STATS 310 (MATH 230) Lecture notes (ongoing, to be updated)

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 $\mathcal{A}$ 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

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

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.)

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. In measure theory, $\mu(A)$ is thought of as a measure of the 'size' of A. When $\mu(A)$ is small, we think of A as a 'small' set. Following this line of thought, we think of $\mu(A \Delta B)$ as a kind of distance between two sets A and B (where $A \Delta B$ is the symmetric difference of A and B). If this is small, then A and B are 'almost the same set'. The following useful result shows that if μ is a probability measure, then any element of $\mathcal{F}$ can be arbitrarily well-approximated by elements of any generating algebra.

THEOREM 1.2.6. *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, take any $A_1, A_2, \ldots \in \mathcal{G}$. 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 $B' := \cup_{i=1}^{n} B_i$. It is not hard to see that

$$A' \Delta B' \subseteq \bigcup_{i=1}^{n} (A_i \Delta B_i).$$

Thus,

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

Again, it is not hard to check that $A \Delta B' \subseteq (A \setminus A') \cup (A' \Delta B')$. Therefore

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

Thus, $A \in \mathcal{G}$. This proves that $\mathcal{G}$ is a σ-algebra, and hence contains $\sigma(\mathcal{A}) = \mathcal{F}$. □

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. □

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. □

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 $E \subseteq \Omega$,*

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

PROOF. By monotonicity of ϕ, the left side dominates the right. For the opposite inequality, 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, 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,

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

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

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

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

$$\mu^*(A) := \inf\left\{\sum_{i=1}^{\infty} \mu(A_i) : A_1, A_2, \ldots \in \mathcal{A},\ A \subseteq \bigcup_{i=1}^{\infty} A_i\right\}. \tag{1.5.1}$$

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 $\{A_i\}_{i\geq 1}$ 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 1.4.3, 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 B is the disjoint union of A and $B \setminus A$, 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 $(a_i, b_i + 2^{-i}\epsilon)$, $i \ge 1$. 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 given by Corollary 1.6.5 is called the Lebesgue measure on the real line. The outer measure defined by the formula (1.5.1) in this case is called Lebesgue outer measure, and the σ-algebra induced by this outer measure is called the Lebesgue σ-algebra.

Exercise 1.6.7. Prove that for any $-\infty < a \leq b < \infty$, $\lambda([a, b]) = b - a$.

Exercise 1.6.8. 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.

Exercise 1.6.9. Let A be a Lebesgue measurable subset of $\mathbb{R}$ and let $x \in \mathbb{R}$. Let $A + x$ denote the set $\{a + x : a \in A\}$. Prove that $A + x$ is also Lebesgue measurable, and has the same Lebesgue measure as A.

1.7. Example of a non-measurable set

In this section we will describe a standard construction of a subset of $\mathbb{R}$ that is not Lebesgue measurable (and hence not Borel measurable). The critical ingredient is the

axiom of choice. Indeed, it can be shown that such a set cannot be produced without invoking the axiom of choice.

Define an equivalence relation $\sim$ on $[0,1]$ as: $x \sim y$ if $x - y \in \mathbb{Q}$. By the axiom of choice, there is a subset A of $[0,1]$ consisting of exactly one element from each equivalence class. We claim that A is not Lebesgue measurable. To prove this, let $q_1, q_2, \ldots$ be an enumeration of the rationals in $[-1,1]$, and define

$$B := \bigcup_{i=1}^{\infty} (A + q_i).$$

First, note that if $i \neq j$, then $(A + q_i) \cap (A + q_j) = \emptyset$. This is a simple consequence of the definition of A. Next, take any $x \in [0,1]$. Then there is some $a \in A$ such that $x - a \in \mathbb{Q}$. But $x - a \in [-1,1]$. Thus, $x \in A + q_i$ for some i. This implies that $B \supseteq [0,1]$. On the other hand, it is clear that $B \subseteq [-1,2]$. Now, if A is Lebesgue measurable, then by Exercise 1.6.9, so is B, and thus

$$\lambda(B) = \sum_{i=1}^{\infty} \lambda(A + q_i) = \sum_{i=1}^{\infty} \lambda(A) = \begin{cases} \infty & \text{if } \lambda(A) > 0, \\ 0 & \text{if } \lambda(A) = 0. \end{cases}$$

But since $[0,1] \subseteq B \subseteq [-1,2]$, we have $1 \leq \lambda(B) \leq 3$, which gives us a contradiction.

It can be shown that the Lebesgue σ-algebra is strictly bigger than the Borel σ-algebra. That is, there is a Lebesgue measurable set that is not Borel measurable. But this is a much more difficult result than the above example.

1.8. Completion of a measure space

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 which is complete. The smallest such extension is called the completion of the measure space $(\Omega, \mathcal{F}, \mu)$.

Exercise 1.8.1. Prove that the completion $(\Omega, \mathcal{F}_0, \mu_0)$ of an arbitrary measure space $(\Omega, \mathcal{F}, \mu)$ can be obtained as follows.

(1) Let Z be the set of all subsets of the zero μ-measure sets in $\mathcal{F}$.
(2) Let $\mathcal{F}_0 := \{A \cup B : A \in \mathcal{F},\ B \in Z\}$.
(3) If $C \in \mathcal{F}_0$ equals $A \cup B$ for some $A \in \mathcal{F}$ and $B \in Z$, let $\mu_0(C) := \mu(A)$. (First, show that this well-defined.)

Exercise 1.8.2. Prove that the completion of the Borel σ-algebra of $\mathbb{R}$ (or $\mathbb{R}^n$) obtained via the above prescription is the Lebesgue σ-algebra.

Recall that Lebesgue measure is defined on the Lebesgue 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 the functions $\liminf_{n\to\infty} f_n$ 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 any 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 $f \le g$ everywhere, then $\int f d\mu \le \int g d\mu$.*

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

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

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). *Suppose that $\{f_n\}_{n \geq 1}$ is 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 next exercise generalizes Lemma 2.3.2.

EXERCISE 2.3.5. 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 2.3.6. *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. □

EXERCISE 2.3.7 (Change of variable formula). Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ be a measure space and $(\Omega_2, \mathcal{F}_2)$ be a measurable space. Let $g : \Omega_1 \to \Omega_2$ and $f : \Omega_2 \to \mathbb{R}$ be measurable functions such that $f \circ g$ is integrable. Let ν be the measure on Ω_2 induced by g, that is, $\nu(A) := \mu_1(g^{-1}(A))$ for each $A \in \mathcal{F}_2$. Then show that f is integrable with respect to ν, and

$$\int_{\Omega_1} f \circ g d\mu_1 = \int_{\Omega_2} f d\nu.$$

(Hint: Start with $f = 1_A$ and use Proposition 2.3.6 and the monotone convergence theorem to generalize.)

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 2.3.6, 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, $|\int f d\mu| \leq \int |f| d\mu$.

The reader may be wondering 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.)

The following exercise has many applications in probability theory.

EXERCISE 2.4.7. If $f_1, f_2, \ldots$ are measurable functions from Ω into $[0, \infty]$, show that $\int \sum f_i d\mu = \sum \int f_i d\mu$.

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 show that $\int f d\mu \geq \int g d\mu$, 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 the collection of finite disjoint unions of 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. It suffices to show that

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

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 union of $A_{i,n}$ over $i \in I$. Now, if $y \in A_{i,n} \cap A_{j,n}$ for some distinct $i, j \in I$, then since $x \in (A_{i,1} \times \cdots \times A_{i,n-1}) \cap (A_{j,1} \times \cdots \times A_{j,n-1})$, we get that $(x, y) \in (A_{i,1} \times \cdots \times A_{i,n}) \cap (A_{j,1} \times \cdots \times A_{j,n})$,

which is impossible. Thus, the sets $A_{i,n}$, as i ranges over I, are disjoint. This gives

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

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 2.4.7 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 $A \in \mathcal{F}_3$ 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 $E \in \mathcal{F}_1 \times \mathcal{F}_2$ such that $E_x \in \mathcal{F}_2$, 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 let μ_1 and μ_2 be 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 2.3.6.

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 2.3.6, 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,

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

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 $A_1 \times A_2 \times \cdots$ such that $A_i = \Omega_i$ for all but finitely many i.

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.

3.4. Kolmogorov's extension theorem

While infinite dimensional product measures are sufficient for many purposes, sometimes we need to construct non-product measures on infinite dimensional spaces. The fundamental tool for such constructions, known as Kolmogorov's existence theorem (also as Kolmogorov's consistency theorem, or the Daniell–Kolmogorov theorem), is stated below.. The version given here works only for Euclidean spaces. There is a more general version that works for any complete separable metric space — see Exercise 12.3.4 in Chapter 12.

THEOREM 3.4.1 (Kolmogorov's extension theorem). *Take any $d \geq 1$. Suppose that for each n, we have a probability measure μ_n on $((\mathbb{R}^d)^n, \mathcal{B}((\mathbb{R}^d)^n))$ such that the collection $\{\mu_n\}_{n\geq 1}$ satisfies the following consistency property: For each n, if $(X_1, \ldots, X_{n+1})$ is a random array with law μ_{n+1}, then $(X_1, \ldots, X_n)$ has law μ_n. Then there is a unique*

probability measure μ on the product σ-algebra of $(\mathbb{R}^d)^{\mathbb{N}}$ such that if $(X_1, X_2, \ldots) \sim \mu$, then for all n, $(X_1, \ldots, X_n) \sim \mu_n$.

To prove this result, we first need to establish that every probability measure on Euclidean space has a property known as 'regularity'.

LEMMA 3.4.2 (Regularity of probability measures). *Let μ be a probability measure on $\mathbb{R}^n$. Then for any $B \in \mathcal{B}(\mathbb{R}^n)$,*

$$\begin{aligned}\mu(B) &= \inf\{\mu(U) : U \supseteq B,\ U \text{ open}\} \\ &= \sup\{\mu(C) : C \subseteq B,\ C \text{ closed}\} \\ &= \sup\{\mu(K) : K \subseteq B,\ K \text{ compact}\}.\end{aligned}$$

PROOF. Let $\mathcal{C}$ be the collection of Borel sets B for which the first and second identities displayed above hold true. It is easy to see that $\mathcal{C}$ contains all closed sets, since for closed set C can be written as the decreasing intersection of the open sets

$$U_k := \{x \in \mathbb{R}^n : |x - y| < 1/k \text{ for some } y \in C\},$$

where $|x-y|$ denotes the Euclidean norm of $x-y$. Thus, the proof of the first two identities will be complete if we can show that $\mathcal{C}$ is a σ-algebra.

It is easy to see that $\emptyset \in \mathcal{C}$. Next, if $B \in \mathcal{C}$, then it is not hard to see that $B^c \in \mathcal{C}$ using the observation that if $C \subseteq B$ is a closed set and $U \supseteq B$ is an open set, then $C^c \supseteq B^c$ is an open set and $U^c \subseteq B^c$ is a closed set. Next, take any $B_1, B_2, \ldots \in \mathcal{C}$ and let B be their union. Take any $\epsilon > 0$. Find closed sets $C_1 \subseteq B_1, C_2 \subseteq B_2, \ldots$ such that $\mu(C_i) \geq \mu(B_i) - 2^{-i}\epsilon$ for each i, and open sets $U_1 \supseteq B_1, U_2 \supseteq B_2, \ldots$ such that $\mu(U_i) \leq \mu(B_i) + 2^{-i}\epsilon$ for each i. Let $U := \cup_{i \geq 1} U_i$, so that U is an open set, and

$$\mu(U) - \mu(B) \leq \sum_{i=1}^{\infty} (\mu(U_i) - \mu(B_i)) \leq \epsilon.$$

Similarly, letting $C := \cup_{i \geq 1} C_i$, we have

$$\mu(B) - \mu(C) \leq \sum_{i=1}^{\infty} (\mu(B_i) - \mu(C_i)) \leq \epsilon.$$

The problem is, C may not be closed. But this is easily resolved by taking k large enough so that $\mu(C) - \mu(\cup_{i=1}^{k} C_i) \leq \epsilon$. Then $C' := \cup_{i=1}^{k} C_i$ is a closed set contained in B, and $\mu(B) - \mu(C') \leq 2\epsilon$. This proves that $\mathcal{C}$ is a σ-algebra.

Finally, to prove the third identity, take any $B \in \mathcal{B}(\mathbb{R}^n)$, any $\epsilon > 0$, and a closed set $C \subseteq B$ such that $\mu(B) - \mu(C) \leq \epsilon$. Let B_r denote the closed ball of radius r centered at the origin. Then $\mu(C \cap B_r) \to \mu(C)$ as $r \to \infty$, and $C \cap B_r$ is a compact subset of C. Thus, it is possible to find a compact set $K \subseteq C$ such that $\mu(B) - \mu(K) \leq 2\epsilon$. □

PROOF OF THEOREM 3.4.1. We will use Carathéodory's extension theorem (Theorem 1.5.1) to prove both the existence and the uniqueness of μ. Let $\mathcal{A}$ be the algebra of cylinder sets of $(\mathbb{R}^d)^{\mathbb{N}}$, that is, sets of the form

$$\{\omega \in (\mathbb{R}^d)^{\mathbb{N}} : (\omega_0, \ldots, \omega_{k-1}) \in B\}$$

for some k and some $B \in \mathcal{B}((\mathbb{R}^d)^k)$. Note that the consistency of the family $\{\mu_n\}_{n\geq 1}$ automatically implies the existence of a finitely additive probability measure μ on $\mathcal{A}$ defined in the the natural way. To apply Carathédory's theorem, we only have to show that μ is countably additive on $\mathcal{A}$.

Let $A_1, A_2, \ldots$ be a disjoint sequence of elements of $\mathcal{A}$ such that their union, A, is also in $\mathcal{A}$. We need to show that $\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$. Note that for any n, the finite additivity of μ implies that

$$\mu(A) = \mu(B_n) + \sum_{i=1}^{n} \mu(A_i),$$

where $B_n := A \setminus \cup_{i=1}^{n} A_i$. Thus, we only need to show that $\mu(B_n) \to 0$ as $n \to \infty$. We will prove this by contradiction. Suppose that $\mu(B_n) \not\to 0$. Then, since this is a decreasing sequence, it must converge to a positive real number ϵ. Under this assumption, we will show that $\cap_{n\geq 1} B_n \neq \emptyset$, which will give a contradiction, because we know that the intersection is empty.

Since the B_n's are cylinder sets, there exist integers $m_1 \leq m_2 \leq \cdots$, and Borel sets $D_1 \subseteq (\mathbb{R}^d)^{m_1}, D_2 \subseteq (\mathbb{R}^d)^{m_2}, \ldots$ such that for each n,

$$B_n = D_n \times \mathbb{R}^d \times \mathbb{R}^d \times \cdots .$$

Moreover, since $B_n \supseteq B_{n+1}$ for each n, we must have that

$$D_{n+1} \subseteq D_n \times (\mathbb{R}^d)^{m_{n+1}-m_n} \tag{3.4.1}$$

for each n.

Now recall the uniform positive lower bound ϵ on $\mu(B_n)$. By Lemma 3.4.2, we can find a compact set $K_n \subseteq D_n$ for each n, such that

$$\mu_{m_n}(D_n \setminus K_n) \leq 2^{-n-1}\epsilon.$$

Define

$$\begin{aligned} K_n' := (K_1 \times (\mathbb{R}^d)^{m_n - m_1}) \cap (K_2 \times (\mathbb{R}^d)^{m_n - m_2}) \cap \cdots \\ \cdots \cap (K_{n-1} \times (\mathbb{R}^d)^{m_n - m_{n-1}}) \cap K_n. \end{aligned}$$

Then K_n', being a closed subset of K_n, is compact. Moreover,

$$\begin{aligned} \mu_{m_n}(D_n \setminus K_n') &\leq \sum_{i=1}^{n} \mu_{m_n}(D_n \setminus (K_i \times (\mathbb{R}^d)^{m_n - m_i})) \\ &\leq \sum_{i=1}^{n} \mu_{m_n}((D_i \times (\mathbb{R}^d)^{m_n - m_i}) \setminus (K_i \times (\mathbb{R}^d)^{m_n - m_i})) \\ &= \sum_{i=1}^{n} \mu_{m_i}(D_i \setminus K_i) \leq \sum_{i=1}^{n} 2^{-i-1}\epsilon \leq \frac{\epsilon}{2}. \end{aligned}$$

Since $\mu_{m_n}(D_n) = \mu(B_n) \geq \epsilon$, this shows that $\mu_{m_n}(K_n') \geq \epsilon/2$ for each n. In particular, each K_n' is nonempty. Now, from the definition of K_n', it is easy to see that $K_{n+1}' \subseteq K_n' \times (\mathbb{R}^d)^{m_{n+1}-m_n}$ for each n. Thus, if we choose $(x_1^n, \ldots, x_{m_n}^n) \in K_n'$, then for each $i < n$, $(x_1^n, \ldots, x_{m_i}^n) \in K_i'$. In particular, $(x_1^n, \ldots, x_{m_1}^n) \in K_1'$ for each n, and so, by

the compactness, we can choose a subsequence of $\{(x_1^n, \ldots, x_{m_1}^n)\}_{n\geq 1}$ converging to some $(x_1, \ldots, x_{m_1}) \in K_1'$. Passing to a further subsequence, we can have $(x_1^n, \ldots, x_{m_2}^n)$ converging to some $(x_1, \ldots, x_{m_2}) \in K_2'$, and so on. Then, by the standard diagonal argument, we can find a subsequence along which $(x_1^n, \ldots, x_{m_i}^n)$ converges to some $(x_1, \ldots, x_{m_i}) \in K_i'$ for each i. Consider the vector $x = (x_1, x_2, \ldots) \in (\mathbb{R}^d)^{\mathbb{N}}$. We claim that $x \in \cap_{n\geq 1} B_n$. Indeed, this is true, because for each n,

$$\begin{aligned} x &= (x_1, \ldots, x_{m_n}, x_{m_n+1}, \ldots) \in K_n' \times \mathbb{R}^d \times \mathbb{R}^d \times \cdots \\ &\subseteq D_n \times \mathbb{R}^d \times \mathbb{R}^d \times \cdots = B_n. \end{aligned}$$

This completes the proof of the theorem. □

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,

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

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

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

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 $\lambda \ge 0$ 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 $\{f_n\}_{n\ge1}$ is a Cauchy sequence in the L^p norm for some $1 \le p \le \infty$, then there is function $f \in L^p(\Omega, \mathcal{F}, \mu)$, and a subsequence $\{f_{n_k}\}_{k\ge1}$, 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 $\{f_{n_k}\}_{k\ge1}$ such that $\|f_{n_k} - f_{n_{k+1}}\|_{L^p} \le 2^{-k}$ for every k. Define the event

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

Then by Markov's inequality,

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

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 $|f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| \le 2^{-k/2}$ for all sufficiently large k. From this, it follows that $\{f_{n_k}(\omega)\}_{k\ge1}$ 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,

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

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 $\{f_{n_k}\}_{k\ge 1}$ as before. Then for each k, $|f_{n_k} - f_{n_{k+1}}| \le 2^{-k}$ a.e. Therefore, if we define

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

then $\mu(E^c) = 0$. For any $\omega \in E$, $\{f_{n_k}(\omega)\}_{k\ge 1}$ 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 $\{f_n\}_{n\ge 1}$ that is Cauchy in L^p. Then note that by Lemma 4.4.5, there is a subsequence $\{f_{n_k}\}_{k\ge 1}$ 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

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

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 2.3.6). 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

$$(f, g) := \int f g d\mu. \tag{4.4.1}$$

It is easy to verify that this is indeed an inner product on $L^2(\Omega, \mathcal{F}, \mu)$, 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 generated by the random variables $X_1, \ldots, X_n$.

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 1.2.6.)

Exercise 5.1.4. Let $\{X_i\}_{i \in I}$ be a finite or countable collection of real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Show that any $A \in \sigma(\{X_i\}_{i \in I})$ can be expressed as $X^{-1}(B)$ for some measurable set $B \subseteq \mathbb{R}^I$, where $X : \Omega \to \mathbb{R}^I$ is the map $X(\omega) = (X_i(\omega))_{i \in I}$.

EXERCISE 5.1.5. 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 $X = f(X_1, \ldots, X_n)$. Hint: Start with indicator random variables.

EXERCISE 5.1.6. Extend the previous exercise to σ-algebras generated by countably 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*

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

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 2.3.5, 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 analogues of normal distributions. The discrete analogues 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.

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 $X \sim Poi(\lambda)$.

CHAPTER 6

Expectation, variance, and other functionals

This chapter introduces a number of functionals associated with random variables — or, more accurately, with *laws* of random variables. The list includes expectation, variance, covariance, moments, moment generating function, and characteristic function.

6.1. 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 6.1.1. 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 6.1.2. Let S be a measurable space and let X be an S-valued random variable (meaning that X is a measurable map from some probability space into S). Then μ_X can be defined as usual, that is, $\mu_X(B) := \mathbb{P}(X \in B)$. Let $g : S \to \mathbb{R}$ be a measurable map. Generalize the previous exercise to this setting.

Exercise 6.1.3. 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 6.1.4. 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 6.1.5. Compute the means 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 6.1.6. 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

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

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 6.1.7. 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 6.1.8. 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 6.1.7 is the following. It is proved easily by the monotone convergence theorem.

EXERCISE 6.1.9. 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.

6.2. 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 6.2.1. 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 6.2.2 (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 6.2.3 (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 6.2.4. 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 6.2.5. *For any $X_1, \ldots, X_n$ defined on the same space,*

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

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 $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. □

6.3. 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 6.3.1. 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

$$\sum_{k=0}^{\infty} \frac{|t|^k}{k!} \mathbb{E}|X|^k < \infty, \tag{6.3.1}$$

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

EXERCISE 6.3.2. Carry out the above verification.

EXERCISE 6.3.3. 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 6.3.4. *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 6.3.5. 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 $e^{-t^2/2\sigma^2}$.

The above bound is actually not optimal. The following exercise gives the correct asymptotic behavior for the normal tail.

EXERCISE 6.3.6. If $X \sim N(0, 1)$, show that for any $t > 0$,

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{e^{-t^2/2}}{\sqrt{2\pi}} \leq \mathbb{P}(X \geq t) \leq \frac{e^{-t^2/2}}{t\sqrt{2\pi}}.$$

(Hint: For the upper bound, use the inequality $\int_t^\infty e^{-x^2/2} dx \leq \int_t^\infty (x/t) e^{-x^2/2} dx$. For the lower bound, use $\int_t^\infty e^{-x^2/2} dx = \int_t^\infty x^{-1}(x e^{-x^2/2}) dx$, use integration by parts, and finally use a similar trick as above to bound the second term coming out of the integration by parts.)

EXERCISE 6.3.7. If X is a nonnegative random variable, prove that for any $p > 0$,

$$\mathbb{E}(X^p) = \int_0^\infty p t^{p-1} \mathbb{P}(X \geq t) dt.$$

6.4. 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 6.4.1. *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 6.4.2. Show that the characteristic function can be a written as a power series in t if (6.3.1) holds.

EXERCISE 6.4.3. 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 6.4.4. *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 6.1.6), 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 6.4.5. Compute the characteristic functions of the Bernoulli, binomial, geometric, Poisson, uniform, exponential and Gamma distributions.

6.5. 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 6.5.1. *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, 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 7

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.

7.1. Definition

DEFINITION 7.1.1. 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 7.1.2. 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 7.1.3. 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 7.1.4. 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 7.1.5. 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 7.1.6. 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 7.1.7. 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 7.1.8. 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 7.1.9. 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 7.1.10. 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 Exercise 7.1.9.)

EXERCISE 7.1.11. Let $\{X_i\}_{i \in I}$ and $\{Y_j\}_{j \in J}$ be two collections of random variables defined on the same probability space. Suppose that for any finite $F \subseteq I$ and $G \subseteq J$, the collections $\{X_i\}_{i \in F}$ and $\{Y_j\}_{j \in G}$ are independent. Then prove that $\{X_i\}_{i \in I}$ and $\{Y_j\}_{j \in J}$ are independent. (Hint: Use Exercise 7.1.9.)

EXERCISE 7.1.12. 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 $T \sim Geo(p)$.

7.2. 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 7.2.1. *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 2.3.6. 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 $\mathbb{E}(X_nY_n) = \mathbb{E}(X_n)\mathbb{E}(Y_n)$, 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 $X_nY_n \uparrow XY$. Therefore by the monotone convergence theorem,

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

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 7.2.2. If $X_1, X_2, \ldots, X_n$ are independent integrable random variables, show that the product $X_1X_2\cdots X_n$ is also integrable and

$$\mathbb{E}(X_1X_2\cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\cdots\mathbb{E}(X_n).$$

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 7.2.3. 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 7.2.4. 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 7.2.5. Give an example of three random variables X_1, X_2, X_3 that are pairwise independent but not independent.

7.3. 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 7.3.1 (The second Borel–Cantelli lemma). *If $\{A_n\}_{n\geq 1}$ is a sequence of independent events 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 7.3.2. 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 7.3.3. 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 $X_1, X_2, \ldots$.

7.4. 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 7.4.1 (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 7.1.10 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 7.1.9, 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 7.4.2. 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 7.4.3. 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 7.4.4. 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.

7.5. 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 7.5.1. *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 5.1.5, 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

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

On the other hand,

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

since the derivative of the map $x \mapsto x(1 - x)$ is bounded by 1 in $[0, 1]$. Combining, we get $\mathbb{P}(A)(1 - \mathbb{P}(A)) < 2\epsilon$. 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 $f(\omega) \le t$. 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 7.5.1.

COROLLARY 7.5.2 (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 7.5.1 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 7.5.2.

EXERCISE 7.5.3. 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 7.5.1 and the above discussion, prove that N equals a constant in $\{0, 1, 2, \ldots\} \cup \{\infty\}$ almost surely.

Another consequence of Theorem 7.5.1 is the following result.

COROLLARY 7.5.4 (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 7.5.1. □

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

EXERCISE 7.5.5. 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.

7.6. 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 7.6.1. Prove that the c.d.f. of a random vector uniquely characterizes its law (that is, a multivariate analogue of Proposition 5.3.1).

EXERCISE 7.6.2. 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 7.6.3. If $X_1, \ldots, X_n$ are independent random variables, show that the cumulative distribution function and the characteristic function of the random vector $(X_1, \ldots, X_n)$ can be written as products of one-dimensional distribution functions and characteristic functions.

EXERCISE 7.6.4. 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.

EXERCISE 7.6.5. Let X be an $\mathbb{R}^n$-valued random vector and $U \subseteq \mathbb{R}^n$ be an open set such that $\mathbb{P}(X \in U) = 1$. Suppose that the c.d.f. F of X is n times differentiable in U. Let

$$f(x_1, \ldots, x_n) := \frac{\partial^n F}{\partial x_1 \cdots \partial x_n}$$

for $(x_1, \ldots, x_n) \in U$, and let $f \equiv 0$ outside U. Prove that f is the p.d.f. of X. (Hint: First show that for any rectangle $R \subseteq U$, $\mathbb{P}(X \in R)$ equals the integral of f over R. Then extend this to any Borel subset of U using the π-λ theorem.)

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 7.6.6. 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 7.6.7. 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 7.6.8. 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, $AX + a \sim N_m(a + A\mu, A\Sigma A^T)$.

7.7. 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 7.6.2 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 7.7.1. *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 7.7.2. 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 7.7.3. 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 7.7.4. 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 7.7.5. 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 7.7.6. 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 7.7.7. 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 8

Convergence of random variables

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

8.1. Four notions of convergence

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

DEFINITION 8.1.1. 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 8.1.2. 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 8.1.3. 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 8.1.4. 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.

8.2. Interrelations between the four notions

The four notions of convergence defined above are inter-related in interesting ways.

PROPOSITION 8.2.1. *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 8.2.2. Give a counterexample to show that convergence in probability does not imply almost sure convergence.

Although convergence in probability does not imply almost sure convergence, it does imply that there is a subsequence that converges almost surely.

PROPOSITION 8.2.3. *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 8.2.1, $X_{n_k} \to Y$ 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 8.2.4. *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

$$\begin{aligned} F_{X_n}(t) &= \mathbb{P}(X_n \le t) \\ &\le \mathbb{P}(X \le t + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon), \end{aligned}$$

which proves that $\limsup F_{X_n}(t) \le F_X(t + \epsilon)$. Since this is true for any $\epsilon > 0$ and F_X is right continuous, this gives $\limsup F_{X_n}(t) \le F_X(t)$. A similar argument gives $\liminf F_{X_n}(t) \ge F_X(t)$. □

EXERCISE 8.2.5. 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 8.2.6. 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 8.2.7 (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 8.2.8. *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 8.2.9. *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 8.2.10. 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.

8.3. Uniform integrability

Under a certain condition known as uniform integrability, almost sure convergence implies L^1 convergence. We say that a sequence of random variables $\{X_n\}_{n\geq 1}$ is uniformly integrable if for any $\epsilon > 0$, there is some $K > 0$ such that for all n,

$$\mathbb{E}(|X_n|; |X_n| > K) \leq \epsilon.$$

PROPOSITION 8.3.1. *If a uniformly integrable sequence of random variables $\{X_n\}_{n\geq 1}$ converges almost surely to a limit random variable X as $n \to \infty$, then X is integrable and $X_n \to X$ in L^1.*

PROOF. Take any ϵ, and find K such that $\mathbb{E}(|X_n|; |X_n| > K) \leq \epsilon$ for all n. This implies, in particular, that

$$\mathbb{E}|X_n| \leq \mathbb{E}(|X_n|; |X_n| > K) + \mathbb{E}(|X_n|; |X_n| \leq K) \leq \epsilon + K.$$

Therefore by Fatou's lemma, $\mathbb{E}|X| \leq \epsilon + K < \infty$. So, by the dominated convergence theorem, $\lim_{L\to\infty} \mathbb{E}(|X|; |X| > L) = 0$. This implies that by increasing K if necessary, we can ensure that $\mathbb{E}(|X|; |X| > K)$ is also bounded by ϵ. Define a function

$$\phi(x) := \begin{cases} x & \text{if } -K \leq x \leq K, \\ K & \text{if } x > K, \\ -K & \text{if } x < -K. \end{cases}$$

Note that ϕ is bounded and continuous. Therefore by the dominated convergence theorem, $\mathbb{E}|\phi(X_n) - \phi(X)| \to 0$. But

$$\begin{aligned}\mathbb{E}|X_n - X| &\leq \mathbb{E}|X_n - \phi(X_n)| + \mathbb{E}|\phi(X_n) - \phi(X)| + \mathbb{E}|\phi(X) - X| \\ &\leq \mathbb{E}(|X_n|; |X_n| > K) + \mathbb{E}|\phi(X_n) - \phi(X)| + \mathbb{E}(|X|; |X| > K) \\ &\leq 2\epsilon + \mathbb{E}|\phi(X_n) - \phi(X)|.\end{aligned}$$

Thus, $\limsup_{n\to\infty} \mathbb{E}|X_n - X| \leq 2\epsilon$. Since this holds for any ϵ, we conclude that $X_n \to X$ in L^1. □

The following proposition gives a useful criterion for checking uniform integrability.

PROPOSITION 8.3.2. *If* $\sup_{n\geq 1}\mathbb{E}|X_n|^p$ *is finite for some* $p > 1$*, then the sequence* $\{X_n\}_{n\geq 1}$ *is uniformly integrable.*

PROOF. Note that for any K,

$$\mathbb{E}(|X_n|; |X_n| > K) \leq K^{-(p-1)}\mathbb{E}|X_n|^p.$$

Since $\mathbb{E}|X_n|^p$ is uniformly bounded, the above bound can be made uniformly small by choosing K large enough. □

Uniform integrability has the following equivalent formulation, which is sometimes useful.

PROPOSITION 8.3.3. *A sequence* $\{X_n\}_{n\geq 1}$ *is uniformly integrable if and only if* $\sup_{n\geq 1}\mathbb{E}|X_n| < \infty$ *and for for all* $\epsilon > 0$*, there is some* $\delta > 0$ *such that for any event* A *with* $\mathbb{P}(A) < \delta$*, we have* $\mathbb{E}(|X_n|; A) < \epsilon$ *for all* n*.*

PROOF. Suppose that $\{X_n\}_{n\geq 1}$ is uniformly integrable. First, choose $a > 0$ such that

$$\sup_{n\geq 1}\mathbb{E}(|X_n|; |X_n| > a) \leq 1.$$

Then for any n,

$$\mathbb{E}|X_n| = \mathbb{E}(|X_n|; |X_n| \leq a) + \mathbb{E}(|X_n|; |X_n| > a) \leq a + 1,$$

which shows that $\sup_{n\geq 1}\mathbb{E}|X_n| < \infty$. Next, for any $a > 0$, any event A, and any n,

$$\begin{aligned}\mathbb{E}(|X_n|; A) &= \mathbb{E}(|X_n|; A \cap \{|X_n| \leq a\}) + \mathbb{E}(|X_n|; A \cap \{|X_n| > a\}) \\ &\leq a\mathbb{P}(A) + \mathbb{E}(|X_n|; |X_n| > a).\end{aligned}$$

By uniform integrability, the right side can be made uniformly small by choosing a large enough and $\mathbb{P}(A)$ small enough.

Conversely, suppose that the alternative criterion holds. Then for any $a > 0$,

$$a\mathbb{P}(|X_n| > a) \leq \mathbb{E}(|X_n|; |X_n| > a) \leq \sup_{n\geq 1}\mathbb{E}|X_n| < \infty.$$

Thus, $\mathbb{P}(|X_n| > a)$ can be made uniformly small by choosing a large enough. Thus, $\mathbb{E}(|X_n|; |X_n| > a)$ can be made uniformly small by choosing a large enough, which proves uniform integrability. □

An interesting corollary of the above result is the following. We will have occasion to use it later.

COROLLARY 8.3.4. *For any integrable random variable* X*, and for any* $\epsilon > 0$*, there is some* $\delta > 0$ *such that* $\mathbb{E}(|X|; A) < \epsilon$ *whenever* $\mathbb{P}(A) < \epsilon$*.*

PROOF. By the dominated convergence theorem, $\mathbb{E}(|X|; |X| > K) \to 0$ as $K \to \infty$. Thus, the sequence $X, X, X, \ldots$ is uniformly integrable. Now apply Proposition 8.3.3. □

The following proposition is also sometimes useful.

PROPOSITION 8.3.5. *If a sequence of random variables converges in L^1, then it is uniformly integrable.*

PROOF. Suppose that $X_n \to X$ in L^1. Take any $\epsilon > 0$ and $M > 0$. By Corollary 8.3.4, there is some $\delta > 0$ such that if $\mathbb{P}(A) < \delta$, then $\mathbb{E}(|X|; A) \leq \epsilon/4$. Now, note that

$$\begin{aligned}
&\mathbb{E}(|X_n|; |X_n| > M) - \mathbb{E}(|X|; |X| > M/2) \\
&\leq \mathbb{E}(|X_n - X|; |X_n| > M) + \mathbb{E}(|X|(1_{\{|X_n|>M\}} - 1_{\{|X|>M/2\}})) \\
&\leq \mathbb{E}|X_n - X| + \mathbb{E}(|X|; |X_n - X| > M/2).
\end{aligned}$$

Since $X_n \to X$ in L^1, we can find n_0 large enough so that if $n \geq n_0$, then $\mathbb{E}|X_n - X| \leq \epsilon/4$ and $\mathbb{P}(|X_n - X| > M/2) < \delta$ (the second assertion follows by Proposition 8.2.8). Thus, for $n \geq n_0$,

$$\mathbb{E}(|X_n|; |X_n| > M) \leq \mathbb{E}(|X|; |X| > M/2) + \epsilon/2.$$

Now choose M so large that $\mathbb{E}(|X|; |X| > M/2) < \epsilon/2$, and also $\mathbb{E}(|X_n|; |X_n| > M) < \epsilon$ for all $n < n_0$. Then for all $n \geq 1$, $\mathbb{E}(|X_n|; |X_n| > M) \leq \epsilon$. □

8.4. 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 8.4.1 (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 6.2.5 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 8.4.2. *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 8.4.3 (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 8.4.4 (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 8.4.5 (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.

8.5. 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 8.5.1 (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 tends to ∞.*

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 6.1.9 and the fact that the X_i's are identically distributed,

$$\sum_{i=1}^{\infty} \mathbb{P}(X_i \neq Y_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(X_i \geq i) = \sum_{i=1}^{\infty} \mathbb{P}(X_1 \geq i) \leq \mathbb{E}(X_1) < \infty.$$

Therefore by the first Borel–Cantelli lemma, $\mathbb{P}(X_i \neq Y_i \text{ i.o.}) = 0$. Thus, 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 $\operatorname{Cov}(Y_i, Y_j) = 0$ for any $i \neq j$. Thus for any n, by Theorem 8.4.1,

$$\mathbb{P}(|Z_{k_n}| > \epsilon) \leq \frac{1}{\epsilon^2 k_n^2} \sum_{i=1}^{k_n} \operatorname{Var}(Y_i).$$

Therefore

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|Z_{k_n}| > \epsilon) &\leq \sum_{n=1}^{\infty} \frac{1}{\epsilon^2 k_n^2} \sum_{i=1}^{k_n} \operatorname{Var}(Y_i) \\ &= \frac{1}{\epsilon^2} \sum_{i=1}^{\infty} \operatorname{Var}(Y_i) \sum_{n\,:\,k_n \geq i} \frac{1}{k_n^2}\end{aligned}$$

It is easy to see that there is some $\beta > 1$, depending only on α, such that $k_{n+1}/k_n \geq \beta$ for all n large enough. Therefore for large enough n,

$$\sum_{n\,:\,k_n \geq i} \frac{1}{k_n^2} \leq \frac{1}{i^2} \sum_{n=0}^{\infty} \beta^{-n} \leq \frac{C}{i^2},$$

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,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|Z_{k_n}| > \epsilon) &\leq \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\operatorname{Var}(Y_i)}{i^2} \leq \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(Y_i^2)}{i^2} \\ &= \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(X_i^2; X_i < i)}{i^2} = \frac{C}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathbb{E}(X_1^2; X_1 < i)}{i^2} \\ &\leq \frac{C}{\epsilon^2} \mathbb{E}\left(X_1^2 \sum_{i > X_1} \frac{1}{i^2} \right) \leq \frac{C'}{\epsilon^2} \mathbb{E}(X_1),\end{aligned}$$

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 $k_n < m \leq k_{n+1}$, then

$$\frac{k_n}{k_{n+1}} \frac{T_{k_n}}{k_n} = \frac{T_{k_n}}{k_{n+1}} \leq \frac{T_m}{m} \leq \frac{T_{k_{n+1}}}{k_n} = \frac{T_{k_{n+1}}}{k_{n+1}} \frac{k_{n+1}}{k_n}.$$

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

$$\frac{\mu}{\alpha} \leq \liminf_{m\to\infty} \frac{T_m}{m} \leq \limsup_{m\to\infty} \frac{T_m}{m} \leq \alpha\mu \quad \text{a.s.}$$

Since $\alpha > 1$ is arbitrary, this completes the proof. □

EXERCISE 8.5.2. Using Exercise 7.3.2, 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

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \text{ has a finite limit as } n \to \infty \right) = 0.$$

EXERCISE 8.5.3. If $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}(X_1) = \infty$, 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 8.5.4. 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 8.5.5. 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 each X_i is of the form $f(Y_i, \ldots, Y_{i+m})$ for some measurable function $f : \mathbb{R}^{m+1} \to \mathbb{R}$.

THEOREM 8.5.6. *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 7.1.5) 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 8.5.7 (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 8.5.8 (Random matrices). Let $\{X_{ij}\}_{1\le i\le j<\infty}$ 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 $\lambda_{n,1} \ge \cdots \ge \lambda_{n,n}$ be the eigenvalues of W_n, repeated by multiplicities. For any integer $k \ge 1$, 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 8.4.1. The tail bound is strong enough to prove almost sure convergence.)

EXERCISE 8.5.9. If all the random graphs in Exercise 8.4.5 are defined on the same probability space, show that the convergence is almost sure.

8.6. 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 8.6.1. A sequence of random variables $\{X_n\}_{n\ge1}$ is called a tight family if for any ϵ, there exists some K such that $\sup_n \mathbb{P}(|X_n| \ge K) \le \epsilon$.

EXERCISE 8.6.2. If $X_n \to X$ in distribution, show that $\{X_n\}_{n\ge1}$ is a tight family.

EXERCISE 8.6.3. If $\{X_n\}_{n\ge1}$ is a tight family and $\{c_n\}_{n\ge1}$ is a sequence of constants tending to 0, show that $c_nX_n \to 0$ in probability.

A partial converse of Exercise 8.6.2 is the following theorem.

THEOREM 8.6.4 (Helly's selection theorem). *If $\{X_n\}_{n\ge1}$ is a tight family, then there is a subsequence $\{X_{n_k}\}_{k\ge1}$ that converges in distribution.*

PROOF. Let F_n be the c.d.f. of X_n. By the standard diagonal argument, there is subsequence $\{n_k\}_{k\ge1}$ 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. □

8.7. An alternative characterization of weak convergence

The following result gives an important equivalent criterion for convergence in distribution.

PROPOSITION 8.7.1. *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

$$\limsup_{n\to\infty} F_{X_n}(t) \le F_X(t).$$

A similar argument shows that for any $\epsilon > 0$,

$$\liminf_{n\to\infty} F_{X_n}(t) \ge F_X(t-\epsilon).$$

Since t is a continuity point of F_X, this proves that $\liminf F_{X_n}(t) \ge F_X(t)$. 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 8.6.2 there exists K such that $\mathbb{P}(|X_n| \ge K) \le \epsilon$ for all n. Choose K so large that we also have $\mathbb{P}(|X| \ge K) \le \epsilon$. Let M be a number such that $|f(x)| \le M$ for all x.

Since f is uniformly continuous in $[-K, K]$, there is some $\delta > 0$ such that $|f(x)-f(y)| \le \epsilon$ whenever $|x-y| \le \delta$ 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 $x_1 \le x_2 \le \cdots \le x_m \in [-K, K]$ such that each x_i is a continuity point of F_X, $x_1 = -K$, $x_m = K$, and $x_{i+1} - x_i \le \delta$ for each i. Now note that

$$\begin{aligned}\mathbb{E}f(X_n) = \mathbb{E}(f(X_n); X_n > K) + \mathbb{E}(f(X_n); X_n \le -K)\\ + \sum_{i=1}^{m-1} \mathbb{E}(f(X_n); x_i < X_n \le x_{i+1}),\end{aligned}$$

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 8.7.1 is the following result.

PROPOSITION 8.7.2. *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. □

8.8. 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 8.8.1. *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 $|\phi(t)| \le 1$ 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 6.5.1,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{\mathrm{i}(y-x)t - \theta t^2} dt &= \sqrt{\frac{\pi}{\theta}} \int_{-\infty}^{\infty} e^{\mathrm{i}(2\theta)^{-1/2}(y-x)s} \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds \\ &= \sqrt{\frac{\pi}{\theta}} e^{-(y-x)^2/4\theta}. \end{aligned}$$

Therefore

$$f_\theta(x) = \int_{-\infty}^{\infty} \frac{e^{-(y-x)^2/4\theta}}{\sqrt{4\pi\theta}} d\mu(y).$$

But by Proposition 7.7.1, 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 6.2.1), 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 8.7.1. □

An immediate corollary of the above theorem is the following important fact.

COROLLARY 8.8.2. *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 8.8.1, $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$ for every bounded continuous g. Therefore by Proposition 8.7.1, X and Y have the same law. □

EXERCISE 8.8.3. Let ϕ be the characteristic function of a random variable X. If $|\phi(t)| = 1$ for all t, show that X is a degenerate random variable — that is, there is a constant c such that $\mathbb{P}(X = c) = 1$. (Hint: Start by showing that $|\phi(t)|^2$ is the characteristic function of $X - X'$, where X' has the same law as X and is independent of X. Then apply Corollary 8.8.2.)

EXERCISE 8.8.4. Let X be a non-degenerate random variable. If aX and bX have the same distribution for some positive a and b, show that $a = b$. (Hint: Use the previous exercise.)

EXERCISE 8.8.5. Let X be any random variable. If $X + a$ and $X + b$ have the same distribution for some $a, b \in \mathbb{R}$, show that $a = b$.

EXERCISE 8.8.6. Suppose that a random variable X has the same distribution as the sum of n i.i.d. copies of X, for some $n \geq 2$. Prove that $X = 0$ with probability one. (Hint: Prove that the characteristic function takes value in a finite set and must therefore be equal to 1 everywhere.)

Another important corollary of Theorem 8.8.1 is the following simplified inversion formula.

COROLLARY 8.8.7. *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 8.8.1 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 8.8.8. *Let X be an integer-valued random variable with characteristic function ϕ. Then for any $x \in \mathbb{Z}$,*

$$\mathbb{P}(X = x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} \phi(t) dt.$$

PROOF. Let μ be the law of X. Then note that by Fubini's theorem,

$$\begin{aligned}
\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} \phi(t) dt &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \int_{-\infty}^{\infty} e^{-itx} e^{ity} d\mu(y) dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} e^{it(y-x)} dt d\mu(y) \\
&= \sum_{y \in \mathbb{Z}} \mathbb{P}(X = y) \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{it(y-x)} dt \right) \\
&= \mathbb{P}(X = x),
\end{aligned}$$

where the last identity holds because $x, y \in \mathbb{Z}$ in the previous step. □

8.9. 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 8.9.1 (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 8.7.1. 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 6.4.3), 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 6.4.4 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 8.8.2, and we get a contradiction by the second assertion of Proposition 8.7.1 applied to this X and Y. □

EXERCISE 8.9.2. 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. (Hint: It suffices to show that if $t_n \to t$, then $\phi_n(t_n) \to \phi(t)$. Deduce this from Lévy's continuity theorem.)

EXERCISE 8.9.3. 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.)

8.10. 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 8.10.1 (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 8.10.2. *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 8.10.3. *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 8.10.2

$$\left| e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{x^2}{2} \right| \leq \frac{|x|^3}{6}.$$

On the other hand, by Lemma 8.10.2 we also have

$$\left|e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{x^2}{2}\right| \le |e^{\mathrm{i}x} - 1 - \mathrm{i}x| + \frac{x^2}{2}$$
$$\le \frac{x^2}{2} + \frac{x^2}{2} = x^2.$$

The proof is completed by combining the two bounds. □

LEMMA 8.10.4. *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

$$\left|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i\right| \le \sum_{i=1}^n \left|a_1 \cdots a_{i-1} b_i \cdots b_n - a_1 \cdots a_i b_{i+1} \cdots b_n\right|$$
$$= \sum_{i=1}^n |a_1 \cdots a_{i-1}(b_i - a_i) b_{i+1} \cdots b_n|$$
$$\le \sum_{i=1}^n |a_i - b_i|,$$

where the last inequality holds because $|a_i|$ and $|b_i|$ are ≤ 1 for each i. □

We are now ready to prove Theorem 8.10.1.

PROOF OF THEOREM 8.10.1. 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 6.5.1), 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 8.10.4, 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 8.10.3,

$$n\left|\phi_{X_1}(t/\sqrt{n}) - \left(1 - \frac{t^2}{2n}\right)\right| = n\left|\mathbb{E}\left(e^{\mathrm{i}tX_1/\sqrt{n}} - 1 - \frac{\mathrm{i}tX_1}{\sqrt{n}} + \frac{t^2 X_1^2}{2n}\right)\right|$$
$$\le \mathbb{E}\min\left\{t^2 X_1^2, \frac{|t|^3 |X_1|^3}{6\sqrt{n}}\right\}.$$

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 8.10.5. Give a counterexample to show that the i.i.d. assumption in Theorem 8.10.1 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 8.10.6. Let $X_n \sim Bin(n,p)$. Prove a central limit theorem for X_n. (Hint: Use Exercise 7.7.4 and the CLT for i.i.d. random variables.)

EXERCISE 8.10.7. Let $X_n \sim Gamma(n,\lambda)$. Prove a central limit theorem for X_n. (Hint: Use Exercise 7.7.5 and the CLT for i.i.d. random variables.)

EXERCISE 8.10.8. 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 8.10.9 (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 $B_{k_n} \subseteq \{1, \ldots, n\}$. 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 8.6.3, $(R_n - R'_n)/\sqrt{n} \to 0$ 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 8.10.10. Let $X_1, X_2, \ldots$ 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 $\sum_{i=2}^n Y_i$.

8.11. 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 8.11.1 (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 $\mu_{n,i} := \mathbb{E}(X_{n,i})$, $\sigma_{n,i}^2 := \mathrm{Var}(X_{n,i})$, and*

$$s_n^2 := \sum_{i=1}^{k_n} \sigma_{n,i}^2.$$

Suppose that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{k_n} \mathbb{E}((X_{n,i} - \mu_{n,i})^2; |X_{n,i} - \mu_{n,i}| \geq \epsilon s_n) = 0. \tag{8.11.1}$$

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 (8.11.1) 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 8.11.2. *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 8.10.4. □

PROOF OF THEOREM 8.11.1. 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 (8.11.1) becomes

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}(X_{n,i}^2; |X_{n,i}| \geq \epsilon) = 0. \tag{8.11.2}$$

Note that for any $\epsilon > 0$ and any n,

$$\begin{aligned}\max_{1\leq i\leq k_n} \sigma_{n,i}^2 &= \max_{1\leq i\leq k_n} (\mathbb{E}(X_{n,i}^2; |X_{n,i}| < \epsilon) + \mathbb{E}(X_{n,i}^2; |X_{n,i}| \geq \epsilon)) \\ &\leq \epsilon^2 + \max_{1\leq i\leq k_n} \mathbb{E}(X_{n,i}^2; |X_{n,i}| \geq \epsilon).\end{aligned}$$

Therefore by (8.11.2),

$$\limsup_{n\to\infty} \max_{1\leq i\leq k_n} \sigma_{n,i}^2 \leq \epsilon^2.$$

Since ϵ is arbitrary, this shows that

$$\lim_{n\to\infty} \max_{1\leq i\leq k_n} \sigma_{n,i}^2 = 0. \tag{8.11.3}$$

Let $S_n := \sum_{i=1}^{n} X_i$. Then by Exercise 7.7.7,

$$\phi_{S_n}(t) = \prod_{i=1}^{k_n} \phi_{X_{n,i}}(t).$$

By Corollary 8.10.3,

$$\begin{aligned}\left|\phi_{X_{n,i}}(t) - 1 + \frac{t^2\sigma_{n,i}^2}{2}\right| &= \left|\mathbb{E}\left(e^{\mathrm{i}tX_{n,i}} - 1 - \mathrm{i}tX_{n,i} + \frac{t^2X_{n,i}^2}{2}\right)\right| \\ &\leq \mathbb{E}\min\left\{t^2X_{n,i}^2, \frac{|t|^3|X_{n,i}|^3}{6}\right\}.\end{aligned}$$

Now fix t. Equation (8.11.3) tells us that $t^2\sigma_{n,i}^2 \le 2$ for each i when n is sufficiently large. Thus by Lemma 8.10.4,

$$\begin{aligned}\left|\phi_{S_n}(t) - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| &\le \sum_{i=1}^{k_n}\left|\phi_{X_i}(t) - 1 + \frac{t^2\sigma_{n,i}^2}{2}\right| \\ &\le \sum_{i=1}^{k_n}\mathbb{E}\min\left\{t^2X_{n,i}^2, \frac{|t|^3|X_{n,i}|^3}{6}\right\}.\end{aligned}$$

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 (8.11.2),

$$\limsup_{n\to\infty}\left|\phi_{S_n}(t) - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{2}\right)\right| \le \frac{|t|^3\epsilon}{6}.$$

Since this holds for any $\epsilon > 0$, the lim sup on the left must be equal to zero. But by Corollary 8.11.2 and equation (8.11.3),

$$\begin{aligned}\limsup_{n\to\infty}\left|e^{-t^2/2} - \prod_{i=1}^{k_n}\left(1 - \frac{t^2\sigma_{n,i}^2}{8}\right)\right| &\le \limsup_{n\to\infty}\frac{1}{2}\sum_{i=1}^{k_n}t^4\sigma_{n,i}^4 \\ &\le \limsup_{n\to\infty}\frac{t^4\max_{1\le i\le k_n}\sigma_{n,i}^2}{8}\sum_{i=1}^{k_n}\sigma_{n,i}^2 \\ &= \limsup_{n\to\infty}\frac{t^4\max_{1\le i\le k_n}\sigma_{n,i}^2}{8} = 0.\end{aligned}$$

Thus, $\phi_{S_n}(t) \to e^{-t^2/2}$ as $n \to \infty$. By Lévy's continuity theorem and Proposition 6.5.1, 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 8.11.3 (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 $\delta > 0$,*

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}|X_i - \mu_i|^{2+\delta} = 0, \tag{8.11.4}$$

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,

$$\begin{aligned}\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}((X_i - \mu_i)^2; |X_i - \mu_i| \geq \epsilon s_n) &\leq \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left(\frac{|X_i - \mu_i|^{2+\delta}}{(\epsilon s_n)^{\delta}} \right) \\ &= \frac{1}{\epsilon^{\delta} s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}|X_i - \mu_i|^{2+\delta}.\end{aligned}$$

The Lyapunov condition (8.11.4) implies that this upper bound tends to zero as $n \to \infty$, which completes the proof. □

EXERCISE 8.11.4. Suppose that $X_n \sim Bin(n, p_n)$, where $\{p_n\}_{n=1}^{\infty}$ is a sequence of constants such that $np_n(1-p_n) \to \infty$. Prove a CLT for X_n.

EXERCISE 8.11.5. 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.

8.12. Stable laws

The central limit theorem gives the limiting distribution for the normalized sums of i.i.d. random variables with finite second moment. What if the second moment is not finite? 'Stable laws' are a class of distributions that characterize all possible limits of this sort.

DEFINITION 8.12.1. A real-valued random variable Y is said to have a 'stable law' if there are i.i.d. random variables $\{X_n\}_{n\geq 1}$ and sequences of real constants $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, with $a_n > 0$ for each n, such that

$$\frac{X_1 + \cdots + X_n}{a_n} - b_n \xrightarrow{d} Y.$$

By the central limit theorem, the standard normal distribution is a stable law. The goal of this section is to characterize the set of all stable laws. We need the following lemma.

