Stats 310A (Math 230A) Lecture notes

Sourav Chatterjee

Contents

CHAPTER 1

Measures

The mathematical foundation of probability theory if measure theoretic. In this chapter we will build some of the basic measure theoretic tools that are needed for probability theory.

1.1. Measurable spaces

DEFINITION 1.1.1. Let Ω be a set. A σ-algebra $\mathcal{F}$ on Ω is a collection of subsets such that

(1) $\emptyset \in \mathcal{F}$ (where $\emptyset$ denotes the empty set),
(2) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (where A^c denotes the complement of A in Ω), and
(3) if $A_1, A_2, \ldots$ is a countable collection of sets in $\mathcal{F}$, then

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

If the countable union condition is replaced by a finite union condition, then $\mathcal{F}$ is called an algebra instead of a σ-algebra. Note that σ-algebras are also closed under finite unions, since we can append empty sets to convert a finite collection into a countable collection.

DEFINITION 1.1.2. A pair $(\Omega, \mathcal{F})$, where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω, is called a measurable space.

EXERCISE 1.1.3. Prove that a σ-algebra is closed under countable intersections.

EXERCISE 1.1.4. For any set Ω, show that the power set of Ω is a σ-algebra on Ω.

EXERCISE 1.1.5. Prove that the intersection of any arbitrary collection of σ-algebras on a set is a σ-algebra.

EXERCISE 1.1.6. If Ω is a set and $\mathcal{A}$ is any collection of subsets of Ω, show that there is a 'smallest' σ-algebra $\mathcal{F}$ containing $\mathcal{A}$, in the sense that any σ-algebra $\mathcal{G}$ containing $\mathcal{F}$ must also contain $\mathcal{F}$. (Hint: Use the previous two exercises.)

The above exercise motivates the following definition.

DEFINITION 1.1.7. Let Ω be a set and $\mathcal{A}$ be a collection of subsets of Ω. The smallest σ-algebra containing $\mathcal{A}$ is called the σ-algebra generated by $\mathcal{A}$, and is denoted by $\sigma(\mathcal{A})$.

The above definition makes it possible to define the following important class of σ-algebras.

DEFINITION 1.1.8. Let Ω be a set endowed with a topology. The Borel σ-algebra on Ω is the σ-algebra generated by the collection of open subsets of Ω. It is sometimes denoted by $\mathcal{B}(\Omega)$.

In particular, the Borel σ-algebra on the real line $\mathbb{R}$ is the σ-algebra generated by all open subsets of $\mathbb{R}$.

EXERCISE 1.1.9. Prove that the Borel σ-algebra on $\mathbb{R}$ is also generated by the set of all open intervals (or half-open intervals, or closed intervals). (Hint: Show that any open set is a countable union of open — or half-open, or closed — intervals.)

EXERCISE 1.1.10. Show that intervals of the form (x, ∞) also generated $\mathcal{B}(\mathbb{R})$, as do intervals of the form $(-\infty, x)$.

EXERCISE 1.1.11. Let $(\Omega, \mathcal{F})$ be a measurable space and take any $\Omega' \in \mathcal{F}$. Consider the set $\mathcal{F}' := \{A \cap \Omega' : A \in \mathcal{F}\}$. Show that this is a σ-algebra. Moreover, show that if $\mathcal{F}$ is generated by a collection of sets $\mathcal{A}$, then $\mathcal{F}'$ is generated by the collection $\mathcal{A}' := \{A \cap \Omega' : A \in \mathcal{A}\}$. (The σ-algebra $\mathcal{F}'$ is called the restriction of $\mathcal{F}$ to Ω'.)

EXERCISE 1.1.12. As a corollary of the above exercise, show that if Ω is a topological space and Ω' is a Borel subset of Ω endowed with the topology inherited from Ω, then the restriction of $\mathcal{B}(\Omega)$ to Ω' equals the Borel σ-algebra of Ω'.

1.2. Measure spaces

A measurable space endowed with a measure is called a measure space. The definition of measure is as follows.

DEFINITION 1.2.1. Let $(\Omega, \mathcal{F})$ be a measurable space. A measure μ on this space is a function from $\mathcal{F}$ into $[0, \infty]$ such that

(1) $\mu(\emptyset) = 0$, and
(2) if $A_1, A_2, \ldots$ is a countable sequence of disjoint sets in $\mathcal{F}$, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

The triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space.

The second condition is known as the countable additivity condition. Note that finite additivity is also valid, since we can append empty sets to a finite collection to make it countable.

EXERCISE 1.2.2. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. For any $A, B \in \mathcal{F}$ such that $A \subseteq B$, show that $\mu(A) \leq \mu(B)$. For any $A_1, A_2, \ldots \in \mathcal{F}$, show that $\mu(\cup A_i) \leq \sum \mu(A_i)$. (These are known as the monotonicity and countable subadditivity properties of measures. Hint: Rewrite the union as a disjoint union of $B_1, B_2, \ldots$, where $B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1})$. Then use countable additivity.)

DEFINITION 1.2.3. If a measure μ on a measurable space $(\Omega, \mathcal{F})$ satisfies $\mu(\Omega) = 1$, then it is called a probability measure, and the triple $(\Omega, \mathcal{F}, \mu)$ is called a probability space. In this case, elements of $\mathcal{F}$ are often called 'events'.

EXERCISE 1.2.4. If $(\Omega, \mathcal{F}, \mu)$ is a probability space and $A \in \mathcal{F}$, show that $\mu(A^c) = 1 - \mu(A)$.

EXERCISE 1.2.5. If $(\Omega, \mathcal{F}, \mu)$ is a measure space and $A_1, A_2, \ldots \in \mathcal{F}$ is an increasing sequence of events (meaning that $A_1 \subseteq A_2 \subseteq \cdots$), prove that $\mu(\cup A_n) = \lim \mu(A_n)$. Moreover, if μ is a probability measure, and if $A_1, A_2, \ldots$ is a decreasing sequence of events (meaning that $A_1 \supseteq A_2 \supseteq \cdots$), prove that $\mu(\cap A_n) = \lim \mu(A_n)$. Lastly, show that the second assertion need not be true if μ is not a probability measure. (Hint: For the first, rewrite the union as a disjoint union and apply countable additivity. For the second, write the intersection as the complement of a union and apply the first part.)

1.3. Dynkin's π-λ theorem

DEFINITION 1.3.1. Let Ω be a set. A collection $\mathcal{P}$ of subsets of Ω is called a π-system if it is closed under finite intersections.

DEFINITION 1.3.2. Let Ω be a set. A collection $\mathcal{L}$ of subsets of Ω is called a λ-system (or Dynkin system) if $\Omega \in \mathcal{L}$ and $\mathcal{L}$ is closed under taking complements and countable disjoint unions.

EXERCISE 1.3.3. Show that $\mathcal{L}$ is a λ-system if and only if

(1) $\Omega \in \mathcal{L}$,
(2) if $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{L}$, and
(3) if $A_1, A_2, \ldots \in \mathcal{L}$ and $A_i \subseteq A_{i+1}$ for each i, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{L}.$$

LEMMA 1.3.4. *If a λ-system is also a π-system, then it is a σ-algebra.*

PROOF. Let $\mathcal{L}$ be a λ-system which is also a π-system. Then $\mathcal{L}$ is closed under complements by definition, and clearly, $\emptyset \in \mathcal{L}$. Suppose that $A_1, A_2, \ldots \in \mathcal{L}$. Let $B_1 = A_1$, and for each $i \geq 2$, let

$$B_i = A_i \cap A_1^c \cap A_2^c \cap \cdots \cap A_{i-1}^c.$$

Since $\mathcal{L}$ is a λ-system, each A_i^c is in $\mathcal{L}$. Therefore, since $\mathcal{L}$ is also a π-system, each $B_i \in \mathcal{L}$. By construction, $B_1, B_2, \ldots$ are disjoint sets and

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

Thus $\cup A_i \in \mathcal{L}$. This shows that $\mathcal{L}$ is a σ-algebra. □

THEOREM 1.3.5 (Dynkin's π-λ theorem). *Let Ω be a set. Let $\mathcal{P}$ be a π-system of subsets of Ω, and let $\mathcal{L} \supseteq \mathcal{P}$ be a λ-system of subsets of Ω. Then $\mathcal{L} \supseteq \sigma(\mathcal{P})$.*

PROOF. Since the intersection of all λ-systems containing $\mathcal{P}$ is again a λ-system, we may assume that $\mathcal{L}$ is the smallest λ-system containing $\mathcal{P}$.

Take any $A \in \mathcal{P}$. Let

$$\mathcal{G}_A := \{B \in \mathcal{L} : A \cap B \in \mathcal{L}\}. \tag{1.3.1}$$

Then $\mathcal{G}_A \supseteq \mathcal{P}$ since $\mathcal{P}$ is a π-system and $\mathcal{P} \subseteq \mathcal{L}$. Clearly, $\Omega \in \mathcal{G}_A$. If $B \in \mathcal{G}_A$, then $B^c \in \mathcal{L}$ and

$$A \cap B^c = (A^c \cup (A \cap B))^c \in \mathcal{L},$$

since A^c and $A \cap B$ are disjoint elements of $\mathcal{L}$ and $\mathcal{L}$ is a λ-system. Thus, $B^c \in \mathcal{G}_A$. If $B_1, B_2, \ldots$ are disjoint sets in $\mathcal{G}_A$, then

$$A \cap (B_1 \cup B_2 \cup \cdots) = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \in \mathcal{L},$$

again since $\mathcal{L}$ is a λ-system. Thus, $\mathcal{G}_A$ is a λ-system containing $\mathcal{P}$. By the minimality of $\mathcal{L}$, this shows that $\mathcal{G}_A = \mathcal{L}$. In particular, if $A \in \mathcal{P}$ and $B \in \mathcal{L}$, then $A \cap B \in \mathcal{L}$.

Next, for $A \in \mathcal{L}$, let $\mathcal{G}_A$ be defined as in (1.3.1). By the deduction in the previous paragraph, $\mathcal{G}_A \supseteq \mathcal{P}$. As before, $\mathcal{G}_A$ is a λ-system. Thus, $\mathcal{G}_A = \mathcal{L}$. In particular, $\mathcal{L}$ is a π-system. By Lemma 1.3.4, this completes the proof. □

An important corollary of the π-λ theorem is the following result about uniqueness of measures.

THEOREM 1.3.6. *Let $\mathcal{P}$ be a π-system. If μ_1 and μ_2 are measures on $\sigma(\mathcal{P})$ that agree on $\mathcal{P}$, and there is a sequence $A_1, A_2, \ldots \in \mathcal{P}$ such that A_n increases to Ω and $\mu_1(A_n)$ and $\mu_2(A_n)$ are both finite for every n, then $\mu_1 = \mu_2$ on $\sigma(\mathcal{P})$.*

PROOF. Take any $A \in \mathcal{P}$ such that $\mu_1(A) = \mu_2(A) < \infty$. Let

$$\mathcal{L} := \{B \in \sigma(\mathcal{P}) : \mu_1(A \cap B) = \mu_2(A \cap B)\}.$$

Clearly, $\Omega \in \mathcal{L}$. If $B \in \mathcal{L}$, then

$$\begin{aligned} \mu_1(A \cap B^c) &= \mu_1(A) - \mu_1(A \cap B) \\ &= \mu_2(A) - \mu_2(A \cap B) = \mu_2(A \cap B^c), \end{aligned}$$

and hence $B^c \in \mathcal{L}$. If $B_1, B_2, \ldots \in \mathcal{L}$ are disjoint and B is their union, then

$$\begin{aligned}\mu_1(A \cap B) &= \sum_{i=1}^{\infty} \mu_1(A \cap B_i) \\ &= \sum_{i=1}^{\infty} \mu_2(A \cap B_i) = \mu_2(A \cap B),\end{aligned}$$

and therefore $B \in \mathcal{L}$. This shows that $\mathcal{L}$ is a λ-system. Therefore by the π-λ theorem, $\mathcal{L} = \sigma(\mathcal{P})$. In other words, for every $B \in \sigma(\mathcal{P})$ and $A \in \mathcal{P}$ such that $\mu_1(A) < \infty$, $\mu_1(A \cap B) = \mu_2(A \cap B)$. By the given condition, there is a sequence $A_1, A_2, \ldots \in \mathcal{P}$ such that $\mu_1(A_n) < \infty$ for every n and $A_n \uparrow \Omega$. Thus, for any $B \in \sigma(\mathcal{P})$,

$$\mu_1(B) = \lim_{n\to\infty} \mu_1(A_n \cap B) = \lim_{n\to\infty} \mu_2(A_n \cap B) = \mu_2(B).$$

This completes the proof of the theorem. □

Another very useful application of the π-λ theorem is in the proof of the following result.

THEOREM 1.3.7. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, and let $\mathcal{A}$ be an algebra of sets generating $\mathcal{F}$. Then for any $A \in \mathcal{F}$ and any $\epsilon > 0$, there is some $B \in \mathcal{A}$ such that $\mu(A \Delta B) < \epsilon$, where $A\Delta B$ is the symmetric difference of A and B.*

PROOF. Let $\mathcal{G}$ be the collection of all $A \in \mathcal{F}$ for which the stated property holds. Clearly, $\mathcal{A} \subseteq \mathcal{G}$. In particular, $\Omega \in \mathcal{G}$. Take $A \in \mathcal{G}$ and any $\epsilon > 0$. Find $B \in \mathcal{A}$ such that $\mu(A\Delta B) < \epsilon$. Since $A^c \Delta B^c = A\Delta B$, we get $\mu(A^c \Delta B^c) < \epsilon$. Thus, $A^c \in \mathcal{G}$. Finally, suppose that $A_1, A_2, \ldots \in \mathcal{G}$ are disjoint. Let $A := \cup A_i$ and take any $\epsilon > 0$. Since μ is a probability measure, there is some n large enough such that

$$\mu\left(A \setminus \bigcup_{i=1}^{n} A_i\right) < \frac{\epsilon}{2}.$$

For each i, find $B_i \in \mathcal{A}$ such that $\mu(A_i \Delta B_i) < 2^{-i-1}\epsilon$. Let $A' := \cup_{i=1}^n A_i$ and $B' := \cup_{i=1}^n B_i$. Then

$$\mu(A'\Delta B') \le \sum_{i=1}^{n} \mu(A_i \Delta B_i) \le \frac{\epsilon}{2}.$$

Therefore

$$\mu(A\Delta B') \le \mu(A \setminus A') + \mu(A'\Delta B') < \epsilon.$$

Thus, $A \in \mathcal{G}$. This proves that $\mathcal{G}$ is a λ-system. Since any algebra is automatically a π-system, the π-λ theorem implies that $\mathcal{G} = \mathcal{F}$. □

1.4. Outer measures

DEFINITION 1.4.1. Let Ω be any set and let 2^Ω denote its power set. A function $\phi : 2^\Omega \to [0, \infty]$ is called an outer measure if it satisfies the following conditions:

(1) $\phi(\emptyset) = 0$.
(2) $\phi(A) \leq \phi(B)$ whenever $A \subseteq B$.
(3) For any $A_1, A_2, \ldots \subseteq \Omega$,

$$\phi\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \phi(A_i).$$

Note that there is no σ-algebra in the definition of an outer measure. In fact, we will show below that an outer measure generates its own σ-algebra.

DEFINITION 1.4.2. If ϕ is an outer measure on a set Ω, a subset $A \subseteq \Omega$ is called ϕ-measurable if for all $B \subseteq \Omega$,

$$\phi(B) = \phi(B \cap A) + \phi(B \cap A^c).$$

Note that A is ϕ-measurable if and only if

$$\phi(B) \geq \phi(B \cap A) + \phi(B \cap A^c)$$

for every B, since the opposite inequality follows by subadditivity. The following result is the most important fact about outer measures.

THEOREM 1.4.3. *Let Ω be a set and ϕ be an outer measure on Ω. Let $\mathcal{F}$ be the collection of all ϕ-measurable subsets of Ω. Then $\mathcal{F}$ is a σ-algebra and ϕ is a measure on $\mathcal{F}$.*

The proof of Theorem 1.4.3 is divided into a sequence of lemmas.

LEMMA 1.4.4. *The collection $\mathcal{F}$ is an algebra.*

PROOF. Clearly, $\emptyset \in \mathcal{F}$, since $\phi(\emptyset) = 0$. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ by the definition of $\mathcal{F}$. If $A, B \in \mathcal{F}$, let $D := A \cup B$ and note that by subadditivity of ϕ, for any E we have

$$\begin{aligned}
&\phi(E \cap D) + \phi(E \cap D^c) \\
&= \phi((E \cap A) \cup (E \cap B \cap A^c)) + \phi(E \cap A^c \cap B^c) \\
&\leq \phi(E \cap A) + \phi(E \cap B \cap A^c) + \phi(E \cap A^c \cap B^c).
\end{aligned}$$

But since $B \in \mathcal{F}$,

$$\phi(E \cap B \cap A^c) + \phi(E \cap A^c \cap B^c) = \phi(E \cap A^c).$$

Thus,

$$\phi(E \cap D) + \phi(E \cap D^c) \leq \phi(E \cap A) + \phi(E \cap A^c).$$

But since $A \in \mathcal{F}$, the right side equals $\phi(E)$. This completes the proof. $\square$

LEMMA 1.4.5. *If $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint and $E \subseteq \Omega$, then*

$$\phi(E \cap (A_1 \cup \cdots \cup A_n)) = \sum_{i=1}^{n} \phi(E \cap A_i).$$

PROOF. For each j, let $B_j := A_1 \cup \cdots \cup A_j$. Since $A_n \in \mathcal{F}$,

$$\phi(E \cap B_n) = \phi(E \cap B_n \cap A_n) + \phi(E \cap B_n \cap A_n^c).$$

But since $A_1, \ldots, A_n$ are disjoint, $B_n \cap A_n^c = B_{n-1}$ and $B_n \cap A_n = A_n$. Thus,

$$\phi(E \cap B_n) = \phi(E \cap A_n) + \phi(E \cap B_{n-1}).$$

The proof is now completed by induction. □

A consequence of the last two lemmas is the following.

LEMMA 1.4.6. *If $A_1, A_2, \ldots$ is a sequence of sets in $\mathcal{F}$ increasing to a set $A \subseteq \Omega$, then for any $E \subseteq \Omega$,*

$$\phi(E \cap A) \leq \lim_{n\to\infty} \phi(E \cap A_n).$$

PROOF. Let $B_n := A_n \cap (A_1 \cup \cdots \cup A_{n-1})^c$ for each n, so that the sets $B_1, B_2, \ldots$ are disjoint and $A_n = B_1 \cup \cdots \cup B_n$ for each n. By Lemma 1.4.4, $B_n \in \mathcal{F}$ for each n. Thus, by Lemma 1.4.5,

$$\phi(E \cap A_n) = \sum_{i=1}^{n} \phi(E \cap B_i).$$

Consequently,

$$\lim_{n\to\infty} \phi(E \cap A_n) = \sum_{i=1}^{\infty} \phi(E \cap B_i).$$

Since ϕ is countably subadditive (by Lemma 1.5.2), this completes the proof of the lemma. □

We are now ready to prove Theorem 1.4.3.

PROOF OF THEOREM 1.4.3. Let $A_1, A_2, \ldots \in \mathcal{F}$ and let $A := \cup A_i$. For each n, let

$$B_n := \bigcup_{i=1}^{n} A_i.$$

Take any $E \subseteq \Omega$ and any n. By Lemma 1.4.4, $B_n \in \mathcal{F}$ and hence

$$\phi(E) = \phi(E \cap B_n) + \phi(E \cap B_n^c).$$

By monotonicity of ϕ, $\phi(E \cap B_n^c) \geq \phi(E \cap A^c)$. Thus,

$$\phi(E) \geq \phi(E \cap B_n) + \phi(E \cap A^c).$$

By Lemma 1.4.6,

$$\lim_{n\to\infty} \phi(E \cap B_n) \geq \phi(E \cap A).$$

Thus, $A \in \mathcal{F}$. Together with Lemma 1.4.4, this shows that $\mathcal{F}$ is a σ-algebra.

To show that ϕ is a measure on $\mathcal{F}$, take any disjoint collection of sets $A_1, A_2, \ldots \in \mathcal{F}$. Let B be the union of these sets, and let $B_n = A_1 \cup \cdots \cup A_n$. By the monotonicity of ϕ and Lemma 1.4.5,

$$\phi(B) \geq \phi(B_n) = \sum_{i=1}^{n} \phi(A_i).$$

Letting $n \to \infty$, we get $\phi(B) \geq \sum \phi(A_i)$. On the other hand, subadditivity of ϕ gives the opposite inequality. This shows that ϕ is a measure on $\mathcal{F}$ and completes the proof. □

1.5. Carathéodory's extension theorem

Let Ω be a set and $\mathcal{A}$ be an algebra of subsets of Ω. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a measure on $\mathcal{A}$ if

(1) $\mu(\emptyset) = 0$, and
(2) if $A_1, A_2, \ldots$ is a countable collection of disjoint elements of $\mathcal{A}$ such that their union is also in $\mathcal{A}$, then

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

A measure μ on an algebra $\mathcal{A}$ is called σ-finite if there is a countable family of sets $A_1, A_2, \ldots \in \mathcal{A}$ such that $\mu(A_i) < \infty$ for each i, and $\Omega = \cup A_i$.

THEOREM 1.5.1 (Carathéodory's extension theorem). *If $\mathcal{A}$ is an algebra of subsets of a set Ω and μ is a measure on $\mathcal{A}$, then μ has an extension to $\sigma(\mathcal{A})$. Moreover, if μ is σ-finite on $\mathcal{A}$, then the extension is unique.*

The plan of the proof is to construct an outer measure on Ω that agrees with μ on $\mathcal{A}$. Then Theorem 1.4.3 will give the required extension. The outer measure is defined as follows. For each $A \subseteq \Omega$, let

$$\mu^*(A) := \inf\Big\{\sum_{i=1}^{\infty} \mu(A_i) : A_1, A_2, \ldots \in \mathcal{A},\ A \subseteq \bigcup_{i=1}^{\infty} A_i\Big\}.$$

The next lemma establishes that μ^* is an outer measure on Ω.

LEMMA 1.5.2. *The functional μ^* is an outer measure on Ω.*

PROOF. It is obvious from the definition of μ^* that $\mu^*(\emptyset) = 0$ and μ^* is monotone. For proving subadditivity, take any sequence of sets $\{A_i\}_{i=1}^{\infty}$ and let $A := \cup A_i$. Fix some $\epsilon > 0$, and for each i, let $\{A_{ij}\}_{j=1}^{\infty}$ be a collection of elements of $\mathcal{A}$ such that $A_i \subseteq \cup_j A_{ij}$ and

$$\sum_{j=1}^{\infty} \mu(A_{ij}) \leq \mu^*(A_i) + 2^{-i}\epsilon.$$

Then $\{A_{ij}\}_{i,j=1}^{\infty}$ is a countable cover for A, and so

$$\begin{aligned}\mu^*(A) &\leq \sum_{i,j=1}^{\infty} \mu(A_{ij}) \\ &\leq \sum_{i=1}^{\infty}(\mu^*(A_i) + 2^{-i}\epsilon) = \epsilon + \sum_{i=1}^{\infty} \mu^*(A_i).\end{aligned}$$

Since ϵ is arbitrary, this completes the proof of the lemma. □

The next lemma shows that μ^* is a viable candidate for an extension.

LEMMA 1.5.3. *For $A \in \mathcal{A}$, $\mu^*(A) = \mu(A)$.*

PROOF. Take any $A \in \mathcal{A}$. By definition, $\mu^*(A) \leq \mu(A)$. Conversely, take any $A_1, A_2, \ldots \in \mathcal{A}$ such that $A \subseteq \cup A_i$. Then $A = \cup(A \cap A_i)$. Note that each $A \cap A_i \in \mathcal{A}$, and their union is A, which is also in $\mathcal{A}$. It is easy to check that countable subadditivity (Exercise 1.2.2) continues to be valid for measures on algebras, provided that the union belongs to the algebra. Thus, we get

$$\mu(A) \leq \sum_{i=1}^{\infty} \mu(A \cap A_i) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

This shows that $\mu(A) \leq \mu^*(A)$. □

We are now ready to prove Carathéodory's extension theorem.

PROOF OF THEOREM 1.5.1. Let $\mathcal{A}^*$ be the set of all μ^*-measurable sets. By Theorem 1.5.1, we know that $\mathcal{A}^*$ is a σ-algebra and that μ^* is a measure on $\mathcal{A}^*$. We now claim that $\mathcal{A} \subseteq \mathcal{A}^*$. To prove this, take any $A \in \mathcal{A}$ and $E \subseteq \Omega$. Let $A_1, A_2, \ldots$ be any sequence of elements of $\mathcal{A}$ that cover E. Then $\{A \cap A_i\}_{i=1}^{\infty}$ is a cover for $E \cap A$ and $\{A^c \cap A_i\}_{i=1}^{\infty}$ is a cover for $E \cap A^c$. Consequently,

$$\begin{aligned}\mu^*(E \cap A) + \mu^*(E \cap A^c) &\leq \sum_{i=1}^{\infty}(\mu(A \cap A_i) + \mu(A^c \cap A_i)) \\ &= \sum_{i=1}^{\infty} \mu(A_i).\end{aligned}$$

Taking infimum over all choices of $\{A_i\}_{i=1}^{\infty}$, this shows that $\mu^*(E \cap A) + \mu^*(E \cap A^c) \leq \mu^*(E)$, which means that $A \in \mathcal{A}^*$. Thus, $\mathcal{A} \subseteq \mathcal{A}^*$. This proves the existence part of the theorem. If μ is σ-finite on $\mathcal{A}$, then the uniqueness of the extension follows from Theorem 1.3.6. □

1.6. Construction of Lebesgue measure

Let $\mathcal{A}$ be the set of all subsets of $\mathbb{R}$ that are finite disjoint unions of half-open intervals of the form $(a, b] \cap \mathbb{R}$, where $-\infty \le a \le b \le \infty$. Here we write $(a, b] \cap \mathbb{R}$ to ensure that the interval is (a, ∞) if $b = \infty$. If $a = b$, the interval is empty.

EXERCISE 1.6.1. Show that $\mathcal{A}$ is an algebra of subsets of $\mathbb{R}$.

EXERCISE 1.6.2. Show that the algebra $\mathcal{A}$ generates the Borel σ-algebra of $\mathbb{R}$.

Define a functional $\lambda : \mathcal{A} \to \mathbb{R}$ as:

$$\lambda\left(\bigcup_{i=1}^{n}(a_i, b_i] \cap \mathbb{R}\right) := \sum_{i=1}^{n}(b_i - a_i),$$

where remember that $(a_1, b_1], \ldots, (a_n, b_n]$ are disjoint. In other words, λ measures the length of an element of $\mathcal{A}$, as understood in the traditional sense. It is obvious that λ is finitely additive on $\mathcal{A}$ (that is, $\lambda(A_1 \cup \cdots \cup A_n) = \lambda(A_1) + \cdots + \lambda(A_n)$ when $A_1, \ldots, A_n$ are disjoint elements of $\mathcal{A}$). It is also obvious that λ is monotone, that is, $\lambda(A) \le \lambda(B)$ when $A \subseteq B$ (just observe that A is the disjoint union of B and $A \setminus B$, and apply finite additivity).

LEMMA 1.6.3. *For any $A_1, \ldots, A_n \in \mathcal{A}$ and any $A \subseteq A_1 \cup \cdots \cup A_n$, $\lambda(A) \le \sum \lambda(A_i)$.*

PROOF. Let $B_1 = A_1$ and $B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1})$ for $2 \le i \le n$. Then $B_1, \ldots, B_n$ are disjoint and their union is the same as the union of $A_1, \ldots, A_n$. Therefore by the finite additivity and monotonicity of λ,

$$\lambda(A) = \sum_{i=1}^{n}\lambda(A \cap B_i) \le \sum_{i=1}^{n}\lambda(B_i) \le \sum_{i=1}^{n}\lambda(A_i),$$

where the last inequality holds because $B_i \subseteq A_i$ for each i. □

PROPOSITION 1.6.4. *The functional λ defined above is a σ-finite measure on $\mathcal{A}$.*

PROOF. Suppose that an element $A \in \mathcal{A}$ is a countable disjoint union of elements $A_1, A_2, \ldots \in \mathcal{A}$. We have to show that

$$\lambda(A) = \sum_{i=1}^{\infty}\lambda(A_i). \tag{1.6.1}$$

It suffices to show this when $A = (a, b] \cap \mathbb{R}$ and $A_i = (a_i, b_i] \cap \mathbb{R}$ for each i, since each element of $\mathcal{A}$ is a finite disjoint union of such intervals. There is nothing to prove if $a = b$, so assume that $a < b$.

First suppose that $-\infty < a < b < \infty$. Take any $\delta > 0$ such that $a + \delta < b$, and take any $\epsilon > 0$. Then the closed interval $[a + \delta, b]$ is contained

in the union of $(a_i, b_i + 2^{-i}\epsilon)$, $i \geq 1$. To see this, take any $x \in [a+\delta, b]$. Then $x \in (a, b]$, and hence $x \in (a_i, b_i]$ for some i. Thus, $x \in (a_i, b_i + 2^{-i}\epsilon)$.

Since $[a+\delta, b]$ is compact, it is therefore contained in the union of finitely many $(a_i, b_i + 2^{-i}\epsilon)$. Consequently, there exists k such that

$$(a+\delta, b] \subseteq \bigcup_{i=1}^{k} (a_i, b_i + 2^{-i}\epsilon].$$

Thus, by Lemma 1.6.3,

$$b - a - \delta \leq \sum_{i=1}^{k} (b_i + 2^{-i}\epsilon - a_i) \leq \epsilon + \sum_{i=1}^{\infty} (b_i - a_i).$$

Since this holds for any ϵ and δ, we get

$$b - a \leq \sum_{i=1}^{\infty} (b_i - a_i). \tag{1.6.2}$$

On other hand, for any k, finite additivity and monotonicity of λ implies that

$$b - a = \lambda(A) \geq \sum_{i=1}^{k} \lambda(A_i) = \sum_{i=1}^{k} (b_i - a_i).$$

Thus,

$$b - a \geq \sum_{i=1}^{\infty} (b_i - a_i), \tag{1.6.3}$$

which proves (1.6.1) when a and b are finite. If either a or b is infinite, we find finite a', b' such that $(a', b'] \subseteq (a, b] \cap \mathbb{R}$. Repeating the above steps, we arrive at the inequality

$$b' - a' \leq \sum_{i=1}^{\infty} (b_i - a_i).$$

Since this hold for any finite $a' > a$ and $b' < b$, we recover (1.6.2), and (1.6.3) continues to hold as before. This completes the proof of countable additivity of λ. The σ-finiteness is trivial. □

COROLLARY 1.6.5. *The functional λ has a unique extension to a measure on $\mathcal{B}(\mathbb{R})$.*

PROOF. By Exercise 1.6.2, the algebra $\mathcal{A}$ generates the Borel σ-algebra. The existence and uniqueness of the extension now follows by Proposition 1.6.4 and Carathéodory's extension theorem. □

DEFINITION 1.6.6. The unique extension of λ to $\mathcal{B}(\mathbb{R})$ given by Corollary 1.6.5 is called the Lebesgue measure on the real line.

EXERCISE 1.6.7. Define Lebesgue measure on $\mathbb{R}^n$ for general n by considering disjoint unions of products of half-open intervals, and carrying out a similar procedure as above.

1.7. Completion of σ-algebras

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Sometimes, it is convenient if for any set $A \in \mathcal{F}$ with $\mu(A) = 0$, any subset B of A is automatically also in $\mathcal{F}$. This is because in probability theory, we sometimes encounter events of probability zero which are not obviously measurable. If $\mathcal{F}$ has this property, then it is called a complete σ-algebra for the measure μ. If a σ-algebra is not complete, we may want to find a bigger σ-algebra on which μ has an extension, and which is complete for μ. This way, any event and any function that was measurable with respect to the original σ-algebra remains measurable with respect to the new σ-algebra, and we additionally gain the completeness property. The following proposition shows that such a completion can always be obtained.

PROPOSITION 1.7.1. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then there exists a σ-algebra $\mathcal{F}' \supseteq \mathcal{F}$, and an extension of μ to $\mathcal{F}'$, such that $\mathcal{F}'$ is a complete σ-algebra for μ.*

PROOF. Define an outer measure μ^* and a σ-algebra $\mathcal{A}^*$ as in the proof of Carathéodory's extension theorem. Since $\mathcal{F}$ is, in particular, an algebra, the proof of Carathéodory's theorem shows that $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{A}^*$, and μ^* is an extension of μ to $\mathcal{A}^*$. Therefore it suffices to show that $\mathcal{A}^*$ is complete for μ^*. To see this, take any $A \in \mathcal{A}^*$ such that $\mu^*(A) = 0$, and any set $B \subseteq A$. By Lemma 1.5.2, we know that μ^* is a monotone functional. Therefore, for any $D \subseteq \Omega$, $\mu^*(D \cap B) \leq \mu^*(A) = 0$, and $\mu^*(D \cap B^c) \leq \mu^*(D)$. Consequently,

$$\mu^*(D) \geq \mu^*(D \cap B) + \mu^*(D \cap B^c).$$

As observed in the proof of Carathéodory's theorem, this implies that $B \in \mathcal{A}^*$. Thus, $\mathcal{A}^*$ is complete for μ^*. □

The completion of the Borel σ-algebra of $\mathbb{R}$ (or $\mathbb{R}^n$) obtained via the above prescription is called the Lebesgue σ-algebra. Note that Lebesgue measure is actually defined on this larger σ-algebra. We will, however, work with the Borel σ-algebra most of the time. When we say that a function defined on $\mathbb{R}$ is 'measurable', we will mean Borel measurable unless otherwise mentioned. On the other hand, abstract probability spaces on which we will define our random variables (measurable maps), will usually be taken to be complete.

CHAPTER 2

Measurable functions and integration

In this chapter, we will define measurable functions, Lebesgue integration, and the basic properties of integrals.

2.1. Measurable functions

DEFINITION 2.1.1. Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. A function $f : \Omega \to \Omega'$ is called measurable if $f^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{F}'$. Here, as usual, $f^{-1}(A)$ denotes the set of all $x \in \Omega$ such that $f(x) \in A$.

EXERCISE 2.1.2. If $(\Omega_i, \mathcal{F}_i)$ are measurable spaces for $i = 1, 2, 3$, $f : \Omega_1 \to \Omega_2$ is a measurable function, and $g : \Omega_2 \to \Omega_3$ is a measurable function, show that $f \circ g : \Omega_1 \to \Omega_3$ is a measurable function.

The main way to check that a function is measurable is the following lemma.

LEMMA 2.1.3. *Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces and $f : \Omega \to \Omega'$ be a function. Suppose that there is a set $\mathcal{A} \subseteq \mathcal{F}'$ that generates $\mathcal{F}'$, and suppose that $f^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{A}$. Then f is measurable.*

PROOF. It is easy to verify that the set of all $B \subseteq \Omega'$ such that $f^{-1}(B) \in \mathcal{F}$ is a σ-algebra. Since this set contains $\mathcal{A}$, it must also contains the σ-algebra generated by $\mathcal{A}$, which is $\mathcal{F}'$. Thus, f is measurable. □

Essentially all functions that arise in practice are measurable. Let us now see why that is the case by identifying some large classes of measurable functions.

PROPOSITION 2.1.4. *Suppose that Ω and Ω' are topological spaces, and $\mathcal{F}$ and $\mathcal{F}'$ are their Borel σ-algebras. Then any continuous function from Ω into Ω' is measurable.*

PROOF. Use Lemma 2.1.3 with $\mathcal{A} =$ the set of all open subsets of Ω'. □

A combination of Exercise 2.1.2 and Proposition 2.1.4 shows, for example, that sums and products of real-valued measurable functions are measurable, since addition and multiplication are continuous maps from $\mathbb{R}^2$ to $\mathbb{R}$, and if $f, g : \Omega \to \mathbb{R}$ are measurable, then $(f, g) : \Omega \to \mathbb{R}^2$ is measurable with respect to $\mathcal{B}(\mathbb{R}^2)$ (easy to show).

