), and the second inequality in (12.22) therefore yields \( 1 \leq \liminf_{n \to \infty} \frac{k(n)^2}{n} \).

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 (12.21), and we readily get (12.16) upon combining this last conclusion with the convergence (12.19).

To complete the proof note that \( \mathbb{E}[(X^\pm)^2] < \infty \) since \( \mathbb{E}[|X|^2] = \mathbb{E}[(X^+)^2] + \mathbb{E}[(X^-)^2] \). Thus, it holds that
\[ \lim_{n \to \infty} \frac{\sum_{k=1}^{n} X_k^\pm}{n} = \mathbb{E}[X^\pm] \quad \text{a.s.} \]  
(12.23)

since the rvs \( \{X^\pm, X_k^\pm, k = 1, 2, \ldots\} \) form an i.i.d. sequence of rvs with finite second moments. The desired result (12.16) automatically follows since
\[ X_n = X_n^+ - X_n^-, \quad n = 1, 2, \ldots \]
and \( \mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] \).

\[ \Box \]

12.5 The Central Limit Theorem

The Central Limit Theorem completes the Law of Large Numbers, in that it provides some indication as to the rate at which convergence takes place.

**Theorem 12.5.1** Assume the rvs \( \{X, X_n, n = 1, 2, \ldots\} \) to be i.i.d. rvs with \( \mathbb{E}[|X|^2] < \infty \). Then, we have
\[ \sqrt{n} \left( \frac{S_n}{n} - \mathbb{E}[X] \right) \Rightarrow_n \sqrt{\text{Var}[X]} : U \]  
(12.24)

where \( U \) is standard zero-mean unit-variance Gaussian rv.
12.5. THE CENTRAL LIMIT THEOREM

Proof. Fix \( n = 1, 2, \ldots \) and \( \theta \) in \( \mathbb{R} \). This time, as in the proof of Theorem 12.3.1 we get

\[
\mathbb{E} \left[ e^{i\theta \sqrt{n} \left( \frac{\hat{S}_n}{n} - \mathbb{E}[X] \right)} \right] = \left( \mathbb{E} \left[ e^{i\frac{\theta}{\sqrt{n}} \left( X - \mathbb{E}[X] \right)} \right] \right)^n
\]

under the enforced independence.

Using Theorem 8.7.1 (with \( k = 2 \) and \( x = X - \mathbb{E}[X] \)), we get

\[
e^{i\theta(X - \mathbb{E}[X])} = 1 + i\theta(X - \mathbb{E}[X]) - \frac{\theta^2}{2} (X - \mathbb{E}[X])^2
\]

(12.25)

and taking expectations yields

\[
\mathbb{E} \left[ e^{i\theta(X - \mathbb{E}[X])} \right] = 1 - \frac{\theta^2}{2} \cdot \text{Var} \, [X] - \frac{\theta^2}{2} \cdot C_2(\theta)
\]

(12.26)

with

(12.27) \( C_2(\theta) \equiv \mathbb{E} \left[ \int_0^{X - \mathbb{E}[X]} (X - \mathbb{E}[X] - t) \left( e^{i\theta t} - 1 \right) dt \right] \).

Substituting \( \theta \) by \( \frac{\theta}{\sqrt{n}} \) in this last relation leads to

\[
\mathbb{E} \left[ e^{i\frac{\theta}{\sqrt{n}} (X - \mathbb{E}[X])} \right] = 1 - \frac{\theta^2}{2n} \cdot \text{Var} \, [X] - \frac{\theta^2}{2n} \cdot C_2 \left( \frac{\theta}{\sqrt{n}} \right)
\]

so that

\[
\mathbb{E} \left[ e^{i\theta \sqrt{n} \left( \frac{\hat{S}_n}{n} - \mathbb{E}[X] \right)} \right] = \left( 1 - \frac{\theta^2}{2n} \cdot \text{Var} \, [X] - \frac{\theta^2}{2n} \cdot C_2 \left( \frac{\theta}{\sqrt{n}} \right) \right)^n.
\]

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

\[
\lim_{n \to \infty} n \left( \frac{\theta^2}{2n} \cdot \text{Var} \, [X] - \frac{\theta^2}{2} \cdot C_2 \left( \frac{\theta}{\sqrt{n}} \right) \right) = \frac{\theta^2}{2} \cdot \text{Var} \, [X]
\]
It follows that
\[
\lim_{n \to \infty} E \left[ e^{i \theta \sqrt{n} \left( \frac{S_n}{n} - E[X] \right)} \right] = e^{-\frac{x^2}{2} \cdot \Var[X]}
\]
This completes the proof of (12.24).