LEMMA 8.12.2. *Suppose that $X_n \xrightarrow{d} X$, where X is non-degenerate. Let $\alpha_n > 0$ and $\beta_n \in \mathbb{R}$ be such that $\alpha_n X_n + \beta_n \xrightarrow{d} Y$ for some non-degenerate Y. Then there exist $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $\alpha_n \to \alpha$, $\beta_n \to \beta$, and $Y \stackrel{d}{=} \alpha X + \beta$.*

PROOF. Let ϕ_n be the characteristic function of X_n and ψ_n be the characteristic function of $\alpha_n X_n + \beta_n$, so that

$$\psi_n(t) = e^{\mathrm{i}t\beta_n} \phi(\alpha_n t) \tag{8.12.1}$$

for all t. Let ϕ be the characteristic function of X and ψ be the characteristic function of Y, so that ϕ_n converges pointwise to ϕ and ψ_n converges pointwise to ψ. If $\alpha_n \to 0$ through a subsequence, (8.12.1) shows that $|\phi(t)| = 1$ for all t. Similarly, if $\alpha_n \to \infty$ through a subsequence, (8.12.1) shows that $|\psi(t)| = 1$ for all t. By Exercise 8.8.3, these are contradictions to the assumption that X and Y are non-degenerate random variables. Thus, the sequence $\{\alpha_n\}_{n\geq 1}$ is bounded away from 0 and ∞.

Now let α and α' be two subsequential limits of of $\{\alpha_n\}_{n\geq 1}$. Note that these are positive and finite, by the conclusion from the previous paragraph. By (8.12.1), it follows that for all t,

$$|\phi(\alpha t)| = |\psi(t)| = |\phi(\alpha' t)|.$$

Suppose $\alpha' > \alpha$. The above equation shows that for all t, $|\phi(t)| = |\phi(at)|$, where $a = \alpha/\alpha'$. Iterating this, we get $|\phi(t)| = \lim_{n\to\infty} |\phi(a^n t)| = |\phi(0)| = 1$. Again by Exercise 8.8.3, this implies that X is a degenerate random variable, which contradicts our assumption. Thus, $\alpha := \lim_{n\to\infty} \alpha_n$, exists and is in $(0, \infty)$.

The last sentence of the previous paragraph, and the uniform convergence of ϕ_n to ϕ in bounded neighborhoods, allow us to conclude that there is some $t_0 > 0$ small enough such that for $t \in (-t_0, t_0)$, $|\phi_n(\alpha_n t)| > 1/2$ for all n. Thus, by equation (8.12.1), we get that $e^{\mathrm{i}t\beta_n}$ converges uniformly to $\psi(t)/\phi(\alpha t)$ in $(-t_0, t_0)$. This shows that $\{\beta_n\}_{n\geq 1}$ cannot have a subsequence $\{\beta_{n_k}\}_{k\geq 1}$ such that $|\beta_{n_k}| \to \infty$, because otherwise we can take $t_k = \pi/\beta_{n_k}$ and have $e^{\mathrm{i}\beta_{n_k} t_k} = -1$ for all k, whereas $t_k \to 0$ and $\psi(0)/\phi(0) = 1$.

Thus, $\{\beta_n\}_{n\geq 1}$ must be a bounded sequence. Let β and β' be limit points of this sequence. Then $e^{\mathrm{i}t\beta} = \psi(t)/\phi(\alpha t) = e^{\mathrm{i}t\beta'}$ for all $t \in (-t_0, t_0)$, which implies that $\beta = \beta'$. Thus, $\beta := \lim_{n\to\infty} \beta_n$ exists and is finite.

This completes the proofs of the first two claims of the lemma. The last claim is now obvious. □

The next result gives an intrinsic characterization of stable laws.

THEOREM 8.12.3. *A non-degenerate random variable Y has a stable law if and only if for each n, there exist $A_n > 0$ and $B_n \in \mathbb{R}$ such that*

$$Y \stackrel{d}{=} \frac{Y_1 + \cdots + Y_n}{A_n} - B_n,$$

where $Y_1, \ldots, Y_n$ are i.i.d. random variables with the same law as Y.

PROOF. If Y has the stated property, then it is obvious that Y is stable. Conversely, suppose that Y is stable. Let X_n, a_n and b_n be as in Definition 8.12.1. Let

$$Z_n := \frac{X_1 + \cdots + X_n}{a_n} - b_n.$$

Take any $k \geq 1$. For $1 \leq j \leq k$, let

$$Z_n^{(j)} := \frac{X_{n(j-1)+1} + \cdots + X_{nj}}{a_n} - b_n,$$

so that

$$Z_{nk} = \frac{a_n}{a_{nk}} \sum_{j=1}^{k} Z_n^{(j)} + \frac{k a_n}{a_{nk}} b_n - b_{nk}.$$

Now, as k remains fixed and $n \to \infty$,

$$Z_{nk} \xrightarrow{d} Y, \quad \sum_{j=1}^{k} Z_n^{(j)} \xrightarrow{d} Y_1 + \cdots + Y_n.$$

By Lemma 8.12.2, this completes the proof. □

The next result shows that the constant A_n in Theorem 8.12.3 must necessarily have a specific form.

THEOREM 8.12.4. *In the setting of Theorem 8.12.3, there exists a number $\alpha \in (0, 2]$ such that $A_n = n^{1/\alpha}$ for all n.*

We need the following lemma for the proof of Theorem 8.12.4. A random variable X is called 'symmetric' if $X \stackrel{d}{=} -X$.

LEMMA 8.12.5. *Let $X_1, \ldots, X_n$ be i.i.d. symmetric random variables with c.d.f. F. Then for any $t > 0$,*

$$\mathbb{P}(|X_1 + \cdots + X_n| > t) \geq \frac{1}{2}(1 - e^{-2n(1-F(t))}).$$

PROOF. Let $Y_i := |X_i|$ and $s_i := \text{sign}(X_i)$, where the sign is taken to be 1 or -1 with equal probability if $X_i = 0$. It is a simple consequence of the symmetry of X_i that s_i and Y_i are independent. Let I be the minimum i such that $|X_i| = \max_{1 \leq j \leq n} |X_j|$. Let $s := (s_1, \ldots, s_n)$ and $Y := (Y_1, \ldots, Y_n)$. Let $s' = (s'_1, \ldots, s'_n)$ be the vector defined as $s'_i := -s_i$ if $i \neq I$ and $s'_i = s_i$ if $i = I$. We claim that $(s, Y) \stackrel{d}{=} (s', Y)$. To see this, note that by the independence of s and $X := (X_1, \ldots, X_n)$, we have that for any $a \in \{-1, 1\}^n$ and $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\begin{aligned}
\mathbb{P}(s' = a, Y \in A) &= \sum_{i=1}^{n} \mathbb{P}(s' = a, Y \in A, I = i) \\
&= \sum_{i=1}^{n} \mathbb{P}(s_j = -a_j \text{ for } j \neq i, s_i = a_i, Y \in A, I = i) \\
&= \sum_{i=1}^{n} \mathbb{P}(s_j = -a_j \text{ for } j \neq i, s_i = a_i) \mathbb{P}(Y \in A, I = i) \\
&= 2^{-n} \sum_{i=1}^{n} \mathbb{P}(Y \in A, I = i) = 2^{-n} \mathbb{P}(Y \in A),
\end{aligned}$$

and this is equal to $\mathbb{P}(s = a, Y \in A)$ by a similar argument.

Now let $M := X_I$ and $T := \sum_{j \neq I} X_j$. Then note that $M = s_I Y_I$ and $T = \sum_{j \neq I} s_j Y_j$. This implies that (M, T) has the same law as $(M, -T)$, since the function that maps (s, Y) to (M, T), maps (s', Y) to $(M, -T)$. Thus, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(M > t) &\leq \mathbb{P}(M > t, T \geq 0) + \mathbb{P}(M > t, T \leq 0) \\
&= \mathbb{P}(M > t, T \geq 0) + \mathbb{P}(M > t, -T \geq 0) \\
&= 2\mathbb{P}(M > t, T \geq 0).
\end{aligned}$$

Thus, by symmetry,

$$\begin{aligned}\mathbb{P}(|X_1+\cdots+X_n|>t) &= 2\mathbb{P}(X_1+\cdots+X_n>t)\\ &= 2\mathbb{P}(M+T>t)\\ &\geq 2\mathbb{P}(M>t, T\geq 0)\geq \mathbb{P}(M>t).\end{aligned}$$

To complete the proof, note that again by symmetry,

$$\mathbb{P}(M>t)=\frac{1}{2}\mathbb{P}(|M|>t)=\mathbb{P}(\max_{1\leq i\leq n}|X_i|>t).$$

If t is a point of continuity of F, then

$$\begin{aligned}\mathbb{P}(\max_{1\leq i\leq n}|X_i|>t) &= 1-\mathbb{P}(\max_{1\leq i\leq n}|X_i|\leq t)\\ &= 1-(\mathbb{P}(|X_1|\leq t))^n = 1-(1-2(1-F(t)))^n.\end{aligned}$$

By the inequality $1-x\leq e^{-x}$, this proves the lemma when t is a point of continuity of F. If t is not a point of continuity, then the result can be obtained by taking a sequence of continuity points decreasing to t. □

Proof of Theorem 8.12.4. Let $Z := Y_1 - Y_2$, and let $Z_1, Z_2, \ldots$ be i.i.d. random variables with the same law as Z. Then note that for any m and n,

$$\begin{aligned}A_{m+n}Z &\stackrel{d}{=} \sum_{i=1}^{m+n} Z_i = \sum_{i=1}^{m} Z_i + \sum_{i=m+1}^{n} Z_i\\ &\stackrel{d}{=} A_m Z_1 + A_n Z_2. \end{aligned} \tag{8.12.2}$$

Thus, for any $t>0$,

$$\begin{aligned}\mathbb{P}(Z>t) &= \mathbb{P}(A_{m+n}Z > A_{m+n}t)\\ &\geq \mathbb{P}(A_mZ_1 > A_{m+n}t \text{ and } Z_2\geq 0)\geq \frac{1}{2}\mathbb{P}(A_mZ_1>A_{m+n}t),\end{aligned}$$

where the last inequality holds because Z_2 is symmetrically distributed around zero. Since $Z_1 \stackrel{d}{=} Z$ and Z is nondegenerate and symmetric, this proves that

$$B := \sup_{1\leq n\leq k}\frac{A_n}{A_k} < \infty. \tag{8.12.3}$$

Now, by (8.12.2) and induction, we have that for any r and k,

$$A_{rk}Z \stackrel{d}{=} A_rZ_1 + A_rZ_2 + \cdots + A_rZ_k \stackrel{d}{=} A_rA_kZ.$$

By Exercise 8.8.4, this proves that

$$A_{rk} = A_rA_k. \tag{8.12.4}$$

In particular, $A_{r^k} = A_r^k$. This, together with (8.12.3), implies that $A_r\geq 1$ for all r. By Exercise 8.8.6, $A_r\neq 1$ for all r. Thus, $A_r>1$ for all r.

Now take any $r,s\geq 2$. Since $A_r, A_s>1$, there are positive real numbers α and β such that $A_r = r^{1/\alpha}$ and $A_s = s^{1/\beta}$. We claim that $\alpha=\beta$. To prove this claim, take any k and let $m=s^k$. For k large enough, there is some j such that $n=r^j$ satisfies $n\leq m\leq rn$.

Then by (8.12.4),

$$A_m = A_s^k = A^{k/\beta} = m^{1/\beta} \leq (rn)^{1/\beta} = r^{1/\beta} A_n^{\alpha/\beta}.$$

But by (8.12.3) and the fact that $m \geq n$, we have $A_m \geq B^{-1} A_n$. Thus, $B^{-1} A_n \leq r^{1/\beta} A_n^{\alpha/\beta}$. Since $A_r > 1$, it is impossible that this holds as $k \to \infty$ unless $\alpha \geq \beta$. Similarly, $\beta \geq \alpha$. This proves the claim. Thus, there is some $\alpha > 0$ such that $A_n = n^{1/\alpha}$ for all n.

Lastly, we prove that $\alpha \leq 2$. Suppose not. Let $S_n := Z_1 + \cdots + Z_n$, so that $S_n \stackrel{d}{=} A_n Z$. Thus, there exists $t > 0$ such that for all n,

$$\mathbb{P}(|S_n| > tA_n) = \mathbb{P}(|Z| > t) < \frac{1}{4}.$$

By Lemma 8.12.5, this shows that $B := \sup_{n \geq 1} n(1 - F(tA_n)) < \infty$, where F is the c.d.f. of Z. Take any $x \geq t$. Then there is some n such that $n^{1/\alpha} \leq x/t \leq (n+1)^{1/\alpha}$. For any such n,

$$1 - F(x) \leq 1 - F(tn^{1/\alpha}) = 1 - F(tA_n) \leq \frac{B}{n} \leq \frac{2B}{n+1} \leq \frac{2Bt^\alpha}{x^\alpha}.$$

Thus,

$$\mathbb{E}(Z^2) = \int_0^\infty 2x\mathbb{P}(|Z| \geq x)dx \leq t^2 + \int_t^\infty 4Bt^\alpha x^{1-\alpha} dx,$$

which is finite since $\alpha > 2$. But then, by the central limit theorem, $(Z_1 + \cdots + Z_n)/n^{1/\alpha}$ cannot converge in distribution as $n \to \infty$. This gives a contradiction which proves that $\alpha \leq 2$. □

Exercise 8.12.6. Show that if X is a symmetric stable random variable, it must have characteristic function $\phi(t) = e^{-\theta|t|^\alpha}$ for some $\theta > 0$ and $\alpha \in (0, 2]$.

CHAPTER 9

Conditional expectation and martingales

In this chapter we will learn about the measure theoretic definition of conditional expectation, about martingales and their properties, and some applications.

9.1. Conditional expectation

Conditional probability and conditional expectations have straightforward definitions when we are conditioning on events of nonzero probability, but problems arise when we try to condition on events of probability zero — for example, when trying to compute the distribution of Y given $X = x$, where X is a continuous random variable. The following exercise gives an example.

EXERCISE 9.1.1. Let (X, Y) be a point distributed uniformly in the unit disk in $\mathbb{R}^2$. What is the distribution of Y given $X = 0$? Show that the answer depends on how we calculate $\mathbb{P}(Y \in A | X = 0)$. One way is to take the limit, as $\epsilon \to 0$, of $\mathbb{P}(Y \in A | X \in (-\epsilon, \epsilon))$. Another way is to interpret the event $\{X = 0\}$ as $\{\Theta \in \{\pi/2, 3\pi/2\}\}$, where (R, Θ) is the representation of (X, Y) in polar coordinates, with Θ ranging in $[0, 2\pi)$. In this second approach, $\mathbb{P}(Y \in A | X = 0)$ is the limit, as $\epsilon \to 0$, of $\mathbb{P}(Y \in A | \Theta \in (\pi/2 - \epsilon, \pi/2 + \epsilon) \cup (3\pi/2 - \epsilon, 3\pi/2 + \epsilon))$. Show that the two answers are different. (This is an instance of the 'Borel–Kolmogorov paradox', which uses a different example.)

The measure-theoretic approach to probability gives us a way of avoiding such problematic issues. The key idea is to define the conditional expectation of a random variable, not by conditioning on an event, but by conditioning on a σ-algebra. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a real-valued random variable defined on this space. Suppose that X is integrable (that is, $\mathbb{E}|X|$ is finite). Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}(X|\mathcal{G})$, is a $\mathcal{G}$-measurable integrable random variable Y such that for any $B \in \mathcal{G}$, $\mathbb{E}(X; B) = \mathbb{E}(Y; B)$. It is not clear from the definition whether such a random variable exists, but if it does, then almost sure uniqueness is not difficult to prove.

LEMMA 9.1.2. *If Y and Z are two random variables that qualify as $\mathbb{E}(X|\mathcal{G})$, then $Y = Z$ a.s.*

PROOF. Let $B := \{\omega : Y(\omega) > Z(\omega)\}$. Note that $B \in \mathcal{G}$. Since $\mathbb{E}(X; B) = \mathbb{E}(Y; B) = \mathbb{E}(Z; B)$, we get $\mathbb{E}((Y - Z); B) = 0$. But $(Y - Z)1_B$ is a nonnegative random variable. Thus, $(Y - Z)1_B = 0$ a.s., which implies that $Y \leq Z$ a.s. Similarly, $Y \geq Z$ a.s. □

The definition of conditional expectation implies that we cannot have uniqueness holding everywhere; almost everywhere is the best that we can hope for. However, we

will generally treat conditional expectation as if it is a uniquely defined random variable, because differences on null sets usually do not matter.

We will now show that conditional expectation always exists. To do this, we will follow the usual route — that is, first show it for simple random variables, then use monotonicity to show it for all nonnegative random variables, and finally use positive and negative parts to cover all integrable random variables.

LEMMA 9.1.3. *If X is a simple random variable, then $\mathbb{E}(X|\mathcal{G})$ exists.*

PROOF. Since X is simple, it is square-integrable. Let $\mathcal{H}$ denote the Hilbert space $L^2(\Omega, \mathcal{G}, \mathbb{P})$. Let $\{Y_n\}_{n\geq 1}$ be a sequence of elements of $\mathcal{H}$ such that

$$\lim_{n\to\infty} \mathbb{E}(X - Y_n)^2 = s := \inf_{Y\in\mathcal{H}} \mathbb{E}(X - Y)^2. \tag{9.1.1}$$

Take any m and n. Let $Z := (Y_m + Y_n)/2$. Recall the parallelogram identity $(a-b)^2 + (a+b)^2 = 2a^2 + 2b^2$. Taking $a = (X - Y_m)/2$ and $b = (X - Y_n)/2$, we get

$$\mathbb{E}(X - Z)^2 + \frac{1}{4}\mathbb{E}(Y_m - Y_n)^2 = \frac{1}{2}\mathbb{E}(X - Y_m)^2 + \frac{1}{2}\mathbb{E}(X - Y_n)^2.$$

Since $Z \in \mathcal{H}$, $\mathbb{E}(X - Z)^2 \geq s$. Thus,

$$\frac{1}{4}\mathbb{E}(Y_m - Y_n)^2 \leq \frac{1}{2}\mathbb{E}(X - Y_m)^2 + \frac{1}{2}\mathbb{E}(X - Y_n)^2 - s.$$

By (9.1.1), this proves that $\{Y_n\}_{n\geq 1}$ is a Cauchy sequence in $\mathcal{H}$. Since $\mathcal{H}$ is a Hilbert space, this sequence has a $\mathcal{G}$-measurable L^2 limit Y. Clearly,

$$\mathbb{E}(X - Y)^2 = \lim_{n\to\infty} \mathbb{E}(X - Y_n)^2 = s.$$

Now take any $Z \in \mathcal{H}$. Then for any $\lambda \in \mathbb{R}$, $Y + \lambda Z \in \mathcal{H}$, and so by the above property of Y, $\mathbb{E}(X - Y - \lambda Z)^2$ is minimized when $\lambda = 0$. But this is just a quadratic polynomial in λ, and so the condition that it is minimized at 0 implies that the coefficient of λ, namely $\mathbb{E}((X - Y)Z)$, is zero. Taking $Z = 1_B$ for any $B \in \mathcal{G}$ proves that $Y = \mathbb{E}(X|\mathcal{G})$. □

LEMMA 9.1.4. *If X_1 and X_2 are simple random variables and $a, b \in \mathbb{R}$, then $\mathbb{E}(aX_1 + bX_2|\mathcal{G}) = a\mathbb{E}(X_1|\mathcal{G}) + b\mathbb{E}(X_2|\mathcal{G})$ a.s.*

PROOF. It is easy to see from the definition that $a\mathbb{E}(X_1|\mathcal{G}) + b\mathbb{E}(X_2|\mathcal{G})$ is a valid candidate for $\mathbb{E}(aX_1 + bX_2|\mathcal{G})$, and therefore the result follows by the uniqueness of conditional expectation. □

LEMMA 9.1.5. *If X is a nonnegative simple random variable, then $\mathbb{E}(X|\mathcal{G})$ is also a nonnegative random variable.*

PROOF. Let $Y = \mathbb{E}(X|\mathcal{G})$. Let $B := \{\omega : Y(\omega) < 0\}$. Then $B \in \mathcal{G}$, and hence $\mathbb{E}(Y; B) = \mathbb{E}(X; B)$. Since X is nonnegative, $\mathbb{E}(X; B) \geq 0$. Thus, $\mathbb{E}(Y; B) \geq 0$. But this implies that $Y \geq 0$ a.s. □

LEMMA 9.1.6. *If X is $[0, \infty]$-valued, $\mathbb{E}(X|\mathcal{G})$ exists and is $[0, \infty]$-valued. If X is integrable, then so is $\mathbb{E}(X|\mathcal{G})$, and these random variables have the same expected value. Lastly, if X_1 and X_2 are $[0, \infty]$-valued random variables and $a, b \geq 0$, then $\mathbb{E}(aX_1+bX_2|\mathcal{G}) = a\mathbb{E}(X_1|\mathcal{G}) + b\mathbb{E}(X_2|\mathcal{G})$ a.s.*

PROOF. By Proposition 2.3.6, we can find a sequence of nonnegative simple functions $\{X_n\}_{n\geq 1}$ increasing to X. By Lemma 9.1.3, $Y_n := \mathbb{E}(X_n|\mathcal{G})$ exists for every n. By Lemma 9.1.4, $\mathbb{E}(X_n - X_{n-1}|\mathcal{G}) = Y_n - Y_{n-1}$ for every n. Therefore by Lemma 9.1.5, $\{Y_n\}_{n\geq 1}$ is also an increasing sequence of nonnegative random variables. Let Y be the pointwise limit of this sequence. Then for any $B \in \mathcal{G}$, the monotone convergence theorem gives us

$$\begin{aligned}\mathbb{E}(X;B) &= \lim_{n\to\infty} \mathbb{E}(X_n;B)\\ &= \lim_{n\to\infty} \mathbb{E}(Y_n;B) = \mathbb{E}(Y;B).\end{aligned}$$

Thus, $Y = \mathbb{E}(X|\mathcal{G})$. Clearly, Y is $[0,\infty]$-valued. If $\mathbb{E}(X) < \infty$, then $\mathbb{E}(Y) = \mathbb{E}(Y;\Omega) = \mathbb{E}(X;\Omega) = \mathbb{E}(X)$ is also finite. Linearity is proved exactly as in Lemma 9.1.4. □

Finally, we arrive at the main result of this section.

THEOREM 9.1.7. *For any integrable random variable X, $\mathbb{E}(X|\mathcal{G})$ exists and is unique almost everywhere. Moreover, $|\mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}(|X||\mathcal{G})$ a.s.*

PROOF. We have already established uniqueness in Lemma 9.1.2. To prove existence, let X^+ and X^- denote the positive and negative parts of X. Then by Lemma 9.1.6, $Y_1 := \mathbb{E}(X^+|\mathcal{G})$ and $Y_2 := \mathbb{E}(X^-|\mathcal{G})$ exist. Let $Y := Y_1 - Y_2$. Since X is integrable, so are X^+ and X^-. So by the last assertion of Lemma 9.1.6, Y_1 and Y_2 are integrable, and thus, so is Y. Therefore for any $B \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}(X;B) &= \mathbb{E}(X^+;B) - \mathbb{E}(X^-;B)\\ &= \mathbb{E}(Y_1;B) - \mathbb{E}(Y_2;B) = \mathbb{E}((Y_1 - Y_2);B) = \mathbb{E}(Y;B).\end{aligned}$$

Thus, Y qualifies as $\mathbb{E}(X|\mathcal{G})$. Finally, by the linearity of conditional expectation for nonnegative random variables, $|Y| \leq Y_1 + Y_2 = \mathbb{E}(X^+|\mathcal{G}) + \mathbb{E}(X^-|\mathcal{G}) = \mathbb{E}(|X||\mathcal{G})$ a.s. □

Thus, we have established the existence and uniqueness of the conditional expectation of an integrable random variable given a σ-algebra. One can similarly define the conditional probability of an event, as $\mathbb{P}(A|\mathcal{G}) := \mathbb{E}(1_A|\mathcal{G})$. Another common notation is that if X and Y are two random variables, then $\mathbb{E}(Y|\sigma(X))$ is written as $\mathbb{E}(Y|X)$.

EXERCISE 9.1.8. Show that if A and B are two events with $\mathbb{P}(B) > 0$, and $\mathcal{G}$ is the σ-algebra generated by B (that is, $\mathcal{G} = \{\emptyset, B, B^c, \Omega\}$), then

$$\mathbb{P}(A|\mathcal{G}) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

EXERCISE 9.1.9. If X is random variable that takes value in some finite or countable set $\mathcal{X}$, then for any other random variable Y, $\mathbb{E}(Y|X) = g(X)$ a.s., where g is defined as

$$g(x) = \mathbb{E}(Y|X = x) := \frac{\mathbb{E}(Y; X = x)}{\mathbb{P}(X = x)},$$

when $\mathbb{P}(X = x) > 0$. When $\mathbb{P}(X = x) = 0$, $g(x)$ may be defined arbitrarily.

EXERCISE 9.1.10. If (X, Y) is a pair of real-valued random variables having a joint probability density function f, then $\mathbb{E}(Y|X) = g(X)$ a.s., where g is defined as

$$g(x) = \mathbb{E}(Y|X = x) := \int_{\mathbb{R}} y f(y|x) dy,$$

where $f(y|x)$ is the conditional probability density function of Y given $X = x$, defined as

$$f(y|x) := \frac{f(x,y)}{\int_{\mathbb{R}} f(x,z)dz}$$

where the denominator is nonzero. If the denominator is zero, $f(y|x)$ may be defined arbitrarily.

EXERCISE 9.1.11. Let $\Omega = [0,1)$, equipped with its Borel σ-algebra, so that a real-valued random variable X defined on Ω is just a measurable map from $[0,1)$ into $\mathbb{R}$. Take any $n \geq 1$. Let $\mathcal{F}_n$ be the σ-algebra generated by the intervals $[i/n, (i+1)/n)$, $0 \leq i \leq n-1$. What is $\mathbb{E}(X|\mathcal{F}_n)$?

EXERCISE 9.1.12. If (X,Y) is distributed uniformly on the unit disk in $\mathbb{R}^2$ and A is a Borel subset of $\mathbb{R}$, compute $\mathbb{P}(Y \in A|X)$.

EXERCISE 9.1.13. In the above exercise, let (R,Θ) be the representation of (X,Y) in polar coordinates, with Θ ranging in $[0,2\pi)$. Compute $\mathbb{P}(Y \in A|\Theta)$. (Observe that in the measure-theoretic formulation, unlike our attempt in Exercise 9.1.1, the 'conditional probability of $Y \in A$ given the event $X = 0$' has no meaning, and so there is no contradiction. We can use the random variable $\mathbb{P}(Y \in A|X)$ to compute $\mathbb{P}(Y \in A, X \in B)$ for any B, by integrating $\mathbb{P}(Y \in A|X)$ over the set $\{X \in B\}$. Similarly, $\mathbb{P}(Y \in A|\Theta)$ can be used to compute $\mathbb{P}(Y \in A, \Theta \in B)$ for any B, by integrating $\mathbb{P}(Y \in A|\Theta)$ over the set $\{\Theta \in B\}$.)

9.2. Basic properties of conditional expectation

It is clear from the definition that if conditional expectation exists, it must be linear. That is,

$$\mathbb{E}(aX_1 + bX_2|\mathcal{G}) = a\mathbb{E}(X_1|\mathcal{G}) + b\mathbb{E}(X_2|\mathcal{G}) \text{ a.s.}$$

This holds because the right side is easily seen to be a valid candidate for the conditional expectation of $aX_1 + bX_2$ given $\mathcal{G}$, and we know that conditional expectation is unique.

Conditional expectation behaves nicely under independence. If X is independent of a σ-algebra $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s. To see this, simply note that for any $A \in \mathcal{G}$, $\mathbb{E}(X;A) = \mathbb{E}(X)\mathbb{P}(A) = \mathbb{E}(\mathbb{E}(X)1_A)$.

On the other hand, if X is $\mathcal{G}$-measurable, it is clear that $\mathbb{E}(X|\mathcal{G}) = X$ a.s., since X satisfies the defining property of $\mathbb{E}(X|\mathcal{G})$.

By Lemma 9.1.6, the conditional expectation of a nonnegative random variable is a nonnegative random variable. Together with linearity, this implies that conditional expectation is monotone. That is, if $X \leq Y$ a.s., then $\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})$. Moreover, by the monotone convergence theorem, this implies that if $\{X_n\}_{n\geq 1}$ is a sequence of nonnegative random variables increasing to a limit X, then

$$\mathbb{E}(X|\mathcal{G}) = \lim_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) \text{ a.s.}$$

One may say that this is the monotone convergence theorem for conditional expectation, or simply the conditional monotone convergence theorem. An immediate consequence of the conditional monotone convergence theorem is conditional Fatou's lemma, which says that if

$\{X_n\}_{n\geq 1}$ is a sequence of nonnegative random variables, then

$$\mathbb{E}(\liminf_{n\to\infty} X_n|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) \text{ a.s.}$$

To prove this, let $Y_n := \inf_{k\geq n} X_k$. Then Y_n increases to $Y := \liminf_{n\to\infty} X_n$, and $Y_n \geq X_n$ for each n. Therefore by the conditional monotone convergence theorem and the monotonicity of conditional expectation,

$$\mathbb{E}(Y|\mathcal{G}) = \lim_{n\to\infty} \mathbb{E}(Y_n|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) \text{ a.s.}$$

Similarly, we have the conditional dominated convergence theorem: Suppose that $X_n \to X$ a.s. and $|X_n|$ is uniformly bounded by an integrable random variable Z. Then $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ a.s. as $n \to \infty$. To prove this, first note that $\{X_n + Z\}_{n\geq 1}$ is a sequence of nonnegative random variables converging to $X + Z$, and so by the conditional Fatou's lemma,

$$\mathbb{E}(X + Z|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(X_n + Z|\mathcal{G}) \text{ a.s.},$$

which gives

$$\mathbb{E}(X|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) \text{ a.s.}$$

Similarly, $Z - X_n$ is a sequence of nonnegative random variables converging to $Z - X$, so again by the conditional Fatou's lemma,

$$\mathbb{E}(Z - X|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbb{E}(Z - X_n|\mathcal{G}) \text{ a.s.},$$

which gives

$$\mathbb{E}(X|\mathcal{G}) \geq \limsup_{n\to\infty} \mathbb{E}(X_n|\mathcal{G}) \text{ a.s.}$$

Combining the two inequalities, we get the desired result. Under the hypotheses of the conditional dominated convergence theorem, we can also prove that $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ in L^1, as follows. First, note that for each n, $|\mathbb{E}(X_n|\mathcal{G})| \leq \mathbb{E}(|X_n||\mathcal{G}) \leq \mathbb{E}(Z|\mathcal{G})$. By the definition of conditional expectation, $\mathbb{E}(\mathbb{E}(Z|\mathcal{G})) = \mathbb{E}(Z) < \infty$. Lastly, we know that $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ a.s. So, applying the L^1 assertion of the dominated convergence theorem, we get that $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$ in L^1.

The above properties of conditional expectation are all analogues of properties of unconditional expectation. Conditional expectation has a few properties that have no unconditional analogue. One example is the tower property, which says that if we have two σ-algebras $\mathcal{G}' \subset \mathcal{G}$, then

$$\mathbb{E}(X|\mathcal{G}') = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{G}') \text{ a.s.}$$

To prove this, take any $B \in \mathcal{G}'$. Then $B \in \mathcal{G}$, and hence

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G}); B) = \mathbb{E}(X; B) = \mathbb{E}(\mathbb{E}(X|\mathcal{G}'); B).$$

Thus, $\mathbb{E}(X|\mathcal{G}')$ satisfies the defining property of the conditional expectation of $\mathbb{E}(X|\mathcal{G})$ given $\mathcal{G}'$, which proves the claim.

A second example is the property that if Y is $\mathcal{G}$-measurable and XY is integrable (in addition to X being integrable), then

$$\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}) \text{ a.s.} \tag{9.2.1}$$

To prove this, let us first prove the weaker statement

$$\mathbb{E}(XY) = \mathbb{E}(Y\mathbb{E}(X|\mathcal{G})). \tag{9.2.2}$$

Note that this follows from the definition of conditional expectation if $Y = 1_B$ for some $B \in \mathcal{G}$. Therefore it holds for any simple $\mathcal{G}$-measurable Y. Next, if X and Y are both nonnegative, we can take a sequence of nonnegative, simple, $\mathcal{G}$-measurable random variables increasing to Y and apply the monotone convergence theorem on both sides to get (9.2.2). In particular, this shows that if X and Y are nonnegative and XY is integrable, then so is $Y\mathbb{E}(X|\mathcal{G})$.

Finally, in the general case, note that the integrability of XY implies that X^+Y^+, X^+Y^-, X^-Y^+ and X^-Y^- are all integrable. By linearity, this gives us (9.2.2) for any X and Y such that X and XY are integrable, and Y is $\mathcal{G}$-measurable.

Next, to get (9.2.1), replace Y by $Y1_B$ in (9.2.2), where $B \in \mathcal{G}$. Since the integrability of XY implies that of $XY1_B$, we get

$$\mathbb{E}(XY; B) = \mathbb{E}(Y\mathbb{E}(X|\mathcal{G}); B).$$

Since this holds for any $B \in \mathcal{G}$, we get (9.2.1).

The following result is another basic property of conditional expectation that is often useful.

PROPOSITION 9.2.1. *Let S and T be two measurable spaces. Let X be an S-valued random variable and Y be a T-valued random variable, defined on the same probability space, such that X and Y are independent. Let $\phi : S \times T \to \mathbb{R}$ be a measurable function (with respect to the product σ-algebra) such that $\phi(X, Y)$ is an integrable random variable. For each $x \in S$, define*

$$\psi(x) := \mathbb{E}(\phi(x, Y)).$$

Then ψ is a measurable function, $\psi(X)$ is an integrable random variable and $\psi(X) = \mathbb{E}(\phi(X, Y)|X)$ a.s.

PROOF. Take any $A \in \sigma(X)$. Then $A = X^{-1}(B)$ for some B in the σ-algebra of S. Let μ be the law of X and ν be the law of Y. Then by Exercise 6.1.2, the integrability of $\phi(X, Y)$, and Fubini's theorem,

$$\begin{aligned}\mathbb{E}(\phi(X, Y); A) &= \int_S \int_T \phi(x, y) 1_B(x) d\nu(y) d\mu(x) \\ &= \int_B \int_T \phi(x, y) d\nu(y) d\mu(x) = \int_B \psi(x) d\mu(x) \\ &= \int_S \psi(x) 1_B(x) d\mu(x) = \mathbb{E}(\psi(X); A).\end{aligned}$$

This shows that $\psi(X)$ is indeed a version of $\mathbb{E}(\phi(X, Y)|X)$. The measurability and ψ and the integrability of $\psi(X)$ are also consequences of Fubini's theorem. □

In the following exercises, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space on which all our random variables are defined, and $\mathcal{G}$ is an arbitrary sub-σ-algebra of $\mathcal{F}$.

EXERCISE 9.2.2. If $X_n \to X$ in L^1, prove that $\mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G})$.

EXERCISE 9.2.3. On the unit interval with Lebesgue measure, let $X(\omega) = \omega$ and for an arbitrary positive integer n let $Y(\omega) = n\omega - [n\omega]$, where $[x]$ denotes the largest integer $\leq x$. What is $\mathbb{E}(X|Y)$? Explain your answer.

EXERCISE 9.2.4. Suppose $\mathbb{E}(X_i)$ exists for $i = 1, 2$ and $\mathbb{E}(X_1; A) \leq \mathbb{E}(X_2; A)$ for all $A \in \mathcal{F}$. Show that $\mathbb{P}(X_1 \leq X_2) = 1$.

EXERCISE 9.2.5. Suppose X and Y are square integrable. Show that for every sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, $\mathbb{E}[X\mathbb{E}(Y|\mathcal{G})] = \mathbb{E}[\mathbb{E}(X|\mathcal{G})Y]$.

EXERCISE 9.2.6. Let $\mathrm{Var}(X|\mathcal{F}) = \mathbb{E}(X^2|\mathcal{F}) - \mathbb{E}(X|\mathcal{F})^2$. Prove that

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X|\mathcal{F})) + \mathrm{Var}(\mathbb{E}(X|\mathcal{F})).$$

EXERCISE 9.2.7. Suppose that $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathbb{E}(X|Y) = Y$ a.s. and $\mathbb{E}(Y|X) = X$ a.s. Prove that $X = Y$ a.s.

EXERCISE 9.2.8. Let S be a random variable with $\mathbb{P}(S > t) = \exp(-t)$ for $t > 0$. Find $\mathbb{E}(S|\max\{S, t\})$ and $\mathbb{E}(S|\min\{S, t\})$ for each $t > 0$.

EXERCISE 9.2.9. Suppose θ is 1 or 0 with probabilities p and $1 - p$, respectively, and independently of θ, X and Y are independent and identically distributed. Let $Z = (Z_1, Z_2)$, where $Z_1 = \theta X + (1-\theta)Y$ and $Z_2 = (1-\theta)X + \theta Y$. Show that Z and θ are independent. Then, find an explicit expression for $\mathbb{E}(g(X, Y)|Z)$, for an arbitrary bounded Borel-measurable function g.

EXERCISE 9.2.10. Suppose that $\mathbb{E}(X_n)$ exists for all n, $\mathbb{E}(X)$ exists and $X_1 \leq X_2 \leq \cdots \to X$ with probability one. Show that $\mathbb{E}(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ a.e. on $\{\mathbb{E}(X_1|\mathcal{G}) > -\infty\}$ as $n \to \infty$. Hint: Consider $B = \{\mathbb{E}(X_1|\mathcal{G}) \geq -b\}$ and the random variables $X_n 1_B$.

EXERCISE 9.2.11. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F} = \sigma(\cup_{i=1}^{\infty} \mathcal{F}_i)$. Let $X \in L^1(\mathcal{F})$. Let $\mathcal{A}$ be an algebra that generates the σ-algebra $\mathcal{F}$. First, show that for any $\epsilon > 0$ there exists a random variable Y that is a finite linear combination of indicator variables of sets in $\mathcal{A}$ such that $\mathbb{E}|X - Y| < \epsilon$. (Hint: See Theorem 1.2.6.) Using only this result and the basic properties of conditional expectations show that $E(X|\mathcal{F}_n) - X \to 0$ in L^1 as $n \to \infty$.

EXERCISE 9.2.12. Suppose that $0 \leq X_n \to 0$ in probability, $X_n \leq Y$ with probability one for all n, and $\mathbb{E}(Y) < \infty$. Prove that $\mathbb{E}(X_n|\mathcal{G}) \to 0$ in probability as $n \to \infty$. Hint: Break up $\mathbb{E}(X_n|\mathcal{G})$ into three parts according as $X_n < \epsilon$, $\epsilon \leq X_n < a$, or $X_n > a$.

EXERCISE 9.2.13. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two independent sub-σ-algebras of $\mathcal{F}$. Let $\mathcal{G} := \sigma(\mathcal{G}_1 \cup \mathcal{G}_2)$. If an integrable random variable X has the property that $\sigma(\sigma(X) \cup \mathcal{G}_1)$ is independent of $\mathcal{G}_2$, show that $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}_1)$ a.s.

9.3. Jensen's inequality for conditional expectation

The following result is known as Jensen's inequality for conditional expectation. Its proof is a bit more involved than the proof of ordinary Jensen's inequality.

THEOREM 9.3.1 (Jensen's inequality for conditional expectation). *Let X be an integrable random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let I be an interval containing the range of X, and let $\phi : I \to \mathbb{R}$ be a convex function such that $\phi(X)$ is integrable. Then for any σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, $\mathbb{E}(\phi(X)|\mathcal{G}) \geq \phi(\mathbb{E}(X|\mathcal{G}))$ a.s.*

For the proof, we need a couple of lemmas about convex functions.

LEMMA 9.3.2. *The supremum of any collection of convex functions defined on an interval is convex.*

PROOF. Let $\mathcal{A}$ be a collection of convex functions defined on an interval I. Let $g(x) := \sup_{f\in\mathcal{A}} f(x)$. Take any $x, y \in I$, with $x \leq y$, and any $t \in [0, 1]$. Then for any $f \in \mathcal{A}$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \leq tg(x) + (1-t)g(y).$$

Taking supremum over $f \in \mathcal{A}$ on the left, we get

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

Thus, g is convex. □

LEMMA 9.3.3. *Let I be an interval and $\phi : I \to \mathbb{R}$ be a convex function. Then there are sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ such that $\phi(x) = \sup_{n\geq 1}(a_n x + b_n)$ for all $x \in I$.*

PROOF. Take any $x \in \mathbb{Q} \cap I$. By Exercise 4.2.5, there are numbers a_x and b_x such that $\phi(y) \geq a_x y + b_x$ for all $y \in I$, and $\phi(x) = a_x x + b_x$. This shows that if we define

$$g(x) := \sup_{y\in\mathbb{Q}\cap I} (a_y x + b_y)$$

for each $x \in I$, then $\phi(x) = g(x)$ for each $x \in \mathbb{Q} \cap I$. By Lemma 9.3.2, g is convex, and hence continuous. Thus, ϕ and g are two continuous functions that agree on a dense subset of I. Therefore ϕ and g must agree everywhere on I. □

PROOF OF THEOREM 9.3.1. Let a_n and b_n be as in Lemma 9.3.3. Then for any n,

$$\mathbb{E}(\phi(X)|\mathcal{G}) \geq \mathbb{E}(a_n X + b_n|\mathcal{G}) = a_n\mathbb{E}(X|\mathcal{G}) + b_n \text{ a.s.}$$

Implicit in the above display is that we have chosen and fixed some versions of the conditional expectations displayed on the two sides. Since the above inequality holds almost surely for any given n, they hold simultaneously for all n with probability one. By the monotonicity of conditional expectation, it follows that $\mathbb{E}(X|\mathcal{G}) \in I$ a.s. Thus, with probability one,

$$\mathbb{E}(\phi(X)|\mathcal{G}) \geq \sup_{n\geq 1}(a_n\mathbb{E}(X|\mathcal{G}) + b_n) = \phi(\mathbb{E}(X|\mathcal{G}))a.s.,$$

completing the proof of the theorem. □

9.4. Martingales

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An increasing sequence of σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ contained in $\mathcal{F}$ is called a filtration. A sequence of random variables $\{X_n\}_{n\geq 0}$ is said to be adapted to this filtration if for each n, X_n is $\mathcal{F}_n$-measurable. An adapted

sequence is called a martingale if for each n, X_n is integrable, and

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n \text{ a.s.}$$

Note that by the tower property, $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$ a.s. whenever $n \geq m$. Also note that $\mathbb{E}(X_n) = \mathbb{E}(X_0)$ for all n.

EXAMPLE 9.4.1 (Simple random walk). Perhaps the simplest example of a nontrivial martingale is a random walk with mean zero increments. Let $X_1, X_2, \ldots$ be independent, integrable random variables with $\mathbb{E}(X_i) = 0$ for all i. Let $S_0 := 0$ and $S_n := X_1 + \cdots + X_n$. Let $\mathcal{F}_n$ be the σ-algebra generated by $X_1, \ldots, X_n$ for $n \geq 1$, and let $\mathcal{F}_0$ be the trivial σ-algebra. Then it is easy to see that $\{S_n\}_{n\geq 0}$ is a martingale adapted to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. To see this, first note that that $\mathbb{E}|X_i| < \infty$ for each i, the triangle inequality shows that S_n is integrable for each n. It is obvious that S_n is $\mathcal{F}_n$-measurable. Next, by linearity of conditional expectation, and the fact that S_{n-1} is $\mathcal{F}_{n-1}$ measurable, we have

$$\begin{aligned}\mathbb{E}(S_n|\mathcal{F}_{n-1}) &= \mathbb{E}(S_{n-1} + X_n|\mathcal{F}_{n-1}) \\ &= S_{n-1} + \mathbb{E}(X_n|\mathcal{F}_{n-1}).\end{aligned}$$

But, since X_n is independent of $\mathcal{F}_{n-1}$, $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = \mathbb{E}(X_n) = 0$. This proves the martingale property.

EXERCISE 9.4.2. Let $X_1, X_2, \ldots$ be square-integrable independent random variables, and let $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 := \mathrm{Var}(X_i)$. Show that

$$Z_n := \left(\sum_{i=1}^n (X_i - \mu_i)\right)^2 - \sum_{i=1}^n \sigma_i^2$$

is a martingale adapted to $\mathcal{F}_n$.

EXERCISE 9.4.3. Let $X_1, X_2, \ldots$ be i.i.d. random variables and $S_n := X_1 + \cdots + X_n$. Suppose that for some $\theta \in \mathbb{R} \setminus \{0\}$, $m(\theta) := \mathbb{E}(e^{\theta X_1})$ is finite. Then show that

$$M_n := \frac{e^{\theta S_n}}{m(\theta)^n}$$

is a martingale adapted to $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$.

EXERCISE 9.4.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration of sub-σ-algebras of $\mathcal{F}$. Let X be an integrable random variable defined on Ω. Let $Y_n := \mathbb{E}(X|\mathcal{F}_n)$. Show that $\{Y_n\}_{n\geq 0}$ is a martingale adapted to $\{\mathcal{F}_n\}_{n\geq 0}$.

9.5. Stopping times

A random variable T taking value in $\{0, 1, 2, \ldots\} \cup \{\infty\}$ is called a stopping time for a filtration $\{\mathcal{F}_n\}_{n\geq 0}$ if for each finite n, then the event $\{T = n\}$ is $\mathcal{F}_n$-measurable.

A large class of stopping times arise in the following way. Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration of σ-algebras. Let $\{X_n\}_{n\geq 0}$ be a sequence of real-valued random variables adapted to this filtration — that is, for each n, X_n is $\mathcal{F}_n$-measurable. Let A be a Borel subset of $\mathbb{R}$. Let $T := \inf\{n : X_n \in A\}$, where the infimum is interpreted as ∞ if $X_n \notin A$ for all n. This is a

stopping time for the filtration $\{\mathcal{F}_n\}_{n\geq 0}$, because for any finite n,

$$\{T = n\} = \{X_n \in A\} \cap \{X_j \notin A \text{ for all } j < n\},$$

which is an element of $\mathcal{F}_n$ by the given conditions.

EXERCISE 9.5.1. Let $\{S_n\}_{n\geq 0}$ be a simple symmetric random walk on $\mathbb{Z}$ starting at the origin. That is, $S_0 = 0$, and for each n, $S_{n+1} - S_n$ is 1 or -1 with equal probability, irrespective of past events. Let $\mathcal{F}_n$ be the σ-algebra generated by $S_0, \ldots, S_n$. Take any integers $a < 0 < b$, and let $T_{a,b} := \min\{n : S_n = a \text{ or } S_n = b\}$. Show that $T_{a,b}$ is a stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$.

EXERCISE 9.5.2. In the above exercise, show that $\mathbb{E}(T_{a,b}) < \infty$, and hence that $T_{a,b}$ is finite a.s. Hint: Consider the first occurrence of a continuous sequence of $a + b$ positive jumps of the walk, and use it to obtain a random upper bound on $T_{a,b}$.

EXERCISE 9.5.3. Let S_n be as above. Let $T := \max\{n : S_n \geq n\}$. Show that $T < \infty$ a.s., and that T is *not* a stopping time. Hint: For the first part, use the strong law of large numbers. For the second, produce an argument by contradiction to show that $\{T = 0\} \notin \mathcal{F}_0$.

EXERCISE 9.5.4. Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration. Show that a random variable T taking value in $\{0, 1, 2, \ldots\} \cup \{\infty\}$ is a stopping time with respect to this filtration if and only if $\{T > n\} \in \mathcal{F}_n$ for all $n \geq 0$, and also if and only if $\{T \leq n\} \in \mathcal{F}_n$ for all $n \geq 0$.

Given any stopping time T, there is an associated σ-algebra called the stopped σ-algebra, denoted by $\mathcal{F}_T$. It is defined as

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T = n\} \in \mathcal{F}_n \text{ for all } n\}.$$

Intuitively, $\mathcal{F}_T$ encodes 'all the information up to time T'. Alternatively, an event is in $\mathcal{F}_T$ if and only if we can determine whether it is true by observing what happens up to time T.

EXERCISE 9.5.5. Show that a stopping time T is always measurable with respect to the stopped σ-algebra $\mathcal{F}_T$.

EXERCISE 9.5.6. Give an example to show that the σ-algebra generated by a stopping time T may be a proper subset of $\mathcal{F}_T$.

EXERCISE 9.5.7. Let $T_{a,b}$ be as in Exercise 9.5.1. Take any $k \geq 1$, and let A be the event that by time $T_{a,b}$, the random walk visits 0 exactly k times. Show that $A \in \mathcal{F}_{T_{a,b}}$.

EXERCISE 9.5.8. Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let T be a stopping time with respect to this filtration. The *stopped random variable* X_T is defined as $X_T(\omega) := X_{T(\omega)}(\omega)$. Show that X_T is $\mathcal{F}_T$-measurable.

An important observation is that if S and T are stopping times such that $S \leq T$ always, then $\mathcal{F}_S \subseteq \mathcal{F}_T$. To see this, take any $A \in \mathcal{F}_S$. Then for any n,

$$A \cap \{T = n\} = A \cap \{S \leq n\} \cap \{T = n\},$$

since $S \leq T$. But $A \cap \{S \leq n\} \in \mathcal{F}_n$ since $A \in \mathcal{F}_S$, and $\{T = n\} \in \mathcal{F}_n$ since T is a stopping time. Therefore $A \cap \{T = n\} \in \mathcal{F}_n$.

Note that a random variable T that is identically equal to a positive integer n is trivially a stopping time. For this T, it is easy to see that $\mathcal{F}_T = \mathcal{F}_n$.

A stopping time is called bounded if there is a constant c such that $T \leq c$ always.

9.6. Optional stopping theorem

One of the most important results relating martingales and stopping times is the optional stopping theorem. The simplest version of this theorem goes as follows.

THEOREM 9.6.1 (Optional stopping theorem). *Let $\{X_n\}_{n\geq 0}$ be a martingale adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let S and T be bounded stopping times for this filtration, such that $S \leq T$ always. Then X_S and X_T are integrable, and $\mathbb{E}(X_T|\mathcal{F}_S) = X_S$ a.s. In particular, $\mathbb{E}(X_T) = \mathbb{E}(X_0)$.*

PROOF. Since T is a bounded stopping time, there is an integer n such that $S \leq T \leq n$ always. To prove integrability of X_S and X_T, simply note that $|X_T|$ and $|X_S|$ are both bounded by $|X_0| + \cdots + |X_n|$. Take any $A \in \mathcal{F}_S$. Then note that by the given conditions,

$$\mathbb{E}(X_n; A) = \sum_{i=0}^{n} \mathbb{E}(X_n; \{T = i\} \cap A).$$

But $\{T = i\} \cap A \in \mathcal{F}_i$ since $A \in \mathcal{F}_S \subseteq \mathcal{F}_T$, and $\mathbb{E}(X_n|\mathcal{F}_i) = X_i$ since $\{X_j\}_{j\geq 0}$ is a martingale. So,

$$\begin{aligned}\mathbb{E}(X_n; \{T = i\} \cap A) &= \mathbb{E}(\mathbb{E}(X_n|\mathcal{F}_i); \{T = i\} \cap A) \\ &= \mathbb{E}(X_i; \{T = i\} \cap A).\end{aligned}$$

Summing over i, we get

$$\mathbb{E}(X_n; A) = \sum_{i=0}^{n} \mathbb{E}(X_i; \{T = i\} \cap A) = \mathbb{E}(X_T; A).$$

But by the same argument, $\mathbb{E}(X_n; A) = \mathbb{E}(X_S; A)$. Therefore $\mathbb{E}(X_T; A) = \mathbb{E}(X_S; A)$ for all $A \in \mathcal{F}_S$. By Exercise 9.5.8, X_S is $\mathcal{F}_S$-measurable. Thus, $\mathbb{E}(X_T|\mathcal{F}_S) = X_S$ a.s. Considering the special case $S \equiv 0$, we get $\mathbb{E}(X_T|\mathcal{F}_0) = X_0$, which gives $\mathbb{E}(X_T) = \mathbb{E}(X_0)$. □

Although most stopping times that occur in practice are unbounded, there is a simple trick to apply the optional stopping theorem is a great variety of situations. Let T be a stopping time with respect to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Take any $n \geq 0$, and let $T \wedge n$ denote the minimum of T and n. That is, $T \wedge n$ is a random variable defined as $T \wedge n(\omega) := \min\{T(\omega), n\}$. Then $T \wedge n$ is a stopping time. To see this, note that for any $k \geq 0$,

$$\{T \wedge n > k\} = \begin{cases} \{T > k\} & \text{if } n > k, \\ \emptyset & \text{if } n \leq k, \end{cases}$$

and apply Exercise 9.5.4. But note that $T \wedge n$ is also bounded. Thus, we can apply the optional stopping time with $T \wedge n$, and later, take $n \to \infty$ to recover T, because $\lim_{n\to\infty} T \wedge n = T$ on the set $\{T < \infty\}$.

To see how this works, consider the following. Let $\{S_n\}_{n\geq 0}$ be a simple symmetric random walk on $\mathbb{Z}$, starting at the origin. Take any $a < 0 < b$, and let $T := \min\{n : S_n =$

a or $S_n = b\}$. Let $\mathcal{F}_n$ be the σ-algebra generated by $S_0, \ldots, S_n$. As noted in Section 9.4, $\{S_n\}_{n\geq 0}$ is a martingale adapted to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Take any n. Then $T \wedge n$ is a bounded stopping time, and hence, by the optional stopping theorem,

$$\mathbb{E}(S_{T\wedge n}) = \mathbb{E}(S_0) = 0. \tag{9.6.1}$$

By Exercise 9.5.2, $T < \infty$ a.s. Thus, for almost every ω, $T(\omega) \wedge n = T(\omega)$ for all sufficiently large n, and hence

$$\lim_{n\to\infty} S_{T\wedge n}(\omega) = \lim_{n\to\infty} S_{T(\omega)\wedge n}(\omega) = S_{T(\omega)}(\omega) = S_T(\omega).$$

In other words, $S_{T\wedge n} \to S_T$ a.s. as $n \to \infty$. Note also that S_n remains in the interval $[a, b]$ up to time T, and therefore, $|S_{T\wedge n}| \leq \max\{|a|, b\}$ for all n. Combining this with (9.6.1) and the dominated convergence theorem, we get that $\mathbb{E}(S_T) = 0$. But S_T can take only two values, a and b. Thus,

$$0 = \mathbb{E}(S_T) = a\mathbb{P}(S_T = a) + b(1 - \mathbb{P}(S_T = a)),$$

which gives

$$\mathbb{P}(S_T = a) = \frac{b}{b-a}.$$

Thus, the chance that the walk exits the interval $[a, b]$ through a is $b/(b-a)$, and the chance that it exits through b is $-a/(b-a)$. (This is known as the *gambler's ruin* problem: If a gambler starts with x dollars, and wins or loses 1 dollar with equal probability at each turn, then what is the chance that he will reach a threshold of y dollars before hitting 0 (that is, getting ruined)? It is easy to use the above formula to show that in this case, the chance is x/y.)