EXERCISE 2.1.5. Following the above sketch, show that sums and products of measurable functions are measurable.

EXERCISE 2.1.6. Show that any right-continuous of left-continuous function $f : \mathbb{R} \to \mathbb{R}$ is measurable.

EXERCISE 2.1.7. Show that any monotone function $f : \mathbb{R} \to \mathbb{R}$ is measurable.

EXERCISE 2.1.8. If Ω is a topological space endowed with its Borel σ-algebra, show that any lower- or upper-semicontinuous $f : \Omega \to \mathbb{R}$ is measurable.

Often, we will have occasion to consider measurable functions that take value in the set $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$, equipped with the σ-algebra generated by all intervals of the form $[a, b]$ where $-\infty \leq a \leq b \leq \infty$.

PROPOSITION 2.1.9. *Let $(\Omega, \mathcal{F})$ be a measurable space and let $\{f_n\}_{n\geq 1}$ be a sequence of measurable functions from Ω into $\mathbb{R}^*$. Let $g(\omega) := \inf_{n\geq 1} f_n(\omega)$ and $h(\omega) := \sup_{n\geq 1} f_n(\omega)$ for $\omega \in \Omega$. Then g and h are also measurable functions.*

PROOF. For any $t \in \mathbb{R}^*$, $g(\omega) \geq t$ if and only if $f_n(\omega) \geq t$ for all n. Thus,

$$g^{-1}([t, \infty]) = \bigcap_{n=1}^{\infty} f_n^{-1}([t, \infty]),$$

which shows that $g^{-1}([t, \infty]) \in \mathcal{F}$. It is straightforward to verify that sets of the form $[t, \infty]$ generate the σ-algebra of $\mathbb{R}^*$. Thus, g is measurable. The proof for h is similar. □

The following exercises are useful consequences of Proposition 2.1.9.

EXERCISE 2.1.10. If $\{f_n\}_{n\geq 1}$ is a sequence of $\mathbb{R}^*$-valued measurable functions defined on the same measure space, show that $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$ are also measurable. (Hint: Write the lim sup as an infimum of suprema and the lim inf as a supremum of infima.)

EXERCISE 2.1.11. If $\{f_n\}_{n\geq 1}$ is a sequence of $\mathbb{R}^*$-valued measurable functions defined on the same measure space, and $f_n \to f$ pointwise, show that f is measurable. (Hint: Under the given conditions, $f = \limsup_{n\to\infty} f_n$.)

EXERCISE 2.1.12. If $\{f_n\}_{n\geq 1}$ is a sequence of $[0, \infty]$-valued measurable functions defined on the same measure space, show that $\sum f_n$ is measurable.

EXERCISE 2.1.13. If $\{f_n\}_{n\geq 1}$ is a sequence of measurable $\mathbb{R}^*$-valued functions defined on the same measurable space, show that the set of all ω where $\lim f_n(\omega)$ exists is a measurable set.

It is sometimes useful to know that Exercise 2.1.11 has a generalization to functions taking value in arbitrary separable metric spaces. In that setting, however, the proof using lim sup does not work, and a different argument is needed. This is given in the proof of the following result.

PROPOSITION 2.1.14. *Let $(\Omega, \mathcal{F})$ be a measurable space and let S be a separable metric space endowed with its Borel σ-algebra. If $\{f_n\}_{n\geq 1}$ is a sequence of measurable functions from Ω into S that converge pointwise to a limit function f, then f is also measurable.*

PROOF. Let $B(x,r)$ denote the open ball with center x and radius r in S. Since S is separable, any open set is a countable union of open balls. Therefore by Lemma 2.1.3 it suffices to check that $f^{-1}(B) \in \mathcal{F}$ for any open ball B. Take such a ball $B(x,r)$. For any $\omega \in \Omega$, if $f(\omega) \in B(x,r)$, then there is some large enough integer k such that $f(\omega) \in B(x, r-k^{-1})$. Since $f_n(\omega) \to f(\omega)$, this implies that $f_n(\omega) \in B(x, r-k^{-1})$ for all large enough n. On the other hand, if there exists $k \geq 1$ such that $f_n(\omega) \in B(x, r-k^{-1})$ for all large enough n, then $f(\omega) \in B(x,r)$. Thus, we have shown that $f(\omega) \in B(x,r)$ if and only if there is some integer $k \geq 1$ such that $f_n(\omega) \in B(x, r-k^{-1})$ for all large enough n. In set theoretic notation, this statement can be written as

$$f^{-1}(B(x,r)) = \bigcup_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} f_n^{-1}(B(x, r-k^{-1})).$$

Since each f_n is measurable, and σ-algebras are closed under countable unions and intersections, this shows that $f^{-1}(B(x,r)) \in \mathcal{F}$, completing the proof. □

Measurable maps generate σ-algebras of their own, which are important for various purposes.

DEFINITION 2.1.15. Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces and let $f : \Omega \to \Omega'$ be a measurable function. The σ-algebra generated by f is defined as

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

EXERCISE 2.1.16. Verify that $\sigma(f)$ in the above definition is indeed a σ-algebra.

2.2. Lebesgue integration

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. This space will be fixed throughout this section. Let $f : \Omega \to \mathbb{R}^*$ be a measurable function, where $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, \infty\}$, as defined in the previous section. Our goal in this section is to define the Lebesgue integral

$$\int_\Omega f(\omega) d\mu(\omega).$$

The definition comes in several steps. For $A \in \mathcal{F}$, define the indicator function 1_A as

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Suppose that

$$f = \sum_{i=1}^{n} a_i 1_{A_i} \tag{2.2.1}$$

for some $a_1, \ldots, a_n \in [0, \infty)$ and disjoint sets $A_1, \ldots, A_n \in \mathcal{F}$. Such functions are called 'nonnegative simple functions'. For a nonnegative simple function f as above, define

$$\int_\Omega f(\omega) d\mu(\omega) := \sum_{i=1}^{n} a_i \mu(A_i). \tag{2.2.2}$$

A subtle point here is that the same simple function can have many different representations like (2.2.1). It is not difficult to show that all of them yield the same answer in (2.2.2).

Next, take any measurable $f : \Omega \to [0, \infty]$. Let $\mathrm{SF}_+(f)$ be the set of all nonnegative simple functions g such that $g(\omega) \leq f(\omega)$ for all ω. Define

$$\int_\Omega f(\omega) d\mu(\omega) := \sup_{g \in \mathrm{SF}_+(f)} \int_\Omega g(\omega) d\mu(\omega).$$

It is not difficult to prove that if f is a nonnegative simple function, then this definition gives the same answer as (2.2.2), so there is no inconsistency.

Finally, take an arbitrary measurable $f : \Omega \to \mathbb{R}^*$. Define $f_+(\omega) = \max\{f(\omega), 0\}$ and $f_-(\omega) = -\min\{f(\omega), 0\}$. Then f_+ and f_- are nonnegative measurable functions (easy to show), and $f = f_+ - f_-$. We say that the integral of f is defined when the integrals of at least one of f_+ and f_- is finite. In this situation, we define

$$\int_\Omega f(\omega) d\mu(\omega) := \int_\Omega f_+(\omega) d\mu(\omega) - \int_\Omega f_-(\omega) d\mu(\omega).$$

Often, we will simply write $\int f d\mu$ for the integral of f.

Definition 2.2.1. A measurable function $f : \Omega \to \mathbb{R}^*$ is called integrable if $\int f_+ d\mu$ and $\int f_- d\mu$ are both finite.

Note that for the integral $\int f d\mu$ to be defined, the function f need not be integrable. As defined above, it suffices to have at least one of $\int f_+ d\mu$ and $\int f_- d\mu$ finite.

Exercise 2.2.2. Show that if f is an integrable function, then $\{\omega : f(\omega) = \infty \text{ or } -\infty\}$ is a set of measure zero.

Sometimes, we will need to integrate a function f over a measurable subset $S \subseteq \Omega$ rather than the whole of Ω. This is defined simply by making f zero outside S and integrating the resulting function. That is, we define

$$\int_S f d\mu := \int_\Omega f 1_S d\mu,$$

provided that the right side is defined. Here and everywhere else, we use the convention $\infty \cdot 0 = 0$, which is a standard convention in measure theory and probability. The integral over S can also be defined in a different way, by considering S itself as a set endowed with a σ-algebra and a measure. To be precise, let $\mathcal{F}_S$ be the restriction of $\mathcal{F}$ to S, that is, the set of all subsets of S that belong to $\mathcal{F}$. Let μ_S be the restriction of μ to $\mathcal{F}_S$. Let f_S be the restriction of f to S. It is easy to see that $(S, \mathcal{F}_S, \mu_S)$ is a measure space and f_S is a measurable function from this space into $\mathbb{R}^*$. The integral of f over the set S can then be defined as the integral of f_S on the measure space $(S, \mathcal{F}_S, \mu_S)$, provided that the integral exists. When $S = \emptyset$, the integral can be defined to be zero.

EXERCISE 2.2.3. Show that the two definitions of $\int_S f d\mu$ discussed above are the same, in the sense that one is defined if and only if the other is defined, and in that case they are equal.

2.3. The monotone convergence theorem

The monotone convergence theorem is a fundamental result of measure theory. We will prove this result in this section. First, we need two lemmas. Throughout, $(\Omega, \mathcal{F}, \mu)$ denotes a measure space.

LEMMA 2.3.1. *If $f, g : \Omega \to [0, \infty]$ are two measurable functions such that $f \leq g$ everywhere, then $\int f d\mu \leq \int g d\mu$.*

PROOF. If $s \in \mathrm{SF}_+(f)$, then s is also in $\mathrm{SF}_+(g)$. Thus, the definition of $\int g d\mu$ takes supremum over a larger set than $\int f d\mu$. This proves the inequality. □

LEMMA 2.3.2. *Let $s : \Omega \to [0, \infty)$ be a measurable simple function. For each $S \in \mathcal{F}$, let $\nu(S) := \int_S s d\mu$. Then ν is a measure on $(\Omega, \mathcal{F})$.*

PROOF. Suppose that

$$s = \sum_{i=1}^{n} a_i 1_{A_i}.$$

Since $\nu(\emptyset) = 0$ by definition, it suffices to show countable additivity of ν. Let $S_1, S_2, \ldots$ be a sequence of disjoint sets in $\mathcal{F}$, and let S be their union.

Then

$$\begin{aligned}\nu(S) &= \sum_{i=1}^{n} a_i \mu(A_i \cap S) \\ &= \sum_{i=1}^{n} a_i \Big(\sum_{j=1}^{\infty} \mu(A_i \cap S_j) \Big).\end{aligned}$$

Since an infinite series with nonnegative summands may be rearranged any way we like without altering the result, this gives

$$\nu(S) = \sum_{j=1}^{\infty} \sum_{i=1}^{n} a_i \mu(A_i \cap S_j) = \sum_{j=1}^{\infty} \nu(S_j).$$

This proves the countable additivity of ν. □

THEOREM 2.3.3 (Monotone convergence theorem). *Let $\{f_n\}_{n\geq 1}$ be a sequence of measurable functions from Ω into $[0,\infty]$, which are increasing pointwise to a limit function f. Then f is measurable, and $\int f d\mu = \lim \int f_n d\mu$.*

PROOF. The measurability of f has already been established earlier (Exercise 2.1.11). Since $f \geq f_n$ for every n, Lemma 2.3.1 gives $\int f d\mu \geq \lim \int f_n d\mu$, where the limit on the right exists because the integrals on the right are increasing with n. For the opposite inequality, take any $s \in \mathrm{SF}_+(f)$. Define

$$\nu(S) := \int_S s d\mu,$$

so that by Lemma 2.3.2, ν is a measure on Ω. Take any $\alpha \in (0,1)$. Let

$$S_n := \{\omega : \alpha s(\omega) \leq f_n(\omega)\}.$$

It is easy to see that these sets are measurable and increasing with n. Moreover, note that any $\omega \in \Omega$ belongs to S_n for all sufficiently large n, because $f_n(\omega)$ increases to $f(\omega)$ and $\alpha s(\omega) < f(\omega)$ (unless $f(\omega) = 0$, in which case $\alpha s(\omega) = f_n(\omega) = 0$ for all n). Thus, Ω is the union of the increasing sequence $S_1, S_2, \ldots$. Since ν is a measure, this shows that

$$\int s d\mu = \nu(\Omega) = \lim_{n\to\infty} \nu(S_n) = \lim_{n\to\infty} \int_{S_n} s d\mu.$$

But $\alpha s \leq f_n$ on S_n. Moreover, it is easy to see that

$$\int_{S_n} \alpha s d\mu = \alpha \int_{S_n} s d\mu$$

since s is a simple function. Thus, by Lemma 2.3.1,

$$\alpha \int_{S_n} s d\mu = \int_{S_n} \alpha s d\mu \leq \int_{S_n} f_n d\mu \leq \int_{\Omega} f_n d\mu.$$

Combining the last two steps, we get $\alpha \int s d\mu \leq \lim \int f_n d\mu$. Since $\alpha \in (0,1)$ is arbitrary, this shows that $\int s d\mu \leq \lim \int f_n d\mu$. Taking supremum over s, we get the desired result. □

EXERCISE 2.3.4. Produce a counterexample to show that the nonnegativity condition in the monotone convergence theorem cannot be dropped.

The following exercise has many applications in probability theory.

EXERCISE 2.3.5. If $f_1, f_2, \ldots$ are measurable functions from Ω into $[0,\infty]$, show that $\int \sum f_i d\mu = \sum \int f_i d\mu$.

The next exercise generalizes Lemma 2.3.2.

EXERCISE 2.3.6. Let $f : \Omega \to [0,\infty]$ be a measurable function. Show that the functional $\nu(S) := \int_S f d\mu$ defined on $\mathcal{F}$ is a measure.

The following result is often useful in applications of the monotone convergence theorem.

PROPOSITION 2.3.7. *Given any measurable $f : \Omega \to [0,\infty]$, there is a sequence of nonnegative simple functions increasing pointwise to f.*

PROOF. Take any n. If $k2^{-n} \leq f(\omega) < (k+1)2^{-n}$ for some integer $0 \leq k < n2^n$, let $f_n(\omega) = k2^{-n}$. If $f(\omega) \geq n$, let $f_n(\omega) = n$. It is easy to check that the sequence $\{f_n\}_{n\geq 1}$ is a sequence of nonnegative simple functions that increase pointwise to f. □

2.4. Linearity of the Lebesgue integral

A basic result about Lebesgue integration, which is not quite obvious from the definition, is that the map $f \mapsto \int f d\mu$ is linear. This is a little surprising, because the Lebesgue integral is defined as a supremum, and functionals that are defined as suprema or infima are rarely linear. In the following, when defining the sum of two functions, we adopt the convention that $\infty - \infty = 0$. Typically such a convention can lead to inconsistencies, but if the functions are integrable then such occurrences will only take place on sets of measure zero and will not cause any problems.

PROPOSITION 2.4.1. *If f and g are two integrable functions from Ω into $\mathbb{R}^*$, then for any $\alpha, \beta \in \mathbb{R}$, the function $\alpha f + \beta g$ is integrable and $\int(\alpha f + \beta g)d\mu = \alpha \int f d\mu + \beta \int g d\mu$. Moreover, if f and g are measurable functions from Ω into $[0,\infty]$ (but not necessarily integrable), then $\int(f+g)d\mu = \int f d\mu + \int g d\mu$, and for any $\alpha \in \mathbb{R}$, $\int \alpha f d\mu = \alpha \int f d\mu$.*

PROOF. It is easy to check that additivity of the integral holds for nonnegative simple functions. Take any measurable $f, g : \Omega \to [0,\infty]$. Then by Proposition 2.3.7, there exist sequences of nonnegative simple functions

$\{u_n\}_{n\geq 1}$ and $\{v_n\}_{n\geq 1}$ increasing to f and g pointwise. Then $u_n + v_n$ increases to $f+g$ pointwise. Thus, by the linearity of integral for nonnegative simple functions and the monotone convergence theorem,

$$\begin{aligned}\int (f+g)d\mu &= \lim_{n\to\infty} \int (u_n + v_n)d\mu \\ &= \lim_{n\to\infty} \left(\int u_n d\mu + \int v_n d\mu \right) = \int f d\mu + \int g d\mu.\end{aligned}$$

Next, take any $\alpha \geq 0$ and any measurable $f : \Omega \to [0,\infty]$. Any element of $\mathrm{SF}_+(\alpha f)$ must be of the form αg for some $g \in \mathrm{SF}_+(f)$. Thus

$$\int \alpha f d\mu = \sup_{g\in \mathrm{SF}_+(f)} \int \alpha g d\mu = \alpha \sup_{g\in \mathrm{SF}_+(f)} \int g d\mu = \alpha \int f d\mu.$$

If $\alpha \leq 0$, then $(\alpha f)_+ \equiv 0$ and $(\alpha f)_- = -\alpha f$. Thus,

$$\int \alpha f d\mu = -\int (-\alpha f) d\mu = \alpha \int f d\mu.$$

This completes the proofs of the assertions about nonnegative functions. Next, take any integrable f and g. It is easy to see that a consequence of integrability is that set where f or g take infinite values is a set of measure zero. The only problem in defining the function $f+g$ is that at some ω, it may happen that one of $f(\omega)$ and $g(\omega)$ is ∞ and the other is $-\infty$. At any such point, define $(f+g)(\omega) = 0$. Then $f+g$ is defined everywhere, is measurable, and $(f+g)_+ \leq f_+ + g_+$. Therefore, by the additivity of integration for nonnegative functions and the integrability of f and g, $\int (f+g)_+ d\mu$ is finite. Similarly, $\int (f+g)_- d\mu$ is finite. Thus, $f+g$ is integrable. Now,

$$(f+g)_+ - (f+g)_- = f+g = f_+ - f_- + g_+ - g_-,$$

which can be written as

$$(f+g)_+ + f_- + g_- = (f+g)_- + f_+ + g_+.$$

Note that the above identity holds even if one or more of the summands on either side are infinity. Thus, by additivity of integration for nonnegative functions,

$$\int (f+g)_+ d\mu + \int f_- d\mu + \int g_- d\mu = \int (f+g)_- d\mu + \int f_+ d\mu + \int g_+ d\mu,$$

which rearranges to give $\int (f+g)d\mu = \int f d\mu + \int g d\mu$. Finally, if f is integrable and $\alpha \geq 0$, then $(\alpha f)_+ = \alpha f_+$ and $(\alpha f)_- = \alpha f_-$, which shows that αf is integrable and $\int \alpha f d\mu = \alpha \int f d\mu$. Similarly, $\alpha \leq 0$, then $(\alpha f)_+ = -\alpha f_-$ and $(\alpha f)_- = -\alpha f_+$, and the proof can be completed as before. □

An immediate consequence of Proposition 2.4.1 is that if f and g are measurable real-valued functions such that $f \geq g$ everywhere, then $\int f d\mu \geq \int g d\mu$. This can be seen easily by writing $f = (f-g)+g$, and observing that $f-g$ is nonnegative. Another immediate consequence is that integrability

of f is equivalent to the condition that $\int |f| d\mu$ is finite, since $|f| = f_+ + f_-$. Finally, another inequality that we will often use, which is an easy consequence of the triangle inequality, the definition of the Lebesgue integral, and the fact that $|f| = f_+ + f_-$, is that for any $f : \Omega \to \mathbb{R}^*$ such that $\int f d\mu$ is defined, $|\int f d\mu| \leq \int |f| d\mu$.

At this point, the reader may wonder about the connection between the Lebesgue integral defined above and the Riemann integral taught in undergraduate analysis classes. The following exercises clarify the relationship between the two.

EXERCISE 2.4.2. Let $[a, b]$ be a finite interval, and let $f : [a, b] \to \mathbb{R}$ be a bounded measurable function. If f is Riemann integrable, show that it is also Lebesgue integrable and that the two integrals are equal.

EXERCISE 2.4.3. Give an example of a Lebesgue integrable function on a finite interval that is not Riemann integrable.

EXERCISE 2.4.4. Generalize Exercise 2.4.2 to higher dimensions.

EXERCISE 2.4.5. Consider the integral

$$\int_0^\infty \frac{\sin x}{x} dx.$$

Show that this makes sense as a limit of integrals (both Lebesgue and Riemann) from 0 to a as $a \to \infty$. However, show that the above integral is not defined as an integral in the Lebesgue sense.

Lebesgue integration with respect to the Lebesgue measure on $\mathbb{R}$ (or $\mathbb{R}^n$) shares many properties of Riemann integrals. The following is one example.

EXERCISE 2.4.6. Let $f : \mathbb{R} \to \mathbb{R}$ be a Lebesgue integrable function. Take any $a \in \mathbb{R}$ and let $g(x) := f(x + a)$. Then show that $\int g d\lambda = \int f d\lambda$, where λ is Lebesgue measure. (Hint: First prove this for simple functions.)

2.5. Fatou's lemma and dominated convergence

In this section we will establish two results about sequences of integrals that are widely used in probability theory. Throughout, $(\Omega, \mathcal{F}, \mu)$ is a measure space.

THEOREM 2.5.1 (Fatou's lemma). *Let $\{f_n\}_{n\geq 1}$ be a sequence of measurable functions from Ω into $[0, \infty]$. Then $\int (\liminf f_n) d\mu \leq \liminf \int f_n d\mu$.*

PROOF. For each n, let $g_n := \inf_{m\geq n} f_m$. Then, as $n \to \infty$, g_n increases pointwise to $g := \liminf_{n\to\infty} f_n$. Moreover, $g_n \leq f_n$ everywhere and g_n and g are measurable by Exercise 2.1.10. Therefore by the monotone convergence theorem,

$$\int g d\mu = \lim_{n\to\infty} \int g_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu,$$

which completes the proof. □

THEOREM 2.5.2 (Dominated convergence theorem). *Let $\{f_n\}_{n\geq 1}$ be a sequence of measurable functions from Ω into $\mathbb{R}^*$ that converge pointwise to a limit function f. Suppose that there exists an integrable function h such that for each n, $|f_n| \leq h$ everywhere. Then $\lim \int f_n d\mu = \int f d\mu$. Moreover, we also have the stronger result $\lim \int |f_n - f| d\mu = 0$.*

PROOF. We know that f is measurable since it is the limit of a sequence of measurable functions (Exercise 2.1.11). Moreover, $|f| \leq h$ everywhere. Thus, f_n+h and $f+h$ are nonnegative integrable functions and $f_n+h \to f+h$ pointwise. Therefore by Fatou's lemma, $\int (f+h)d\mu \leq \liminf \int (f_n+h)d\mu$. Since all integrals are finite (due to the integrability of h and the fact that $|f_n|$ and $|f|$ are bounded by h), this gives $\int f d\mu \leq \liminf \int f_n d\mu$.

Now replace f_n by $-f_n$ and f by $-f$. All the conditions still hold, and therefore the same deduction shows that $\int (-f)d\mu \leq \liminf \int (-f_n)d\mu$, which is the same as $\int f d\mu \geq \limsup \int f_n d\mu$. Thus, we get $\int f d\mu = \lim \int f_n d\mu$.

Finally, to show that $\lim \int |f_n - f| d\mu = 0$, observe that $|f_n - f| \to 0$ pointwise, and $|f_n - f|$ is bounded by the integrable function $2h$ everywhere. Then apply the first part. □

EXERCISE 2.5.3. Produce a counterexample to show that the nonnegativity condition in Fatou's lemma cannot be dropped.

EXERCISE 2.5.4. Produce a counterexample to show that the domination condition of the dominated convergence theorem cannot be dropped.

A very important application of the dominated convergence theorem is in giving conditions for differentiating under the integral sign.

PROPOSITION 2.5.5. *Let I be an open subset of $\mathbb{R}$, and let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $f : I \times \Omega \to \mathbb{R}$ satisfies the following conditions:*

(i) *$f(x, \omega)$ is an integrable function of ω for each $x \in I$,*
(ii) *for all $\omega \in \Omega$, the derivative f_x of f with respect to x exists for all $x \in I$, and*
(iii) *there is an integrable function $h : \Omega \to \mathbb{R}$ such that $|f_x(x, \omega)| \leq h(\omega)$ for all $x \in I$ and $\omega \in \Omega$.*

Then for all $x \in I$,

$$\frac{d}{dx} \int_\Omega f(x, \omega) d\mu(\omega) = \int_\Omega f_x(x, \omega) d\mu(\omega).$$

PROOF. Take any $x \in I$ and a sequence $x_n \to x$. Without loss of generality, assume that $x_n \neq x$ and $x_n \in I$ for each n. Define

$$g_n(\omega) := \frac{f(x, \omega) - f(x_n, \omega)}{x - x_n}.$$

Since the derivative of f with respect to x is uniformly bounded by the function h, it follows that $|g_n(\omega)| \leq h(\omega)$. Since h is integrable and $g_n(\omega) \to f_x(x,\omega)$ for each ω, the dominated convergence theorem gives us

$$\lim_{n\to\infty} \int_\Omega g_n(\omega)d\mu(\omega) = \int_\Omega f_x(x,\omega)d\mu(\omega).$$

But

$$\int_\Omega g_n(\omega)d\mu(\omega) = \frac{\int_\Omega f(x,\omega)d\mu(\omega) - \int_\Omega f(x_n,\omega)d\mu(\omega)}{x - x_n},$$

which proves the claim. □

A slight improvement of Proposition 2.5.5 is given in Exercise 2.6.8 later. Similar slight improvements of the monotone convergence theorem and the dominated convergence theorem are given in Exercise 2.6.5.

2.6. The concept of almost everywhere

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. An event $A \in \mathcal{F}$ is said to 'happen almost everywhere' if $\mu(A^c) = 0$. Almost everywhere is usually abbreviated as a.e. Sometimes, in probability theory, we say 'almost surely (a.s.)' instead of a.e. When the measure μ may not be specified from the context, we say μ-a.e. If f and g are two functions such that the $\{\omega : f(\omega) = g(\omega)\}$ is an a.e. event, we say that $f = g$ a.e. Similarly, if $\{f_n\}_{n\geq 1}$ is a sequence of functions such that $\{\omega : \lim_{n\to\infty} f_n(\omega) = f(\omega)\}$ is an a.e. event, we say $f_n \to f$ a.e. We have already seen an example of a class of a.e. events in Exercise 2.2.2. The following is another result of the same type.

PROPOSITION 2.6.1. *Let $f : \Omega \to [0,\infty]$ be a measurable function. Then $\int f d\mu = 0$ if and only if $f = 0$ a.e.*

PROOF. First suppose that $\mu(\{\omega : f(\omega) > 0\}) = 0$. Take any nonnegative measurable simple function g such that $g \leq f$ everywhere. Then $g = 0$ whenever $f = 0$. Thus, $\mu(\{\omega : g(\omega) > 0\}) = 0$. Since g is a simple function, this implies that $\int g d\mu = 0$. Taking supremum over all such g gives $\int f d\mu = 0$. Conversely, suppose that $\mu(\{\omega : f(\omega) > 0\}) > 0$. Then

$$\begin{aligned}\mu(\{\omega : f(\omega) > 0\}) &= \mu\Big(\bigcup_{n=1}^\infty \{\omega : f(\omega) > n^{-1}\}\Big) \\ &= \lim_{n\to\infty} \mu(\{\omega : f(\omega) > n^{-1}\}) > 0,\end{aligned}$$

where the second equality follows from the observation that the sets in the union form an increasing sequence. Therefore for some n, $\mu(A_n) > 0$, where $A_n := \{\omega : f(\omega) > n^{-1}\}$. But then

$$\int f d\mu \geq \int f 1_{A_n} d\mu \geq \int n^{-1} 1_{A_n} d\mu = n^{-1}\mu(A_n) > 0.$$

This completes the proof of the proposition. □

The following exercises are easy consequences of the above proposition.

EXERCISE 2.6.2. If f, g are integrable functions on Ω such that $f = g$ a.e., then show that $\int f d\mu = \int g d\mu$ and $\int |f - g| d\mu = 0$. Conversely, show that if $\int |f - g| d\mu = 0$, then $f = g$ a.e.

EXERCISE 2.6.3. Suppose that f and g are two integrable functions on Ω such that $f \geq g$ a.e. Then $\int f d\mu \geq \int g d\mu$, and equality holds if and only if $f = g$ a.e.

Broadly speaking, the idea is that 'almost everywhere' and 'everywhere' can be treated as basically the same thing. There are some exceptions to this rule of thumb, but in most situations it is valid. Often, when a function or a limit is defined almost everywhere, we will treat it as being defined everywhere by defining it arbitrarily (for example, equal to zero) on the set where it is undefined. The following exercises show how to use this rule of thumb to get slightly better versions of the convergence theorems of this chapter.

EXERCISE 2.6.4. If $\{f_n\}_{n \geq 1}$ is a sequence of measurable functions and f is a function such that $f_n \to f$ a.e., show that there is a measurable function g such that $g = f$ a.e. Moreover, if the σ-algebra is complete, show that f itself is measurable.

EXERCISE 2.6.5. Show that in the monotone convergence theorem and the dominated convergence theorem, it suffices to have $f_n \to f$ a.e. and f measurable.

EXERCISE 2.6.6. Let $f : \Omega \to I$ be a measurable function, where I is an interval in $\mathbb{R}^*$. The interval I is allowed to be finite or infinite, open, closed or half-open. In all cases, show that $\int f d\mu \in I$ if μ is a probability measure. (Hint: Use Exercise 2.6.3.)

EXERCISE 2.6.7. If $f_1, f_2, \ldots$ are measurable functions from Ω into $\mathbb{R}^*$ such that $\sum \int |f_i| d\mu < \infty$, then show that $\sum f_i$ exists a.e., and $\int \sum f_i d\mu = \sum \int f_i d\mu$.

EXERCISE 2.6.8. Show that in Proposition 2.5.5, every occurrence of 'for all $\omega \in \Omega$' can be replaced by 'for almost all $\omega \in \Omega$'.

CHAPTER 3

Product spaces

This chapter is about the construction of product measure spaces. Product measures are necessary for defining sequences of independent random variables, that form the backbone of many probabilistic models.

3.1. Finite dimensional product spaces

Let $(\Omega_1, \mathcal{F}_1), \ldots, (\Omega_n, \mathcal{F}_n)$ be measurable spaces. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be the Cartesian product of $\Omega_1, \ldots, \Omega_n$. The product σ-algebra $\mathcal{F}$ on Ω is defined as the σ-algebra generated by sets of the form $A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$. It is usually denoted by $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

PROPOSITION 3.1.1. *Let Ω and $\mathcal{F}$ be as above. If each Ω_i is endowed with a σ-finite measure μ_i, then there is a unique measure μ on Ω which satisfies*

$$\mu(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mu_i(A_i) \tag{3.1.1}$$

for each $A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n$.

PROOF. It is easy to check that sets of the form $A_1 \times \cdots \times A_n$ (sometimes called 'rectangles') form an algebra. Therefore by Carathéodory's theorem, it suffices to show that μ defined through (3.1.1) is a countably additive measure on this algebra.

We will prove this by induction on n. It is true for $n = 1$ by the definition of a measure. Suppose that it holds for $n - 1$. Take any rectangular set $A_1 \times \cdots \times A_n$. Suppose that this set is a disjoint union of $A_{i,1} \times \cdots \times A_{i,n}$ for $i = 1, 2, \ldots$, where $A_{i,j} \in \mathcal{F}_j$ for each i and j.

Take any $x \in A_1 \times \cdots \times A_{n-1}$. Let I be the collection of indices i such that $x \in A_{i,1} \times \cdots \times A_{i,n-1}$. Take any $y \in A_n$. Then $(x, y) \in A_1 \times \cdots \times A_n$, and hence $(x, y) \in A_{i,1} \cdots \times \cdots A_{i,n}$ for some i. In particular, $x \in A_{i,1} \times \cdots \times A_{i,n-1}$ and hence $i \in I$. Also, $y \in A_{i,n}$. Thus,

$$A_n \subseteq \bigcup_{i \in I} A_{i,n}.$$

On the other hand, if $y \in A_{i,n}$ for some $i \in I$, then $(x, y) \in A_{i,1} \times \cdots \times A_{i,n}$. Thus, $(x, y) \in A_1 \times \cdots \times A_n$, and therefore $y \in A_n$. This shows that A_n is

the disjoint union of $A_{i,n}$ for $i \in I$, and so

$$\mu_n(A_n) = \sum_{i \in I} \mu_n(A_{i,n}).$$

Finally, if $x \notin A_1 \times \cdots \times A_{n-1}$ and $x \in A_{i,1} \times \cdots \times A_{i,n-1}$ for some i, then $A_{i,n}$ must be empty, because otherwise $(x, y) \in A_{i,1} \times \cdots \times A_{i,n}$ for any $y \in A_{i,n}$ and so $(x, y) \in A_1 \times \cdots \times A_n$, which implies that $x \in A_1 \times \cdots \times A_{n-1}$. Therefore we have proved that

$$1_{A_1 \times \cdots \times A_{n-1}}(x)\mu_n(A_n) = \sum_{i=1}^{\infty} 1_{A_{i,1} \times \cdots \times A_{i,n-1}}(x)\mu_n(A_{i,n}).$$

Let $\mu' := \mu_1 \times \cdots \times \mu_{n-1}$, which exists by the induction hypothesis. Integrating both sides with respect to the measure μ' on $\Omega_1 \times \cdots \times \Omega_{n-1}$, and applying Exercise 2.3.5 to the right, we get

$$\mu'(A_1 \times \cdots \times A_{n-1})\mu_n(A_n) = \sum_{i=1}^{\infty} \mu'(A_{i,1} \times \cdots \times A_{i,n-1})\mu_n(A_{i,n}).$$

The desired result now follows by applying the induction hypothesis to μ' on both sides. □

The measure μ of Proposition 3.1 is called the product of $\mu_1, \ldots, \mu_n$, and is usually denoted by $\mu_1 \times \cdots \times \mu_n$.

EXERCISE 3.1.2. Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be σ-finite measure spaces for $i = 1, 2, 3$. Show that $(\mu_1 \times \mu_2) \times \mu_3 = \mu_1 \times \mu_2 \times \mu_3 = \mu_1 \times (\mu_2 \times \mu_3)$.

EXERCISE 3.1.3. Let S be a separable metric space, endowed with its Borel σ-algebra. Then $S \times S$ comes with its product topology, which defines its own Borel σ-algebra. Show that this is the same as the product σ-algebra on $S \times S$.

EXERCISE 3.1.4. Let S be as above, and let $(\Omega, \mathcal{F})$ be a measurable space. If f and g are measurable functions from Ω into S, show that $(f, g) : \Omega \to S \times S$ is a measurable function.

EXERCISE 3.1.5. Let S and Ω be as above. If ρ is the metric on S, show that $\rho : S \times S \to \mathbb{R}$ is a measurable map.

EXERCISE 3.1.6. Let $(\Omega, \mathcal{F})$ be a measurable space and let S be a complete separable metric space endowed with its Borel σ-algebra. Let $\{f_n\}_{n \geq 1}$ be a sequence of measurable functions from Ω into S. Show that

$$\{\omega \in S : \lim_{n \to \infty} f_n(\omega) \text{ exists}\}$$

is a measurable subset of Ω. (Hint: Try to write this set as a combination of countable unions and intersections of measurable sets, using the Cauchy criterion for convergence on complete metric spaces.)

3.2. Fubini's theorem

We will now learn how to integrate with respect to product measures. The natural idea is to compute an integral with respect to a product measure as an iterated integral, integrating with respect to one coordinate at a time. This works under certain conditions. The conditions are given by Fubini's theorem, stated below. We need a preparatory lemma.