12.6 The Central Limit Theorem – An application

We are still in the setting of Theorem 12.5.1. We can rephrase (12.24) as
\[
\lim_{n \to \infty} P \left[ \sqrt{n} \left( \frac{S_n}{n} - E[X] \right) \leq x \right] = P \left[ \sqrt{\Var[X]} \cdot U \leq x \right], \quad x \in \mathbb{R}.
\]
(12.28)
as we recall that every point in \( \mathbb{R} \) is a point of continuity for the rv \( U \) (or \( \sqrt{\Var[X]} \cdot U \)).

It follows that
\[
\lim_{n \to \infty} P \left[ \left| \sqrt{n} \left( \frac{S_n}{n} - E[X] \right) \right| \leq x \right] = P \left[ \sqrt{\Var[X]} \cdot U \leq x \right] - P \left[ \sqrt{\Var[X]} \cdot U \leq -x \right] = \Phi \left( \frac{x}{\sqrt{\Var[X]}} \right) - \Phi \left( -\frac{x}{\sqrt{\Var[X]}} \right)
\]
(12.29)
Fix \( x \geq 0 \) and \( n = 1, 2, \ldots \): We have
\[
\left| \sqrt{n} \left( \frac{S_n}{n} - E[X] \right) \right| \leq x
\]
if and only if
\[
-x \leq \sqrt{n} \left( \frac{S_n}{n} - E[X] \right) \leq x
\]
if and only if
\[
E[X] \in \left[ \frac{S_n}{n} - \frac{x}{\sqrt{n}}, \frac{S_n}{n} + \frac{x}{\sqrt{n}} \right].
\]
Thus, if we think of 
\[ \hat{X}_n = \frac{S_n}{n}, \quad n = 1, 2, \ldots \]
as an estimate of \( E[X] \) 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
\[ \lim_{n \to \infty} \hat{X}_n = E[X] \quad a.s. \]
The calculations above show via (12.29) that
\[
\lim_{n \to \infty} \mathbb{P} \left[ E[X] \in \left[ \hat{X}_n - \frac{x}{\sqrt{n}}, \hat{X}_n + \frac{x}{\sqrt{n}} \right] \right] = 2\Phi \left( \frac{x}{\sqrt{\text{Var}[X]}} \right) - 1, \quad x \geq 0.
\]
(12.30)
In other words, for large \( n \), the unknown value \( E[X] \) lies in a symmetric interval centered at the estimate \( \hat{X}_n \) (obtained from the observed data \( X_1, \ldots, X_n \)) of width \( \frac{2x}{\sqrt{n}} \) with a probability approximately given by
\[ 2\Phi \left( \frac{x}{\sqrt{\text{Var}[X]}} \right) - 1, \]
the accuracy of this approximation improving with increasing \( n \). With \( \alpha \) in \((0, 1)\) given, we can ensure that
\[ \mathbb{P} \left[ E[X] \in \left[ \hat{X}_n - \frac{x}{\sqrt{n}}, \hat{X}_n + \frac{x}{\sqrt{n}} \right] \right] \simeq 1 - \alpha \]
for large \( n \) if we select \( x \geq 0 \) such that
\[ 2\Phi \left( \frac{x}{\sqrt{\text{Var}[X]}} \right) - 1 = 1 - \alpha, \]
or equivalently,
\[ \Phi \left( \frac{x}{\sqrt{\text{Var}[X]}} \right) = 1 - \frac{\alpha}{2}. \]
With \( \lambda \) in \((0, 1)\) let \( z_\lambda \) denote the unique solution to the nonlinear equation
\[ 1 - \Phi(x) = \lambda, \quad x \in \mathbb{R}. \]
Equivalently,
\[ P \{ U > x \} = \lambda, \quad x \in \mathbb{R}. \]

With this notation we see that the \textit{random} interval
\[
\left[ \frac{S_n}{n} - \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\text{Var}[X]}, \frac{S_n}{n} + \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\text{Var}[X]} \right]
\]
is known as the \textit{confidence interval} for estimating \( \mathbb{E}[X] \) on the basis data \( X_1, \ldots, X_n \) with confidence \((1 - \alpha)\%\)