EXERCISE 9.6.2. In the above setting, compute $\mathbb{E}(T)$ using the martingale $S_n^2 - n$.

EXERCISE 9.6.3. If the random walk is *biased* — that is, the chance of a positive step is $p \neq 1/2$, show that $(q/p)^{S_n}$ is a martingale, where $q = 1-p$. Using this compute $\mathbb{P}(S_T = a)$.

EXERCISE 9.6.4. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with negative mean and a nonzero probability of being positive, and finite moment generating function m. Show that there exists $\theta^* > 0$ such that $m(\theta^*) = 1$, and that $\{e^{\theta^* S_n}\}_{n\geq 0}$ is a martingale, where $S_n = X_1 + \cdots + X_n$ and $S_0 = 0$. Check that the martingale in the previous exercise is a special case of this one when $p < 1/2$. Hint: For the first part, first show that m is differentiable everywhere, then show that $m'(0) < 0$, and finally show that $m(\theta) \to \infty$ as $\theta \to \infty$.

EXERCISE 9.6.5. Let all notation be as in the previous exercise. Let $M := \max_{n\geq 0} S_n$. Show that M is finite a.s. Next, using the martingale from the previous exercise, prove that for any $x \in (0, \infty)$, $\mathbb{P}(M \geq x) \leq e^{-\theta^* x}$. Hint: Use the stopping time $T := \inf\{n \geq 0 : S_n \geq x\}$ and find an upper bound on $\mathbb{P}(T < \infty)$.

EXERCISE 9.6.6. Let S_n be the total assets of an insurance company at the end of year n. In year n, premiums totaling $c > 0$ are received and claims ζ_n are paid where $\zeta_n \sim N(\mu, \sigma^2)$ and $\mu < c$. To be precise, if $\xi_n = c - \zeta_n$, then $S_n = S_{n-1} + \xi_n$. The

company is ruined if its assets drop to 0 or less. Show that if $S_0 > 0$ is nonrandom, then $\mathbb{P}(\text{ruin}) \leq \exp(-2(c-\mu)S_0/\sigma^2)$.

EXERCISE 9.6.7 (Wald's equation). Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables with mean μ and let $S_n := X_1 + \cdots + X_n$. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ for $n \geq 1$, and let $\mathcal{F}_0$ be the trivial σ-algebra. Let T be a stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. If $\mathbb{E}(T) < \infty$, show that $\mathbb{E}(S_T) = \mu\mathbb{E}(T)$. Hint: Show that for all n, $|S_{T\wedge n}|$ is bounded by $\sum_{i=1}^{\infty} |X_i| 1_{\{T\geq i\}}$, and that this random variable is integrable.

EXERCISE 9.6.8. Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables with mean 0 and let $S_n := X_1 + \cdots + X_n$. Take any $x > 0$, and let $T := \inf\{n \geq 0 : S_n \geq x\}$. Prove that $\mathbb{E}(T) = \infty$. Hint: Use the previous exercise.

9.7. Submartingales and supermartingales

Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration of σ-algebras. A sequence of integrable random variables $\{X_n\}_{n\geq 0}$ adapted to this filtration is called a submartingale if for each n, we have

$$X_n \leq \mathbb{E}(X_{n+1}|\mathcal{F}_n) \text{ a.s.},$$

and a supermartingale if for each n, we have

$$X_n \geq \mathbb{E}(X_{n+1}|\mathcal{F}_n) \text{ a.s.}$$

A simple way to produce a submartingale is to apply a convex function to a martingale. Indeed, if $\{Y_n\}_{n\geq 0}$ is a martingale and ϕ is a convex function defined on an interval containing the ranges of the Y_n's, such that $\phi(Y_n)$ is integrable for all n, then by the conditional Jensen inequality, we have

$$\mathbb{E}(\phi(Y_{n+1})|\mathcal{F}_n) \geq \phi(\mathbb{E}(Y_{n+1}|\mathcal{F}_n)) = \phi(Y_n).$$

One can also produce submartingales from other submartingales by applying functions that are both convex and non-decreasing. Indeed, if $\{X_n\}_{n\geq 0}$ is a submartingale and ϕ is such a function, and $\phi(X_n)$ is integrable for all n, then

$$\mathbb{E}(\phi(X_{n+1})|\mathcal{F}_n) \geq \phi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \geq \phi(X_n).$$

Similarly, supermartingales can be produced by applying concave functions to martingales, or concave non-increasing functions to supermartingales.

Many results for martingales have analogous versions for submartingales and supermartingales. The following exercises contain some of these.

EXERCISE 9.7.1 (Optional stopping theorem for submartingales and supermartingales). Let $\{X_n\}_{n\geq 0}$ be a submartingale adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let S and T be bounded stopping times for this filtration, such that $S \leq T$ always. Then X_S and X_T are integrable, and $\mathbb{E}(X_T|\mathcal{F}_S) \geq X_S$ a.s. In particular, $\mathbb{E}(X_T) \geq \mathbb{E}(X_0)$. Prove also the analogous result for supermartingales.

There is also a way to extract a martingale out of a submartingale or supermartingale — or, in fact, any adapted sequence. This procedure, known as the 'Doob decomposition',

goes as follows. Let $\{X_n\}_{n\geq 0}$ be a sequence of integrable random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Define $M_0 := 0$ and for each $n \geq 1$,

$$M_n := \sum_{k=0}^{n-1} (X_{k+1} - \mathbb{E}(X_{k+1}|\mathcal{F}_k)),$$

and

$$A_n := \sum_{k=0}^{n-1} (\mathbb{E}(X_{k+1}|\mathcal{F}_k) - X_k).$$

Then note that by definition,

$$X_n = X_0 + M_n + A_n. \tag{9.7.1}$$

It is easy to see that M_n and A_n are both integrable for each n. Next, note that M_n is $\mathcal{F}_n$-measurable for each n, and

$$\mathbb{E}(M_n|\mathcal{F}_{n-1}) = \sum_{k=0}^{n-2} (X_{k+1} - \mathbb{E}(X_{k+1}|\mathcal{F}_k)) = M_{n-1}.$$

Thus, $\{M_n\}_{n\geq 0}$ is a martingale adapted to $\{\mathcal{F}_n\}_{n\geq 0}$. Finally, note that $\{A_n\}_{n\geq 0}$ is not only an adapted process, but actually, A_n is $\mathcal{F}_{n-1}$-measurable for each n. Such processes are called 'predictable'. Thus, the Doob decomposition (9.7.1) expresses any adapted process as the sum of the initial value, plus a martingale, plus a predictable process. Additionally, if $\{X_n\}_{n\geq 0}$ is a submartingale, we have that $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ for each n, and hence $\{A_n\}_{n\geq 0}$ is a nonnegative and increasing process. Similarly, if $\{X_n\}_{n\geq 0}$ is a supermartingale, $\{A_n\}_{n\geq 0}$ is a nonpositive and decreasing process.

The following construction is another method of obtaining submartingales, supermartingales and martingales from arbitrary adapted processes, which is often useful in applications.

PROPOSITION 9.7.2. *Let $\{X_n\}_{n\geq 0}$ be a sequence of integrable random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let T be a stopping time for this filtration. Suppose that for each n, $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ a.s. on the set $\{T > n\}$, that is,*

$$\mathbb{P}(\{\mathbb{E}(X_{n+1}|\mathcal{F}_n) < X_n\} \cap \{T > n\}) = 0.$$

Then $\{X_{T\wedge n}\}_{n\geq 0}$ is a submartingale adapted to $\{\mathcal{F}_n\}_{n\geq 0}$. Similarly, if we have that $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ a.s. on $\{T > n\}$, then $\{X_{T\wedge n}\}_{n\geq 0}$ is a supermartingale, and if $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ a.s. on $\{T > n\}$, then $\{X_{T\wedge n}\}_{n\geq 0}$ is a martingale.

(The process $\{X_{T\wedge n}\}_{n\geq 0}$ is often called the 'stopped' version of the process $\{X_n\}_{n\geq 0}$, since it progresses as X_n up to time T, and then gets frozen at X_T.)

PROOF. Take any $n \geq 0$. By Exercise 9.5.8, $X_{T\wedge n}$ is $\mathcal{F}_{T\wedge n}$-measurable. Since $T \wedge n \leq n$, and $T \wedge n$ and n are both stopping times, we have that $\mathcal{F}_{T\wedge n} \subseteq \mathcal{F}_n$. Thus, $X_{T\wedge n}$ is $\mathcal{F}_n$-measurable. Also, clearly, $|X_{T\wedge n}| \leq |X_0| + \cdots + |X_n|$, which shows that $X_{T\wedge n}$ is integrable.

Now note that

$$\begin{aligned}\mathbb{E}(X_{T\wedge(n+1)}|\mathcal{F}_n) &= \sum_{i=0}^{n} \mathbb{E}(X_{T\wedge(n+1)}1_{\{T=i\}}|\mathcal{F}_n) + \mathbb{E}(X_{T\wedge(n+1)}1_{\{T>n\}}|\mathcal{F}_n)\\ &= \sum_{i=0}^{n} \mathbb{E}(X_i 1_{\{T=i\}}|\mathcal{F}_n) + \mathbb{E}(X_{n+1}1_{\{T>n\}}|\mathcal{F}_n)\\ &= \sum_{i=0}^{n} X_i 1_{\{T=i\}} + 1_{\{T>n\}}\mathbb{E}(X_{n+1}|\mathcal{F}_n),\end{aligned}$$

where the last identity holds because X_i and $1_{\{T=i\}}$ are $\mathcal{F}_i$-measurable for each i. But

$$\sum_{i=0}^{n} X_i 1_{\{T=i\}} = X_T 1_{\{T\le n\}},$$

and by the assumed property, $1_{\{T>n\}}\mathbb{E}(X_{n+1}|\mathcal{F}_n) \ge 1_{\{T>n\}}X_n$ a.s. Thus,

$$\mathbb{E}(X_{T\wedge(n+1)}|\mathcal{F}_n) \ge X_T 1_{\{T\le n\}} + X_n 1_{\{T>n\}} = X_{T\wedge n}.$$

This proves the claim for submartingales. The proofs for supermartingales and martingales are exactly the same. □

The above result is sometimes used in the following way. Let $\{X_n\}_{n\ge0}$ be a sequence of integrable random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\ge0}$, starting at $X_0 \equiv x$ for some nonrandom $x > 0$. Suppose that T is a stopping time with respect to this filtration, with the property that there is some $a \le x$ such that $X_{T\wedge n} \ge a$ a.s., and there is some $b > 0$ such that $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \le X_n - b$ a.s. on the set $\{T > n\}$. Then we have the bound

$$\mathbb{E}(T) \le \frac{x-a}{b}, \tag{9.7.2}$$

by the following argument. Define

$$Y_n := X_n + bn.$$

Then by the stated property, a.s. on the set $\{T > n\}$,

$$\begin{aligned}\mathbb{E}(Y_{n+1}|\mathcal{F}_n) &= \mathbb{E}(X_{n+1}|\mathcal{F}_n) + b(n+1)\\ &\le X_n - b + b(n+1) = X_n + bn = Y_n.\end{aligned}$$

Therefore by Proposition 9.7.2, $\{Y_{T\wedge n}\}_{n\ge0}$ is a supermartingale. In particular,

$$\mathbb{E}(Y_{T\wedge n}) \le \mathbb{E}(Y_{T\wedge 0}) = \mathbb{E}(Y_0) = x.$$

But

$$\mathbb{E}(Y_{T\wedge n}) = \mathbb{E}(X_{T\wedge n}) + b\mathbb{E}(T\wedge n)$$

Since $X_{T\wedge n} \ge a$ a.s., combining the last two displays shows that

$$\mathbb{E}(T\wedge n) \le \frac{x-a}{b}.$$

Taking $n \to \infty$ and applying the monotone convergence theorem, we get (9.7.2). The following exercises demonstrate how this technique can be applied.

EXERCISE 9.7.3. A gambler wins or loses 1 dollar with equal probability at each turn of a game. However, if his total assets at any point of time are bigger than some amount a, he has to pay a tax of b dollars for playing that turn. Suppose that the gambler starts playing with assets totaling x dollars. Let T be the first time at which his assets go below a dollars. Prove that $\mathbb{E}(T) \leq (x-a)/b$. In particular, deduce that T is finite a.s.

EXERCISE 9.7.4. Consider the following random walk on the integers. The walk starts at some odd integer $x > 3$. At each step, if the walk is at some y, then it goes to either $p-2$ or $p+2$ with equal probability, where p is the largest prime divisor of y. Let T be the first time the walk hits a prime number or 1. Prove that $\mathbb{E}(T) \leq C \log x$, where C is a constant that does not depend on x. In particular, deduce that $T < \infty$ a.s. (Note that if X_T is a prime, then either $X_T + 2$ or $X_T - 2$ is also a prime, which means that X_T is a twin prime.)

9.8. Optimal stopping and Snell envelopes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_N$ be a filtration of sub-σ-algebras of $\mathcal{F}$. Let $X_1, \ldots, X_N$ be integrable random variables defined on Ω, not necessarily adapted to this filtration. The 'finite horizon optimal stopping problem' in this setting is the problem of finding the stopping time T (with respect to this filtration) that maximizes $\mathbb{E}(X_T)$. The idea is that $\mathcal{F}_n$ encodes 'information up to time n', and X_n is the reward obtained if we execute some action at time n, where our decision to execute the action is based solely on information available up to time n. We want to devise a strategy for executing the action at an opportune time, so that the expected reward is maximized. Surprisingly, this problem has a solution at this complete level of generality, without any further assumptions.

The first step is to simplify the problem by defining $Y_n := \mathbb{E}(X_n|\mathcal{F}_n)$ and proving the following lemma.

LEMMA 9.8.1. *For any stopping time T for the filtration $\{\mathcal{F}_n\}_{1\leq n\leq N}$ (taking value in $\{1, \ldots, N\}$), $\mathbb{E}(X_T) = \mathbb{E}(Y_T)$.*

PROOF. Note that

$$\mathbb{E}(Y_T) = \sum_{n=1}^{N} \mathbb{E}(Y_n; T = n) = \sum_{n=1}^{N} \mathbb{E}(X_n; T = n),$$

because $\{T = n\} \in \mathcal{F}_n$ and $Y_n = \mathbb{E}(X_n|\mathcal{F}_n)$. But the last expression is just $\mathbb{E}(X_T)$. □

Next, we define a sequence of random variables known as the 'Snell envelope' of $Y_1, \ldots, Y_N$. Let $V_N := Y_N$, and define, by backward induction,

$$V_n := \max\{Y_n, \mathbb{E}(V_{n+1}|\mathcal{F}_n)\} \tag{9.8.1}$$

for $n = N-1, N-2, \ldots, 1$. Note that this well-defined, because each V_n is integrable (easy to show by backward induction). Also, clearly, V_n is $\mathcal{F}_n$-measurable. Observe that $\{V_n\}_{1\leq n\leq N}$ is a supermartingale adapted to $\{\mathcal{F}_n\}_{1\leq n\leq N}$, because it is a trivial consequence of the above definition that $V_n \geq \mathbb{E}(V_{n+1}|\mathcal{F}_n)$ a.s. Note also that $V_n \geq Y_n$ a.s. for each n. In fact, the following holds (which is why it's called an 'envelope'):

EXERCISE 9.8.2. Prove that $\{V_n\}_{1\leq n\leq N}$ is the 'smallest' supermartingale dominating $\{Y_n\}_{1\leq n\leq N}$, in the sense that if $\{U_n\}_{1\leq n\leq N}$ is any other supermartingale adapted to $\{\mathcal{F}_n\}_{1\leq n\leq N}$ such that $U_n \geq Y_n$ a.s. for all n, then $U_n \geq V_n$ a.s. for all n.

Finally, define

$$\tau := \min\{n : V_n = Y_n\}.$$

Note that $\tau \in \{1, \ldots, N\}$, since there is always at least one n (namely, N) such that $V_n = Y_n$. Note also that τ is a stopping time with respect to $\{\mathcal{F}_n\}_{1\leq n\leq N}$, since

$$\{\tau = n\} = \{V_n = Y_n\} \cap \{V_k \neq Y_k \text{ for all } k < n\}.$$

We now arrive at the main result of this section, which is that τ is the solution of the optimal stopping problem mentioned above.

THEOREM 9.8.3. *The stopping time τ maximizes $\mathbb{E}(X_\tau)$ among all stopping times adapted to the filtration $\{\mathcal{F}_n\}_{1\leq n\leq N}$. Moreover, this maximum value is equal to $\mathbb{E}(V_1)$.*

PROOF. Let T be any other stopping time. By Lemma 9.8.1, it suffices to show that $\mathbb{E}(Y_\tau) \geq \mathbb{E}(Y_T)$. By the definitions of V_n and τ, we have that on the set $\{\tau > n\}$, $V_n = \mathbb{E}(V_{n+1}|\mathcal{F}_n)$ a.s. Therefore, by Proposition 9.7.2, $\{V_{\tau\wedge n}\}_{1\leq n\leq N}$ is a martingale adapted to $\{\mathcal{F}_n\}_{1\leq n\leq N}$. Moreover, $Y_\tau = V_\tau$ by the definition of τ. Consequently,

$$\mathbb{E}(Y_\tau) = \mathbb{E}(V_\tau) = \mathbb{E}(V_{\tau\wedge N}) = \mathbb{E}(V_{\tau\wedge 1}) = \mathbb{E}(V_1).$$

But, since $\{V_n\}_{1\leq n\leq N}$ is a supermartingale, T is a bounded stopping time, and $V_n \geq Y_n$ for all n, the optional stopping theorem implies that

$$\mathbb{E}(V_1) \geq \mathbb{E}(V_T) \geq \mathbb{E}(Y_T).$$

This completes the proof. □

Let us now work out an application of the above method. Suppose that there are $N \geq 2$ candidates appearing for an interview in a certain order. Let r_i be the rank of candidate i among all candidates (with rank 1 being the best). Ideally, the employer would want to hire the candidate with rank 1. The problem is that at any time n, the employer only knows $r_1^n, \ldots, r_n^n$, where r_i^n is the rank of candidate i *among the first n candidates*. That is, the employer has no knowledge about the candidates who are going to appear for the interview at later times. Also, the employer is constrained by the rule that he has to either immediately offer the job to a candidate after the interview, or let them go without the possibility of making an offer later. There are various problems of this type, where you have to make a decision at some point of time based in prior information, but are not allowed to change your decision later.

To inject randomness, let us make the quite reasonable assumption that the $(r_1, \ldots, r_N)$ is uniformly distributed over the set of all permutations of $\{1, \ldots, n\}$. Since the employer knows only $r_1^n, \ldots, r_n^n$ at time n, let $\mathcal{F}_n$ be the σ-algebra generated by this vector. Clearly, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_N$. We want to find a stopping time τ with respect to this filtration that maximizes $\mathbb{P}(r_\tau = 1)$. To put this into the framework of Theorem 9.8.3, let $X_n := 1_{\{r_n=1\}}$, so that $\mathbb{P}(r_T = 1) = \mathbb{E}(X_T)$ for any stopping time T.

First, let us compute $Y_n = \mathbb{E}(X_n|\mathcal{F}_n)$. If $r_n^n \neq 1$, then clearly $r_n \neq 1$. On the other hand, given $r_n = 1$, the remaining r_i's form a uniform random permutation of $2, \ldots, n$. In particular, given $r_n = 1$, all possible rankings of the first $n-1$ candidates are equally likely. Consequently, for any permutation $\sigma_1, \ldots, \sigma_{n-1}$ of $2, \ldots, n$,

$$\mathbb{P}(r_1^n = \sigma_1, \ldots, r_{n-1}^n = \sigma_{n-1}, r_n^n = 1 | r_n = 1) = \frac{1}{(n-1)!}.$$

By Bayes' rule, this gives

$$\begin{aligned}
&\mathbb{P}(r_n = 1 | r_1^n = \sigma_1, \ldots, r_{n-1}^n = \sigma_{n-1}, r_n^n = 1) \\
&= \frac{\mathbb{P}(r_1^n = \sigma_1, \ldots, r_{n-1}^n = \sigma_{n-1}, r_n^n = 1 | r_n = 1)\mathbb{P}(r_n = 1)}{\mathbb{P}(r_1^n = \sigma_1, \ldots, r_{n-1}^n = \sigma_{n-1}, r_n^n = 1)} \\
&= \frac{(1/(n-1)!)(1/N)}{(1/n!)} = \frac{n}{N}.
\end{aligned}$$

The above argument shows that

$$Y_n = \mathbb{E}(X_n|\mathcal{F}_n) = \mathbb{P}(r_n = 1|\mathcal{F}_n) = \frac{n}{N} 1_{\{r_n^n = 1\}}.$$

In other words, $Y_n = n/N$ if candidate n is the best among the first n candidates, and 0 otherwise. The fact that makes it easy to apply Theorem 9.8.3 in this problem is the following.

LEMMA 9.8.4. *For each $2 \leq n \leq N$, Y_n is independent of $\mathcal{F}_{n-1}$. Moreover, $\mathbb{P}(Y_n = n/N) = 1/n$.*

PROOF. Since Y_n can take only two values, 0 and n/N, it suffices to show that the random variable $\mathbb{P}(Y_n = n/N|\mathcal{F}_{n-1})$, which is the same as $\mathbb{P}(r_n^n = 1|\mathcal{F}_n)$, is actually nonrandom. Take any permutation $\sigma_1, \ldots, \sigma_{n-1}$ of $1, \ldots, n-1$. Then

$$\begin{aligned}
&\mathbb{P}(r_n^n = 1 | r_1^{n-1} = \sigma_1, \ldots, r_{n-1}^{n-1} = \sigma_{n-1}) \\
&= \frac{\mathbb{P}(r_n^n = 1, r_1^{n-1} = \sigma_1, \ldots, r_{n-1}^{n-1} = \sigma_{n-1})}{\mathbb{P}(r_1^{n-1} = \sigma_1, \ldots, r_{n-1}^{n-1} = \sigma_{n-1})} \\
&= \frac{\mathbb{P}(r_n^n = 1, r_1^{n-1} = \sigma_1 + 1, \ldots, r_{n-1}^{n-1} = \sigma_{n-1} + 1)}{\mathbb{P}(r_1^{n-1} = \sigma_1, \ldots, r_{n-1}^{n-1} = \sigma_{n-1})} = \frac{1/n!}{1/(n-1)!} = \frac{1}{n}.
\end{aligned}$$

Thus, $\mathbb{P}(Y_n = n/N|\mathcal{F}_{n-1}) \equiv 1/n$. This completes the proof of the lemma. □

Now let $V_1, \ldots, V_N$ be defined by backward induction as in (9.8.1), starting with $V_N = Y_N$. Lemma 9.8.4 yields the following corollary.

COROLLARY 9.8.5. *For $1 \leq n \leq N-1$, $\mathbb{E}(V_{n+1}|\mathcal{F}_n)$ is equal to a nonrandom quantity v_n^N.*

PROOF. We will prove the claim by backward induction on n. By Lemma 9.8.4,

$$\mathbb{E}(V_N|\mathcal{F}_{N-1}) = \mathbb{E}(Y_N|\mathcal{F}_{N-1}) = \mathbb{E}(Y_N) = \frac{1}{N}.$$

So, the claim holds for $n = N-1$. Suppose that it holds for some n. Then

$$\mathbb{E}(V_n|\mathcal{F}_{n-1}) = \mathbb{E}(\max\{Y_n, \mathbb{E}(V_{n+1}|\mathcal{F}_n)\}|\mathcal{F}_{n-1}).$$

But, by the induction hypothesis, $\mathbb{E}(V_{n+1}|\mathcal{F}_n)$ is equal to some nonrandom quantity v_n^N. Thus,

$$\mathbb{E}(V_n|\mathcal{F}_{n-1}) = \mathbb{E}(\max\{Y_n, v_n^N\}|\mathcal{F}_{n-1}).$$

But by Lemma 9.8.4, Y_n is independent of $\mathcal{F}_{n-1}$, and therefore, so is $\max\{Y_n, v_n^N\}$. Thus, we get that $v_{n-1}^N := \mathbb{E}(V_n|\mathcal{F}_{n-1})$ is nonrandom. □

Let us now evaluate the quantities $v_1^N, \ldots, v_{N-1}^N$. Already in the proof of Corollary 9.8.5, we have seen that $v_{N-1}^N = 1/N$, and also that for each $n \le N-1$,

$$v_{n-1}^N = \mathbb{E}(\max\{Y_n, v_n^N\}). \tag{9.8.2}$$

This shows that $v_1^N \ge v_2^N \ge \cdots \ge v_{N-1}^N = 1/N$. Let

$$t_N := \min\{n \le N-1 : n/N \ge v_n^N\},$$

which is well defined since $v_{N-1}^N = 1/N \le (N-1)/N$ (because $N \ge 2$). By the decreasing nature of v_n^N, it follows that $n/N < v_n^N$ for all $n < t_N$ and $n/N \ge v_n^N$ for all $n \ge t_N$. In particular, for any $n < t_N$, $Y_n \le n/N < v_n^N$, and therefore, (9.8.2) shows that $v_{n-1}^N = v_n^N$. On the other hand, for $t_N \le n \le N-1$, (9.8.2) and Lemma 9.8.4 imply that

$$\begin{aligned} v_{n-1}^N &= \frac{n}{N}\mathbb{P}(Y_n = n/N) + v_n^N \mathbb{P}(Y_n = 0) \\ &= \frac{1}{N} + \left(1 - \frac{1}{n}\right) v_n^N. \end{aligned}$$

Using this, and the fact that $v_{N-1}^N = 1/N$, it is now easy to prove by backward induction that for $t_N - 1 \le n \le N-1$,

$$v_n^N = \frac{1}{N} + \frac{n}{N} \sum_{k=n+1}^{N-1} \frac{1}{k}. \tag{9.8.3}$$

Thus, t_N is the unique integer $n \in \{1, \ldots, N-1\}$ such that

$$\frac{1}{N} + \frac{n-1}{N} \sum_{k=n}^{N-1} \frac{1}{k} > \frac{n}{N} \ge \frac{1}{N} + \frac{n}{N} \sum_{k=n+1}^{N-1} \frac{1}{k}.$$

It is easy to see from these inequalities that asymptotically as $N \to \infty$, $t_N \sim N/e$. By Theorem 9.8.3, we now deduce that the optimal stopping rule τ in this problem is given by

$$\tau = \min\{n \ge t_N : Y_n = n/N\} = \min\{n \ge t_N : r_n^n = 1\}.$$

That is, we should choose the first candidate since time t_N who is better than everyone who came before him. Moreover, with this optimal rule, the probability of choosing the best candidate is v_1^N, which is equal to $v_{t_N-1}^N$. By the characterization of t_N given above, and the formula (9.8.3), this value is asymptotically $1/e$.

Exercise 9.8.6. Let $X_1, X_2, \ldots, X_N$ be i.i.d. integrable random variables. Let $K > 0$ be a given constant, and let $S_n := X_1 + \cdots + X_n$ and $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Consider all stopping times T for the filtration $\{\mathcal{F}_n\}_{1 \le n \le N}$, which are allowed to take values in $\{1, \ldots, N\} \cup \{\infty\}$. Stopping at time T gets us a reward $S_T - K$ if T is finite, and 0 if $T = \infty$. We want to maximize the expected reward. (This comes from 'American options',

where an asset has price S_n dollars on day n, and the holder of the asset has to pay a fee of K dollars to sell the asset at any time before time N, which is the 'expiration date' for the option. The holder has no obligation to sell, and if the option is not exercised by time N, then the reward is zero — this is the case $T = \infty$. The 'sum of i.i.d.' is a toy model. In reality, more complex models are used.) Prove the following:

(1) Show that the problem can be reformulated as the problem of finding a stopping time T taking values in $\{1, \ldots, N\}$ that maximizes $\mathbb{E}((S_T - K)^+)$. Having found such a T, what would be an optimal strategy for the original problem?

(2) Define a sequence of functions $w_0, w_1, \ldots$ from $\mathbb{R}$ into $\mathbb{R}$ as follows: Let $w_0(x) := (x - K)^+$, and for each $n \geq 1$, let

$$w_n(x) := \max\{(x - K)^+, \mathbb{E}(w_{n-1}(x + X_1))\}.$$

Finally, let $T := \min\{1 \leq n \leq N : (S_n - K)^+ = w_{N-n}(S_n)\}$. Show that T is an optimal stopping time for the revised problem described in part (1).

(3) Show that each w_n is a convex nondecreasing function.

9.9. Almost sure convergence of martingales

Martingales, submartingales and supermartingales often have nice convergence properties. In this section, we will derive a criterion for almost sure convergence due to Doob. A key ingredient in the proof is Doob's upcrossing lemma. If $\{X_n\}_{n\geq 0}$ is any sequence of random variables, and $[a, b]$ is a bounded interval, the upcrossings of $[a, b]$ by $\{X_n\}_{n\geq 0}$ are defined as follows. Let S_1 be the first n such that $X_n \leq a$. Let T_1 be the first n after S_1 such that $X_n \geq b$. Let S_2 be the first n after T_1 such that $X_n \leq a$, and T_2 be the first n after S_2 such that $X_n \geq b$. And we keep going like this. That is, taking $T_0 = -1$, we have that for all $k \geq 1$,

$$S_k = \inf\{n > T_{k-1} : X_n \leq a\},$$
$$T_k = \inf\{n > S_k : X_n \geq b\}.$$

In the above, we follow the convention that the infimum of an empty set is infinity. The intervals $\{S_k, S_k + 1, \ldots, T_k\}$, for all k such that S_k is finite, are called the upcrossings of $[a, b]$ by $\{X_n\}_{n\geq 0}$.

LEMMA 9.9.1 (Upcrossing lemma). *Let $\{X_n\}_{n\geq 0}$ be a submartingale sequence adapted to some filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Take any interval $[a, b]$ with $a < b$. Let U_m be the number of upcrossings of this interval by the sequence $\{X_n\}_{n\geq 0}$ that are completed by time m. Then*

$$\mathbb{E}(U_m) \leq \frac{\mathbb{E}(X_m - a)^+ - \mathbb{E}(X_0 - a)^+}{b - a}.$$

PROOF. For each n, let $Y_n := a + (X_n - a)^+$. That is, $Y_n = a$ if $X_n \leq a$ and $Y_n = X_n$ if $X_n > a$. Note that $x \mapsto a + (x - a)^+$ is a convex non-decreasing function. Thus, $\{Y_n\}_{n\geq 0}$ is also a submartingale. Moreover, it has the same upcrossings of $[a, b]$ as the sequence $\{X_n\}_{n\geq 0}$. Note that that if $k, \ldots, k + l$ is an upcrossing, then $Y_{k+l} - Y_k \geq b - a$, and moreover, for any $k \leq j \leq k + l$, $Y_j - Y_k \geq 0$ (the second property is not valid for the process $\{X_n\}_{n\geq 0}$). Thus, if we add up $Y_{k+l} - Y_k$ for all upcrossings $k, \ldots, k + l$ that are

completed by time m, and also $Y_m - Y_k$ for an incomplete upcrossing $k, \ldots, m$ at the end, the sum will be at least $(b-a)U_m$.

Let us now represent the above sum in a different way. For each $n \geq 1$, let Z_n be the indicator of the event that the indices n and $n-1$ are both in the same upcrossing. The crucial observation is that Z_n is $\mathcal{F}_{n-1}$-measurable. To see this, note that there are four possibilities: (a) $Y_{n-1} \leq a$. In this case, $n-1$ is part of an upcrossing (since the process is always above a between upcrossings), and n is also a part of the same upcrossing. (b) $a < Y_{n-1} < b$, and $n-1$ is part of an upcrossing. In this case, n is part of the same upcrossing. (c) $a < Y_{n-1} < b$, but $n-1$ is not in an upcrossing. In this case, obviously, n and $n-1$ are not both in the same upcrossing. (d) $Y_{n-1} \geq b$. In this case, n and $n-1$ cannot be in the same upcrossing. Thus, whether $Z_n = 1$ or 0 is completely determined by $Y_0, \ldots, Y_{n-1}$, and hence, Z_n is $\mathcal{F}_{n-1}$-measurable. Now, the sum defined in the previous paragraph can be written as

$$\sum_{n=1}^{m} Z_n(Y_n - Y_{n-1}).$$

Since Z_n is $\mathcal{F}_{n-1}$-measurable and $\{0,1\}$-valued, and $\{Y_n\}_{n\geq 0}$ is a submartingale (which implies that $\mathbb{E}(Y_n|\mathcal{F}_{n-1}) - Y_{n-1} \geq 0$), we have

$$\begin{aligned}\mathbb{E}(Z_n(Y_n - Y_{n-1})) &= \mathbb{E}[\mathbb{E}(Z_n(Y_n - Y_{n-1})|\mathcal{F}_{n-1})] \\ &= \mathbb{E}[Z_n(\mathbb{E}(Y_n|\mathcal{F}_{n-1}) - Y_{n-1})] \\ &\leq \mathbb{E}[\mathbb{E}(Y_n|\mathcal{F}_{n-1}) - Y_{n-1}] = \mathbb{E}(Y_n - Y_{n-1}).\end{aligned}$$

Thus,

$$(b-a)\mathbb{E}(U_m) \leq \sum_{n=1}^{m} \mathbb{E}(Y_n - Y_{n-1}) = \mathbb{E}(Y_m - Y_0).$$

Since $Y_n = a + (X_n - a)^+$, this completes the proof. □

Using the upcrossing lemma, we will now prove the following important result.

THEOREM 9.9.2 (Submartingale convergence theorem). *Let $\{X_n\}_{n\geq 0}$ be a submartingale adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Suppose that $\sup_{n\geq 0} \mathbb{E}(X_n^+)$ is finite. Then there is a random variable X such that $X_n \to X$ a.s. and $\mathbb{E}|X| < \infty$.*

PROOF. Take any interval $[a,b]$ with $a < b$. Let U_m be the number of upcrossings of this interval by the sequence $\{X_n\}_{n\geq 0}$ that are completed by time m. Then $\{U_m\}_{m\geq 0}$ is an increasing sequence of random variables, whose limit U is the total number of upcrossings of this interval. But by the upcrossing lemma and the condition that $\sup_{n\geq 0} \mathbb{E}(X_n^+) < \infty$, we get that $\mathbb{E}(U_m)$ is uniformly bounded. Thus by the monotone convergence theorem, $\mathbb{E}(U) < \infty$ and hence $U < \infty$ a.s. But this is true for any $[a,b]$, and so it holds almost surely for all intervals with rational endpoints. This implies that $\liminf_{n\to\infty} X_n$ cannot be strictly less than $\limsup_{n\to\infty} X_n$, because otherwise there would exist an interval with rational endpoints which is upcrossed infinitely many times. This proves the existence of the limit X.

By Fatou's lemma, $\mathbb{E}(X^+) \leq \liminf_{n\to\infty} \mathbb{E}(X_n^+) < \infty$. On the other hand, since $\{X_n\}_{n\geq 0}$ is a submartingale sequence,

$$\mathbb{E}(X_n^-) = \mathbb{E}(X_n^+) - \mathbb{E}(X_n) \leq \mathbb{E}(X_n^+) - \mathbb{E}(X_0),$$

which proves the uniform boundedness of $\mathbb{E}(X_n^-)$. Thus again by Fatou's lemma, $\mathbb{E}(X^-) < \infty$ and hence $\mathbb{E}|X| < \infty$. □

EXERCISE 9.9.3. Prove that a nonnegative supermartingale converges almost surely to an integrable random variable.

EXERCISE 9.9.4 (Pólya's urn). Consider an urn which initially has one black ball and one white ball. At each turn, a ball is picked uniformly at random from the urn, and replaced back into the urn with an additional ball of the same color. Let W_n be the fraction of white balls at time n (so that $W_0 = 1/2$). Prove that $\{W_n\}_{n\geq 0}$ is a martingale with respect to a suitable filtration, and show that it converges almost surely to a limit. Show that the limit is uniformly distributed in $[0, 1]$. Hint: For the last part, prove by induction that each W_n is uniformly distributed on a certain set depending on n.

EXERCISE 9.9.5. Show by a counterexample that the convergence in the submartingale convergence theorem need not hold in L^1. Hint: Consider the martingale $\{S_{T\wedge n}\}_{n\geq 0}$, where $\{S_n\}_{n\geq 0}$ is a simple symmetric random walk starting at 0, and T is the first time to hit 1.

9.10. Lévy's downwards convergence theorem

A second consequence of the upcrossing lemma is the backwards martingale convergence theorem, stated below. This is also known as Lévy's downwards convergence theorem.

THEOREM 9.10.1 (Backwards martingale convergence theorem, or, Lévy's downwards convergence theorem). *Let X be an integrable random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_0 \supseteq \mathcal{F}_1 \supseteq \cdots$ be a decreasing sequence of sub-σ-algebras of $\mathcal{F}$. Let $\mathcal{F}^* = \cap_{n\geq 0}\mathcal{F}_n$. Then $\mathbb{E}(X|\mathcal{F}_n)$ converges to $\mathbb{E}(X|\mathcal{F}^*)$ a.s. and in L^1 as $n \to \infty$.*

PROOF. Let $X_n := \mathbb{E}(X|\mathcal{F}_n)$. Take any interval $[a, b]$ with $a < b$. For each n, let U_n be the number of complete upcrossings of this interval by the finite sequence $X_n, X_{n-1}, \ldots, X_0$. Note that this sequence is a martingale with respect to the filtration $\mathcal{F}_n, \mathcal{F}_{n-1}, \ldots, \mathcal{F}_0$. Therefore by the upcrossing lemma,

$$\mathbb{E}(U_n) \leq \frac{\mathbb{E}(X_0 - a)^+}{b - a}.$$

Since $X_0 = \mathbb{E}(X|\mathcal{F}_0)$ is integrable, the quantity on the right is finite. Moreover, it does not depend on n. Thus, $\sup_{n\geq 0} \mathbb{E}(U_n) < \infty$. But U_n increases to a limit as $n \to \infty$, and if this limit is finite, then the sequence $\{X_n\}_{n\geq 0}$ cannot attain infinitely many values in both the intervals $(-\infty, a]$ and $[b, \infty)$. Therefore, with probability one, this cannot happen for any interval with rational endpoints. This implies that X_n converges almost surely to a limit X^*.

The next step is to show that the sequence $\{X_n\}_{n\geq 0}$ is uniformly integrable. Since

$$|X_n| = |\mathbb{E}(X|\mathcal{F}_n)| \leq \mathbb{E}(|X||\mathcal{F}_n),$$

we have

$$\begin{aligned}\mathbb{E}(|X_n|;|X_n|>K) &\leq \mathbb{E}(\mathbb{E}(|X||\mathcal{F}_n);\mathbb{E}(|X||\mathcal{F}_n)>K)\\ &= \mathbb{E}(|X|;\mathbb{E}(|X||\mathcal{F}_n)>K). \qquad (9.10.1)\end{aligned}$$

Take any $\epsilon > 0$. By Corollary 8.3.4, there is some $\delta > 0$ such that for any event A with $\mathbb{P}(A) < \delta$, we have $\mathbb{E}(|X|;A) < \epsilon$. Fixing K, let A_n be the event $\{\mathbb{E}(|X||\mathcal{F}_n) > K\}$. By Markov's inequality,

$$\mathbb{P}(A_n) \leq \frac{\mathbb{E}(\mathbb{E}(|X||\mathcal{F}_n))}{K} = \frac{\mathbb{E}|X|}{K},$$

which can be made less than δ by choosing K large enough. Therefore, by (9.10.1),

$$\mathbb{E}(|X_n|;|X_n|>K) \leq \mathbb{E}(|X|;A_n) \leq \epsilon.$$

This proves the uniform integrability of $\{X_n\}_{n\geq 0}$. By Proposition 8.3.1, we conclude that $X_n \to X^*$ in L^1. This implies that for any $B \in \mathcal{F}^*$,

$$\mathbb{E}(X^*;B) = \lim_{n\to\infty} \mathbb{E}(X_n;B) = \mathbb{E}(X;B),$$

where the last identity holds because $X_n = \mathbb{E}(X|\mathcal{F}_n)$ and $B \in \mathcal{F}_n$ for every n. So, to complete the proof we only have to show that X^* is $\mathcal{F}^*$-measurable. To show this, take any n. Since $\mathcal{F}_m \subseteq \mathcal{F}_n$ for any $m \geq n$, we have that X_m is $\mathcal{F}_n$-measurable for any $m \geq n$. Therefore X^* is $\mathcal{F}_n$-measurable. But since this is true for all n, X^* must be $\mathcal{F}^*$-measurable. □

EXERCISE 9.10.2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. integrable random variables. Let $S_n := X_1 + \cdots + X_n$, and $\mathcal{G}_n := \sigma(S_n, S_{n+1}, S_{n+2}, \ldots)$. Using Exercise 9.2.13, show that $\mathbb{E}(X_1|\mathcal{G}_n) = \mathbb{E}(X_1|S_n) = \mathbb{E}(X_i|S_n)$ for any $i \leq n$. From this, deduce that $\mathbb{E}(X_1|\mathcal{G}_n) = S_n/n$. Using this fact, the downwards convergence theorem, and Kolmogorov's zero-one law, show that S_n/n converges a.s. and in L^1 to $\mathbb{E}(X_1)$ as $n \to \infty$.

9.11. De Finetti's theorem

Let $\{X_n\}_{n\geq 1}$ be an infinite sequence of real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} := \sigma(X_1, X_2, \ldots)$. By Exercise 5.1.6, any $A \in \mathcal{G}$ can be expressed as $X^{-1}(B)$ for some measurable set $B \subseteq \mathbb{R}^{\mathbb{N}}$, where $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ is the map $X(\omega) = (X_n(\omega))_{n\geq 1}$. We will say that A is invariant under permutations of the first n coordinates if B is invariant under permutations of the first n coordinates (for some choice of B such that $A = X^{-1}(B)$) — that is, for any $(x_1, x_2, \ldots) \in B$, and any permutation π of $1, \ldots, n$, $(x_{\pi(1)}, \ldots, x_{\pi(n)}, x_{n+1}, x_{n+2}, \ldots)$ is also in B. Let $\mathcal{E}_n$ be the σ-algebra consisting of all $A \in \mathcal{G}$ that are invariant under permutations of the first n coordinates (it is easy to see that this is a σ-algebra). Let

$$\mathcal{E} := \bigcap_{n=1}^{\infty} \mathcal{E}_n$$

be the σ-algebra generated by all X that are invariant under permutations of any finitely many coordinates. This is known as the 'exchangeable σ-algebra' of the sequence $\{X_n\}_{n\geq 1}$.

EXERCISE 9.11.1 (Hewitt–Savage zero-one law, alternative version). Recall the Hewitt–Savage zero-one law proved in Chapter 7 (Corollary 7.5.4). Prove the following equivalent

version. If $\{X_n\}_{n\geq 0}$ is a sequence of i.i.d. random variables, show that the exchangeable σ-algebra $\mathcal{E}$ is trivial (meaning that for any $A \in \mathcal{E}$, $\mathbb{P}(A)$ is either 0 or 1).

An infinite sequence of random variables $\{X_n\}_{n\geq 1}$ is called 'exchangeable' if its joint distribution remains invariant under permutations of any finitely many coordinates. To be more precise, the condition means that for any bijection $\pi : \mathbb{N} \to \mathbb{N}$ that fixes all but finitely many coordinates, the sequences $(X_{\pi(1)}, X_{\pi(2)}, \ldots)$ and $(X_1, X_2, \ldots)$ induce the same probability measure on $\mathbb{R}^{\mathbb{N}}$.

The following fundamental result for infinite exchangeable sequences is called De Finetti's theorem.

THEOREM 9.11.2 (De Finetti's theorem). *Let $\{X_n\}_{n\geq 1}$ be an infinite exchangeable sequence of real-valued random variables. Then they are independent and identically distributed conditional on the exchangeable σ-algebra $\mathcal{E}$, in the sense that for any k and any Borel sets $A_1, \ldots, A_k \subseteq \mathbb{R}$,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_k \in A_k | \mathcal{E}) = \prod_{i=1}^{k} \mathbb{P}(X_i \in A_i | \mathcal{E}) = \prod_{i=1}^{k} \mathbb{P}(X_1 \in A_i | \mathcal{E}) \quad a.s.$$

We need several lemmas to prove this theorem.

LEMMA 9.11.3. *If a measurable function $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is invariant under permutations of the first n coordinates, then $f(X_1, X_2, \ldots)$ is an $\mathcal{E}_n$-measurable random variable.*

PROOF. For any Borel set $A \subseteq \mathbb{R}$, it is easy to check that the set

$$\{\omega : f(X_1(\omega), X_2(\omega), \ldots) \in A\}$$

is in $\mathcal{E}_n$, because it is equal to $X^{-1}(B)$, where $B = f^{-1}(A)$, and B is invariant under permutations of the first n coordinates. □

LEMMA 9.11.4. *For any n, any bounded measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}$, and any permutation π of $1, \ldots, n$,*

$$\mathbb{E}(\varphi(X_{\pi(1)}, \ldots, X_{\pi(n)}) | \mathcal{E}_n) = \mathbb{E}(\varphi(X_1, \ldots, X_n) | \mathcal{E}_n) \quad a.s.$$

PROOF. Let Y and Z denote the two conditional expectations displayed above, that have to be shown to be equal. Take any $A \in \mathcal{E}_n$, so that $A = X^{-1}(B)$ for some measurable set $B \subseteq \mathbb{R}^{\mathbb{N}}$ that is invariant under permutations of the first n coordinates. Let $f := 1_B$. Then

$$\begin{aligned}
&\mathbb{E}(Y; A) = \mathbb{E}(\varphi(X_{\pi(1)}, \ldots, X_{\pi(n)}) f(X_1, X_2, \ldots)) \\
&= \mathbb{E}(\varphi(X_{\pi(1)}, \ldots, X_{\pi(n)}) f(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}, X_{n+1}, X_{n+2}, \ldots)) \\
&= \mathbb{E}(\varphi(X_1, \ldots, X_n) f(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots)) \\
&= \mathbb{E}(Z; A),
\end{aligned}$$

where the second identity holds because f is invariant under permutations of the first n coordinates, and the third identity holds because $X_1, X_2, \ldots$ are exchangeable. □

LEMMA 9.11.5. *For any bounded measurable function* $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \to \mathbb{E}(\varphi(X_1)|\mathcal{E})$$

a.s. as $n \to \infty$.

PROOF. By the downwards convergence theorem,

$$\mathbb{E}(\varphi(X_1)|\mathcal{E}_n) \to \mathbb{E}(\varphi(X_1)|\mathcal{E})$$

a.s. as $n \to \infty$. By Lemma 9.11.4, $\mathbb{E}(\varphi(X_1)|\mathcal{E}_n) = \mathbb{E}(\varphi(X_i)|\mathcal{E}_n)$ a.s. for any $1 \le i \le n$. Thus,

$$\begin{aligned}\mathbb{E}(\varphi(X_1)|\mathcal{E}_n) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\varphi(X_i)|\mathcal{E}_n) \quad \text{a.s.} \\ &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)\middle|\mathcal{E}_n\right) \quad \text{a.s.}\end{aligned}$$

But by Lemma 9.11.3, $n^{-1}\sum_{i=1}^{n} \varphi(X_i)$ is $\mathcal{E}_n$-measurable. Thus,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)\middle|\mathcal{E}_n\right) = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \quad \text{a.s.}$$

This completes the proof of the lemma. □

LEMMA 9.11.6. *For any* k, *and any bounded measurable functions* $\varphi_1, \ldots, \varphi_k : \mathbb{R}^k \to \mathbb{R}$,

$$\mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E}_n) - \prod_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} \varphi_i(X_j)\right) \to 0$$

a.s. as $n \to \infty$.

PROOF. Take any $n \ge k$. By Lemma 9.11.4,

$$\mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E}_n) = \mathbb{E}(\varphi_1(X_{i_1})\cdots\varphi_k(X_{i_k})|\mathcal{E}_n) \quad \text{a.s.} \tag{9.11.1}$$

for any distinct $i_1, \ldots, i_k \in \{1, \ldots, n\}$. Let $(n)_k := n(n-1)\cdots(n-k+1)$, and define

$$Y_n := \frac{1}{(n)_k} \sum_{\substack{1 \le i_1, \ldots, i_k \le n \\ \text{distinct}}} \varphi_1(X_{i_1})\cdots\varphi_k(X_{i_k})$$

and

$$Z_n := \frac{1}{n^k} \sum_{1 \le i_1, \ldots, i_k \le n} \varphi_1(X_{i_1})\cdots\varphi_k(X_{i_k}) = \prod_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} \varphi_i(X_j)\right).$$

Since the φ_i's are bounded functions, a simple calculation shows that $Y_n - Z_n \to 0$ as $n \to \infty$. By equation (9.11.1), $\mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E}_n) = \mathbb{E}(Y_n|\mathcal{E}_n)$, and by Lemma 9.11.3, Y_n is $\mathcal{E}_n$-measurable. Combining these observations completes the proof. □

Finally, we are ready to prove De Finetti's theorem.

PROOF OF THEOREM 9.11.2. Take any k, and any bounded measurable functions $\varphi_1, \ldots, \varphi_k : \mathbb{R}^k \to \mathbb{R}$. By the downwards convergence theorem,

$$\mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E}_n) \to \mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E})$$

a.s. as $n \to \infty$. By Lemma 9.11.5,

$$\prod_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n}\varphi_i(X_j)\right) \to \prod_{i=1}^{k}\mathbb{E}(\varphi_i(X_1)|\mathcal{E})$$

a.s. as $n \to \infty$. The above two displays, together with Lemma 9.11.6, give

$$\mathbb{E}(\varphi_1(X_1)\cdots\varphi_k(X_k)|\mathcal{E}) = \prod_{i=1}^{k}\mathbb{E}(\varphi_i(X_1)|\mathcal{E}) \text{ a.s.}$$

As a special case of the above identity, we get $\mathbb{E}(\varphi_i(X_i)|\mathcal{E}) = \mathbb{E}(\varphi_i(X_1))$ a.s. for each i. This completes the proof. □

EXERCISE 9.11.7. Let $X_1, X_2, \ldots$ be an exchangeable infinite sequence of $\{0,1\}$-valued random variables. Prove that:

(1) $n^{-1}\sum_{i=1}^{n} X_i$ converges a.s. to a limit Θ as $n \to \infty$.
(2) Let μ be the law of Θ. Then for any n and any $x_1, \ldots, x_n \in \{0,1\}$, show that

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_{[0,1]} \theta^k(1-\theta)^{n-k}d\mu(\theta),$$

where $k = \sum_{i=1}^{n} x_i$. In other words, the law of the infinite sequence $(X_1, X_2, \ldots)$ is a mixture of i.i.d. distributions. (This kind of result is often cited as justification for Bayesian modeling.)

9.12. Lévy's upwards convergence theorem

The following, result, known as Lévy's upwards convergence theorem, complements the downwards convergence theorem from Section 9.10.

THEOREM 9.12.1 (Lévy upwards convergence theorem). *Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration of σ-algebras and let $\mathcal{F}^* := \sigma(\cup_{n=0}^{\infty}\mathcal{F}_n)$. Then for any integrable random variable X, $\mathbb{E}(X|\mathcal{F}_n) \to \mathbb{E}(X|\mathcal{F}^*)$ a.s. and in L^1.*

PROOF. Let $X_n := \mathbb{E}(X|\mathcal{F}_n)$. Note that the sequence $\{X_n\}_{n\geq 0}$ is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Moreover, by Jensen's inequality for conditional expectation, $\mathbb{E}|X_n| \leq \mathbb{E}|X|$ for all n. Therefore by Theorem 9.9.2, there is an integrable random variable Y such that $X_n \to Y$ a.s. By a similar argument as in the proof of the backward martingale convergence theorem, we see that $\{X_n\}_{n\geq 0}$ is uniformly integrable. Thus, $X_n \to Y$ in L^1. Finally, we need to show that $Y = \mathbb{E}(X|\mathcal{F}^*)$ a.s. To prove this, take any $A \in \cup_{n\geq 1}\mathcal{F}_n$. Then $A \in \mathcal{F}_n$ for some n, and so

$$\mathbb{E}(X;A) = \mathbb{E}(X_n;A).$$

But $X_n = \mathbb{E}(X_m|\mathcal{F}_n)$ for any $m \geq n$. Thus, $\mathbb{E}(X;A) = \mathbb{E}(X_m;A)$ for any $m \geq n$. Since $X_m \to Y$ in L^1 as $m \to \infty$, this shows that $\mathbb{E}(X;A) = \mathbb{E}(Y;A)$. It is now a simple exercise using the π-λ theorem to show that $\mathbb{E}(X;A) = \mathbb{E}(Y;A)$ for all $A \in \mathcal{F}^*$, which proves that $Y = \mathbb{E}(X|\mathcal{F}^*)$. □

EXERCISE 9.12.2. Let X be a Borel measurable, integrable map from $[0,1)$ into $\mathbb{R}$. Let $X_n : [0,1) \to \mathbb{R}$ be defined as follows. For each n and $0 \leq i < 2^n$, and for each x in the

interval $[2^{-n}i, 2^{-n}(i+1))$, let $X_n(x)$ be the average value of X in that interval. Show that $X_n \to X$ pointwise almost everywhere on $[0,1)$. Hint: Use Exercise 9.1.11 and Exercise 9.4.4, along with the martingale convergence theorem.

9.13. L^p convergence of martingales

In this section we discuss the necessary and sufficient conditions for L^p convergence of martingales, for all $p \geq 1$. The cases $p > 1$ and $p = 1$ are somewhat different. Let us first discuss $p > 1$. An important ingredient in the L^p convergence theorem is the following inequality, known as Doob's maximal inequality for submartingales.