LEMMA 3.2.1. *Let $(\Omega_i, \mathcal{F}_i)$ be measurable spaces for $i = 1, 2, 3$. Let $f : \Omega_1 \times \Omega_2 \to \Omega_3$ be a measurable function. Then for each $x \in \Omega_1$, the map $y \mapsto f(x, y)$ is measurable on Ω_2.*

PROOF. Take any $A \in \Omega_3$ and $x \in \Omega_1$. Let $B := f^{-1}(A)$ and

$$B_x := \{y \in \Omega_2 : (x, y) \in B\} = \{y \in \Omega_2 : f(x, y) \in A\}.$$

We want to show that $B_x \in \mathcal{F}_2$. For the given x, let $\mathcal{G}$ be the set of all $E \in \Omega_1 \times \Omega_2$ such that $E_x \in \Omega_2$, where $E_x := \{y \in \Omega_2 : (x, y) \in E\}$. An easy verification shows that $\mathcal{G}$ is a σ-algebra. Moreover, it contains every rectangular set. Thus, $\mathcal{G} \supseteq \mathcal{F}_1 \times \mathcal{F}_2$, which shows in particular that $B_x \in \mathcal{F}_2$ for every $x \in \Omega_1$, since $B \in \mathcal{F}_1 \times \mathcal{F}_2$ due to the measurability of f. □

THEOREM 3.2.2 (Fubini's theorem). *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be two σ-finite measure spaces. Let $\mu = \mu_1 \times \mu_2$, and let $f : \Omega_1 \times \Omega_2 \to \mathbb{R}^*$ be a measurable function. If f is either nonnegative or integrable, then the map $x \mapsto \int_{\Omega_2} f(x, y) d\mu_2(y)$ on Ω_1 and the map $y \mapsto \int_{\Omega_1} f(x, y) d\mu_1(x)$ on Ω_2 are well-defined and measurable (when set equal to zero if the integral is undefined). Moreover, we have*

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f(x, y) d\mu(x, y) &= \int_{\Omega_1} \int_{\Omega_2} f(x, y) d\mu_2(y) d\mu_1(x) \\ &= \int_{\Omega_2} \int_{\Omega_1} f(x, y) d\mu_1(x) d\mu_2(y),\end{aligned}$$

Finally, if either of

$$\int_{\Omega_1} \int_{\Omega_2} |f(x, y)| d\mu_2(y) d\mu_1(x) \text{ or } \int_{\Omega_2} \int_{\Omega_1} |f(x, y)| d\mu_1(x) d\mu_2(y)$$

is finite, then f is integrable.

PROOF. First, suppose that $f = 1_A$ for some $A \in \mathcal{F}_1 \times \mathcal{F}_2$. Then notice that for any $x \in \Omega_1$,

$$\int_{\Omega_2} f(x, y) d\mu_2(y) = \mu_2(A_x),$$

where $A_x := \{y \in \Omega_2 : (x, y) \in A\}$. The integral is well-defined by Lemma 3.2.1. We will first prove that $x \mapsto \mu_2(A_x)$ is a measurable map, and its integral equals $\mu(A)$. This will prove Fubini's theorem for this f.

Let $\mathcal{L}$ be the set of all $E \in \mathcal{F}_1 \times \mathcal{F}_2$ such that the map $x \mapsto \mu_2(E_x)$ is measurable and integrates to $\mu(E)$. We will now show that $\mathcal{L}$ is a λ-system. We will first prove this under the assumption that μ_1 and μ_2 are finite measures. Clearly $\Omega_1 \times \Omega_2 \in \mathcal{L}$. If $E_1, E_2, \ldots \in \mathcal{L}$ are disjoint, and E is their union, then for any x, E_x is the disjoint union of $(E_1)_x, (E_2)_x, \ldots$, and hence

$$\mu_2(E_x) = \sum_{i=1}^{\infty} \mu_2((E_i)_x).$$

By Exercise 2.1.12, this shows that $x \mapsto \mu_2(E_x)$ is measurable. The monotone convergence theorem and the fact that $E_i \in \mathcal{L}$ for each i show that

$$\int_{\Omega_1} \mu_2(E_x) d\mu_1(x) = \sum_{i=1}^{\infty} \int_{\Omega_1} \mu_2((E_i)_x) d\mu_1(x) = \sum_{i=1}^{\infty} \mu(E_i) = \mu(E).$$

Thus, $E \in \mathcal{L}$, and therefore $\mathcal{L}$ is closed under countable disjoint unions.

Finally, take any $E \in \mathcal{L}$. Since μ_1 and μ_2 are finite measures, we have

$$\mu_2((E^c)_x) = \mu_2((E_x)^c) = \mu_2(\Omega_2) - \mu_2(E_x),$$

which proves that $x \mapsto \mu_2((E^c)_x)$ is measurable. It also proves that

$$\begin{aligned} \int_{\Omega_1} \mu_2((E^c)_x) d\mu_1(x) &= \mu_1(\Omega_1)\mu_2(\Omega_2) - \int_{\Omega_1} \mu_2(E_x) d\mu_1(x) \\ &= \mu(\Omega) - \mu(E) = \mu(E^c). \end{aligned}$$

Thus, $\mathcal{L}$ is a λ-system. Since it contains the π-system of all rectangles, which generates $\mathcal{F}_1 \times \mathcal{F}_2$, we have now established the claim that for any $E \in \mathcal{F}_1 \times \mathcal{F}_2$, the map $x \mapsto \mu_2(E_x)$ is measurable and integrates to $\mu(E)$, provided that μ_1 and μ_2 are finite measures.

Now suppose that μ_1 and μ_2 are σ-finite measures. Let $\{E_{n,1}\}_{n \geq 1}$ and $\{E_{n,2}\}_{n \geq 2}$ be sequences of measurable sets of finite measure increasing to Ω_1 and Ω_2, respectively. For each n, let $E_n := E_{n,1} \times E_{n,2}$, and define the functionals $\mu_{n,1}(A) := \mu_1(A \cap E_{n,1})$, $\mu_{n,2}(B) := \mu_2(B \cap E_{n,2})$ and $\mu_n(E) := \mu(E \cap E_n)$ for $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$ and $E \in \mathcal{F}_1 \times \mathcal{F}_2$. It is easy to see that these are finite measures, increasing to μ_1, μ_2 and μ.

If $f : \Omega_1 \to [0, \infty]$ is a measurable function, it is easy see that

$$\int_{\Omega_1} f(x) d\mu_{n,1}(x) = \int_{\Omega_1} f(x) 1_{E_{n,1}}(x) d\mu_1(x), \tag{3.2.1}$$

where we use the convention $\infty \cdot 0 = 0$ on the right. To see this, first note that it holds for indicator functions by the definition of $\mu_{n,1}$. From this, pass to simple functions by linearity and then to nonnegative measurable functions by the monotone convergence theorem and Proposition 2.3.7.

Next, note that for any $E \in \mathcal{F}_1 \times \mathcal{F}_2$ and any $x \in \Omega_1$, $\mu_2(E_x)$ is the increasing limit of $\mu_{n,2}(E_x) 1_{E_{n,1}}(x)$ as $n \to \infty$. Firstly, this shows that $x \mapsto \mu_2(E_x)$ is a measurable map. Moreover, for each n, $\mu_n = \mu_{n,1} \times \mu_{n,2}$,

as can be seen by verifying on rectangles and using the uniqueness of product measures on σ-finite spaces. Therefore by the monotone convergence theorem and equation (3.2.1),

$$\begin{aligned}\int_{\Omega_1} \mu_2(E_x) d\mu_1(x) &= \lim_{n\to\infty} \int_{\Omega_1} \mu_{n,2}(E_x) 1_{E_{n,1}}(x) d\mu_1(x) \\ &= \lim_{n\to\infty} \int_{\Omega_1} \mu_{n,2}(E_x) d\mu_{n,1}(x) \\ &= \lim_{n\to\infty} \mu_n(E) = \mu(E).\end{aligned}$$

This completes the proof of Fubini's theorem for all indicator functions. By linearity, this extends to all simple functions. Using the monotone convergence theorem and Proposition 2.3.7, it is straightforward the extend the result to all nonnegative measurable functions.

Now take any integrable $f : \Omega_1 \times \Omega_2 \to \mathbb{R}^*$. Applying Fubini's theorem to f_+ and f_-, we get

$$\begin{aligned}\int f d\mu &= \int_{\Omega_1} \int_{\Omega_2} f_+(x,y) d\mu_2(y) d\mu_1(x) - \int_{\Omega_1} \int_{\Omega_2} f_-(x,y) d\mu_2(y) d\mu_1(x) \\ &= \int_{\Omega_1} g(x) d\mu_1(x) - \int_{\Omega_1} h(x) d\mu_1(x),\end{aligned}$$

where

$$g(x) := \int_{\Omega_2} f_+(x,y) d\mu_2(y), \quad h(x) := \int_{\Omega_2} f_-(x,y) d\mu_2(y).$$

By Fubini's theorem for nonnegative functions and the integrability of f, it follows that g and h are integrable functions. Thus,

$$\int f d\mu = \int_{\Omega_1} (g(x) - h(x)) d\mu_1(x),$$

where we have adopted the convention that $\infty - \infty = 0$, as in Proposition 2.4.1.

Now, at any x where at least one of $g(x)$ and $h(x)$ is finite,

$$g(x) - h(x) = \int_{\Omega_2} (f_+(x,y) - f_-(x,y)) d\mu_2(y) = \int_{\Omega_2} f(x,y) d\mu_2(y). \quad (3.2.2)$$

On the other hand, the set where $g(x)$ and $h(x)$ are both infinite is a set of measure zero, and therefore we may define

$$\int_{\Omega_2} f(x,y) d\mu_2(y) = 0$$

on this set, so that again (3.2.2) is valid under the convention $\infty - \infty = 0$. (The construction ensures, among other things, that $x \mapsto \int_{\Omega_2} f(x,y) d\mu_2(y)$ is a measurable function.) This concludes the proof of Fubini's theorem for integrable functions. The final assertion of the theorem follows easily from Fubini's theorem for nonnegative functions. □

EXERCISE 3.2.3. Produce a counterexample to show that the integrability condition in Fubini's theorem is necessary, in that otherwise the order of integration cannot be interchanged even if the two iterated integrals make sense.

EXERCISE 3.2.4. Compute the Lebesgue measure of the unit disk in $\mathbb{R}^2$ by integrating the indicator function of the disk using Fubini's theorem.

3.3. Infinite dimensional product spaces

Suppose now that $\{(\Omega_i, \mathcal{F}_i, \mu_i)\}_{i=1}^{\infty}$ is a countable sequence of σ-finite measure spaces. Let $\Omega = \Omega_1 \times \Omega_2 \times \cdots$. The product σ-algebra $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots$ is defined to be the σ-algebra generated by all sets of the form $A_1 \times A_2 \times \cdots$, where $A_i = \Omega_i$ for all but finitely many i. Such sets generalize the notion of rectangles to infinite product spaces.

Although the definition of the product σ-algebra is just as before, the existence of a product measure is a slightly trickier question. In particular, we need that each μ_i is a probability measure to even define the infinite product measure on rectangles in a meaningful way. To see why this condition is necessary, suppose that the measure spaces are all the same. Then if $\mu_i(\Omega_i) > 1$, each rectangle must have measure ∞, and if $\mu_i(\Omega_i) < 1$, each rectangle must have measure zero. To avoid these trivialities, we need to have $\mu_i(\Omega_i) = 1$.

THEOREM 3.3.1. *Let all notation be as above. Suppose that $\{\mu_i\}_{i=1}^{\infty}$ are probability measures. Then there exists a unique probability measure μ on $(\Omega, \mathcal{F})$ that satisfies*

$$\mu(A_1 \times A_2 \times \cdots) = \prod_{i=1}^{\infty} \mu_i(A_i)$$

for every rectangle $A_1 \times A_2 \times \cdots$.

PROOF. First note that by the existence of finite dimensional product measures, the measure $\nu_n := \mu_1 \times \cdots \times \mu_n$ is defined for each n.

Next, define $\Omega^{(n)} := \Omega_{n+1} \times \Omega_{n+2} \times \cdots$. Let $A \in \mathcal{F}$ be called a cylinder set if it is of the form $B \times \Omega^{(n)}$ for some n and some $B \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Define $\mu(A) := \nu_n(B)$. It is easy to see that this is a well-defined functional on $\mathcal{A}$, and that it satisfies the product property for rectangles.

Let $\mathcal{A}$ be the collection of all cylinder sets. It is not difficult to check that $\mathcal{A}$ is an algebra, and that μ is finitely additive and σ-finite on $\mathcal{A}$. Therefore by Carathéodory's theorem, we only need to check that μ is countably additive on $\mathcal{A}$. Suppose that $A \in \mathcal{A}$ is the union of a sequence of disjoint sets $A_1, A_2, \ldots \in \mathcal{A}$. For each n, let $B_n := A \setminus (A_1 \cup \cdots \cup A_n)$. Since $\mathcal{A}$ is an algebra, each $B_n \in \mathcal{A}$. Since μ is finitely additive on $\mathcal{A}$, $\mu(A) = \mu(B_n) + \mu(A_1) + \cdots + \mu(A_n)$ for each n. Therefore we have to show that $\lim \mu(B_n) = 0$. Suppose that this is not true. Since $\{B_n\}_{n \geq 1}$ is a

decreasing sequence of sets, this implies that there is some $\epsilon > 0$ such that $\mu(B_n) \geq \epsilon$ for all n. Using this, we will now get a contradiction to the fact that $\cap B_n = \emptyset$.

For each n, let $\mathcal{A}^{(n)}$ be the algebra of all cylinder sets in $\Omega^{(n)}$, and let $\mu^{(n)}$ be defined on $\mathcal{A}^{(n)}$ using $\mu_{n+1}, \mu_{n+2}, \ldots$, the same way we defined μ on $\mathcal{A}$. For any n, m, and $(x_1, \ldots, x_m) \in \Omega_1 \times \cdots \times \Omega_m$, define

$$B_n(x_1, \ldots, x_m) := \{(x_{m+1}, x_{m+2}, \ldots) \in \Omega^{(m)} : (x_1, x_2, \ldots) \in B_n\}.$$

Since B_n is a cylinder set, it is of the form $C_n \times \Omega^{(m)}$ for some m and some $C_n \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$. Therefore by Lemma 3.2.1, $B_n(x_1) \in \mathcal{A}^{(1)}$ for any $x_1 \in \Omega_1$, and by Fubini's theorem, the map $x_1 \mapsto \mu^{(1)}(B_n(x_1))$ is measurable. (Although we have not yet shown that $\mu^{(1)}$ is a measure on the product σ-algebra of $\Omega^{(1)}$, it is evidently a measure on the σ-algebra $\mathcal{G}$ of all sets of the form $D \times \Omega^{(m)} \subseteq \Omega^{(1)}$, where $D \in \mathcal{F}_2 \times \cdots \times \mathcal{F}_m$. Moreover, $B_n \in \mathcal{F}_1 \times \mathcal{G}$. This allows us to use Fubini's theorem to reach the above conclusion.) Thus, the set

$$F_n := \{x_1 \in \Omega_1 : \mu^{(1)}(B_n(x_1)) \geq \epsilon/2\}$$

is an element of $\mathcal{F}_1$. But again by Fubini's theorem,

$$\begin{aligned}
\mu(B_n) &= \int \mu^{(1)}(B_n(x_1)) d\mu_1(x_1) \\
&= \int_{F_n} \mu^{(1)}(B_n(x_1)) d\mu_1(x_1) + \int_{F_n^c} \mu^{(1)}(B_n(x_1)) d\mu_1(x_1) \\
&\leq \mu_1(F_n) + \frac{\epsilon}{2}.
\end{aligned}$$

Since $\mu(B_n) \geq \epsilon$, this shows that $\mu_1(F_n) \geq \epsilon/2$. Since $\{F_n\}_{n \geq 1}$ is a decreasing sequence of sets, this shows that $\cap F_n \neq \emptyset$. Choose a point $x_1^* \in \cap F_n$.

Now note that $\{B_n(x_1^*)\}_{n \geq 1}$ is a decreasing sequence of sets in $\mathcal{F}^{(1)}$, such that $\mu^{(1)}(B_n(x_1^*)) \geq \epsilon/2$ for each n. Repeating the above argument for the product space $\Omega^{(1)}$ and the sequence $\{B_n(x_1^*)\}_{n \geq 1}$, we see that there exists $x_2^* \in \Omega_2$ such that $\mu^{(2)}(B_n(x_1^*, x_2^*)) \geq \epsilon/4$ for every n.

Proceeding like this, we obtain a point $x = (x_1^*, x_2^*, \ldots) \in \Omega$ such that for any m and n,

$$\mu^{(m)}(B_n(x_1^*, \ldots, x_m^*)) \geq 2^{-m}\epsilon.$$

Take any n. Since B_n is a cylinder set, it is of the form $C_n \times \Omega^{(m)}$ for some m and some $C_n \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$. Since $\mu^{(m)}(B_n(x_1^*, \ldots, x_m^*)) > 0$, there is some $(x_{m+1}, x_{m+2}, \ldots) \in \Omega^{(m)}$ such that $(x_1^*, \ldots, x_m^*, x_{m+1}, x_{m+2}, \ldots) \in B_n$. But by the form of B_n, this implies that $x \in B_n$. Thus, $x \in \cap B_n$. This completes the proof. □

Having constructed products of countably many probability spaces, one may now wonder about uncountable products. Surprisingly, this is quite simple, given that we know how to handle the countable case. Suppose that $\{(\Omega_i, \mathcal{F}_i, \mu_i)\}_{i \in I}$ is an arbitrary collection of probability spaces. Let

$\Omega := \times_{i \in I} \Omega_i$, and let $\mathcal{F} := \times_{i \in I} \mathcal{F}_i$ be defined using rectangles as in the countable case. Now take any countable set $J \subseteq I$, and let $\mathcal{F}_J := \times_{j \in J} \mathcal{F}_j$. Consider a set $A \in \mathcal{F}$ of the form $\{(\omega_i)_{i \in I} : (\omega_j)_{j \in J} \in B\}$, where B is some element of $\mathcal{F}_J$. Let $\mathcal{G}$ be the collection of all such A, as B varies in $\mathcal{F}_J$ and J varies over all countable subsets of I. It is not hard to check that $\mathcal{G}$ is in fact a σ-algebra. Moreover, it is contained in $\mathcal{F}$, and it contains all rectangular sets. Therefore $\mathcal{G} = \mathcal{F}$. Thus we can define μ on this σ-algebra simply using the definition of the product measure on countable product spaces.

Sometimes showing that some function is measurable with respect to an infinite product σ-algebra can be somewhat tricky. The following exercise gives such an example, which arises in percolation theory.

EXERCISE 3.3.2. Take any $d \geq 1$ and consider the integer lattice $\mathbb{Z}^d$ with the nearest-neighbor graph structure. Let E denote the set of edges. Take the two-point set $\{0, 1\}$ with its power set σ-algebra, and consider the product space $\{0, 1\}^E$. Given $\omega = (\omega_e)_{e \in E} \in \{0, 1\}^E$, define a subgraph of $\mathbb{Z}^d$ as follows: Keep an edge e if $\omega_e = 1$, and delete it otherwise. Let $N(\omega)$ be the number of connected components of this subgraph. Prove that $N : \{0, 1\}^E \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ is a measurable function.

CHAPTER 4

Norms and inequalities

The main goal of this chapter is to introduce L^p spaces and discuss some of their properties. In particular, we will discuss some inequalities for L^p spaces that are useful in probability theory.

4.1. Markov's inequality

The following simple but important inequality is called Markov's inequality in the probability literature.

THEOREM 4.1.1 (Markov's inequality). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to [0, \infty]$ be a measurable function. Then for any $t > 0$,*

$$\mu(\{\omega : f(\omega) \geq t\}) \leq \frac{1}{t} \int f d\mu.$$

PROOF. Take any $t > 0$. Define a function g as

$$g(\omega) := \begin{cases} 1 & \text{if } f(\omega) \geq t, \\ 0 & \text{if } f(\omega) < t. \end{cases}$$

Then $g \leq t^{-1} f$ everywhere. Thus,

$$\int g d\mu \leq \frac{1}{t} \int f d\mu.$$

The proof is completed by observing that $\int g d\mu = \mu(\{\omega : f(\omega) \geq t\})$ by the definition of Lebesgue integral for simple functions. □

4.2. Jensen's inequality

Jensen's inequality is another basic tool in probability theory. Unlike Markov's inequality, this inequality holds for probability measures only.

First, let us recall the definition of a convex function. Let I be an interval in $\mathbb{R}^*$. The interval may be finite or infinite, open, closed or half-open. A function $\phi : I \to \mathbb{R}^*$ is called convex if for all $x, y \in I$ and $t \in [0, 1]$,

$$\phi(tx + (1 - t)y) \leq t\phi(x) + (1 - t)\phi(y).$$

EXERCISE 4.2.1. If ϕ is differentiable, show that ϕ is convex if and only if ϕ' is an increasing function.

EXERCISE 4.2.2. If ϕ is twice differentiable, show that ϕ is convex if and only if ϕ'' is nonnegative everywhere.

EXERCISE 4.2.3. Show that the functions $\phi(x) = |x|^\alpha$ for $\alpha \geq 1$ and $\phi(x) = e^{\theta x}$ for $\theta \in \mathbb{R}$, are all convex.

An important property of convex functions is that they are continuous in the interior of their domain.

EXERCISE 4.2.4. Let I be an interval and $\phi : I \to \mathbb{R}$ be a convex function. Prove that ϕ is continuous at every interior point of I, and hence that ϕ is measurable.

Another important property of convex functions is that have at least one tangent at every interior point of their domain.

EXERCISE 4.2.5. Let I be an interval and $\phi : I \to \mathbb{R}$ be a convex function. Then for any x in the interior of I, show that there exist $a, b \in \mathbb{R}$ such that $\phi(x) = ax + b$ and $\phi(y) \geq ay + b$ for all $y \in I$. Moreover, if ϕ is nonlinear in every neighborhood of x, show that a and b can be chosen such that $\phi(y) > ay + b$ for all $y \neq x$.

THEOREM 4.2.6 (Jensen's inequality). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $f : \Omega \to \mathbb{R}^*$ be an integrable function. Let I be an interval containing the range of f, and let $\phi : I \to \mathbb{R}$ be a convex function. Let $m := \int f d\mathbb{P}$. Then $\phi(m) \leq \int \phi \circ f d\mathbb{P}$, provided that $\phi \circ f$ is also integrable. Moreover, if ϕ is nonlinear in every open neighborhood of m, then equality holds in the above inequality if and only if $f = m$ a.e.*

PROOF. Integrability implies that f is finite a.e. Therefore we can replace f by a function that is finite everywhere, without altering either side of the claimed inequality. By Exercise 2.6.6, $m \in I$. If m equals either endpoint of I, then it is easy to show that $f = m$ a.e., and there is nothing more to prove. So assume that m is in the interior of I. Then by Exercise 4.2.5, there exist $a, b \in \mathbb{R}$ such that $ax + b \leq \phi(x)$ for all $x \in I$ and $am + b = \phi(m)$. Thus,

$$\int_\Omega \phi(f(x))\, d\mathbb{P}(x) \geq \int_\Omega (af(x) + b)\, d\mathbb{P}(x) = am + b = \phi(m),$$

which is the desired inequality. If ϕ is nonlinear in every open neighborhood of m, then by the second part of Exercise 4.2.5 we can guarantee that $\phi(x) > ax + b$ for all $x \neq m$. Thus, by Exercise 2.6.3, equality can hold in the above display if and only if $f = m$ a.e. □

Jensen's inequality is often used to derive inequalities for functions of real numbers. An example is the following.

EXERCISE 4.2.7. If $x_1, \ldots, x_n \in \mathbb{R}$ and $p_1, \ldots, p_n \in [0, 1]$ are numbers such that $\sum p_i = 1$, and ϕ is a convex function on an interval containing the

x_i's, use Jensen's inequality to show that $\phi(\sum x_i p_i) \leq \sum \phi(x_i) p_i$. (Strictly speaking, Jensen's inequality is not really needed for this proof; it can done by simply using the definition of convexity and induction on n.)

A function ϕ is called concave if $-\phi$ is convex. Clearly, the opposite of Jensen's inequality holds for concave functions. An important concave function is the logarithm, whose concavity can be verified by simply noting that its second derivative is negative everywhere on the positive real line. A consequence of the concavity of the logarithm is Young's inequality, which we will use later in this chapter to prove Hölder's inequality.

EXERCISE 4.2.8 (Young's inequality). If x and y are positive real numbers, and $p, q \in (1, \infty)$ are number such that $1/p + 1/q = 1$, show that

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

(Hint: Take the logarithm of the right side and apply the definition of concavity.)

4.3. The first Borel–Cantelli lemma

The first Borel–Cantelli lemma is an important tool for proving limit theorems in probability theory. In particular, we will use it in the next section to prove the completeness of L^p spaces.

THEOREM 4.3.1 (The first Borel–Cantelli lemma). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $A_1, A_2, \ldots \in \mathcal{F}$ be events such that $\sum \mu(A_n) < \infty$. Then $\mu(\{\omega : \omega \in \text{infinitely many } A_n\text{'s}\}) = 0$.*

PROOF. It is not difficult to see that in set theoretic notation,

$$\{\omega : \omega \in \text{infinitely many } A_n\text{'s}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Since the inner union on the right is decreasing in n,

$$\begin{aligned}\mu(\{\omega : \omega \in \text{infinitely many } A_n\text{'s}\}) &\leq \inf_{n \geq 1} \mu\Big(\bigcup_{k=n}^{\infty} A_k\Big) \\ &\leq \inf_{n \geq 1} \sum_{k=n}^{\infty} \mu(A_k) = \lim_{n \to \infty} \sum_{k=n}^{\infty} \mu(A_k).\end{aligned}$$

The finiteness of $\sum \mu(A_n)$ shows that the limit on the right equals zero, completing the proof. □

EXERCISE 4.3.2. Produce a counterexample to show that the converse of the first Borel–Cantelli lemma is not true, even for probability measures.

4.4. L^p spaces and inequalities

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, to be fixed throughout this section. Let $f : \Omega \to \mathbb{R}^*$ be a measurable function. Given any $p \in [1, \infty)$, the L^p norm of f is defined as

$$\|f\|_{L^p} := \left(\int |f|^p d\mu \right)^{1/p}.$$

The space $L^p(\Omega, \mathcal{F}, \mu)$ (or simply $L^p(\Omega)$ or $L^p(\mu)$) is the set of all measurable $f : \Omega \to \mathbb{R}^*$ such that $\|f\|_{L^p}$ is finite. In addition to the above, there is also the L^∞ norm, defined as

$$\|f\|_{L^\infty} := \inf\{K \in [0, \infty] : |f| \leq K \text{ a.e.}\}.$$

The right side is called the 'essential supremum' of the function $|f|$. It is not hard to see that $|f| \leq \|f\|_{L^\infty}$ a.e. As before, $L^\infty(\Omega, \mathcal{F}, \mu)$ is the set of all f with finite L^∞ norm.

It turns out that for any $1 \leq p \leq \infty$, $L^p(\Omega, \mathcal{F}, \mu)$ is actually a vector space and the L^p is actually a norm on this space, provided that we first quotient out the space by the equivalence relation of being equal almost everywhere. Moreover, these norms are complete (that is, Cauchy sequences converge). The only thing that is obvious is that $\|\alpha f\|_{L^p} = |\alpha| \|f\|_{L^p}$ for any $\alpha \in \mathbb{R}$ and any p and f. Proving the other claims, however, requires some work, which we do below.

THEOREM 4.4.1 (Hölder's inequality). *Take any measurable $f, g : \Omega \to \mathbb{R}^*$, and any $p \in [1, \infty]$. Let q be the solution of $1/p + 1/q = 1$. Then $\|fg\|_{L^1} \leq \|f\|_{L^p} \|g\|_{L^q}$.*

PROOF. If $p = 1$, then $q = \infty$. The claimed inequality then follows simply as

$$\int |fg| d\mu \leq \|g\|_{L^\infty} \int |f| d\mu = \|f\|_{L^1} \|g\|_{L^\infty}.$$

If $p = \infty$, then $q = 1$ and the proof is exactly the same. So assume that $p \in (1, \infty)$, which implies that $q \in (1, \infty)$. Let $\alpha := 1/\|f\|_{L^p}$ and $\beta := 1/\|g\|_{L^q}$, and let $u := \alpha f$ and $v := \beta g$. Then $\|u\|_{L^p} = \alpha \|f\|_{L^p} = 1$ and $\|v\|_{L^q} = \beta \|g\|_{L^q} = 1$. Now, by Young's inequality,

$$|uv| \leq \frac{|u|^p}{p} + \frac{|v|^q}{q}$$

everywhere. Integrating both sides, we get

$$\|uv\|_{L^1} \leq \frac{\|u\|_{L^p}^p}{p} + \frac{\|v\|_{L^q}^q}{q} = \frac{1}{p} + \frac{1}{q} = 1.$$

It is now easy to see that this is precisely the inequality that we wanted to prove. □

EXERCISE 4.4.2. If $p, q \in (1, \infty)$ and $f \in L^p(\mu)$ and $g \in L^q(\mu)$, then show that Hölder's inequality becomes an equality if and only if there exist $\alpha, \beta \geq 0$, not both of them zero, such that $\alpha|f|^p = \beta|g|^q$ a.e.

Using Hölder's inequality, it is now easy to prove that the L^p norms satisfy the triangle inequality.

THEOREM 4.4.3 (Minkowski's inequality). *For any $1 \leq p \leq \infty$ and any measurable $f, g : \Omega \to \mathbb{R}^*$, $\|f + g\|_{L^p} \leq \|f\|_{L^p} + \|g\|_{L^p}$.*

PROOF. If the right side is infinite, there is nothing to prove. So assume that the right side is finite. First, consider the case $p = \infty$. Since $|f| \leq \|f\|_{L^\infty}$ a.e. and $|g| \leq \|g\|_{L^\infty}$ a.e., it follows that $|f + g| \leq \|f\|_{L^\infty} + \|g\|_{L^\infty}$ a.e., which completes the proof by the definition of essential supremum.

On the other hand, if $p = 1$, then the claimed inequality follows trivially from the triangle inequality and the additivity of integration.

So let us assume that $p \in (1, \infty)$. First, observe that by Exercise 4.2.7,

$$\left|\frac{f+g}{2}\right|^p \leq \frac{|f|^p + |g|^p}{2},$$

which shows that $\|f + g\|_{L^p}$ is finite. Next note that by Hölder's inequality,

$$\begin{aligned}\int |f+g|^p d\mu &= \int |f+g||f+g|^{p-1} d\mu \\ &\leq \int |f||f+g|^{p-1} d\mu + \int |g||f+g|^{p-1} d\mu \\ &\leq \|f\|_{L^p} \||f+g|^{p-1}\|_{L^q} + \|g\|_{L^p} \||f+g|^{p-1}\|_{L^q},\end{aligned}$$

where q solves $1/p + 1/q = 1$. But

$$\||f+g|^{p-1}\|_{L^q} = \left(\int |f+g|^{(p-1)q} d\mu\right)^{1/q} = \left(\int |f+g|^p d\mu\right)^{1/q}.$$

Combining, and using the finiteness of $\|f + g\|_{L^p}$, we get

$$\|f+g\|_{L^p} = \left(\int |f+g|^p d\mu\right)^{1-1/q} \leq \|f\|_{L^p} + \|g\|_{L^p},$$

which is what we wanted to prove. □

EXERCISE 4.4.4. If $p \in (1, \infty)$, show that equality holds in Minkowski's inequality if and only if $f = \lambda g$ for some $\lambda \geq 0$ or $g = 0$.

Minkowski's inequality shows, in particular, that $L^p(\Omega, \mathcal{F}, \mu)$ is a vector space, and the L^p norm satisfies two of the three required properties of a norm on this space. The only property that it does not satisfy is that f may be nonzero even if $\|f\|_{L^p} = 0$. But this is not a serious problem, since by Proposition 2.6.1, the vanishing of the L^p norm implies that $f = 0$ a.e. More generally, $\|f - g\|_{L^p} = 0$ if and only if $f = g$ a.e. This shows that if we quotient out $L^p(\Omega, \mathcal{F}, \mu)$ by the equivalence relation of being equal a.e.,

the resulting quotient space is a vector space where the L^p norm is indeed a norm. Since we already think of two functions which are equal a.e. as effectively the same function, we will not worry about this technicality too much and continue to treat our definition of L^p space as a vector space with L^p norm.

The fact that is somewhat nontrivial, however, is that the L^p norm is complete. That is, any sequence of functions that is Cauchy in the L^p norm converges to a limit in L^p space. A first step towards the proof is the following lemma, which is important in its own right.

LEMMA 4.4.5. *If $\{f_n\}_{n\geq 1}$ is a Cauchy sequence in the L^p norm for some $1 \leq p \leq \infty$, then there is function $f \in L^p(\Omega, \mathcal{F}, \mu)$, and a subsequence $\{f_{n_k}\}_{k\geq 1}$, such that $f_{n_k} \to f$ a.e. as $k \to \infty$.*

PROOF. First suppose that $p \in [1, \infty)$. It is not difficult to see that using the Cauchy criterion, we can extract a subsequence $\{f_{n_k}\}_{k\geq 1}$ such that $\|f_{n_k} - f_{n_{k+1}}\|_{L^p} \leq 2^{-k}$ for every k. Define the event

$$A_k := \{\omega : |f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| \geq 2^{-k/2}\}.$$

Then by Markov's inequality,

$$\begin{aligned}\mu(A_k) &= \mu(\{\omega : |f_{n_k}(\omega) - f_{n_{k+1}}(\omega)|^p \geq 2^{-kp/2}\}) \\ &\leq 2^{kp/2} \int |f_{n_k} - f_{n_{k+1}}|^p d\mu \\ &= 2^{kp/2} \|f_{n_k} - f_{n_{k+1}}\|_{L^p}^p \leq 2^{-kp/2}.\end{aligned}$$

Thus, $\sum \mu(A_k) < \infty$. Therefore by the first Borel–Cantelli lemma, $\mu(B) = 0$, where $B := \{\omega : \omega \in \text{infinitely many } A_k\text{'s}\}$. If $\omega \in B^c$, then ω belongs to only finitely many of the A_k's. This means that $|f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| \leq 2^{-k/2}$ for all sufficiently large k. From this, it follows that $\{f_{n_k}(\omega)\}_{k\geq 1}$ is a Cauchy sequence of real numbers. Define $f(\omega)$ to be the limit of this sequence. For $\omega \in B$, define $f(\omega) = 0$. Then f is measurable and $f_{n_k} \to f$ a.e. Moreover, by the a.e. version of Fatou's lemma,

$$\int |f|^p d\mu \leq \liminf_{k\to\infty} \int |f_{n_k}|^p d\mu = \liminf_{k\to\infty} \|f_{n_k}\|_{L^p}^p < \infty,$$

since a Cauchy sequence of real numbers must be bounded. Thus, f has finite L^p norm.