Note that this analysis is predicated on knowing the variance \( \text{Var}[X] \). When this value is unknown, we replace \( \text{Var}[X] \) by the \textit{sample variance} \( S_n^2 \) given by
\[
S_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left( X_k - \frac{1}{n} \sum_{\ell=1}^{n} X_\ell \right)^2, \quad n = 2, 3, \ldots
\]

\section*{12.7 Poisson convergence}

The setting is as 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.,
\[
P \{ X_{k,n}(p_n) = 1 \} = 1 - P \{ X_{k,n}(p_n) = 0 \} = p_n, \quad k = 1, \ldots, n
\]

Write
\[
S_n = \sum_{k=1}^{n} X_k(p_n), \quad n = 1, 2, \ldots
\]

\textbf{Theorem 12.7.1} \ Assume there exists \( \lambda > 0 \) such that
\begin{equation}
\lim_{n \to \infty} np_n = \lambda. \tag{12.31}
\end{equation}

Then, we have
\begin{equation}
S_n \xrightarrow{n} \Pi(\lambda) \tag{12.32}
\end{equation}

where \( \Pi(\lambda) \) denotes a Poisson rv with parameter \( \lambda \).

The convergence (12.32) can be restated as
\begin{equation}
\lim_{n \to \infty} P \{ S_n = k \} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots \tag{12.33}
\end{equation}

We give two proofs of this important result.
12.7. POISSON CONVERGENCE

**Proof 1** The first proof uses the characterization of weak convergence for integer-valued rvs given in Theorem 11.4.1: Fix \( n = 1, 2, \ldots \). Under the independence assumptions, the rv \( S_n \) is a binomial rv \( \text{Bin}(n; p_n) \). Thus, Fix \( k = 0, 1, \ldots \) For every integer \( n \) such that \( k \leq n \) we have

\[
\Pr[S_n = k] = \binom{n}{k} p_n^k (1 - p_n)^{n-k}
\]

\[
= \frac{n!}{k!(n-k)!} \cdot p_n^k (1 - p_n)^{n-k}
\]

\[
= \frac{1}{k!} \left( \frac{p_n}{1 - p_n} \right)^k \cdot \frac{n!}{(n-k)!} \cdot (1 - p_n)^n
\]

\[
= \frac{1}{k!} \left( \frac{np_n}{1 - p_n} \right)^k \cdot \frac{n!}{n^k(n-k)!} \cdot (1 - p_n)^n.
\]

(12.34)

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 (12.31) 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 (12.33) as we make use of Theorem 11.4.1.

\[\blacksquare\]

**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 \( \theta \) in \( \mathbb{R} \) we get

\[
\mathbb{E} \left[ e^{i\theta S_n} \right] = \mathbb{E} \left[ e^{i\theta \sum_{k=1}^{n} X_k(p_n)} \right]
\]

\[
= \mathbb{E} \left[ \prod_{k=1}^{n} e^{i\theta X_k(p_n)} \right]
\]

\[
= \prod_{k=1}^{n} \mathbb{E} \left[ e^{i\theta X_k(p_n)} \right]
\]
\[ (12.35) \]

\[ = \left( 1 - p_n + p_n e^{i\theta} \right)^n \]

\[ = \left( 1 - p_n \left( 1 - e^{i\theta} \right) \right)^n. \]

Under (12.31) we get that

\[ \lim_{n \to \infty} np_n \left( 1 - e^{i\theta} \right) = \lambda \left( 1 - e^{i\theta} \right). \]

Thus,

\[ \lim_{n \to \infty} \mathbb{E} \left[ e^{i\theta S_n} \right] = e^{-\lambda \left( 1 - e^{i\theta} \right)}, \quad \theta \in \mathbb{R} \]

and the conclusion (12.32) follows since

\[ \mathbb{E} \left[ e^{i\theta \Pi(\lambda)} \right] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \cdot e^{ik\theta} \]

\[ = \left( \sum_{k=0}^{\infty} \frac{1}{k!} \left( \lambda e^{i\theta} \right)^k \right) e^{-\lambda} = e^{-\lambda \left( 1 - e^{i\theta} \right)}, \quad \theta \in \mathbb{R} \]

as we use Theorem 11.4.2.
Part III

APPENDICES
Chapter 13

Limits in $\mathbb{R}$

We begin with a few 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 $\{a_n, \, n = 1, 2, \ldots\}$.