THEOREM 9.13.1 (Doob's maximal inequality). *Let $\{X_i\}_{0\leq i\leq n}$ be a submartingale sequence adapted to a filtration $\{\mathcal{F}_i\}_{0\leq i\leq n}$. Let $M_n := \max_{0\leq i\leq n} X_i$. Then for any $t > 0$,*

$$\mathbb{P}(M_n \geq t) \leq \frac{\mathbb{E}(X_n; M_n \geq t)}{t}.$$

PROOF. For each $0 \leq i \leq n$, let A_i be the event that $X_j < t$ for $j < i$ and $X_i \geq t$. Then the event $\{M_n \geq t\}$ is the union of the disjoint events $A_0, \ldots, A_n$. Thus,

$$\mathbb{E}(X_n; M_n \geq t) = \sum_{i=0}^{n} \mathbb{E}(X_n; A_i).$$

But A_i is $\mathcal{F}_i$-measurable and $\mathbb{E}(X_n|\mathcal{F}_i) = X_i$. Thus,

$$\mathbb{E}(X_n; M_n \geq t) = \sum_{i=0}^{n} \mathbb{E}(X_i; A_i).$$

But if A_i happens, then $X_i \geq t$. Thus,

$$\mathbb{E}(X_n; M_n \geq t) \geq t \sum_{i=0}^{n} \mathbb{P}(A_i) = t\mathbb{P}(M_n \geq t),$$

which is what we wanted. □

Doob's maximal inequality implies the following inequality, known as Doob's L^p inequality.

THEOREM 9.13.2 (Doob's L^p inequality). *Let $\{X_i\}_{0\leq i\leq n}$ be a martingale or nonnegative submartingale adapted to a filtration $\{\mathcal{F}_i\}_{0\leq i\leq n}$. Let $M_n := \max_{0\leq i\leq n} |X_i|$. Then for any $p > 1$,*

$$\mathbb{E}(M_n^p) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}|X_n|^p.$$

PROOF. Under the given assumptions, $\{|X_i|\}_{0\leq i\leq n}$ is a submartingale sequence. Therefore by Doob's maximal inequality,

$$\mathbb{P}(M_n \geq t) \leq \frac{\mathbb{E}(|X_n|; M_n \geq t)}{t}$$

for any $t > 0$. Therefore, by Exercise 6.3.7 and Fubini's theorem,

$$\begin{aligned}\mathbb{E}(M_n^p) &= \int_0^\infty pt^{p-1}\mathbb{P}(M_n \geq t)dt \\ &\leq \int_0^\infty pt^{p-2}\mathbb{E}(|X_n|; M_n \geq t)dt \\ &= \mathbb{E}\left(|X_n| \int_0^\infty pt^{p-2} 1_{\{M_n \geq t\}} dt\right) \\ &= \frac{p}{p-1}\mathbb{E}(|X_n| M_n^{p-1}).\end{aligned}$$

By Hölder's inequality,

$$\mathbb{E}(|X_n| M_n^{p-1}) \leq (\mathbb{E}|X_n|^p)^{1/p}(\mathbb{E}(M_n^p))^{(p-1)/p}.$$

Putting together the above two displays and rearranging, we get the desired inequality. □

Doob's L^p inequality implies the following necessary and sufficient condition for L^p convergence of martingales.

THEOREM 9.13.3. *Let $\{X_n\}_{n\geq 0}$ be a martingale or nonnegative submartingale adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Suppose that $\sup_{n\geq 0}\mathbb{E}|X_n|^p < \infty$ for some $p > 1$. Then there is a random variable X defined on the same probability space such that $X_n \to X$ in L^p and also almost surely. Conversely, if X_n converges the some X in L^p, then $\sup_{n\geq 0}\mathbb{E}|X_n|^p < \infty$.*

PROOF. The converse implication is trivial, so let us prove the forward direction only. Since the L^p norm of X_n is uniformly bounded, so is the L^1 norm. Therefore, by the submartingale convergence theorem, there exists a random variable X such that $X_n \to X$ a.s. By Fatou's lemma,

$$\mathbb{E}|X|^p = \mathbb{E}(\liminf_{n\to\infty} |X_n|^p) \leq \liminf_{n\to\infty} \mathbb{E}|X_n|^p < \infty.$$

Thus, $X \in L^p$. For each n, let

$$Z_n := \sup_{0\leq m\leq n} |X_m - X|^p,$$

and

$$Z := \lim_{n\to\infty} Z_n = \sup_{n\geq 0} |X_n - X|^p.$$

By Jensen's inequality,

$$|X_n - X|^p = 2^p \left|\frac{X_n - X}{2}\right|^p \leq 2^p \frac{|X_n|^p + |X|^p}{2}.$$

Thus, by Doob's L^p inequality (Theorem 9.13.2),

$$\mathbb{E}(Z_n) \leq \left(\frac{p}{p-1}\right)^p 2^{p-1}(\mathbb{E}|X_n|^p + \mathbb{E}|X|^p).$$

But, by the monotone convergence theorem, $\mathbb{E}(Z) = \lim_{n\to\infty}\mathbb{E}(Z_n)$, and so, $\mathbb{E}(Z) < \infty$. But Z dominates $|X_n - X|^p$ for each n, and $|X_n - X|^p \to 0$ a.s. Thus, by the dominated convergence theorem, we get the desired result. □

EXERCISE 9.13.4. Let $\{X_n\}_{n\geq 0}$ be a martingale sequence that is uniformly bounded in L^2. Show that for any $n \leq m$,

$$\mathbb{E}(X_n - X_m)^2 = \sum_{i=n}^{m-1} \mathbb{E}(X_i - X_{i+1})^2,$$

and using this, give a direct proof that X_n converges in L^2 to some limit random variable.

For the necessary and sufficient condition for L^1 convergence of martingales, recall the definition of uniform integrability from Section 8.3 of Chapter 8.

THEOREM 9.13.5. *A submartingale converges in L^1 if and only if it is uniformly integrable, and in this situation, it also converges a.s. to the same limit.*

PROOF. Suppose that $\{X_n\}_{n\geq 0}$ is a uniformly integrable submartingale. By the definition of uniform integrability, it is obvious that $\sup_{n\geq 1} \mathbb{E}|X_n| < \infty$. Therefore, by the submartingale convergence theorem, there exists X such that $X_n \to X$ a.s. But then by Proposition 8.3.1, it follows that $X_n \to X$ in L^1. Conversely, suppose that $\{X_n\}_{n\geq 0}$ is a submartingale converging in L^1 to some limit X. Then by Proposition 8.3.5, $\{X_n\}_{n\geq 0}$ is uniformly integrable. □

EXERCISE 9.13.6. Let $\{X_n\}_{n\geq 0}$ be a martingale or a nonnegative submartingale. If $\sup_{n\geq 0} \mathbb{E}|X_n|^p < \infty$ for some $p > 1$, show that the sequence $\{|X_n|^p\}_{n\geq 0}$ is uniformly integrable.

EXERCISE 9.13.7. Let $\{X_n\}_{n\geq 0}$ be a square-integrable martingale. Define

$$\langle X\rangle_n := \sum_{i=1}^{n} \mathbb{E}((X_i - X_{i-1})^2|\mathcal{F}_{i-1}).$$

Show that the sequence $\{\langle X\rangle_n\}_{n\geq 0}$ is a nonnegative, increasing, and predictable. (It is called the 'predictable quadratic variation' of X_n.) Show that $X_n^2 - \langle X\rangle_n$ is a martingale. Finally, let $\langle X\rangle_\infty := \lim_{n\to\infty}\langle X\rangle_n$, and show that X_n converges a.s. on the set $\{\langle X\rangle_\infty < \infty\}$.

EXERCISE 9.13.8 (Galton–Watson branching process). The Galton–Watson branching process is defined as follows. At generation 0, there is a single organism, which gives birth to a random number of offspring, which all belong to generation 1. Then, each organism in generation 1 independently gives birth to a random number of offspring, which are the members of generation 2, and so on. The assumptions are that the individuals give birth independently, and the distribution of the number of offspring per individual is the same throughout. Mathematically, if Z_n denotes the number of individuals in generation n, then

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i},$$

where $X_{n,i}$ are i.i.d. random variables from the offspring distribution, which are independent of everything up to generation n. Suppose that the mean number of offspring per individual, μ, is finite. Then:

(1) Show that $M_n := \mu^{-n} Z_n$ is a martingale adapted to a suitable filtration.

(2) Verifying the required conditions, conclude that $\lim_{n\to\infty} M_n$ exists a.s. and is integrable. From this, show that if $\mu < 1$, then $Z_n \to 0$ a.s.
(3) If $\mu = 1$, use the above fact and the fact that Z_n is integer-valued to show that $Z_n \to 0$ a.s. when the number of offspring is not always exactly equal to 1.
(4) Suppose that the variance of the offspring distribution, σ^2, is finite. Then show that $\mathbb{E}(Z_{n+1}^2|\mathcal{F}_n) = \mu^2 Z_n^2 + \sigma^2 Z_n$.
(5) Use the above result to show that if $\mu \neq 1$,

$$N_n := M_n^2 - \frac{\sigma^2}{\mu^{n+1}} \frac{\mu^n - 1}{\mu - 1} M_n$$

is a martingale.
(6) Using the above, show that $\sup_{n\geq 1} \mathbb{E}(M_n^2) < \infty$ if $\mu > 1$ and $\sigma^2 < \infty$. From this, conclude that $\mathbb{P}(\lim_{n\to\infty} M_n > 0) > 0$, that is, there is positive probability that Z_n behaves like a positive multiple of μ^n as $n \to \infty$.

EXERCISE 9.13.9. In Exercise 9.12.2, if $X \in L^p$ for some $p \geq 1$, show that $X_n \to X$ in L^p.

EXERCISE 9.13.10. Using the above exercise, show that $C[0,1]$ is a dense subset of $L^2[0,1]$. (Hint: First show that step functions are dense, and then approximate continuous functions by dense functions.)

EXERCISE 9.13.11. A trigonometric polynomial on $[0,1]$ is a function of the form

$$\sum_{n=0}^{N} a_n \cos 2\pi nx + \sum_{n=1}^{N} b_n \sin 2\pi nx$$

for some N and some coefficients $a_0, \ldots, a_n$ and $b_1, \ldots, b_N$.

(1) Show that the trigonometric polynomials form an algebra over the real numbers — that is, they are closed under addition, multiplication, and scalar multiplication.
(2) Take any $0 < a < b < 1$, and any N. Define a function $f_N : [0,1] \to [0,\infty)$ as

$$f_N(x) := \int_a^b \sqrt{N} \left(\frac{1 + \cos 2\pi(u-x)}{2} \right)^N du.$$

Show that f_N is uniformly bounded above by a constant that does not depend on N, and that as $N \to \infty$, $f_N(x) \to 0$ if $x \notin [a,b]$, and $f_N(x) \to c$ if $x \in (a,b)$, where c is a positive real number.
(3) Using the above and Exercise 9.12.2, show that if $g \in L^1[0,1]$ is orthogonal to $\sin 2\pi nx$ and $\cos 2\pi nx$ for all n (under the L^2 inner product), then $g = 0$ a.e. on $[0,1]$.

9.14. Almost supermartingales

A sequence of random variables $\{X_n\}_{n\geq 0}$ adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$ is called a nonnegative 'almost supermartingale' if each X_n is a nonnegative integrable random variable, and there are two sequences of adapted nonnegative integrable random variables $\{\xi_n\}_{n\geq 0}$ and $\{\zeta_n\}_{n\geq 0}$ such that for each n,

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n + \xi_n - \zeta_n \quad \text{a.s.}$$

The following theorem gives a criterion under which the above almost supermartingale converges a.s. It is sometimes called the 'Robbins–Siegmund almost supermartingale theorem'.

THEOREM 9.14.1. *Let X_n, ξ_n and ζ_n be as above. Then with probability one on the set $\{\sum_{n=0}^{\infty}\xi_n < \infty\}$, $\lim_{n\to\infty} X_n$ exists and is finite, and $\sum_{n=0}^{\infty}\zeta_n < \infty$.*

PROOF. Let

$$Y_n := X_n - \sum_{k=0}^{n-1}(\xi_k - \zeta_k).$$

Then note that

$$\begin{aligned}\mathbb{E}(Y_{n+1}|\mathcal{F}_n) &= \mathbb{E}(X_{n+1}|\mathcal{F}_n) - \sum_{k=0}^{n}(\xi_k - \zeta_k)\\ &\le X_n + \xi_n - \zeta_n - \sum_{k=0}^{n}(\xi_k - \zeta_k)\\ &= X_n - \sum_{k=0}^{n-1}(\xi_k - \zeta_k) = Y_n.\end{aligned}$$

Thus, $\{Y_n\}_{n\ge 0}$ is a supermartingale. Take any $a > 0$. Let $\tau := \inf\{n : \sum_{k=0}^{n}\xi_k \ge a\}$. Note that τ is allowed to be infinity. By the optional stopping theorem, $\{Y_{\tau\wedge n}\}$ is a supermartingale adapted to $\{\mathcal{F}_n\}_{n\ge 0}$. Moreover, since the X_n's and ζ_n's are nonnegative random variables,

$$Y_{\tau\wedge n} = X_{\tau\wedge n} - \sum_{k=0}^{\tau\wedge n-1}(\xi_k - \zeta_k) > -a.$$

By the submartingale convergence theorem, and supermartingale that is uniformly bounded below must converge a.s. to an integrable limit. Therefore, we get that $\lim_{n\to\infty} Y_{\tau\wedge n}$ exists and is finite a.s. This shows that $\lim_{n\to\infty} Y_n$ exists and is finite a.s. on the set $\{\sum_{k=0}^{\infty}\xi_k < a\}$. Taking union over $a = 1, 2, \ldots$, we conclude that $\lim_{n\to\infty} Y_n$ exists and is finite a.s. on the set $\{\sum_{k=0}^{\infty}\xi_k < \infty\}$.

Now, if Y_n converges and $\sum_{k=0}^{\infty}\xi_k$ is finite, then the definition of Y_n shows that

$$\sup_{n\ge 0}\left(X_n + \sum_{k=0}^{n-1}\zeta_k\right) = \sup_{n\ge 0}\left(Y_n + \sum_{k=0}^{n-1}\xi_k\right) < \infty.$$

This implies that $\sum_{k=0}^{\infty}\zeta_k < \infty$, and hence,

$$\lim_{n\to\infty} X_n = \lim_{n\to\infty} Y_n - \sum_{k=0}^{\infty}(\xi_k - \zeta_k)$$

exists. This completes the proof of the theorem. □

Striking applications of almost supermartingales can be found in the literature on stochastic approximation. The following is a special case of the celebrated Robbins–Monro algorithm from this literature.

Suppose that there is an unknown measurable function $f : \mathbb{R} \to \mathbb{R}$ with a single zero at some $x^* \in \mathbb{R}$, and our goal is to find x^*. This kind of problem often comes up in

optimization, because maxima and minima of functions are characterized by the zeros of their derivatives. The problem is that whenever we query the value of f at some point x, we can only recover $f(x)$ up to some random error with mean zero but possibly a substantial variance. The following amazing algorithm allows us to converge to x^* even under such unfavorable conditions.

Start with $X_0 = x_0$ for some arbitrary $x_0 \in \mathbb{R}$. At step n of the algorithm (starting with $n = 0$), we query the value of f at X_n, and obtain a value Y_n, where Y_n is a function of X_n and some extra randomness that is independent of past events, and $\mathbb{E}(Y_n|X_n) = f(X_n)$. Define

$$X_{n+1} := X_n - a_n Y_n,$$

where $\{a_n\}_{n\geq 0}$ is a sequence of constants satisfying

$$\sum_{n=0}^{\infty} a_n^2 < \infty, \quad \sum_{n=0}^{\infty} a_n = \infty. \tag{9.14.1}$$

For example, we can take $a_n = n^{-3/4}$. The following theorem shows that under mild conditions, X_n converges almost surely to x^* as $n \to \infty$.

THEOREM 9.14.2 (Convergence of Robbins–Monro algorithm). *In the above setting, suppose that $|f|$ is uniformly bounded, and that for any $\epsilon > 0$,*

$$\inf_{\epsilon<x<1/\epsilon} f(x^* + x) > 0, \qquad \sup_{-1/\epsilon<x<-\epsilon} f(x^* + x) < 0. \tag{9.14.2}$$

Further, suppose that $\mathrm{Var}(Y_n|X_n)$ is uniformly bounded above by a constant. Then $X_n \to x^*$ almost surely.

PROOF. Define $Z_n := (X_n - x^*)^2$. Let $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$. Since Y_n is a function of X_n and some extra randomness that is independent of the past, $\mathbb{E}(Y_n|\mathcal{F}_n) = \mathbb{E}(Y_n|X_n)$ and $\mathrm{Var}(Y_n|\mathcal{F}_n) = \mathrm{Var}(Y_n|X_n)$. In particular,

$$\mathbb{E}(Y_n - f(X_n)|\mathcal{F}_n) = 0.$$

Thus,

$$\begin{aligned}
&\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = \mathbb{E}((X_{n+1} - x^*)^2|\mathcal{F}_n) \\
&= \mathbb{E}((X_n - x^* - a_n(Y_n - f(X_n)) - a_n f(X_n))^2|\mathcal{F}_n) \\
&= (X_n - x^*)^2 + a_n^2\mathbb{E}((Y_n - f(X_n))^2|\mathcal{F}_n) + a_n^2 f(X_n)^2 \\
&\qquad - 2a_n(X_n - x^*)\mathbb{E}(Y_n - f(X_n)|\mathcal{F}_n) + 2a_n^2 f(X_n)\mathbb{E}(Y_n - f(X_n)|\mathcal{F}_n) \\
&\qquad - 2a_n(X_n - x^*)f(X_n) \\
&= Z_n + a_n^2\mathrm{Var}(Y_n|X_n) + a_n^2 f(X_n)^2 - 2a_n|X_n - x^*||f(X_n)|,
\end{aligned}$$

where in the last step we used the fact (implied by (9.14.2)) that $X_n - x^*$ and $f(X_n)$ have the same sign. Let σ^2 be a uniform upper bound on $\mathrm{Var}(Y_n|X_n)$ and B be a uniform upper bound on $|f|$. Then the above expression shows that

$$\mathbb{E}(Z_{n+1}|\mathcal{F}_n) \leq Z_n + \xi_n - \zeta_n,$$

where

$$\xi_n = a_n^2(\sigma^2 + B^2), \quad \zeta_n = 2a_n|X_n - x^*||f(X_n)|.$$

Note that ξ_n and ζ_n are nonnegative and $\mathcal{F}_n$-measurable. Thus, this expresses Z_n as an almost supermartingale. But here the ξ_n's are constants, and by assumption (9.14.1), $\sum_{n=0}^{\infty} \xi_n < \infty$. Therefore, by the Robbins–Siegmund theorem (Theorem 9.14.1), $\lim_{n\to\infty} Z_n$ exists and is finite a.s., and $\sum_{n=0}^{\infty} \zeta_n < \infty$ a.s. Let

$$D := \lim_{n\to\infty} \sqrt{Z_n} = \lim_{n\to\infty} |X_n - x^*|.$$

Suppose that $D \neq 0$ at some sample point. Then from assumption (9.14.2) we deduce that $f(X_n)$ must stay bounded away from 0 as $n \to \infty$, and also that $|X_n - x^*| \to D > 0$. Since $\sum_{n=0}^{\infty} a_n = \infty$, this shows that at such a sample point, we cannot have $\sum_{n=0}^{\infty} \zeta_n < \infty$. This proves that $D = 0$ a.s., that is, $X_n \to x^*$ a.s. □

The following exercise is a simple fact from real analysis that will be needed in the subsequent problems.

Exercise 9.14.3. Let $\{a_n\}_{n\geq 0}$ be an increasing sequence of positive real numbers. Then show that

$$\sum_{n=0}^{\infty} \frac{a_{n+1} - a_n}{a_{n+1}} = \infty, \quad \sum_{n=0}^{\infty} \frac{a_{n+1} - a_n}{a_{n+1}^2} < \infty.$$

(Hint: For the second claim, simply compare with the integral of $1/x^2$. For the first, separately consider the cases $\liminf a_n/a_{n+1} = 0$ and > 0, and compare with the integral of $1/x$ for the latter.)

Exercise 9.14.4 (Strong law of large numbers for martingales). Let $\{S_n\}_{n\geq 0}$ be a mean zero, square-integrable martingale adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let $\{c_n\}_{n\geq 1}$ be a predictable and increasing sequence of random variables (that is, c_n is $\mathcal{F}_{n-1}$-measurable and $c_n \leq c_{n+1}$ for each n). Let $\sigma_n^2 := \mathrm{Var}(S_n|\mathcal{F}_{n-1})$. Then, show that $S_n/c_n \to 0$ a.s. on the set

$$\left\{\sum_{n=0}^{\infty} \frac{\sigma_n^2}{c_n^2} < \infty \text{ and } \lim_{n\to\infty} c_n = \infty\right\}.$$

(Hint: Define $Z_n := S_n^2/c_n^2$ and show that it is an almost supermartingale, and then apply the Robbins–Siegmund theorem.)

Exercise 9.14.5. Let $X_0, X_1, \ldots$ be a sequence of i.i.d. square-integrable random variables with mean zero. Let $S_n := \sum_{i=0}^{n} X_i X_{i+1}$. Using the SLLN for martingales, prove that $S_n/n \to 0$ a.s.

Exercise 9.14.6. Let $\{X_n\}_{n\geq 0}$ be a sequence of $\{0,1\}$-valued random variables adapted to a filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let $p_n := \mathbb{P}(X_n = 1|\mathcal{F}_{n-1})$ (with $p_0 := \mathbb{P}(X_0 = 1)$) and let $S_n := \sum_{k=0}^{n} X_k$. Then show that the limit

$$\lim_{n\to\infty} \frac{S_n}{\sum_{k=0}^{n} p_k}$$

exists a.s. Moreover, show that the limit is equal to 1 almost surely on the set $\{\sum_{n=0}^{\infty} p_n = \infty\}$. (Hint: Define a suitable almost supermartingale.)

CHAPTER 10

Ergodic theory

Ergodic theory is the study of measure preserving transforms and their properties. It is useful in probability theory for the study of random processes whose laws are invariant under various groups of transformations. This chapter gives an introduction to the basic notions and results of ergodic theory and some applications in probability.

10.1. Measure preserving transforms

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable map $\varphi : \Omega \to \Omega$ is called 'measure preserving' if $\mathbb{P}(\varphi^{-1}A) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$. (In ergodic theory, the standard practice is to write $\varphi^{-1}A$ instead of $\varphi^{-1}(A)$, $\varphi^2(A)$ instead of $\varphi(\varphi(A))$, as so on.) The following lemma is often useful for proving that a given map is measure preserving.

LEMMA 10.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\varphi : \Omega \to \Omega$ be a measurable map. Suppose that $\mathbb{P}(\varphi^{-1}A) = \mathbb{P}(A)$ for all A in a π-system $\mathcal{P}$ that generates $\mathcal{F}$. Then φ is measure preserving.*

PROOF. By Dynkin's π-λ theorem, it suffices to check that the set $\mathcal{L}$ of all $A \in \mathcal{F}$ such that $\mathbb{P}(\varphi^{-1}A) = \mathbb{P}(A)$ is a λ-system. First, note that $\varphi^{-1}\Omega = \Omega$. Thus, $\Omega \in \mathcal{L}$. Next, if $A \in \mathcal{L}$, then

$$\mathbb{P}(\varphi^{-1}A^c) = \mathbb{P}((\varphi^{-1}A)^c) = 1 - \mathbb{P}(\varphi^{-1}A) = 1 - \mathbb{P}(A) = \mathbb{P}(A^c),$$

which shows that $A^c \in \mathcal{L}$. Finally, if $A_1, A_2, \ldots$ are disjoint elements of $\mathcal{L}$, and $A = \cup_{i=1}6\infty A_i$, then

$$\begin{aligned}\mathbb{P}(\varphi^{-1}A) &= \mathbb{P}\Big(\bigcup_{i=1}^{\infty} \varphi^{-1}A_i\Big) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(\varphi^{-1}A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \mathbb{P}(A).\end{aligned}$$

Thus, $\mathcal{L}$ is a λ-system. □

EXAMPLE 10.1.2 (Bernoulli shift). Let $\Omega = \{0,1\}^{\mathbb{N}}$ be the set of all infinite sequences of 0's and 1's. Let $\mathcal{F}$ be the product σ-algebra on this space, and $\mathbb{P}$ be the product of $Ber(p)$ measures. The 'shift operator' on this space is defined as

$$\varphi(\omega_0, \omega_1, \ldots) = (\omega_1, \omega_2, \ldots).$$

We now show that this map is measure preserving. Let $\mathcal{A}$ be the collection of all sets of the form

$$\{\omega : \omega_0 = x_0, \ldots, \omega_n = x_n\}$$

for some $n \geq 0$ and some $x_0, \ldots, x_n \in \{0,1\}$. Clearly, this is a π-system that generates the product σ-algebra. Let A denote the set displayed above. Then $\varphi^{-1}A$ is the set of all sequences $(\omega_0, \omega_1, \ldots)$ such that $(\omega_1, \omega_2, \ldots) \in A$, that is, the set of all ω such that $\omega_1 = x_0, \omega_2 = x_1, \ldots \omega_{n+1} = x_n$. Under the product measure, both A and $\varphi^{-1}A$ have the same measure. By Lemma 10.1.1, this shows that φ is measure preserving.

EXAMPLE 10.1.3 (Rotations of the circle). Let $\Omega = [0,1)$ with the Borel σ-algebra and Lebesgue measure. Take any $\theta \in [0,1)$, and let

$$\varphi(x) := x + \theta \pmod 1 = \begin{cases} x + \theta & \text{if } x + \theta < 1, \\ x + \theta - 1 & \text{if } x + \theta \geq 1. \end{cases}$$

One way to view φ is to consider $[0,1)$ as the unit circle in the complex plane via the map $x \mapsto e^{2\pi i x}$, which makes φ a 'rotation by angle θ'. We claim that φ is measure preserving. To see this, simply note that the map φ is a bijection with

$$\varphi^{-1}(x) = x - \theta \pmod 1 = \begin{cases} x - \theta & \text{if } x - \theta \geq 0, \\ x - \theta + 1 & \text{if } x - \theta < 0. \end{cases}$$

and therefore, for any interval $[a,b] \subseteq [0,1)$,

$$\varphi^{-1}([a,b]) = \begin{cases} [a-\theta, b-\theta] & \text{if } a \geq \theta, \\ [0, b-\theta] \cup [1+a-\theta, 1] & \text{if } a < \theta \leq b, \\ [1+a-\theta, 1+b-\theta] & \text{if } b < \theta. \end{cases}$$

In all three cases, the Lebesgue measure of $\varphi^{-1}([a,b])$ equals $b-a$. Since the set of closed intervals is a π-system that generates the Borel σ-algebra on $[0,1)$, Lemma 10.1.1 shows that φ is measure preserving.

EXERCISE 10.1.4. Show that the map $\varphi(x) := 2x \pmod 1$ preserves Lebesgue measure on $[0,1)$.

Measure preserving transforms possess the following important property.

PROPOSITION 10.1.5. *If φ is a measure preserving map on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then for any integrable random variable X defined on this space, $X \circ \varphi$ has the same distribution as X, and hence, has the same expected value as X.*

PROOF. Simply note that for any Borel set $A \subseteq \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(X \circ \varphi \in A) &= \mathbb{P}(\{\omega : \phi(\omega) \in X^{-1}(A)) \\ &= \mathbb{P}(\varphi^{-1}X^{-1}(A)) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A), \end{aligned}$$

where the second-to-last inequality holds by the measure preserving property. □

10.2. Ergodic transforms

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\varphi : \Omega \to \Omega$ be a measure preserving transform. The 'invariant σ-algebra' of φ is the set of all $A \in \mathcal{F}$ such that $\varphi^{-1}A = A$. It is easy to see that this is indeed a σ-algebra. The transform φ is called 'ergodic' for the probability measure $\mathbb{P}$ if its invariant σ-algebra $\mathcal{I}$ is trivial, that is, $\mathbb{P}(A) \in \{0,1\}$ for all $A \in \mathcal{I}$.

EXAMPLE 10.2.1 (Ergodicity of Bernoulli shift). Recall the Bernoulli shift operator φ defined in Example 10.1.2, and the associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We claim that φ is ergodic. To see this, take any $A \in \mathcal{I}$, where $\mathcal{I}$ is the invariant σ-algebra of φ. Define a sequence of random variables $(X_n)_{n\geq 0}$ as $X_n(\omega) := \omega_n$. Then, by the definition of $\mathbb{P}$, these random variables are i.i.d. $Ber(p)$. The invariance of A means that for any k,

$$\begin{aligned} A &= \varphi^{-k} A \\ &= \{\omega : (\omega_k, \omega_{k+1}, \ldots) \in A\} \\ &= \{\omega : (X_k(\omega), X_{k+1}(\omega), \ldots) \in A\}. \end{aligned}$$

Thus, $A \in \sigma(X_k, X_{k+1}, \ldots)$. Since this holds for any k, we conclude that A is an element of the tail σ-algebra of the sequence $(X_n)_{n\geq 0}$. Thus, by Kolmogorov's zero-one law, $\mathbb{P}(A) \in \{0, 1\}$, and hence, φ is ergodic.

EXAMPLE 10.2.2 (Ergodicity of rotations). Recall the rotation map φ from Example 10.1.3. We will now show that φ is ergodic if θ is irrational, but not ergodic if θ is rational.

First, suppose that θ is rational. Then $\theta = m/n$ for some integers $0 \leq m < n$. Let

$$A := \bigcup_{i=0}^{n-1} \left[\frac{i}{n}, \frac{i}{n} + \frac{1}{2n}\right).$$

Now note that for any $i \geq m$,

$$\varphi^{-1}\left[\frac{i}{n}, \frac{i}{n} + \frac{1}{2n}\right) = \left[\frac{i-m}{n}, \frac{i-m}{n} + \frac{1}{2n}\right),$$

and for any $i < m$,

$$\varphi^{-1}\left[\frac{i}{n}, \frac{i}{n} + \frac{1}{2n}\right) = \left[\frac{i-m+n}{n}, \frac{i-m+n}{n} + \frac{1}{2n}\right),$$

From this, it is easy to see that $\varphi^{-1}A = A$. But the Lebesgue measure of A is $1/2$. Thus, φ is not ergodic if θ is rational.

Next, suppose that θ is irrational. Take any A such that $A = \varphi^{-1}A$. Then note that for any integer k, Proposition 10.1.5 gives

$$c_k := \int_0^1 1_A(x) e^{2\pi \mathrm{i} k x} dx = \int_0^1 1_A(\varphi(x)) e^{2\pi \mathrm{i} k \varphi(x)} dx.$$

But the invariance of A implies that

$$1_A(\varphi(x)) = 1_{\varphi^{-1}A}(x) = 1_A(x),$$

and the fact that k is an integer implies that

$$e^{2\pi \mathrm{i} k \varphi(x)} = e^{2\pi \mathrm{i} k (x+\theta)},$$

because $\varphi(x)$ is either $x + \theta$ or $x + \theta - 1$, and $e^{-2\pi \mathrm{i} k} = 1$. Thus,

$$c_k = e^{2\pi \mathrm{i} k \theta} c_k.$$

Since θ is irrational, $k\theta$ is not an integer for any nonzero integer k, and so, the above identity implies that $c_k = 0$ for all $k \neq 0$. Consequently, the inner product of $1_A - c_0$ with

$\cos 2\pi nx$ or $\sin 2\pi nx$, for any n, is zero. By Exercise 9.13.11, this implies that $1_A = c_0$ a.e., which shows that the Lebesgue measure of A must be 0 or 1.

EXERCISE 10.2.3. Prove that the map φ from Exercise 10.1.4 is ergodic.

EXERCISE 10.2.4. If $\mathcal{I}$ is the invariant σ-algebra of a measure preserving transform φ, show that an $\mathbb{R}^*$-valued random variable Y is $\mathcal{I}$-measurable if and only if $Y \circ \varphi = Y$.

EXERCISE 10.2.5. Give an example of an ergodic measure preserving transformation φ such that φ^2 is not ergodic.

EXERCISE 10.2.6. Let $d \geq 1$ and $\Omega = \{0,1\}^E$, where E is the set of edges of $\mathbb{Z}^d$. Let Ω be endowed with the product σ-algebra. Let $p \in (0,1)$ and let μ be the infinite product of Bernoulli(p) measures on Ω. That is, μ is the law of a collection of i.i.d. Bernoulli(p) random variables, one attached to each edge of $\mathbb{Z}^d$. For each $x \in \mathbb{Z}^d$, let T_x denote the map on Ω that translates by x; that is, if $\omega' = T_x(\omega)$, then $\omega'_e = \omega_{e+x}$ for each edge e, where $e + x$ is the translation of e by x. Prove that for any x, T_x is measure preserving, and if $x \neq 0$, then T_x is ergodic.

EXERCISE 10.2.7. Consider a collection of i.i.d. Bernoulli(p) random variables attached to edges of $\mathbb{Z}^d$, as in the preceding problem. Let X_e denote the variable attached to edge e. Let us say that an edge is 'open' if $X_e = 1$ and 'closed' otherwise. This is the classical edge percolation model. Say that two vertices are connected if there is a sequence of open edges connecting one to the other. This is an equivalence relation that divides up $\mathbb{Z}^d$ into a set of disjoint connected 'clusters'. Using the previous problem, show that the number of infinite clusters is almost surely a constant N, belonging to $\{0,1,2,\ldots\} \cup \{\infty\}$.

EXERCISE 10.2.8. In the previous problem, suppose that N is finite but greater than one. Then show that there is a set $B \subseteq \mathbb{Z}^d$ such that B intersects more than one infinite cluster with positive probability. Using this, produce an argument that leads to a contradiction, and thus establish that $N \in \{0,1\} \cup \{\infty\}$. (Actually, the value ∞ is also impossible, but that is much harder to prove. This is a famous result in percolation theory.)

EXERCISE 10.2.9. Let $\varphi(x) = 4x(1-x)$ for $x \in [0,1]$. Prove that φ preserves the probability measure on $[0,1]$ with density $1/(\pi\sqrt{x(1-x)})$.

10.3. Birkhoff's ergodic theorem

The following is one of the fundamental results of ergodic theory, known as Birkhoff's ergodic theorem, and also known as the pointwise ergodic theorem.

THEOREM 10.3.1 (Birkhoff's ergodic theorem). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and φ be a measure preserving transform on Ω. Let X be an integrable real-valued random variable defined on Ω. Let $\mathcal{I}$ be the invariant σ-algebra of φ. Then*

$$\frac{1}{n}\sum_{m=0}^{n-1} X \circ \varphi^m \to \mathbb{E}(X|\mathcal{I})$$

almost surely and in L^1 as $n \to \infty$. In particular, if φ is ergodic, then the averages on the left converge to $\mathbb{E}(X)$ a.s. and in L^1.

The first step in the proof of Birkhoff's ergodic theorem is the following result, known as the maximal ergodic theorem.

THEOREM 10.3.2 (Maximal ergodic theorem). *Let X and φ be as in Birkhoff's ergodic theorem. For each $k \geq 0$, let $X_k := X \circ \varphi^k$, $S_k := X_0 + \cdots + X_k$, and*

$$M_k := \max\{0, S_0, S_1, \ldots, S_k\}.$$

Then for any k, $\mathbb{E}(X; M_k > 0) \geq 0$.

PROOF. Take any $\omega \in \Omega$. If $j \leq k$, then $S_j(\varphi\omega) \leq M_k(\varphi\omega)$. Thus,

$$X(\omega) + M_k(\varphi\omega) \geq X(\omega) + S_j(\varphi\omega) = S_{j+1}(\omega).$$

Thus,

$$X(\omega) \geq S_{j+1}(\omega) - M_k(\varphi\omega)$$

for $j = 0, \ldots, k$. But also,

$$X(\omega) \geq S_0(\omega) - M_k(\varphi\omega)$$

since $S_0 = X$ and $M_k \geq 0$ everywhere. Thus,

$$\begin{aligned}\mathbb{E}(X; M_k > 0) &\geq \int_{\{\omega: M_k(\omega) > 0\}} (\max\{S_0(\omega), \ldots, S_k(\omega)\} - M_k(\varphi\omega)) d\mathbb{P}(\omega) \\ &= \int_{\{\omega: M_k(\omega) > 0\}} (M_k(\omega) - M_k(\varphi\omega)) d\mathbb{P}(\omega).\end{aligned}$$

Now, by Proposition 10.1.5,

$$\int_\Omega (M_k(\omega) - M_k(\varphi\omega)) d\mathbb{P}(\omega) = 0.$$

On the other hand, the nonnegativity of M_k implies that

$$\int_{\{\omega: M_k(\omega) = 0\}} (M_k(\omega) - M_k(\varphi\omega)) d\mathbb{P}(\omega) \leq 0.$$

By the last three displays, we see that $\mathbb{E}(X; M_k > 0)$ must be nonnegative. □

We are now ready to prove Birkhoff's ergodic theorem.

PROOF OF THEOREM 10.3.1. First, we claim that it suffices to prove the theorem under the assumption that $\mathbb{E}(X|\mathcal{I}) = 0$ a.s. To see this, suppose that the theorem holds under this assumption. Let $Y := \mathbb{E}(X|\mathcal{I})$. Since Y is $\mathcal{I}$-measurable, it satisfies $Y \circ \varphi^m = Y$ for any m, by Exercise 10.2.4. Let $Z := X - Y$, so that $\mathbb{E}(Z|\mathcal{I}) = 0$ a.s. Thus, by the assumed version of the theorem,

$$\frac{1}{n} \sum_{m=0}^{n-1} Z \circ \varphi^m \to 0$$

a.s. and in L^1. But $Z \circ \varphi^m = X \circ \varphi^m - Y \circ \varphi^m = X \circ \varphi^m - Y$, and thus,

$$\frac{1}{n} \sum_{m=0}^{n-1} Z \circ \varphi^m = \frac{1}{n} \sum_{m=0}^{n-1} X \circ \varphi^m - \mathbb{E}(X|\mathcal{I}).$$

This gives us the full version of the theorem. Thus, let us henceforth work under the assumption that $\mathbb{E}(X|\mathcal{I}) = 0$ a.s. Our goal, then, is to show that $S_n/(n+1) \to 0$ a.s. and

in L^1, where S_n is as in the maximal ergodic theorem. Define

$$\overline{X} := \limsup_{n\to\infty} \frac{S_n}{n+1}.$$

From this definition, it is clear that $\overline{X} \circ \varphi = \overline{X}$. By Exercise 10.2.4, this shows that $\overline{X}$ is $\mathcal{I}$-measurable.

Next, fix some $\epsilon > 0$, and define

$$D := \{\omega : \overline{X}(\omega) > \epsilon\}.$$

Since $\overline{X}$ is $\mathcal{I}$-measurable, we conclude that $D \in \mathcal{I}$. Define

$$X^* := (X - \epsilon)1_D,$$

and define, in analogy with S_k and M_k, the random variables

$$S_k^*(\omega) := X^*(\omega) + X^*(\varphi\omega) + \cdots + X^*(\varphi^k\omega),$$
$$M_k^*(\omega) := \max\{0, S_0^*(\omega), \ldots, S_k^*(\omega)\}.$$

For each k, define the set

$$F_k := \{\omega : M_k^*(\omega) > 0\},$$

and let $F := \cup_{k\geq 0} F_k$. We claim that $F = D$. To see this, first take any $\omega \in F$. Then $M_k^*(\omega) > 0$ for some k, and hence $S_j^*(\omega) > 0$ for some j. Thus, $X^*(\varphi^i\omega) > 0$ for some i. In particular, this shows that $\varphi^i\omega \in D$ for some i. Since $D \in \mathcal{I}$, this shows that $\omega \in D$. Thus, $F \subseteq D$. Conversely, take any $\omega \in D$. Then $\varphi^j\omega \in D$ for all j, and so,

$$X^*(\varphi^j\omega) = (X(\varphi^j\omega) - \epsilon)1_D(\varphi^j\omega) = X(\varphi^j\omega) - \epsilon.$$

Thus,

$$S_k^*(\omega) = S_k(\omega) - (k+1)\epsilon$$

for all k. Since $\omega \in D$, we have

$$\limsup_{k\to\infty} \frac{S_k(\omega)}{k+1} > \epsilon.$$

Thus, the previous display shows that $S_k^*(\omega) > 0$ for some k, and hence, $M_k^*(\omega) > 0$ for some k. In other words, $\omega \in F$. Thus, $D \subseteq F$, completing the proof of our claim that $F = D$.

Since M_k^* is an increasing sequence, so is F_k. Thus, $X^*1_{F_k} \to X^*1_F$ pointwise as $k \to \infty$. Moreover, $|X^*1_{F_k}| \leq |X^*| \leq |X| + \epsilon$ for all k, and X is integrable. So, by the dominated convergence theorem, we get that

$$\lim_{k\to\infty} \mathbb{E}(X^*; F_k) = \mathbb{E}(X^*; F).$$

By the maximal ergodic theorem, $\mathbb{E}(X^*; F_k) \geq 0$ for any k. Since $F = D$, this shows that

$$\mathbb{E}(X^*; F) = \mathbb{E}(X^*; D) \geq 0.$$

But

$$\mathbb{E}(X^*; D) = \mathbb{E}((X - \epsilon)1_D) = \mathbb{E}(\mathbb{E}(X|\mathcal{I}); D) - \epsilon\mathbb{P}(D).$$

Since $\mathbb{E}(X|\mathcal{I}) = 0$ a.s., this shows that $-\epsilon\mathbb{P}(D) \geq 0$. Thus, $\mathbb{P}(D) = 0$. Looking back at the definition of D, we conclude that $\limsup S_n/(n+1) \leq 0$. Replacing X by $-X$ throughout the proof, we conclude that $\liminf S_n/(n+1) \geq 0$. This completes the proof of the a.s. convergence of S_n/n to 0.

It remains to prove convergence in L^1. Fix any $M > 0$. Define $X'_M := X1_{\{|X|\leq M\}}$ and $X''_M := X - X'_M$. Then by the almost sure part of the theorem,

$$\frac{1}{n}\sum_{m=0}^{n-1} X'_M \circ \varphi^m \to \mathbb{E}(X'_M|\mathcal{I})$$

a.s. as $n \to \infty$. Since the random variable on the right is uniformly bounded by M in absolute value, the dominated convergence theorem implies that the convergence also holds in L^1. Now, since $\mathbb{E}(X|\mathcal{I}) = 0$ a.s., we have

$$\begin{aligned}&\mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X \circ \varphi^m\right| \\ &\leq \mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X'_M \circ \varphi^m - \mathbb{E}(X'_M|\mathcal{I})\right| + \mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M \circ \varphi^m - \mathbb{E}(X''_M|\mathcal{I})\right|.\end{aligned}$$

We have already shown that the first term on the right tends to zero as $n \to \infty$. On the other hand,

$$\mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M \circ \varphi^m - \mathbb{E}(X''_M|\mathcal{I})\right| \leq \frac{1}{n}\sum_{m=0}^{n-1} \mathbb{E}|X''_M \circ \varphi^m| + \mathbb{E}|\mathbb{E}(X''_M|\mathcal{I})|.$$

By Proposition 10.1.5, $\mathbb{E}|X''_M \circ \varphi^m| = \mathbb{E}|X''_M|$ for any m, and by Jensen's inequality,

$$\mathbb{E}|\mathbb{E}(X''_M|\mathcal{I})| \leq \mathbb{E}(\mathbb{E}(|X''_M||\mathcal{I})) = \mathbb{E}|X''_M|.$$

Combining everything, we see that

$$\limsup_{n\to\infty} \mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X \circ \varphi^m\right| \leq 2\mathbb{E}|X''_M|.$$

But M is arbitrary, and by the dominated convergence theorem, $\mathbb{E}|X''_M| \to 0$ as $M \to \infty$. This proves the L^1 part of the theorem. □

Exercise 10.3.3. In Birkhoff's ergodic theorem, show that if $X \in L^p$ for some $p > 1$, then the convergence happens in L^p. (The special case of $p = 2$ is known as 'Von Neumann's ergodic theorem' or the 'mean ergodic theorem'. It has a simpler direct proof using Hilbert spaces.)

Exercise 10.3.4. Let φ be the rotation map from Example 10.1.3. If $\theta \in [0, 1)$ is irrational, show that for any Borel set $A \subseteq [0, 1)$, for a.e. $x \in [0, 1)$ we have

$$\lim_{n\to\infty} \frac{1}{n}|\{0 \leq m \leq n-1 : \varphi^m x \in A\}| = \lambda(A),$$

where $\lambda(A)$ is the Lebesgue measure of A.

EXERCISE 10.3.5. In the previous exercise, suppose that $A = [a, b)$ is a half-open interval contained in $[0, 1)$. Then show that the claim actually for all $x \in [0, 1)$ (instead of almost all), by the following steps:

(1) Take any k so large that $b - a > 2/k$. Let $A_k := [a + 1/k, b - 1/k)$. Applying the previous exercise to the set A_k, find a set $\Omega_k \subseteq [0, 1)$ of Lebesgue measure 1 such that the conclusion of the previous exercise holds for all $x \in \Omega_k$.
(2) Deduce that Ω_k is dense in $[0, 1)$.
(3) Given any $x \in [0, 1)$, find $y_k \in \Omega_k$ such that $|x - y_k| < 1/k$. Show that if $\varphi^m y_k \in A_k$, then $\varphi^m x \in A$.
(4) Using the above, conclude that for all large enough k,
$$\liminf_{n\to\infty} \frac{1}{n}|\{0 \le m \le n - 1 : \varphi^m x \in [a, b)\}| \ge b - a - \frac{2}{k}.$$
(5) Working with the complement of $[a, b)$, deduce the opposite inequality with lim sup. Combining the two, deduce that the conclusion of the previous exercise holds for all $x \in [0, 1)$ if $A = [a, b)$.

10.4. Stationary sequences

Recall that a sequence of real-valued random variables $(X_n)_{n\ge0}$ is called a 'stationary sequence' if for any n and k, the law of $(X_0, \dots, X_n)$ is the same as that of $(X_k, X_{k+1}, \dots, X_{k+n})$. (See Definition 8.5.4 and the discussion following it.) The following lemma connects stationary sequences of random variables with measure preserving maps.

LEMMA 10.4.1. *Let $(X_n)_{n\ge0}$ be a stationary sequence of real-valued random variables. Let μ be the law of $(X_n)_{n\ge0}$ on $\mathbb{R}^{\mathbb{N}}$. Let φ be the shift operator on $\mathbb{R}^{\mathbb{N}}$, that is, $\varphi(\omega_0, \omega_1, \dots) = (\omega_1, \omega_2, \dots)$. Then φ preserves μ.*

PROOF. Take any rectangular set
$$A = \{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_{i_1} \in B_1, \dots, \omega_{i_k} \in B_k\},$$
where $i_1, \dots, i_k$ are distinct indices and $B_1, \dots, B_k$ are Borel subsets of $\mathbb{R}$. Then
$$\begin{aligned}\mu(\varphi^{-1}A) &= \mathbb{P}((X_0, X_1, \dots) \in \varphi^{-1}A)\\ &= \mathbb{P}(\varphi(X_0, X_1, \dots) \in A)\\ &= \mathbb{P}((X_1, X_2, \dots) \in A)\\ &= \mathbb{P}(X_{i_1+1} \in B_1, X_{i_2+1} \in B_2, \dots, X_{i_k+1} \in B_k).\end{aligned}$$
But by stationarity, the last expression equals
$$\mathbb{P}(X_{i_1} \in B_1, X_{i_2} \in B_2, \dots, X_{i_k} \in B_k),$$
which is equal to $\mu(A)$. This proves that φ preserves μ. □

Now let X denote the sequence $(X_n)_{n\ge0}$. Let $\mathcal{I}$ be the invariant σ-algebra of the shift operator on $\mathbb{R}^{\mathbb{N}}$. Then $\mathcal{J} := X^{-1}(\mathcal{I})$ is a sub-σ-algebra of $\mathcal{F}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space on which X is defined. Let us call this the 'invariant σ-algebra of X'. A stationary sequence is called 'ergodic' if its invariant σ-algebra is trivial.

As a corollary of Birkhoff's ergodic theorem, we get the following theorem for infinite stationary sequences of real-valued random variables.

THEOREM 10.4.2. *Let $X = (X_n)_{n\geq 0}$ be a stationary sequence of integrable real-valued random variables, with invariant σ-algebra $\mathcal{J}$. Then*

$$\frac{1}{n}\sum_{m=0}^{n-1} X_m \to \mathbb{E}(X_0|\mathcal{J})$$

almost surely and in L^1 as $n \to \infty$. In particular, if X is ergodic, then the averages on the left converge to $\mathbb{E}(X_0)$ a.s. and in L^1.

PROOF. Let μ be the law of X on $\mathbb{R}^{\mathbb{N}}$, and let φ be the shift operator on $\mathbb{R}^{\mathbb{N}}$. Define a random variable $Y : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ as $Y(\omega_0, \omega_1, \ldots) := \omega_0$. Under μ, the law of Y is the same as that of X_0. In particular, Y is integrable. Therefore, by Lemma 10.4.1 and Birkhoff's ergodic theorem,

$$\frac{1}{n}\sum_{m=0}^{n-1} Y \circ \varphi^m \to \mathbb{E}(Y|\mathcal{I})$$

almost surely and in L^1 as $n \to \infty$, where $\mathcal{I}$ is the invariant σ-algebra of φ. Let A be a measurable subset of $\mathbb{R}^{\mathbb{N}}$ of μ-measure 1 on which the above convergence takes place. Then $1 = \mu(A) = \mathbb{P}(X \in A)$, which implies that with $\mathbb{P}$-probability 1,

$$\frac{1}{n}\sum_{m=0}^{n-1} Y \circ \varphi^m \circ X \to Z \circ X,$$

where $Z := \mathbb{E}(Y|\mathcal{I})$. Now note that $Y \circ \varphi^m \circ X = X_m$. Thus, to prove the 'almost sure' part of the theorem, it suffices to show that $Z \circ X = \mathbb{E}(X|\mathcal{J})$. Take any $E \in \mathcal{J}$. Then $E = X^{-1}(F)$ for some $F \in \mathcal{I}$. Thus,

$$\mathbb{E}(Z \circ X; E) = \int_E Z \circ X d\mathbb{P}$$

By the general change of variable formula for Lebesgue integrals (Exercise 2.3.7),

$$\int_E Z \circ X d\mathbb{P} = \int_F Z d\mu = \mathbb{E}(Z; F) = \mathbb{E}(Y; F),$$

where the last identity holds because $Z = \mathbb{E}(Y|\mathcal{I})$ and $F \in \mathcal{I}$. But again, by change of variable,

$$\mathbb{E}(Y; F) = \int_F Y d\mu = \int_E Y \circ X d\mathbb{P} = \mathbb{E}(X_0; E),$$

where the last identity holds because $Y \circ X = X_0$. This shows that $Z \circ X = \mathbb{E}(X_0|\mathcal{J})$, and completes the proof of the almost sure part of the theorem. For the L^1 part, simply note that

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X_m - \mathbb{E}(X_0|\mathcal{J})\right| &= \int \left|\frac{1}{n}\sum_{m=0}^{n-1} Y \circ \varphi^m \circ X - Z \circ X\right| d\mathbb{P} \\ &= \int \left|\frac{1}{n}\sum_{m=0}^{n-1} Y \circ \varphi^m - Z\right| d\mu,\end{aligned}$$

which tends to zero as $n \to \infty$. This completes the proof of the theorem. □

EXERCISE 10.4.3. Use Theorem 10.4.2 to give an alternative proof of the SLLN for m-dependent stationary sequences (Theorem 8.5.6).

EXERCISE 10.4.4. A number $x \in [0,1)$ is said to be normal in base b if the digits in its base b expansion have the property that for any finite sequence $d_1, \ldots, d_k \in \{0, 1, \ldots, b-1\}$, the fraction of times the pattern $d_1 d_2 \cdots d_k$ occurs in the first n digits of x tends to b^{-k} as $n \to \infty$. Show that if x is chosen uniformly at random from $[0,1)$, then its base b digits are i.i.d. random variables uniformly distributed in $\{0, 1, \ldots, b-1\}$. Using this, and the previous exercise, prove that almost every number $x \in [0,1)$ is normal in base b. Then use this to show that almost all $x \in [0,1)$ is a normal number, meaning that it's normal in every basis. (It's a famous open problem to show that some "naturally occurring" element of $[0,1)$, such as $1/\sqrt{2}$ or $1/e$, is normal. As of now, we can construct only very artificial examples of normal numbers, even though almost all numbers are normal.)

EXERCISE 10.4.5. Let $X_0, X_1, \ldots$ be a sequence of jointly normal random variables (meaning that any finitely many of them have a joint normal distribution), such that for each i, $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$, and for each $i \neq j$, $\mathrm{Cov}(X_i, X_j) = \rho$, where $\rho \in [-1, 1]$ is some fixed constant. Show that this sequence is stationary, and that $\rho \geq 0$. Then, let $\mathcal{J}$ denote the invariant σ-algebra of this sequence, and evaluate the distribution of $\mathbb{E}(X_0|\mathcal{J})$. (This gives an example of a stationary sequence where the limit in Theorem 10.4.2 is not a constant.)

CHAPTER 11

Markov chains

In this chapter, we define and prove the existence of Markov chains on Euclidean spaces and discuss some of their basic properties.

11.1. The Ionescu-Tulcea existence theorem

A Markov chain is defined by its transition kernel. A transition kernel, or transition probability, is defined as follows.

DEFINITION 11.1.1. Let $(S, \mathcal{S})$ be a measurable space. A function $K : S \times \mathcal{S} \to \mathbb{R}$ is called a 'Markov transition probability' or 'Markov transition kernel' if

(1) for each $x \in S$, $A \mapsto K(x, A)$ is probability measure on $(S, \mathcal{S})$, and
(2) for each $A \in \mathcal{S}$, $x \mapsto K(x, A)$ is a measurable function.

DEFINITION 11.1.2. Let $(S, \mathcal{S})$ be a measurable space. Let $\{X_n\}_{n\geq 0}$ be a sequence of S-valued random variables adapted to a filtration of σ-algebras $\{\mathcal{F}_n\}_{n\geq 0}$. Let K be a Markov transition kernel on S and σ be a probability measure on σ. We say that $\{X_n\}_{n\geq 0}$ is a time-homogeneous Markov chain with kernel K and initial distribution σ if $X_0 \sim \sigma$, and for all $n \geq 0$ and $B \in \mathcal{S}$, $\mathbb{P}(X_{n+1} \in B|\mathcal{F}_n) = K(X_n, B)$ a.s.