Next, suppose that $p = \infty$. Extract a subsequence $\{f_{n_k}\}_{k\geq 1}$ as before. Then for each k, $|f_{n_k} - f_{n_{k+1}}| \leq 2^{-k}$ a.e. Therefore, if we define

$$E := \{\omega : |f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| \leq 2^{-k} \text{ for all } k\},$$

then $\mu(E^c) = 0$. For any $\omega \in E$, $\{f_{n_k}(\omega)\}_{k\geq 1}$ is a Cauchy sequence of real numbers. Define $f(\omega)$ to be the limit of this sequence. For $\omega \in E^c$, define

$f(\omega) = 0$. Then f is measurable and $f_{n_k} \to f$ a.e. Moreover, on E,

$$|f| \le |f_{n_1}| + \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| \le |f_{n_1}| + \sum_{k=1}^{\infty} 2^{-k},$$

which shows that f has finite L^∞ norm. □

THEOREM 4.4.6 (Riesz–Fischer theorem). *For any $1 \le p \le \infty$, the L^p norm on $L^p(\Omega, \mathcal{F}, \mu)$ is complete.*

PROOF. Take any sequence $\{f_n\}_{n\ge1}$ that is Cauchy in L^p. Then by Lemma 4.4.5, there is a subsequence $\{f_{n_k}\}_{k\ge1}$ that converges a.e. to a function f which is also in L^p. First, suppose that $p \in [1, \infty)$. Take any $\epsilon > 0$, and find N such that for all $m, n \ge N$, $\|f_n - f_m\|_{L^p} < \epsilon$. Take any $n \ge N$. Then by the a.e. version of Fatou's lemma,

$$\begin{aligned}\int |f_n - f|^p d\mu &= \int \lim_{k\to\infty} |f_n - f_{n_k}|^p d\mu \\ &\le \liminf_{k\to\infty} \int |f_n - f_{n_k}|^p d\mu \le \epsilon^p.\end{aligned}$$

This shows that $f_n \to f$ in the L^p norm. Next, suppose that $p = \infty$. Take any $\epsilon > 0$ and find N as before. Let

$$E := \{\omega : |f_n(\omega) - f_m(\omega)| \le \epsilon \text{ for all } m, n \ge N\},$$

so that $\mu(E^c) = 0$. Take any $n \ge N$. Then for any ω such that $\omega \in E$ and $f_{n_k}(\omega) \to f(\omega)$,

$$|f_n(\omega) - f(\omega)| = \lim_{k\to\infty} |f_n(\omega) - f_{n_k}(\omega)| \le \epsilon.$$

This shows that $\|f_n - f\|_{L^\infty} \le \epsilon$, completing the proof that $f_n \to f$ in the L^∞ norm. □

The following fact about L^p spaces of probability measures is very important in probability theory. It does not hold for general measures.

THEOREM 4.4.7 (Monotonicity of L^p norms for probability measures). *Suppose that μ is a probability measure. Then for any measurable $f : \Omega \to \mathbb{R}^*$ and any $1 \le p \le q \le \infty$, $\|f\|_{L^p} \le \|f\|_{L^q}$.*

PROOF. If $p = q = \infty$, there is nothing to prove. So let p be finite. If $q = \infty$, then

$$\int |f|^p d\mu \le \|f\|_{L^\infty}^p \mu(\Omega) = \|f\|_{L^\infty}^p,$$

which proves the claim. So let q be also finite. First assume that $\|f\|_{L^p}$ and $\|f\|_{L^q}$ are both finite. Applying Jensen's inequality with the convex function $\phi(x) = |x|^{q/p}$, we get the desired inequality

$$\left(\int |f|^p d\mu\right)^{q/p} \le \int |f|^q d\mu.$$

Finally, let us drop the assumption of finiteness of $\|f\|_{L^p}$ and $\|f\|_{L^q}$. Take a sequence of nonnegative simple functions $\{g_n\}_{n\geq 1}$ increasing pointwise to $|f|$ (which exists by Proposition 2.3.7). Since μ is a probability measure, $\|g_n\|_{L^p}$ and $\|g_n\|_{L^q}$ are both finite. Therefore $\|g_n\|_{L^p} \leq \|g_n\|_{L^q}$ for each n. We can now complete the proof by applying the monotone convergence theorem to both sides. □

EXERCISE 4.4.8. When $\Omega = \mathbb{R}$ and μ is the Lebesgue measure, produce a function f whose L^2 norm is finite but L^1 norm is infinite.

The space L^2 holds a special status among all the L^p spaces, because it has a natural rendition as a Hilbert space with respect to the inner product

$$(f, g) := \int f g d\mu.$$

It is easy to verify that this is indeed an inner product on the vector space $L^2(\Omega, \mathcal{F}, \mu)$, and the norm generated by this inner product is the L^2 norm (after quotienting out by the equivalence relation of a.e. equality, as usual). The completeness of the L^2 norm guarantees that this is indeed a Hilbert space. The Cauchy–Schwarz inequality on this Hilbert space is a special case of Hölder's inequality with $p = q = 2$.

CHAPTER 5

Random variables

In this chapter we will study the basic properties of random variables and related functionals.

5.1. Definition

If $(\Omega, \mathcal{F})$ is a measurable space, a measurable function from Ω into $\mathbb{R}$ or $\mathbb{R}^*$ is called a random variable. More generally, if $(\Omega', \mathcal{F}')$ is another measurable space, then a measurable function from Ω into Ω' is called a Ω'-valued random variable defined on Ω. Unless otherwise mentioned, we will assume that any random variable that we are talking about is real-valued.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable defined on Ω. It is the common convention in probability theory to write $\mathbb{P}(X \in A)$ instead of $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$. Similarly, we write $\{X \in A\}$ to denote the set $\{\omega \in \Omega : X(\omega) \in A\}$. Similarly, if X and Y are two random variables, the event $\{X \in A, Y \in B\}$ is the set $\{\omega : X(\omega) \in A, Y(\omega) \in B\}$.

Another commonly used convention is that if $f : \mathbb{R} \to \mathbb{R}$ is a measurable function and X is a random variable, $f(X)$ denotes the random variable $f \circ X$. The σ-algebra generated by a random variable X is denoted by $\sigma(X)$, and if $\{X_i\}_{i \in I}$ is a collection of random variables defined on the same probability space, then the σ-algebra $\sigma(\{X_i\}_{i \in I})$ generated by the collection $\{X_i\}_{i \in I}$ is defined to be the σ-algebra generated by the union of the sets $\sigma(X_i)$, $i \in I$.

Exercise 5.1.1. If $X_1, \ldots, X_n$ are random variables defined on the same probability space, and $f : \mathbb{R}^n \to \mathbb{R}$ is a measurable function, show that the random variable $f(X_1, \ldots, X_n)$ is measurable with respect to the σ-algebra $\sigma(X_1, \ldots, X_n)$.

Exercise 5.1.2. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. Show that the random variables $\sup X_n$, $\inf X_n$, $\limsup X_n$ and $\liminf X_n$ are all measurable with respect to $\sigma(X_1, X_2, \ldots)$.

Exercise 5.1.3. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any $A \in \sigma(X_1, X_2, \ldots)$ and any $\epsilon > 0$, show that there is some $n \geq 1$ and some $B \in \sigma(X_1, \ldots, X_n)$ such that $\mathbb{P}(A \Delta B) < \epsilon$. (Hint: Use Theorem 1.3.7.)

EXERCISE 5.1.4. If $X_1, \ldots, X_n$ are random variables defined on the same probability space, and X is a random variable that is measurable with respect to $\sigma(X_1, \ldots, X_n)$, then there is a measurable function $f : \mathbb{R}^n \to \mathbb{R}$ such that $X = f(X_1, \ldots, X_n)$.

EXERCISE 5.1.5. Extend the previous exercise to σ-algebras generated by arbitrarily many random variables.

5.2. Cumulative distribution function

DEFINITION 5.2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable defined on Ω. The cumulative distribution function of X is a function $F_X : \mathbb{R} \to [0, 1]$ defined as

$$F_X(t) := \mathbb{P}(X \leq t).$$

Often, the cumulative distribution function is simply called the distribution function of X, or abbreviated as the c.d.f. of X.

PROPOSITION 5.2.2. *Let $F : \mathbb{R} \to [0, 1]$ be a function that is non-decreasing, right-continuous, and satisfies*

$$\lim_{t \to -\infty} F(t) = 0 \quad \textit{and} \quad \lim_{t \to \infty} F(t) = 1. \tag{5.2.1}$$

Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X defined on this space such that F is the cumulative distribution function of X. Conversely, the cumulative distribution function of any random variable X has the above properties.

PROOF. Take $\Omega = (0, 1)$, $\mathcal{F} =$ the restriction of $\mathcal{B}(\mathbb{R})$ to $(0, 1)$, and $\mathbb{P} =$ the restriction of Lebesgue measure to $(0, 1)$, which is a probability measure. For $\omega \in \Omega$, define $X(\omega) := \inf\{t \in \mathbb{R} : F(t) \geq \omega\}$. Since $F(t) \to 1$ as $t \to \infty$ and $F(t) \to 0$ as $t \to -\infty$, X is well-defined on Ω. Now, if $\omega \leq F(t)$ for some ω and t, then by the definition of X, $X(\omega) \leq t$. Conversely, if $X(\omega) \leq t$, then there is a sequence $t_n \downarrow t$ such that $F(t_n) \geq \omega$ for each n. By the right-continuity of F, this implies that $F(t) \geq \omega$. Thus, we have shown that $X(\omega) \leq t$ if and only if $F(t) \geq \omega$. By either Exercise 2.1.6 or Exercise 2.1.7, F is measurable. Hence, the previous sentence shows that X is also measurable, and moreover that $\mathbb{P}(X \leq t) = \mathbb{P}((0, F(t)]) = F(t)$. Thus, F is the c.d.f. of the random variable X.

Conversely, if F is the c.d.f. of a random variable, it is easy to show that F is right-continuous and satisfies (5.2.1) by the continuity of probability measures under increasing unions and decreasing intersections. Monotonicity of F follows by the monotonicity of probability measures. □

Because of Proposition 5.2.2, any function satisfying the three conditions stated in the statement of the proposition is called a cumulative distribution function (or just distribution function or c.d.f.).

The following exercise is often used in proofs involving cumulative distribution functions.

EXERCISE 5.2.3. Show that any cumulative distribution function can have only countably many points of discontinuity. As a consequence, show that the set of continuity points is a dense subset of $\mathbb{R}$.

5.3. The law of a random variable

Any random variable X induces a probability measure μ_X on the real line, defined as

$$\mu_X(A) := \mathbb{P}(X \in A).$$

This generalizes easily to random variables taking value in other spaces. The probability measure μ_X is called the law of X.

PROPOSITION 5.3.1. *Two random variables have the same cumulative distribution function if and only if they have the same law.*

PROOF. If X and Y have the same law, it is clear that they have the same c.d.f. Conversely, let X and Y be two random variables that have the same distribution function. Let $\mathcal{A}$ be the set of all Borel sets $A \subseteq \mathbb{R}$ such that $\mu_X(A) = \mu_Y(A)$. It is easy to see that $\mathcal{A}$ is a λ-system. Moreover, $\mathcal{A}$ contains all intervals of the form $(a, b]$, $a, b \in \mathbb{R}$, which is a π-system that generates the Borel σ-algebra on $\mathbb{R}$. Therefore, by Dynkin's π-λ theorem, $\mathcal{A} \supseteq \mathcal{B}(\mathbb{R})$, and hence $\mu_X = \mu_Y$. □

The above proposition shows that there is a one-to-one correspondence between cumulative distribution functions and probability measures on $\mathbb{R}$. Moreover, the following is true.

EXERCISE 5.3.2. If X and Y have the same law, and $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, show that $g(X)$ and $g(Y)$ also have the same law.

5.4. Probability density function

Suppose that f is a nonnegative integrable function on the real line, such that

$$\int_{-\infty}^{\infty} f(x)dx = 1,$$

where $dx = d\lambda(x)$ denotes integration with respect to the Lebesgue measure λ defined in Section 1.6, and the range of the integral denotes integration over the whole real line. Although this is Lebesgue integration, we retain the notation of Riemann integration for the sake of familiarity. By Exercise 2.3.6, f defines a probability measure ν on $\mathbb{R}$ as

$$\nu(A) := \int_A f(x)dx \tag{5.4.1}$$

for each $A \in \mathcal{B}(\mathbb{R})$. The function f is called the probability density function (p.d.f.) of the probability measure ν.

To verify that a given function f is the p.d.f. of a random variable, it is not necessary to check (5.4.1) for every Borel set A. It suffices that it holds for a much smaller class, as given by the following proposition.

PROPOSITION 5.4.1. *A function f is the p.d.f. of a random variable X if and only if* (5.4.1) *is satisfied for every set A of the form $[a, b]$ where a and b are continuity points of the c.d.f. of X.*

PROOF. One implication is trivial. For the other, let F be the c.d.f. of X and suppose that (5.4.1) is satisfied for every set A of the form $[a, b]$ where a and b are continuity points of F. We then claim that (5.4.1) holds for every $[a, b]$, even if a and b are not continuity points of F. This is easily established using Exercise 5.2.3 and the dominated convergence theorem. Once we know this, the result can be completed by the π-λ theorem, observing that the set of closed intervals is a π-system, and the set of all Borel sets A for which the identity (5.4.1) holds is a λ-system. □

The following exercise relates the p.d.f. and the c.d.f. It is simple consequence of the above proposition.

EXERCISE 5.4.2. If f is a p.d.f. on $\mathbb{R}$, show that the function

$$F(x) := \int_{-\infty}^{x} f(y)dy \tag{5.4.2}$$

is the c.d.f. of the probability measure ν defined by f as in (5.4.1). Conversely, if F is a c.d.f. on $\mathbb{R}$ for which there exists a nonnegative measurable function f satisfying (5.4.2) for all x, show that f is a p.d.f. which generates the probability measure corresponding to F.

The next exercise establishes that the p.d.f. is essentially unique when it exists.

EXERCISE 5.4.3. If f and g are two probability density functions for the same probability measure on $\mathbb{R}$, show that $f = g$ a.e. with respect to Lebesgue measure.

Because of the above exercise, we will generally treat the probability density function of a random variable X as a unique function (although it is only unique up to almost everywhere equality), and refer to it as 'the p.d.f.' of X.

5.5. Some standard densities

An important example of a p.d.f. is the density function for the normal (or Gaussian) distribution with mean parameter $\mu \in \mathbb{R}$ and standard deviation parameter $\sigma > 0$, given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

If a random variable X has this distribution, we write $X \sim N(\mu, \sigma^2)$. The special case of $\mu = 0$ and $\sigma = 1$ is known as the standard normal distribution.

EXERCISE 5.5.1. Verify that the p.d.f. of the standard normal distribution is indeed a p.d.f., that is, its integral over the real line equals 1. Then use a change of variable to prove that the normal p.d.f. is indeed a p.d.f. for any μ and σ. (Hint: Square the integral and pass to polar coordinates.)

EXERCISE 5.5.2. If $X \sim N(\mu, \sigma^2)$, show that for any $a, b \in \mathbb{R}$, $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Another class of densities that occur quite frequently are the exponential densities. The exponential distribution with rate parameter λ has p.d.f.

$$f(x) = \lambda e^{-\lambda x} 1_{\{x \geq 0\}},$$

where $1_{\{x \geq 0\}}$ denotes the function that is 1 when $x \geq 0$ and 0 otherwise. It is easy to see that this is indeed a p.d.f., and its c.d.f. is given by

$$F(x) = (1 - e^{-\lambda x}) 1_{\{x \geq 0\}}.$$

If X is a random variable with this c.d.f., we write $X \sim Exp(\lambda)$.

EXERCISE 5.5.3. If $X \sim Exp(\lambda)$, show that for any $a > 0$, $aX \sim Exp(\lambda/a)$.

The Gamma distribution with rate parameter $\lambda > 0$ and shape parameter $r > 0$ has probability density function

$$f(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} 1_{\{x \geq 0\}},$$

where Γ denotes the standard Gamma function:

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$$

(Recall that $\Gamma(r) = (r-1)!$ if r is a positive integer.) If a random variable X has this distribution, we write $X \sim Gamma(r, \lambda)$.

Yet another important class of distributions that have densities are uniform distributions. The uniform distribution on an interval $[a, b]$ has the probability density that equals $1/(b-a)$ in this interval and 0 outside. If a random variable X has this distribution, we write $X \sim Unif[a, b]$.

5.6. Standard discrete distributions

Random variables that have continuous c.d.f. are known as continuous random variables. A discrete random variable is a random variable that can only take values in a finite or countable set. Note that it is possible that a random variable is neither continuous nor discrete. The law of a discrete random variable is characterized by its probability mass function (p.m.f.), which gives the probabilities of attaining the various values in its range. It is not hard to see that the p.m.f. uniquely determines the c.d.f. and hence the law. Moreover, any nonnegative function on a finite or countable subset

of $\mathbb{R}$ that adds up to 1 is a p.m.f. for a probability measure. The simplest example of a discrete random variable is a Bernoulli random variable with probability parameter p, which can take values 0 or 1, with p.m.f.

$$f(x) = (1-p)1_{\{x=0\}} + p1_{\{x=1\}}.$$

If X has this p.m.f., we write $X \sim Ber(p)$. A generalization of the Bernoulli distribution is the binomial distribution with parameters n and p, whose p.m.f. is

$$f(x) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} 1_{\{x=k\}}.$$

If X has this p.m.f., we write $X \sim Bin(n,p)$. Note that when $n = 1$, this is simply the Bernoulli distribution with parameter p.

The binomial distributions are, in some sense, discrete analogs of normal distributions. The discrete analogs of exponential distributions are geometric distributions. The geometric distribution with parameter p has p.m.f.

$$f(x) = \sum_{k=1}^{\infty} (1-p)^{k-1} p 1_{\{x=k\}}.$$

If a random variable X has this p.m.f., we write $X \sim Geo(p)$. Again, the reader probably knows already that this distribution models the waiting time for the first head in a sequence of coin tosses where the chance of heads is p.

EXERCISE 5.6.1. If $X_1, X_2, \ldots$ is a sequence of independent $Ber(p)$ random variables where $p \in (0,1)$, and $T := \min\{k : X_k = 1\}$, give a complete measure theoretic proof of the fact that $T \sim Geo(p)$.

Finally, a very important class of probability distributions are the Poisson distributions. The Poisson distribution with parameter $\lambda > 0$ is has p.m.f.

$$f(x) = \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} 1_{\{x=k\}}.$$

If X has this distribution, we write $X \sim Poi(\lambda)$.

5.7. Expected value

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable defined on Ω. The expected value (or expectation, or mean) of X, denoted by $\mathbb{E}(X)$, is simply the integral $\int X d\mathbb{P}$, provided that the integral exists. A random variable X is called integrable if it is integrable as a measurable function, that is, if $\mathbb{E}|X| < \infty$. A notation that is often used is that if X is a random variable and A is an event (defined on the same space), then

$$\mathbb{E}(X; A) := \mathbb{E}(X 1_A).$$

The following exercises show how to compute expected values in practice.

EXERCISE 5.7.1. If a random variable X has law μ_X, show that for any measurable $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) d\mu_X(x),$$

in the sense that one side exists if and only if the other does, in then the two are equal. (Hint: Start with simple functions.)

EXERCISE 5.7.2. If a random variable X takes values in a countable or finite set A, prove that for any $g : A \to \mathbb{R}$, $g(X)$ is also a random variable, and

$$\mathbb{E}(g(X)) = \sum_{a \in A} g(a)\mathbb{P}(X = a),$$

provided that at least one of $\sum g(a)_+\mathbb{P}(X = a)$ and $\sum g(a)_-\mathbb{P}(X = a)$ is finite.

EXERCISE 5.7.3. If X has p.d.f. f and $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, shows that $\mathbb{E}(g(X))$ exists if and only if the integral

$$\int_{-\infty}^{\infty} g(x)f(x)dx$$

exists in the Lebesgue sense, and in that case the two quantities are equal.

EXERCISE 5.7.4. Compute the expected values of normal, exponential, Gamma, uniform, Bernoulli, binomial, geometric and Poisson random variables.

We will sometimes make use of the following application of Fubini's theorem to compute expected values of integrals of functions of random variables.

EXERCISE 5.7.5. Let X be a random variable and let $f : \mathbb{R}^2 \to \mathbb{R}$ be a measurable function such that

$$\int_{-\infty}^{\infty} \mathbb{E}|f(t, X)|dt < \infty.$$

Then show that

$$\int_{-\infty}^{\infty} \mathbb{E}(f(t, X))dt = \mathbb{E}\bigg(\int_{-\infty}^{\infty} f(t, X)dt\bigg),$$

in the sense that both sides exist and are equal. (Of course, here $f(t, X)$ denotes the random variable $f(t, X(\omega))$.)

A very important representation of the expected value of a nonnegative random variable is given by the following exercise, which is a simple consequence of the previous one. It gives a way of calculating the expected value of a nonnegative random variable if we know its law.

EXERCISE 5.7.6. If X is a nonnegative random variable, prove that

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq t)dt.$$

A corollary of the above exercise is the following fact, which shows that the expected value is a property of the law of a random variable rather than the random variable itself.

EXERCISE 5.7.7. Prove that if two random variables X and Y have the same law, then $\mathbb{E}(X)$ exists if and only if $\mathbb{E}(Y)$ exists, and in this case they are equal.

An inequality related to Exercise 5.7.6 is the following. It is proved easily by the monotone convergence theorem.

EXERCISE 5.7.8. If X is a nonnegative random variable, prove that

$$\sum_{n=1}^\infty \mathbb{P}(X \geq n) \leq \mathbb{E}(X) \leq \sum_{n=0}^\infty \mathbb{P}(X \geq n),$$

with equality on the left if X is integer-valued.

5.8. Variance and covariance

The variance of X is defined as

$$\mathrm{Var}(X) := \mathbb{E}(X^2) - (\mathbb{E}(X))^2,$$

provided that $\mathbb{E}(X^2)$ is finite. Note that by the monotonicity of Hölder norms for probability measures, the finiteness of $\mathbb{E}(X^2)$ automatically implies the finiteness of $\mathbb{E}|X|$ and in particular the existence of $\mathbb{E}(X)$.

EXERCISE 5.8.1. Compute the variances of the normal, exponential, Gamma, uniform, Bernoulli, binomial, geometric and Poisson distributions.

It is not difficult to verify that

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2,$$

where $X - \mathbb{E}(X)$ denotes the random variable obtained by subtracting off the constant $\mathbb{E}(X)$ from X at each ω. Note that for any $a, b \in \mathbb{R}$, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$. When the variance exists, an important property of the expected value is that it is the constant a that minimizes $\mathbb{E}(X - a)^2$. This follows from the easy-to-prove identity

$$\mathbb{E}(X - a)^2 = \mathrm{Var}(X) + (\mathbb{E}(X) - a)^2.$$

The square-root of $\mathrm{Var}(X)$ is known as the standard deviation of X. A simple consequence of Minkowski's inequality is that the standard deviation of $X + Y$, where X and Y are two random variables defined on the same probability space, is bounded by the sum of the standard deviations of X and Y. A simple consequence of Markov's inequality is the following result,

which shows that X is likely to be within a few standard deviations from the mean.

THEOREM 5.8.2 (Chebychev's inequality). *Let X be any random variable with $\mathbb{E}(X^2) < \infty$. Then for any $t > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

PROOF. By Markov's inequality,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{t^2},$$

and recall that $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$. □

Chebychev's inequality is an example of what is known as a 'tail bound' for a random variable. Tail bounds are indispensable tools in modern probability theory.

Another very useful inequality involving the $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$ is the following. It gives a kind of inverse for Chebychev's inequality.

THEOREM 5.8.3 (Paley–Zygmund inequality). *Let X be a nonnegative random variable with $\mathbb{E}(X^2) < \infty$. Then for any $t \in [0, \mathbb{E}(X))$,*

$$\mathbb{P}(X > t) \geq \frac{(\mathbb{E}(X) - t)^2}{\mathbb{E}(X^2)}.$$

PROOF. Take any $t \in [0, \mathbb{E}(X))$. Let $Y := (X - t)_+$. Then

$$0 \leq \mathbb{E}(X - t) \leq \mathbb{E}(X - t)_+ = \mathbb{E}(Y) = \mathbb{E}(Y; Y > 0).$$

By the Cauchy–Schwarz inequality for L^2 space,

$$(\mathbb{E}(Y; Y > 0))^2 \leq \mathbb{E}(Y^2)\mathbb{P}(Y > 0) = \mathbb{E}(Y^2)\mathbb{P}(X > t).$$

The proof is completed by observing that $Y^2 \leq X^2$. □

The covariance of two random variables X and Y defined on the same probability space is defined as

$$\mathrm{Cov}(X, Y) := \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

provided that both X, Y and XY are integrable. Notice that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. It is straightforward to show that

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

From this, it follows by the Cauchy–Schwarz inequality for L^2 space that if $X, Y \in L^2$, then

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

In fact, the covariance itself is an inner product, on the Hilbert space obtained by quotienting L^2 by the subspace consisting of all a.e. constant random variables. In particular, the covariance is a bilinear functional on

L^2, and $\mathrm{Cov}(X, a) = 0$ for any random variable X and constant a (viewed as a random variable).

The correlation between X and Y is defined as

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

provided that the variances are nonzero. By the Cauchy–Schwarz inequality, the correlation always lies between -1 and 1. If the correlation is zero, we say that the random variables are uncorrelated.

EXERCISE 5.8.4. Show that $\mathrm{Corr}(X, Y) = 1$ if and only if $Y = aX + b$ for some $a > 0$ and $b \in \mathbb{R}$, and $\mathrm{Corr}(X, Y) = -1$ if and only if $Y = aX + b$ for some $a < 0$ and $b \in \mathbb{R}$.

An important formula involving the covariance is the following.

PROPOSITION 5.8.5. *For any $X_1, \ldots, X_n$ defined on the same space,*

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i,j=1}^n \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).\end{aligned}$$

PROOF. This is a direct consequence of the bilinearity of covariance. The second identity follows from the observations that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ and $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. □

5.9. Moments and moment generating function

For any positive integer k, the kth moment of a random variable X is defined as $\mathbb{E}(X^k)$, provided that this expectation exists.

EXERCISE 5.9.1. If $X \sim N(0, 1)$, show that

$$\mathbb{E}(X^k) = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ (k-1)!! & \text{if } k \text{ is even,} \end{cases}$$

where $(k-1)!! := (k-1)(k-3)\cdots 5 \cdot 3 \cdot 1$.

The moment generating function of X is defined as

$$m_X(t) := \mathbb{E}(e^{tX})$$

for $t \in \mathbb{R}$. Note that the moment generating function is allowed to take infinite values. The moment generating function derives its name from the fact that

$$m_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}(X^k)$$

whenever

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}|X|^k < \infty, \tag{5.9.1}$$

as can be easily verified using the monotone convergence theorem and the dominated convergence theorem.

EXERCISE 5.9.2. Carry out the above verification.

EXERCISE 5.9.3. If $X \sim N(\mu, \sigma^2)$, show that $m_X(t) = e^{t\mu + t^2\sigma^2/2}$.

Moments and moment generating functions provide tail bounds for random variables that are more powerful than Chebychev's inequality.

PROPOSITION 5.9.4. *For any random variable X, any $t > 0$, and any $p > 0$,*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|^p}{t^p}.$$

Moreover, for any $t \in \mathbb{R}$ and $\theta \geq 0$,

$$\mathbb{P}(X \geq t) \leq e^{-\theta t} m_X(\theta), \quad \mathbb{P}(X \leq t) \leq e^{\theta t} m_X(-\theta).$$

PROOF. All of these inequalities are simple consequences of Markov's inequality. For the first inequality, observe that $|X| \geq t$ if and only if $|X|^p \geq t^p$ and apply Markov's inequality. For the second and third, observe that $X \geq t$ if and only if $e^{\theta X} \geq e^{\theta t}$, and $X \leq t$ if and only if $e^{-\theta X} \geq e^{-\theta t}$. □

Often, it is possible to get impressive tail bounds in a given problem by optimizing over θ or p in the above result.

EXERCISE 5.9.5. If $X \sim N(0, \sigma^2)$, use the above procedure to prove that for any $t \geq 0$, $\mathbb{P}(X \geq t)$ and $\mathbb{P}(X \leq -t)$ are bounded by $e^{-t^2/2\sigma^2}$.

5.10. Characteristic function

Another important function associated with a random variable X is its characteristic function ϕ_X, defined as

$$\phi_X(t) := \mathbb{E}(e^{\mathrm{i}tX}),$$

where $\mathrm{i} = \sqrt{-1}$. Until now we have only dealt with expectations of random variables, but this is not much different. The right side of the above expression is defined simply as

$$\mathbb{E}(e^{\mathrm{i}tX}) := \mathbb{E}(\cos tX) + \mathrm{i}\mathbb{E}(\sin tX),$$

and the two expectations on the right always exist because cos and sin are bounded functions. In fact, the expected value of any complex random variable can be defined in the same manner, provided that the expected values of the real and imaginary parts exist and are finite.

PROPOSITION 5.10.1. *If X and Y are integrable random variables and $Z = X + \mathrm{i}Y$, and $\mathbb{E}(Z)$ is defined as $\mathbb{E}(X) + \mathrm{i}\mathbb{E}(Y)$, then $|\mathbb{E}(Z)| \leq \mathbb{E}|Z|$.*

PROOF. Let $\alpha := \mathbb{E}(Z)$ and $\bar{\alpha}$ denote the complex conjugate of α. It is easy to check the linearity of expectation holds for complex random variables. Thus,

$$\begin{aligned} |\mathbb{E}(Z)|^2 = \bar{\alpha}\mathbb{E}(Z) &= \mathbb{E}(\bar{\alpha}Z) \\ &= \mathbb{E}(\Re(\bar{\alpha}Z)) + \mathrm{i}\mathbb{E}(\Im(\bar{\alpha}Z)), \end{aligned}$$

where $\Re(z)$ and $\Im(z)$ denote the real and imaginary parts of a complex number z. Since $|\mathbb{E}(Z)|^2$ is real, the above identity show that $|\mathbb{E}(Z)|^2 = \mathbb{E}(\Re(\bar{\alpha}Z))$. But $\Re(\bar{\alpha}Z) \leq |\bar{\alpha}Z| = |\alpha||Z|$. Thus, $|\mathbb{E}(Z)|^2 \leq |\alpha|\mathbb{E}|Z|$, which completes the proof. □

The above proposition shows in particular that $|\phi_X(t)| \leq 1$ for any random variable X and any t.

EXERCISE 5.10.2. Show that the characteristic function can be a written as a power series in t if (5.9.1) holds.

EXERCISE 5.10.3. Show that the characteristic function of any random variable is a uniformly continuous function. (Hint: Use the dominated convergence theorem.)

Perhaps somewhat surprisingly, the characteristic function also gives a tail bound. This bound is not very useful, but we will see at least one fundamental application in a later chapter.

PROPOSITION 5.10.4. *Let X be a random variable with characteristic function ϕ_X. Then for any $t > 0$,*

$$\mathbb{P}(|X| \geq t) \leq \frac{t}{2} \int_{-2/t}^{2/t} (1 - \phi_X(s))ds.$$

PROOF. Note that for any $a > 0$,

$$\begin{aligned} \int_{-a}^{a} (1 - \phi_X(s))ds &= \int_0^a (2 - \phi_X(s) - \phi_X(-s))ds \\ &= \int_0^a \mathbb{E}(2 - e^{\mathrm{i}sX} - e^{-\mathrm{i}sX})ds \\ &= 2\int_0^a \mathbb{E}(1 - \cos sX)ds. \end{aligned}$$

By Fubini's theorem (specifically, Exercise 5.7.5), expectation and integral can be interchanged above, giving

$$\int_{-a}^{a} (1 - \phi_X(s))ds = 2a\mathbb{E}\left(1 - \frac{\sin aX}{aX}\right),$$

interpreting $(\sin x)/x = 1$ when $x = 0$. Now notice that

$$1 - \frac{\sin aX}{aX} \geq \begin{cases} 1/2 & \text{when } |X| \geq 2/a, \\ 0 & \text{always.} \end{cases}$$

Thus,

$$\mathbb{E}\left(1 - \frac{\sin aX}{aX}\right) \geq \frac{1}{2}\mathbb{P}(|X| \geq 2/a).$$

Taking $a = 2/t$, this proves the claim. □

EXERCISE 5.10.5. Compute the characteristic functions of the Bernoulli, binomial, geometric, Poisson, uniform, exponential and Gamma distributions.

5.11. Characteristic function of the normal distribution

The characteristic function of a standard normal random variable will be useful for us in the proof of the central limit theorem later. The calculation of this characteristic function is not entirely trivial; the standard derivation involves contour integration. The complete details are given below.

PROPOSITION 5.11.1. *If $X \sim N(0,1)$, then $\phi_X(t) = e^{-t^2/2}$.*

PROOF. Note that

$$\begin{aligned} \phi_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\mathrm{i}tx} e^{-x^2/2}\, dx \\ &= \frac{e^{-t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mathrm{i}t)^2/2}\, dx. \end{aligned}$$

Take any $R > 0$. Let C be the contour in the complex plane that forms the boundary of the box with vertices $-R, R, R - \mathrm{i}t, -R - \mathrm{i}t$. Since the map $z \mapsto e^{-z^2/2}$ is entire,

$$\oint_C e^{-z^2/2}\, dz = 0.$$

Let C_1 be the part of C that lies on the real line, going from left to right. Let C_2 be the part that is parallel to C_1 going from left to right. Let C_3 and C_4 be the vertical parts, going from top to bottom. It is easy to see that as $R \to \infty$,

$$\oint_{C_3} e^{-z^2/2}\, dz \to 0 \quad \text{and} \quad \oint_{C_4} e^{-z^2/2}\, dz \to 0.$$

Thus, as $R \to \infty$,

$$\oint_{C_1} e^{-z^2/2}\, dz - \oint_{C_2} e^{-z^2/2}\, dz \to 0.$$

Also, as $R \to \infty$,

$$\oint_{C_1} e^{-z^2/2}\, dz \to \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi},$$

and

$$\oint_{C_2} e^{-z^2/2}\,dz \to \int_{-\infty}^{\infty} e^{-(x-\mathrm{i}t)^2/2}\,dx.$$

This completes the proof. □

As a final remark of this chapter, we note that by Exercise 5.3.2, the expectation, variance, moments, moment generating function, and characteristic function of a random variable are all determined by its law. That is, if two random variables have the same law, then the above functionals are also the same for the two.

CHAPTER 6

Independence

A central idea of probability theory, which distinguishes it from measure theory, is the notion of independence. The reader may be already familiar with independent random variables from undergraduate probability classes. In this chapter we will bring the concept of independence into the measure-theoretic framework and derive some important consequences.

6.1. Definition

DEFINITION 6.1.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and suppose that $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are sub-σ-algebras of $\mathcal{F}$. We say that $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are independent if for any $A_1 \in \mathcal{G}_1$, $A_2 \in \mathcal{G}_2$, $\ldots$, $A_n \in \mathcal{G}_n$,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i).$$

More generally, an arbitrary collection $\{\mathcal{G}_i\}_{i \in I}$ of sub-σ-algebras are called independent if any finitely many of them are independent.

The independence of σ-algebras is used to define the independence of random variables and events.

DEFINITION 6.1.2. A collection of events $\{A_i\}_{i \in I}$ in $\mathcal{F}$ are said to be independent if the σ-algebras $\mathcal{G}_i := \{\emptyset, A_i, A_i^c, \Omega\}$ generated by the A_i's are independent. Moreover, an event A is said to be independent of a σ-algebra $\mathcal{G}$ is the σ-algebras $\{\emptyset, A, A^c, \Omega\}$ and $\mathcal{G}$ are independent.

EXERCISE 6.1.3. Show that a collection of events $\{A_i\}_{i \in I}$ are independent if and only if for any finite $J \subseteq I$,

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j).$$

DEFINITION 6.1.4. A collection of random variables $\{X_i\}_{i \in I}$ defined on Ω are said to be independent if $\{\sigma(X_i)\}_{i \in I}$ are independent sub-σ-algebras of $\mathcal{F}$. Moreover, a random variable X is said to be independent of a σ-algebra $\mathcal{G}$ if the σ-algebras $\sigma(X)$ and $\mathcal{G}$ are independent.

A particularly important definition in probability theory is the notion of an independent and identically distributed (i.i.d.) sequence of random

variables. A sequence $\{X_i\}_{i=1}^{\infty}$ is said to be i.i.d. if the X_i's are independent and all have the same distribution.

We end this section with a sequence of important exercises about independent random variables and events.

EXERCISE 6.1.5. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables defined on the same probability space. If X_{n+1} is independent of $\sigma(X_1, \ldots, X_n)$ for each n, prove that the whole collection is independent.

EXERCISE 6.1.6. If $X_1, \ldots, X_n$ are independent random variables and $f : \mathbb{R}^n \to \mathbb{R}$ is a measurable function, show that the law of $f(X_1, \ldots, X_n)$ is determined by the laws of $X_1, \ldots, X_n$.