A sequence $a : \mathbb{N}_0 \to \mathbb{R}$ converges to $a^*$ in $\mathbb{R}$ if for every $\varepsilon > 0$, there exists an integer $n^*(\varepsilon)$ such that

$$ |a_n - a^*| \leq \varepsilon, \quad n \geq n^*(\varepsilon). $$

(13.1)

We shall write $\lim_{n \to \infty} a_n = a^*$, and refer to the scalar $a^*$ 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 unboundedly 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 occurrences is as follows: We write $\lim_{n \to \infty} a_n = \infty$ to signify that for every $M > 0$, there exists a finite integer $n^*(M)$ in $\mathbb{N}_0$ such that

$$ a_n > M, \quad n \geq n^*(M). $$

(13.2)

It is natural to define $\lim_{n \to \infty} a_n = -\infty$ if $\lim_{n \to \infty} (-a_n) = \infty$.

If there exists $a^*$ in $\mathbb{R} \cup \{-\infty, \infty\}$ such that $\lim_{n \to \infty} a_n = a^*$, 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 $a^*$ is an element of $\mathbb{R}$ (thus finite).

Applying the definition (13.1) requires that the limit be known. Often this information is not available, and yet the need remains to check 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 \( a^* \) in \( \mathbb{R} \), then for every \( \varepsilon > 0 \), there exists a finite integer \( n^*(\varepsilon) \) such that (13.1) holds, hence for \( n, m \geq n^*(\varepsilon) \) we have
\[
|a_n - a_m| \leq |a_n - a^*| + |a^* - a_m| \leq \varepsilon + \varepsilon = 2\varepsilon.
\]

This observation is turned into the following definition.

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 \( n^*(\varepsilon) \) such that
\[
|a_n - a_m| \leq \varepsilon, \quad m, n \geq n^*(\varepsilon).
\]

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 for convergence of the sequence in \( \mathbb{R} \).

**Theorem 13.0.2 (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 criterion for convergence which does not require knowledge of the limit.

**13.1 Two important facts**

In addition to the Cauchy convergence criterion, here are two facts that are often found useful in studying convergence.

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.

**Theorem 13.1.1** 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).
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 \( n_k < n_{k+1}, \ k = 1, 2, \ldots \). This forces \( \lim_{k \to \infty} n_k = \infty \).

The sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) is said to be bounded if there exists some \( B > 0 \) such that \( \sup(|a_n|, n = 1, 2, \ldots) \leq B \).

**Theorem 13.1.2 (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 \( \lim_{k \to \infty} a_{n_k} = a^* \) for some \( a^* \) in \( \mathbb{R} \).

### 13.2 Accumulation points

Since not all sequences converge, it is important to understand how non-convergence occurs.

An accumulation point for the sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) is defined as any \( a^* \) in \( \mathbb{R} \cup \{ \pm \infty \} \) such that \( \lim_{k \to \infty} a_{n_k} = a^* \) for some subsequence \( \mathbb{N}_0 \to \mathbb{R} : k \to a_{n_k} \).

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 \( \mathbb{R} \cup \{ \pm \infty \} \)), in which case there is a smallest and a largest. The next definition formalizes this observation. Given a sequence \( a : \mathbb{N}_0 \to \mathbb{R} \), the quantities

\[
\bar{A} = \limsup_{n \to \infty} a_n = \inf_{n \geq 1} \left( \sup_{m \geq n} a_m \right)
\]

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

\[
\bar{A}_n = \sup_{m \geq n} a_m \quad \text{and} \quad \underline{A}_n = \inf_{m \geq n} a_m
\]

(13.4)

Note that \( \underline{A}_n \leq \bar{A}_n \), and the sequences \( n \to \bar{A}_n \) and \( n \to \underline{A}_n \) are non-increasing and non-decreasing, respectively. Therefore, \( \bar{A} = \lim_{n \to \infty} \bar{A}_n \) and \( \underline{A} = \lim_{n \to \infty} \underline{A}_n \) both exist, but are possibly infinite. Moreover, we always have \( \underline{A} \leq \bar{A} \).

**Theorem 13.2.1** Consider a sequence \( a : \mathbb{N}_0 \to \mathbb{R} \). If it converges to \( a^* \), then \( \bar{A} = A = a^* \). Conversely, if \( \bar{A} = A = a^* \) for some \( a^* \) in \( \mathbb{R} \cup \{\pm \infty\} \), then the sequence converges to \( a^* \).