Often, the initial distribution σ is taken to be the point mass at some point $x_0 \in S$. In that case, we say that x_0 is the initial state of the chain. The following theorem shows that given any kernel and any initial distribution on any state space, it is possible to construct a Markov chain with the given kernel and initial distribution.

THEOREM 11.1.3 (Ionescu-Tulcea theorem). *Given any Markov transition kernel K on any state space $(S, \mathcal{S})$ as above, and any probability measure σ on $(S, \mathcal{S})$, it is possible to construct a Markov chain on S with kernel K and initial distribution σ.*

We need the following lemma.

LEMMA 11.1.4. *Let $(S, \mathcal{S})$ be a measurable space and K be a Markov transition kernel on this space. For any $n \geq 2$ and any bounded measurable $f : S^n \to \mathbb{R}$, the map*

$$g(x_1, \ldots, x_{n-1}) := \int_S f(x_1, \ldots, x_n) K(x_{n-1}, dx_n)$$

is well-defined, bounded and measurable.

PROOF. First of all note that by Lemma 3.2.1, for any given $x_1, \ldots, x_{n-1}$, the map $x_n \mapsto f(x_1, \ldots, x_n)$ is bounded and measurable. This proves that g is well-defined. Next, note that if we can prove the claim when $f = 1_B$ for any measurable set $B \subseteq S^n$, the general claim will follow by the usual route of going from indicator functions to simple

functions and then to general measurable functions. Thus, let us prove it when $f = 1_B$. Suppose first that $B = B_1 \times \cdots \times B_n$ for some $B_1, \ldots, B_n \in \mathcal{S}$. Then

$$g(x_1, \ldots, x_{n-1}) = 1_{B_1}(x_1) \cdots 1_{B_{n-1}}(x_{n-1}) K(x_{n-1}, B_n),$$

which is measurable by the properties of K. Now consider the set of all measurable B for which the claim is true. Using the properties of K, it is not hard to see that this is a λ-system, and by the above deduction, it contains the π-system of all rectangular sets. The proof is now completed by invoking the π-λ theorem. □

Proof of Theorem 11.1.3. By Lemma 11.1.4, the following iterated integral is well-defined as a functional on $\mathcal{S}^n$:

$$\mu_n(B) := \iint \cdots \int 1_B(x_1, \ldots, x_n) K(x_{n-1}, dx_n) K(x_{n-2}, dx_{n-1}) \cdots \\ \cdots K(x_1, dx_2) d\sigma(x_1).$$

Using the properties of K and the monotone convergence theorem, it is not hard to show that μ_n is a probability measure on S^n and moreover, that the family $\{\mu_n\}_{n \geq 1}$ is consistent, in the sense that for any n and any $A \in \mathcal{S}^n$, $\mu_{n+1}(A \times S) = \mu_n(A)$. We will now show that there is a probability measure μ on the infinite product space such that μ_n is the marginal of μ on S^n for each n.

Note that we cannot apply Kolmogorov's extension theorem, since S is not a Euclidean space or a Polish space. Instead, consider the algebra $\mathcal{A}$ of all cylinder sets, that is, sets of the form $B = A \times S \times S \times \cdots$ for some $A \in \mathcal{S}^n$ for some n. Define $\mu(B) = \mu_n(A)<$ which is well-defined due to the consistency of the family $\{\mu_n\}_{n \geq 1}$. Then μ is finitely additive on $\mathcal{A}$. By Carathéordory's extension theorem, we only need to show that μ is countably additive on $\mathcal{A}$.

As in the proof of Theorem 3.3.1, this can be reduced to showing that for any sequences of set $\{A_n\}_{n \geq 1}$ in $\mathcal{A}$ decreasing to $\emptyset$, $\mu(A_n) \to 0$. So, let us take any such sequence, and suppose that $\mu(A_n) \geq \varepsilon > 0$ for all n.

Take any $B \in \mathcal{A}$. Take any n such that $B = A \times S \times S \times \cdots$ for some $A \in \mathcal{S}^n$. For $1 \leq k \leq n-1$, define the "conditional probability of the chain being in B given that the first k coordinates are $x_1, \ldots, x_k$" to be

$$\mu(B | x_1, \ldots, x_k) := \int \cdots \int 1_A(x_1, \ldots, x_n) K(x_{n-1}, dx_n) \cdots K(x_k, dx_{k+1}).$$

Additionally, define $\mu(B|x_1, \ldots, x_k) = 1_A(x_1, \ldots, x_n)$ for $k \geq n$. It is easy to see that this definition does not depend on our choice of n, is a measurable function on S^k (by Lemma 11.1.4), and that for all $1 \leq j < k < \infty$ and $x_1, \ldots, x_k \in S$,

$$\mu(B | x_1, \ldots, x_j) \\ = \int \cdots \int \mu(B | x_1, \ldots, x_k) K(x_{k-1}, dx_k) \cdots K(x_j, dx_{j+1}), \tag{11.1.1}$$

and moreover,

$$\mu(B) = \int \mu(B|x) d\sigma(x). \tag{11.1.2}$$

Now, given our sequence $\{A_n\}_{n\geq 1}$, define

$$m_k(x_1,\ldots,x_k) := \lim_{n\to\infty} \mu(A_n|x_1,\ldots,x_k).$$

The limit exists because $\{A_n\}_{n\geq 1}$ is a decreasing sequence, and is measurable because it is the limit of a sequence of measurable functions. Analogously, define $m_0 := \lim_{n\to\infty}\mu(A_n)$. By (11.1.1), (11.1.2), and the dominated convergence theorem, we get that for any $1 \leq j < k$,

$$m_j(x_1,\ldots,x_j) = \int \cdots \int m_k(x_1,\ldots,x_k)K(x_{k-1},dx_k)\cdots K(x_j,dx_{j+1},$$

and

$$m_0 = \int m_1(x)d\sigma(x).$$

By assumption, $m_0 > 0$. Thus, there exists x_1 such that $m_1(x_1) > 0$. Therefore, there exists x_2 such that $m_2(x_1,x_2) > 0$, and so on. Let $x := (x_1,x_2,\ldots)$. We claim that $x \in \cap_{n=1}^{\infty} A_n$. To see this, take any n. Then $A_n = B \times S \times S \times \cdots$ for some $B \in S^N$, for some N (depending on A_n). Since $m_N(x_1,\ldots,x_N) > 0$, we have $\mu(A_k|x_1,\ldots,x_N) > 0$ for all k. In particular, $\mu(A_n|x_1,\ldots,x_N) > 0$. But

$$\mu(A_n|x_1,\ldots,x_N) = 1_B(x_1,\ldots,x_N).$$

Thus, $x \in B \times S \times S \times \cdots = A_n$. This proves that $x \in \cap_{n=1}^{\infty} A_n$, giving us the desired contradiction. Thus, μ extends to a probability measure on the infinite product σ-algebra.

Finally, to produce an infinite Markov chain with transition kernel K and initial distribution σ, take the space $S^{\mathbb{N}}$ with its product σ-algebra, and let $X_1, X_2, \ldots$ be the coordinate maps. Let $\mathcal{F}_n := \sigma(X_1,\ldots,X_n)$. Then note that for any measurable $D \subseteq S^n$ and $B \subseteq S$,

$$\begin{aligned}
&\int_{S^{n+1}} 1_D(x_1,\ldots,x_n)K(x_n,B)d\mu_n(x_1,\ldots,x_n,x_{n+1}) \\
&= \int_{S^n} 1_D(x_1,\ldots,x_n)K(x_n,B)d\mu_n(x_1,\ldots,x_n) \\
&= \int_{S^{n+1}} 1_{D\times B}(x_1,\ldots,x_n,x_{n+1})K(x_n,dx_{n+1})d\mu_n(x_1,\ldots,x_n) \\
&= \mu_{n+1}(D\times B) = \int_{S^{n+1}} 1_D(x_1,\ldots,x_n)1_B(x_{n+1})d\mu_{n+1}(x_1,\ldots,x_n,x_{n+1}).
\end{aligned}$$

This shows that $\mathbb{P}(X_{n+1} \in B|\mathcal{F}_n) = K(X_n,B)$. It is clear that $X_1 \sim \sigma$. This completes the proof. □

11.2. Markov chains on countable state spaces

In this section, we will study some basic properties of Markov chains on countable state spaces. Let S be a finite or countable set and let $\{X_n\}_{n\geq 0}$ be a Markov chain on S. The matrix $P = (p_{ij})_{i,j\in S}$ defined as

$$p_{ij} = \mathbb{P}(X_1 = j|X_0 = i)$$

is called 'transition matrix' of the chain. It is easy to see that the entries of P are nonnegative and each row sums to 1. Let $P^{(n)}$ be the transition matrix from time 0 to time n. That is,

the $(i,j)^{\text{th}}$ entry of $P^{(n)}$ is

$$p_{ij}^{(n)} = \mathbb{P}(X_n = j | X_0 = i).$$

This is sometimes called the 'n-step transition matrix'.

PROPOSITION 11.2.1 (Chapman–Kolmogorov formula). *For any n,*

$$P^{(n)} = P^n,$$

where the right side denotes the n^{th} power of P under matrix multiplication.

PROOF. First note that by applying the definition of Markov chain in the third step below,

$$\begin{aligned}
&\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) \\
&= \mathbb{P}(X_0 = i_0) \prod_{m=1}^{n} \mathbb{P}(X_m = i_m | X_0 = i_0, \ldots, X_{m-1} = i_{m-1}) \\
&= \mathbb{P}(X_0 = i_0) \prod_{m=1}^{n} \mathbb{P}(X_m = i_m | X_{m-1} = i_{m-1}) \\
&= \mathbb{P}(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}.
\end{aligned}$$

This shows that

$$\begin{aligned}
&\mathbb{P}(X_n = j | X_0 = i) \\
&= \sum_{i_1, \ldots, i_{n-1} \in S} \mathbb{P}(X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = j | X_0 = i) \\
&= \sum_{i_1, \ldots, i_{n-1} \in S} p_{i i_1} p_{i_1 i_2} \cdots p_{i_{n-1} j}.
\end{aligned}$$

But the last expression is just the $(i,j)^{\text{th}}$ element of P^n. Thus, $P^{(n)} = P^n$, as we wanted to show. □

A state $i \in S$ is called 'recurrent' if

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 | X_0 = i) = 1,$$

and 'transient' if the above probability is strictly less than 1. The following theorem gives an easily checkable condition for recurrence and transience.

THEOREM 11.2.2. *Let $\{X_n\}_{n \geq 0}$ be a time-homogeneous Markov chain on a finite or countably infinite state space S. Then a state $i \in S$ is recurrent if*

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty,$$

and transient if this sum is finite. Moreover, if i recurrent, the chain started at i returns to i infinitely many times with probability 1, and if i is transient, then the chain started at i returns to i only finitely many times with probability 1.

PROOF. Fix some state i. For proving the above theorem, it is convenient to define a random variable N which is the number of $n \geq 1$ such that $X_n = i$. (Note that we are ignoring $n = 0$. So if the chain starts from i but never comes back to i, then $N = 0$. Also

note that N is allowed to be ∞, in case the Markov chain visits i infinitely many times.) The monotone convergence theorem for conditional expectation implies that

$$\begin{aligned}\mathbb{E}(N|X_0 = i) &= \mathbb{E}\left(\sum_{n=1}^{\infty} 1_{\{X_n=i\}} \middle| X_0 = i\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(X_n = i|X_0 = i) = \sum_{n=1}^{\infty} p_{ii}^{(n)}. \end{aligned} \tag{11.2.1}$$

Define

$$p = \mathbb{P}(X_n = i \text{ for some } n \geq 1|X_0 = i).$$

The identity (11.2.1) implies that to prove the theorem, we have to show that $p < 1$ if $\mathbb{E}(N|X_0 = i) < \infty$ and $p = 1$ if $\mathbb{E}(N|X_0 = i) = \infty$. To prove this, it suffices to show that

$$\mathbb{E}(N|X_0 = i) = \sum_{k=1}^{\infty} p^k. \tag{11.2.2}$$

Note that by the monotone convergence theorem for conditional expectation,

$$\begin{aligned}\mathbb{E}(N|X_0 = i) &= \mathbb{E}\left(\sum_{k=1}^{\infty} 1_{\{N\geq k\}} \middle| X_0 = i\right) \\ &= \sum_{k=1}^{\infty} \mathbb{P}(N \geq k|X_0 = i).\end{aligned}$$

Thus, to prove (11.2.2), it suffices to show that

$$\mathbb{P}(N \geq k|X_0 = i) = p^k. \tag{11.2.3}$$

By the definition of p, this holds for $k = 1$. Let us assume that this holds for $k - 1$. For any $n \geq 1$ and any $j_1, \ldots, j_n \in S$ such that $j_n = i$ and exactly $k - 2$ of the other j_m's are equal to i, let $A_{j_1,\ldots,j_n}$ be the event $\{X_1 = j_1, \ldots, X_n = j_n\}$. It is not hard to see that the event $\{N \geq k - 1\}$ is the union of these events. Moreover, these events are disjoint. Since $\{N \geq k\} \subseteq \{N \geq k - 1\}$, this gives

$$\begin{aligned}&\mathbb{P}(N \geq k|X_0 = i) = \mathbb{P}(\{N \geq k\} \cap \{N \geq k - 1\}|X_0 = i) \\ &= \sum \mathbb{P}(\{N \geq k\} \cap A_{j_1,\ldots,j_n}|X_0 = i) \\ &= \sum \mathbb{P}(N \geq k|A_{j_1,\ldots,j_n} \cap \{X_0 = i\})\mathbb{P}(A_{j_1,\ldots,j_n}|X_0 = i)\end{aligned}$$

where the sum is over all n and $j_1, \ldots, j_n$ as above. Now, for any m and $i_1, \ldots, i_m$, let $B_{i_1,\ldots,i_m}$ be the event $\{X_{n+1} = i_1, \ldots, X_{n+m} = i_m\}$. Using the definition of Markov chain and the fact that $j_n = i$, it is easy to see that

$$\begin{aligned}&\mathbb{P}(B_{i_1,\ldots,i_m}|A_{j_1,\ldots,j_n} \cap \{X_0 = i\}) \\ &= \mathbb{P}(B_{i_1,\ldots,i_m}|X_n = i) = p_{ii_1}p_{i_1i_2} \cdots p_{i_{m-1}i_m}.\end{aligned}$$

Now, $\mathbb{P}(N \geq k|A_{j_1,\ldots,j_n} \cap \{X_0 = i\})$ is just the sum of the above quantity over all choices of m and $i_1, \ldots, i_m$ such that $i_m = i$ and $i_l \neq i$ for $l < m$. But on the other hand, the sum of $p_{ii_1}p_{i_1i_2} \cdots p_{i_{m-1}i_m}$ over the same set of m and $i_1, \ldots, i_m$ equals p. Thus, for any such

$j_1, \dots, j_n$,

$$\mathbb{P}(N \geq k | A_{j_1,\dots,j_n} \cap \{X_0 = i\}) = p.$$

Therefore,

$$\begin{aligned}\mathbb{P}(N \geq k | X_0 = i) &= \sum \mathbb{P}(N \geq k | A_{j_1,\dots,j_n} \cap \{X_0 = i\}) \mathbb{P}(A_{j_1,\dots,j_n} | X_0 = i) \\ &= p \sum \mathbb{P}(A_{j_1,\dots,j_n} | X_0 = i) \\ &= p\mathbb{P}(N \geq k-1 | X_0 = i).\end{aligned}$$

Since $\mathbb{P}(N \geq k-1 | X_0 = i) = p^{k-1}$, this completes the induction step.

To complete the proof of the final claim, simply observe that by (11.2.2), it follows that if i is recurrent, then $\mathbb{P}(N \geq k | X_0 = i) = 1$ for each k, which implies that $\mathbb{P}(N = \infty | X_0 = i) = 1$, and conversely, if i is transient, then $\mathbb{E}(N | X_0 = i) < \infty$, which implies that $\mathbb{P}(N < \infty | X_0 = i) = 1$. □

Theorem 11.2.2 gives necessary and sufficient conditions for recurrence and transience for a given state. What if we want such information for *all* states? Fortunately, we do not usually have to check separately for each state, as we will see below.

A state j of a Markov chain is said to be *accessible* from a state i if $p_{ij}^{(n)} > 0$ for some $n \geq 1$. (Note that we are leaving out $n = 0$.) A Markov chain on a finite state space is called *irreducible* if every state j is accessible from every state i.

Let us say that a state j is accessible from a state i in n steps if $p_{ij}^{(n)} > 0$. The following lemma is often useful for checking irreducibility.

LEMMA 11.2.3. *Take any $k \geq 1$, $i_0, i_1, \dots, i_k \in S$ and $n_1, \dots, n_k \geq 1$. Suppose that i_t is accessible from i_{t-1} in n_t steps, for $t = 1, \dots, k$. Let $n = n_1 + \cdots + n_k$. Then i_k is accessible from i_0 in n steps.*

PROOF. By the Chapman–Kolmogorov formula,

$$\begin{aligned}P^{(n)} &= P^n = P^{n_1} P^{n_2} \cdots P^{n_k} \\ &= P^{(n_1)} P^{(n_2)} \cdots P^{(n_k)}.\end{aligned}$$

Therefore

$$\begin{aligned}p_{i_0 i_k}^{(n)} &= \sum_{j_1,\dots,j_{k-1}} p_{i_0 j_1}^{(n_1)} p_{j_1 j_2}^{(n_2)} \cdots p_{j_{k-1} i_k}^{(n_k)} \\ &\geq p_{i_0 i_1}^{(n_1)} p_{i_1 i_2}^{(n_2)} \cdots p_{i_{k-1} i_k}^{(n_k)},\end{aligned}$$

which proves the claim. □

The following proposition shows that any two states that are accessible from each other have to be either both recurrent or both transient.

PROPOSITION 11.2.4. *If states i and j are accessible from each other, then either both i and j are recurrent, or both i and j are transient. Consequently, if a time-homogeneous irreducible Markov chain on a finite or countably infinite state space has one recurrent state, then all states are recurrent. Similarly, if one state is transient, then all states are transient.*

PROOF. Take any two states i and j such that j is accessible from i and i is accessible from j. Thus, $p_{ij}^{(l)} > 0$ and $p_{ji}^{(m)} > 0$ for some l and m. By the Chapman–Kolmogorov formula (which continues to hold on countable state spaces),

$$p_{jj}^{(n)} \geq p_{ji}^{(m)} p_{ii}^{(n-m-l)} p_{ij}^{(l)}$$

for any $n > m + l$. Summing both sides over n, we see that the recurrence of i implies the recurrence of j (by Theorem 11.2.2). Similarly, the recurrence of j implies the recurrence of i. For an irreducible chain, every state is accessible from every other state. Thus, either every state is recurrent or every state is transient. □

Because of Proposition 11.2.4, an irreducible chain as a whole may be called recurrent or transient. In the next section, we will see important examples of recurrent and transient chains.

11.3. Pólya's recurrence theorem

Simple symmetric random walk on $\mathbb{Z}^d$ is a Markov chain with state space $\mathbb{Z}^d$, and kernel $K(x, \cdot)$ = uniform distribution on the $2d$ nearest neighbors of x. That is, if the walk is at $x \in \mathbb{Z}^d$ at time n, it jumps to a uniformly chosen neighbor at time $n + 1$. It is easy to see that this chain is irreducible, and so we can ask whether the walk is recurrent or transient without mentioning the starting state. We will now prove the following famous result.

THEOREM 11.3.1 (Pólya's recurrence theorem). *Simple symmetric random walk on $\mathbb{Z}^d$ is recurrent if $d = 1$ or $d = 2$, and transient if $d \geq 3$.*

PROOF. Let r_n denote the probability that the chain returns to the origin at time n, given that it started from the origin at time 0. The number of ways this can happen is counted as follows. First, suppose that the walk takes n_i steps in the i^{th} coordinate direction (either forward or backward). Here $n_1 + \cdots + n_d = n$. The number of ways of allocating these steps is

$$\frac{n!}{n_1! n_2! \cdots n_d!}.$$

Having associated each step of the walk to a coordinate direction, we now have to decide whether the move is forward or backward. If the walk has to return to the origin at time n, the number of forward steps in any coordinate direction must be equal to the number of backward steps. For direction i, this allocation can be done in $\binom{n_i}{n_i/2}$ ways, provided that n_i is even. Otherwise, this is zero. Combining, we see that the number of paths that return to the origin at time n is

$$\sum_{\substack{n_1,\ldots,n_d \text{ even}, \\ n_1+\cdots+n_d=n}} \frac{n!}{n_1! n_2! \cdots n_d!} \prod_{i=1}^{d} \binom{n_i}{n_i/2}. \tag{11.3.1}$$

Therefore, r_n equals the above quantity times $(2d)^{-n}$. When $d = 1$, this gives

$$r_{2n} = \binom{2n}{n} 2^{-2n},$$

for any positive integer n. By Stirling's formula, this is asymptotically

$$\frac{\sqrt{2\pi}(2n)^{2n+\frac{1}{2}}e^{-2n}2^{-2n}}{(\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n})^2} = \frac{1}{\sqrt{\pi n}}.$$

In other words,

$$\lim_{n\to\infty} \frac{r_{2n}}{(\pi n)^{-1/2}} = 1.$$

By the ratio test for summability of a series, this proves that

$$\sum_{n=1}^{\infty} r_{2n} = \infty.$$

Therefore by Theorem 11.2.2, simple symmetric random walk on $\mathbb{Z}$ is recurrent.

Next, let us consider $d = 2$. By the formula (11.3.1), we get that if n is even, then

$$r_n = \sum_{\substack{k,l \text{ even},\\ k+l=n}} \frac{n!4^{-n}}{((k/2)!(l/2)!)^2}.$$

This can be rewritten as

$$r_{2n} = \sum_{k=0}^{n} \frac{(2n)!4^{-2n}}{(k!(n-k)!)^2} = \binom{2n}{n}4^{-2n}\sum_{k=0}^{n}\binom{n}{k}^2.$$

Now, since

$$\binom{n}{k} = \binom{n}{n-k},$$

it follows that

$$\sum_{k=0}^{n}\binom{n}{k}^2$$

is the coefficient of x^n in the expansion of $(1+x)^n(1+x)^n$. But the coefficient of x^n in $(1+x)^{2n}$ is $\binom{2n}{n}$. Thus,

$$\sum_{k=0}^{n}\binom{n}{k}^2 = \binom{2n}{n}.$$

This shows that

$$r_{2n} = \binom{2n}{n}^2 4^{-2n}.$$

But this is just the square of r_{2n} in the one-dimensional case. Thus, this is asymptotic to $(\pi n)^{-1}$. Again by the ratio test, this proves the divergence of $\sum_{n=1}^{\infty} r_{2n}$ and hence the recurrence of the simple symmetric random walk on $\mathbb{Z}^2$.

Finally, let us turn our attention to the case $d \geq 3$. First, note that any path of length $2n$ that starts and ends at the origin can be extended to a path of length $2n+2$ by adding two steps at the end, the first one moving one step away from the origin in a fixed direction and the next one moving back to the origin. Since this gives a one-to-one map from the set of path of length $2n$ that start and end at the origin, into the set of paths of length $2n+2$ that start and end at the origin, this shows that if N_{2n} is the number of paths of length $2n$ that start and end the origin, then $N_{2n} \leq N_{2n+2}$. Consequently, N_{2n} is a non-decreasing function of n. But $r_{2n} = N_{2n}(2d)^{-2n}$. Thus, for any $k \geq 1$,

$$r_{2n} \leq N_{2n+2k}(2d)^{-2n} = r_{2n+2k}(2d)^{2k}.$$

Therefore,

$$\begin{aligned}\sum_{n=1}^{\infty} r_{2n} &= \sum_{m=1}^{\infty} (r_{2dm-2d+2} + r_{2dm-2d+4} + \cdots + r_{2dm}) \\ &\leq \sum_{m=1}^{\infty} r_{2dm}((2d)^{2d-2} + (2d)^{2d-4} + \cdots + 1)\end{aligned}$$

Thus, to prove $\sum_{n=1}^{\infty} r_{2n} < \infty$, it suffices to show that

$$\sum_{n=1}^{\infty} r_{2dn} < \infty.$$

Now from (11.3.1), we get

$$\begin{aligned}r_{2dn} &= \sum_{\substack{k_1,\ldots,k_d \\ k_1+\cdots+k_d=dn}} \frac{(2dn)!(2d)^{-2dn}}{(k_1!k_2!\cdots k_d!)^2} \\ &= \binom{2dn}{dn} 2^{-2dn} \sum_{\substack{k_1,\ldots,k_d \\ k_1+\cdots+k_d=dn}} \left(\frac{(dn)!d^{-dn}}{k_1!k_2!\cdots k_d!}\right)^2. \qquad (11.3.2)\end{aligned}$$

Take any nonnegative integers $k_1, \ldots, k_d$ that sum to dn. Writing $k_i!$ as the product of $1, 2, \ldots, k_i$, we can write $k_1!k_2!\cdots k_d!$ as a product of integers between 1 and dn, possibly with many repeats. Using the condition $k_1 + \cdots + k_d = dn$, little bit of thought shows that in this product we can replace the integers bigger then n by integers less than n in such a way that the end result is exactly equal to $(n!)^d$. (If some k_i is less than n, it 'falls short' of contributing $n - k_i$ terms to $n!$. On the other hand, if some k_i is bigger than n, it contributes $k_i - n$ extra terms that are all bigger than n. Since $k_1 + \cdots + k_d = dn$, the number of missing contributions exactly equals the number excess terms.) This procedure shows that

$$k_1!k_2!\cdots k_d! \geq (n!)^d.$$

Thus,

$$\begin{aligned}\sum_{\substack{k_1,\ldots,k_d \\ k_1+\cdots+k_d=dn}} \left(\frac{(dn)!d^{-dn}}{k_1!k_2!\cdots k_d!}\right)^2 &\leq \frac{(dn)!d^{-dn}}{(n!)^d} \sum_{\substack{k_1,\ldots,k_d \\ k_1+\cdots+k_d=dn}} \frac{(dn)!d^{-dn}}{k_1!k_2!\cdots k_d!} \\ &= \frac{(dn)!d^{-dn}}{(n!)^d},\end{aligned}$$

where the last identity was obtained by observing that the sum in the previous line is the sum of a multinomial probability mass function. By Stirling's formula, the above quantity is asymptotic to

$$\frac{\sqrt{2\pi}(dn)^{dn+\frac{1}{2}}e^{-dn}d^{-dn}}{(\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n})^d} = \frac{\sqrt{d}}{(2\pi)^{(d-1)/2}n^{(d-1)/2}}.$$

On the other hand, by our calculations in the one-dimensional case, the term

$$\binom{2dn}{dn} 2^{-2dn}$$

in (11.3.2) is asymptotic to $(\pi dn)^{-1/2}$. Thus, r_{2dn} is bounded above by a quantity that is asymptotic to $C(d)n^{-d/2}$, where $C(d)$ is a constant that depends only on d. Since

$$\sum_{n=1}^{\infty} n^{-d/2} < \infty$$

when $d \geq 3$, this proves the summability of r_{2dn}, and so by our previous discussion, the summability of r_{2n}. Since $r_n = 0$ when n is odd, the completes the proof of transience of simple symmetric random walk in $d \geq 3$. □

11.4. Markov chains on finite state spaces

Let $\{X_n\}_{n\geq 0}$ be a time-homogeneous Markov chain on a finite state space S with transition matrix P. A probability measure on S is simply a set of values $\mu = (\mu_i)_{i\in S}$ such that $\mu_i \geq 0$ for each i, and $\sum_{i\in S} \mu_i = 1$. We will think of μ as a row vector in $\mathbb{R}^N$, where N is the size of S. This deviates from the usual convention of thinking of any vector as a column vector, but it is the usual practice in the Markov chain world, and is convenient for various purposes.

A probability distribution μ on S is called an *invariant distribution* (also called a stationary distribution or an equilibrium distribution) for the above Markov chain if $\mu P = \mu$. In other words, for each $j \in S$,

$$\sum_{i\in S} \mu_i p_{ij} = \mu_j.$$

The significance of an invariant distribution is that if the distribution of X_0 is an invariant distribution, then the distribution of each X_n is the same as that of X_0. To see this, suppose that $X_0 \sim \mu$, where μ is an invariant distribution. Let $\mu^{(n)}$ be the distribution of X_n. Then note that

$$\begin{aligned}\mu_j^{(n)} = P(X_n = j) &= \sum_{i\in S} P(X_n = j|X_0 = i)P(X_0 = i)\\ &= \sum_{i\in S} P(X_n = j|X_0 = i)\mu_i.\end{aligned}$$

By the Chapman–Kolmogorov formula, this gives

$$\mu^{(n)} = \mu P^n.$$

But $\mu P = P$. Thus, $\mu^{(n)} = \mu P^n = \mu$ for any n. The following result shows that a time-homogeneous Markov chain on a finite state space always has at least one invariant distribution.

THEOREM 11.4.1. *Let $\{X_n\}$ be a time-homogeneous Markov chain on a finite state space S, with transition matrix P. Then there is at least one invariant distribution μ for this chain.*

PROOF. Take any probability distribution μ_0 on S. For each n, define

$$\mu_n = \frac{\mu_0 + \mu_0 P + \mu_0 P^2 + \cdots + \mu_0 P^n}{n+1}.$$

It is an easy consequence of the Chapman–Kolmogorov formula (as observed above) that $\mu_0 P^k$ is the distribution of X_k if $X_0 \sim \mu_0$. Since the average of a set of probability distributions is again a probability distribution, μ_n must be a probability measure on S. Now note that

$$\begin{aligned}
&\mu_n - \mu^{(n)} P \\
&= \frac{\mu_0 + \mu_0 P + \mu_0 P^2 + \cdots + \mu_0 P^n}{n+1} \\
&\qquad - \frac{\mu_0 P + \mu_0 P^2 + \mu_0 P^2 + \cdots + \mu_0 P^{n+1}}{n+1} \\
&= \frac{\mu_0 - \mu_0 P^{n+1}}{n+1}.
\end{aligned}$$

Since μ_0 and $\mu_0 P^{n+1}$ are both probability measures, the last expression tends to the zero vector as $n \to \infty$. Thus,

$$\lim_{n\to\infty} (\mu_n - \mu_n P) = 0$$

where the 0 on the right is the zero vector. Finally, note that since $\{\mu_n\}_{n\geq 0}$ is a bounded sequence (because each vector is a probability measure), it has a convergent subsequence. Let μ be a limit of such a subsequence. The above equation proves that $\mu P = \mu$. The set of probability measures on S is a closed set, which implies that μ must be a probability measure. Thus, μ is an invariant distribution. □

Let $\{X_n\}_{n\geq 1}$ be a time-homogeneous Markov chain on a finite state space S, with transition matrix P. Let μ be an invariant distribution for P. The Doeblin condition is a condition under which we can identify the limit of P^n as $n \to \infty$, together with a rate of convergence.

Let M be the square matrix whose rows are all equal to μ. We say that P satisfies the *Doeblin condition* if for some $k \geq 1$ and some $\epsilon \in (0, 1)$,

$$P^k \geq \epsilon M, \tag{11.4.1}$$

where the inequality means that each entry of the matrix on the left is greater than or equal to the corresponding entry of the matrix on the right. Under this condition, we can prove that P^n converges to M as $n \to \infty$. In other words, $P(X_n = j | X_0 = i)$ converges to μ_j as $n \to \infty$, irrespective of the value of i.

To state Doeblin's theorem with a rate of convergence, we need to define a matrix norm. There are many matrix norms. The one that will be useful for us is the following. For a matrix $A = (a_{ij})_{i,j\in S}$, define

$$\|A\| = \max_{i\in S} \sum_{j\in S} |a_{ij}|.$$

It is not hard to check that for any two matrices A and B,

$$\|A + B\| \leq \|A\| + \|B\|, \quad \|AB\| \leq \|A\|\|B\|.$$

We are now ready to state Doeblin's theorem.

THEOREM 11.4.2 (Doeblin's theorem). *Suppose that P satisfies the Doeblin condition (11.4.1) for some $k \geq 1$ and some $\epsilon \in (0,1)$. Then for any n,*

$$\|P^n - M\| \leq 2(1-\epsilon)^{[n/k]},$$

where $[n/k]$ denotes the integer part of n/k.

PROOF. Define

$$Q = \frac{P^k - \epsilon M}{1-\epsilon}.$$

Since P^k and M are both stochastic matrices, it follows that the each row of Q sums to 1. Moreover, by the Doeblin condition, the entries of Q are nonnegative. Thus, Q is a stochastic matrix. Now, note that

$$P^k = (1-\epsilon)Q + \epsilon M.$$

Also note that since μ is a probability distribution, $\mu P = P$, and P is a stochastic matrix, we have the identities

$$M^2 = M, \quad MP = M, \quad PM = M.$$

This implies that $QM = MQ = M$. Consequently, any product of a sequence of Q's and M's that contains at least one M must be equal to M. Thus, for any $m \geq 1$, the distributive law for matrix products gives

$$\begin{aligned} P^{km} &= ((1-\epsilon)Q + \epsilon M)^m \\ &= (1-\epsilon)^m Q^m + \sum_{k=1}^{m} \binom{m}{k} (1-\epsilon)^{m-k}\epsilon^k M \\ &= (1-\epsilon)^m Q^m + (1-(1-\epsilon)^m)M. \end{aligned}$$

Now take any $n \geq 1$. Let $m = [n/k]$, so that $n = km + r$, where $0 \leq r \leq k-1$. Then

$$\begin{aligned} P^n - M &= P^r(P^{km} - M) \\ &= (1-\epsilon)^m P^r (Q^m - M). \end{aligned}$$

Since P^r, Q and M are stochastic matrices, this gives

$$\|P^n - M\| \leq (1-\epsilon)^m \|P^r\|(\|Q^m\| + \|M\|) \leq 2(1-\epsilon)^m,$$

which completes the proof. □

The *period* of a state $i \in S$ is defined to be greatest common divisor of the set of all $n \geq 1$ such that $p_{ii}^{(n)} > 0$. If there is no such n, the period is ∞. A Markov chain is called *aperiodic* if all of its states have period 1. The following theorem is the main result of this section.

THEOREM 11.4.3. *Let $\{X_n\}_{n\geq 0}$ be an irreducible aperiodic Markov chain on a finite state space S, with transition matrix P. Then P has a unique invariant distribution μ. Moreover, if M is the square matrix whose rows are all equal to μ, then $P^n \to M$ as $n \to \infty$.*

The key idea is to show that any irreducible aperiodic chain on a finite state space satisfies the Doeblin condition. The first step is to prove the following lemma.

LEMMA 11.4.4. *For any state i, $p_{ii}^{(n)} > 0$ for all sufficiently large n.*

PROOF. By irreducibility, we know that $P(X_n = i|X_0 = i) > 0$ for some $n \geq 1$. Let A be the set of all such n. Since the g.c.d. of A is 1 (because the chain is aperiodic), there must exist $a_1, \ldots, a_k \in A$ and $u_1, \ldots, u_k \in \mathbb{Z}$ for some $k \geq 1$ such that

$$u_1 a_1 + \cdots + u_k a_k = 1.$$

Define a positive integer

$$m = a_1 + \cdots + a_k.$$

Take any $n \geq 1$. Let q and r be the quotient and remainder when n is divided by m. Then

$$\begin{aligned} n = qm + r &= qm + r(u_1 a_1 + \cdots + u_k a_k) \\ &= (q + ru_1)a_1 + \cdots (q + ru_k)a_k. \end{aligned}$$

Since r is always bounded above by $m - 1$, and $q \to \infty$ as $n \to \infty$, the above identity shows that any sufficiently large n is expressible as a linear combination of $a_1, \ldots, a_k$ with positive integer coefficients. The claim now follows from Lemma 11.2.3. □

By irreducibility, the above lemma generalizes to the following.

LEMMA 11.4.5. *There is some n_0 such that for any $n \geq n_0$, the entries of P^n are all strictly positive.*

PROOF. Take any state i. By Lemma 11.4.4, there is some m such that $p_{ii}^{(n)} > 0$ for all $n \geq m$. By irreducibility and the finiteness of the state space, there is some $k \geq 1$ such that any j is accessible from i in less than k steps. Let $m' = m + k$. Take any $n \geq m'$ and any state j. Then there is a number $r < k$ such that j is accessible from i in r steps. Also, since $n - r > n - k \geq m$, the state i is accessible from itself in $n - r$ steps. Therefore by Lemma 11.2.3, j is accessible from i in n steps. Thus, we have shown that any state is accessible from state i in n steps whenever n is sufficiently large. From this it follows that when n is sufficiently large, $p_{ij}^{(n)} > 0$ for all i and j, which is what we wanted to prove. □

By the finiteness of the state space, Lemma 11.4.5 immediately yields the following corollary, because if the entries of P^n are all strictly positive, then for any given matrix M, there is some $\epsilon > 0$ such that $P^n \geq \epsilon M$.

COROLLARY 11.4.6. *Any irreducible aperiodic Markov chain on a finite state space satisfies the Doeblin condition.*

It is now easy to finish the proof of Theorem 11.4.3. By Theorem 11.4.1, there is at least one invariant distribution μ. Let M be the matrix whose rows are all equal to μ. By the above corollary, $P^n \to M$ as $n \to \infty$. To prove the uniqueness of the invariant distribution, let ν be another invariant distribution. Then on the one hand, $\nu P^n = \nu$ for every n. On the other hand,

$$\lim_{n \to \infty} \nu P^n = \nu M = \mu,$$

since ν is a probability distribution and each column of M is a scalar multiple of the vector of all 1's. Thus, $\mu = \nu$.

EXAMPLE 11.4.7 (Shuffling cards). Consider a deck of n cards, shuffled according to the following rule. At each step, the card on the top is moved to a uniformly chosen location in the deck (which may be the top position too). This is known as the 'top-to-random' shuffle. Here the state space is the space S_n of all permutations of $1, \ldots, n$.

First, note that any state can be accessed from itself in 1 step, which implies that the chain is aperiodic. Next, note that any state can be accessed from any other state by a sequence of 'allowed' moves. Therefore by Lemma 11.2.3, the chain is irreducible. Thus, by Theorem 11.4.3, this Markov chain has a unique invariant distribution μ, and irrespective of the starting state, the distribution of X_n tends to μ as $n \to \infty$.

To complete the discussion, let us identify the invariant distribution μ. We claim that μ is the uniform distribution on S_n. To see this, note that any state $\sigma \in S_n$ can be accessed from exactly n other states $\pi_1, \ldots, \pi_n$ in 1 step, and each of these n states has probability $1/n!$ under the uniform distribution. Moreover, the transition from π_i to σ in one step has probability $1/n$. Thus, if X_0 is uniformly distributed on S_n, then

$$P(X_1 = \sigma) = \sum_{i=1}^{n} P(X_1 = \sigma | X_0 = \pi_i) P(X_0 = \pi_i) = \frac{1}{n!}.$$

Since this holds for any σ, X_1 is again uniformly distributed on S_n. Thus, the uniform distribution is an invariant distribution for this chain. The claim now follows by the uniqueness of the invariant distribution.

EXAMPLE 11.4.8 (Random walk on a graph). Let $G = (V, E)$ be a finite, simple, undirected graph with vertex set V and edge E. A simple random walk $\{X_n\}_{n \geq 0}$ on G is a Markov chain on V that evolves according to the following rule. If the chain is at a vertex v, it chooses one of its neighbors uniformly at random and moves there. In general, simple random walks may not be aperiodic (for example, if the graph is bipartite). There is an easy fix for this problem. We just make the walk stay where it is with some fixed probability $p > 0$, and jump to a uniformly chosen neighbor with probability $1 - p$. This is known as a *lazy* simple random walk on G with holding probability p. Lazy walks are alway aperiodic.

Recall that a graph is called connected if any vertex can be reached from any other by moving along a sequence of edges. By Lemma 11.2.3, a simple random walk (or a lazy simple random walk) on a connected graph is irreducible.

Combining the above observations, we get that a lazy simple random walk on a finite connected graph is irreducible and aperiodic. Therefore by Theorem 11.4.3, there is a unique invariant distribution and the distribution of X_n converges to this invariant distribution from any starting state.

It remains to identify the invariant distribution. Recall that the degree of a vertex v, denoted by d_v, is the number of neighbors of v. We claim the invariant distribution is given by

$$\mu_v = \frac{d_v}{\sum_{u \in V} d_u}.$$

Let us verify this by direct calculation. First, note that this is indeed a probability distribution, since $\mu_v \geq 0$ for all v and $\sum_{v \in V} \mu_v = 1$. Next, let p_{uv} be the probability of

transition from u to v. That is,

$$p_{uv} = \begin{cases} p & \text{if } u = v, \\ (1-p)/d_u & \text{if } v \text{ is a neighbor of } u, \\ 0 & \text{otherwise.} \end{cases}$$

Then note that for any v,

$$\sum_{u \in V} \mu_u p_{uv} = p\mu_v + (1-p) \sum_{u \in N_v} \frac{\mu_u}{d_u},$$

where N_v is the set of neighbors of v. Plugging in the formula for μ_u, we see that the right side equals

$$\begin{aligned} & p\mu_v + (1-p) \sum_{u \in N_v} \frac{1}{d_u} \frac{d_u}{\sum_{w \in V} d_w} \\ &= p\mu_v + (1-p) \sum_{u \in N_v} \frac{1}{\sum_{w \in V} d_w} \\ &= p\mu_v + (1-p) \frac{d_v}{\sum_{w \in V} d_w} = p\mu_v + (1-p)\mu_v = \mu_v. \end{aligned}$$

Thus, μ is indeed the unique invariant distribution for this Markov chain.

CHAPTER 12

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.

12.1. 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 8.7.1.

DEFINITION 12.1.1. 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 12.1.2. Prove that the assertion of Exercise 6.1.1 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 12.1.3. 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 12.1.4. 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 12.1.5. Generalize the above exercise to $\mathbb{R}^n$-valued continuous functions on $[0,\infty)$.

EXERCISE 12.1.6. 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 (that is, it's a measurable map from the sample space into the real line).

EXERCISE 12.1.7. 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 12.1.8. 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 12.1.7 and the measurability assertion from Fubini's theorem.)

EXERCISE 12.1.9. Let X be a $C[0,1]$-valued random variable. Prove that $\max_{0\le t\le 1} X(t)$ is a random variable.

12.2. 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 12.2.1 (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 $F \subseteq S$,*

$$\limsup_{n\to\infty} \mu_n(F) \le \mu(F).$$

(e) *For every open set $V \subseteq S$,*

$$\liminf_{n\to\infty} \mu_n(V) \ge \mu(V).$$

(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

$$f(x) = \rho(x, F) := \inf_{y\in F} \rho(x, y). \tag{12.2.1}$$

Note that for any $x, x' \in S$ and $y \in F$, the triangle inequality gives $\rho(x, y) \le \rho(x, x') + \rho(x', y)$. Taking infimum over y gives $f(x) \le \rho(x, x') + f(x')$. Similarly, $f(x') \le \rho(x, x') + f(x)$. Thus,

$$|f(x) - f(x')| \le \rho(x, x'). \tag{12.2.2}$$

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 $[0, 1]$, $1_F \le g_k$ everywhere, and $g_k \to 1_F$ pointwise as $k \to \infty$. Therefore by (c),

$$\limsup_{n\to\infty} \mu_n(F) \le \limsup_{n\to\infty} \int g_k d\mu_n = \int g_k d\mu$$

for any k, and hence by the dominated convergence theorem,

$$\limsup_{n\to\infty} \mu_n(F) \le \lim_{k\to\infty} \int g_k d\mu = \mu(F),$$

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 $\mu(\partial A_t) > 0$ 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 12.2.2. *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, $\mu(F) \leq \nu(F)$ and $\nu(F) \leq \mu(F)$ for every closed set F. Thus, by Theorem 1.3.6, $\mu = \nu$. □

EXERCISE 12.2.3. 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 12.2.4. Let $\{X_n\}_{n=1}^\infty$ be a sequence of $C[0,1]$-valued random variables converging weakly to some X. Then prove that

$$\max_{0\leq t\leq 1} X_n(t) \xrightarrow{d} \max_{0\leq t\leq 1} X(t).$$

Exercise 12.2.6 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 12.2.5. 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 12.2.6. 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 12.2.7. 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. $t \in [0,1]$.

EXERCISE 12.2.8. Suppose that S and S' are Polish spaces and $h : S \to S'$ is a measurable map. Let $X_1, X_2, \ldots$ be a sequence of S-valued random variables converging

weakly to an S-valued random variable X. Let D be the set of discontinuity points of h. If $\mathbb{P}(X \in D) = 0$, show that $h(X_1), h(X_2), \ldots$ converge weakly to $h(X)$.

12.3. Tightness and Prokhorov's theorem

The appropriate generalization of the notion of tightness to probability measures on Polish spaces is the following.

DEFINITION 12.3.1. 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 12.3.2. *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 $\mu_{n_i}(V_i) \leq 1 - \epsilon$. There cannot exist k such that $n_i = k$ for infinitely many i, because then $\mu_k(V_i) \to 1$ as $i \to \infty$. Thus, $n_i \to \infty$ as $i \to \infty$. Therefore by the portmanteau lemma, for each i,

$$\mu(V_i) \leq \liminf_{j\to\infty} \mu_{n_j}(V_i) \leq \liminf_{j\to\infty} \mu_{n_j}(V_j) \leq 1 - \epsilon.$$

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. □

EXERCISE 12.3.3. Using Theorem 12.3.2, prove that any probability measure on a Polish space is regular, in the sense of Lemma 3.4.2.

EXERCISE 12.3.4. Using the previous exercise, prove the analogue of Kolmogorov's extension theorem (Theorem 3.4.1) for Polish spaces.

EXERCISE 12.3.5. Using the previous exercise, prove the existence of Markov chains with any given kernel on a Polish space (i.e., the analogue of Theorem 11.1.3).

Helly's selection theorem generalizes to Polish spaces. The generalization is known as Prokhorov's theorem.

THEOREM 12.3.6 (Prokhorov's theorem). *Let (S, ρ) be a Polish space. Suppose that $\{\mu_n\}_{n=1}^{\infty}$ 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 12.3.7. *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 12.3.8. *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 12.3.9. *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 (12.2.1) for any A. It is not hard to see (using (12.2.2), 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 12.3.8, 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 12.3.7,

$$\alpha(C) \leq \alpha(C_1) + \alpha(C_2) \leq \beta(V_1) + \beta(V_2).$$

Taking supremum over C completes the proof. □

LEMMA 12.3.10. *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 12.3.9,

$$\alpha(C) \leq \beta(V_1 \cup \cdots \cup V_n) \leq \sum_{i=1}^{n} \beta(V_i) \leq \sum_{i=1}^{\infty} \beta(V_i).$$

Taking supremum over C completes the proof. □

LEMMA 12.3.11. *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 $A_1, A_2, \ldots$ contained in 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 12.3.10,

$$\mu(A) \leq \beta\left(\bigcup_{i=1}^{\infty} V_i\right) \leq \sum_{i=1}^{\infty} \beta(V_i)$$
$$\leq \sum_{i=1}^{\infty} (\mu(A_i) + 2^{-i}\epsilon) = \epsilon + \sum_{i=1}^{\infty} \mu(A_i).$$

Since ϵ is arbitrary, this completes the proof. □

LEMMA 12.3.12. *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 12.3.7,

$$\beta(V) \geq \alpha(C_1 \cup C_2) = \alpha(C_1) + \alpha(C_2).$$

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 12.3.6. 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 12.3.12 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 12.3.13. 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.

Exercise 12.3.14. Let S be the set of all functions from $\mathbb{Z}$ into $\mathbb{R}$. Equip S with the topology of pointwise convergence. Prove that this topology is metrizable, and under a compatible metric, it is a Polish space. If $(X_i)_{i\in\mathbb{Z}}$ is a sequence of real-valued random variables defined on some probability space, show that the whole sequence can be viewed as an S-valued random variable. If $(X_n)_{n\geq 1}$ is a sequence of S-valued random variables, where $X_n = (X_{n,i})_{i\in\mathbb{Z}}$, prove that it is tight if and only if for each $i \in \mathbb{Z}$, the sequence of real-valued random variables $(X_{n,i})_{n\geq 1}$ is tight.

Exercise 12.3.15. Let $\{\mu_n\}_{n\geq 1}$ be a tight family of probability measures on a Polish space S. Suppose that any convergent subsequence converges to the same limit μ. Using Prokhorov's theorem, prove that $\mu_n \to \mu$ weakly. (Hint: Use contradiction.)

EXERCISE 12.3.16 (The Lévy–Prokhorov metric). Let (S, ρ) be a Polish space. For any $A \subseteq S$ and $\epsilon > 0$, let

$$A^\epsilon := \{x \in S : \rho(x, y) < \epsilon \text{ for some } y \in A\}.$$

For any two probability measures μ and ν on S, define

$$d(\mu, \nu) := \inf\{\epsilon > 0 : \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon \text{ for all } A \in \mathcal{B}(S)\},$$

where $\mathcal{B}(S)$ is the Borel σ-algebra of S. Prove that d is a metric on the set of probability measures on S, and $\mu_n \to \mu$ weakly if and only if $d(\mu_n, \mu) \to 0$. (Thus, d metrizes weak convergence of probability measures.)

12.4. 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 12.4.1 (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 12.4.2. *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 $2 \leq i \leq n$. 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 $\operatorname{diam}(A_i) \leq \operatorname{diam}(B_i) \leq \epsilon$ for $1 \leq i \leq n$. Finally, note that $\partial A_i \subseteq \partial B_1 \cup \cdots \cup \partial B_i$ for $1 \leq i \leq n$, 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 12.4.1. For each $j \geq 1$, choose sets $A_0^j, \ldots, A_{k_j}^j$ satisfying the conditions of Lemma 12.4.2 with $\epsilon = 2^{-j}$ and μ as in the statement of Theorem 12.4.1.

Next, for each j, find n_j such that if $n \geq n_j$, then

$$\mu_n(A_i^j) \geq (1 - 2^{-j})\mu(A_i^j) \tag{12.4.1}$$

for all $0 \leq i \leq k_j$. 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 $n \geq n_1$, and find j such that $n_j \leq n < n_{j+1}$. For $0 \leq i \leq k_j$, 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 $1 \leq i \leq k_j$ and some arbitrary probability measure on S of $i = 0$. This can be done because A_i^j is nonempty for $i \geq 1$. 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 (12.4.1), $p_{n,i} \geq 0$. 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 $A \in \mathcal{B}(S)$,

$$\begin{aligned}\mathbb{P}(X_n \in A) &= \mathbb{P}(U > 2^{-j}) \sum_{i=0}^{k_j} \mathbb{P}(X \in A_i^j)\mathbb{P}(Y_{n,i} \in A) \\ &\quad + \mathbb{P}(U \le 2^{-j})\mathbb{P}(Z_n \in A) \\ &= (1 - 2^{-j}) \sum_{i=0}^{k_j} \mu(A_i^j)\mu_{n,i}(A) + 2^{-j} \sum_{i=0}^{k_j} \mu_{n,i}(A) p_{n,i} \\ &= \sum_{i=0}^{k_j} \mu_n(A_i^j)\mu_{n,i}(A) = \sum_{i=0}^{k_j} \mu_n(A \cap A_i^j) = \mu_n(A).\end{aligned}$$

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, $\rho(X_n, X) \le \epsilon$, because $\operatorname{diam}(A_i^j) \le \epsilon$. This proves that $X_n \to X$ a.s. □

12.5. 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 $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\rho(X_n, X) \ge \epsilon) = 0.$$

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 12.5.1. *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 $|f(x) - f(y)| \le \epsilon$ whenever $\rho(x, y) \le \delta$. Thus,

$$\mathbb{P}(|f(X_n) - f(X)| > \epsilon) \le \mathbb{P}(\rho(X_n, X) > \delta),$$

which shows that $f(X_n) \to f(X)$ in probability. Since f is bounded, Proposition 8.2.9 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 12.5.2. *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 12.5.1 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 12.5.3 (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 8.2.9, 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. □

12.6. Multivariate inversion formula

The inversion formula for characteristic functions of random vectors is analogous to the univariate formula that was presented in Theorem 8.8.1. 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 12.6.1. *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*

$$f_\theta(x) := \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-\mathrm{i}t\cdot x - \theta|t|^2} \phi(t) dt.$$

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 8.8.1, 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 8.8.2, the above theorem has the following corollary about random vectors.

COROLLARY 12.6.2. *Two random vectors have the same law if and only if they have the same characteristic function.*

PROOF. Same as the proof of Corollary 8.8.2, using Theorem 12.6.1 and Corollary 12.2.2 instead of Theorem 8.8.1 and Proposition 8.7.1. □

The above corollary has the following important consequences.

COROLLARY 12.6.3. *Let $(X_1, \ldots, X_n)$ is a random vector with characteristic function ϕ. Let ϕ_i be the characteristic function of X_i. Then $\phi(t_1, \ldots, t_n) = \phi_1(t_1) \cdots \phi_n(t_n)$ for all $t_1, \ldots, t_n$ if and only if $X_1, \ldots, X_n$ are independent.*

PROOF. Let $Y_1, \ldots, Y_n$ be independent random variables, with Y_i having the same distribution as X_i for each i. Let ψ be the characteristic function of the random vector $(Y_1, \ldots, Y_n)$. By Exercise 7.6.3, $\psi(t_1, \ldots, t_n) = \phi_1(t_1) \cdots \phi_n(t_n)$. The claim now follows by Corollary 12.6.2. □

COROLLARY 12.6.4. *Let X and Y be real-valued random variables defined on the same probability space. Then X and Y are independent if and only if $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$ for any bounded continuous $f, g : \mathbb{R} \to \mathbb{R}$.*

PROOF. If X and Y are independent, we know that by the product rule for expectation, $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$ for any bounded continuous $f, g : \mathbb{R} \to \mathbb{R}$. Conversely, suppose that this criterion holds. Considering real and imaginary parts, we can assume that the criterion holds even if f and g are complex valued. Then taking $f(x) = e^{\mathrm{i}sx}$ and $g(y) = e^{\mathrm{i}ty}$ for arbitrary $s, t \in \mathbb{R}$, we get that the characteristic function of (X, Y) is the product of the characteristic functions of X and Y. Therefore by Corollary 12.6.3, X and Y are independent. □

EXERCISE 12.6.5. Prove a multivariate version of Corollary 8.8.7.