EXERCISE 6.1.7. If $\{X_i\}_{i \in I}$ is a collection of independent random variables and $\{A_i\}_{i \in I}$ is a collection of measurable subsets of $\mathbb{R}$, show that the events $\{X_i \in A_i\}$, $i \in I$ are independent.

EXERCISE 6.1.8. If $\{F_i\}_{i \in I}$ is a family of cumulative distribution functions, show that there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and independent random variables $\{X_i\}_{i \in I}$ defined on Ω such that for each i, F_i is the c.d.f. of X_i. (Hint: Use product spaces.)

(The above exercise allows us to define arbitrary families of independent random variables on the same probability space. The usual convention in probability theory is to always have a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in the background, on which all random variables are defined. For convenience, this probability space is usually assumed to be complete.)

EXERCISE 6.1.9. If $\mathcal{A}$ is a π-system of sets and B is an event such that B and A are independent for every $A \in \mathcal{A}$, show that B and $\sigma(\mathcal{A})$ are independent.

EXERCISE 6.1.10. If $\{X_i\}_{i \in I}$ is a collection of independent random variables, then show that for any disjoint subsets $J, K \subseteq I$, the σ-algebras generated by $\{X_i\}_{i \in J}$ and $\{X_i\}_{i \in K}$ are independent. (Hint: Use the previous exercise.)

6.2. Expectation of a product under independence

A very important property of independent random variables is the following. Together with the above exercise, it gives a powerful computational tool for probabilistic models.

PROPOSITION 6.2.1. *If X and Y are independent random variables such that X and Y are integrable, then the product XY is also integrable and $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. The identity also holds if X and Y are nonnegative but not necessarily integrable.*

PROOF. First, suppose that $X = \sum_{i=1}^k a_i 1_{A_i}$ and $Y = \sum_{j=1}^m b_j 1_{B_j}$ for some nonnegative a_i's and b_j's, and measurable A_i's and B_j's. Without loss of generality, assume that all the a_i's are distinct and all the b_j's are distinct. Then each $A_i \in \sigma(X)$ and each $B_j \in \sigma(Y)$, because $A_i = X^{-1}(\{a_i\})$ and $B_j = Y^{-1}(\{b_j\})$. Thus, for each i and j, A_i and B_j are independent. This gives

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{i=1}^k \sum_{j=1}^m a_i b_j \mathbb{E}(1_{A_i} 1_{B_j}) = \sum_{i=1}^k \sum_{j=1}^m a_i b_j \mathbb{P}(A_i \cap B_j) \\ &= \sum_{i=1}^k \sum_{j=1}^m a_i b_j \mathbb{P}(A_i)\mathbb{P}(B_j) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Next take any nonnegative X and Y that are independent. Construct nonnegative simple random variables X_n and Y_n increasing to X and Y, using the method from the proof of Proposition 2.3.7. From the construction, it is easy to see that each X_n is $\sigma(X)$-measurable and each Y_n is $\sigma(Y)$-measurable. Therefore X_n and Y_n are independent, and hence $\mathbb{E}(X_n Y_n) = \mathbb{E}(X_n)\mathbb{E}(Y_n)$, since we have already proved this identity for nonnegative simple random variables. Now note that since $X_n \uparrow X$ and $Y_n \uparrow Y$, we have $X_n Y_n \uparrow XY$. Therefore by the monotone convergence theorem,

$$\mathbb{E}(XY) = \lim_{n\to\infty} \mathbb{E}(X_n Y_n) = \lim_{n\to\infty} \mathbb{E}(X_n)\mathbb{E}(Y_n) = \mathbb{E}(X)\mathbb{E}(Y)$$

Finally, take any independent X and Y. It is easy to see that X_+ and X_- are $\sigma(X)$-measurable, and Y_+ and Y_- are $\sigma(Y)$-measurable. Therefore

$$\begin{aligned}\mathbb{E}|XY| &= \mathbb{E}((X_+ + X_-)(Y_+ + Y_-)) \\ &= \mathbb{E}(X_+Y_+) + \mathbb{E}(X_+Y_-) + \mathbb{E}(X_-Y_+) + \mathbb{E}(X_-Y_-) \\ &= \mathbb{E}(X_+)\mathbb{E}(Y_+) + \mathbb{E}(X_+)\mathbb{E}(Y_-) + \mathbb{E}(X_-)\mathbb{E}(Y_+) + \mathbb{E}(X_-)\mathbb{E}(Y_-) \\ &= \mathbb{E}|X|\mathbb{E}|Y|.\end{aligned}$$

Since $\mathbb{E}|X|$ and $\mathbb{E}|Y|$ are both finite, this shows that XY is integrable. Repeating the steps in the above display starting with $\mathbb{E}(XY)$ instead of $\mathbb{E}|XY|$, we get $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. □

The following exercise generalizes the above proposition. It is provable easily by induction, starting with the case $n = 2$.

EXERCISE 6.2.2. If $X_1, X_2, \ldots, X_n$ are independent integrable random variables, show that the product $X_1 X_2 \cdots X_n$ is also integrable and

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\cdots\mathbb{E}(X_n).$$

Moreover, show that the identity also holds if the X_i's are nonnegative but not necessarily integrable.

The following are some important consequences of the above exercise.

EXERCISE 6.2.3. If X and Y are independent integrable random variables, show that $\mathrm{Cov}(X, Y) = 0$. In other words, independent random variables are uncorrelated.

EXERCISE 6.2.4. Give an example of a pair of uncorrelated random variables that are not independent.

A collection of random variables $\{X_i\}_{i\in I}$ is called pairwise independent if for any two distinct $i, j \in I$, X_i and X_j are independent.

EXERCISE 6.2.5. Give an example of three random variables X_1, X_2, X_3 that are pairwise independent but not independent.

6.3. The second Borel–Cantelli lemma

Let $\{A_n\}_{n\geq 1}$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The event $\{A_n \text{ i.o.}\}$ denotes the set of all ω that belong to infinitely many of the A_n's. Here 'i.o.' means 'infinitely often'. In this language, the first Borel–Cantelli says that if $\sum \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$. By Exercise 4.3.2, we know that the converse of the lemma is not true. The second Borel–Cantelli lemma, stated below, says that the converse is true if we additionally impose the condition that the events are independent. Although not as useful as the first lemma, it has some uses.

THEOREM 6.3.1 (The second Borel–Cantelli lemma). *If $\{A_n\}_{n\geq 1}$ is a sequence of independent events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\sum \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.*

PROOF. Let B denote the event $\{A \text{ i.o.}\}$. Then B^c is the set of all ω that belong to only finitely many of the A_n's. In set theoretic notation,

$$B^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c.$$

Therefore

$$\mathbb{P}(B^c) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big).$$

Take any $1 \leq n \leq m$. Then by independence,

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) &\leq \mathbb{P}\Big(\bigcap_{k=n}^{m} A_k^c\Big) \\ &= \prod_{k=n}^{m} (1 - \mathbb{P}(A_k)).\end{aligned}$$

By the inequality $1 - x \leq e^{-x}$ that holds for all $x \geq 0$, this gives

$$\mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) \leq \exp\Big(-\sum_{k=n}^{m} \mathbb{P}(A_k)\Big).$$

Since this holds for any $m \geq n$, we get

$$\mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) \leq \exp\Big(-\sum_{k=n}^{\infty} \mathbb{P}(A_k)\Big) = 0,$$

where the last equality holds because $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \infty$. This shows that $\mathbb{P}(B^c) = 0$ and completes the proof of the theorem. □

EXERCISE 6.3.2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables such that $\mathbb{E}|X_1| = \infty$. Prove that $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$.

EXERCISE 6.3.3. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of integer-valued random variables. Take any m and any sequence of integers $k_1, k_2, \ldots, k_m$ such that $\mathbb{P}(X_1 = k_i) > 0$ for each i. Prove that with probability 1, there are infinitely many occurrences of the sequence $k_1, \ldots, k_m$ in a realization of $X_1, X_2, \ldots$.

6.4. The Kolmogorov zero-one law

Let $X_1, X_2, \ldots$ be a sequence of random variables defined on the same probability space. The tail σ-algebra generated by this family is defined as

$$\mathcal{T}(X_1, X_2, \ldots) := \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots).$$

The following result is often useful for proving that some random variable is actually a constant.

THEOREM 6.4.1 (Kolmogorov's zero-one law). *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{T}$ is the tail σ-algebra of this sequence, then for any $A \in \mathcal{T}$, $\mathbb{P}(A)$ is either 0 or 1.*

PROOF. Take any n. Since $A \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ and the X_i's are independent, it follows by Exercise 6.1.10 that the event A is independent of the σ-algebra $\sigma(X_1, \ldots, X_n)$. Let

$$\mathcal{A} := \bigcup_{n=1}^{\infty} \sigma(X_1, \ldots, X_n).$$

It is easy to see that $\mathcal{A}$ is an algebra, and $\sigma(\mathcal{A}) = \sigma(X_1, X_2, \ldots)$. From the first paragraph we know that A is independent of B for every $B \in \mathcal{A}$. Therefore by Exercise 6.1.9, A is independent of $\sigma(X_1, X_2, \ldots)$. In particular, A is independent of itself, which implies that $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$. This proves that $\mathbb{P}(A)$ is either 0 or 1. □

Another way to state Kolmogorov's zero-one law is to say that the tail σ-algebra of a sequence of independent random variables is 'trivial', in the sense that any event in it has probability 0 or 1. Trivial σ-algebras have the following useful property.

EXERCISE 6.4.2. If $\mathcal{G}$ is a trivial σ-algebra for a probability measure $\mathbb{P}$, show that any random variable X that is measurable with respect to $\mathcal{G}$ must be equal to a constant almost surely.

In particular, if $\{X_n\}_{n=1}^\infty$ is a sequence of independent random variables and X is a random variable that is measurable with respect to the tail σ-algebra of this sequence, then show that there is some constant c such that $X = c$ a.e. Some important consequences are the following.

EXERCISE 6.4.3. If $\{X_n\}_{n=1}^\infty$ is a sequence of independent random variables, show that the random variables $\limsup X_n$ and $\liminf X_n$ are equal to constants almost surely.

EXERCISE 6.4.4. Let $\{X_n\}_{n=1}^\infty$ be a sequence of independent random variables. Let $S_n := X_1 + \cdots + X_n$, and let $\{a_n\}_{n=1}^\infty$ be a sequence of constants increasing to infinity. Then show that $\limsup S_n/a_n$ and $\liminf S_n/a_n$ are constants almost surely.

6.5. Zero-one laws for i.i.d. random variables

For i.i.d. random variables, one can prove a zero-one law that is quite a bit stronger than Kolmogorov's zero-one law.

THEOREM 6.5.1. *Let $X = \{X_i\}_{i\in I}$ be a countable collection of i.i.d. random variables. Suppose that $f : \mathbb{R}^I \to \mathbb{R}$ is a measurable function with the following property. There is a collection Γ of one-to-one maps from I into I, such that*

(i) *$f(\omega^\gamma) = f(\omega)$ for each $\gamma \in \Gamma$ and $\omega \in \mathbb{R}^I$, where $\omega_i^\gamma := \omega_{\gamma(i)}$, and*
(ii) *for any finite $J \subseteq I$, there is some $\gamma \in \Gamma$ such that $\gamma(J) \cap J = \emptyset$.*

Then the random variable $f(X)$ equals a constant almost surely.

PROOF. First, assume that f is the indicator function of a measurable set $E \subseteq \mathbb{R}^I$. Let A be the event that $X \in E$. Take any $\epsilon > 0$. Let $\{i_1, i_2, \ldots\}$ be an enumeration of I. By Exercise 5.1.3, there is some n and some $B \in \sigma(X_{i_1}, \ldots, X_{i_n})$ such that $\mathbb{P}(A \Delta B) < \epsilon$, where $\mathbb{P}$ is the probability measure on the space where X is defined. By Exercise 5.1.4, there is some measurable function $g : \mathbb{R}^n \to \mathbb{R}$ such that $1_B = g(X_{i_1}, \ldots, X_{i_n})$. Then $\mathbb{P}(A\Delta B) < \epsilon$ is equivalent to saying that $\mathbb{E}|1_A - 1_B| < \epsilon$. Using condition (ii), find $\gamma \in \Gamma$ such that $\{\gamma(i_1), \ldots, \gamma(i_n)\}$ does not intersect $\{i_1, \ldots, i_n\}$. Let $Y := X^\gamma$, $Z := f(Y)$ and $W := g(Y_{i_1}, \ldots, Y_{i_n})$. Since the X_i's are i.i.d. and γ is an injection, Y has the same law as X. Therefore

$$\mathbb{E}|Z - W| = \mathbb{E}|f(X) - g(X_{i_1}, \ldots, X_{i_n})| = \mathbb{E}|1_A - 1_B| < \epsilon.$$

But by condition (i), $f(X) = f(Y)$. Therefore

$$\mathbb{E}|1_B - W| \le \mathbb{E}|1_B - 1_A| + \mathbb{E}|Z - W| < 2\epsilon.$$

On the other hand, notice that 1_B and W are independent and identically distributed random variables. Moreover, both are $\{0,1\}$-valued. Therefore

$$\begin{aligned} 2\epsilon > \mathbb{E}|1_B - W| \geq \mathbb{E}(1_B - W)^2 &= \mathbb{E}(1_B + W - 21_B W) \\ &= 2\mathbb{P}(B) - 2\mathbb{P}(B)\mathbb{E}(W) = 2\mathbb{P}(B)(1 - \mathbb{P}(B)). \end{aligned}$$

On the other hand,

$$\begin{aligned} |\mathbb{P}(A)(1-\mathbb{P}(A)) - \mathbb{P}(B)(1-\mathbb{P}(B))| &\leq |\mathbb{P}(A) - \mathbb{P}(B)| \\ &\leq \mathbb{E}|1_A - 1_B| < \epsilon, \end{aligned}$$

since the derivative of the map $x \mapsto x(1-x)$ is bounded by 1 in $[0,1]$. Combining, we get $\mathbb{P}(A)(1-\mathbb{P}(A)) < 3\epsilon$. Since this is true for any $\epsilon > 0$, we must have that $\mathbb{P}(A) = 0$ or 1. In particular, $f(X)$ is a constant.

Now take an arbitrary measurable f with the given properties. Take any $t \in \mathbb{R}$ and let E be the set of all $\omega \in \mathbb{R}^I$ such that $f(\omega) \leq t$. Then the function 1_E also satisfies the hypotheses of the theorem, and so we can apply the first part to conclude that $1_E(X)$ is a constant. Since this holds for any t, we may conclude that $f(X)$ is a constant. □

The following result is a useful consequence of Theorem 6.5.1.

COROLLARY 6.5.2 (Zero-one law for translation-invariant events). *Let $\mathbb{Z}^d$ be the d-dimensional integer lattice, for some $d \geq 1$. Let E be the set of nearest-neighbor edges of this lattice, and let $\{X_e\}_{e\in E}$ be a collection of i.i.d. random variables. Let Γ be the set of all translations of E. Then any measurable function f of $\{X_e\}_{e\in E}$ which is translation-invariant, in the sense that $f(\{X_e\}_{e\in E}) = f(\{X_{\gamma(e)}\}_{e\in E})$ for any $\gamma \in \Gamma$, must be equal to a constant almost surely. The same result holds if edges are replaced by vertices.*

PROOF. Clearly Γ satisfies property (ii) in Theorem 6.5.1 for any chosen enumeration of the edges. Property (i) is satisfied by the hypothesis of the corollary. □

The following exercise demonstrates an important application of Corollary 6.5.2.

EXERCISE 6.5.3. Let $\{X_e\}_{e\in E}$ be a collection of i.i.d. $Ber(p)$ random variables, for some $p \in [0,1]$. Define a random subgraph of $\mathbb{Z}^d$ by keeping an edge e if $X_e = 1$ and deleting it if $X_e = 0$. Let N be the number of infinite connected components of this random graph. (These are known as infinite percolation clusters.) Exercise 3.3.2 shows that N is a random variable. Using Theorem 6.5.1 and the above discussion, prove that N equals a constant in $\{0,1,2,\ldots\} \cup \{\infty\}$ almost surely.

Another consequence of Theorem 6.5.1 is the following result.

COROLLARY 6.5.4 (Hewitt–Savage zero-one law). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Let f be a measurable function of this sequence that has the property that $f(X_1, X_2, \ldots) = f(X_{\sigma(1)}, X_{\sigma(2)}, \ldots)$ for any σ that permutes finitely many of the indices. Then $f(X_1, X_2, \ldots)$ must be equal to a constant almost surely.*

PROOF. Here Γ is the set of all permutations of the positive integers that fix all but finitely many indices. Then clearly Γ satisfies the hypotheses of Theorem 6.5.1. □

A typical application of the Hewitt–Savage zero-one law is the following.

EXERCISE 6.5.5. Suppose that we have m boxes, and an infinite sequence of balls are dropped into the boxes independently and uniformly at random. Set this up as a problem in measure-theoretic probability, and prove that with probability one, each box has the maximum number of balls among all boxes infinitely often.

6.6. Random vectors

A random vector is simply an $\mathbb{R}^n$-valued random variable for some positibe integer n. In this way, it generalizes the notion of a random variable to multiple dimensions. The cumulative distribution function of a random vector $X = (X_1, \ldots, X_n)$ is defined as

$$F_X(t_1, \ldots, t_n) = \mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots, X_n \leq t_n),$$

where $\mathbb{P}$ denotes the probability measure on the probability space on which X is defined. The probability density function of X, if it exists, is a measurable function $f : \mathbb{R}^n \to [0, \infty)$ such that for any $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\mathbb{P}(X \in A) = \int_A f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

where $dx_1 \cdots dx_n$ denotes integration with respect to Lebesgue measure on $\mathbb{R}^n$. The characteristic function is defined as

$$\phi_X(t_1, \ldots, t_n) := \mathbb{E}(e^{\mathrm{i}(t_1 X_1 + \cdots + t_n X_n)}).$$

The law μ_X of X the probability measure on $\mathbb{R}^n$ induced by X, that is, $\mu_X(A) := \mathbb{P}(X \in A)$. The following exercises are basic.

EXERCISE 6.6.1. If $X_1, \ldots, X_n$ are independent random variables with laws $\mu_1, \ldots, \mu_n$, then show that the law of the random vector $(X_1, \ldots, X_n)$ is the product measure $\mu_1 \times \cdots \times \mu_n$.

EXERCISE 6.6.2. If $X_1, \ldots, X_n$ are independent random variables, show that the cumulative distribution function and the characteristic function of $(X_1, \ldots, X_n)$ can be written as products of one-dimensional distribution functions and characteristic functions.

EXERCISE 6.6.3. If $X_1, \ldots, X_n$ are independent random variables, and each has a probability density function, show that $(X_1, \ldots, X_n)$ also has a p.d.f. and it is given by a product formula.

The mean vector of a random vector $X = (X_1, \ldots, X_n)$ is the vector $\mu = (\mu_1, \ldots, \mu_n)$, where $\mu_i = \mathbb{E}(X_i)$, assuming that the means exist. The covariance matrix of a random vector $X = (X_1, \ldots, X_n)$ is the $n \times n$ matrix $\Sigma = (\sigma_{ij})_{i,j=1}^n$, where $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, provided that these covariances exist.

EXERCISE 6.6.4. Prove that the covariance matrix of any random vector is a positive semi-definite matrix.

An important kind of random vector is the multivariate normal (or Gaussian) random vector. Given $\mu \in \mathbb{R}^n$ and a strictly positive definite matrix Σ of order n, the multivariate normal distribution with mean μ and covariance matrix Σ is the probability measure on $\mathbb{R}^n$ with probability density function

$$\frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

If X has this distribution, we write $X \sim N_n(\mu, \Sigma)$.

EXERCISE 6.6.5. Show that the above formula indeed describes a p.d.f. of a probability law on $\mathbb{R}^n$, and that this law has mean vector μ and covariance matrix Σ.

EXERCISE 6.6.6. Let $X \sim N_n(\mu, \Sigma)$, and let m be a positive integer $\leq n$. Show that for any $a \in \mathbb{R}^m$ and any $m \times n$ matrix A of full rank, $AX + a \sim N_m(a + A\mu, A\Sigma A^T)$.

6.7. Convolutions

Given two probability measures μ_1 and μ_2 on $\mathbb{R}$, their convolution $\mu_1 * \mu_2$ is defined to be the push-forward of $\mu_1 \times \mu_2$ under the addition map from $\mathbb{R}^2$ to $\mathbb{R}$. That is, if $\phi(x, y) := x + y$, then for any $A \in \mathcal{B}(\mathbb{R})$,

$$\mu_1 * \mu_2(A) := \mu_1 \times \mu_2(\phi^{-1}(A)).$$

Exercise 6.6.1 shows that in the language of random variables, the convolution of two probability measures has the following description: If X and Y are independent random variables with laws μ_1 and μ_2, then $\mu_1 * \mu_2$ is the law of $X + Y$.

PROPOSITION 6.7.1. *Let X and Y be independent random variables. Suppose that Y has probability density function g. Then the sum $Z := X + Y$ has probability density function $h(z) = \mathbb{E}g(z - X)$.*

PROOF. Let μ_1 be the law of X and μ_2 be the law of Y. Let $\mu := \mu_1 \times \mu_2$. Then the discussion preceding the statement of the proposition shows that for any $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\mathbb{P}(Z \in A) &= \mu_1 * \mu_2(A) = \mu(\phi^{-1}(A)) \\ &= \int_{\mathbb{R}^2} 1_{\phi^{-1}(A)}(x,y) d\mu(x,y).\end{aligned}$$

By Fubini's theorem, this integral equals

$$\int_{\mathbb{R}} \int_{\mathbb{R}} 1_{\phi^{-1}(A)}(x,y) d\mu_2(y) d\mu_1(x).$$

But for any x,

$$\begin{aligned}\int_{\mathbb{R}} 1_{\phi^{-1}(A)}(x,y) d\mu_2(y) &= \int_{\mathbb{R}} 1_A(x+y) d\mu_2(y) \\ &= \int_{\mathbb{R}} 1_A(x+y) g(y) dy \\ &= \int_{\mathbb{R}} 1_A(z) g(z-x) dz = \int_A g(z-x) dz,\end{aligned}$$

where the last step follows by the translation invariance of Lebesgue measure (Exercise 2.4.6). Thus again by Fubini's theorem,

$$\begin{aligned}\mathbb{P}(Z \in A) &= \int_{\mathbb{R}} \int_A g(z-x) dz d\mu_1(x) \\ &= \int_A \int_{\mathbb{R}} g(z-x) d\mu_1(x) dz \\ &= \int_A \mathbb{E} g(z-X) dz.\end{aligned}$$

This proves the claim. □

EXERCISE 6.7.2. If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, prove that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

EXERCISE 6.7.3. As a consequence of the above exercise, prove that any linear combination of independent normal random variables is normal with the appropriate mean and variance.

Often, we need to deal with n-fold convolutions rather than the convolution of just two probability measures. The following exercises are two useful results about n-fold convolutions.

EXERCISE 6.7.4. If $X_1, X_2, \ldots$ is a sequence of independent $Ber(p)$ random variables, and $S_n := \sum_{i=1}^n X_i$, give a complete measure theoretic proof of the fact that $S_n \sim Bin(n,p)$.

EXERCISE 6.7.5. Use induction on n and the above convolution formula to prove that if $X_1, \ldots, X_n$ are i.i.d. $Exp(\lambda)$ random variables, then $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$.

EXERCISE 6.7.6. If $X_1, \ldots, X_n$ are independent random variables in L^2, show that $\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$.

EXERCISE 6.7.7. If $X_1, X_2, \ldots, X_n$ are independent random variables and $S = \sum X_i$, show that the moment generating function m_S and the characteristic function ϕ_S are given by the product formulas $m_S(t) = \prod m_{X_i}(t)$ and $\phi_S(t) = \prod \phi_{X_i}(t)$.

CHAPTER 7

Convergence of random variables

This chapter discusses various notions of convergence of random variables, laws of large numbers, and central limit theorems.

7.1. Four notions of convergence

Random variables can converge to limits in various ways. Four prominent definitions are the following.

DEFINITION 7.1.1. A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to converge to a random variable X almost surely if all of these random variables are defined on the same probability space, and $\lim X_n = X$ a.e. This is often written as $X_n \to X$ a.s.

DEFINITION 7.1.2. A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to converge in probability to a random variable X if all of these random variables are defined on the same probability space, and for each $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

This is usually written as $X_n \xrightarrow{p} X$ or $X_n \to X$ in probability. If X is a constant, then $\{X_n\}_{n\geq 1}$ need not be all defined on the same probability space.

DEFINITION 7.1.3. For $p \in [1, \infty]$, sequence of random variables $\{X_n\}_{n\geq 1}$ is said to converge in L^p to a random variable X if all of these random variables are defined on the same probability space, and

$$\|X_n - X\|_{L^p} = 0.$$

This is usually written as $X_n \xrightarrow{L^p} X$ or $X_n \to X$ in L^p. If X is a constant, then $\{X_n\}_{n\geq 1}$ need not be all defined on the same probability space.

DEFINITION 7.1.4. For each n, let X_n be a random variable with cumulative distribution function F_{X_n}. Let X be a random variable with c.d.f. F. Then X_n is said to converge in distribution to X if for any t where F_X is continuous,

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t).$$

This is usually written as $X_n \xrightarrow{d} X$, or $X_n \xrightarrow{\mathcal{D}} X$, or $X_n \Rightarrow X$, or $X_n \rightharpoonup X$, or $X_n \to X$ in distribution. Convergence in distribution is sometimes called convergence in law or weak convergence.

7.2. Interrelations between the four notions

The four notions of convergence defined above are inter-related in interesting ways.

PROPOSITION 7.2.1. *Almost sure convergence implies convergence in probability.*

PROOF. Let X_n be a sequence converging a.s. to X. Take any $\epsilon > 0$. Since $X_n \to X$ a.s.,

$$\begin{aligned} 1 &= \mathbb{P}\Big(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}\{|X_k - X| \le \epsilon\}\Big) \\ &= \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty}\{|X_k - X| \le \epsilon\}\Big) \\ &\le \lim_{n\to\infty} \mathbb{P}(|X_n - X| \le \epsilon), \end{aligned}$$

which proves the claim. □

EXERCISE 7.2.2. Give a counterexample to show that convergence in probability does not imply almost sure convergence.

Although convergence in probability does not imply convergence almost surely, it does imply that there is a subsequence that converges almost surely.

PROPOSITION 7.2.3. *If $\{X_n\}_{n\ge1}$ is a sequence of random variables converging in probability to a limit X, then there is a subsequence $\{X_{n_k}\}_{k\ge1}$ converging almost surely to X.*

PROOF. Since $X_n \to X$ in probability, it is not hard to see that there is a subsequence $\{X_{n_k}\}_{k\ge1}$ such that for each k, $\mathbb{P}(|X_{n_k} - X_{n_{k+1}}| > 2^{-k}) \le 2^{-k}$. Therefore by the Borel–Cantelli lemma, $\mathbb{P}(|X_{n_k} - X_{n_{k+1}}| > 2^{-k} \text{ i.o.}) = 0$. However, if $|X_{n_k}(\omega) - X_{n_{k+1}}(\omega)| > 2^{-k}$ happens only finitely many times for some ω, then $\{X_{n_k}(\omega)\}_{k\ge1}$ is a Cauchy sequence. Let $Y(\omega)$ denote the limit. On the set where this does not happen, define $Y = 0$. Then Y is a random variable, and $X_{n_k} \to Y$ a.s. Then by Proposition 7.2.1, $X_{n_k} \to Y$ in probability. But, for any $\epsilon > 0$ and any k,

$$\mathbb{P}(|X - Y| \ge \epsilon) \le \mathbb{P}(|X - X_{n_k}| \ge \epsilon/2) + \mathbb{P}(|Y - X_{n_k}| \ge \epsilon/2).$$

Sending $k \to \infty$, this shows that $\mathbb{P}(|X - Y| \ge \epsilon) = 0$. Since this holds for every $\epsilon > 0$, we get $X = Y$ a.s. This completes the proof. □

Next, let us connect convergence in probability and convergence in distribution.

PROPOSITION 7.2.4. *Convergence in probability implies convergence in distribution.*

PROOF. Let X_n be a sequence of random variables converging in probability to a random variable X. Let t be a continuity point of F_X. Take any $\epsilon > 0$. Then

$$\begin{aligned} F_{X_n}(t) &= \mathbb{P}(X_n \leq t) \\ &\leq \mathbb{P}(X \leq t + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon), \end{aligned}$$

which proves that $\limsup F_{X_n}(t) \leq F_X(t + \epsilon)$. Since this is true for any $\epsilon > 0$ and F_X is right continuous, this gives $\limsup F_{X_n}(t) \leq F_X(t)$. A similar argument gives $\liminf F_{X_n}(t) \geq F_X(t)$. □

EXERCISE 7.2.5. Show that the above proposition is not valid if we demanded that $F_{X_n}(t) \to F_X(t)$ for all t, instead of just the continuity points of F_X.

EXERCISE 7.2.6. If $X_n \to c$ in distribution, where c is a constant, show that $X_n \to c$ in probability.

The following result combines weak convergence and convergence in probability is a way that is useful for many purposes.

PROPOSITION 7.2.7 (Slutsky's theorem). *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then $X_n + Y_n \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$.*

PROOF. Let F be the c.d.f. of $X + c$. Let t be a continuity point of F. For any $\epsilon > 0$,

$$\mathbb{P}(X_n + Y_n \leq t) \leq \mathbb{P}(X_n + c \leq t + \epsilon) + \mathbb{P}(Y_n - c < -\epsilon).$$

If $t + \epsilon$ is also a continuity point of F, this shows that

$$\limsup_{n\to\infty} \mathbb{P}(X_n + Y_n \leq t) \leq F(t + \epsilon).$$

By Exercise 5.2.3 and the right-continuity of F, this allows us to conclude that

$$\limsup_{n\to\infty} \mathbb{P}(X_n + Y_n \leq t) \leq F(t).$$

Next, take any $\epsilon > 0$ such that $t - \epsilon$ is a continuity point of F. Since

$$\mathbb{P}(X_n + c \leq t - \epsilon) \leq \mathbb{P}(X_n + Y_n \leq t) + \mathbb{P}(Y_n - c > \epsilon),$$

we get

$$\liminf_{n\to\infty} \mathbb{P}(X_n + Y_n \leq t) \geq F(t - \epsilon).$$

By Exercise 5.2.3 and the continuity of F at t, this gives

$$\liminf_{n\to\infty} \mathbb{P}(X_n + Y_n \leq t) \geq F(t).$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}(X_n + Y_n \leq t) = F(t)$$

for every continuity point t of F, and hence $X_n + Y_n \xrightarrow{d} X + c$. The proof of $X_n Y_n \xrightarrow{d} cX$ is similar, with a slight difference in the case $c = 0$. □

Finally, let us look at the relation between L^p convergence and convergence in probability.

PROPOSITION 7.2.8. *For any $p > 0$, convergence in L^p implies convergence in probability.*

PROOF. Suppose that $X_n \to X$ in L^p. Take any $\epsilon > 0$. By Markov's inequality,

$$\begin{aligned}\mathbb{P}(|X_n - X| > \epsilon) &= \mathbb{P}(|X_n - X|^p > \epsilon^p) \\ &\leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p},\end{aligned}$$

which proves the claim. □

The converse of the above proposition holds under an additional assumption.

PROPOSITION 7.2.9. *If $X_n \to X$ in probability and there is some constant c such that $|X_n| \leq c$ a.s. for each n, then $X_n \to X$ in L^p for any $p \in [1, \infty)$.*

PROOF. It is easy to show from the given condition that $|X| \leq c$ a.s. Take any $\epsilon > 0$. Then

$$\begin{aligned}\mathbb{E}|X_n - X|^p &\leq \mathbb{E}(|X_n - X|^p; |X_n - X| > \epsilon) + \epsilon^p \\ &\leq (2c)^p \mathbb{P}(|X_n - X| > \epsilon) + \epsilon^p.\end{aligned}$$

Sending $n \to \infty$, we get $\limsup \mathbb{E}|X_n - X|^p \leq \epsilon^p$. Since ϵ is arbitrary, this completes the proof. □

Interestingly, there is no direct connection between convergence in L^p and almost sure convergence.

EXERCISE 7.2.10. Take any $p > 0$. Give counterexamples to show that almost sure convergence does not imply L^p convergence, and L^p convergence does not imply almost sure convergence.

7.3. The weak law of large numbers

The weak law of large numbers is a fundamental result of probability theory. Perhaps the best way to state the result is to state a quantitative version. It says that the average of a finite collection of random variables is close to the average of the expected values with high probability if the average of the covariances is small. This allows wide applicability in a variety of problems.

THEOREM 7.3.1 (Weak law of large numbers). *Let $X_1, \ldots, X_n$ be L^2 random variables defined on the same probability space. Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_{ij} := \mathrm{Cov}(X_i, X_j)$. Then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2 n^2}\sum_{i,j=1}^{n} \sigma_{ij}.$$

PROOF. Apply Chebychev's inequality, together with the formula given by Proposition 5.8.5 for the variance of a sum of random variables. □

An immediate corollary is the following theorem, which is traditionally known as the L^2 weak law of large numbers.

COROLLARY 7.3.2. *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of uncorrelated random variables with common mean μ and uniformly bounded finite second moment, then $n^{-1}\sum_{i=1}^{n} X_i$ converges in probability to μ as $n \to \infty$.*

Actually, the above theorem holds true even if the second moment is not finite, provided that the sequence is i.i.d. Since this is a simple consequence of the strong law of large numbers that we will prove later, we will not worry about it here.

EXERCISE 7.3.3 (An occupancy problem). Let n balls be dropped uniformly and independently at random into n boxes. Let N_n be the number of empty boxes. Prove that $N_n/n \to e^{-1}$ in probability as $n \to \infty$. (Hint: Write N_n as a sum of indicator variables.)

EXERCISE 7.3.4 (Coupon collector's problem). Suppose that there are n types of coupons, and a collector wants to obtain at least one of each type. Each time a coupon is bought, it is one of the n types with equal probability. Let T_n be the number of trials needed to acquire all n types. Prove that $T_n/(n \log n) \to 1$ in probability as $n \to \infty$. (Hint: Let τ_k be the number of trials needed to acquire k distinct types of coupons. Prove that $\tau_k - \tau_{k-1}$ are independent geometric random variables with different means, and T_n is the sum of these variables.)

EXERCISE 7.3.5 (Erdős–Rényi random graphs). Define an undirected random graph on n vertices by putting an edge between any two vertices with probability p and excluding the edge with probability $1-p$, all edges independent. This is known as the Erdős–Rényi $G(n,p)$ random graph. First, formulate the model in the measure theoretic framework using independent Bernoulli random variables. Next, show that if $T_{n,p}$ is the number of triangles in this random graph, then $T_{n,p}/n^3 \to p^3/6$ in probability as $n \to \infty$, if p remains fixed.

7.4. The strong law of large numbers

The strong law of large numbers is the almost sure version of the weak law. The best version of the strong law was proved by Etemadi.

THEOREM 7.4.1 (Etemadi's strong law of large numbers). *Let $\{X_n\}_{n\geq 1}$ be a sequence of pairwise independent and identically distributed random variables, with $\mathbb{E}|X_1| < \infty$. Then $n^{-1}\sum_{i=1}^n X_i$ converges almost surely to $\mathbb{E}(X_1)$ as $n \to \infty$.*

PROOF. Splitting each X_i into its positive and negative parts, we see that it suffices to prove the theorem for nonnegative random variables. So assume that the X_i's are nonnegative random variables.

The next step is to truncate the X_i's to produce random variables that are more well-behaved with respect to variance computations. Define $Y_i := X_i 1_{\{X_i<i\}}$. We claim that it suffices to show that $n^{-1}\sum_{i=1}^n Y_i \to \mu$ a.s., where $\mu := \mathbb{E}(X_1)$.