Note that 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 \( \bar{A} \leq \bar{B} \) 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

\[
0 \leq p_n \leq 1, \quad n = 1, 2, \ldots
\]

(13.5)

If we show that

\[
1 \leq \liminf_{n \to \infty} p_n,
\]

(13.6)

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 (13.5), whence

\[
\liminf_{n \to \infty} p_n = \limsup_{n \to \infty} p_n = 1
\]

upon using (13.6). In a similar vein, if we show \( \limsup_{n \to \infty} p_n = 0 \), then we necessarily have convergence of the sequence with \( \lim_{n \to \infty} p_n = 0 \).

### 13.3 Cesaro convergence

With any sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) we associate the Cesaro sequence \( \bar{a} : \mathbb{N}_0 \to \mathbb{R} \) given by

\[
\bar{a}_n = \frac{1}{n} (a_1 + \ldots + a_n), \quad n = 1, 2, \ldots
\]

**Theorem 13.3.1** (Cesaro convergence) If the sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) converges to \( a^* \), then the Cesaro sequence \( \bar{a} : \mathbb{N}_0 \to \mathbb{R} \) also converges with limit \( a^* \).
**Proof.** First we assume the convergent sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) to have a finite limit \( a^* \) in \( \mathbb{R} \). Note that

\[
\bar{a}_n - a^* = \frac{1}{n} \sum_{k=1}^{n} (a_k - a^*), \quad n = 1, 2, \ldots
\]

Now, for every \( \varepsilon > 0 \), there exists an integer \( n^*(\varepsilon) \) such that

\[
|a_n - a^*| \leq \frac{\varepsilon}{2}, \quad n \geq n^*(\varepsilon).
\]

On that range, with \( B(\varepsilon) = \sum_{k=1}^{n^*(\varepsilon)} |a_k - a^*| \), we have

\[
|\bar{a}_n - a^*| \leq \frac{1}{n} \sum_{k=1}^{n} |a_k - a^*| = \frac{1}{n} \sum_{k=1}^{n^*(\varepsilon)} |a_k - a^*| + \frac{1}{n} \sum_{k=n^*(\varepsilon)+1}^{n} |a_k - a^*| \\
\leq \frac{B(\varepsilon)}{n} + \frac{n - n^*(\varepsilon)}{n} \cdot \varepsilon \\
\leq \frac{B(\varepsilon)}{n} + \varepsilon
\]

(13.7)

Since \( \lim_{n \to \infty} \frac{1}{n} = 0 \), for every \( \varepsilon > 0 \), there exists a finite integer \( n^{**}(\varepsilon) \) such that

\[
\frac{1}{n} \leq \frac{\varepsilon}{B(\varepsilon)}, \quad n \geq n^{**}(\varepsilon).
\]

Just take \( n^{**}(\varepsilon) = \lceil \frac{B(\varepsilon)}{\varepsilon} \rceil \). As a result,

\[
|\bar{a}_n - a^*| \leq \varepsilon + \varepsilon = 2\varepsilon, \quad n \geq \max(n^*(\varepsilon), n^{**}(\varepsilon))
\]

and the proof is now complete since \( \varepsilon \) is arbitrary. We leave it as an exercise to show the result when \( a^* = \pm \infty \).

However, the converse is not true: Take the sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) given by

\[
a_n = (-1)^n, \quad n = 1, 2, \ldots
\]

This sequence does not converge and yet \( \lim_{n \to \infty} \bar{a}_n = 0 \). This example nicely illustrate 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: Consider the sequence \( a_n : \mathbb{N}_0 \to \mathbb{R} \) given by

\[
a_n = (-1)^k, \quad 2^k \leq n < 2^{k+1}
\]

It is plain that \( \lim \inf_{n \to \infty} a_n = -1 \) while \( \lim \sup_{n \to \infty} a_n = 1 \), and so the sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) does not converge. However, it is also not Cesaro convergent.

### 13.4 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 \( n \textsuperscript{th} \) 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 \( a : \mathbb{N}_0 \to \mathbb{R} \).

It is customary to say that the series \( \sum_{n=1}^{\infty} a_n \) converges if the sequence \( s : \mathbb{N}_0 \to \mathbb{R} \) converges to some \( s^* \) in \( \mathbb{R} \), in which case we often write \( \sum_{n=1}^{\infty} a_n \) as its limit.