EXERCISE 12.6.6. Let X be a real-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let ϕ be a characteristic function. Prove that X is independent of $\mathcal{G}$ and has characteristic function ϕ if and only if $\mathbb{E}(e^{\mathrm{i}tX}; A) = \phi(t)\mathbb{P}(A)$ for all $t \in \mathbb{R}$ and all $A \in \mathcal{G}$. (Hint: Show that X and 1_A are independent random variables for all $A \in \mathcal{G}$.)

12.7. 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 12.7.1 (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 12.6.2. □

EXERCISE 12.7.2. 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 $(X_{n,1}, \ldots, X_{n,m})$ converges weakly to $(X_1, \ldots, X_m)$ as random vectors on S^n.

12.8. 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 12.8.1 (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. □

12.9. 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 12.9.1 (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 7.6.8, $Y \sim N(0, t^T \Sigma t)$. 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. □

12.10. Independence on Polish spaces

Polish space valued random variables are said to be independent if the σ-algebras generated by the random variables are independent. All of the measure-theoretic results and exercises about independent real-valued random variables from Chapter 7 remain valid for independent Polish space valued random variables.

The following result is sometimes useful for proving independence of Polish space valued random variables. Although the result seems almost obvious, the proof needs some work. We will use it later to prove results about Brownian motion.

PROPOSITION 12.10.1. *Let (S, ρ) be a Polish space. Let $\{X_n\}_{n \geq 1}$ be a sequence of S-valued random variables converging almost surely to a random variable X. Suppose that each X_n is independent of a σ-algebra $\mathcal{G}$. Then so is X.*

PROOF. Take any bounded continuous $f : S \to \mathbb{R}$, and any $A \in \mathcal{G}$. Then

$$\mathbb{E}(f(X_n) 1_A) = \mathbb{E}(f(X_n)) \mathbb{P}(A)$$

for each n. By the almost sure version of the dominated convergence theorem (Exercise 2.6.5), $\mathbb{E}(f(X_n) 1_A) \to \mathbb{E}(f(X) 1_A)$ and $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ as $n \to \infty$. Thus, $\mathbb{E}(f(X) 1_A) = \mathbb{E}(f(X)) \mathbb{P}(A)$.

Now, for any bounded continuous $g, h : \mathbb{R} \to \mathbb{R}$, the composition $g \circ f$ is a bounded continuous function from S into $\mathbb{R}$, and the random variable $h(1_A)$ is just a linear transformation of 1_A. Thus, by the above deduction, $\mathbb{E}(g(f(X)) h(1_A)) = \mathbb{E}(g(f(X))) \mathbb{E}(h(1_A))$. So by Corollary 12.6.4, the random variables $f(X)$ and 1_A are independent.

Now take any closed set $F \subseteq S$, and let $f(x) := \max\{\rho(x, F), 1\}$, where $\rho(x, F)$ is the distance from x to F as defined in (12.2.1). We have already seen in the proof of the Portmanteau lemma that $x \mapsto \rho(x, F)$ is continuous. Thus, f is a continuous map. Therefore $f(X)$ and 1_A are independent. In particular, the events $\{f(X) = 0\}$ and A are independent. But $f(X) = 0$ if and only if $X \in F$. Thus, the events $\{X \in F\}$ and A are

independent. Since this holds for every closed F, the random variable X is independent of the event A (for example, by Exercise 7.1.9). And since this holds for all $A \in \mathcal{G}$, X and $\mathcal{G}$ are independent. $\square$

CHAPTER 13

Brownian motion

This chapter introduces an important probabilistic object known as Brownian motion. We will construct Brownian motion, see how it arises as a scaling limit of random walks (Donsker's theorem), and prove a number of results about it.

13.1. The spaces $C[0,1]$ and $C[0,\infty)$

Recall the spaces $C[0,1]$ and $C[0,\infty)$ defined in Exercises 12.1.3 and 12.1.4. In this section we will study some probabilistic aspects of these Polish spaces.

DEFINITION 13.1.1. 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 13.1.2. *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 $C[0,1]$. Let $\mathcal{B}$ be the Borel σ-algebra of $C[0,1]$ and $\mathcal{F}$ be the σ-algebra generated by the finite dimensional projection maps. Note that by the definition of $\mathcal{F}$, each projection map $\pi_{t_1,\ldots,t_n}$ is measurable from $(C[0,1],\mathcal{F})$ into $\mathbb{R}^n$. 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. 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 $C[0,1]$, and therefore measurable from $\mathbb{R}^n$ into $(C[0,1],\mathcal{B})$. 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 measurable from $(C[0,1],\mathcal{F})$ into $(C[0,1],\mathcal{B})$. Now let $\{t_1,t_2,\ldots\}$ be a dense subset of $[0,1]$ chosen in such a way that

$$\lim_{n\to\infty} \max_{0\le t\le 1} \min_{1\le i\le n} |t - t_i| = 0. \tag{13.1.1}$$

(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 (13.1.1) of the sequence $\{t_n\}_{n=1}^{\infty}$, it is not hard to show that $f_n \to f$ in the topology of $C[0,1]$. Thus, the map $\xi_{t_1,\ldots,t_n}$ converges pointwise

to the identity map as $n \to \infty$. So by Proposition 2.1.14, it follows that the identity map from $(C[0,1], \mathcal{F})$ into $(C[0,1], \mathcal{B})$ is measurable. This proves that $\mathcal{F} \supseteq \mathcal{B}$. 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 (13.1.1) with maximum over $t \in [0,K]$, and then impose that the condition should hold for every K. The rest of the argument is left as an exercise for the reader. □

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 13.1.3. *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 13.1.2, $\mathcal{A}$ 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 13.1.4. *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 13.1.3, there can be only one such limit point. The result now follows by Exercise 12.3.15. □

13.2. 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 13.2.1. *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 12.3.2, 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. □

13.3. 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 $1 \leq i \leq n$, 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 13.3.1 (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 $0 < t_1 < \cdots < t_m \leq 1$, the random vector $(B(t_1), \ldots, B(t_m))$ has a multivariate normal distribution with mean vector zero and covariance structure given by $\mathrm{Cov}(B(t_i), B(t_j)) = \min\{t_i, t_j\}$.*

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 13.3.2. *As $n \to \infty$, the finite-dimensional distributions of B_n converge weakly to the limits described in Theorem 13.3.1.*

PROOF. Since $B_n(0) = 0$ for all n, $B_n(0) \to 0$ in distribution. Take any m and any $0 < t_1 < t_2 < \cdots < t_m \leq 1$. Let $k_i := \lfloor nt_i \rfloor$, 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 $0 \leq i \leq m-1$,

$$W_{n,i+1} - W_{n,i} \xrightarrow{d} N(0, t_{i+1} - t_i).$$

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 12.7.2, 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 the random variables $Z_1, \ldots, Z_m$ are independent and $Z_i \sim N(0, t_i - t_{i-1})$ for each i. Now notice that for each n and i,

$$\begin{aligned} |W_{n,i} - B_n(t_i)| &= \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{k_i} X_j - \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{k_i} X_j + (nt_i - k_i) X_{k_i+1} \right) \right| \\ &\leq \frac{|X_{k_i+1}|}{\sqrt{n}}, \end{aligned}$$

where the right side is interpreted as 0 if $k_i = n$. Thus, for any $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}(|W_{n,i} - B_n(t_i)| > \epsilon) &\leq \mathbb{P}(|X_{k_i+1}| > \epsilon\sqrt{n}) \\ &= \mathbb{P}(|X_1| > \epsilon\sqrt{n}), \end{aligned}$$

which tends to zero as $n \to \infty$. 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 13.3.3. *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 13.3.4. *Let S_n be as in Lemma 13.3.3. Then for any $t > 3$, there is some n_0 such that for all $n \geq n_0$,*

$$\mathbb{P}(\max_{1\leq i\leq n} |S_i| \geq t\sqrt{n}) \leq 6e^{-t^2/8}.$$

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 6.3.5, we have the tail bound $\mathbb{P}(|Z| \geq t/2) \leq 2e^{-t^2/8}$. □

We are now ready to prove tightness.

LEMMA 13.3.5. *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 $\lceil n\delta \rceil + 2 \leq k_r - k_{r-1} \leq 2(\lceil n\delta \rceil + 2)$. 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 13.3.4, 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,

$$\begin{aligned}
\mathbb{P}(\omega_{B_n}(\delta) > \eta) &\leq \mathbb{P}(E) \leq \mathbb{P}(E') \\
&\leq C_1 r e^{-C_2\eta^2/\delta} \leq \frac{C_1 e^{-C_2\eta^2/\delta}}{\delta},
\end{aligned}$$

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 13.2.1 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 13.3.1. Note that by Lemma 13.3.5 and Prokhorov's theorem, any subsequence of $\{B_n\}_{n=1}^{\infty}$ has a weakly convergent subsequence. By Proposition 13.1.3

and Lemma 13.3.2, any two such weak limits must be the same. This suffices to prove that the whole sequences converges. □

EXERCISE 13.3.6. 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 $S_n := \sum_{i=1}^n X_i$, with $S_0 = 0$. Prove that the random variables

$$\frac{1}{\sqrt{n}} \max_{0 \le i \le n} S_i$$

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. In fact, the first sequence converges in law to $\max_{0 \le t \le 1} B(t)$, and the second sequence converges in law to $\int_0^1 1_{\{B(t) \ge 0\}} dt$. (Hint: Use Donsker's theorem and Exercises 12.2.4 and 12.2.6. The second one is technically more challenging.)

EXERCISE 13.3.7. Prove a version of Donsker's theorem for sums of stationary m-dependent sequences. (Hint: The key challenge is to generalize Lemma 13.3.3. The rest goes through as in the i.i.d. case, applying the CLT for stationary m-dependent sequences that we derived earlier.)

EXERCISE 13.3.8. Let $X_1, X_2, \ldots$ be i.i.d. standard Cauchy random variables. Let $S_n := \sum_{i=1}^n X_i$, and let $Z_n(t) := S_j/n$ when $t = j/n$. Linearly interpolate to define $Z_n(t)$ for all t. Show that the finite dimensional distributions of Z_n converge, but the random functions Z_n do not converge in law on $C[0, \infty)$. (That is, you need to show that $\{Z_n\}_{n \ge 1}$ is not a tight family in $C[0, \infty)$.)

13.4. 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 13.4.2 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 $k \le t < k + 1$ for some integer $k \ge 0$, define

$$B(t) := \sum_{j=1}^{k-1} B^j(1) + B^k(t - k).$$

The random function B is called standard Brownian motion. We often write B_t instead of $B(t)$.

EXERCISE 13.4.1. Check that B defined above is indeed a $C[0, \infty)$-valued random variable, and that for any n and any $0 \le t_1 \le \cdots \le t_n < \infty$, the vector $(B(t_1), \ldots, B(t_n))$

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 13.4.2. 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.

EXERCISE 13.4.3. Let B be standard Brownian motion. Prove that

(1) $B(t) \sim N(0, t)$ for any t.
(2) $B(t) - B(s) \sim N(0, t - s)$ for any $0 \leq s \leq t$.
(3) For any n and $0 \leq t_1 \leq \cdots \leq t_n$, the random variables $B(t_1), B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$ are independent. (This is known as the 'independent increments' property.)

Conversely, if B is a random continuous function with the above properties, prove that its law is the Wiener measure.

EXERCISE 13.4.4 (Invariance under sign change). If B is standard Brownian motion, prove that $-B$ is again standard Brownian motion.

EXERCISE 13.4.5 (Invariance under time translation). Let B be standard Brownian motion. Take any $s \geq 0$. Define $W(t) := B(s + t) - B(s)$ for $t \geq 0$. Prove that W is standard Brownian motion.

EXERCISE 13.4.6 (Invariance under time scaling). Let B be standard Brownian motion. Take any $c > 0$. Define $W(t) := c^{-1/2} B(ct)$. Prove that W is standard Brownian motion.

EXERCISE 13.4.7 (Lévy's construction of Brownian motion). Construct a sequence of random elements $X_1, X_2, \ldots$ of $C[0, 1]$ as follows. Let $X_1(0) = 0$, $X_1(1) \sim N(0, 1)$, and define $X_1(t)$ for all other t by linear interpolation. Then, define $X_2(1/2) = X_1(1/2) + Z$, where $Z \sim N(0, 1/4)$. Also let $X_2(1) = X_1(1)$ and $X_2(0) = 0$. Having defined $X_2(t)$ at $t = 0, 1/2, 1$, define it everywhere else by linear interpolation. In general, having defined X_k, define X_{k+1} as follows. If $t = j2^{-k}$ for some even j, let $X_{k+1}(t) = X_k(t)$. If $t = j2^{-k}$ for some odd j, let

$$X_{k+1}(t) = X_k(t) + \text{ independent } N(0, 2^{-k-1}).$$

For all other t, define $X_{k+1}(t)$ by linear interpolation. Then:

(1) Prove that the finite dimensional distributions of the sequence $\{X_n\}_{n \geq 1}$ converge to those of Brownian motion.
(2) Prove that
$$\sum_{n=1}^{\infty} \mathbb{E}\|X_{n+1} - X_n\|_{[0,1]} < \infty.$$
(3) Using the above, prove that $\{X_n\}_{n \geq 1}$ is almost surely a Cauchy sequence in $C[0, 1]$, and therefore has a limit. Prove that the limit is Brownian motion.

EXERCISE 13.4.8. Let $(B(t))_{0 \leq t \leq 1}$ be standard Brownian motion on the time interval $[0, 1]$. Let $W(t) = B(t) - tB(1)$. This process is known as the 'Brownian bridge'. What are

the finite dimensional distributions of the Brownian bridge? Prove that if W is a Brownian bridge, and Z is an independent $N(0,1)$ random variable, and $U(t) = W(t) + tZ$, then U is standard Brownian motion. Consequently, deduce that if B and W are as above, then the random function W and the random variable $B(1)$ are independent.

13.5. An application of Donsker's theorem

Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables with mean 0 and variance 1. Let $S_n = \sum_{i=1}^{n} X_i$, with $S_0 = 0$. What is the limiting distribution of $\max_{0\le i\le n} S_i/\sqrt{n}$? By Exercise 13.3.6, we know that there is a limiting distribution and it does not depend on the law of the X_i's. It is therefore enough to figure out the limiting distribution in one case.

For convenience, we choose the X_i's to be 1 or -1 with equal probability, so that S_n is a simple symmetric random walk on $\mathbb{Z}$. Let M_n be the maximum of $S_0, \ldots, S_n$. Since $S_0 = 0$, we have $M_n \ge 0$. Take any integer $a > 0$. Let $\mathcal{P}$ be the set of all possible values of the vector $(S_0, \ldots, S_n)$. Let $\mathcal{A}$ be the subset consisting of all $(s_0, \ldots, s_n) \in \mathcal{P}$ such that $\max_{0\le i\le n} s_i \ge a$ and $s_n < a$. Let $\mathcal{B}$ be the subset consisting of all $(s_0, \ldots, s_n) \in \mathcal{P}$ such that $\max_{0\le i\le n} s_i \ge a$ and $s_n > a$.

The first key observation is that $\mathcal{A}$ and $\mathcal{B}$ have the same size. To prove this, we exhibit a bijection ϕ from $\mathcal{A}$ onto $\mathcal{B}$. Take any $(s_0, \ldots, s_n) \in \mathcal{A}$. Define $(s_0', \ldots, s_n') = \phi(s_0, \ldots, s_n)$ as follows. Let $m \le n$ be the smallest index such that $s_m = a$. This exists by the definition of $\mathcal{A}$. Define $(s_0', \ldots, s_n')$ by 'reflecting $(s_0, \ldots, s_n)$ across level a beyond the time point m', as

$$s_i' := \begin{cases} s_i & \text{if } i \le m, \\ a - (s_i - a) & \text{if } i > m. \end{cases}$$

Since $s_n' = 2a - s_n$ and $s_n < a$, we have $s_n' > a$. Thus, ϕ is indeed a map from $\mathcal{A}$ into $\mathcal{B}$. Now, it is clear from the definition that m is the smallest number such that $s_m' = a$. Moreover, $s_i = s_i'$ for $i \le m$, and $s_i = a - (s_i' - a)$ for $i > m$. Thus, if we define ϕ as exactly as above on $\mathcal{B}$, then $\phi(s_0', \ldots, s_n') = (s_0, \ldots, s_n)$. This proves that ϕ is a bijection between $\mathcal{A}$ and $\mathcal{B}$, and hence $|\mathcal{A}| = |\mathcal{B}|$.

Now let $\mathcal{C}$ be the set of all $(s_0, \ldots, s_n) \in \mathcal{P}$ such that $\max_{0\le i\le n} s_i \ge a$ and $s_n = a$. Then

$$\mathbb{P}(M_n \ge a) = \frac{|\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}|}{|\mathcal{P}|}.$$

But the sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are disjoint, and $|\mathcal{A}| = |\mathcal{B}|$. Therefore

$$\mathbb{P}(M_n \ge a) = \frac{2|\mathcal{B}| + |\mathcal{C}|}{|\mathcal{P}|}.$$

Now we make the second crucial observation in the proof, which is that $\mathcal{B}$ is just the set of all $(s_0, \ldots, s_n)$ such that $s_n > a$, and $\mathcal{C}$ is just the set of all $(s_0, \ldots, s_n)$ such that $s_n = a$. Thus,

$$\mathbb{P}(M_n \ge a) = 2\mathbb{P}(S_n > a) + \mathbb{P}(S_n = a).$$

Using this identity and the central limit theorem (or just Stirling's approximation) it is easy to show that for any $x \ge 0$,

$$\lim_{n\to\infty} \mathbb{P}(M_n \ge x\sqrt{n}) = 2\mathbb{P}(Z \ge x),$$

where $Z \sim N(0,1)$. Since $2\mathbb{P}(Z \geq x) = \mathbb{P}(|Z| \geq x)$, we arrive at the following result.

PROPOSITION 13.5.1. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean 0 and variance 1. Then as $n \to \infty$,*

$$\frac{1}{\sqrt{n}} \max_{0 \leq i \leq n} \sum_{i=1}^{n} X_i \xrightarrow{d} |Z|,$$

where $Z \sim N(0,1)$. In particular, $\max_{0 \leq t \leq 1} B(t)$ has the same law as $|Z|$.

Incidentally, the argument used above is known as the 'reflection principle'. It has important extensions to other settings, especially in higher dimensions. An important corollary of the above proposition is the following.

COROLLARY 13.5.2. *Let B be standard Brownian motion. Then for all $x > 0$,*

$$\mathbb{P}(\max_{0 \leq t \leq 1} |B(t)| \geq x) \leq 4e^{-x^2/2}.$$

PROOF. Since $-B$ is again a standard Brownian motion (by Exercise 13.4.4), Proposition 13.5.1 gives

$$\begin{aligned}\mathbb{P}(\max_{0 \leq t \leq 1} |B(t)| \geq x) &\leq \mathbb{P}(\max_{0 \leq t \leq 1} B(t) \geq x) + \mathbb{P}(\max_{0 \leq t \leq 1} (-B(t)) \geq x) \\ &= 2\mathbb{P}(\max_{0 \leq t \leq 1} B(t) \geq x) = 4\mathbb{P}(Z \geq x).\end{aligned}$$

The proof is now completed by the normal tail bound from Exercise 6.3.5. □

EXERCISE 13.5.3. Let B be standard Brownian motion. Given $t > 0$, figure out the law of $\max_{0 \leq s \leq t} B(s)$.

EXERCISE 13.5.4. Let B be standard Brownian motion. Given $a > 0$, let $T := \inf\{t : B(t) \geq a\}$. Prove that T is a random variable and that it is finite almost surely. Then compute the probability density function of T. (Hint: Use the previous problem.)

13.6. Law of large numbers for Brownian motion

The following result is known as the law of large numbers for Brownian motion.

PROPOSITION 13.6.1. *Let B be standard Brownian motion. As $t \to \infty$, $B(t)/t \to 0$ almost surely.*

PROOF. Since $B(1), B(2) - B(1), B(3) - B(2), \ldots$ are i.i.d. standard normal random variables, $B(n)/n \to 0$ almost surely as $n \to \infty$ by the strong law of large numbers. Now if $n \leq t \leq n+1$, then

$$\frac{|B(t)|}{t} \leq \frac{|B(t)|}{n} \leq \frac{|B(n)|}{n} + \frac{1}{n} \max_{n \leq s \leq n+1} |B(s) - B(n)|.$$

Therefore it suffices to show that $M_n/n \to 0$ almost surely as $n \to \infty$, where $M_n := \max_{n \leq s \leq n+1} |B(s) - B(n)|$. Take any n. Let $W(t) := B(n+t) - B(n)$ for $t \geq 0$. By Exercise 13.4.5, W is again standard Brownian motion. Therefore by Corollary 13.5.2,

$$\mathbb{P}(M_n \geq x) = \mathbb{P}(\max_{0 \leq t \leq 1} |W(t)| \geq x) \leq 4e^{-x^2/2}.$$

Thus,

$$\sum_{n=1}^{\infty} \mathbb{P}(M_n \geq \sqrt{n}) \leq \sum_{n=1}^{\infty} 4e^{-n/2} < \infty.$$

By the first Borel–Cantelli lemma, this proves that $M_n/n \to 0$ almost surely. □

EXERCISE 13.6.2 (Time inversion). Let B be standard Brownian motion. Define $W(t) := tB(1/t)$ for $t > 0$, and $W(0) = 0$. Prove that W is again standard Brownian motion. (Hint: You will have to use the law of large numbers for continuity at zero.)

EXERCISE 13.6.3. Let B be standard Brownian motion. Given $a > 0$, let $T := \sup\{t : B(t) \geq at\}$. Prove that T is a random variable and that it is finite almost surely. Then compute the probability density function of T. (Hint: Use Exercise 13.5.4 and time inversion.)

EXERCISE 13.6.4. Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with mean zero and variance one. Let $S_n = X_1 + \cdots + X_n$. For each $\epsilon > 0$, let $N(\epsilon)$ be the smallest number such that $S_n/n < \epsilon$ for all $n > N(\epsilon)$. Compute the limiting distribution of $\epsilon^2 N(\epsilon)$ as $\epsilon \to 0$. (Hint: Use the previous exercise.)

EXERCISE 13.6.5. Prove that with probability one,

$$\limsup_{t\to\infty} \frac{B(t)}{\sqrt{t}} = \infty, \quad \liminf_{t\to\infty} \frac{B(t)}{\sqrt{t}} = -\infty.$$

13.7. Nowhere differentiability of Brownian motion

We know that Brownian motion is continuous, but is it differentiable? The following result of Paley, Wiener and Zygmund shows that with probability one, the Brownian path is nowhere differentiable. Note that this is much stronger than proving that it is not differentiable at a given point with probability one, since there are uncountably many points. A minor technical point is that it is not clear that the set of nowhere differentiable continuous functions is a Borel subset of $C[0, \infty)$. However, since we only wish to show that it has probability zero, it suffices to work with the completion of the Borel σ-algebra (see Proposition **??**) and prove that the set of nowhere differentiable continuous functions is a subset of a Borel set of probability zero.

THEOREM 13.7.1. *With probability one, standard Brownian motion is nowhere differentiable.*

PROOF. If we prove that B is nowhere differentiable in $[0, 1]$ with probability one (where differentiability at the endpoints is one-sided), then it follows that with probability one, B is not differentiable in $[k, k + 1]$ for any positive integer k, and therefore nowhere differentiable on $[0, \infty)$. So let us prove that B is nowhere differentiable on $[0, 1]$.

Suppose that in a particular realization, B is differentiable at a point $t \in [0, 1]$. Then

$$\limsup_{h\to 0} \frac{|B(t + h) - B(t)|}{h} < \infty.$$

By the boundedness of B in $[0,1]$, this implies that

$$\sup_{s\in[0,1]\setminus\{t\}} \frac{|B(s)-B(t)|}{|s-t|} < \infty.$$

Thus, there is some positive integer M (depending on the realization of B) such that for all $s \in [0,1]$,

$$|B(s)-B(t)| \le M|s-t|.$$

Fix M and t as above. Take any $n \ge 4$. Then for any $0 \le k \le n-1$,

$$\begin{aligned}|B(k/n)-B((k+1)/n)| &\le |B(k/n)-B(t)| + |B(t)-B((k+1)/n)| \\ &\le M|(k/n)-t| + M|((k+1)/n)-t|.\end{aligned}$$

Choose $0 \le j \le n-3$ such that $|j/n - t| \le 3/n$. (This can be done, for example, by choosing j such that $j/n \le t \le (j+1)/n$ if $t \le 1-3/n$, and $j = n-3$ otherwise.) Then by the above inequality, the quantities $|B(j/n)-B((j+1)/n)|$, $|B((j+1)/n)-B((j+2)/n)|$ and $|B((j+2)/n)-B((j+3)/n)|$ are all bounded above by $11M/n$.

For any positive integers $M \ge 1$, $n \ge 4$, and $0 \le j \le n-3$, let $E_{M,n,j}$ be the event described in the previous sentence. Thus, we have shown that if B has a point of differentiability in $[0,1]$, then there is some positive integer M such that for all $n \ge 4$, there is some $0 \le j \le n-3$ for which the event $E_{M,n,j}$ happens. In other words,

$$\{B \text{ has a point of differentiability in } [0,1]\} \subseteq \bigcup_{M=1}^{\infty} \bigcap_{n=4}^{\infty} \bigcup_{j=0}^{n-3} E_{M,n,j}. \tag{13.7.1}$$

The set on the right is clearly measurable. So it suffices to prove that it has probability zero. To see this, note that

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{M=1}^{\infty} \bigcap_{n=4}^{\infty} \bigcup_{j=0}^{n-3} E_{M,n,j}\Big) &\le \sum_{M=1}^{\infty} \mathbb{P}\Big(\bigcap_{n=4}^{\infty} \bigcup_{j=0}^{n-3} E_{M,n,j}\Big) \\ &\le \sum_{M=1}^{\infty} \inf_{n\ge 4} \mathbb{P}\Big(\bigcup_{j=0}^{n-3} E_{M,n,j}\Big) \\ &\le \sum_{M=1}^{\infty} \inf_{n\ge 4} \sum_{j=0}^{n-3} \mathbb{P}(E_{M,n,j}).\end{aligned}$$

By independence of increments and the fact that $B((i+1)/n) - B(i/n)$ is a $N(0,1/n)$ random variable for any i, it is easy to see that for any given M, n, and j as above,

$$\mathbb{P}(E_{M,n,j}) \le Cn^{-3/2},$$

where C depends only on M (and not on n or j). Thus, for any given M,

$$\inf_{n\ge 4} \sum_{j=0}^{n-3} \mathbb{P}(E_{M,n,j}) = 0.$$

This proves that the event on the right in (13.7.1) indeed has probability zero. □

EXERCISE 13.7.2. For any function $f : [0,1] \to \mathbb{R}$, the total variation of f is defined as

$$V(f) := \sup_{n \geq 1} \sup_{0 = x_0 \leq x_1 \leq \cdots \leq x_n = 1} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|.$$

Prove that Brownian motion on the time interval $[0,1]$ almost surely has infinite total variation.

13.8. The Brownian filtration

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A continuous-time filtration $\{\mathcal{F}_t\}_{t \geq 0}$ is a collection of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$. For example, if B is standard Brownian motion and we let $\mathcal{F}_t$ be the σ-algebra generated by the collection of random variables $\{B(s)\}_{0 \leq s \leq t}$, then $\{\mathcal{F}_t\}_{t \geq 0}$ is a filtration. The following exercise points out a subtle but important fact.

EXERCISE 13.8.1. Viewing the restriction of B to $[0,t]$ as a $C[0,t]$-valued random variable, prove that the σ-algebra $\mathcal{F}_t$ is the same as the σ-algebra generated by this random variable. Prove that this holds also for $t = \infty$, that is, the σ-algebras generated by B and the collection $\{B(t)\}_{t \geq 0}$ are the same. (Hint: Note that $B|_{[0,t]}$ is the limit of a sequence of polygonal approximations, each of which is $\mathcal{F}_t$-measurable. Then use Proposition 2.1.14. The converse is easy.)

It turns out that it is much more useful to consider a slight enlargement of $\mathcal{F}_t$, defined as

$$\mathcal{F}_t^+ := \bigcap_{s > t} \mathcal{F}_s.$$

Clearly, $\{\mathcal{F}_t^+\}_{t \geq 0}$ is also a filtration. This is called the right-continuous filtration of Brownian motion, or simply the Brownian filtration. The important property of this filtration is that it is *right-continuous*, meaning that

$$\mathcal{F}_t^+ = \bigcap_{s > t} \mathcal{F}_s^+.$$

To see this, just note that

$$\mathcal{F}_t^+ = \bigcap_{s > t} \mathcal{F}_s = \bigcap_{s > t} \bigcap_{u > s} \mathcal{F}_u = \bigcap_{s > t} \mathcal{F}_u^+.$$

EXAMPLE 13.8.2. The following is an example of an event in $\mathcal{F}_t^+$. Take any t and let

$$\tau := \inf\{s > t : B(s) = B(t)\}.$$

Then the event $\{\tau = t\}$ means that there is some sequence t_n strictly decreasing to t such that $B(t_n) = B(t)$ for each n. So for any $\epsilon > 0$,

$$\{\tau = t\} = \bigcap_{n \geq 1/\epsilon} E_n,$$

where E_n is the event that $B(t+u) = B(t)$ for some $u \in (0, 1/n]$. Note that by the continuity of B,

$$\begin{aligned} E_n &= \bigcup_{k=n}^{\infty} \{B(t+u) = B(t) \text{ for some } u \in [1/k, 1/n]\} \\ &= \bigcup_{k=n}^{\infty} \bigcap_{j=1}^{\infty} \{|B(t+u) - B(t)| < 1/j \text{ for some } u \in [1/k, 1/n] \cap \mathbb{Q}\}. \end{aligned}$$

This shows that $E_n \in \mathcal{F}_{t+1/n}$. Thus, for $n \geq 1/\epsilon$, $E_n \in \mathcal{F}_{t+1/n} \subseteq \mathcal{F}_{t+\epsilon}$. Since this is true for every $\epsilon > 0$, we see that $\{\tau = t\} \in \mathcal{F}_t^+$.

To see that $\mathcal{F}_t^+$ may not be equal to $\mathcal{F}_t$, consider the special case $t = 0$. Suppose that the probability space on which B is defined is simply $C[0,1]$ endowed with its Borel σ-algebra and the Wiener measure, and B is the identity map on this space. Since $B(0) = 0$, $\mathcal{F}_0$ is the trivial σ-algebra consisting of only Ω and $\emptyset$. But clearly, the event $\{\tau = 0\}$, as a subset of this space, is neither Ω nor $\emptyset$.

13.9. Markov property of Brownian motion

The following result is known as the Markov property of Brownian motion. It is the precursor of a more powerful property, known as the strong Markov property, that we will discuss in the next section.

THEOREM 13.9.1. *Let B be standard Brownian motion. Take any $s \geq 0$. Define $W(t) := B(s+t) - B(s)$ for $t \geq 0$. Then W is a standard Brownian motion and is independent of the σ-algebra $\mathcal{F}_s^+$.*

PROOF. We have already seen that W is a standard Brownian motion in Exercise 13.4.5. So it only remains to prove the independence. We will first prove that W and $\mathcal{F}_s$ are independent. Take any $m, n \geq 1$, any $0 \leq t_1 < \cdots < t_n < \infty$, and any $0 \leq s_1 < \cdots < s_m \leq s$. By the independence of increments of Brownian motion, it follows that the random vectors $(B(s_1), \ldots, B(s_m))$ and $(W(t_1), \ldots, W(t_n))$ are independent. Therefore by Exercise 7.1.11, the σ-algebras generated by $\{B(t)\}_{0 \leq t \leq s}$ and $\{W(t)\}_{t \geq 0}$ are independent. But by Exercise 13.8.1, these are the σ-algebras $\mathcal{F}_s$ and $\sigma(W)$. This proves the independence of W and $\mathcal{F}_s$.

Next, let us prove the independence of W and $\mathcal{F}_s^+$. Take a sequence $\{s_n\}_{n \geq 1}$ strictly decreasing to s. For each n, define $W_n(t) = B(s_n + t) - B(s_n)$ for $t \geq 0$. Then by the previous paragraph, W_n and $\mathcal{F}_{s_n}$ are independent. Since $\mathcal{F}_s^+ \subseteq \mathcal{F}_{s_n}$, we get that W_n and $\mathcal{F}_s^+$ are independent for each n. But $W_n \to W$ in $C[0, \infty)$. Therefore by Proposition 12.10.1, W is independent of $\mathcal{F}_s^+$. □

An interesting consequence of the Markov property is Blumenthal's zero-one law, which says that any event in $\mathcal{F}_0^+$ must have probability 0 or 1 with respect to Wiener measure. Indeed, take any $A \in \mathcal{F}_0^+$. Then due to the Markov property of B, the event A is independent of $\sigma(B)$. On the other hand, since $\mathcal{F}_0^+ \subseteq \sigma(B)$, A is an element of $\sigma(B)$. Therefore A is independent of itself, and hence $\mathbb{P}(A)$ must be 0 or 1.

Blumenthal's zero-one law is a basic fact about Brownian motion, with a number of important consequences. Let us work out a few of these. First, let

$$\tau^+ := \inf\{t > 0 : B(t) > 0\},$$
$$\tau^- := \inf\{t > 0 : B(t) < 0\},$$
$$\tau := \inf\{t > 0 : B(t) = 0\}.$$

It is easy to see (just as we did in the previous section) that the event $\{\tau^+ = 0\}$ is in $\mathcal{F}_0^+$. Now for any $t > 0$,

$$\mathbb{P}(\tau^+ \le t) \ge \mathbb{P}(B(t) > 0) = \frac{1}{2}.$$

Since $\{\tau^+ = 0\}$ is the decreasing limit of $\{\tau^+ \le 1/n\}$ as $n \to \infty$, this shows that $\mathbb{P}(\tau^+ = 0) \ge 1/2$. Thus, by Blumenthal's zero-one law, $\mathbb{P}(\tau^+ = 0) = 1$. By the same argument, $\mathbb{P}(\tau^- = 0) = 1$. But if $\tau^+ = 0$ and $\tau^- = 0$, then B attains both positive and negative values in arbitrarily small neighborhoods of 0. By the continuity of B, this implies that $\tau = 0$. Thus, $\mathbb{P}(\tau = 0) = 1$.

Combined with the Markov property, the above result implies the more general statement that for any given t, $\inf\{s > t : B(s) = B(t)\} = t$ almost surely. Note that this does not imply that there will not exist any exceptional t for which this fails in a given realization of B. To see this, note that $B(1) \neq 0$ with probability 1. Thus, if we let $T := \sup\{t \in [0, 1] : B(t) = 0\}$, then $T < 1$ almost surely. Now note that $\inf\{t > T : B(t) = B(T)\} > 1 > T$.

Similarly, $\inf\{s > t : B(s) > B(t)\} = t$ and $\inf\{s > t : B(s) < B(t)\} = t$ almost surely. As a consequence, we get the following result.

PROPOSITION 13.9.2. *With probability one, Brownian motion has no interval of increase or decrease (that is, no nontrivial interval where Brownian motion is increasing or decreasing monotonically).*

PROOF. For a given interval $[a, b]$, where $a < b$, we know from the discussion preceding the proposition that with probability one, B cannot be monotonically increasing or decreasing in $[a, b]$. Therefore with probability one, this cannot happen in any interval with rational endpoints. But if there is a nontrivial interval where B is monotonically increasing or decreasing, then there is a subinterval with rational endpoints where this happens. □

EXERCISE 13.9.3. Prove that with probability one, every local maximum of Brownian motion is a strict local maximum. (Hint: First prove that the values of the maxima in two given disjoint intervals are different almost surely. Then take intersection over intervals with rational endpoints.)

EXERCISE 13.9.4. Prove that with probability one, the set of local maxima of Brownian motion is countable and dense. (Hint: Use Proposition 13.9.2 and Exercise 13.9.3.)

EXERCISE 13.9.5. For any fixed $s > 0$, show that the random function $V(t) := B(s - t) - B(s)$, defined for $0 \le t \le s$, is a standard Brownian motion on the time interval $[0, s]$. (This is called 'time reversal'.)

EXERCISE 13.9.6. Prove that the for any $s \in [0,1]$,

$$\mathbb{P}\big(\max_{0\le t\le s} B(t) = \max_{s\le t\le 1} B(t)\big) = 0.$$

Using this, prove that with probability one, Brownian motion has a unique point of global maximum in the time interval $[0,1]$.

EXERCISE 13.9.7. Let T be the unique point at which $B(t)$ attains its global maximum in the time interval $[0,1]$. Compute the probability distribution function and the probability density function of T. (This is known as the 'arcsine distribution'. Hint: Try to write $\mathbb{P}(T \le s)$ as a probability involving the maxima of two independent Brownian motions. You will have to apply time reversal and the Markov property.)

13.10. Law of the iterated logarithm for Brownian motion

The following theorem is known as the law of the iterated logarithm (LIL) for Brownian motion. It gives a precise measure of the maximum fluctuations of $B(t)$ as $t \to \infty$ or $t \to 0$. The proof makes several crucial uses of the Markov property.

THEOREM 13.10.1. *Let B be standard Brownian motion. Then with probability one,*

$$\limsup_{t\to\infty} \frac{B(t)}{\sqrt{2t\log\log t}} = 1, \quad \liminf_{t\to\infty} \frac{B(t)}{\sqrt{2t\log\log t}} = -1.$$

The same statements hold if we replace $t \to \infty$ by $t \to 0$, and $\log\log t$ by $\log\log(1/t)$.

PROOF. Using time inversion, it is not difficult to check that it suffices to prove the result only for $t \to \infty$. Also, the lim inf result follows by changing B to $-B$, so it suffices to prove only for lim sup. Let $h(t) := \sqrt{2t\log\log t}$. Take any $\delta > 0$. Let $t_n := (1+\delta)^n$. Let A_n be the event

$$\sup_{s\in[t_n,t_{n+1}]} \frac{B(s)}{h(s)} > 1+\delta.$$

Let $M(t) := \max_{0\le s\le t} B(s)$. Note that $h(t)$ is an increasing function of t if t is sufficiently large. So if n is sufficiently large and A_n happens, then there is some $s \in [t_n, t_{n+1}]$ such that

$$1+\delta < \frac{B(s)}{h(s)} \le \frac{B(s)}{h(t_n)},$$

which implies that $M(t_{n+1}) \ge (1+\delta)h(t_n)$. But we know the exact distribution of $M(t_{n+1})$. Therefore, by Exercise 6.3.5,

$$\begin{aligned}\mathbb{P}(A_n) &\le \mathbb{P}(M(t_{n+1}) \ge (1+\delta)h(t_n)) \\ &\le 2e^{-(1+\delta)^2 h(t_n)^2/2t_{n+1}} \\ &= e^{-(1+\delta)\log\log t_n} = (n\log(1+\delta))^{-(1+\delta)}.\end{aligned}$$

Since this upper bound is summable in n, the first Borel–Cantelli lemma tells us that with probability one, A_n^c happens for all sufficiently large n. But that implies

$$\limsup_{t\to\infty} \frac{B(t)}{h(t)} \le 1+\delta \text{ a.s.}$$

Since this holds for all $\delta > 0$, we can take a countable sequence of δ's tending to zero and conclude that

$$\limsup_{t\to\infty} \frac{B(t)}{h(t)} \le 1 \text{ a.s.} \tag{13.10.1}$$

Next, take any $\delta \in (0,1)$ and let $s_n := \delta^{-n}$. Let $X_n := B(s_n) - B(s_{n-1})$ and let E_n be the event $X_n \ge (1-\delta)h(s_n)$. Let

$$a_n := \frac{(1-\delta)h(s_n)}{\sqrt{s_n - s_{n-1}}} = \sqrt{2(1-\delta)\log\log(\delta^{-n})}. \tag{13.10.2}$$

Then by Exercise 6.3.6,

$$\mathbb{P}(E_n) \ge \left(\frac{1}{a_n} - \frac{1}{a_n^3}\right)\frac{e^{-a_n^2/2}}{\sqrt{2\pi}}.$$

From the expression (13.10.2) for a_n, it follows that $\sum_{n=1}^{\infty}\mathbb{P}(B_n) = \infty$. Since the random variables $X_1, X_2, \ldots$ are independent, so are the events $E_1, E_2, \ldots$. Therefore by the second Borel–Cantelli lemma, E_n occurs infinitely often with probability one. In other words, with probability one,

$$\frac{B(s_n) - B(s_{n-1})}{h(s_n)} \ge 1 - \delta \quad \text{i.o.} \tag{13.10.3}$$

But by (13.10.1) and the fact that $-B$ is also a standard Brownian motion, we have that

$$\begin{aligned}\limsup_{n\to\infty} \frac{-B(s_{n-1})}{h(s_n)} &= \limsup_{n\to\infty} \frac{-B(s_{n-1})}{h(s_{n-1})}\frac{h(s_{n-1})}{h(s_n)} \\ &= \sqrt{\delta}\limsup_{n\to\infty}\frac{-B(s_{n-1})}{h(s_{n-1})} \le \sqrt{\delta}.\end{aligned}$$

Thus, with probability one, $-B(s_{n-1})/h(s_n) \le 2\sqrt{\delta}$ for all sufficiently large n. Plugging this into (13.10.3), we get that with probability one,

$$\frac{B(s_n)}{h(s_n)} \ge 1 - \delta - 2\sqrt{\delta} \quad \text{i.o.} \tag{13.10.4}$$

Since $\delta > 0$ is arbitrary, this shows that $\limsup_{t\to\infty} B(t)/h(t) \ge 1$ a.s. □

EXERCISE 13.10.2. Prove that as $t \to \infty$, $B(t)/\sqrt{2t\log\log t} \to 0$ in probability but does not tend to zero almost surely.

By a refinement of the above proof, one can prove the following stronger version of the law of the iterated logarithm for Brownian motion. We will use it later to prove the law of the iterated logarithm for sums of i.i.d. random variables.

THEOREM 13.10.3. *Let B be standard Brownian motion. Then with probability one, for all sequences $\{t_n\}_{n\ge1}$ such that $t_n \to \infty$ and $t_{n+1}/t_n \to 1$,*

$$\limsup_{n\to\infty} \frac{B(t_n)}{\sqrt{2t_n\log\log t_n}} = 1, \quad \liminf_{t\to\infty} \frac{B(t_n)}{\sqrt{2t_n\log\log t_n}} = -1.$$

The same statements hold if we replace $t_n \to \infty$ by $t_n \to 0$, and $\log\log t_n$ by $\log\log(1/t_n)$.

(Note that the crucial point in this theorem is that t_n is allowed to be random.)

PROOF. As before, it suffices to prove only for lim sup and only for $t_n \to \infty$. Take a realization of B where the conclusion of the LIL is satisfied, and let $\{t_n\}_{n\geq 1}$ be any sequence such that $t_n \to \infty$ and $t_{n+1}/t_n \to 1$. Then immediately from Theorem 13.10.1 we have

$$\limsup_{n\to\infty} \frac{B(t_n)}{\sqrt{2t_n \log\log t_n}} \leq \limsup_{t\to\infty} \frac{B(t)}{\sqrt{2t\log\log t}} = 1.$$

The opposite inequality requires a bit more work. First, let δ, s_n and E_n be as in the second half of the proof of Theorem 13.10.1. Let D_n be the event

$$\min_{s_n\leq t\leq s_{n+1}} (B(t) - B(s_n)) \geq -\sqrt{s_n}.$$

By the Markov property of Brownian motion, it is easy to see that the events E_n and D_n are independent. Also, by Proposition 13.5.1, it is easy to check that $\mathbb{P}(D_n) \geq c(\delta) > 0$ for some constant $c(\delta)$ that depends on δ but not on n. Therefore from our previously obtained lower bound on $\mathbb{P}(E_n)$, we get

$$\sum_{n=1}^{\infty} \mathbb{P}(E_{2n} \cap D_{2n}) = \infty.$$

But again, note that E_n and D_n are determined by the process

$$W_n(t) := B(s_{n-1} + t) - B(s_{n-1}),$$

while E_k and D_k are determined by $\{B(t)\}_{t\leq s_{n-1}}$ for $k \leq n-2$. So by successive applications of the Markov property, we see that the events $\{E_{2n} \cap D_{2n}\}_{n\geq 1}$ are independent. Therefore with probability one, these events occur infinitely often.

Now if $E_n \cap D_n$ occurs for some n, then we have

$$\min_{s_n\leq t\leq s_{n+1}} B(t) \geq B(s_n) - \sqrt{s_n}.$$

Combining with (13.10.4), we get that with probability one,

$$\min_{s_n\leq t\leq s_{n+1}} B(t) \geq (1 - \delta - 2\sqrt{\delta})h(s_n) - \sqrt{s_n} \text{ i.o.}$$

Take a realization of B that satisfies the above property for all δ in a sequence tending to zero. Take any sequence $\{t_n\}_{n\geq 1}$ such that $t_n \to \infty$ and $t_{n+1}/t_n \to 1$. For each k, let $n = n(k)$ be the smallest number such that $s_k \leq t_n < s_{k+1}$. Then by the above property of the particular realization of B, there exist infinitely many n such that

$$B(t_n) \geq (1 - \delta - 2\sqrt{\delta})h(s_k) - \sqrt{s_k}.$$

As $k \to \infty$, we have $n = n(k) \to \infty$, and so $t_n/t_{n-1} \to 1$. This means, in particular, that $t_n/s_k \to 1$ as $k \to \infty$, because otherwise t_{n-1} would be between s_k and t_n infinitely often. Thus,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{B(t_n)}{h(t_n)} &\geq \limsup_{k\to\infty} \frac{B(t_{n(k)})}{h(t_{n(k)})} \\
&\geq \lim_{k\to\infty} \left((1 - \delta - 2\sqrt{\delta}) \frac{h(s_k)}{h(t_{n(k)})} - \frac{\sqrt{s_k}}{h(t_{n(k)})} \right) \\
&= 1 - \delta - 2\sqrt{\delta}.
\end{aligned}$$

Since this holds for all δ in a sequence tending to zero, this completes the proof of the theorem. □

13.11. Stopping times for Brownian motion

Let B be standard Brownian motion. A $[0,\infty]$-valued random variable T, defined on the same probability space as B, is said to be a stopping time for B if for every $t \in [0,\infty)$, the event $\{T \le t\}$ is in $\mathcal{F}_t^+$. Interestingly, the right-continuity of the filtration implies that it does not matter whether we put $T \le t$ or $T < t$ in the definition.

EXERCISE 13.11.1. Show that a random variable T is a stopping time for Brownian motion if and only if for every $t \ge 0$, the event $\{T < t\}$ is in $\mathcal{F}_t^+$.

Any constant is trivially a stopping time. A slightly less trivial example is the following.

EXAMPLE 13.11.2. Let

$$T = \inf\{t \ge 0 : B(t) = a\},$$

where $a \in \mathbb{R}$ is some given number. Then T is a stopping time, because

$$\begin{aligned}\{T \le t\} &= \bigcup_{s\in[0,t]} \{B(s) = a\} \\ &= \bigcap_{n\ge 1} \bigcup_{s\in[0,t]\cap\mathbb{Q}} \{|B(s) - a| \le 1/n\},\end{aligned}$$

where both identities hold because B is continuous. The last event is clearly in $\mathcal{F}_t$, and therefore in $\mathcal{F}_t^+$. By the continuity of B, it also follows that T can be alternately defined as $\inf\{t \ge 0 : B(t) \ge a\}$ if $a \ge 0$ and as $\inf\{t \ge 0 : B(t) \le a\}$ if $a \le 0$.

EXERCISE 13.11.3. If T and S are stopping times for Brownian motion, prove that $\min\{S,T\}$, $\max\{S,T\}$ and $S+T$ are also stopping times.

Associated with any stopping time T is a σ-algebra known as the stopped σ-algebra of T. It is defined as

$$\mathcal{F}_T^+ := \{A \in \sigma(B) : A \cap \{T \le t\} \in \mathcal{F}_t^+ \text{ for all } t \ge 0\}.$$

EXERCISE 13.11.4. Prove that if we defined $\mathcal{F}_T^+$ using $\{T < t\}$ instead of $\{T \le t\}$, the definition would remain the same.

Intuitively, $\mathcal{F}_T^+$ contains the record of all information up to the random time T and infinitesimally beyond. For example, T itself is $\mathcal{F}_T^+$-measurable. To see this, note that for any $s,t \ge 0$,

$$\{T \le s\} \cap \{T \le t\} = \{T \le \min\{s,t\}\} \in \mathcal{F}^+_{\min\{s,t\}} \subseteq \mathcal{F}_t^+.$$

Thus, $\{T \le s\} \in \mathcal{F}_T^+$ for any $s \ge 0$. This proves that T is $\mathcal{F}_T^+$-measurable.

An important and often useful fact is the following.

PROPOSITION 13.11.5. *If S and T are stopping times for Brownian motion, such that $S \le T$ always, then $\mathcal{F}_S^+ \subseteq \mathcal{F}_T^+$.*

PROOF. Take any $t \ge 0$ and any $A \in \mathcal{F}_S^+$. Since $S \le T$, we have

$$A \cap \{T \le t\} = A \cap \{S \le t\} \cap \{T \le t\}.$$

Since $A \in \mathcal{F}_S^+$, we have $A \cap \{S \le t\} \in \mathcal{F}_t^+$. Since T is a stopping time, $\{T \le t\} \in \mathcal{F}_t^+$. Thus, $A \cap \{T \le t\} \in \mathcal{F}_t^+$. Since this holds for every t, we conclude that $A \in \mathcal{F}_T^+$. □

The next result is also often useful for technical purposes.

PROPOSITION 13.11.6. *Let $\{T_n\}_{n\ge1}$ be a sequence of stopping times (for the Brownian filtration) decreasing to a stopping time T. Then $\mathcal{F}_T^+ = \cap_{n\ge1}\mathcal{F}_{T_n}^+$.*

PROOF. By Proposition 13.11.5, $\mathcal{F}_T^+ \subseteq \cap_{n\ge1}\mathcal{F}_{T_n}^+$. To prove the opposite inclusion, take any $A \in \cap_{n\ge1}\mathcal{F}_{T_n}^+$. Then for any $t \ge 0$,

$$\begin{aligned} A \cap \{T < t\} &= A \cap \left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{T_m < t\} \right) \\ &= \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} (A \cap \{T_m < t\}). \end{aligned}$$

But $A \cap \{T_m < t\} \in \mathcal{F}_t^+$ for each m. Thus, $A \cap \{T < t\} \in \mathcal{F}_t^+$, and so by Exercise 13.11.4, $A \in \mathcal{F}_T^+$. □

Using the above results, we now prove a very important fact.

PROPOSITION 13.11.7. *Let T be a stopping time for Brownian motion. Define the 'stopped process'*

$$U(t) := \begin{cases} B(t) & \text{if } t \le T, \\ B(T) & \text{if } t > t. \end{cases}$$

Then U is $\mathcal{F}_T^+$-measurable.

PROOF. For each $n \ge 1$, define

$$T_n := \begin{cases} (m+1)2^{-n} & \text{if } m2^{-n} \le T < (m+1)2^{-n}, \\ \infty & \text{if } T = \infty. \end{cases}$$

We claim that each T_n is a stopping time. To see this, take any $t \ge 0$. Let m be such that $m2^{-n} \le t < (m+1)2^{-n}$. Then

$$\{T_n \le t\} = \{T_n \le m2^{-n}\} = \{T < m2^{-n}\} \in \mathcal{F}_{m2^{-n}}^+ \subseteq \mathcal{F}_t^+.$$

Moreover, T_n decreases to T as $n \to \infty$. Therefore by Propositions 13.11.5 and 13.11.6, $\mathcal{F}_{T_n}^+$ decreases to $\mathcal{F}_T^+$.

Now let U_n be the stopped process for T_n. Note that T_n can take only countably many values. Let t be one of these values. Let $U^{(t)}$ be Brownian motion stopped at time t. It is easy to see that $U^{(t)}$ is $\mathcal{F}_t^+$-measurable. Also, $\{T_n = t\}$ is $\mathcal{F}_t^+$-measurable. Therefore, for any Borel subset A of $C[0,\infty)$, the set

$$\{U_n \in A\} \cap \{T_n = t\} = \{U^{(t)} \in A\} \cap \{T_n = t\}$$

is $\mathcal{F}_t^+$ measurable. From this, it follows that U_n is $\mathcal{F}_{T_n}^+$-measurable. Thus, for any n, U_m is $\mathcal{F}_{T_n}^+$-measurable for all $m \ge n$. Now, $U_m \to U$ in $C[0,\infty)$ as $m \to \infty$. Therefore by Proposition 2.1.14, U is $\mathcal{F}_{T_n}^+$-measurable. Since this holds for every n, Proposition 13.11.6 implies that U is $\mathcal{F}_T^+$-measurable. □

EXERCISE 13.11.8. If T is a stopping time for Brownian motion, prove that the 'stopped' random variable $B(T)$ is $\mathcal{F}_T^+$-measurable.

EXERCISE 13.11.9. Generalize Proposition 13.11.7 and the above exercise to any continuous process adapted to the Brownian filtration.

13.12. The strong Markov property

The following result is known as the strong Markov property of Brownian motion.