To see why this suffices, notice that by Exercise 5.7.8 and the fact that the X_i's are identically distributed,

$$\begin{aligned}\sum_{i=1}^{\infty}\mathbb{P}(X_i \neq Y_i) &\leq \sum_{i=1}^{\infty}\mathbb{P}(X_i \geq i)\\ &= \sum_{i=1}^{\infty}\mathbb{P}(X_1 \geq i) \leq \mathbb{E}(X_1) < \infty.\end{aligned}$$

Therefore by the first Borel–Cantelli lemma, $\mathbb{P}(X_i \neq Y_i \text{ i.o.}) = 0$. Therefore, it suffices to prove that $n^{-1}\sum_{i=1}^n Y_i \to \mu$ a.s.

Next, note that $\mathbb{E}(Y_i) - \mu = \mathbb{E}(X_i; X_i \geq i) \to 0$ as $i \to \infty$ by the dominated convergence theorem. Therefore $n^{-1}\sum_{i=1}^n \mathbb{E}(Y_i) \to \mu$ as $n \to \infty$. Thus, it suffices to prove that $Z_n \to 0$ a.s., where

$$Z_n := \frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}(Y_i)).$$

Take any $\alpha > 1$. Let $k_n := [\alpha^n]$, where $[x]$ denotes the integer part of a real number x. The penultimate step in Etemadi's proof is to show that for any choice of $\alpha > 1$, $Z_{k_n} \to 0$ a.s.

To show this, take any $\epsilon > 0$. Recall that the X_i's are pairwise independent. Therefore so are the Y_i's and hence $\mathrm{Cov}(Y_i, Y_j) = 0$ for any $i \neq j$. Thus for any n, by Theorem 7.3.1,

$$\mathbb{P}(|Z_{k_n}| > \epsilon) \leq \frac{1}{\epsilon^2 k_n^2}\sum_{i=1}^{k_n}\mathrm{Var}(Y_i).$$

Therefore

$$\sum_{n=1}^{\infty} \mathbb{P}(|Z_{k_n}| > \epsilon) \le \sum_{n=1}^{\infty} \frac{1}{\epsilon^2 k_n^2} \sum_{i=1}^{k_n} \mathrm{Var}(Y_i)$$
$$= \frac{1}{\epsilon^2} \sum_{i=1}^{\infty} \mathrm{Var}(Y_i) \sum_{n\,:\,k_n \ge i} \frac{1}{k_n^2}$$

It is easy to see that there is some $\beta > 1$, depending only on α, such that $k_{n+1}/k_n \ge \beta$ for all n large enough. Therefore for large enough n,

$$\sum_{n\,:\,k_n \ge i} \frac{1}{k_n^2} \le \frac{1}{i^2} \sum_{n=0}^{\infty} \beta^{-n} \le \frac{C}{i^2},$$

where C depends only on α. Increasing C if necessary, the inequality can be made to hold for all n. Therefore by the monotone convergence theorem,

$$\sum_{n=1}^{\infty} \mathbb{P}(|Z_{k_n}| > \epsilon) \le \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathrm{Var}(Y_i)}{i^2} \le \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(Y_i^2)}{i^2}$$
$$= \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(X_i^2; X_i < i)}{i^2} = \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(X_1^2; X_1 < i)}{i^2}$$
$$\le \frac{C}{\epsilon^2} \mathbb{E}\biggl(X_1^2 \sum_{i > X_1} \frac{1}{i^2} \biggr) \le \frac{C'}{\epsilon^2} \mathbb{E}(X_1),$$

where C' is some other constant depending only on α. By the first Borel–Cantelli lemma, this shows that $\mathbb{P}(|Z_{k_n}| > \epsilon \text{ i.o.}) = 0$. Since this holds for any $\epsilon > 0$, it follows that $Z_{k_n} \to 0$ a.s. as $n \to \infty$.

The final step of the proof is to deduce that $Z_n \to 0$ a.s. For each n, let $T_n := \sum_{i=1}^n Y_i$. Take any m. If $k_n < m \le k_{n+1}$, then

$$\frac{k_n}{k_{n+1}} \frac{T_{k_n}}{k_n} = \frac{T_{k_n}}{k_{n+1}} \le \frac{T_m}{m} \le \frac{T_{k_{n+1}}}{k_n} = \frac{T_{k_{n+1}}}{k_{n+1}} \frac{k_{n+1}}{k_n}.$$

Let $m \to \infty$, so that k_n also tends to infinity. Since $k_{n+1}/k_n \to \alpha$ and $T_{k_n}/k_n \to \mu$ a.s., the above inequalities imply that

$$\frac{\mu}{\alpha} \le \liminf_{m\to\infty} \frac{T_m}{m} \le \limsup_{m\to\infty} \frac{T_m}{m} \le \alpha\mu \quad \text{a.s.}$$

Since $\alpha > 1$ is arbitrary, this completes the proof. □

Exercise 7.4.2. Using Exercise 6.3.2, show that if $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables such that $\mathbb{E}|X_1| = \infty$, then

$$\mathbb{P}\biggl(\frac{1}{n} \sum_{i=1}^n X_i \text{ has a finite limit as } n \to \infty \biggr) = 0.$$

Moreover, if the X_i's are nonnegative, show that $n^{-1} \sum_{i=1}^n X_i \to \infty$ a.s.

Although the strong law of large numbers is formulated for i.i.d. random variables, it can sometimes be applied even when the random variables are not independent. An useful case is the case of stationary m-dependent sequences.

DEFINITION 7.4.3. A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ is called stationary if for any n and m, the random vector $(X_1, \ldots, X_n)$ has the same law as $(X_{m+1}, \ldots, X_{m+n})$.

DEFINITION 7.4.4. A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ is called m-dependent if for any n, the collections $\{X_i\}_{i=1}^{n}$ and $\{X_i\}_{i=n+m+1}^{\infty}$ are independent.

Let $\{Y_i\}_{i=1}^{\infty}$ be an i.i.d. sequence. An example of a sequence which is 1-dependent and stationary is $\{Y_iY_{i+1}\}_{i=1}^{\infty}$. More generally, an example of an m-dependent stationary sequence is a sequence like $\{X_i\}_{i=1}^{\infty}$ where $X_i = f(Y_i, \ldots, Y_{i+m})$ and $f : \mathbb{R}^{m+1} \to \mathbb{R}$ is a measurable function.

THEOREM 7.4.5. *If $\{X_n\}_{n=1}^{\infty}$ is a stationary m-dependent sequence for some finite m, and $\mathbb{E}|X_1| < \infty$, then $n^{-1}\sum_{i=1}^{n} X_i$ converges a.s. to $\mathbb{E}(X_1)$ as $n \to \infty$.*

PROOF. As in Etemadi's proof of the strong law, we may break up each X_i into its positive and negative parts and prove the result separately for the two, since the positive and negative parts also give stationary m-dependent sequences. Let us therefore assume without loss of generality that the X_i's are nonnegative random variables. For each $k \geq 1$, let

$$Y_k := \frac{1}{m} \sum_{i=m(k-1)+1}^{mk} X_i.$$

By stationarity and m-dependence, it is easy to see (using Exercise 6.1.5, for instance), that $Y_1, Y_3, Y_5, \ldots$ is a sequence of i.i.d. random variables, and $Y_2, Y_4, Y_6, \ldots$ is another sequence of i.i.d. random variables. Moreover $\mathbb{E}(Y_1) = \mathbb{E}(X_1)$. Therefore by the strong law of large numbers, the averages

$$A_n := \frac{1}{n}\sum_{i=1}^{n} Y_{2i-1}, \quad B_n := \frac{1}{n}\sum_{i=1}^{n} Y_{2i}$$

both converge a.s. to $\mathbb{E}(X_1)$. Therefore the average

$$C_n := \frac{A_n + B_n}{2} = \frac{1}{2n}\sum_{i=1}^{2n} Y_i$$

also converges a.s. to $\mathbb{E}(X_1)$. But note that

$$C_n = \frac{1}{2nm}\sum_{i=1}^{2nm} X_i.$$

Now take any $n \geq 2m$ and let k be an integer such that $2mk \leq n < 2m(k+1)$. Since the X_i's are nonnegative,

$$C_k \leq \frac{1}{2mk} \sum_{i=1}^{n} X_i \leq \frac{2m(k+1)}{2mk} C_{k+1}.$$

Since C_k and C_{k+1} both converge a.s. to $\mathbb{E}(X_1)$ as $k \to \infty$, and $2mk/n \to 1$ as $n \to \infty$, this completes the proof. □

Sometimes strong laws of large numbers can be proved using only moment bounds and the first Borel–Cantelli lemma. The following exercises give such examples.

Exercise 7.4.6 (SLLN under bounded fourth moment). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables with mean zero and uniformly bounded fourth moment. Prove that $n^{-1} \sum_{i=1}^{n} X_i \to 0$ a.s. (Hint: Use a fourth moment version of Chebychev's inequality.)

Exercise 7.4.7 (Random matrices). Let $\{X_{ij}\}_{1 \leq i \leq j < \infty}$ be a collection of i.i.d. random variables with mean zero and all moments finite. Let $X_{ji} := X_{ij}$ if $j > i$. Let W_n be the $n \times n$ symmetric random matrix whose (i, j)th entry is $n^{-1/2} X_{ij}$. A matrix like W_n is called a Wigner matrix. Let $\lambda_{n,1} \geq \cdots \geq \lambda_{n,n}$ be the eigenvalues of W_n, repeated by multiplicities. For any integer $k \geq 1$, show that

$$\frac{1}{n} \sum_{i=1}^{n} \lambda_{n,i}^k - \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \lambda_{n,i}^k \right) \to 0 \quad \text{a.s. as } n \to \infty.$$

(Hint: Express the sum as the trace of a power of W_n, and use Theorem 7.3.1. The tail bound is strong enough to prove almost sure convergence.)

Exercise 7.4.8. If all the random graphs in Exercise 7.3.5 are defined on the same probability space, show that the convergence is almost sure.

7.5. Tightness and Helly's selection theorem

Starting in this section, we will gradually move towards the proof of the central limit theorem, which is one of the most important basic results in probability theory. For this, we will first have to develop our understanding of weak convergence to a more sophisticated level. A concept that is closely related to convergence in distribution is the notion of tightness, which we study in this section.

Definition 7.5.1. A sequence of random variables $\{X_n\}_{n \geq 1}$ is called a tight family if for any ϵ, there exists some K such that $\sup_n \mathbb{P}(|X_n| \geq K) \leq \epsilon$.

Exercise 7.5.2. If $X_n \to X$ in distribution, show that $\{X_n\}_{n \geq 1}$ is a tight family.

Exercise 7.5.3. If $\{X_n\}_{n \geq 1}$ is a tight family and $\{c_n\}_{n \geq 1}$ is a sequence of constants tending to 0, show that $c_n X_n \to 0$ in probability.

A partial converse of Exercise 7.5.2 is the following theorem.

THEOREM 7.5.4 (Helly's selection theorem). *If $\{X_n\}_{n\geq 1}$ is a tight family, then there is a subsequence $\{X_{n_k}\}_{k\geq 1}$ that converges in distribution.*

PROOF. Let F_n be the c.d.f. of X_n. By the standard diagonal argument, there is subsequence $\{n_k\}_{k\geq 1}$ of positive integers such that

$$F_*(q) := \lim_{k\to\infty} F_{n_k}(q)$$

exists for every rational number q. For each $x \in \mathbb{R}$, define

$$F(x) := \inf_{q\in\mathbb{Q},\, q>x} F_*(q).$$

Then F_* and F are non-decreasing functions. From tightness, it is easy to argue that $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$. Now, for any x, there is a sequence of rationals $q_1 > q_2 > \cdots$ decreasing to x, such that $F_*(q_n) \to F(x)$. Then $F(q_{n+1}) \leq F_*(q_n)$ for each n, and hence

$$F(x) = \lim_{n\to\infty} F_*(q_n) \geq \lim_{n\to\infty} F(q_{n+1}),$$

which proves that F is right-continuous. Thus, by Proposition 5.2.2, F is a cumulative distribution function.

We claim that F_{n_k} converges weakly to F. To show this, let x be a continuity point of F. Take any rational number $q > x$. Then $F_{n_k}(x) \leq F_{n_k}(q)$ for all k. Thus,

$$\limsup_{k\to\infty} F_{n_k}(x) \leq \lim_{k\to\infty} F_{n_k}(q) = F_*(q).$$

Since this holds for all rational $q > x$,

$$\limsup_{k\to\infty} F_{n_k}(x) \leq F(x).$$

Next, take any $y < x$, and take any rational number $q \in (y, x)$. Then

$$\liminf_{k\to\infty} F_{n_k}(x) \geq \lim_{k\to\infty} F_{n_k}(q) = F_*(q).$$

Since this holds for all rational $q \in (y, x)$,

$$\liminf_{k\to\infty} F_{n_k}(x) \geq F(y).$$

Since this holds for all $y < x$ and x is a continuity point of F, this completes the proof. □

7.6. An alternative characterization of weak convergence

The following result gives an important equivalent criterion for convergence in distribution.

PROPOSITION 7.6.1. *A sequence of random variables $\{X_n\}_{n\geq 1}$ converges to a random variable X in distribution if and only if*

$$\lim_{n\to\infty} \mathbb{E}f(X_n) = \mathbb{E}f(X)$$

for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$. In particular, two random variables X and Y have the same law if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$ for all bounded continuous f.

PROOF. First, suppose that $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every bounded continuous function f. Take any continuity point t of F_X. Take any $\epsilon > 0$. Let f be the function that is 1 below t, 0 above $t + \epsilon$, and goes down linearly from 1 to 0 in the interval $[t, t+\epsilon]$. Then f is a bounded continuous function, and so

$$\begin{aligned}\limsup_{n\to\infty} F_{X_n}(t) &\leq \limsup_{n\to\infty} \mathbb{E}f(X_n) \\ &= \mathbb{E}f(X) \leq F_X(t+\epsilon).\end{aligned}$$

Since this is true for all $\epsilon > 0$ and F_X is right-continuous, this gives

$$\limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t).$$

A similar argument shows that for any $\epsilon > 0$,

$$\liminf_{n\to\infty} F_{X_n}(t) \geq F_X(t-\epsilon).$$

Since t is a continuity point of F_X, this proves that $\liminf F_{X_n}(t) \geq F_X(t)$. Thus, $X_n \to X$ in distribution.

Conversely, suppose that $X_n \to X$ in distribution. Let f be a bounded continuous function. Take any $\epsilon > 0$. By Exercise 7.5.2 there exists K such that $\mathbb{P}(|X_n| \geq K) \leq \epsilon$ for all n. Choose K so large that we also have $\mathbb{P}(|X| \geq K) \leq \epsilon$. Let M be a number such that $|f(x)| \leq M$ for all x.

Since f is uniformly continuous in $[-K, K]$, there is some $\delta > 0$ such that $|f(x)-f(y)| \leq \epsilon$ whenever $|x-y| \leq \delta$ and $x, y \in [-K, K]$. By Exercise 5.2.3, we may assume that $-K$ and K are continuity points of F_X, and we can pick out a set of points $x_1 \leq x_2 \leq \cdots \leq x_m \in [-K, K]$ such that each x_i is a continuity point of F_X, $x_1 = -K$, $x_m = K$, and $x_{i+1} - x_i \leq \delta$ for each i. Now note that

$$\begin{aligned}\mathbb{E}f(X_n) = \mathbb{E}(f(X_n); X_n > K) + \mathbb{E}(f(X_n); X_n \leq -K) \\ + \sum_{i=1}^{m-1} \mathbb{E}(f(X_n); x_i < X_n \leq x_{i+1}),\end{aligned}$$

which implies that

$$\begin{aligned}&\left|\mathbb{E}f(X_n) - \sum_{i=1}^{m-1} f(x_i)\mathbb{P}(x_i < X_n \le x_{i+1})\right| \\ &\le M\epsilon + \sum_{i=1}^{m-1} \mathbb{E}(|f(X_n) - f(x_i)|; x_i < X_n \le x_{i+1}) \\ &\le M\epsilon + \epsilon \sum_{i=1}^{m-1} \mathbb{P}(x_i < X_n \le x_{i+1}) \le (M+1)\epsilon.\end{aligned}$$

A similar argument gives

$$\left|\mathbb{E}f(X) - \sum_{i=1}^{m-1} f(x_i)\mathbb{P}(x_i < X \le x_{i+1})\right| \le (M+1)\epsilon.$$

Since $x_1, \ldots, x_m$ are continuity points of F_X and $X_n \to X$ in distribution,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(x_i < X_n \le x_{i+1}) &= \lim_{n\to\infty} (F_{X_n}(x_{i+1}) - F_{X_n}(x_i)) \\ &= F_X(x_{i+1}) - F_X(x_i) = \mathbb{P}(x_i < X \le x_{i+1})\end{aligned}$$

for each i. Combining the above observations, we get

$$\limsup_{n\to\infty} |\mathbb{E}f(X_n) - \mathbb{E}f(X)| \le 2(M+1)\epsilon.$$

Since ϵ was arbitrary, this completes the proof. □

An immediate consequence of Proposition 7.6.1 is the following result.

PROPOSITION 7.6.2. *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables converging in distribution to a random variable X, then for any continuous $f : \mathbb{R} \to \mathbb{R}$, $f(X_n) \xrightarrow{d} f(X)$.*

PROOF. Take any bounded continuous $g : \mathbb{R} \to \mathbb{R}$. Then $g \circ f$ is also a bounded continuous function. Therefore $\mathbb{E}(g \circ f(X_n)) \to \mathbb{E}(g \circ f(X))$, which shows that $f(X_n) \to f(X)$ in distribution. □

7.7. Inversion formulas

We know how to calculate the characteristic function of a probability law on the real line. The following inversion formula allows to go backward, and calculate expectations of bounded continuous functions using the characteristic function.

THEOREM 7.7.1. *Let X be a random variable with characteristic function ϕ. For each $\theta > 0$, define*

$$f_\theta(x) := \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx - \theta t^2} \phi(t) dt.$$

Then for any bounded continuous $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(X)) = \lim_{\theta \to 0} \int_{-\infty}^{\infty} g(x) f_\theta(x) dx.$$

PROOF. Let μ be the law of X, so that

$$\phi(t) = \int_{-\infty}^{\infty} e^{\mathrm{i}ty} d\mu(y).$$

Since $|\phi(t)| \leq 1$ for all t, we may apply Fubini's theorem to conclude that

$$f_\theta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\mathrm{i}(y-x)t - \theta t^2} dt d\mu(y).$$

But by Proposition 5.11.1,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{\mathrm{i}(y-x)t - \theta t^2} dt &= \sqrt{\frac{\pi}{\theta}} \int_{-\infty}^{\infty} e^{\mathrm{i}(2\theta)^{-1/2}(y-x)s} \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds \\ &= \sqrt{\frac{\pi}{\theta}} e^{-(y-x)^2/4\theta}. \end{aligned}$$

Therefore

$$f_\theta(x) = \int_{-\infty}^{\infty} \frac{e^{-(y-x)^2/4\theta}}{\sqrt{4\pi\theta}} d\mu(y).$$

But by Proposition 6.7.1, the above formula shows that $f_\theta(x)$ is the p.d.f. of $X + Z_\theta$, where $Z_\theta \sim N(0, 2\theta)$. Thus, we get

$$\int_{-\infty}^{\infty} g(x) f_\theta(x) dx = \mathbb{E}(g(X + Z_\theta)).$$

Since $\mathrm{Var}(Z_\theta) = 2\theta$ (Exercise 5.8.1), it follows by Chebychev's inequality that $Z_\theta \to 0$ in probability as $\theta \to 0$. Since g is a bounded continuous function, the proof can now be completed using Slutsky's theorem and Proposition 7.6.1. □

An immediate corollary of the above theorem is the following important fact.

COROLLARY 7.7.2. *Two random variables have the same law if and only if they have the same characteristic function.*

PROOF. If two random variables have the same law, then they obviously have the same characteristic function. Conversely, suppose that X and Y have the same characteristic function. Then by Theorem 7.7.1, $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$ for every bounded continuous g. Therefore by Proposition 7.6.1, X and Y have the same law. □

Another important corollary of Theorem 7.7.1 is the following simplified inversion formula.

COROLLARY 7.7.3. *Let X be a random variable with characteristic function ϕ. Suppose that*

$$\int_{-\infty}^{\infty} |\phi(t)|dt < \infty.$$

Then X has a probability density function f, given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} \phi(t) dt.$$

PROOF. Recall from the proof of Theorem 7.7.1 that f_θ is the p.d.f. of $X + Z_\theta$, where $Z_\theta \sim N(0, 2\theta)$. If ϕ is integrable, then it is easy to see by the dominated convergence theorem that for every x,

$$f(x) = \lim_{\theta \to 0} f_\theta(x).$$

Moreover, the integrability of ϕ also shows that for any θ and x,

$$|f_\theta(x)| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\phi(t)|dt < \infty.$$

Therefore by the dominated convergence theorem, for any $-\infty < a \leq b < \infty$,

$$\int_a^b f(x)dx = \lim_{\theta \to 0} \int_a^b f_\theta(x)dx.$$

Therefore if a and b are continuity points of the c.d.f. of X, then Slutsky's theorem implies that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

By Proposition 5.4.1, this completes the proof. □

For integer-valued random variables, a different inversion formula is often useful.

THEOREM 7.7.4. *Let X be an integer-valued random variable with characteristic function ϕ. Then for any $x \in \mathbb{Z}$,*

$$\mathbb{P}(X = x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\mathrm{i}tx} \phi(t) dt.$$

PROOF. Let μ be the law of X. Then note that by Fubini's theorem,

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\mathrm{i}tx} \phi(t) dt &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \int_{-\infty}^{\infty} e^{-\mathrm{i}tx} e^{\mathrm{i}ty} d\mu(y) dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} e^{\mathrm{i}t(y-x)} dt d\mu(y) \\
&= \sum_{y \in \mathbb{Z}} \mathbb{P}(X = y) \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\mathrm{i}t(y-x)} dt \right) \\
&= \mathbb{P}(X = x),
\end{aligned}$$

where the last identity holds because $x, y \in \mathbb{Z}$ in the previous step. □

7.8. Lévy's continuity theorem

In this section we will prove Lévy's continuity theorem, which asserts that convergence in distribution is equivalent to pointwise convergence of characteristic functions.

THEOREM 7.8.1 (Lévy's continuity theorem). *A sequence of random variables $\{X_n\}_{n\geq 1}$ converges in distribution to a random variable X if and only if the sequence of characteristic functions $\{\phi_{X_n}\}_{n\geq 1}$ converges to the characteristic function ϕ_X pointwise.*

PROOF. If $X_n \to X$ in distribution, then $\phi_{X_n}(t) \to \phi_X(t)$ for every t by Proposition 7.6.1. Conversely, suppose that $\phi_{X_n}(t) \to \phi_X(t)$ for every t. Take any $\epsilon > 0$. Recall that ϕ_X is a continuous function (Exercise 5.10.3), and $\phi_X(0) = 1$. Therefore we can choose a number a so small that $|\phi_X(s) - 1| \leq \epsilon/2$ whenever $|s| \leq a$. Consequently,

$$\frac{1}{a}\int_{-a}^{a}(1-\phi_X(s))ds \leq \epsilon.$$

Therefore, since $\phi_{X_n} \to \phi_X$ pointwise, the dominated convergence theorem shows that

$$\lim_{n\to\infty}\frac{1}{a}\int_{-a}^{a}(1-\phi_{X_n}(s))ds = \frac{1}{a}\int_{-a}^{a}(1-\phi_X(s))ds \leq \epsilon.$$

Let $t := 2/a$. Then by Proposition 5.10.4 and the above inequality,

$$\limsup_{n\to\infty}\mathbb{P}(|X_n| \geq t) \leq \epsilon.$$

Thus, $\mathbb{P}(|X_n| \geq t) \leq 2\epsilon$ for all large enough n. This allows us to choose K large enough such that $\mathbb{P}(|X_n| \geq K) \leq 2\epsilon$ for all n. Since ϵ is arbitrary, this proves that $\{X_n\}_{n\geq 1}$ is a tight family.

Suppose that X_n does not converge in distribution to X. Then there is a bounded continuous function f such that $\mathbb{E}f(X_n) \not\to \mathbb{E}f(X)$. Passing to a subsequence if necessary, we may assume that there is some $\epsilon > 0$ such that $|\mathbb{E}f(X_n) - \mathbb{E}f(X)| \geq \epsilon$ for all n. By tightness, there exists some subsequence $\{X_{n_k}\}_{k\geq 1}$ that converges in distribution to a limit Y. Then $\mathbb{E}f(X_{n_k}) \to \mathbb{E}f(Y)$ and hence $|\mathbb{E}f(Y) - \mathbb{E}f(X)| \geq \epsilon$. But by the first part of this theorem and the hypothesis that $\phi_{X_n} \to \phi_X$ pointwise, we have that $\phi_Y = \phi_X$ everywhere. Therefore Y and X must have the same law by Lemma 7.7.2, and we get a contradiction by the second assertion of Proposition 7.6.1 applied to this X and Y. □

EXERCISE 7.8.2. If a sequence of characteristic functions $\{\phi_n\}_{n\geq 1}$ converges pointwise to a characteristic function ϕ, prove that the convergence is uniform on any bounded interval.

EXERCISE 7.8.3. If a sequence of characteristic functions $\{\phi_n\}_{n\geq 1}$ converges pointwise to some function ϕ, and ϕ is continuous at zero, prove that ϕ is also a characteristic function. (Hint: Prove tightness and proceed from there.)

7.9. The central limit theorem for i.i.d. sums

Broadly speaking, a central limit theorem is any theorem that states that a sequence of random variables converges weakly to a limit random variable that has a continuous distribution (usually Gaussian). The following central limit theorem suffices for many problems.

THEOREM 7.9.1 (CLT for i.i.d. sums). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Then the random variable*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges weakly to the standard Gaussian distribution as $n \to \infty$.

We need a few simple lemmas to prepare for the proof. The main argument is based on Lévy's continuity theorem.

LEMMA 7.9.2. *For any $x \in \mathbb{R}$ and $k \geq 0$,*

$$\left| e^{\mathrm{i}x} - \sum_{j=0}^{k} \frac{(\mathrm{i}x)^j}{j!} \right| \leq \frac{|x|^{k+1}}{(k+1)!}$$

PROOF. This follows easily from Taylor expansion, noting that all derivatives of the map $x \mapsto e^{\mathrm{i}x}$ are uniformly bounded by 1 in magnitude. □

COROLLARY 7.9.3. *For any $x \in \mathbb{R}$,*

$$\left| e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{x^2}{2} \right| \leq \min\left\{ x^2, \frac{|x|^3}{6} \right\}.$$

PROOF. By Lemma 7.9.2

$$\left| e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{x^2}{2} \right| \leq \frac{|x|^3}{6}.$$

On the other hand, by Lemma 7.9.2 we also have

$$\begin{aligned} \left| e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{x^2}{2} \right| &\leq |e^{\mathrm{i}x} - 1 - \mathrm{i}x| + \frac{x^2}{2} \\ &\leq \frac{x^2}{2} + \frac{x^2}{2} = x^2. \end{aligned}$$

The proof is completed by combining the two bounds. □

LEMMA 7.9.4. *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be complex numbers such that $|a_i| \le 1$ and $|b_i| \le 1$ for each i. Then*

$$\left|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i\right| \le \sum_{i=1}^n |a_i - b_i|.$$

PROOF. Writing the difference of the products as a telescoping sum and applying the triangle inequality, we get

$$\begin{aligned}\left|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i\right| &\le \sum_{i=1}^n \left|a_1 \cdots a_{i-1} b_i \cdots b_n - a_1 \cdots a_i b_{i+1} \cdots b_n\right| \\ &= \sum_{i=1}^n |a_1 \cdots a_{i-1}(b_i - a_i) b_{i+1} \cdots b_n| \\ &\le \sum_{i=1}^n |a_i - b_i|,\end{aligned}$$

where the last inequality holds because $|a_i|$ and $|b_i|$ are bounded by 1 for each i. □

We are now ready to prove Theorem 7.9.1.

PROOF OF THEOREM 7.9.1. Replacing X_i by $(X_i - \mu)/\sigma$, let us assume without loss of generality that $\mu = 0$ and $\sigma = 1$. Let $S_n := n^{-1/2} \sum_{i=1}^n X_i$. Take any $t \in \mathbb{R}$. By Lévy's continuity theorem and the formula for the characteristic function of the standard normal distribution (Proposition 5.11.1), it is sufficient to show that $\phi_{S_n}(t) \to e^{-t^2/2}$ as $n \to \infty$, where ϕ_{S_n} is the characteristic function of S_n. By the i.i.d. nature of the summands,

$$\phi_{S_n}(t) = \prod_{i=1}^n \phi_{X_i}(t/\sqrt{n}) = (\phi_{X_1}(t/\sqrt{n}))^n.$$

Therefore by Lemma 7.9.4, when n is so large that $t^2 \le 2n$,

$$\left|\phi_{S_n}(t) - \left(1 - \frac{t^2}{2n}\right)^n\right| \le n\left|\phi_{X_1}(t/\sqrt{n}) - \left(1 - \frac{t^2}{2n}\right)\right|.$$

Thus, we only need to show that the right side tends to zero as $n \to \infty$. To prove this, note that by Corollary 7.9.3,

$$\begin{aligned} n\left|\phi_{X_1}(t/\sqrt{n}) - \left(1 - \frac{t^2}{2n}\right)\right| &= n\left|\mathbb{E}\left(e^{\mathrm{i}tX_1/\sqrt{n}} - 1 - \frac{\mathrm{i}tX_1}{\sqrt{n}} + \frac{t^2 X_1^2}{2n}\right)\right| \\ &\le \mathbb{E}\min\left\{t^2 X_1^2, \frac{|t|^3 |X_1|^3}{6\sqrt{n}}\right\}.\end{aligned}$$

By the finiteness of $\mathbb{E}(X_1^2)$ and the dominated convergence theorem, the above expectation tends to zero as $n \to \infty$. □

EXERCISE 7.9.5. Give a counterexample to show that the i.i.d. assumption in Theorem 7.9.1 cannot be replaced by the assumption of identically distributed and pairwise independent.

In the following exercises, 'prove a central limit theorem for X_n' means 'prove that

$$\frac{X_n - a_n}{b_n} \xrightarrow{d} N(0,1)$$

as $n \to \infty$, for some appropriate sequences of constants a_n and b_n'.

EXERCISE 7.9.6. Let $X_n \sim Bin(n,p)$. Prove a central limit theorem for X_n. (Hint: Use Exercise 6.7.4 and the CLT for i.i.d. random variables.)

EXERCISE 7.9.7. Let $X_n \sim Gamma(n,\lambda)$. Prove a central limit theorem for X_n. (Hint: Use Exercise 6.7.5 and the CLT for i.i.d. random variables.)

EXERCISE 7.9.8. Suppose that $X_n \sim Gamma(n,\lambda_n)$, where $\{\lambda_n\}_{n=1}^\infty$ is any sequence of positive constants. Prove a CLT for X_n.

Just like the strong law of large numbers, one can prove a central limit theorem for a stationary m-dependent sequence of random variables.

THEOREM 7.9.9 (CLT for stationary m-dependent sequences). *Suppose that $X_1, X_2, \ldots$ is a stationary m-dependent sequence of random variables with mean μ and finite second moment. Let*

$$\sigma^2 := \mathrm{Var}(X_1) + 2\sum_{i=2}^{m+1} \mathrm{Cov}(X_1, X_i).$$

Then the random variable

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges weakly to the standard Gaussian distribution as $n \to \infty$.

PROOF. The proof of this result uses the technique of 'big blocks and little blocks', which is also useful for other things. Without loss of generality, assume that $\mu = 0$. Take any $r \geq m$. Divide up the set of positive integers into 'big blocks' size r, with intermediate 'little blocks' of size m. For example, the first big block is $\{1, \ldots, r\}$, which is followed by the little block $\{r+1, \ldots, r+m\}$. Enumerate the big blocks as $B_1, B_2, \ldots$ and the little blocks as $L_1, L_2, \ldots$. Let

$$Y_j := \sum_{i \in B_j} X_i, \quad Z_j := \sum_{i \in L_j} X_i.$$

Note that by stationarity and m-dependence, $Y_1, Y_2, \ldots$ is a sequence of i.i.d. random variables and $Z_1, Z_2, \ldots$ is also a sequence of i.i.d. random variables.

Let k_n be the largest integer such that $B_{k_n} \subseteq \{1, \dots, n\}$. Let $S_n := \sum_{i=1}^n X_i$, $T_n := \sum_{j=1}^{k_n} Y_j$, and $R_n := S_n - T_n$. Then by the central limit theorem for i.i.d. sums and the fact that $k_n \sim n/(r+m)$ as $n \to \infty$, we have

$$\frac{T_n}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_r^2),$$

where

$$\sigma_r^2 = \frac{\mathrm{Var}(Y_1)}{r+m}.$$

Now, it is not hard to see that if we let

$$R_n' := \sum_{j=1}^{k_n} Z_j,$$

then $\{R_n - R_n'\}_{n \geq 1}$ is a tight family of random variables. Therefore by Exercise 7.5.3, $(R_n - R_n')/\sqrt{n} \to 0$ in probability. But again by the CLT for i.i.d. variables,

$$\frac{R_n'}{\sqrt{n}} \xrightarrow{d} N(0, \tau_r^2),$$

where

$$\tau_r^2 = \frac{\mathrm{Var}(Z_1)}{r+m}.$$

Therefore by Slutsky's theorem, $R_n/\sqrt{n}$ also has the same weak limit.

Now let ϕ_n be the characteristic function of $S_n/\sqrt{n}$ and $\psi_{n,r}$ be the characteristic function of $T_n/\sqrt{n}$. Then for any t,

$$\begin{aligned} |\phi_n(t) - \psi_{n,r}(t)| &= |\mathbb{E}(e^{\mathrm{i}tS_n/\sqrt{n}} - e^{\mathrm{i}tT_n/\sqrt{n}})| \\ &\leq \mathbb{E}|e^{\mathrm{i}tR_n/\sqrt{n}} - 1| \end{aligned}$$

Letting $n \to \infty$ on both sides and using the observations made above, we get

$$\limsup_{n \to \infty} |\phi_n(t) - e^{-t^2\sigma_r^2/2}| \leq \mathbb{E}|e^{\mathrm{i}t\xi_r} - 1|,$$

where $\xi_r \sim N(0, \tau_r^2)$. Now send $r \to \infty$. It is easy to check that $\sigma_r \to \sigma$ and $\tau_r \to 0$, and complete the proof from there. □

Exercise 7.9.10. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables. For each $i \geq 2$, let

$$Y_i := \begin{cases} 1 & \text{if } X_i \geq \max\{X_{i-1}, X_{i+1}\}, \\ 0 & \text{if not.} \end{cases}$$

In other words, Y_i is 1 if and only if the original sequence has a local maximum at i. Prove a central limit theorem $\sum_{i=2}^n Y_i$.

7.10. The Lindeberg–Feller central limit theorem

In some applications, the CLT for i.i.d. sums does not suffice. For sums of independent random variables, the most powerful result available in the literature is the following theorem.

THEOREM 7.10.1 (Lindeberg–Feller CLT). *Let $\{k_n\}_{n\geq 1}$ be a sequence of positive integers increasing to infinity. For each n, let $\{X_{n,i}\}_{1\leq i\leq k_n}$ is a collection of independent random variables. Let $\mu_{n,i} := \mathbb{E}(X_{n,i})$, $\sigma^2_{n,i} := \mathrm{Var}(X_{n,i})$, and*

$$s_n^2 := \sum_{i=1}^{k_n} \sigma^2_{n,i}.$$

Suppose that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{k_n} \mathbb{E}((X_{n,i}-\mu_{n,i})^2; |X_{n,i}-\mu_{n,i}| \geq \epsilon s_n) = 0. \tag{7.10.1}$$

Then the random variable

$$\frac{\sum_{i=1}^{k_n}(X_{n,i}-\mu_{n,i})}{s_n}$$

converges in distribution to the standard Gaussian law as $n\to\infty$.