This amounts to the following: For every \( \varepsilon > 0 \) there exists a finite integer \( n^*(\varepsilon) \) such that

\[
|s_n - s^*| < \varepsilon, \quad n \geq n^*(\varepsilon).
\]

The series \( s : \mathbb{N}_0 \to \mathbb{R} \) 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 since

\[
\left| \sum_{k=n+1}^{m} a_k \right| \leq \sum_{k=n+1}^{m} |a_k|, \quad m = n + 1, \ldots
\]

However, the converse is not true as is easily seen through the example

\[
a_n = \frac{(-1)^n}{n}, \quad n = 1, 2, \ldots
\]

A series which is convergent in the usual sense but not absolutely convergent is said to be conditionally convergent.
When the sequence \( a : \mathbb{N}_0 \rightarrow \mathbb{R} \) assumes only non-negative values, i.e., \( a_n \geq 0 \) for all \( n = 1, 2, \ldots \), then the sequence \( s : \mathbb{N}_0 \rightarrow \mathbb{R}_+ \) of partial sums is non-decreasing, so that \( \lim_{n \rightarrow \infty} s_n \) always exists, possibly infinite. When this limit is finite, it is easy to establish the following fact.

**Lemma 13.4.1** For any sequence \( a : \mathbb{N}_0 \rightarrow \mathbb{R} \) whose sequence of partial sums converges in \( \mathbb{R} \), we have \( \lim_{n \rightarrow \infty} a_n = 0 \)

**Proof.** Since the sequence of partial sums \( s : \mathbb{N}_0 \rightarrow \mathbb{R} \) converges in \( \mathbb{R} \), it is a Cauchy sequence: For every \( \varepsilon > 0 \), there exists a finite integer \( n^*(\varepsilon) \) such that

\[
|s_n - s_m| \leq \varepsilon, \quad n, m \geq n^*(\varepsilon).
\]

Selecting \( m = n + 1 \) with \( n \geq n^*(\varepsilon) \), we get \( |a_{n+1}| = |s_n - s_{n+1}| \leq \varepsilon \) whenever \( n \geq n^*(\varepsilon) \), and the conclusion \( \lim_{n \rightarrow \infty} a_n = 0 \) follows.

Many tests exist to check the convergence of series. The most basic one is the Comparison Test given next.

**Theorem 13.4.1** (Comparison Test) Consider two sequences \( a, b : \mathbb{N}_0 \rightarrow \mathbb{R}_+ \) 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 \( \rho \) is the series associated with the sequence \( a : \mathbb{N}_0 \rightarrow \mathbb{R} \) given by

\[
a_n = \rho^n, \quad n = 1, 2, \ldots
\]

It well known that

\[
s_n = a_1 + \ldots + a_n = \begin{cases} 
\frac{\rho}{1-\rho} (1 - \rho^n) & \text{if } \rho \neq 1 \\
 n & \text{if } \rho = 1
\end{cases}
\]
Therefore,
\[
\lim_{n \to \infty} s_n = \frac{\rho}{1 - \rho} \text{ if } |\rho| < 1.
\]
This observation constitutes the basis for two criteria for convergence of series, namely the criteria of Cauchy and d’Alembert, also known as the Root Test and Ratio Test, respectively.

**Theorem 13.4.2 (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 13.4.3 (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 \( 1 < R \).

### 13.5 Power series

In a number of places we shall need to understand the behavior of series that belong to the class of power series. With any sequence \( a : \mathbb{N}_0 \to \mathbb{R} \) we associate the formal power series
\[
\sum_{n=1}^{\infty} a_n z^n, \quad z \in \mathbb{C}.
\]

A natural question arises as to when such formal series are in fact convergent. In particular, we define the domain of convergence of the power series as the set \( \mathcal{C} \) given by
\[
\mathcal{C} = \{ z \in \mathbb{C} : \sum_{n=0}^{\infty} |a_n||z|^n < \infty \}.
\]

This region is determined by the asymptotic behavior of the sequence \( a : \mathbb{N}_0 \to \mathbb{R} \). This is the content of the following well-known result which is a consequence of the Root Test (applied to the sequence \( \{a_n z^n, \ n = 0, 1, \ldots \} \)).

**Theorem 13.5.1** With
\[
R = \limsup_{n \to \infty} \sqrt[n]{|a_n|},
\]
we have $\sum_{n=1}^{\infty} |a_n||z|^n < \infty$ if $|z| < R^{-1}$ and $\sum_{n=1}^{\infty} |a_n||z|^n = \infty$ if $R^{-1} < |z|$.

The open disk $\{ z \in \mathbb{C} : |z| < R^{-1} \}$ is therefore contained in $\mathbb{C}$, and $R^{-1}$ is known as the radius of convergence of the power series.