THEOREM 13.12.1. *Let B be standard Brownian motion and let T be a stopping time for B. Define $W(t) := B(T+t) - B(T)$ for all $t \geq 0$ if $T < \infty$. If $T = \infty$, let W be an independent Brownian motion. Then W is a standard Brownian motion and is independent of the σ-algebra $\mathcal{F}_T^+$.*

PROOF. First, let us prove the claim under the assumption that T takes values in a countable set $\{t_1, t_2, \ldots\}$, which may include ∞. Take any $A \in \mathcal{F}_T^+$ and any E in the Borel σ-algebra of $C[0,\infty)$. For each n such that t_n is finite, define a process $W_n(t) := B(t_n + t) - B(t_n)$. If $t_n = \infty$, let W_n be an independent Brownian motion. If $T = t_n$, then $W = W_n$. Thus,

$$\{W \in E\} \cap A \cap \{T = t_n\} = \{W_n \in E\} \cap A \cap \{T = t_n\}.$$

Now note that if t_n is finite, then $A \cap \{T = t_n\} = A \cap \{T \leq t_n\} \cap \{T = t_n\}$. Since $A \in \mathcal{F}_T^+$, $A \cap \{T \leq t_n\} \in \mathcal{F}_{t_n}^+$. Also, $\{T = t_n\} = \{T \leq t_n\} \setminus \{T < t_n\} \in \mathcal{F}_{t_n}^+$ (by Exercise 13.11.1). Thus, $A \cap \{T = t_n\} \in \mathcal{F}_{t_n}^+$. On the other hand, by the Markov property, W_n is a standard Brownian motion and is independent of $\mathcal{F}_{t_n}^+$. Thus,

$$\begin{aligned}\mathbb{P}(\{W_n \in E\} \cap A \cap \{T = t_n\}) &= \mathbb{P}(W_n \in E)\mathbb{P}(A \cap \{T = t_n\}) \\ &= \mathbb{P}(B \in E)\mathbb{P}(A \cap \{T = t_n\}).\end{aligned}$$

If $t_n = \infty$, then also the above identity holds. Thus, for any n,

$$\mathbb{P}(\{W \in E\} \cap A \cap \{T = t_n\}) = \mathbb{P}(B \in E)\mathbb{P}(A \cap \{T = t_n\}).$$

Summing over n gives

$$\mathbb{P}(\{W \in E\} \cap A) = \mathbb{P}(B \in E)\mathbb{P}(A).$$

Taking $A = \Omega$ gives $\mathbb{P}(W \in E) = \mathbb{P}(B \in E)$. Thus, W is a Brownian motion, and is independent of $\mathcal{F}_T^+$.

Now take a general stopping time T. Let T_n be defined as in the proof of Proposition 13.11.7. Define $W_n(t) := B(T_n + t) - B(T_n)$ if $T_n < \infty$, and let W_n be an independent Brownian motion if $T_n = \infty$. Since $T_n = \infty$ implies $T = \infty$, it also implies that $T_m = \infty$ for all m. Thus, we can take W_n to be the same independent Brownian motion for every n if $T = \infty$.

Now, since T_n can take only countably many values, we know by the previous deduction that W_n is a standard Brownian motion and that W_n and $\mathcal{F}_{T_n}^+$ are independent. Since $T_n \geq T$, Proposition 13.11.5 implies that $\mathcal{F}_{T_n}^+ \supseteq \mathcal{F}_T^+$. Thus, W_n is independent of $\mathcal{F}_T^+$ for each n. But note that $T_n \to T$ as $n \to \infty$, and hence $W_n \to W$ in $C[0,\infty)$. Thus W is also a standard Brownian motion, and by Proposition 12.10.1, W is independent of $\mathcal{F}_T^+$. □

The strong Markov property can sometimes be used to show that certain random variables are not stopping times. The following is an example.

EXAMPLE 13.12.2. Let

$$T := \sup\{t \in [0,1] : B(t) = 0\}.$$

Although it is intuitively clear that T is not a stopping time, it is not clear how to prove it directly from the definition. The strong Markov property of Brownian motion allows us to give an indirect proof. Since $\mathbb{P}(B(1) = 0) = 0$, the continuity of B implies that $\mathbb{P}(T < 1) = 1$. Let $W(t) := B(T+t) - B(T)$ for $t \geq 0$. Note that W does not change sign in the nonempty interval $(0, 1-T)$. Thus, W cannot be Brownian motion. So we conclude that T is not a stopping time.

EXERCISE 13.12.3. Let $T := \sup\{t : B(t) = t\}$. Prove that T is not a stopping time.

Another interesting application of the strong Markov property is the reflection principle for Brownian motion.

PROPOSITION 13.12.4. *Let B be standard Brownian motion. Take any $a > 0$. Let $T := \inf\{t : B(t) = a\}$. Define a new process $\widetilde{B}$ as*

$$\widetilde{B}(t) := \begin{cases} B(t) & \text{if } t \leq T, \\ 2a - B(t) & \text{if } t > T. \end{cases}$$

Then $\widetilde{B}$ is also a standard Brownian motion.

PROOF. Define a map $\phi : C[0,\infty) \times C[0,\infty) \times [0,\infty) \to C[0,\infty)$ as

$$\phi(f,g,t)(s) := \begin{cases} f(s) & \text{if } s \leq t, \\ f(s) + g(s-t) - g(0) & \text{if } s > t. \end{cases}$$

It is easy to see that ϕ is continuous and hence measurable.

Now let $W(t) := B(T+t) - B(T)$ for $t \geq 0$. Let U be the Brownian motion B stopped at time T, as defined in Proposition 13.11.7. By Proposition 13.11.7, U is $\mathcal{F}_T^+$-measurable. We have also seen earlier that T is $\mathcal{F}_T^+$-measurable. By the strong Markov property, W is a standard Brownian motion and is independent of $\mathcal{F}_T^+$. Finally, recall that $-W$ is also a standard Brownian motion. Thus, the joint law of (U, W, T) is the same as that of $(U, -W, T)$. But $B = \phi(U, W, T)$ and $\widetilde{B} = \phi(U, -W, T)$, where ϕ is the map defined above. Thus, $\widetilde{B}$ and B must have the same law. □

As an application of the reflection principle, let us now calculate the joint law of $(M(t), B(t))$, where $M(t) = \max_{0 \leq s \leq t} B(s)$. Take any $t \geq 0$ and $a > 0$. Let $T := \inf\{s : B(s) = a\}$. Then for any $b < a$, the event $\{M(t) \geq a, B(t) \leq b\}$ happens if and only if $\widetilde{B}(t) \geq 2a - b$, where $\widetilde{B}$ is the reflected process. Thus, for $b < a$,

$$\mathbb{P}(M(t) \geq a,\ B(t) \leq b) = \mathbb{P}(\widetilde{B}(t) \geq 2a - b) = \mathbb{P}(\sqrt{t}Z \geq 2a - b),$$

where $Z \sim N(0,1)$. Now note that $M(t) > 0$ almost surely, and also $B(t) < M(t)$ almost surely by time reversal and Blumenthal's zero-one law. Therefore $\mathbb{P}((M(t), B(t)) \in U) = 1$,

where $U := \{(a,b) : a > 0, -\infty < b < a\}$. So by Exercise 7.6.5, $(M(t), B(t))$ has a probability density function, whose value at a point $(a,b) \in U$ is given by

$$\begin{aligned} -\frac{\partial^2}{\partial a \partial b}\mathbb{P}(\sqrt{t}Z \geq 2a - b) &= -\frac{\partial^2}{\partial a \partial b}\int_{(2a-b)/\sqrt{t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= -\frac{\partial}{\partial a}\frac{1}{\sqrt{2\pi t}} e^{-(2a-b)^2/2t} \\ &= \sqrt{\frac{2}{\pi}}\frac{(2a-b)}{t^{3/2}} e^{-(2a-b)^2/2t}. \end{aligned}$$

EXERCISE 13.12.5. Let B be standard Brownian motion and let $M(t)$ be the maximum of B up to time t. Prove that for any t, $M(t) - B(t)$ has the same law as $|B(t)|$.

13.13. Multidimensional Brownian motion

Let d be a positive integer. A d-dimensional standard Brownian motion B is simply a d-tuple $(B_1, \ldots, B_d)$ of i.i.d. standard Brownian motions. We think of $B(t)$ as the trajectory of a particle moving in $\mathbb{R}^d$, where the coordinates perform independent Brownian motions. A multidimensional Brownian motion B generates its own right continuous filtration $\{\mathcal{F}_t^+\}_{t\geq 0}$; the definition is just the same as in the unidimensional case. It is not hard to see that $\mathcal{F}_t^+$ is the σ-algebra generated by the union of the σ-algebras $\mathcal{F}_{i,t}^+$, $i = 1, \ldots, d$, where $\{\mathcal{F}_{i,t}^+\}_{t\geq 0}$ is the right continuous filtration of B_i.

A stopping time for multidimensional Brownian motion is a nonnegative random variable T such that for each $t \geq 0$, $\{T \leq t\} \in \mathcal{F}_t^+$. The results of Section 13.11 remain valid for such stopping times, with identical proofs.

Note that d-dimensional Brownian motion is a random variable taking value in $C[0,\infty)^d$, which is a Polish space under the product topology. Since Proposition 12.10.1 works for any Polish space, the proof of the Markov property easily carries over to the multidimensional setting. For the same reason, the proof of the strong Markov property also carries over.

In spite of being just a d-tuple of i.i.d. Brownian motions, d-dimensional Brownian motion has many interesting properties that are not shared by one-dimensional Brownian motion. For one thing, the filtration generated by multidimensional Brownian motion is much more complex than the filtration in one dimension. So are the stopping times. The path properties are also richer and more complex. We will get a glimpse of some of this in later chapters.

CHAPTER 14

Martingales in continuous time

In continuous time, a collection of σ-algebras $\{\mathcal{F}_t\}_{t\geq 0}$ is called a filtration if $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$. The filtration is called right-continuous if for any t,

$$\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s.$$

A stopping time T for a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is a $[0,\infty]$-valued random variable such that for any $t \in [0,\infty)$, the event $\{T \leq t\}$ is in $\mathcal{F}_t$. A stopping time T defines a stopped σ-algebra $\mathcal{F}_T$ as

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T \leq t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}.$$

If S and T are stopping times such that $S \leq T$ always, then $\mathcal{F}_S \subseteq \mathcal{F}_T$. The proof is exactly the same as the proof of Proposition 13.11.5. Similarly, if the filtration is right-continuous, and $\{T_n\}_{n\geq 1}$ is a sequence of stopping times decreasing to a stopping time T, then

$$\mathcal{F}_T = \bigcap_{n=1}^{\infty} \mathcal{F}_{T_n}.$$

Again, the proof is exactly identical to the proof of Proposition 13.11.6.

A stopping times is called bounded if it is always bounded above by some deterministic constant.

A collection of random variables $\{X_t\}_{t\geq 0}$, called a stochastic process, is said to be adapted to this filtration for X_t is $\mathcal{F}_t$-measurable for each t. The stochastic process is called right-continuous if for any sample point ω, the map $t \mapsto X_t(\omega)$ is right-continuous.

A stochastic process $\{X_t\}_{t\geq 0}$ adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is called a martingale if $\mathbb{E}|X_t| < \infty$ for each t, and for any $s \leq t$, $\mathbb{E}(X_t|\mathcal{F}_s) = X_s$ a.s.

14.1. Optional stopping theorem in continuous time

The following result extends the optional stopping theorem to continuous time.

THEOREM 14.1.1 (Optional stopping theorem in continuous time). *Suppose that $\{X_t\}_{t\geq 0}$ is a right-continuous martingale adapted to a right-continuous filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Let S and T be bounded stopping times for this filtration. Then X_S and X_T are integrable and $\mathbb{E}(X_T|\mathcal{F}_S) = X_S$ a.s.*

PROOF. Let k be an integer such that $T \leq k-1$ always. For each $n \geq 1$, define

$$T_n := \begin{cases} (m+1)2^{-n} & \text{if } m2^{-n} \leq T < (m+1)2^{-n}, \\ \infty & \text{if } T = \infty. \end{cases}$$

Repeating the argument given in the proof of Proposition 13.11.7, we see that each T_n is a stopping time, $T_n \leq k$ for any n, and T_n decreases to T as $n \to \infty$.

Now fix n and consider the sequence $\{X_{j2^{-n}}\}_{j\geq 0}$. It is easy to verify that this sequence is a discrete time martingale adapted to the filtration $\{\mathcal{F}_{j2^{-n}}\}_{j\geq 0}$. Moreover, we claim that T_n is a stopping time for this filtration, and the stopped σ-algebra of T_n with respect to this filtration is the same as the stopped σ-algebra with respect to the original filtration. To prove the first claim, note that for any j,

$$\{T_n = j2^{-n}\} = \{(j-1)2^{-n} \leq T < j2^{-n}\} \in \mathcal{F}_{j2^{-n}}.$$

For the second claim, first take any $A \in \mathcal{F}$ such that $A \cap \{T_n = j2^{-n}\} \in \mathcal{F}_{j2^{-n}}$ for all j. Take any $t \geq 0$ and find m such that $m2^{-n} \leq t < (m+1)2^{-n}$. Then

$$A \cap \{T_n \leq t\} = A \cap \{T_n \leq m2^{-n}\} \in \mathcal{F}_{m2^{-n}} \subseteq \mathcal{F}_t.$$

Conversely, suppose that $A \cap \{T_n \leq t\} \in \mathcal{F}_t$ for all t. Then clearly for any j,

$$\begin{aligned} A \cap \{T_n = j2^{-n}\} &= (A \cap \{T_n \leq j2^{-n}\}) \setminus (A \cap \{T_n \leq (j-1)2^{-n}\}) \\ &\in \mathcal{F}_{j2^{-n}}. \end{aligned}$$

This proves the second claim. Thus, we can apply the optional stopping theorem for discrete time martingales and deduce that X_{T_n} is integrable and $\mathbb{E}(X_k|\mathcal{F}_{T_n}) = X_{T_n}$ a.s. Since this holds for every n, and $\cap_{n\geq 1}\mathcal{F}_{T_n} = \mathcal{F}_T$, the backwards martingale convergence theorem implies that $X_{T_n} \to \mathbb{E}(X_k|\mathcal{F}_T)$ a.s. and in L^1 as $n \to \infty$. But by the right continuity of $\{X_t\}_{t\geq 0}$, we know that $X_{T_n} \to X_T$ everywhere. Thus, X_T is integrable and $\mathbb{E}(X_k|\mathcal{F}_T) = X_T$ a.s.

But the same logic implies that X_S is integrable and $\mathbb{E}(X_k|\mathcal{F}_S) = X_S$ a.s. Therefore,

$$\mathbb{E}(X_T|\mathcal{F}_S) = \mathbb{E}(\mathbb{E}(X_k|\mathcal{F}_T)|\mathcal{F}_S) = \mathbb{E}(X_k|\mathcal{F}_S) = X_S \text{ a.s.}$$

This completes the proof of the theorem. □

14.2. Doob's L^p inequality in continuous time

We also have a useful extension of Doob's L^p inequality to right-continuous martingales.

THEOREM 14.2.1. *Let $\{X_t\}_{t\geq 0}$ be a right-continuous martingale adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Let $M_t := \max_{0\leq t\leq s} |X_s|$. Then for any $t \geq 0$ and $p > 1$,*

$$\mathbb{E}(M_t^p) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}|X_t|^p.$$

PROOF. Take any $t \geq 0$ and some integer $n \geq 1$. Then $\{X_{kt2^{-n}}\}_{0\leq k\leq 2^n}$ is a martingale adapted to $\{\mathcal{F}_{kt2^{-n}}\}_{0\leq k\leq 2^n}$. Therefore by Doob's L^p inequality for discrete time martingales,

$$\mathbb{E}(\max_{0\leq k\leq 2^n} |X_{kt2^{-n}}|^p) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}|X_t|^p.$$

As $n \to \infty$, the right side stays fixed, whereas the right-continuity of X ensures that the random variable inside the expectation on the left side increases to M_t. Therefore the claim follows by the monotone convergence theorem. □

14.3. Martingales related to Brownian motion

Let B be standard Brownian motion. The Markov property of B gives rise to a class of continuous martingales adapted to the right continuous Brownian filtration.

EXAMPLE 14.3.1. The simplest example of a martingale related to B is B itself. That is, $\mathbb{E}(B(t)|\mathcal{F}_s^+) = B(s)$ a.s. whenever $s \le t$. To see this, define $W(u) := B(s+u) - B(s)$ for $u \ge 0$. Then W is independent of $\mathcal{F}_s^+$, and $B(t) - B(s) = W(t-s)$. Thus,

$$\begin{aligned}\mathbb{E}(B(t)|\mathcal{F}_s^+) &= \mathbb{E}(B(t) - B(s) + B(s)|\mathcal{F}_s^+)\\ &= \mathbb{E}(W(t-s)|\mathcal{F}_s^+) + B(s)\\ &= \mathbb{E}(W(t-s)) + B(s) \text{ a.s.}\\ &= B(s).\end{aligned}$$

EXAMPLE 14.3.2. The process $B(t)^2 - t$ is another martingale related to B. To see this, note that

$$\begin{aligned}&\mathbb{E}(B(t)^2 - t|\mathcal{F}_s^+) = \mathbb{E}[(B(t) - B(s) + B(s))^2|\mathcal{F}_s^+) - t\\ &= \mathbb{E}[(B(t) - B(s))^2 + 2B(s)(B(t) - B(s)) + B(s)^2|\mathcal{F}_s^+] - t\\ &= \mathbb{E}[W(t-s)^2 + 2B(s)W(t-s)|\mathcal{F}_s^+] + B(s)^2 - t.\end{aligned}$$

By the Markov property, $\mathbb{E}(W(t-s)^2|\mathcal{F}_s^+) = \mathbb{E}(W(t-s)^2) = t - s$ a.s. Similarly, since $B(s)$ is $\mathcal{F}_s^+$-measurable and $B(s)W(t-s) \in L^1$,

$$\mathbb{E}(B(s)W(t-s)|\mathcal{F}_s^+) = B(s)\mathbb{E}(W(t-s)|\mathcal{F}_s^+) = 0 \text{ a.s.}$$

Thus, for any $0 \le s \le t$,

$$\mathbb{E}(B(t)^2 - t|\mathcal{F}_s^+) = B(s)^2 - s \text{ a.s.}$$

By a similar procedure, we can find polynomials in $B(t)$ and t of any degree that are martingales with respect to the Brownian filtration.

EXAMPLE 14.3.3. Another class of examples is given by the exponential martingales of Brownian motion. For any $\lambda \in \mathbb{R}$, $e^{\lambda B(t) - \frac{1}{2}\lambda^2 t}$ is a martingale adapted to the Brownian filtration. Again, this is a simple application of the Markov property. Take any $s \le t$ and define $W(u) := B(s+u) - B(s)$, as before. Then

$$\begin{aligned}\mathbb{E}(e^{\lambda B(t) - \frac{1}{2}\lambda^2 t}|\mathcal{F}_s^+) &= e^{\lambda B(s) - \frac{1}{2}\lambda^2 t}\mathbb{E}(e^{\lambda(B(t) - B(s))}|\mathcal{F}_s^+)\\ &= e^{\lambda B(s) - \frac{1}{2}\lambda^2 t}\mathbb{E}(e^{\lambda W(t-s)}|\mathcal{F}_s^+)\\ &= e^{\lambda B(s) - \frac{1}{2}\lambda^2 t}\mathbb{E}(e^{\lambda W(t-s)})\\ &= e^{\lambda B(s) - \frac{1}{2}\lambda^2 t}e^{\frac{1}{2}\lambda^2(t-s)} = e^{\lambda B(s) - \frac{1}{2}\lambda^2 s}.\end{aligned}$$

Let us now work out some applications of the martingales obtained above.

EXAMPLE 14.3.4. Take any $a < 0 < b$. Let $T := \inf\{t : B(t) \notin (a, b)\}$. Then $B(T)$ can take only two possible values, namely, a and b. The optional stopping theorem helps us calculate the respective probabilities, as follows.

Take any $t \ge 0$ and let $T \wedge t$ denote the minimum of T and t. This is the standard notation in this area. By Exercise 13.11.3, $T \wedge t$ is also a stopping time. Moreover, it is a bounded stopping time. Therefore by the optional stopping theorem for right continuous martingales, $\mathbb{E}(B(T \wedge t)) = \mathbb{E}(B(0)) = 0$ for any t.

Now notice that since $T \wedge t \leq T$, $B(T \wedge t) \in [a, b]$. Also, letting $M(t) := \max_{0 \leq s \leq t} B(s)$, we have

$$\begin{aligned}\mathbb{P}(T = \infty) &= \lim_{t \to \infty} \mathbb{P}(T \geq t) \\ &= \lim_{t \to \infty} \mathbb{P}(M(t) \leq b) = 0,\end{aligned}$$

where the last identity follows from the fact that $M(t)$ has the same law as the absolute value of a $N(0, t)$ random variable. Thus, $\mathbb{P}(T < \infty) = 1$, and so $B(T \wedge t) \to B(T)$ a.s. as $t \to \infty$. Combining all of these observations, we see that by the dominated convergence theorem,

$$\mathbb{E}(B(T)) = \lim_{t \to \infty} \mathbb{E}(B(T \wedge t)) = 0.$$

On the other hand,

$$\begin{aligned}\mathbb{E}(B(T)) &= a\mathbb{P}(B(T) = a) + b\mathbb{P}(B(T) = b) \\ &= a\mathbb{P}(B(T) = a) + b(1 - \mathbb{P}(B(T) = a)).\end{aligned}$$

Since the above expression must be equal to zero, we get

$$\mathbb{P}(B(T) = a) = \frac{b}{b - a}, \quad \mathbb{P}(B(T) = b) = \frac{-a}{b - a}.$$

EXAMPLE 14.3.5. Let us now use the martingale $B(t)^2 - t$ to compute $\mathbb{E}(T)$. As before, the optional stopping theorem gives us

$$\mathbb{E}(B(T \wedge t)^2) = \mathbb{E}(T \wedge t)$$

for every t. We have already observed that $B(T \wedge t) \in [a, b]$ for all t, and converges to $B(T)$ as $t \to \infty$. Thus, by the dominated convergence theorem, the right side of the above identity approaches $\mathbb{E}(B(T)^2)$ as $t \to \infty$. On the other hand, the monotone convergence theorem implies that the left side approaches $\mathbb{E}(T)$. Thus,

$$\mathbb{E}(T) = \mathbb{E}(B(T)^2).$$

But we already know the distribution of $B(T)$. This gives us

$$\begin{aligned}\mathbb{E}(T) &= a^2\mathbb{P}(B(T) = a) + b^2\mathbb{P}(B(T) = b) \\ &= a^2 \frac{b}{b - a} + b^2 \frac{-a}{b - a} = -ab.\end{aligned}$$

EXAMPLE 14.3.6. The exponential martingales help us calculate the distribution of the maximum excess of Brownian motion above a linear boundary. Take any $\theta > 0$. Let

$$M := \sup_{t \geq 0} (B(t) - \theta t).$$

By the law of large numbers for Brownian motion, it is easy to see that $M < \infty$ a.s., because otherwise $B(t)$ will have to exceed θt along a sequence of times tending to infinity. Also, $M \geq 0$ since $B(t) - \theta t = 0$ at $t = 0$. We will now use the exponential martingales to derive the distribution of M.

Let $\lambda := 2\theta$ and let $X_t := e^{\lambda B(t) - \frac{1}{2}\lambda^2 t}$, so that X_t is a martingale. Fix any $x > 0$. Let $T := \inf\{t \geq 0 : B(t) \geq \theta t + x\}$. Here we allow T to be infinity, and that is actually an important matter. Clearly, T is a stopping time, and by the optional stopping

theorem, $\mathbb{E}(X_{T\wedge t}) = 1$ for any $t \geq 0$. Now note that for any t, $T \wedge t \leq T$ and hence $B(T \wedge t) \leq \theta(T \wedge t) + x$. Therefore,

$$\begin{aligned} 0 \leq X_{T\wedge t} &\leq e^{\lambda(\theta(T\wedge t)+x)-\frac{1}{2}\lambda^2(T\wedge t)} \\ &= e^{\lambda x+(\lambda\theta-\frac{1}{2}\lambda^2)(T\wedge t)} = e^{2\theta x}. \end{aligned}$$

Now, as $t \to \infty$, $X_{T\wedge t}$ tends to

$$X_T = e^{\lambda B_T-\frac{1}{2}\lambda^2 T} = e^{\lambda(\theta T+x)-\frac{1}{2}\lambda^2 T} = e^{2\theta x}$$

if $T < \infty$. On the other hand, if $T = \infty$, then

$$X_{T\wedge t} = X_t = e^{2\theta(B(t)-\theta t)}$$

for all t, and so by the law of large numbers for Brownian motion, $X_{T\wedge t} \to 0$ as $t \to \infty$. Therefore by the dominated convergence theorem,

$$1 = \mathbb{E}(\lim_{n\to\infty} X_{T\wedge t}) = e^{2\theta x}\mathbb{P}(T < \infty).$$

Thus, $\mathbb{P}(T < \infty) = e^{-2\theta x}$. Consequently, $\mathbb{P}(M \geq x) = e^{-2\theta x}$. That is, M follows an exponential distribution with rate 2θ.

EXERCISE 14.3.7. Let $(B_t)_{t\geq 0}$ be standard Brownian motion starting at zero. For any $b > 0$, show that the process $V_t = \cosh(b|B_t|)e^{-b^2t/2}$ is a martingale adapted to the right continuous Brownian filtration. As an application, take any $a > 0$ and compute the moment generating function of the stopping time $\tau_a := \inf\{t : |B_t| = a\}$.

EXERCISE 14.3.8. Take any $a > 0$ and let $T := \inf\{t \geq 0 : |B(t)| = \sqrt{t+a}\}$. Prove that $\mathbb{E}(T) = \infty$.

EXERCISE 14.3.9. Let B_1 and B_2 be independent standard Brownian motions started at 0. Let $\{\mathcal{F}_t^+\}_{t\geq 0}$ denote the right continuous filtration generated by the pair (B_1, B_2). Prove that for any $\lambda \in \mathbb{R}$, the process

$$X(t) := e^{\lambda B_1(t)} \cos(\lambda B_2(t))$$

is a continuous martingale adapted to this filtration. Let

$$T := \inf\{t \geq 0 : B_1(t) = 1\}.$$

Show that T is a stopping time for the above filtration, and then using the above martingale, compute the characteristic function of $B_2(T)$. (This idenitifies the law of the y coordinate when the x coordinate of a 2D Brownian motion hits 1.)

14.4. Skorokhod's embedding

Skorokhod's embedding is a technique for constructing a randomized stopping time T for Brownian motion such that the stopped random variable $B(T)$ has the same law as any given random variable X with mean zero and finite variance.

Take any $a < 0 < b$. Let $T_{a,b} := \inf\{t \geq 0 : B(t) \notin (a,b)\}$. We derived the law of $B(T)$ in the previous section. Using that, it follows that for any function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[\phi(B(T_{a,b}))] &= \phi(a)\mathbb{P}(B(T) = a) + \phi(b)\mathbb{P}(B(T) = b) \\ &= \frac{\phi(a)b - \phi(b)a}{b-a}. \end{aligned} \tag{14.4.1}$$

Using the above formula, and choosing a and b randomly, we will now construct Skorokhod's embedding.

THEOREM 14.4.1. *Let X be a random variable with mean zero and finite variance. Let μ be the law of X. Define a probability measure ν on the set $(-\infty, 0) \times (0, \infty) \cup \{(0,0)\}$ using the formulas $\nu(\{(0,0)\}) = \mu(\{0\})$, and for any Borel set $A \subseteq (-\infty, 0) \times (0, \infty)$,*

$$\nu(A) := \frac{1}{c} \iint_A (v - u) d\mu(u) d\mu(v),$$

where $c := \int_{(0,\infty)} v d\mu(v)$. Let (U, V) be a pair of random variables with joint law ν, and let B be an independent standard Brownian motion. Let $T := T_{U,V}$, where $T_{a,b}$ is defined as above. Then $B(T)$ has the same law as X. Moreover, $\mathbb{E}(T) = \mathbb{E}(X^2)$.

PROOF. It is not obvious that ν is a probability measure, so let us first prove that. Since $\mathbb{E}(X) = 0$, we have

$$0 = \int_{\mathbb{R}} x d\mu(x) = \int_{(0,\infty)} v d\mu(v) + \int_{(-\infty,0)} u d\mu(u).$$

Thus, the number c can be alternately expressed as $\int_{(-\infty,0)} (-u) d\mu(u)$. By Exercise 2.3.5, ν is a measure. To prove that it is a probability measure, we only have to show that $\nu((-\infty, 0) \times (0, \infty) \cup \{(0,0)\}) = 1$. By construction, $\nu(\{(0,0)\}) = \mu(\{0\})$. On the other hand, by Fubini's theorem,

$$\begin{aligned}\nu((-\infty, 0) \times (0, \infty)) &= \frac{1}{c} \int_{(-\infty,0)} \int_{(0,\infty)} (v - u) d\mu(v) d\mu(u) \\ &= \frac{1}{c} \int_{(-\infty,0)} \int_{(0,\infty)} v d\mu(v) d\mu(u) - \frac{1}{c} \int_{(-\infty,0)} \int_{(0,\infty)} u d\mu(v) d\mu(u) \\ &= \int_{(-\infty,0)} d\mu(u) + \int_{(0,\infty)} d\mu(v) = 1 - \mu(\{0\}).\end{aligned}$$

Thus, ν is a probability measure. Now, we can consider $B(T) = \inf\{t \geq 0 : B(t) \notin (U, V)\}$ as a function of B, U and V. It is not hard to show that it's a measurable function. Therefore by Proposition 9.2.1 and equation (14.4.1), for any bounded continuous function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(B(T))|U, V] = \left(\frac{\phi(U)V - \phi(V)U}{V - U}\right) 1_{\{(U,V) \neq (0,0)\}} + \phi(0) 1_{\{(U,V)=(0,0)\}}.$$

So, to show that $B(T)$ has the same law as X, it suffices to prove that the expected value of the right side equals $\mathbb{E}(\phi(X))$. To show this, note that

$$\begin{aligned}
&\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} \frac{\phi(u)v-\phi(v)u}{v-u}(v-u)d\mu(v)d\mu(u)\\
&=\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} \phi(u)vd\mu(v)d\mu(u)-\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} \phi(v)ud\mu(v)d\mu(u)\\
&=\int_{(-\infty,0)}\phi(u)d\mu(u)+\int_{(0,\infty)}\phi(v)d\mu(v).
\end{aligned}$$

(A slight modification of the above argument also shows integrability, which is left to the reader.) Also, by construction, $\mathbb{P}((U,V)=(0,0))=\mathbb{P}(X=0)$. Thus, $B(T)$ has the same law as X. To show that $\mathbb{E}(T)=\mathbb{E}(X^2)$, we again use Proposition 9.2.1 and the fact that $\mathbb{E}(T_{a,b})=-ab$ (derived in the previous section) to get

$$\mathbb{E}(T|U,V)=-UV$$

and hence

$$\begin{aligned}
\mathbb{E}(T)=-\mathbb{E}(UV)&=-\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} uv(v-u)d\mu(v)d\mu(u)\\
&=-\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} uv^2d\mu(v)d\mu(u)+\frac{1}{c}\int_{(-\infty,0)\times(0,\infty)} u^2vd\mu(v)d\mu(u)\\
&=\int_{(0,\infty)}v^2d\mu(v)+\int_{(-\infty,0)}u^2d\mu(u)=\mathbb{E}(X^2).
\end{aligned}$$

This completes the proof of the theorem. □

14.5. Strassen's coupling

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean zero and variance one. Let $S_0 := 0$ and $S_n := \sum_{i=1}^n X_i$ for $n \geq 1$. Strassen's coupling is a way of constructing the process $\{S_n\}_{n\geq 0}$ and a standard Brownian motion B on the same probability space such that with probability one, $\max_{k\leq n}|S_k - B_k| = o(\sqrt{n})$ as $n \to \infty$. The construction proceeds as follows. Let X be a random variable with the law of the X_i's, and let (U,V) be as in Skorokhod's embedding from the previous section. Let $(U_1,V_1),(U_2,V_2),\ldots$ be a sequence of i.i.d. random pairs with the same law as (U,V). Let B be a standard Brownian motion that is independent of this sequence. Define a sequence of random variables $T_1, T_2, \ldots$ as follows. Let $T_0 := 0$, $T_1 := \inf\{t \geq 0 : B(t) \notin (U_1,V_1)\}$, and for each $n \geq 1$, let

$$T_n := \inf\{t \geq T_{n-1} : B(t) - B(T_{n-1}) \notin (U_n, V_n)\}.$$

The following theorem establishes the coupling property.

THEOREM 14.5.1. *The sequence $\{B(T_n)\}_{n\geq 0}$ has the same law as the sequence $\{S_n\}_{n\geq 0}$. Moreover, the sequence $\{T_n - T_{n-1}\}_{n\geq 1}$ is i.i.d. with mean* 1.

We need a small lemma about measurability.

LEMMA 14.5.2. *Let $T_{a,b}$ be as in the previous section. Then for any $t \in \mathbb{R}$, the maps $(a,b) \mapsto \mathbb{P}(B(T_{a,b}) \leq t)$ and $(a,b) \mapsto \mathbb{P}(T_{a,b} \leq t)$ are measurable.*

PROOF. The value of $\mathbb{P}(B(T_{a,b}) \le t)$ is explicitly given by equation (14.4.1) with $\phi(x) = 1_{\{x \le t\}}$, which shows its measurability as a function of (a, b). The map $(a, b) \mapsto \mathbb{P}(T_{a,b} \le t)$ is increasing in a and decreasing in b. It is not difficult to show that any such map is measurable. □

PROOF OF THEOREM 14.5.1. For each n, let $Y_n := B(T_n) - B(T_{n-1})$. We will show that $Y_1, Y_2, \ldots$ are i.i.d. with the same law as the X_i's. Since the law of a random vector is characterized by its cumulative distribution function (Exercise 7.6.1) and the law of an infinite sequence of random variables is uniquely characterized by its finite dimensional distributions (Theorem 3.3.1), it suffices to show that for any n and any bounded continuous functions $t_1, \ldots, t_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \le t_1, \ldots, X_n \le t_n) = \mathbb{P}(Y_1 \le t_1, \ldots, Y_n \le t_n). \tag{14.5.1}$$

Fix $t_1, \ldots, t_n \in \mathbb{R}$ and let $\phi_i(x) := 1_{\{x \le t_i\}}$ for each i. Suppose that instead of random (U_i, V_i), we had deterministic (u_i, v_i) and defined T_n and Y_n as above with these u_i's and v_i's. Then each T_n would be a stopping time for the Brownian filtration, and $0 = T_0 \le T_1 \le T_2 \le \cdots$. Let $W_n(t) := B(T_n + t) - B(T_n)$. Then by the strong Markov property, each W_n is a standard Brownian motion, and is independent of $\mathcal{F}^+_{T_n}$. Also, note that for each n,

$$T_n - T_{n-1} = \inf\{t \ge 0 : W_{n-1}(t) \notin (u_n, v_n)\},$$

and $Y_1, \ldots, Y_{n-1}$ are $\mathcal{F}^+_{T_{n-1}}$-measurable (because $B(T_i)$ is $\mathcal{F}^+_{T_i}$-measurable for all i and $\mathcal{F}^+_{T_i} \subseteq \mathcal{F}^+_{T_j}$ when $i \le j$). Lastly, note that $Y_n = W_{n-1}(T_n - T_{n-1})$. Thus,

$$\begin{aligned}
&\mathbb{E}(\phi_1(Y_1) \cdots \phi_n(Y_n) | \mathcal{F}^+_{T_{n-1}}) \\
&= \phi_1(Y_1) \cdots \phi_{n-1}(Y_{n-1}) \mathbb{E}[\phi_n(W_n(T_n - T_{n-1})) | \mathcal{F}^+_{T_{n-1}}] \\
&= \phi_1(Y_1) \cdots \phi_{n-1}(Y_{n-1}) \mathbb{E}[\phi_n(W_n(T_n - T_{n-1}))].
\end{aligned}$$

By Lemma 14.5.2 and the observations made above, $\mathbb{E}[\phi_n(W_n(T_n - T_{n-1}))]$ is a measurable function of (u_n, v_n). Let's call it $\psi_n(u_n, v_n)$. Proceeding inductively, we get

$$\mathbb{E}(\phi_1(Y_1) \cdots \phi_n(Y_n)) = \psi_1(u_1, v_1) \cdots \psi_n(u_n, v_n).$$

Now consider the case of random (U_i, V_i), generated as before. Considering the random variables $Y_1, \ldots, Y_n$ as functions of B and $(U_1, V_1), \ldots, (U_n, V_n)$, and applying Proposition 9.2.1 and the above identity, we get

$$\mathbb{E}(\phi_1(Y_1) \cdots \phi_n(Y_n)) = \prod_{i=1}^{n} \mathbb{E}(\psi_i(U_i, V_i)).$$

But by Skorokhod's embedding theorem and our definition of the T_i's,

$$\mathbb{E}(\psi_i(U_i, V_i)) = \mathbb{E}(\phi_i(X_i)).$$

This proves (14.5.1). So we have shown that (14.5.1) holds for every n, which completes the proof of the first claim of the theorem. The proof of the second claim is exactly similar, using $Z_i := T_i - T_{i-1}$ instead of Y_i. □

Strassen's coupling can also be used to give an alternative proof of Donsker's theorem, assuming that we construct Brownian motion by some other means (for example, by Exercise 13.4.7). This is outlined in the following exercises.

EXERCISE 14.5.3. Let $X_1, X_2, \dots$ be i.i.d. random variables with mean zero and variance one. Let $S_n := \sum_{i=1}^n X_i$. Using Strassen's coupling, prove that it is possible to construct $\{S_n\}_{n\geq 1}$ and a standard Brownian motion B on the same probability space such that $n^{-1/2}\max_{0\leq k\leq n}|S_k - B(k)| \to 0$ in probability as $n\to\infty$. (Hint: First, show that $n^{-1}\max_{0\leq k\leq n}|T_k - k| \to 0$ a.s. This is a simple consequence of the fact that $T_n/n \to 1$ a.s. Then use this to control $\max_{0\leq k\leq n}|B(T_k) - B(k)|$.)

EXERCISE 14.5.4. Let B_n be the piecewise linear path from Donsker's theorem. Using the previous exercise, prove that B_n and a standard Brownian motion B can be constructed on the same probability space such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|B_n - B\|_{[0,1]} > \epsilon) = 0.$$

(In other words, $B_n \to B$ in probability on $C[0,1]$.) As a consequence, prove Donsker's theorem.

Incidentally, if we have more information about the X_i's, such as finiteness of higher moments or the moment generating function, it is possible to construct better couplings. For example, under the assumption of finite moment generating function, the famous Komlós–Major–Tusnády embedding constructs $\{S_n\}_{n\geq 0}$ and B on the same space such that $\{(\log n)^{-1}\max_{0\leq k\leq n}|S_k - B(k)|\}_{n\geq 0}$ is a tight family. Proving this is beyond the scope of this discussion.

14.6. The Hartman–Wintner LIL

Combined with the better version of the law of iterated logarithm for Brownian motion (Theorem 13.10.3), Strassen's coupling immediately implies the law of the iterated logarithm for sums of i.i.d. random variables with finite variance. This is known as the Hartman–Wintner LIL.

COROLLARY 14.6.1. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean zero and variance one, and let $S_n := \sum_{i=1}^n X_i$. Then with probability one,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1, \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1.$$

PROOF. Let T_n be as in Theorem 14.5.1, so that the sequence $\{B(T_n)\}_{n\geq 0}$ has the same law as the sequence $\{S_n\}_{n\geq 0}$. From Theorem 14.5.1 and the strong law of large numbers, we get $T_n/n \to 1$ a.s. In particular, $T_{n+1}/T_n \to 1$ a.s. Therefore by Theorem 13.10.3,

$$\limsup_{n\to\infty} \frac{B(T_n)}{\sqrt{2n\log\log n}} = \limsup_{n\to\infty} \frac{B(T_n)}{\sqrt{2T_n\log\log T_n}} = 1 \quad \text{a.s.}$$

Replacing S_n by $-S_n$, we get the result for lim inf. □

CHAPTER 15

Introduction to stochastic calculus

This chapter introduces the basics of stochastic calculus and some applications of these methods. For technical simplicity, we will not attempt to prove the most general versions of the results, but only as much as needed.

15.1. Continuous stochastic processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A collection of real-valued random variables $\{X(t)\}_{t\geq 0}$ is called a stochastic process in continuous time. Such a process is called continuous if for each $\omega \in \Omega$, the map $t \mapsto X(t)(\omega)$ is continuous. We will encounter such processes often in this chapter. A continuous stochastic process X can also be viewed as a $C[0,\infty)$-valued random variable by defining $X(\omega)(t) := X(t)(\omega)$. To see that X is measurable, simply observe that it is the pointwise limit (in $C[0,\infty)$) of a sequence of piecewise linear $C[0,\infty)$-valued random variables, and apply Proposition 2.1.14.

A continuous stochastic process X is said to be adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if for each t, $X(t)$ is $\mathcal{F}_t$-measurable. As above, it is easy to see that the restriction of X to $[0,t]$ is a $C[0,t]$-valued random variable.

15.2. The Itô integral

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which a standard Brownian motion B is defined. Let $\{\mathcal{F}_t^+\}_{t\geq 0}$ be the right continuous filtration of B. Let $\{X(t)\}_{t\geq 0}$ be a stochastic process adapted to this filtration. Given some $t \geq 0$, Itô integration is a way of giving meaning to the integral

$$\int_0^t X(s)dB(s)$$

as a limit of sums like

$$\sum_{i=0}^{n-1} X(s_i)(B(s_{i+1}) - B(s_i))$$

where $0 = s_0 < s_1 < \cdots < s_n = t$, as the mesh size $\max_{0\leq i\leq n-1} |s_{i+1} - s_i|$ tends to zero. Although this is in the spirit of the Riemann–Stieltjes integral $\int_0^t f(s)dg(s)$ of one function f with respect to another function g, the situation here is more complicated. For instance, the Riemann–Stieltjes integral is defined only when g has finite total variation, and by Exercise 13.7.2, we know that Brownian motion does not have finite total variation. Indeed, the partial sum displayed above usually does not converge almost surely as the mesh size tends to zero. Itô's insight, however, was to realize that convergence takes place in L^2 under mild conditions. We will now carry out this construction.

Fix some $t \geq 0$. The first step is to define the Itô integral for simple processes. A stochastic process $X = \{X(s)\}_{s\leq t}$ is called simple if there exist some deterministic M, n

and $0 = s_0 < s_1 < \cdots < s_n = t$ such that for each $i < n$, $X(s) = X(s_i)$ for all $s \in [s_i, s_{i+1})$, and $|X(s_i)| \leq M$ always. Suppose that X is adapted to the Brownian filtration, meaning that for each s, $X(s)$ is $\mathcal{F}_s^+$-measurable. Note that this is equivalent to saying that $X(s_i)$ is $\mathcal{F}_{s_i}^+$-measurable for each i. We define

$$\int_0^t X(s)dB(s) := \sum_{i=0}^{n-1} X(s_i)(B(s_{i+1}) - B(s_i)).$$

For simplicity, let us denote the integral by $I(X)$. It is not hard to see that this is well-defined, in the sense that if we take two different representations of X as a simple process, the value of $I(X)$ would be the same for both.

The Itô integral on simple processes enjoys two important properties. First, it is linear, meaning that if X and Y are two simple processes, and a and b are real numbers, then $aX + bY$ is also a simple process, and $I(aX + bY) = aI(X) + bI(Y)$. This is easy to verify from the definition. The second important property is the Itô isometry property, which says that

$$\mathbb{E}(I(X)^2) = \int_0^t \mathbb{E}(X(s)^2)ds.$$

To see this, note that for any $0 \leq j < i < n$, the facts that X is adapted to the Brownian filtration, X is bounded, and B is a martingale together imply that

$$\begin{aligned}&\mathbb{E}[X(s_i)(B(s_{i+1}) - B(s_i))X(s_j)(B(s_{j+1}) - B(s_j))|\mathcal{F}_{s_i}^+] \\ &= X(s_i)(B(s_{i+1}) - B(s_i))X(s_j)\mathbb{E}[(B(s_{j+1}) - B(s_j))|\mathcal{F}_{s_i}^+] = 0 \text{ a.s.},\end{aligned}$$

and therefore

$$\mathbb{E}[X(s_i)(B(s_{i+1}) - B(s_i))X(s_j)(B(s_{j+1}) - B(s_j))] = 0.$$

Similarly, by the Markov property of Brownian motion,

$$\begin{aligned}\mathbb{E}[X(s_i)^2(B(s_{i+1}) - B(s_i))^2|\mathcal{F}_{s_i}^+] &= X(s_i)^2\mathbb{E}[(B(s_{i+1}) - B(s_i))^2|\mathcal{F}_{s_i}^+] \\ &= X(s_i)^2(s_{i+1} - s_i).\end{aligned}$$

As a consequence of these two observations, we get

$$\begin{aligned}\mathbb{E}(I(X)^2) &= \sum_{i=0}^{n-1} \mathbb{E}(X(s_i)^2)(s_{i+1} - s_i) \\ &= \int_0^t \mathbb{E}(X(s)^2)ds,\end{aligned}$$

which is the Itô isometry. This is called an isometry because of the following reason. Consider the product space $[0, t] \times \Omega$ endowed with the product σ-algebra and the product measure. We may view X as a measurable map $(s, \omega) \mapsto X(s)(\omega)$ from this product space into the real line. Then the right side of the Itô isometry is the squared L^2 norm of X, and the left side is the L^2 norm of $I(X)$. So the map I is an isometry from a subspace of $L^2([0, t] \times \Omega)$ into $L^2(\Omega)$.

We will now extend the definition of I to continuous adapted processes. It is possible to further extend it to square-integrable processes, but we will refrain from doing that to avoid technical complications.

Let $\{X(s)\}_{s\le t}$ be a continuous stochastic process adapted to the Brownian filtration, meaning that $X(s)$ is $\mathcal{F}_s^+$-measurable for each s, and for each $\omega \in \Omega$, the map $s \mapsto X(s)(\omega)$ is continuous. Suppose that

$$\int_0^t \mathbb{E}(X(s)^2)ds < \infty. \tag{15.2.1}$$

(We will see below that the integral makes sense because the continuity of X implies that the map $s \mapsto \mathbb{E}(X(s)^2)$ is measurable.) Under this condition, we want to define $I(X)$. The key observation is the following.

LEMMA 15.2.1. *Let X be as above. Define $X(s,\omega) := X(s)(\omega)$. Then $X \in L^2([0,t]\times\Omega)$, and there is a sequence of simple adapted processes $\{X_n\}_{n\ge1}$ such that $X_n \to X$ in $L^2([0,t]\times\Omega)$.*

PROOF. For each n, define

$$Z_n(s,\omega) := X(kt/n,\omega) \quad \text{if} \quad \frac{kt}{n} \le s < \frac{(k+1)t}{n}. \tag{15.2.2}$$

It is easy to show that each Z_n is measurable. Since X is the pointwise limit of Z_n as $n \to \infty$ (due to the continuity of $s \mapsto X(s,\omega)$ for each ω), Proposition 2.1.14 tells us that X is also measurable. Furthermore, by the assumption (15.2.1) and Fubini's theorem, $X \in L^2([0,t]\times\Omega)$.

Suppose that there is some constant M such that $|X(s,\omega)| \le M$ for all s and ω. Let Z_n be as above. Then $|Z_n(s,\omega)| \le M$ for all n, s and ω and $Z_n \to X$ pointwise. So by the dominated convergence theorem,

$$\lim_{n\to\infty} \|Z_n - X\|_{L^2([0,t]\times\Omega)} = 0. \tag{15.2.3}$$

Let us now drop the assumption that X is bounded. For each positive integer M, define

$$Y_M(s,\omega) := \begin{cases} M & \text{if } X(s,\omega) > M, \\ X(s,\omega) & \text{if } -M \le X(s,\omega) \le M, \\ -M & \text{if } X(s,\omega) < -M. \end{cases} \tag{15.2.4}$$

By the condition (15.2.1) and the dominated convergence theorem, we get that

$$\lim_{M\to\infty} \|Y_M - X\|_{L^2([0,T]\times\Omega)} = 0.$$

But Y_M is a continuous adapted process that is bounded by a constant. Therefore by our previous deduction, we can find a simple adapted process X_M such that

$$\|Y_M - X_M\|_{L^2([0,T]\times\Omega)} \le \frac{1}{M}.$$

The sequence $\{X_M\}_{M\ge1}$ satisfies the requirement of the lemma. □

We are now ready to define the Itô integral $I(X)$ for any continuous adapted process X. Take a sequence of simple processes $\{X_n\}_{n\ge1}$ that converge to X in $L^2([0,t]\times\Omega)$. Then by the Itô isometry and linearity of I, $\{I(X_n)\}_{n\ge1}$ is a Cauchy sequence in $L^2(\Omega)$. We define $I(X)$ to be its L^2 limit. The Itô isometry ensures that this limit does not depend on the choice of the approximating sequence. Furthermore, since $I(X)$ is constructed as an L^2 limit, it is clear that $X \mapsto I(X)$ is linear and the Itô isometry continues to hold. Lastly,

note that $I(X)$ is $\mathcal{F}_t^+$-measurable, since by Lemma 4.4.5, $I(X_n)$ converges almost surely to $I(X)$ along a subsequence, and each $I(X_n)$ is $\mathcal{F}_t^+$-measurable (and then apply Proposition 2.1.14).

The above procedure defines the integral from 0 to t, for any t. What about integration from s to t for some $0 \le s \le t$? We can retrace the exact same steps to define $\int_s^t X(u)dB(u)$ for any continuous adapted process X satisfying

$$\int_s^t \mathbb{E}(X(u)^2)du < \infty.$$

This integral satisfies the Itô isometry

$$\mathbb{E}\left[\left(\int_s^t X(u)dB(u)\right)^2\right] = \int_s^t \mathbb{E}(X(u)^2)du.$$

Moreover, when X satisfies (15.2.1), the above integral also satisfies the natural identity

$$\int_0^t X(u)dB(u) = \int_0^s X(u)dB(u) + \int_s^t X(u)dB(u) \quad \text{a.s.}$$

All of the above can be proved by first considering simple processes and then passing to the L^2 limit. The details are left to the reader.

EXERCISE 15.2.2. Let X and Y be two continuous adapted processes satisfying (15.2.1) for some $t \ge 0$. Suppose that $\int_0^t X(s)dB(s) = \int_0^t Y(s)dB(s)$ a.s. Then prove that with probability one, $X(s) = Y(s)$ for all $s \in [0,t]$.

EXERCISE 15.2.3. Let X and Y be two continuous adapted processes satisfying (15.2.1). Prove that for any $t \ge 0$, the event

$$\left\{\int_0^t X(s)dB(s) \ne \int_0^t Y(s)dB(s)\right\} \cap \{X(s) = Y(s) \text{ for all } s \in [0,t]\}$$

has probability zero. (Hint: Approximate by simple processes.)

EXERCISE 15.2.4. Let $f : [0,t] \to \mathbb{R}$ be a (deterministic) continuous function. Prove that $\int_0^t f(s)dB(s)$ is a normal random variable with mean zero and variance $\int_0^t f(s)^2 ds$.

15.3. The Itô integral as a continuous martingale

Let all notation be as in the previous section. Let $X = \{X(t)\}_{t\ge 0}$ be a continuous stochastic process adapted to the Brownian filtration, and satisfying the square-integrability condition (15.2.1) for all t. We now know how to define

$$Z(t) := \int_0^t X(s)dB(s)$$

for any $t \ge 0$. However, this definition is only in an almost sure sense, since the Itô integral is defined as an L^2 limit. Can we produce versions of $Z(t)$ in a way such that with probability one, $t \mapsto Z(t)$ is continuous? It turns out that this is possible. Moreover, this process is a continuous martingale. We will need to assume that $\mathcal{F}_t^+$ is a complete σ-algebra for every t, which can be easily arranged by replacing it with its completion (see Section 1.8). Henceforth, we will work under this assumption.

THEOREM 15.3.1. *Let X be as above. Then there exists a continuous martingale $\{Z(t)\}_{t\geq 0}$ adapted to the Brownian filtration such that for each $t \geq 0$,*

$$Z(t) = \int_0^t X(s)dB(s) \text{ a.s.}$$

Moreover, if Z' is another such martingale, then $\mathbb{P}(Z(t) = Z'(t) \text{ for all } t) = 1$.

PROOF. Fix some $t \geq 0$. Temporarily, denote the restriction of X to the interval $[0, t]$ also by X. We will first construct a continuous martingale $\{Z(s)\}_{s\leq t}$ such that for each $s \in [0, t]$, $Z(s) = \int_0^s X(u)dB(u)$ a.s. Later, we will patch these martingales together to create a single continuous martingale.

By Lemma 15.2.1, find a sequence of simple adapted processes $\{X_n\}_{n\geq 1}$ that converge to X in $L^2([0, t] \times \Omega)$. For each n and each $s \leq t$, define

$$Z_n(s) := \int_0^s X_n(u)dB(u)$$

using the definition of the Itô integral for simple processes. Then $Z_n = \{Z_n(s)\}_{s\leq t}$ is a continuous adapted process. Moreover, using the Markov property of Brownian motion, it is not difficult to see that Z_n is martingale with respect to the Brownian filtration. Therefore by Doob's L^2 maximal inequality for continuous martingales,

$$\mathbb{E}\left[\sup_{0\leq s\leq t}(Z_n(s) - Z_m(s))^2\right] \leq 4\mathbb{E}[(Z_n(t) - Z_m(t))^2].$$

But by the Itô isometry,

$$\mathbb{E}[(Z_n(t) - Z_m(t))^2] = \|X_n - X_m\|^2_{L^2([0,t]\times\Omega)}.$$

Since $\{X_n\}_{n\geq 1}$ is a Cauchy sequence in $L^2([0, t] \times \Omega)$, this allows us to find a subsequence $\{Z_{n_k}\}_{k\geq 1}$ such that for each k,

$$\mathbb{E}\left[\sup_{0\leq s\leq t}(Z_{n_k}(s) - Z_{n_{k+1}}(s))^2\right] \leq 2^{-k}.$$

A simple application of the first Borel–Cantelli lemma now shows that with probability one, the sequence $\{Z_{n_k}\}_{k\geq 1}$, viewed as a sequence of random continuous functions on $[0, t]$, is Cauchy with respect to the supremum norm. Let Z denote its limit. Being the uniform limit of a sequence of continuous functions, Z is continuous. Since for each s, the restriction of X_{n_k} to $[0, s]$ converges to the restriction of X to $[0, s]$ in $L^2([0, s] \times \Omega)$, we conclude that $Z(s)$ must be a version of $\int_0^s X(u)dB(u)$. Moreover, this also shows that $Z_{n_k}(s) \to Z(s)$ in L^2.

Since each $Z_{n_k}(s)$ is $\mathcal{F}_s^+$-measurable and $Z_{n_k}(s) \to Z(s)$ a.s., the completeness of $\mathcal{F}_s^+$ shows that $Z(s)$ is $\mathcal{F}_s^+$-measurable (see Exercise 2.6.4). Finally, to prove that Z is a martingale, recall that Z_n is a martingale for each n. So for any $0 \leq u \leq s \leq t$ and any k,

$$\mathbb{E}(Z_{n_k}(s)|\mathcal{F}_u^+) = Z_{n_k}(u) \text{ a.s.}$$

We know that the right side converges to $Z(s)$ a.s. as $k \to \infty$. For the left side, recall that $Z_{n_k}(s) \to Z(s)$ in L^2 and hence also in L^1. Therefore by Exercise 9.2.2, $\mathbb{E}(Z_{n_k}(s)|\mathcal{F}_u^+) \to \mathbb{E}(Z(s)|\mathcal{F}_u^+)$ a.s. as $k \to \infty$. Thus, Z is a martingale.