The condition (7.10.1) is commonly known as Lindeberg's condition. The proof of the Lindeberg–Feller CLT similar to the proof of the CLT for i.i.d. sums, but with a few minor additional technical subtleties.

LEMMA 7.10.2. *For any $x_1,\ldots,x_n \in [0,1]$,*

$$\left|\exp\left(-\sum_{i=1}^n x_i\right) - \prod_{i=1}^n (1-x_i)\right| \leq \frac{1}{2}\sum_{i=1}^n x_i^2.$$

PROOF. By Taylor expansion, we have $|e^{-x}-(1-x)| \leq x^2/2$ for any $x\geq 0$. The proof is now easily completed by Lemma 7.9.4. □

PROOF OF THEOREM 7.10.1. Replacing $X_{n,i}$ by $(X_{n,i}-\mu_{n,i})/s_n$, let us assume without loss of generality that $\mu_{n,i}=0$ and $s_n=1$ for each n and i. Then the condition (7.10.1) becomes

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}(X_{n,i}^2; |X_{n,i}|\geq\epsilon) = 0. \tag{7.10.2}$$

Note that for any $\epsilon>0$ and any n,

$$\begin{aligned}\max_{1\leq i\leq k_n} \sigma^2_{n,i} &= \max_{1\leq i\leq k_n} (\mathbb{E}(X_{n,i}^2; |X_{n,i}|<\epsilon) + \mathbb{E}(X_{n,i}^2; |X_{n,i}|\geq\epsilon)) \\ &\leq \epsilon^2 + \max_{1\leq i\leq k_n} \mathbb{E}(X_{n,i}^2; |X_{n,i}|\geq\epsilon).\end{aligned}$$

Therefore by (7.10.2),

$$\limsup_{n\to\infty} \max_{1\le i\le k_n} \sigma_{n,i}^2 \le \epsilon^2.$$

Since ϵ is arbitrary, this shows that

$$\lim_{n\to\infty} \max_{1\le i\le k_n} \sigma_{n,i}^2 = 0. \tag{7.10.3}$$

Let $S_n := \sum_{i=1}^n X_i$. Then by Exercise 6.7.7,

$$\phi_{S_n}(t) = \prod_{i=1}^{k_n} \phi_{X_{n,i}}(t).$$

By Corollary 7.9.3,

$$\begin{aligned}\left|\phi_{X_{n,i}}(t) - 1 + \frac{t^2\sigma_{n,i}^2}{2}\right| &= \left|\mathbb{E}\left(e^{\mathrm{i}tX_{n,i}} - 1 - \mathrm{i}tX_{n,i} + \frac{t^2X_{n,i}^2}{2}\right)\right| \\ &\le \mathbb{E}\min\left\{t^2X_{n,i}^2, \frac{|t|^3|X_{n,i}|^3}{6}\right\}.\end{aligned}$$

Now fix t. Equation (7.10.3) tells us that $t^2\sigma_{n,i}^2 \le 2$ for each i when n is sufficiently large. Thus by Lemma 7.9.4,

$$\begin{aligned}\left|\phi_{S_n}(t) - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| &\le \sum_{i=1}^{k_n}\left|\phi_{X_i}(t) - 1 + \frac{t^2\sigma_{n,i}^2}{2}\right| \\ &\le \sum_{i=1}^{k_n} \mathbb{E}\min\left\{t^2X_{n,i}^2, \frac{|t|^3|X_{n,i}|^3}{6}\right\}.\end{aligned}$$

Take any $\epsilon > 0$. Then

$$\begin{aligned}&\mathbb{E}\min\left\{t^2X_{n,i}^2, \frac{|t|^3|X_{n,i}|^3}{6}\right\} \\ &\le \mathbb{E}(t^2X_{n,i}^2; |X_{n,i}| \ge \epsilon) + \frac{|t|^3}{6}\mathbb{E}(|X_{n,i}|^3; |X_{n,i}| < \epsilon) \\ &\le t^2\mathbb{E}(X_{n,i}^2; |X_{n,i}| \ge \epsilon) + \frac{|t|^3\epsilon}{6}\mathbb{E}(X_{n,i}^2).\end{aligned}$$

Therefore for any fixed t and sufficiently large n,

$$\begin{aligned}\left|\phi_{S_n}(t) - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| &\le t^2\sum_{i=1}^{k_n}\mathbb{E}(X_{n,i}^2; |X_{n,i}| \ge \epsilon) + \frac{|t|^3\epsilon}{6}\sum_{i=1}^{k_n}\sigma_{n,i}^2 \\ &= t^2\sum_{i=1}^{k_n}\mathbb{E}(X_{n,i}^2; |X_{n,i}| \ge \epsilon) + \frac{|t|^3\epsilon}{6}.\end{aligned}$$

Therefore by (7.10.2),

$$\limsup_{n\to\infty}\left|\phi_{S_n}(t) - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| \le \frac{|t|^3\epsilon}{6}.$$

Since this holds for any $\epsilon > 0$, the lim sup on the left must be equal to zero. But by Corollary 7.10.2 and equation (7.10.3),

$$\begin{aligned}\limsup_{n\to\infty}\left|e^{-t^2/2} - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| &\le \limsup_{n\to\infty}\frac{1}{2}\sum_{i=1}^{k_n} t^4\sigma_{n,i}^4 \\ &\le \limsup_{n\to\infty}\frac{t^4\max_{1\le i\le k_n}\sigma_{n,i}^2}{2}\sum_{i=1}^{k_n}\sigma_{n,i}^2 \\ &= \limsup_{n\to\infty}\frac{t^4\max_{1\le i\le k_n}\sigma_{n,i}^2}{2} = 0.\end{aligned}$$

Thus, $\phi_{S_n}(t) \to e^{-t^2/2}$ as $n \to \infty$. By Lévy's continuity theorem and Proposition 5.11.1, this proves that S_n converges weakly to the standard Gaussian distribution. □

A corollary of the Lindeberg–Feller CLT that is useful for sums of independent but not identically distributed random variables is the Lyapunov CLT.

THEOREM 7.10.3 (Lyapunov CLT). *Let $\{X_n\}_{n=1}^\infty$ be a sequence of independent random variables. Let $\mu_i := \mathbb{E}(X_i)$, $\sigma_i^2 := \mathrm{Var}(X_i)$, and $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for some $\delta > 0$,*

$$\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^n \mathbb{E}|X_i - \mu_i|^{2+\delta} = 0, \tag{7.10.4}$$

then the random variable

$$\frac{\sum_{i=1}^n (X_i - \mu_i)}{s_n}$$

converges weakly to the standard Gaussian distribution as $n \to \infty$.

PROOF. To put this in the framework of the Lindeberg–Feller CLT, let $k_n = n$ and $X_{n,i} = X_i$. Then for any $\epsilon > 0$ and any n,

$$\begin{aligned}\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}((X_i - \mu_i)^2; |X_i - \mu_i| \ge \epsilon s_n) &\le \frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\left(\frac{|X_i - \mu_i|^{2+\delta}}{(\epsilon s_n)^\delta}\right) \\ &= \frac{1}{\epsilon^\delta s_n^{2+\delta}}\sum_{i=1}^n \mathbb{E}|X_i - \mu_i|^{2+\delta}.\end{aligned}$$

The Lyapunov condition (7.10.4) implies that this upper bound tends to zero as $n \to \infty$, which completes the proof. □

EXERCISE 7.10.4. Suppose that $X_n \sim Bin(n, p_n)$, where $\{p_n\}_{n=1}^\infty$ is a sequence of constants such that $np_n \to \infty$. Prove a CLT for X_n.

EXERCISE 7.10.5. Let $X_1, X_2, \ldots$ be a sequence of uniformly bounded independent random variables, and let $S_n = \sum_{i=1}^n X_i$. If $\mathrm{Var}(S_n) \to \infty$, show that S_n satisfies a central limit theorem.

CHAPTER 8

Weak convergence on Polish spaces

In this chapter we will develop the framework of weak convergence on complete separable metric spaces, also called Polish spaces. The most important examples are finite dimensional Euclidean spaces and spaces of continuous functions.

8.1. Definition

Let S be a Polish space with metric ρ. The notions of almost sure convergence, convergence in probability and L^p convergence remain the same, with $|X_n - X|$ replaced by $\rho(X_n, X)$. Convergence in distribution is a bit more complicated, since cumulative distribution functions do not make sense in a Polish space. It turns out that the right way to define convergence in distribution on Polish spaces is to generalize the equivalent criterion given in Proposition 7.6.1.

DEFINITION 8.1.1. Let (S, ρ) be a Polish space. A sequence X_n of S-valued random variables is said to converge weakly to an S-valued random variable X if for any bounded continuous function $f : S \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}f(X_n) \to \mathbb{E}f(X).$$

Alternatively, a sequence of probability measure μ_n on S is said to converge weakly to a probability measure μ on S if for every bounded continuous function $f : S \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f d\mu_n = \int_S f d\mu.$$

Just as on the real line, the law of an S-valued random variable X is the probability measure μ_X on S defined as

$$\mu_X(A) := \mathbb{P}(X \in A).$$

Of course, here the σ-algebra on S is the Borel σ-algebra generated by its topology. By the following exercise, it follows that a sequence of random variables on S converge weakly if and only if their laws converge weakly.

EXERCISE 8.1.2. Prove that the assertion of Exercise 5.7.1 holds on Polish spaces.

For any n, the Euclidean space $\mathbb{R}^n$ with the usual Euclidean metric is an example of a Polish space. The following exercises give some other examples of Polish spaces.

EXERCISE 8.1.3. Let $C[0,1]$ be the space of all continuous functions from $[0,1]$ into $\mathbb{R}$, with the metric

$$\rho(f,g) := \sup_{x\in[0,1]} |f(x) - g(x)|.$$

Prove that this is a Polish space. (Often the distance $\rho(f,g)$ is denoted by $\|f-g\|_\infty$ or $\|f-g\|_{[0,1]}$ or $\|f-g\|_{\sup}$, and is called the sup-metric.)

EXERCISE 8.1.4. Let $C[0,\infty)$ be the space of all continuous functions from $[0,\infty)$ into $\mathbb{R}$, with the metric

$$\rho(f,g) := \sum_{j=0}^{\infty} 2^{-j} \frac{\|f-g\|_{[j,j+1]}}{1+\|f-g\|_{[j,j+1]}},$$

where

$$\|f-g\|_{[j,j+1]} := \sup_{x\in[j,j+1]} |f(x) - g(x)|.$$

Prove that this is a Polish space. Moreover, show that $f_n \to f$ on this space if and only if $f_n \to f$ uniformly on compact sets.

EXERCISE 8.1.5. Generalize the above exercise to $\mathbb{R}^n$-valued continuous functions on $[0,\infty)$.

EXERCISE 8.1.6. If X is a $C[0,1]$-valued random variable, show that for any $t \in [0,1]$, $X(t)$ is a real-valued random variable.

EXERCISE 8.1.7. If X is a $C[0,1]$-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, show that the map $(\omega, t) \mapsto X(\omega)(t)$ from $\Omega \times [0,1]$ into $\mathbb{R}$ is measurable with respect to the product σ-algebra. (Hint: Approximate X by a sequence of piecewise linear random functions, and use the previous exercise.)

EXERCISE 8.1.8. If X is a $C[0,1]$-valued random variable, show that for any bounded measurable $f : \mathbb{R} \to \mathbb{R}$, $\int_0^1 f(X(t))dt$ is a real-valued random variable. Also, show that the map $t \mapsto \mathbb{E}f(X(t))$ is measurable. (Hint: Use Exercise 8.1.7 and the measurability assertion from Fubini's theorem.)

EXERCISE 8.1.9. If X is a $C[0,1]$-valued random variable, show that $\max_{0\le t\le 1} X(t)$ is a real-valued random variable.

8.2. The portmanteau lemma

The following result gives an important set of equivalent criteria for weak convergence on Polish spaces. Recall that if (S,ρ) is a metric space, a function $f : S \to \mathbb{R}$ is called Lipschitz continuous if there is some $L < \infty$ such that $|f(x) - f(y)| \le L\rho(x,y)$ for all $x, y \in S$.

PROPOSITION 8.2.1 (Portmanteau lemma). *Let (S,ρ) be a Polish space and let $\{\mu_n\}_{n=1}^\infty$ be a sequence of probability measures on S. The following are equivalent:*

(a) $\mu_n \to \mu$ *weakly.*
(b) $\int f d\mu_n \to \int f d\mu$ *for every bounded and uniformly continuous* f*.*
(c) $\int f d\mu_n \to \int f d\mu$ *for every bounded and Lipschitz continuous* f*.*
(d) *For every closed set* $F \subseteq S$*,*

$$\limsup_{n\to\infty} \mu_n(F) \leq \mu(F).$$

(e) *For every open set* $V \subseteq S$*,*

$$\liminf_{n\to\infty} \mu_n(V) \geq \mu(V).$$

(f) *For every Borel set* $A \subseteq S$ *such that* $\mu(\partial A) = 0$ *(where* ∂A *denotes the topological boundary of* A*),*

$$\lim_{n\to\infty} \mu_n(A) = \mu(A).$$

(g) $\int f d\mu_n \to \int f d\mu$ *for every bounded measurable* $f : S \to \mathbb{R}$ *that is continuous a.e. with respect to the measure* μ*.*

PROOF. It is clear that (a) $\implies$ (b) $\implies$ (c). Suppose that (c) holds. Take any closed set $F \subseteq S$. Define the function

$$f(x) = \rho(x, F) := \inf_{y\in F} \rho(x, y). \tag{8.2.1}$$

Note that for any $x, x' \in S$ and $y \in F$, the triangle inequality gives $\rho(x, y) \leq \rho(x, x') + \rho(x', y)$. Taking infimum over y gives $f(x) \leq \rho(x, x') + f(x')$. Similarly, $f(x') \leq \rho(x, x') + f(x)$. Thus,

$$|f(x) - f(x')| \leq \rho(x, x'). \tag{8.2.2}$$

In particular, f is a Lipschitz continuous function. Since F is closed, it is easy to see that $f(x) = 0$ if and only if $x \in F$. Thus, if we define $g_k(x) := (1 - kf(x))_+$, then g_k is Lipschitz continuous, takes values in $[0, 1]$, $1_F \leq g_k$ everywhere, and $g_k \to 1_F$ pointwise as $k \to \infty$. Therefore by (c),

$$\limsup_{n\to\infty} \mu_n(F) \leq \limsup_{n\to\infty} \int g_k d\mu_n = \int g_k d\mu$$

for any k, and hence by the dominated convergence theorem,

$$\limsup_{n\to\infty} \mu_n(F) \leq \lim_{k\to\infty} \int g_k d\mu = \mu(F),$$

which proves (d). The implication (d) $\implies$ (e) follows simply by recalling that open sets are complements of closed sets. Suppose that (d) and (e) both hold. Take any A such that $\mu(\partial A) = 0$. Let F be the closure of A and V be the interior of A. Then F is closed, V is open, $V \subseteq A \subseteq F$, and

$$\mu(F) - \mu(V) = \mu(F \setminus V) = \mu(\partial A) = 0.$$

Consequently, $\mu(A) = \mu(V) = \mu(F)$. Therefore by (d) and (e),

$$\begin{aligned}\mu(A) = \mu(V) &\le \liminf_{n\to\infty} \mu_n(V) \le \liminf_{n\to\infty} \mu_n(A) \\ &\le \limsup_{n\to\infty} \mu_n(A) \le \limsup_{n\to\infty} \mu_n(F) \le \mu(F) = \mu(A),\end{aligned}$$

which proves (f). Next, suppose that (f) holds. Take any f as in (g), and let D_f be the set of discontinuity points of f. Since f is bounded, we may apply a linear transformation and assume without loss of generality that f takes values in $[0,1]$. For $t \in [0,1]$, let $A_t := \{x : f(x) \ge t\}$. Then by Fubini's theorem, for any probability measure ν on S,

$$\begin{aligned}\int_S f(x)d\nu(x) &= \int_S \int_0^1 1_{\{f(x)\ge t\}} dt d\nu(x) \\ &= \int_0^1 \int_S 1_{\{f(x)\ge t\}} d\nu(x) dt = \int_0^1 \nu(A_t) dt.\end{aligned}$$

Now take any t and $x \in \partial A_t$. Then there is a sequence $y_n \to x$ such that $y_n \notin A_t$ for each n, and there is a sequence $z_n \to x$ such that $z_n \in A_t$ for each n. Thus if $x \notin D_f$, then $f(x) = t$. Consequently, $\partial A_t \subseteq D_f \cup \{x : f(x) = t\}$. Since $\mu(D_f) = 0$, this shows that $\mu(A_t) > 0$ if and only if $\mu(\{x : f(x) = t\}) > 0$. But $\mu(\{x : f(x) = t\})$ can be strictly positive for only countably many t. Thus, $\mu(\partial A_t) = 0$ for all but countably many t. Therefore by (f) and the a.e. version of the dominated convergence theorem (Exercise 2.6.5),

$$\lim_{n\to\infty} \int f d\mu_n = \lim_{n\to\infty} \int_0^1 \mu_n(A_t) dt = \int_0^1 \mu(A_t) dt = \int f d\mu,$$

proving (g). Finally, if (g) holds then (a) is obvious. □

An important corollary of the portmanteau lemma is the following.

COROLLARY 8.2.2. *Let S be a Polish space. If μ and ν are two probability measures on S such that $\int f d\mu = \int f d\nu$ for every bounded continuous $f : S \to \mathbb{R}$, then $\mu = \nu$.*

PROOF. By the given condition, μ converges weakly to ν and ν converges weakly to μ. Therefore by the portmanteau lemma, $\mu(F) \le \nu(F)$ and $\nu(F) \le \mu(F)$ for every closed set F. Thus, by Theorem 1.3.6, $\mu = \nu$. □

EXERCISE 8.2.3. Let (S, ρ) be a Polish space, and let $\{X_n\}_{n=1}^\infty$ be a sequence of S-valued random variables converging in law to a random variable X. Show that for any continuous $f : S \to \mathbb{R}$, $f(X_n) \xrightarrow{d} f(X)$.

EXERCISE 8.2.4. Let $\{X_n\}_{n=1}^\infty$ be a sequence of $C[0,1]$-valued random variables converging weakly to some X. Then prove that $\max_{0\le t\le 1} X_n(t) \xrightarrow{d} \max_{0\le t\le 1} X(t)$.

Exercise 8.2.6 given below requires an application of criterion (g) from the portmanteau lemma. Although criterion (g) implicitly assumes that the

set of discontinuity points is a measurable set of measure zero, it is often not trivial to prove measurability of the set of discontinuity points. The following exercise is helpful.

EXERCISE 8.2.5. Let (S, ρ) be a metric space, and let $f : S \to \mathbb{R}$ be any function. Prove that the set of continuity points of f is measurable. (Hint: For any open set U, define $d(U) := \sup_{x \in U} f(x) - \inf_{x \in U} f(x)$. Let V_ϵ be the union of all open U with $d(U) < \epsilon$. Show that the set of continuity points of f is exactly $\cap_{n \geq 1} V_{1/n}$.)

EXERCISE 8.2.6. In the setting of the previous exercise, suppose further that $\mathbb{P}(X(t) = 0) = 0$ for a.e. $t \in [0, 1]$. Then show that

$$\int_0^1 1_{\{X_n(t) \geq 0\}} dt \xrightarrow{d} \int_0^1 1_{\{X(t) \geq 0\}} dt.$$

(Hint: Use criterion (g) from the portmanteau lemma.)

EXERCISE 8.2.7. In the previous exercise, give a counterexample to show that the conclusion may not be valid without the condition that $\mathbb{P}(X(t) = 0) = 0$ for a.e. $t \in [0, 1]$.

8.3. Tightness and Prokhorov's theorem

The appropriate generalization of the notion of tightness to probability measures on Polish spaces is the following.

DEFINITION 8.3.1. Let (S, ρ) be a Polish space and let $\{\mu_n\}_{n \geq 1}$ be a sequence of probability measures on S. The sequence is called tight if for any ϵ, there is a compact set $K \subseteq S$ such that $\mu_n(K) \geq 1 - \epsilon$ for all n.

The following result was almost trivial on the real line, but requires effort to prove on Polish spaces.

THEOREM 8.3.2. *If a sequence of probability measures on a Polish space is weakly convergent, then it is tight.*

PROOF. Let (S, ρ) be a Polish space and let $\{\mu_n\}_{n=1}^\infty$ be a sequence of probability measures on S that converge weakly to some μ. First, we claim that if $\{V_i\}_{i=1}^\infty$ is any increasing sequence of open sets whose union is S, then for each $\epsilon > 0$ there is some i such that $\mu_n(V_i) > 1 - \epsilon$ for all n. If not, then for each i there exists n_i such that $\mu_{n_i}(V_i) \leq 1 - \epsilon$. But then, by the portmanteau lemma, for each i,

$$\mu(V_i) \leq \liminf_{j \to \infty} \mu_{n_j}(V_i) \leq \liminf_{j \to \infty} \mu_{n_j}(V_{n_j}) \leq 1 - \epsilon.$$

But this is impossible since $V_i \uparrow S$. This proves the claim.

Now take any $\epsilon > 0$ and $k \geq 1$. Recall that separable metric spaces have the Lindëlof property, namely, that any open cover has a countable

subcover. Thus, there is a sequence of open balls $\{B_{k,i}\}_{i=1}^{\infty}$ of radius $1/k$ that cover S. By the above claim, we can choose n_k such that for any n,

$$\mu_n\left(\bigcup_{i=1}^{n_k} B_{k,i}\right) \geq 1 - 2^{-k}\epsilon.$$

Define

$$L := \bigcap_{k=1}^{\infty} \bigcup_{i=1}^{n_k} B_{k,i}.$$

Then for any n,

$$\mu_n(L) \geq 1 - \sum_{k=1}^{\infty} \mu_n\left(\bigcap_{i=1}^{n_k} B_{k,i}^c\right) \geq 1 - \sum_{k=1}^{\infty} 2^{-k}\epsilon = 1 - \epsilon.$$

Now recall that any totally bounded subset of a complete metric space is precompact. By construction, L is totally bounded. Therefore L is precompact, and hence the closure $\overline{L}$ of L is a compact set which satisfies $\mu_n(\overline{L}) \geq 1 - \epsilon$ for all n. □

Helly's selection theorem generalizes to Polish spaces. The generalization is known as Prokhorov's theorem.

THEOREM 8.3.3 (Prokhorov's theorem). *If (S, ρ) is a Polish space and $\{\mu_n\}_{n=1}^{\infty}$ is a tight family of probability measures on S, then there is a subsequence $\{\mu_{n_k}\}_{k=1}^{\infty}$ converging weakly to a probability measure μ as $k \to \infty$.*

There are various proofs of Prokhorov's theorem. The proof given below is a purely measure-theoretic argument. There are other proofs that are more functional analytic in nature. In the proof below, the measure μ is constructed using the technique of outer measures, as follows.

Let $K_1 \subseteq K_2 \subseteq \cdots$ be a sequence of compact sets such that $\mu_n(K_i) \geq 1 - 1/i$ for all n. Such a sequence exists because the family $\{\mu_n\}_{n=1}^{\infty}$ is tight. Let D be a countable dense subset of S, which exists because S is separable. Let $\mathcal{B}$ be the set of all closed balls with centers at elements of D and nonzero rational radii. Then $\mathcal{B}$ is a countable collection. Let $\mathcal{C}$ be the collection of all finite unions of sets that are of the form $B \cap K_i$ for some $B \in \mathcal{B}$ and some i. (In particular, $\mathcal{C}$ contains the empty union, which equals $\emptyset$.) Then $\mathcal{C}$ is also countable. By the standard diagonal argument, there is a subsequence $\{n_k\}_{k=1}^{\infty}$ such that

$$\alpha(C) := \lim_{k \to \infty} \mu_{n_k}(C)$$

exists for every $C \in \mathcal{C}$. For every open set V, define

$$\beta(V) := \sup\{\alpha(C) : C \subseteq V, C \in \mathcal{C}\}.$$

Finally, for every $A \subseteq S$, let

$$\mu(A) := \inf\{\beta(V) : V \text{ open}, A \subseteq V\}.$$

In particular, $\mu(V) = \beta(V)$ if V is open. We will eventually show that μ is an outer measure. The proof requires several steps.

LEMMA 8.3.4. *The functional α is monotone and finitely subadditive on the class $\mathcal{C}$. Moreover, if $C_1, C_2 \in \mathcal{C}$ are disjoint, then $\alpha(C_1 \cup C_2) = \alpha(C_1) + \alpha(C_2)$.*

PROOF. This is obvious from the definition of α and the properties of measures. □

LEMMA 8.3.5. *If F is a closed set, V is an open set containing F, and some $C \in \mathcal{C}$ also contains F, then there exists $D \in \mathcal{C}$ such that $F \subseteq D \subseteq V$.*

PROOF. Since $F \subseteq C$ for some $C \in \mathcal{C}$, F is contained in some K_j. Since F is closed, this implies that F is compact. For each $x \in F$, choose $B_x \in \mathcal{B}$ such that $B_x \subseteq V$ and x belongs to the interior B_x° of B_x. This can be done because V is open and $V \supseteq F$. Then by the compactness of F, there exist finitely many $x_1, \ldots, x_n$ such that

$$F \subseteq \bigcup_{i=1}^n B_{x_i}^\circ \subseteq \bigcup_{i=1}^n B_{x_i} \subseteq V.$$

To complete the proof, take $D = (B_{x_1} \cap K_j) \cup \cdots \cup (B_{x_n} \cap K_j)$. □

LEMMA 8.3.6. *The functional β is finitely subadditive on open sets.*

PROOF. Take any two open sets V_1 and V_2, and any $C \in \mathcal{C}$ such that $C \subseteq V_1 \cup V_2$. Define

$$F_1 := \{x \in C : \rho(x, V_1^c) \geq \rho(x, V_2^c)\},$$
$$F_2 := \{x \in C : \rho(x, V_2^c) \geq \rho(x, V_1^c)\},$$

where $\rho(x, A)$ is defined as in (8.2.1) for any A. It is not hard to see (using (8.2.2), for instance), that $x \mapsto \rho(x, A)$ is a continuous map for any A. Therefore the sets F_1 and F_2 are closed. Moreover, if $x \notin V_1$ and $x \in F_1$, then the definition of F_1 implies that $x \notin V_2$, which is impossible since $F_1 \subseteq C \subseteq V_1 \cup V_2$. Thus, $F_1 \subseteq V_1$. Similarly, $F_2 \subseteq V_2$. Moreover F_1 and F_2 are both subsets of C. Therefore by Lemma 8.3.5, there exist $C_1, C_2 \in \mathcal{C}$ such that $F_1 \subseteq C_1 \subseteq V_1$ and $F_2 \subseteq C_2 \subseteq V_2$. But then $C = F_1 \cup F_2 \subseteq C_1 \cup C_2$, and therefore by Lemma 8.3.4,

$$\alpha(C) \leq \alpha(C_1) + \alpha(C_2) \leq \beta(V_1) + \beta(V_2).$$

Taking supremum over C completes the proof. □

LEMMA 8.3.7. *The functional β is countably subadditive on open sets.*

PROOF. Let $V_1, V_2, \ldots$ be a sequence of open sets and let C be an element of $\mathcal{C}$ that is contained in the union of these sets. Since C is compact, there is some finite n such that $C \subseteq V_1 \cup \cdots \cup V_n$. Then by the definition of β and Lemma 8.3.6,

$$\alpha(C) \leq \beta(V_1 \cup \cdots \cup V_n) \leq \sum_{i=1}^{n} \beta(V_i) \leq \sum_{i=1}^{\infty} \beta(V_i).$$

Taking supremum over C completes the proof. □

LEMMA 8.3.8. *The functional μ is an outer measure.*

PROOF. It is clear from the definition that μ is monotone and satisfies $\mu(\emptyset) = 0$. We only need to show that μ is subadditive. Take any sequence of set $A_1, A_2, \ldots \subseteq S$ and let A be their union. Take any $\epsilon > 0$. For each i, let V_i be an open set containing A_i such that $\beta(V_i) \leq \mu(A_i) + 2^{-i}\epsilon$. Then by Lemma 8.3.7,

$$\begin{aligned}\mu(A) &\leq \beta\left(\bigcup_{i=1}^{\infty} V_i\right) \leq \sum_{i=1}^{\infty} \beta(V_i) \\ &\leq \sum_{i=1}^{\infty} (\mu(A_i) + 2^{-i}\epsilon) = \epsilon + \sum_{i=1}^{\infty} \mu(A_i).\end{aligned}$$

Since ϵ is arbitrary, this completes the proof. □

LEMMA 8.3.9. *For any open V and closed F,*

$$\beta(V) \geq \mu(V \cap F) + \mu(V \cap F^c).$$

PROOF. Choose any $C_1 \in \mathcal{C}$, $C_1 \subseteq V \cap F^c$. Having chosen C_1, choose $C_2 \in \mathcal{C}$, $C_2 \subseteq V \cap C_1^c$. Then C_1 and C_2 are disjoint, $C_1 \cup C_2 \in \mathcal{C}$, and $C_1 \cup C_2 \subseteq V$. Therefore by Lemma 8.3.4,

$$\beta(V) \geq \alpha(C_1 \cup C_2) = \alpha(C_1) + \alpha(C_2).$$

Taking supremum over all choices of C_2, we get

$$\beta(V) \geq \alpha(C_1) + \beta(V \cap C_1^c) \geq \alpha(C_1) + \mu(V \cap F),$$

where the second inequality holds because $V \cap C_1^c$ is an open set that contains $V \cap F$. Now taking supremum over C_1 completes the proof. □

We are finally ready to prove Prokhorov's theorem.

PROOF OF THEOREM 8.3.3. Let $\mathcal{F}$ be the set of all μ-measurable sets, as defined in Section 1.4. Then recall that by Theorem 1.4.3, $\mathcal{F}$ is a σ-algebra and μ is a measure on $\mathcal{F}$. We need to show that (a) $\mathcal{F}$ contains the Borel σ-algebra of S, (b) μ is a probability measure, and (c) μ_{n_k} converges weakly to μ.

To prove (a), take any closed set F and any set $D \subseteq S$. Then for any open $V \supseteq D$, Lemma 8.3.9 gives

$$\beta(V) \geq \mu(V \cap F) + \mu(V \cap F^c) \geq \mu(D \cap F) + \mu(D \cap F^c),$$

where the second inequality holds because μ is monotone. Now taking infimum over V shows that $F \in \mathcal{F}$. Therefore $\mathcal{F}$ contains the Borel σ-algebra. To prove (b), take any i and observe that by the compactness of K_i, K_i can be covered by finitely many elements of $\mathcal{B}$. Consequently, K_i itself is an element of $\mathcal{C}$. Therefore

$$\mu(S) = \beta(S) \geq \alpha(K_i) \geq 1 - \frac{1}{i}.$$

Since this holds for all i, we get $\mu(S) \geq 1$. On the other hand, it is clear from the definition of μ that $\mu(S) \leq 1$. Thus, μ is a probability measure. Finally, to prove (c), notice that for any open set V and any $C \in \mathcal{C}$, $C \subseteq V$,

$$\alpha(C) = \lim_{k\to\infty} \mu_{n_k}(C) \leq \liminf_{k\to\infty} \mu_{n_k}(V),$$

and take supremum over C. □

EXERCISE 8.3.10. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of $\mathbb{R}^d$-valued random vectors. Show that the sequence is tight if and only if for every $\epsilon > 0$, there is some R such that $\mathbb{P}(|X_n| > R) \leq \epsilon$ for all n, where $|X_n|$ is the Euclidean norm of X_n.

8.4. Skorokhod's representation theorem

We know that convergence in law does not imply almost sure convergence. In fact, weak convergence does not even assume that the random variables are defined on the same probability space. It turns out, however, that a certain kind of converse can be proved. It is sometimes useful for proving theorems and constructing random variables.

THEOREM 8.4.1 (Skorokhod's representation theorem). *Let S be a Polish space and let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of probability measures on S that converge weakly to a limit μ. Then it is possible to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sequence of S-valued random variable $\{X_n\}_{n=1}^{\infty}$ on Ω, and another S-valued random variable X on Ω, such that $X_n \sim \mu_n$ for each n, $X \sim \mu$ and $X_n \to X$ almost surely.*

We need a topological lemma about Polish spaces. Recall that the diameter of a set A in a metric space (S, ρ) is defined as

$$\operatorname{diam}(A) = \sup_{x,y\in A} \rho(x, y).$$

LEMMA 8.4.2. *Let S be a Polish space and μ be a probability measure on S. Then for any $\epsilon > 0$, there is a partition $A_0, A_1, \ldots, A_n$ of S into measurable sets such that $\mu(A_0) < \epsilon$, A_0 is open, $\mu(\partial A_i) = 0$ and $\mu(A_i) > 0$ for $1 \leq i \leq n$, and $\operatorname{diam}(A_i) \leq \epsilon$ for $1 \leq i \leq n$.*

PROOF. Let $B(x,r)$ denote the closed ball of radius r centered at a point x in S. Since $\partial B(x,r)$ and $\partial B(x,s)$ are disjoint for any distinct r and s, it follows that for any x, there can be only countably many r such that $\mu(\partial B(x,r)) > 0$. In particular, for each x we can find $r_x \in (0, \epsilon/2)$ such that $\mu(\partial B(x, r_x)) = 0$. Then the interiors of the balls $B(x, r_x)$ form a countable open cover of S. Since S is a separable metric space, it has the property that any open cover has a countable subcover (the Lindelöf property). Thus, there exist $x_1, x_2, \ldots$ such that

$$S = \bigcup_{i=1}^{\infty} B(x_i, r_{x_i}).$$

Now choose n so large that

$$\mu(S) - \mu\left(\bigcup_{i=1}^{n} B(x_i, r_{x_i})\right) < \epsilon.$$

Let $B_i := B(x_i, r_{x_i})$ for $i = 1, \ldots, n$. Define $A_1 = B_1$ and

$$A_i = B_i \setminus (B_1 \cup \cdots \cup B_{i-1})$$

for $2 \le i \le n$. Finally, let $A_0 := S \setminus (B_1 \cup \cdots \cup B_n)$. Then by our choice of n, $\mu(A_0) < \epsilon$. By construction, A_0 is open and $\operatorname{diam}(A_i) \le \operatorname{diam}(B_i) \le \epsilon$ for $1 \le i \le n$. Finally, note that $\partial A_i \subseteq \partial B_1 \cup \cdots \cup \partial B_i$ for $1 \le i \le n$, which shows that $\mu(\partial A_i) = 0$ because $\mu(\partial B_j) = 0$ for every j. Finally, we merge those A_i with A_0 for which $\mu(A_i) = 0$, so that $\mu(A_i) > 0$ for all i that remain. □

PROOF OF THEOREM 8.4.1. For each $j \ge 1$, choose sets $A_0^j, \ldots, A_{k_j}^j$ satisfying the conditions of Lemma 8.4.2 with $\epsilon = 2^{-j}$ and μ as in the statement of Theorem 8.4.1.

Next, for each j, find n_j such that if $n \ge n_j$, then

$$\mu_n(A_i^j) \ge (1 - 2^{-j})\mu(A_i^j) \tag{8.4.1}$$

for all $0 \le i \le k_j$. We can find such n_j by the portmanteau lemma, because $\mu_n \to \mu$ weakly, A_0^j is open, and $\mu(\partial A_i^j) = 0$. Without loss of generality, we can choose $\{n_j\}_{j=1}^{\infty}$ to be a strictly increasing sequence.

Take any $n \ge n_1$, and find j such that $n_j \le n < n_{j+1}$. For $0 \le i \le k_j$, define a probability measure $\mu_{n,i}$ on S as

$$\mu_{n,i}(A) := \frac{\mu_n(A \cap A_i^j)}{\mu_n(A_i^j)}$$

if $\mu_n(A_i^j) > 0$. If $\mu_n(A_i^j) = 0$, define $\mu_{n,i}$ to be some arbitrary probability measure on A_i^j if $1 \le i \le k_j$ and some arbitrary probability measure on S of $i = 0$. This can be done because A_i^j is nonempty for $i \ge 1$. It is easy to

see that $\mu_{n,i}$ is indeed a probability measure by the above construction, and moreover if $i \geq 1$, then $\mu_{n,i}(A_i^j) = 1$. Next, for each $0 \leq i \leq k_j$, let

$$p_{n,i} := 2^j(\mu_n(A_i^j) - (1 - 2^{-j})\mu(A_i^j)).$$

By (8.4.1), $p_{n,i} \geq 0$. Moreover,

$$\begin{aligned}\sum_{i=0}^{k_j} p_{n,i} &= 2^j \sum_{i=0}^{k_j} (\mu_n(A_i^j) - (1 - 2^{-j})\mu(A_i^j)) \\ &= 2^j(\mu_n(S) - (1 - 2^{-j})\mu(S)) = 1.\end{aligned}$$

Therefore the convex combination

$$\nu_n(A) := \sum_{i=0}^{k_j} \mu_{n,i}(A)p_{n,i}$$

also defines a probability measure on S.

Next, let $X \sim \mu$, $Y_n \sim \mu_n$, $Y_{n,i} \sim \mu_{n,i}$, $Z_n \sim \nu_n$, and $U \sim Unif[0,1]$ be independent random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where we take all $n \geq n_1$ and for each n, all $0 \leq i \leq k_j$ where j is the number such that $n_j \leq n < n_{j+1}$. Take any such n and j, and define

$$X_n := 1_{\{U > 2^{-j}\}} \sum_{i=0}^{k_j} 1_{\{X \in A_i^j\}} Y_{n,i} + 1_{\{U \leq 2^{-j}\}} Z_n.$$

Then for any $A \in \mathcal{B}(S)$,

$$\begin{aligned}\mathbb{P}(X_n \in A) &= \mathbb{P}(U > 2^{-j}) \sum_{i=0}^{k_j} \mathbb{P}(X \in A_i^j)\mathbb{P}(Y_{n,i} \in A) \\ &\quad + \mathbb{P}(U \leq 2^{-j})\mathbb{P}(Z_n \in A) \\ &= (1 - 2^{-j}) \sum_{i=0}^{k_j} \mu(A_i^j)\mu_{n,i}(A) + 2^{-j} \sum_{i=0}^{k_j} \mu_{n,i}(A)p_{n,i} \\ &= \sum_{i=0}^{k_j} \mu_n(A_i^j)\mu_{n,i}(A) = \sum_{i=0}^{k_j} \mu_n(A \cap A_i^j) = \mu_n(A).\end{aligned}$$

Thus, $X_n \sim \mu_n$. To complete the proof, we need to show that $X_n \to X$ a.s. To show this, first note that by the first Borel–Cantelli lemma,

$$\mathbb{P}(X \notin A_0^j \text{ and } U > 2^{-j} \text{ for all sufficiently large } j) = 1.$$

If the above event happens, then for all sufficiently large n, $X_n = Y_{n,i}$ for some i such that $X \in A_i^j$. In particular, $\rho(X_n, X) \leq \epsilon$, because $\operatorname{diam}(A_i^j) \leq \epsilon$. This proves that $X_n \to X$ a.s. $\square$

8.5. Convergence in probability on Polish spaces

Let (S, ρ) be a Polish space. A sequence $\{X_n\}_{n=1}^\infty$ of S-valued random variables is said to converge in probability to a random variable X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\rho(X_n, X) \geq \epsilon) = 0.$$

Here we implicitly assumed that all the random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If X is a constant, this assumption can be dropped.

PROPOSITION 8.5.1. *Convergence in probability implies convergence in distribution on Polish spaces.*

PROOF. Let (S, ρ) be a Polish space, and suppose that $\{X_n\}_{n=1}^\infty$ is a sequence of S-valued random variables converging to a random variable X in probability. Take any bounded uniformly continuous function $f : S \to \mathbb{R}$. Take any $\epsilon > 0$. Then there is some $\delta > 0$ such that $|f(x) - f(y)| \leq \epsilon$ whenever $\rho(x, y) \leq \delta$. Thus,

$$\mathbb{P}(|f(X_n) - f(X)| > \epsilon) \leq \mathbb{P}(\rho(X_n, X) > \delta),$$

which shows that $f(X_n) \to f(X)$ in probability. Since f is bounded, Proposition 7.2.9 shows that $f(X_n) \to f(X)$ in L^1. In particular, $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$. Thus, $X_n \to X$ in distribution. □

PROPOSITION 8.5.2. *Let (S, ρ) be a Polish space. A sequence of S-valued random variables $\{X_n\}_{n=1}^\infty$ converges to a constant $c \in S$ in probability if and only if $X_n \to c$ in distribution.*

PROOF. If $X_n \to c$ in probability, then it follows from Proposition 8.5.1 that $X_n \to c$ in distribution. Conversely, suppose that $X_n \to c$ in distribution. Take any ϵ and let $F := \{x \in S : \rho(x, c) \geq \epsilon\}$. Then F is a closed set, and so by assertion (d) in the portmanteau lemma,

$$\limsup_{n\to\infty} \mathbb{P}(\rho(X_n, c) \geq \epsilon) = \limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leq 0,$$

since $c \notin F$. This shows that $X_n \to c$ in probability. □

There is also a version of Slutsky's theorem for Polish spaces.

PROPOSITION 8.5.3 (Slutsky's theorem for Polish spaces). *Let (S, ρ) be a Polish space. Let $\{X_n\}_{n=1}^\infty$ and $\{Y_n\}_{n=1}^\infty$ be two sequences of S-valued random variables, defined on the same probability space, such that $X_n \to X$ in distribution and $Y_n \to c$ in probability, where X is an S-valued random variable and $c \in S$ is a constant. Then, as random variables on $S \times S$, $(X_n, Y_n) \to (X, c)$ in distribution.*

PROOF. There are many metrics that metrize the product topology on $S \times S$. For example, we can use the metric

$$d((x,y),(w,z)) := \rho(x,w) + \rho(y,z).$$

Suppose that $f : S \times S \to \mathbb{R}$ is a bounded and uniformly continuous function. Take any $\epsilon > 0$. Then there is some $\delta > 0$ such that $|f(x,y) - f(w,z)| \leq \epsilon$ whenever $d((x,y),(w,z)) \leq \delta$. Then

$$\mathbb{P}(|f(X_n,Y_n) - f(X_n,c)| > \epsilon) \leq \mathbb{P}(\rho(Y_n,c) > \delta),$$

which implies that $f(X_n,Y_n) - f(X_n,c) \to 0$ in probability. By Proposition 7.2.9, this shows that $f(X_n,Y_n) - f(X_n,c) \to 0$ in L^1. In particular, $\mathbb{E}f(X_n,Y_n) - \mathbb{E}f(X_n,c) \to 0$. On the other hand, $x \mapsto f(x,c)$ is a bounded continuous function on S, and so $\mathbb{E}f(X_n,c) \to \mathbb{E}f(X,c)$. Thus, $\mathbb{E}f(X_n,Y_n) \to \mathbb{E}f(X,c)$. By the portmanteau lemma, this completes the proof. □

8.6. Multivariate inversion formula

The inversion formula for characteristic functions of random vectors is a straightforward analogue of the univariate formula that was presented in Theorem 7.7.1. In the following, $a \cdot b$ denotes the scalar product of two vectors a and b, and $|a|$ denotes the Euclidean norm of a.

THEOREM 8.6.1. *Let X be an n-dimensional random vector with characteristic function ϕ. For each $\theta > 0$, define a function $f_\theta : \mathbb{R}^n \to \mathbb{C}$ as*

$$f_\theta(x) := \frac{1}{2\pi} \int_{\mathbb{R}^n} e^{-\mathrm{i}t\cdot x - \theta|t|^2} \phi(t) dt.$$

Then for any bounded continuous $g : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}(g(X)) = \lim_{\theta \to 0} \int_{\mathbb{R}^n} g(x) f_\theta(x) dx.$$

PROOF. Proceeding exactly as in the proof of Theorem 7.7.1, we can deduce that f_θ is the p.d.f. of a multivariate normal random vector Z_θ with mean zero and i.i.d. components with variance 2θ. It is easy to show that $Z_\theta \to 0$ in probability as $\theta \to 0$, and therefore by Slutsky's theorem for Polish spaces, $X + Z_\theta \to X$ in distribution as $\theta \to 0$. The proof is now completed as before. □

Just like Corollary 7.7.2, the above theorem has the following corollary about random vectors.

COROLLARY 8.6.2. *Two random vectors have the same law if and only if they have the same characteristic function.*

PROOF. Same as the proof of Corollary 7.7.2, using Theorem 8.6.1 and Corollary 8.2.2 instead of Theorem 7.7.1 and Proposition 7.6.1. □

EXERCISE 8.6.3. Prove a multivariate version of Corollary 7.7.3.

8.7. Multivariate Lévy continuity theorem

The multivariate version of Lévy's continuity theorem is a straightforward generalization of the univariate theorem. The only slightly tricky part of the proof is the proof of tightness, since we did not prove a tail bound for random vectors in terms characteristic functions.

THEOREM 8.7.1 (Multivariate Lévy continuity theorem). *A sequence of random vectors $\{X_n\}_{n\geq 1}$ converges in distribution to a random vector X if and only if the sequence of characteristic functions $\{\phi_{X_n}\}_{n\geq 1}$ converges to the characteristic function ϕ_X pointwise.*

PROOF. If $X_n \xrightarrow{d} X$, then it follows from the definition of weak convergence that the characteristic functions converge. Conversely, suppose that $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}^m$, where m is the dimension of the random vectors. Let $X_n^1, \ldots, X_n^m$ denote the coordinates of X_n. Similarly, let $X^1, \ldots, X^m$ be the coordinates of X. Then note that the pointwise convergence of ϕ_{X_n} to ϕ_X automatically implies the pointwise convergence of $\phi_{X_n^i}$ to ϕ_{X^i} for each i. Consequently by Lévy's continuity theorem, $X_n^i \to X^i$ in distribution for each i. In particular, for each i, $\{X_n^i\}_{n=1}^\infty$ is a tight family of random variables. Thus, given any $\epsilon > 0$, there is some $K^i > 0$ such that $\mathbb{P}(X_n^i \in [-K^i, K^i]) \geq 1 - \epsilon/m$ for all n. Let $K = \max\{K^1, \ldots, K^m\}$, and let R be the cube $[-K, K]^m$. Then for any n,

$$\mathbb{P}(X_n \notin R) \leq \sum_{i=1}^m \mathbb{P}(X_n^i \notin [-K, K]) \leq \epsilon.$$

Thus, we have established that $\{X_n\}_{n=1}^\infty$ is a tight family of $\mathbb{R}^m$-valued random vectors. We can complete the proof of the theorem as we did for the original Lévy continuity theorem, using Corollary 8.6.2. □

EXERCISE 8.7.2. Let (S, ρ) be a Polish space. Suppose that for each $1 \leq j \leq m$, $\{X_{n,j}\}_{n=1}^\infty$ is a sequence of S-valued random variables converging weakly to a random variable X_j. Suppose that for each n, the random variables $X_{n,1}, \ldots, X_{n,m}$ are defined on the same probability space and are independent, and the same holds for $(X_1, \ldots, X_m)$. Then show that the S^n-valued random variable $(X_{n,1}, \ldots, X_{n,m})$ converges weakly to $(X_1, \ldots, X_m)$.

8.8. The Cramér–Wold device

The Cramér–Wold device is a simple idea about proving weak convergence of random vectors using weak convergence of random variables. We will use it to prove the multivariate central limit theorem.

PROPOSITION 8.8.1 (Cramér–Wold theorem). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of m-dimensional random vectors and X be another m-dimensional random vector. Then $X_n \xrightarrow{d} X$ if and only if $t \cdot X_n \xrightarrow{d} t \cdot X$ for every $t \in \mathbb{R}^m$.*

PROOF. If $X_n \xrightarrow{d} X$, then $\mathbb{E}f(t \cdot X_n) \to \mathbb{E}f(t \cdot X)$ for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ and every $t \in \mathbb{R}^m$. This shows that $t \cdot X_n \xrightarrow{d} t \cdot X$ for every t. Conversely, suppose that $t \cdot X_n \xrightarrow{d} t \cdot X$ for every t. Then

$$\phi_{X_n}(t) = \mathbb{E}(e^{\mathrm{i}t \cdot X_n}) \to \mathbb{E}(e^{\mathrm{i}t \cdot X}) = \phi_X(t).$$

Therefore, $X_n \xrightarrow{d} X$ by the multivariate Lévy continuity theorem. □

8.9. The multivariate CLT for i.i.d. sums

In this section we will prove a multivariate version of the central limit theorem for sums of i.i.d. random variables. The proof is a simple consequence of the univariate CLT and the Cramér–Wold device. Recall that $N_m(\mu, \Sigma)$ denotes the m-dimensional normal distribution with mean vector μ and covariance matrix Σ.

THEOREM 8.9.1 (Multivariate CLT for i.i.d. sums). *Let $X_1, X_2, \ldots$ be i.i.d. m-dimensional random vectors with mean vector μ and covariance matrix Σ. Then, as $n \to \infty$,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N_m(0, \Sigma).$$

PROOF. Let $Z \sim N_m(0, \Sigma)$. Take any nonzero $t \in \mathbb{R}^m$. Let $Y := t \cdot Z$. Then by Exercise 6.6.6, $Y \sim N(0, t^T \Sigma t)$. Next, let $Y_i := t \cdot (X_i - \mu)$. Then $Y_1, Y_2, \ldots$ are i.i.d. random variables, with $\mathbb{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = t^T \Sigma t$. Therefore by the univariate CLT for i.i.d. sums,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \xrightarrow{d} N(0, t^T \Sigma t).$$

Combining the two, we get

$$t \cdot \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \right) \xrightarrow{d} t \cdot Z.$$

Since this happens for every $t \in \mathbb{R}^m$, the result now follows by the Cramér–Wold device. □

8.10. The spaces $C[0,1]$ and $C[0,\infty)$

Recall the Polish spaces $C[0,1]$ and $C[0,\infty)$ defined in Exercises 8.1.3 and 8.1.4. In this section we will study some probabilistic aspects of these spaces.

DEFINITION 8.10.1. For each n and each $t_1, \ldots, t_n \in [0,1]$, the projection map $\pi_{t_1,\ldots,t_n} : C[0,1] \to \mathbb{R}^n$ is defined as

$$\pi_{t_1,\ldots,t_n}(f) := (f(t_1), \ldots, f(t_n)).$$

The projection maps are defined similarly on $C[0,\infty)$.

It is easy to see that the projection maps are continuous and hence measurable. However, more is true:

PROPOSITION 8.10.2. *The finite-dimensional projection maps generate the Borel σ-algebras of $C[0,1]$ and $C[0,\infty)$.*

PROOF. Let us first consider the case of $C[0,1]$. Given $f \in C[0,1]$ and some n and some $t_1, \ldots, t_n$, we can reconstruct an 'approximation' of f from $\pi_{t_1,\ldots,t_n}(f)$ as the function g which satisfies $g(t_i) = f(t_i)$ for each i, and linearly interpolate when t is between some t_i and t_j. Define g to be constant to the left of the smallest t_i and to the right of the largest t_i, such that continuity is preserved. Moreover, the map $\rho_{t_1,\ldots,t_n}$ that constructs g from $\pi_{t_1,\ldots,t_n}(f)$ is a continuous map from $\mathbb{R}^n$ into $C[0,1]$. Thus, if we let

$$\xi_{t_1,\ldots,t_n} := \rho_{t_1,\ldots,t_n} \circ \pi_{t_1,\ldots,t_n},$$

then $\xi_{t_1,\ldots,t_n}$ is a continuous map from $C[0,1]$ into itself, and moreover, it is measurable with respect to the σ-algebra generated by $\pi_{t_1,\ldots,t_n}$.

Now let $\{t_1, t_2, \ldots\}$ be a dense subset of $[0,1]$ chosen in such a way that

$$\lim_{n\to\infty} \max_{0\le t\le 1} \min_{1\le i\le n} |t - t_i| = 0. \tag{8.10.1}$$

(It is easy to construct such a sequence.) Let $f_n := \xi_{t_1,\ldots,t_n}(f)$. By the uniform continuity of f, the construction of f_n, and the property (8.10.1) of the sequence $\{t_n\}_{n=1}^{\infty}$, it is not hard to show that $f_n \to f$ in the topology of $C[0,1]$. Since each f_n is measurable with respect to the σ-algebra $\mathcal{F}$ generated by the finite-dimensional projection maps, Proposition 2.1.14 shows that the identity map from $(C[0,1], \mathcal{F})$ into $(C[0,1], \mathcal{B}(C[0,1]))$ is measurable. This proves that $\mathcal{F} \supseteq \mathcal{B}(C[0,1])$. We have already observed the opposite inclusion in the sentence preceding the statement of the proposition. This completes the argument for $C[0,1]$.

For $C[0,\infty)$, the argument is exactly the same, except that we have to replace the maximum over $t \in [0,1]$ in (8.10.1) with maximum over $t \in [0,K]$, and then impose that the condition should hold for every K. Finally, we have to observe that any $f \in C[0,\infty)$ is uniformly continuous on any compact interval, and convergence is $C[0,\infty)$ is equivalent to uniform convergence on compact sets. □

Given any probability measure μ on $C[0,1]$ or $C[0,\infty)$, the push-forwards of μ under the projection maps are known as the finite-dimensional distributions of μ. In the language of random variables, if X is a random variable with law μ, then the finite-dimensional distributions of μ are the laws of random vectors like $(X(t_1),\ldots,X(t_n))$, where n and $t_1,\ldots,t_n$ are arbitrary.

PROPOSITION 8.10.3. *On $C[0,1]$ and $C[0,\infty)$, a probability measure is determined by its finite-dimensional distributions.*

PROOF. Given a probability measure, the finite-dimensional distributions determine the probabilities of all sets of the form $\pi^{-1}_{t_1,\ldots,t_n}(A)$, where n and $t_1,\ldots,t_n$ are arbitrary, and $A \in \mathcal{B}(\mathbb{R}^n)$. Let $\mathcal{A}$ denote the collection of all such sets. It is not difficult to see that $\mathcal{A}$ is an algebra. Moreover, by Proposition 8.10.2, $\mathcal{A}$ generates the Borel σ-algebra of $C[0,1]$ (or $C[0,\infty)$). By Theorem 1.3.6, this shows that the finite-dimensional distributions determine the probability measure. □

COROLLARY 8.10.4. *If $\{\mu_n\}_{n=1}^{\infty}$ is a tight family of probability measures on $C[0,1]$ or $C[0,\infty)$, whose finite-dimensional distributions converge to limiting distributions, then the sequence itself converges weakly to a limit. Moreover, the limiting probability measure is uniquely determined by the limiting finite-dimensional distributions.*

PROOF. By Prokhorov's theorem, any subsequence has a further subsequence that converges weakly. By Proposition 8.10.3, there can be only one such limit point. The result can now be proved by a standard easy argument by contradiction. □

8.11. Tightness on $C[0,1]$

In this section we investigate criteria for tightness of sequences of probability measures on $C[0,1]$. Recall that the modulus of continuity of a function $f \in C[0,1]$ is defined as

$$\omega_f(\delta) := \sup\{|f(s) - f(t)| : 0 \le s,t \le 1,\ |s-t| \le \delta\}.$$

Recall that a family of functions $F \subseteq C[0,1]$ is called equicontinuous if for any $\epsilon > 0$, there is some $\delta > 0$ such that for and $f \in F$, $|f(s) - f(t)| \le \epsilon$ whenever $|s-t| \le \delta$. The family f is called uniformly bounded if there is some finite M such that $|f(t)| \le M$ for all $f \in F$ and all $t \in [0,1]$. Finally, recall the Arzelà–Ascoli theorem, which says that a closed set $F \subseteq C[0,1]$ is compact if and only if it is uniformly bounded and equicontinuous.

PROPOSITION 8.11.1. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of $C[0,1]$-valued random variables. The sequence is tight if and only the following two conditions hold:*

(i) *For any $\epsilon > 0$ there is some $a > 0$ such that for all n,*

$$\mathbb{P}(|X_n(0)| > a) \leq \epsilon.$$

(ii) *For any $\epsilon > 0$ and $\eta > 0$, there is some $\delta > 0$ such that for all large enough n (depending on ϵ and η),*

$$\mathbb{P}(\omega_{X_n}(\delta) > \eta) \leq \epsilon.$$

PROOF. First, suppose that the sequence $\{X_n\}_{n=1}^{\infty}$ is tight. Take any $\epsilon > 0$. Then there is some compact $K \subseteq C[0,1]$ such that $\mathbb{P}(X_n \notin K) \leq \epsilon$ for all n. By the Arzelà–Ascoli theorem, there is some finite M such that $|f(t)| \leq M$ for all $f \in K$ and all $t \in [0,1]$. Thus, for any n,

$$\mathbb{P}(|X_n(0)| > M) \leq \mathbb{P}(X_n \notin K) \leq \epsilon.$$

This proves condition (i). Next, take any positive η. Again by the Arzelà–Ascoli theorem, the family K is equicontinuous. Thus, there exists δ such that $\omega_f(\delta) \leq \eta$ for all $f \in K$. Therefore, for any n,

$$\mathbb{P}(\omega_{X_n}(\delta) > \eta) \leq \mathbb{P}(X_n \notin K) \leq \epsilon.$$

This proves condition (ii).

Conversely, suppose that conditions (i) and (ii) hold. We will first assume that (ii) holds for all n. Take any $\epsilon > 0$. Choose a so large that $\mathbb{P}(|X_n(0)| > a) \leq \epsilon/2$ for all n. Next, for each k, choose δ_k so small that

$$\mathbb{P}(\omega_{X_n}(\delta_k) > k^{-1}) \leq 2^{-k-1}\epsilon.$$

Finally, let

$$K := \{f \in C[0,1] : |f(0)| \leq a,\ \omega_f(\delta_k) \leq k^{-1} \text{ for all } k\}.$$

Then by construction, $\mathbb{P}(X_n \notin K) \leq \epsilon$ for all n. Moreover, it is easy to see that K is closed, uniformly bounded, and equicontinuous. Therefore by the Arzelà–Ascoli theorem, K is compact. This proves tightness.

Note that we have proved tightness under the assumption that (ii) holds for all n. Now suppose that (ii) holds only for $n \geq n_0$, where n_0 depend on ϵ and η. By Theorem 8.3.2, any single random variable is tight. This, and the fact that (i) and (ii) hold for any tight family (which we have shown above), allows us to decrease δ sufficiently so that (ii) holds for $n < n_0$ too. □

8.12. Donsker's theorem

In this section, we will prove a 'functional version' of the central limit theorem, that is known as Donsker's theorme or Donsker's invariance principle.

Let us start with a sequence of i.i.d. random variables $\{X_n\}_{n=1}^{\infty}$, with mean zero and variance one. For each n, let us use them to construct a

$C[0,1]$-valued random variable B_n as follows. Let $B_n(0) = 0$. When $t = i/n$ for some $1 \le i \le n$, let

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{i} X_j.$$

Finally, define B_n between $(i-1)/n$ and i/n by linear interpolation.

Donsker's theorem identifies the limiting distribution of B_n as $n \to \infty$. Just like in the central limit theorem, it turns out that the limiting distribution does not depend on the law of the X_i's.

THEOREM 8.12.1 (Donsker's invariance principle). *As $n \to \infty$, the sequence $\{B_n\}_{n=1}^{\infty}$ converges weakly to a $C[0,1]$-valued random variable B. The finite-dimensional distributions of B are as follows: $B(0) = 0$, and for any m and any $0 < t_1 < \cdots < t_m \le 1$, the random vector $(B(t_1), \ldots, B(t_m))$ has a multivariate normal distribution with mean vector zero and covariance structure given by $\operatorname{Cov}(B(t_i), B(t_j)) = \min\{t_i, t_j\}$.*

The limit random variable B is called Brownian motion, and its law is called the Wiener measure on $C[0,1]$. The proof of Donsker's theorem comes in a number of steps. First, we identify the limits of the finite-dimensional distributions.

LEMMA 8.12.2. *As $n \to \infty$, the finite-dimensional distributions of B_n converge weakly to the limits described in Theorem* 8.12.1.

PROOF. Since $B_n(0) = 0$ for all n, $B_n(0) \to 0$ in distribution. Take any m and any $0 < t_1 < t_2 < \cdots < t_m \le 1$. Let $k_i := [nt_i]$, and define

$$W_{n,i} := \frac{1}{\sqrt{n}} \sum_{j=1}^{k_i} X_j.$$

Also let $W_{n,0} = 0$ and $t_0 = 0$. It is a simple exercise to show by the Lindeberg–Feller CLT that for any $0 \le i \le m-1$,

$$W_{n,i+1} - W_{n,i} \xrightarrow{d} N(0, t_{i+1} - t_i).$$

Moreover, for any n, $W_{n,1} - W_{n,0}, W_{n,2} - W_{n,1}, \ldots, W_{n,m} - W_{n,m-1}$ are independent random variables. Therefore by Exercise 8.7.2, the random vector

$$W_n := (W_{n,1} - W_{n,0}, W_{n,2} - W_{n,1}, \ldots, W_{n,m} - W_{n,m-1})$$

converges in distribution to the random vector $Z = (Z_1, \ldots, Z_m)$, where $Z_1, \ldots, Z_m$ are independent random variables and $Z_i \sim N(0, t_i - t_{i-1})$ for each i. Now notice that for each n and i,

$$|W_{n,i} - B_n(t_i)| \le \frac{|X_{k_i+1}|}{\sqrt{n}},$$

where the right side is interpreted as 0 if $k_i = n$. This shows that as $n \to \infty$, $W_{n,i} - B_n(t_i) \to 0$ in probability. Therefore, if we let

$$U_n := (B_n(t_1), B_n(t_2) - B_n(t_1), \ldots, B_n(t_m) - B_n(t_{m-1})),$$

then $W_n - U_n \to 0$ in probability. Thus, $U_n \to Z$ in probability. The claimed result is easy to deduce from this. □

Next, we have to prove tightness. For that, the following maximal inequality is useful.

LEMMA 8.12.3. *Let $X_1, X_2, \ldots$ be independent random variables with mean zero and variance one. For each n, let $S_n := \sum_{i=1}^n X_i$. Then for any $n \geq 1$ and $t \geq 0$,*

$$\mathbb{P}(\max_{1\leq i\leq n} |S_i| \geq t\sqrt{n}) \leq 2\mathbb{P}(|S_n| \geq (t-\sqrt{2})\sqrt{n}).$$

PROOF. Define

$$A_i := \{\max_{1\leq j<i} |S_j| < t\sqrt{n} \leq |S_i|\},$$

where the maximum of the left is interpreted as zero if $i = 1$. Also let

$$B := \{|S_n| \geq (t-\sqrt{2})\sqrt{n}\}.$$

Then

$$\begin{aligned}
\mathbb{P}(\max_{1\leq i\leq n} |S_n| \geq t\sqrt{n}) &= \mathbb{P}\biggl(\bigcup_{i=1}^n A_i\biggr) \\
&= \mathbb{P}\biggl(B \cap \bigcup_{i=1}^n A_i\biggr) + \mathbb{P}\biggl(B^c \cap \bigcup_{i=1}^n A_i\biggr) \\
&\leq \mathbb{P}(B) + \sum_{i=1}^{n-1} \mathbb{P}(A_i \cap B^c).
\end{aligned}$$

Now, $A_i \cap B^c$ implies $|S_n - S_i| \geq \sqrt{2n}$. But this event is independent of A_i. Thus, we have

$$\begin{aligned}
\mathbb{P}(A_i \cap B^c) &\leq \mathbb{P}(A_i \cap \{|S_n - S_i| \geq \sqrt{2n}\}) \\
&= \mathbb{P}(A_i)\mathbb{P}(|S_n - S_i| \geq \sqrt{2n}) \\
&\leq \mathbb{P}(A_i)\frac{n-i}{2n} \leq \frac{1}{2}\mathbb{P}(A_i),
\end{aligned}$$

where the second-to-last inequality follows by Chebychev's inequality. Combining, we get

$$\mathbb{P}(\max_{1\leq i\leq n} |S_n| \geq t\sqrt{n}) \leq \mathbb{P}(B) + \frac{1}{2}\sum_{i=1}^{n-1} \mathbb{P}(A_i).$$

But the A_i's are disjoint events. Therefore,

$$\sum_{i=1}^{n-1} \mathbb{P}(A_i) = \mathbb{P}\biggl(\bigcup_{i=1}^{n-1} A_i\biggr) \leq \mathbb{P}(\max_{1\leq i\leq n} |S_n| \geq t\sqrt{n}).$$

Plugging this upper bound into the right side of the previous display, we get the desired result. □

COROLLARY 8.12.4. *Let S_n be as in Lemma 8.12.3. Then for any $t > 3$, there is some n_0 such that for all $n \geq n_0$,*

$$\mathbb{P}(\max_{1 \leq i \leq n} |S_i| \geq t\sqrt{n}) \leq Ce^{-t^2/8},$$

where C is a constant that does not depend on n.

PROOF. Take any $t > 3$. Then $t - \sqrt{2} \geq t/2$. Let $Z \sim N(0,1)$. Then by the central limit theorem,

$$\lim_{n \to \infty} \mathbb{P}(|S_n| \geq (t - \sqrt{2})\sqrt{n}) = \mathbb{P}(|Z| \geq t - \sqrt{2}) \leq \mathbb{P}(|Z| \geq t/2).$$

To complete the proof, recall that by Exercise 5.9.5, we have the tail bound $\mathbb{P}(|Z| \geq t/2) \leq 2e^{-t^2/8}$. □

We are now ready to prove tightness.

LEMMA 8.12.5. *The sequence $\{B_n\}_{n=1}^{\infty}$ is tight.*

PROOF. Choose any $\eta, \epsilon, \delta > 0$. Note that for any $0 \leq s \leq t \leq 1$ such that $t - s \leq \delta$, we have $\lceil nt \rceil - \lfloor ns \rfloor \leq \lceil n\delta \rceil + 2$. From this it is not hard to see that

$$\omega_{B_n}(\delta) \leq \max\left\{ \frac{|S_l - S_k|}{\sqrt{n}} : 0 \leq k \leq l \leq n,\, l - k \leq \lceil n\delta \rceil + 2 \right\}.$$

Let E be the event that the quantity on the right is greater than η. Let $0 = k_0 \leq k_1 \leq \cdots \leq k_r = n$ satisfy $k_{i+1} - k_i = \lceil n\delta \rceil + 2$ for each $0 \leq i \leq r-2$, and $k_r - k_{r-1} \leq \lceil n\delta \rceil + 2$. Let n be so large that $\lceil n\delta \rceil + 2 \leq 2n\delta$.

Suppose that the event E happens, and let $k \leq l$ be a pair that witnesses that. Then either $k_i \leq k \leq l \leq k_{i+1}$ for some i, or $k_i \leq k \leq k_{i+1} \leq l \leq k_{i+2}$ for some i. In either case, the following event must happen:

$$E' := \left\{ \max_{k_i \leq m \leq k_{i+1}} |S_m - S_{k_i}| > \frac{\eta\sqrt{n}}{3} \text{ for some } i \right\},$$

because if it does not happen, then $|S_l - S_k|$ cannot be greater than $\eta\sqrt{n}$. But by Corollary 8.12.4, we have for any $0 \leq i \leq r - 1$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{k_i \leq m \leq k_{i+1}} |S_m - S_{k_i}| > \frac{\eta\sqrt{n}}{3} \right) \\
&= \mathbb{P}\left(\max_{k_i \leq m \leq k_{i+1}} |S_m - S_{k_i}| > \frac{\eta\sqrt{n}}{3\sqrt{k_{i+1} - k_i}} \sqrt{k_{i+1} - k_i} \right) \\
&\leq \mathbb{P}\left(\max_{k_i \leq m \leq k_{i+1}} |S_m - S_{k_i}| > \frac{\eta}{3\sqrt{2\delta}} \sqrt{k_{i+1} - k_i} \right) \leq C_1 e^{-C_2 \eta^2/\delta},
\end{aligned}$$

where C_1 and C_2 do not depend on n, η or δ, and n is large enough, depending on η and δ. Thus, for all large enough n,

$$\begin{aligned}\mathbb{P}(\omega_{B_n}(\delta) > \eta) &\le \mathbb{P}(E) \le \mathbb{P}(E') \\ &\le C_1 r e^{-C_2\eta^2/\delta} \le \frac{C_1 e^{-C_2\eta^2/\delta}}{\delta},\end{aligned}$$

where the last inequality holds since r is bounded above by a constant multiple of $1/\delta$. Thus, given any $\eta, \epsilon > 0$ we can choose δ such that condition (ii) of Proposition 8.11.1 holds for all large enough n. Condition (i) is automatic. This completes the proof. □

We now have all the ingredients to prove Donsker's theorem.

Proof of Theorem 8.12.1. By Lemma 8.12.5 and Prokhorov's theorem, any subsequence of $\{B_n\}_{n=1}^{\infty}$ has a weakly convergent subsequence. By Proposition 8.10.3 and Lemma 8.12.2, any two such weak limits must be the same. This suffices to prove that the whole sequences converges. □

Exercise 8.12.6. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean zero and variance one. For each n, let $S_n := \sum_{i=1}^n X_i$. Prove that the random variables

$$\frac{1}{\sqrt{n}} \max_{1 \le i \le n} S_i$$

and

$$\frac{1}{n} \sum_{i=1}^n 1_{\{S_i \ge 0\}}$$

converge in law as $n \to \infty$, and the limiting distributions do not depend on the distribution of the X_i's. (Hint: Use Donsker's theorem and Exercises 8.2.4 and 8.2.6. The second one is technically more challenging.)

Exercise 8.12.7. Prove a version of Donsker's theorem for sums of stationary m-dependent sequences. (Hint: The only challenge is to generalize Lemma 8.12.3. The rest goes through as in the i.i.d. case, applying the CLT for stationary m-dependent sequences that we derived earlier.)

8.13. Construction of Brownian motion

The probability measure on $C[0,1]$ obtained by taking the limit $n \to \infty$ in Donsker's theorem is known as the Wiener measure on $C[0,1]$. The statement of Donsker's theorem gives the finite-dimensional distributions of this measure. A random function B with this law is called Brownian motion on the time interval $[0,1]$. Brownian motion on the time interval $[0,\infty)$ is a similar $C[0,\infty)$-valued random variable. One can define it in the spirit of Donsker's theorem (see Exercise 8.13.2 below), but one can also define it more directly using a sequence of independent Brownian motions on $[0,1]$, as follows.

Let $B^1, B^2, \ldots$ be a sequence of i.i.d. Brownian motions on the time interval $[0,1]$. Define a $C[0,\infty)$-valued random variable B as follows. If $k \leq t < k+1$ for some integer $k \geq 0$, define

$$B(t) := \sum_{j=1}^{k-1} B^j(1) + B^k(t-k).$$

EXERCISE 8.13.1. Check that B defined above is indeed a $C[0,\infty)$-valued random variable, and that for any n and any $0 \leq t_1 \leq \cdots \leq t_n < \infty$, $(B(t_1), \ldots, B(t_n))$ is a multivariate Gaussian random vector with mean vector zero and covariance structure given by $\mathrm{Cov}(B(t_i), B(t_j)) = \min\{t_i, t_j\}$.

EXERCISE 8.13.2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean zero and variance one. Define $B_n(t)$ as in Donsker's theorem, but for all $t \in [0,\infty)$, so that B_n is now a $C[0,\infty)$-valued random variable. Prove that B_n converges weakly to the random function B defined above.