Using the above procedure we can construct, for each $n \geq 1$, a continuous martingale $\{Z^n(t)\}_{0\leq t\leq n}$ such that for each $t \in [0, n]$, $Z^n(t) = \int_0^t X(s)dB(s)$ a.s. Take any $1 \leq m \leq n$. Then with probability one, for each rational $t \in [0, m]$, $Z^m(t) = Z^n(t)$. But Z^m and Z^n are continuous processes. So, with probability one, $Z^m(t) = Z^n(t)$ for all $t \in [0, m]$. Thus, if we define $Z(t) := Z^n(t)$ for $n = \lceil t \rceil$, then $\{Z(t)\}_{t\geq 0}$ is a continuous martingale satisfying $Z(t) = \int_0^t X(s)dB(s)$ a.s. for each given t. Uniqueness follows by continuity. □

EXERCISE 15.3.2. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Let $X(t)$ be the continuous martingale representing $\int_0^t f(s)dB(s)$. Prove that X is a centered Gaussian process, meaning that the finite dimensional distributions are jointly normal with mean zero. Show that $\mathrm{Cov}(X(s), X(t)) = \int_0^{s\wedge t} f(u)^2 du$.

15.4. The quadratic variation process

Let B be a standard Brownian motion and X be a continuous stochastic process adapted to the filtration of B, and satisfying the square-integrability condition (15.2.1) for every t. Let

$$Z(t) := \int_0^t X(s)dB(s)$$

be the continuous martingale obtained in the previous section. The quadratic variation process of Z is defined as

$$[Z](t) := \int_0^t X(s)^2 ds.$$

By Exercise 15.2.2, $[Z]$ is indeed determined by Z, in the sense that if Z is alternatively expressed as the stochastic integral of some other process Y, we will end up with the same $[Z]$. Note that $[Z]$ is an adapted, continuous, and non-decreasing stochastic process. The following is one of the main properties of this process.

THEOREM 15.4.1. *Let Z and $[Z]$ be as above. Then the process $Z(t)^2 - [Z](t)$ is a continuous martingale adapted to the Brownian filtration.*

We need the following simple lemma.

LEMMA 15.4.2. *If $\{U_n\}_{n\geq 1}$ is a sequence of real-valued random variables converging in L^2 to U, the $U_n^2 \to U^2$ in L^1.*

PROOF. By the Cauch–Schwarz inequality and the triangle inequality,

$$\begin{aligned}\mathbb{E}|U_n^2 - U^2| &= \mathbb{E}|(U_n - U)(U_n + U)| \\ &\leq \|U_n - U\|_{L^2}\|U_n + U\|_{L^2} \\ &\leq \|U_n - U\|_{L^2}^2 + 2\|U_n - U\|_{L^2}\|U\|_{L^2}.\end{aligned}$$

Since $U_n \to U$ in L^2, the right side tends to 0 as $n \to \infty$. □

PROOF OF THEOREM 15.4.1. Fix $t \geq 0$. Let $\{X_n\}_{n\geq 1}$ be a sequence of simple adapted processes converging to X in $L^2([0, t] \times \Omega)$. Such a sequence exists by Lemma 15.2.1. Let $Z_n(s) := \int_0^s X_n(u)dB(u)$ for each u and let $[Z_n]$ be the quadratic variation process of Z_n.

It is easy to verify by direct computation that for any $s \le t$ and any n,

$$\mathbb{E}((Z_n(t) - Z_n(s))^2 | \mathcal{F}_s^+) = \mathbb{E}\left(\int_s^t X_n(u)^2 du \middle| \mathcal{F}_s^+ \right).$$

Since Z_n is a martingale adapted to the Brownian filtration, this shows that

$$\begin{aligned}
&\mathbb{E}(Z_n(t)^2 | \mathcal{F}_s^+) \\
&= \mathbb{E}((Z_n(t) - Z_n(s))^2 + 2Z_n(s)(Z_n(t) - Z_n(s)) + Z_n(s)^2 | \mathcal{F}_s^+) \\
&= \mathbb{E}\left(\int_s^t X_n(u)^2 du \middle| \mathcal{F}_s^+ \right) + Z_n(s)^2 \\
&= \mathbb{E}([Z_n](t) - [Z_n](s) | \mathcal{F}_s^+) + Z_n(s)^2.
\end{aligned}$$

Since $[Z_n](s)$ is $\mathcal{F}_s^+$-measurable, we get

$$\mathbb{E}(Z_n(t)^2 - [Z_n](t) | \mathcal{F}_s^+) = Z_n(s)^2 - [Z_n](s).$$

Now, by Lemma 15.4.2, it follows that $Z_n(s)^2 \to Z(s)^2$ in L^1 and $[Z_n](s) \to [Z](s)$ in L^1. Combining this information with the above identity, we get

$$\mathbb{E}(Z(t)^2 - [Z](t) | \mathcal{F}_s^+) = Z(s)^2 - [Z](s) \quad \text{a.s.},$$

which completes the proof. □

An important corollary of Theorem 15.4.1 is the following result, which allows us to apply the Itô isometry up to a bounded stopping time. Depending on the situation, one can then hope to extend it to unbounded stopping times.

COROLLARY 15.4.3. *Let X be a continuous adapted process satisfying the square-integrability condition (15.2.1). Let S and T be bounded stopping times for the Brownian filtration, such that $S \le T$ always. Then*

$$\mathbb{E}\left[\left(\int_S^T X(s) dB(s) \right)^2 - \int_S^T X(s)^2 ds \middle| \mathcal{F}_S^+ \right] = 0 \text{ a.s.}$$

PROOF. Let $Z(t)$ be the continuous martingale version of $\int_0^t X(s) dB(s)$. By Theorem 15.4.1, $Z(t)^2 - [Z](t)$ is also a continuous martingale. Therefore by the optional stopping theorem for continuous martingales, we get $\mathbb{E}(Z(T) | \mathcal{F}_S^+) = Z(S)$ and $\mathbb{E}(Z(T)^2 - [Z](T) | \mathcal{F}_S^+) = Z(S)^2 - [Z](S)$. The second identity can be rewritten as

$$\begin{aligned}
\mathbb{E}(Z(T)^2 - Z(S)^2 | \mathcal{F}_S^+) &= \mathbb{E}([Z](T) - [Z](S) | \mathcal{F}_S^+) \\
&= \mathbb{E}\left[\int_S^T X(s)^2 ds \middle| \mathcal{F}_S^+ \right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left(\int_S^T X(s) dB(s) \right)^2 \middle| \mathcal{F}_S^+ \right] &= \mathbb{E}((Z(T) - Z(S))^2 | \mathcal{F}_S^+) \\
&= \mathbb{E}(Z(T)^2 - 2Z(T)Z(S) + Z(S)^2 | \mathcal{F}_S^+) \\
&= \mathbb{E}(Z(T)^2 - Z(S)^2 | \mathcal{F}_S^+).
\end{aligned}$$

Combined with the preceding display, this proves the corollary. □

15.5. Itô integrals for multidimensional Brownian motion

Let B be d-dimensional standard Brownian motion and let $\{X(t)\}_{t\geq 0}$ be a real-valued continuous process adapted to the right continuous filtration of B, and satisfying the condition

$$\int_0^t \mathbb{E}(X(s)^2)ds < \infty$$

for each $t \geq 0$. Then the stochastic integral

$$\int_0^t X(s)dB_i(s)$$

is defined for each i and t, and can be represented as a continuous martingale adapted to the filtration of B. Note that the only difference here is that X not adapted to the filtration of B_i, but the larger filtration generated by B. The definitions of the integral and the continuous martingale are just as before; all steps go through without any problem because the Markov property is valid for multidimensional Brownian motion, as we observed in Section 13.13.

15.6. Itô's formula

Itô's formula is the stochastic analogue of the fundamental theorem of calculus. In its simplest form, it says the following.

THEOREM 15.6.1. *Let B be standard Brownian motion. Take any twice continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that for each $t \geq 0$,*

$$\int_0^t \mathbb{E}(f'(B(s))^2)ds < \infty.$$

Then with probability one, we have that for all $t \geq 0$,

$$f(B(t)) = f(B(0)) + \int_0^t f'(B(s))dB(s) + \frac{1}{2}\int_0^t f''(B(s))ds,$$

where first integral should be interpreted as the continuous martingale given by Theorem 15.3.1.

Note that if B was a function with finite total variation, we would only have the first integral on the right side. The appearance of the second integral is the main distinctive feature of Itô's formula. The reason for the appearance of this term is the following lemma.

LEMMA 15.6.2. *Let B be standard Brownian motion and $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Take any $t \geq 0$ and a partition $0 = s_0 < s_1 < \cdots < s_n = t$. Then the sum*

$$\sum_{i=0}^{n-1} g(B(s_i))(B(s_{i+1}) - B(s_i))^2$$

converges in probability to $\int_0^t g(B(s))ds$ as the mesh size $\max_{0\leq i<n} |s_{i+1} - s_i|$ approaches zero.

PROOF. Let us first prove the lemma under the assumption that g is bounded, in addition to being continuous. Since $s \mapsto g(B(s))$ is a continuous map, the sum

$$\sum_{i=0}^{n} g(B(s_i))(s_{i+1} - s_i)$$

converges to $\int_0^t g(B(s))ds$ as the mesh size approaches zero. So it suffices to prove that

$$\mathbb{E}\left[\left(\sum_{i=0}^{n} g(B(s_i))((B(s_{i+1}) - B(s_i))^2 - (s_{i+1} - s_i))\right)^2\right] \tag{15.6.1}$$

tends to zero as the mesh size goes to zero. When we expand the square and take the expectation inside, the cross terms vanish due to the Markov property of Brownian motion, which implies that for any $j < i$,

$$\mathbb{E}((B(s_{i+1}) - B(s_i))^2 | \mathcal{F}^+_{s_j}) = s_{i+1} - s_i \quad \text{a.s.}$$

Thus, the expectation displayed in (15.6.1) equals

$$\sum_{i=0}^{n-1} \mathbb{E}[g(B(s_i))^2((B(s_{i+1}) - B(s_i))^2 - (s_{i+1} - s_i))^2].$$

There is a constant C such that $|g(x)| \le C$ for all x, and $B(s_{i+1}) - B(s_i) \sim N(0, s_{i+1} - s_i)$, which gives

$$\mathbb{E}[((B(s_{i+1}) - B(s_i))^2 - (s_{i+1} - s_i))^2] = (s_{i+1} - s_i)^2 \mathbb{E}((Z^2 - 1)^2),$$

where $Z \sim N(0, 1)$. Since

$$\begin{aligned}\sum_{i=0}^{n-1} (s_{i+1} - s_i)^2 &\le \max_{0 \le i < n} |s_{i+1} - s_i| \sum_{i=0}^{n-1} (s_{i+1} - s_i) \\ &= t \max_{0 \le i \le n-1} |s_{i+1} - s_i|,\end{aligned}$$

this shows that the quantity displayed in (15.6.1) indeed converges to zero as the mesh size goes to zero.

Next, let us drop the assumption that g is bounded. Take any $\epsilon, \delta > 0$. Find M so large that $\mathbb{P}(\max_{0 \le s \le t} |B(s)| > M) < \delta/2$. Define

$$h(x) := \begin{cases} g(x) & \text{if } |x| \le M, \\ g(M) & \text{if } x > M, \\ g(-M) & \text{if } x < -M. \end{cases}$$

Then h is a bounded continuous function, and so

$$\sum_{i=0}^{n} h(B(s_i))(s_{i+1} - s_i) \xrightarrow{P} \int_0^t h(B(s))ds$$

as the mesh size tends to zero. On the other hand, note that

$$\mathbb{P}\Big(\sum_{i=0}^{n-1} h(B(s_i))(s_{i+1}-s_i) \neq \sum_{i=0}^{n-1} g(B(s_i))(s_{i+1}-s_i)\Big)$$
$$\leq \mathbb{P}(\max_{0\leq s\leq t} |B(s)| > M) < \frac{\delta}{2}.$$

Similarly,

$$\mathbb{P}\Big(\int_0^t h(B(s))ds \neq \int_0^t g(B(s))ds\Big) \leq \mathbb{P}(\max_{0\leq s\leq t} |B(s)| > M) < \frac{\delta}{2}.$$

Combining the above observations, it follows that

$$\limsup \mathbb{P}\Big(\Big|\sum_{i=0}^{n-1} g(B(s_i))(s_{i+1}-s_i) - \int_0^t g(B(s))ds\Big| > \epsilon\Big) \leq \delta,$$

where the lim sup is taken over a sequence of partitions with mesh size tending to zero. Since ϵ and δ are arbitrary, this completes the proof of the lemma. □

We also the analogous lemma for first-order differences.

LEMMA 15.6.3. *Let B be standard Brownian motion and $g : \mathbb{R} \to \mathbb{R}$ be a continuous function such that*

$$\int_0^t \mathbb{E}(g(B(s))^2)ds < \infty.$$

Take any $t \geq 0$ and a partition $0 = s_0 < s_1 < \cdots < s_n = t$. Then the sum

$$\sum_{i=0}^{n-1} g(B(s_i))(B(s_{i+1}) - B(s_i))$$

converges in probability to $\int_0^t g(B(s))dB(s)$ as the mesh size $\max_{0\leq i<n} |s_{i+1} - s_i|$ approaches zero.

PROOF. The proof is immediate from the definition of the Itô integral when g is bounded and continuous. When g is not bounded, we apply the same truncation trick as in the proof of Lemma 15.6.2, and then apply Exercise 15.2.3 to get the required upper bound on $\mathbb{P}(\int_0^t h(B(s))dB(s) \neq \int_0^t g(B(s))dB(s))$. □

With the aid of the above lemmas, we are now ready to prove Itô's formula.

PROOF OF THEOREM 15.6.1. First, fix some $t > 0$. Take a partition $0 = s_0 < s_1 < \cdots < s_n = t$. For each $\delta, M > 0$, define

$$\omega(\delta, M) := \sup_{\substack{x,y\in[-M,M],\\ |x-y|\leq\delta}} |f''(x) - f''(y)|.$$

Then by Taylor approximation,

$$\Big|f(B(s_{i+1})) - f(B(s_i)) - f'(B(s_i))(B(s_{i+1}) - B(s_i))$$
$$- \frac{1}{2}f''(B(s_i))(B(s_{i+1}) - B(s_i))^2\Big| \leq \frac{1}{2}\omega(\delta_B, M_B)(B(s_{i+1}) - B(s_i))^2,$$

where $M_B := \max_{0\le s\le t} |B(s)|$ and $\delta_B := \max_{0\le i<n} |B(s_{i+1}) - B(s_i)|$. Summing over i and applying the triangle inequality, we get

$$\begin{aligned}\Big| f(B(t)) - f(B(0)) - \sum_{i=0}^{n-1} f'(B(s_i))(B(s_{i+1}) - B(s_i)) \\ - \frac{1}{2}\sum_{i=0}^{n-1} f''(B(s_i))(B(s_{i+1}) - B(s_i))^2 \Big| \\ \le \frac{1}{2}\omega(\delta_B, M_B)\sum_{i=0}^{n-1}(B(s_{i+1}) - B(s_i))^2.\end{aligned}$$

Now suppose that we take a sequence of partitions such that the mesh size tend to zero. Then M_B remains fixed, but by the continuity of Brownian motion, $\delta_B \to 0$ a.s. Therefore $\omega(\delta_B, M_B) \to 0$ a.s. By Lemma 15.6.2,

$$\sum_{i=0}^{n-1} f''(B(s_i))(B(s_{i+1}) - B(s_i))^2 \xrightarrow{P} \int_0^t f''(B(s))ds$$

and

$$\sum_{i=0}^{n-1}(B(s_{i+1}) - B(s_i))^2 \xrightarrow{P} t.$$

Finally, by Lemma 15.6.3,

$$\sum_{i=0}^{n-1} f'(B(s_i))(B(s_{i+1}) - B(s_i)) \xrightarrow{L^2} \int_0^t f'(B(s))dB(s).$$

Combining these observations, we have

$$f(B(t)) = f(B(0)) + \int_0^t f'(B(s))dB(s) + \frac{1}{2}\int_0^t f''(B(s))ds \quad \text{a.s.}$$

Therefore, with probability one, this equation holds for all rational t. But if we use the continuous martingale for $\int_0^t f'(B(s))dB(s)$, then all terms in the above display are continuous functions of t. Thus, with probability one, the equation holds for all t. □

Itô's formula is commonly used to evaluate stochastic integrals. For example, let us evaluate $\int_0^t B(s)dB(s)$.

EXAMPLE 15.6.4. To put the problem in the framework of Itô's formula, we choose a function f such that $f'(x) = x$. Let's take $f(x) = x^2/2$. Then Itô's formula gives

$$\frac{B(t)^2}{2} - \frac{B(0)^2}{2} = \int_0^t B(s)dB(s) + \frac{1}{2}\int_0^t ds,$$

which gives

$$\int_0^t B(s)dB(s) = \frac{1}{2}(B(t)^2 - t).$$

Theorem 15.6.1 gives the simplest version of Itô's formula. There are many generalizations. One version that suffices for most purposes is the following.

THEOREM 15.6.5. *Let d be a positive integer and let $B = (B_1, \dots, B_d)$ be d-dimensional standard Brownian motions. Let $f(t, x_1, \dots, x_d)$ be a function from $[0, \infty) \times \mathbb{R}^d$ into $\mathbb{R}$*

which is continuously differentiable in t (where the derivative at 0 is the right derivative) and twice continuously differentiable in $(x_1, \ldots, x_d)$. Suppose that for each $t \geq 0$ and each i,

$$\int_0^t \mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(s, B(s))\right)^2\right] ds < \infty.$$

Then with probability one, we have that for all $t \geq 0$,

$$f(t, B(t)) = f(0, B(0)) + \sum_{i=1}^{d} \int_0^t \frac{\partial f}{\partial x_i}(s, B(s)) dB_i(s) + \int_0^t \left(\frac{\partial f}{\partial t}(s, B(s)) + \frac{1}{2}\sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(s, B(s))\right) ds,$$

where the stochastic integrals are interpreted as continuous martingales.

The proof of this theorem is exactly similar to that of Theorem 15.6.1, except that we need the following enhancement of Lemma 15.6.2.

LEMMA 15.6.6. *Let B be d-dimensional standard Brownian motions, and $g : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}$ be a continuous function. Take any $t \geq 0$ and a partition $0 = s_0 < s_1 < \cdots < s_n = t$. Then for any $1 \leq j \neq k \leq d$, as the mesh size $\max_{0 \leq i < n} |s_{i+1} - s_i|$ approaches zero,*

$$\sum_{i=0}^{n-1} g(s_i, B(s_i))(B_j(s_{i+1}) - B_j(s_i))^2 \xrightarrow{P} \int_0^t g(s, B(s)) ds$$

and

$$\sum_{i=0}^{n-1} g(s_i, B(s_i))(B_j(s_{i+1}) - B_j(s_i))(B_k(s_{i+1}) - B_k(s_i)) \xrightarrow{P} 0.$$

PROOF. The proof goes just as for Lemma 15.6.2, by first assuming that g is bounded and showing convergence in L^2, and then dropping the boundedness assumption using the same trick as in the proof of Lemma 15.6.2. □

PROOF OF THEOREM 15.6.5. The proof is an easy generalization of the proof of Theorem 15.6.1. As in that proof, we use second order Taylor expansion of f in the x coordinates and first order expansion in t, and then use Lemma 15.6.6 to calculate the limits of the second order terms. The mixed second order derivatives do not appear in the limit because of the second limit in Lemma 15.6.6. □

15.7. Stochastic differential equations

Let B be standard (one-dimensional) Brownian motion. Let $Y(t) := e^{B(t)}$. Then by Itô's formula, we have that for any t,

$$\begin{aligned} Y(t) - Y(0) &= \int_0^t e^{B(s)} dB(s) + \frac{1}{2}\int_0^t e^{B(s)} ds \\ &= \int_0^t Y(s) dB(s) + \frac{1}{2}\int_0^t Y(s) ds. \end{aligned}$$

Therefore the process Y satisfies a 'stochastic integral equation'. The related differential equation would be

$$\frac{dY(t)}{dt} = \frac{1}{2}Y(t) + Y(t)\frac{dB(t)}{dt}.$$

However, Brownian motion is not differentiable. So instead, we write

$$dY(t) = \frac{1}{2}Y(t)dt + Y(t)dB(t). \tag{15.7.1}$$

This is known as a 'stochastic differential equation' (s.d.e.). The equation is just shorthand for writing the integral equation displayed above. It should not be interpreted as a differential equation. (Note that we wrote dt before $dB(t)$. Although dt comes after $dB(t)$ in Itô's formula, it is customary to place the dt term before the $dB(t)$ term when writing s.d.e.'s.)

Now consider the following generalization of the stochastic differential equation (15.7.1):

$$dY(t) = \mu Y(t)dt + \sigma Y(t)dB(t), \tag{15.7.2}$$

with $Y(0) = c$, where $c, \mu \in \mathbb{R}$ and $\sigma > 0$ are some given constants. To solve the equation, let us try to guess a solution, and then verify that it works. Suppose that there is some solution of the form $X(t) = f(t, B(t))$ for some smooth function $f(t, x)$. Then by Itô's formula,

$$dX(t) = \left(\frac{\partial f}{\partial t}(t, B(t)) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t, B(t))\right)dt + \frac{\partial f}{\partial x}(t, B(t))dB(t).$$

Thus, we need f to satisfy

$$\frac{\partial f}{\partial x} = \sigma f, \quad \frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial x^2} = \mu f.$$

From the first equation, we see that f must be of the form $A(t)e^{\sigma x}$ for some function A. Plugging this into the second equation, we get

$$A'(t) + \frac{1}{2}\sigma^2 A(t) = \mu A(t).$$

Solving this, we get $A(t) = A(0)e^{\mu t - \frac{1}{2}\sigma^2 t}$. Thus,

$$f(t, x) = A(0)e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma x}.$$

The condition $X(0) = c$ implies that $A(0) = c$. Therefore, our guess for the solution to (15.7.2) is

$$X(t) = ce^{(\mu - \frac{1}{2}\sigma^2)t + \sigma B(t)}. \tag{15.7.3}$$

It is easy to check that this is a continuous adapted process that satisfies (15.2.1). Therefore, we can apply Itô's formula and check that it is indeed a solution of the s.d.e. (15.7.2) with initial condition $X(0) = c$.

EXERCISE 15.7.1. Find a solution for the stochastic differential equation

$$dX(t) = -\frac{1}{2}X(t)dt + \sqrt{1 - X(t)^2}dB(t)$$

with initial condition $X(0) = 0$.

So we now know how to use Itô's formula to produce solutions of stochastic differential equations. But are these solutions unique? The following theorem gives a sufficient condition for uniqueness within the class of adapted continuous solutions. It shows, for example, that the solution (15.7.3) of the equation (15.7.2) is unique. But it has its limitations; for example, it does not apply to Exercise 15.7.1.

THEOREM 15.7.2. *Let B be standard Brownian motion. Let $a(t,x)$ and $b(t,x)$ be continuous functions from $[0,\infty)\times\mathbb{R}$ into $\mathbb{R}$ satisfying the local Lipschitz condition that for any n, there is a number L_n such that for all t and all $x,y\in[-n,n]$,*

$$|a(t,x)-a(t,y)|\le L_n|x-y|,\quad |b(t,x)-b(t,y)|\le L_n|x-y|.$$

Let c be a real number. Consider the s.d.e.

$$dX(t)=a(t,X(t))dt+b(t,X(t))dB(t)$$

with initial condition $X(0)=c$. Let X and Y be continuous adapted processes solving the above equation with the given initial condition. Then with probability one, $X(t)=Y(t)$ for all t.

PROOF. Fix some $n\ge 1$. Let

$$T_n=\inf\{t\ge 0:|X(t)|>n \text{ or } |Y(t)|>n\}.$$

Since X and Y are continuous adapted process, it is not difficult to see that T_n is a stopping time for the Brownian filtration. Let $\widetilde{X}(t):=X(t\wedge T_n)$ and $\widetilde{Y}(t):=Y(t\wedge T_n)$. Since X and Y are solutions of the given s.d.e., we have

$$\begin{aligned}\widetilde{X}(t)&=X(t\wedge T_n)\\&=c+\int_0^{t\wedge T_n}a(s,X(s))ds+\int_0^{t\wedge T_n}b(s,X(s))dB(s),\end{aligned}$$

and a similar identity holds for $\widetilde{Y}(t)$. Therefore by the Cauchy–Schwarz inequality and Corollary 15.4.3,

$$\begin{aligned}\mathbb{E}[(\widetilde{X}(t)-\widetilde{Y}(t))^2]&\le 2\mathbb{E}\left[\left(\int_0^{t\wedge T_n}(a(s,X(s))-a(s,Y(s)))ds\right)^2\right]\\&\quad+2\mathbb{E}\left[\left(\int_0^{t\wedge T_n}(b(s,X(s))-b(s,Y(s)))dB(s)\right)^2\right]\\&\le 2t\mathbb{E}\left[\int_0^{t\wedge T_n}(a(s,X(s))-a(s,Y(s)))^2ds\right]\\&\quad+2\mathbb{E}\left[\int_0^{t\wedge T_n}(b(s,X(s))-b(s,Y(s)))^2ds\right].\end{aligned}$$

By the local Lipschitz condition on a and b, note that $|a(s,X(s))-a(s,Y(s))|$ and $|b(s,X(s))-b(s,Y(s))|$ are both bounded by $L_n|X(s)-Y(s)|$ when $s\le t\wedge T_n$. But if $s\le t\wedge T_n$, then $s=s\wedge T_n$. Thus, putting $K:=2L_n^2t+2L_n^2$, we have

$$\begin{aligned}\mathbb{E}[(\widetilde{X}(t)-\widetilde{Y}(t))^2]&\le K\mathbb{E}\left[\int_0^{t\wedge T_n}(X(s)-Y(s))^2ds\right]\\&=K\mathbb{E}\left[\int_0^{t\wedge T_n}(X(s\wedge T_n)-Y(s\wedge T_n))^2ds\right]\\&\le K\mathbb{E}\left[\int_0^{t}(X(s\wedge T_n)-Y(s\wedge T_n))^2ds\right].\end{aligned}$$

Thus, if we define $f(t) := \mathbb{E}[(\widetilde{X}(t) - \widetilde{Y}(t))^2]$, then f satisfies the recursive inequality

$$f(t) \leq K \int_0^t f(s)ds.$$

Moreover, since K is an increasing function of t, this shows that for all $s \leq t$,

$$f(s) \leq K \int_0^t f(u)du.$$

Also, notice that $f(s) \leq 4n^2$ for all s since $|\widetilde{X}(t) - \widetilde{Y}(t)| \leq 2n$. Therefore we can iterate the recursive inequality and get that for any $j \geq 1$,

$$\begin{aligned} f(t) &\leq K^j \int_0^t \int_0^{s_1} \int_0^{s_2} \cdots \int_0^{s_{j-1}} f(s_j) ds_j ds_{j-1} \cdots ds_1 \\ &\leq K^j \int_0^t \int_0^{s_1} \int_0^{s_2} \cdots \int_0^{s_{j-1}} 4n^2 ds_j ds_{j-1} \cdots ds_1 \\ &\leq \frac{4n^2 K^j}{j!}. \end{aligned}$$

Letting $j \to \infty$, we get $f(t) = 0$. Thus, $X(t \wedge T_n) = Y(t \wedge T_n)$ a.s. Since X and Y are continuous processes, it follows that $T_n \to \infty$ a.s. as $n \to \infty$. This shows that $X(t) = Y(t)$ a.s. Since this holds for any t, the continuities of X and Y imply that with probability one, $X(t) = Y(t)$ for all t. □

EXERCISE 15.7.3. Solve the stochastic differential equation

$$dX(t) = -\frac{1}{1+t}dt + \frac{1}{1+t}dB(t)$$

with initial condition $X(0) = 0$, and show that the solution is unique in the class of adapted continuous solutions.

Theorem 15.7.2 gives a sufficient condition for the uniqueness of a solution. The condition is mild enough to be widely applicable. But what about existence? In the special case of equation (15.7.2), we could produce an explicit solution, but that requires good luck. The following theorem gives a sufficient condition for the existence of a solution adapted to the Brownian filtration (such solutions are known as 'strong solutions'). The condition is similar to that of Theorem 15.7.2, but with the important difference that now we require a and b to be globally Lipschitz. This is a strong and quite restrictive assumption, but not so different than similar conditions for the existence of solutions of ordinary differential equations.

THEOREM 15.7.4. *Let B be standard Brownian motion. Let $a(t,x)$ and $b(t,x)$ be continuous functions from $[0,\infty) \times \mathbb{R}$ into $\mathbb{R}$ satisfying the global Lipschitz conditions*

$$|a(t,x) - a(t,y)| \leq L|x-y|, \quad |b(t,x) - b(t,y)| \leq L|x-y|$$

for all x, y and t, where L is a constant. Let c be a real number. Then there is a continuous adapted process $\{X(t)\}_{t\geq 0}$ that solves the s.d.e.

$$dX(t) = a(t, X(t))dt + b(t, X(t))dB(t),$$

with initial condition $X(0) = c$.

PROOF. We will construct a solution by Picard iteration. First, let $X_0(t) := c$ for all t. Iteratively, define

$$X_{n+1}(t) := c + \int_0^t a(s, X_n(s))ds + \int_0^t b(s, X_n(s))dB(s),$$

where the stochastic integral on the right should be interpreted as a continuous martingale. It is easy to show by induction, using the conditions on a and b, that each X_n is a continuous adapted process satisfying (15.2.1). This shows that the stochastic integral in the above display is well-defined.

For any $t \geq 0$ and $n \geq 1$, define

$$\Delta_n(t) := \mathbb{E}\|X_{n+1} - X_n\|^2_{[0,t]},$$

where the term inside the expectation on the right is the supremum norm of the process $X_{n+1} - X_n$ on the interval $[0, t]$. Also define

$$U_{n+1}(t) := \int_0^t a(s, X_n(s))ds, \quad V_{n+1}(t) := \int_0^t b(s, X_n(s))dB(s).$$

Now fix some $n \geq 1$ and $0 \leq s \leq t$. The Lipschitz condition on a gives us, for any $s' \leq s$,

$$\begin{aligned}
(U_{n+1}(s') - U_n(s'))^2 &\leq \left(\int_0^{s'} (a(u, X_n(u)) - a(u, X_{n-1}(u))du\right)^2 \\
&\leq s' \int_0^{s'} (a(u, X_n(u)) - a(u, X_{n-1}(u)))^2 du \\
&\leq L^2 s' \int_0^{s'} (X_n(u) - X_{n-1}(u))^2 du \\
&\leq L^2 t \int_0^s (X_n(u) - X_{n-1}(u))^2 du.
\end{aligned}$$

Thus,

$$\|U_{n+1} - U_n\|^2_{[0,s]} \leq L^2 t \int_0^s (X_n(u) - X_{n-1}(u))^2 du. \tag{15.7.4}$$

Next, note that by the Itô isometry and the Lipschitz condition on b,

$$\begin{aligned}
\mathbb{E}[(V_{n+1}(s) - V_n(s))^2] &= \int_0^s \mathbb{E}[(b(u, X_n(u)) - b(u, X_{n-1}(u)))^2]du \\
&\leq L^2 \int_0^s \mathbb{E}[(X_n(u) - X_{n-1}(u))^2]du.
\end{aligned}$$

But $V_{n+1} - V_n$ is a continuous martingale. So by Doob's L^2 inequality for continuous martingales (Theorem 14.2.1),

$$\begin{aligned}
\mathbb{E}\|V_{n+1} - V_n\|^2_{[0,s]} &\leq 4\mathbb{E}[(V_{n+1}(s) - V_n(s))^2] \\
&\leq 4L^2 \int_0^s \mathbb{E}[(X_n(u) - X_{n-1}(u))^2]du.
\end{aligned}$$

Combining this with (15.7.4) and using the inequality $(x+y)^2 \le 2x^2 + 2y^2$, we get

$$\begin{aligned}\Delta_n(s) &= \mathbb{E}\|X_{n+1} - X_n\|^2_{[0,s]} \\ &\le 2\mathbb{E}\|U_{n+1} - U_n\|^2_{[0,s]} + 2\mathbb{E}\|V_{n+1} - V_n\|^2_{[0,s]} \\ &\le K\int_0^s \mathbb{E}[(X_n(u) - X_{n-1}(u))^2]du \\ &\le K\int_0^s \Delta_{n-1}(u)du,\end{aligned}$$

where $K := 2L^2t + 8L^2$. Let $C := \Delta_0(t)$. We claim that

$$\Delta_n(s) \le \frac{C(Ks)^n}{n!} \tag{15.7.5}$$

for each $n \ge 1$ and $0 \le s \le t$. This is easily shown by induction. Suppose that this is true for $n-1$. Then the recursive inequality shows that for any $s \le t$,

$$\begin{aligned}\Delta_n(s) &\le K\int_0^s \Delta_{n-1}(u)du \\ &\le K\int_0^s \frac{C(Ku)^{n-1}}{(n-1)!}du \\ &= \frac{C(Ks)^n}{n!}.\end{aligned}$$

Thus, we get

$$\begin{aligned}\sum_{n=1}^\infty \mathbb{E}\|X_{n+1} - X_n\|_{[0,t]} &\le \sum_{n=1}^\infty \sqrt{\Delta_n(t)} \\ &\le \sum_{n=1}^\infty \sqrt{\frac{C(Kt)^n}{n!}} < \infty.\end{aligned}$$

By the monotone convergence theorem, this shows that

$$\sum_{n=1}^\infty \|X_{n+1} - X_n\|_{[0,t]} < \infty \quad \text{a.s.}$$

This, in turn, shows that with probability one, the sequence $\{X_n\}_{n\ge1}$ is Cauchy in $C[0,t]$ under the supremum norm. Therefore with probability one, $\{X_n\}_{n\ge1}$ is Cauchy in $C[0,M]$ for every integer M. Consequently, $\{X_n\}_{n\ge1}$ is Cauchy in $C[0,\infty)$ with probability one. Let X denote the limit. Clearly, X is a continuous adapted process.

Now, for any given t and n,

$$\int_0^t \mathbb{E}[(X_{n+1}(s) - X_n(s))^2]ds \le \int_0^t \Delta_n(s)ds.$$

Using this, and the bound (15.7.5) on $\Delta_n(s)$, it is easy to see that $\{X_n\}_{n\ge1}$ is a Cauchy sequence in $L^2([0,t]\times\Omega)$. By Fatou's lemma, the limit must necessarily be equal to X. This shows that X satisfies the square-integrability condition (15.2.1) for every t. Thus, by the Lipschitz properties of a and b, the following processes are well-defined, continuous and adapted:

$$U(t) := \int_0^t a(s, X(s))ds, \quad V(t) := \int_0^t b(s, X(s))dB(s).$$

To show that X satisfies the required s.d.e., we need to prove that with probability one, $X(t) = c + U(t) + V(t)$ for all t. Since X, U and V are continuous processes, it suffices to show that this holds almost surely for any given t. So let us fix some t. Since $X_n(t) = c + U_n(t) + V_n(t)$ for each n, and $X_n(t) \to X(t)$ a.s. as $n \to \infty$, it suffices to show that $U_n(t) \to U(t)$ and $V_n(t) \to V(t)$ in L^2 (because then there will exist subsequences converging a.s.). To prove this, first observe that

$$\begin{aligned}\mathbb{E}[(U(t) - U_n(t))^2] &\le t \int_0^t \mathbb{E}[(a(s, X(s)) - a(s, X_{n-1}(s)))^2]ds \\ &\le L^2 t \|X - X_{n-1}\|^2_{L^2([0,t]\times\Omega)}.\end{aligned}$$

But we have already seen above that $X_n \to X$ in $L^2([0,t] \times \Omega)$. Thus, $U_n(t) \to U(t)$ in L^2. Next, by the Itô isometry,

$$\begin{aligned}\mathbb{E}[(V(t) - V_n(t))^2] &= \int_0^t \mathbb{E}[(b(s, X(s)) - b(s, X_{n-1}(s)))^2]ds \\ &\le L^2 \|X - X_{n-1}\|^2_{L^2([0,t]\times\Omega)}.\end{aligned}$$

Again, this shows that $V_n(t) \to V(t)$ in L^2. Thus, X is indeed a solution to the required s.d.e. with initial condition $X(0) = c$. □

15.8. Chain rule for stochastic calculus

The Itô formula has a generalization that is often useful for solving stochastic differential equations. Suppose that a continuous adapted process $\{X(t)\}_{t\ge 0}$ satisfies the s.d.e.

$$dX(t) = a(t)dt + b(t)dB(t),$$

where $\{a(t)\}_{t\ge 0}$ and $\{b(t)\}_{t\ge 0}$ are continuous adapted processes, with b satisfying (15.2.1) for each t. Let $Y(t) := f(t, X(t))$, where $f(t, x)$ is a twice continuously differentiable function. The following result identifies the s.d.e. satisfied by the process $\{Y(t)\}_{t\ge 0}$.

THEOREM 15.8.1. *Let X and Y be as above. Suppose that for each $t \ge 0$,*

$$\int_0^t \mathbb{E}\left(\frac{\partial f}{\partial x}(s, X(s))^2 b(s)^2\right) ds < \infty.$$

Then

$$dY(t) = \frac{\partial f}{\partial t}(t, X(t))dt + \frac{\partial f}{\partial x}(t, X(t))dX(t) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t, X(t))(dX(t))^2,$$

where $dX(t) = a(t)dt + b(t)dB(t)$ and $(dX(t))^2$ stands for $b(t)^2 dt$.

PROOF. The proof is almost exactly the same as for Theorem 15.6.1, except that we have to substitute Brownian motion with the process X. A sketch of the proof goes as follows. The details are left to reader. To deal with potential unboundedness of X, we start by 'localizing' X as follows. Choose some large number M, and define

$$T := \inf\{t : |X(t)| > M \text{ or } |a(t)| > M \text{ or } |b(t)| > M\}.$$

Since X, a and b are adapted continuous processes, T is a stopping time. Define the localized processes $\widetilde{X}(t) := X(t \wedge T)$ and $\widetilde{Y}(t) := Y(t \wedge T) = f(t \wedge T, \widetilde{X}(t))$. We will show

that with probability one, for any $t \geq 0$,

$$\widetilde{Y}(t) - \widetilde{Y}(0) = \int_0^{t\wedge T} \frac{\partial f}{\partial x}(s, X(s))b(s)dB(s) + \int_0^{t\wedge T} \left(\frac{\partial f}{\partial t}(s, X(s)) + \frac{\partial f}{\partial x}(s, X(s))a(s) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(s, X(s))b(s)^2 \right) ds. \tag{15.8.1}$$

Suppose that we have shown this. Then let $M \to \infty$. By the continuity of X, a, and b, the stopping time T must tend to infinity. Therefore the left side approaches $Y(t) - Y(0)$ and the right side approaches the sum of the same two integrals, but integrated from 0 to t. This proves the theorem, provided that we can establish equation (15.8.1) for the localized process.

Taking any $t \geq 0$ and $0 = s_0 < s_1 < \cdots < s_n = t$, observe that

$$\widetilde{X}(s_{i+1}) - \widetilde{X}(s_i) = \int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)du + \int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u). \tag{15.8.2}$$

So, for any function $g(t, x)$,

$$\begin{aligned} &\sum_{i=0}^{n-1} g(s_i \wedge T, \widetilde{X}(s_i))(\widetilde{X}(s_{i+1}) - \widetilde{X}(s_i)) \\ &= \sum_{i=0}^{n-1} \left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} g(s_i \wedge T, \widetilde{X}(s_i))a(u)du \right. \\ &\qquad \left. + \int_{s_i\wedge T}^{s_{i+1}\wedge T} g(s_i \wedge T, \widetilde{X}(s_i))b(u)dB(u) \right). \end{aligned}$$

If g is continuous, then using the dominated convergence theorem and Corollary 15.4.3, it is not hard to show that the above random variable converges in probability to

$$\int_0^{t\wedge T} g(s, X(s))a(s)ds + \int_0^{t\wedge T} g(s, X(s))b(s)dB(s)$$

as the mesh size $\max_{0\leq i<n}(s_{i+1} - s_i)$ tends to zero. Similarly,

$$\sum_{i=0}^{n-1} g(s_i \wedge T, \widetilde{X}(s_i))(s_{i+1} \wedge T - s_i \wedge T) \xrightarrow{P} \int_0^{t\wedge T} g(s, X(s))ds.$$

This takes care of the first-order terms in (15.8.1). For the second-order term, we proceed as follows. First, note that

$$\begin{aligned} &(\widetilde{X}(s_{i+1}) - \widetilde{X}(s_i))^2 \\ &= \left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)du \right)^2 + \left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u) \right)^2 \\ &\quad + 2\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)du \right)\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u) \right). \end{aligned} \tag{15.8.3}$$

By the Cauchy–Schwarz inequality,

$$\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)du\right)^2 \le (s_{i+1}-s_i)\int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)^2du.$$

Thus,

$$\begin{aligned}\sum_{i=0}^{n-1}\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} a(u)du\right)^2 &\le \max_{0\le i\le n-1}(s_{i+1}-s_i)\int_0^{t\wedge T} a(u)^2du\\ &\le M^2t\max_{0\le i\le n-1}(s_{i+1}-s_i),\end{aligned}$$

which tends to zero as the mesh size goes to zero. Similarly, using the Cauchy–Schwarz inequality and Corollary 15.4.3, we get an upper bound for L^1 norm of the third term on the right side of (15.8.3), which also tends to zero as the mesh size goes to zero. Combining, we get

$$\begin{aligned}\sum_{i=0}^{n-1} g(s_i\wedge T,\widetilde{X}(s_i))\Bigg[&(\widetilde{X}(s_{i+1})-\widetilde{X}(s_i))^2\\ &-\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^2\Bigg]\xrightarrow{P}0\end{aligned}\tag{15.8.4}$$

as the mesh size goes to zero. Now, by Corollary 15.4.3,

$$\mathbb{E}\left[\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^2-\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)^2du\,\middle|\,\mathcal{F}^+_{s_i\wedge T}\right]=0.$$

Note that

$$\begin{aligned}&\sum_{i=0}^{n-1}\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^4\\ &\le \max_{0\le i<n}\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^2\sum_{i=0}^{n-1}\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^2,\end{aligned}$$

which tends to zero in probability as the mesh size goes to zero, because the expected value of the sum remains bounded (which implies that it forms a tight family) and the maximum goes to zero almost surely by the continuity of the stochastic integral. Similarly,

$$\sum_{i=0}^{n-1}\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)^2du\right)^2\xrightarrow{P}0.$$

With all this information at our disposal, we can now proceed as in the proof of Lemma 15.6.2 to show that

$$\sum_{i=0}^{n-1} g(s_i\wedge T,\widetilde{X}(s_i))\left[\left(\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)dB(u)\right)^2-\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)^2du\right]\xrightarrow{P}0$$

as the mesh size goes to zero. Combined with (15.8.4), this gives

$$\sum_{i=0}^{n-1} g(s_i\wedge T,\widetilde{X}(s_i))\left[(\widetilde{X}(s_{i+1})-\widetilde{X}(s_i))^2-\int_{s_i\wedge T}^{s_{i+1}\wedge T} b(u)^2du\right]\xrightarrow{P}0.$$

It is now easy to complete the argument. □

EXERCISE 15.8.2. Write down the details for all the steps in the proof of Theorem 15.8.1.

A simple way to remember the identity $(dX(t))^2 = b(t)^2 dt$ is to memorize the following rules of thumb:

$$(dt)^2 = 0, \quad dtdB(t) = 0, \quad (dB(t))^2 = dt.$$

Using these rules and the identity $dX(t) = a(t)dt + b(t)dB(t)$, it is easy to 'derive' $(dX(t))^2 = b(t)^2 dt$. Such rules of thumb are particularly useful if we have a process adapted to multidimensional Brownian motion. In this case we have an s.d.e. like

$$dX(t) = a(t)dt + \sum_{i=1}^{d} b_i(t)dB_i(t).$$

As above, we have the for each i,

$$(dt)^2 = 0, \quad dtdB_i(t) = 0, \quad (dB_i(t))^2 = dt.$$

Additionally, we also have $dB_i(t)dB_j(t) = 0$ for $i \neq j$. These rules give

$$(dX(t))^2 = \sum_{i=1}^{d} b_i(t)^2 dt.$$

EXERCISE 15.8.3. Formulate and prove a version of Theorem 15.8.1 for processes adapted to multidimensional Brownian motion.

15.9. The Ornstein–Uhlenbeck process

Consider the stochastic differential equation

$$dX(t) = -\mu X(t)dt + \sigma dB(t) \tag{15.9.1}$$

with initial condition $X(0) = c \in \mathbb{R}$, where μ and σ are strictly positive constants. This looks almost the same as the s.d.e. (15.7.2), except that we do not have an $X(t)$ in front of $dB(t)$, and constant in front of $X(t)dt$ is strictly negative. A simple verification shows that in this problem we cannot find a solution of the form $f(t, B(t))$, so we have to implement a different plan. Note that by Theorems 15.7.4 and 15.7.2, an adapted continuous solution exists and is unique. So we only have to guess the solution and verify that the guess is valid. Taking cue from the deterministic case $\sigma = 0$, we define $Y(t) := e^{\mu t}X(t)$. By Theorem 15.8.1,

$$\begin{aligned} dY(t) &= \mu e^{\mu t}X(t)dt + e^{\mu t}dX(t) \\ &= \sigma e^{\mu t}dB(t). \end{aligned}$$

The initial condition for Y is $Y(0) = c$. Thus,

$$Y(t) = c + \sigma \int_0^t e^{\mu s}dB(s),$$

and therefore,

$$X(t) = ce^{-\mu t} + \sigma \int_0^t e^{\mu(s-t)}dB(s).$$

Since Y is a continuous adapted process, so is X. It is simple to verify that this is indeed a solution of (15.9.1). The process $X(t)$ is known as an *Ornstein–Uhlenbeck process*. Clearly, $\mathbb{E}(X(t)) = ce^{-\mu t}$. By Exercise 15.3.2,

$$\begin{aligned}\operatorname{Cov}(X(s), X(t)) &= \sigma^2 e^{-\mu(s+t)} \int_0^{s\wedge t} e^{2\mu u} du \\ &= \frac{\sigma^2}{2\mu} e^{-\mu(s+t)}(e^{2\mu(s\wedge t)} - 1) \\ &= \frac{\sigma^2}{2\mu}(e^{-\mu|s-t|} - e^{-\mu(s+t)}).\end{aligned}$$

The formulas for the mean and the covariance completely identify the distribution of the Ornstein–Uhlenbeck process. Note that as $t \to \infty$, $X(t)$ converges in distribution to $N(0, \sigma^2/2\mu)$. The special case $\mu = 1$, $\sigma = \sqrt{2}$ is sometimes called the standard Ornstein–Uhlenbeck process. In this case, $X(t) \to N(0,1)$ in distribution as $t \to \infty$.

EXERCISE 15.9.1. Let B be standard Brownian motion. Take three numbers $c \in \mathbb{R}$, $\mu > 0$ and $\sigma > 0$, and define $X(t) := ce^{-\mu t} + \sigma e^{-\mu t} B(e^{2\mu t} - 1)$. Show that X has the same distribution as the Ornstein–Uhlenbeck process defined above.

EXERCISE 15.9.2. Let X be an Ornstein–Uhlenbeck process. Find a function $f(t)$ such that $\limsup_{t\to\infty} X(t)/f(t)$ and $\liminf_{t\to\infty} X(t)/f(t)$ are finite constants almost surely.

15.10. Lévy's characterization of Brownian motion

Let B be standard Brownian motion. We have seen that $B(t)$ and $B(t)^2 - t$ are continuous martingales with respect to the right-continuous filtration generated by B. It turns out that Brownian motion is the only continuous process having these two properties. This is known as Lévy's characterization of Brownian motion.

THEOREM 15.10.1. *Let $\{X(t)\}_{t\geq 0}$ be a continuous stochastic process adapted to a right-continuous filtration $\{\mathcal{F}_t\}_{t\geq 0}$, with $X(0) = 0$. Suppose that $X(t)$ and $X(t)^2 - t$ are martingales adapted to this filtration. Then X is standard Brownian motion.*

PROOF. It is not hard to see that it's sufficient to prove that for any $0 \leq s \leq t$, $X(t) - X(s) \sim N(0, t-s)$ and $X(t) - X(s)$ is independent of $\mathcal{F}_s$. By Exercise 12.6.6, we have to show that for any $A \in \mathcal{F}_s$ and any $\theta \in \mathbb{R}$,

$$\mathbb{E}(e^{i\theta(X(t)-X(s))}; A) = \mathbb{P}(A)e^{-\theta^2(t-s)/2}. \tag{15.10.1}$$

Fix some $s \geq 0$ and a positive integer M, and let

$$T := \inf\{t \geq s : |X(t)| > M\}.$$

By the continuity of X, T is a stopping time for the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, and T is always $\geq s$. For each $t \geq 0$, let

$$Y(t) := X(t \wedge T), \quad Z(t) := X(t \wedge T)^2 - (t \wedge T).$$

Let $\mathcal{G}_t := \mathcal{F}_{t\wedge T}$. Then by the results of Section 13.11 (which hold for any right-continuous filtration), $\{\mathcal{G}_t\}_{t\geq 0}$ is also a right-continuous filtration. Moreover, by the optional stopping

theorem for continuous martingales, $Y(t)$ and $Z(t)$ are martingales adapted to $\mathcal{G}_t$. A consequence of the martingale properties of $Y(t)$ and $Z(t)$ is that for any $0 \le u \le t$,

$$\begin{aligned}\mathbb{E}((Y(t)-Y(u))^2|\mathcal{G}_u) &= \mathbb{E}(Y(t)^2 - 2Y(t)Y(u) + Y(u)^2|\mathcal{G}_u)\\ &= \mathbb{E}(Y(t)^2|\mathcal{G}_u) - Y(u)^2\\ &= \mathbb{E}(Z(t) + t\wedge T|\mathcal{G}_u) - Y(u)^2\\ &= Z(u) + \mathbb{E}(t\wedge T|\mathcal{G}_u) - Y(u)^2\\ &= \mathbb{E}(t\wedge T - u\wedge T|\mathcal{G}_u). \end{aligned} \tag{15.10.2}$$

Now fix some $t \ge s$, $A \in \mathcal{F}_s$, and $\theta \in \mathbb{R}$. Let $f(x) := e^{\mathrm{i}\theta x}$. Take a partition $s = s_0 < s_1 < \cdots < s_n = t$ of the interval $[s,t]$. By the martingale property of Y and the fact that $\mathcal{F}_s \subseteq \mathcal{G}_u$ for any $u \ge s$,

$$\sum_{i=0}^{n-1} \mathbb{E}[(Y(s_{i+1}) - Y(s_i))f'(Y(s_i) - Y(s)); A] = 0. \tag{15.10.3}$$

By (15.10.2),

$$\begin{aligned}&\sum_{i=0}^{n-1} \mathbb{E}[(Y(s_{i+1}) - Y(s_i))^2 f''(Y(s_i) - Y(s)); A]\\ &= \sum_{i=0}^{n-1} \mathbb{E}[(s_{i+1}\wedge T - s_i \wedge T) f''(Y(s_i) - Y(s)); A]\\ &= \mathbb{E}\left[\sum_{i=0}^{n-1} (s_{i+1}\wedge T - s_i \wedge T) f''(Y(s_i) - Y(s)); A\right].\end{aligned}$$

As the mesh size goes to zero, the sum the expectation approaches

$$\int_s^{t\wedge T} f''(Y(u) - Y(s))du.$$

Also, by the boundedness of f'', the sums are uniformly bounded by a constant. Therefore by the dominated convergence theorem,

$$\begin{aligned}&\sum_{i=0}^{n-1} \mathbb{E}[(Y(s_{i+1}) - Y(s_i))^2 f''(Y(s_i) - Y(s)); A]\\ &\to \mathbb{E}\left(\int_s^{t\wedge T} f''(Y(u) - Y(s))du; A\right)\end{aligned} \tag{15.10.4}$$

as the mesh size goes to zero.

Next, let $\delta := \max_{0\le i\le n-1} |Y(s_{i+1}) - Y(s_i)|$, and let

$$\omega(\delta) := \max_{x,y\in\mathbb{R}, |x-y|\le\delta} |f''(x) - f''(y)|.$$

Since f''' is uniformly bounded and Y is a continuous process, $\omega(\delta) \to 0$ as the mesh size goes to zero. Since f'' is uniformly bounded, $\omega(\delta)$ is bounded by a constant. Therefore by the dominated convergence theorem, as the mesh size goes to zero,

$$\mathbb{E}(\omega(\delta)^2) \to 0. \tag{15.10.5}$$

Note that by (15.10.2),

$$\mathbb{E}[(Y(s_{i+1}) - Y(s_i))^2(Y(s_{j+1}) - Y(s_j))^2] \le (s_{i+1} - s_i)(s_{j+1} - s_j)$$

when $i \neq j$. Now, if $T = s$, then $Y(t)$ is unchanging beyond time s, and if $T > s$, then $|Y(t)| \le M$ beyond time s. So in either case, we have that for any i, $|Y(s_{i+1}) - Y(s_i)| \le 2M$. This gives

$$\begin{aligned}\mathbb{E}[(Y(s_{i+1}) - Y(s_i))^4] &\le 4M^2\mathbb{E}[(Y(s_{i+1}) - Y(s_i))^2] \\ &\le 4M^2(s_{i+1} - s_i).\end{aligned}$$

Using the last two displays, we get

$$\mathbb{E}\left[\left(\sum_{i=0}^{n-1}(Y(s_{i+1}) - Y(s_i))^2\right)^2\right] \le 4M^2(t - s) + (t - s)^2.$$

In particular, the above expectation remains bounded as the mesh size tends to zero. Combining this information with (15.10.5) and applying the Cauchy–Schwarz inequality, we get that

$$\mathbb{E}\left(\omega(\delta)\sum_{i=0}^{n-1}(Y(s_{i+1}) - Y(s_i))^2\right) \to 0 \tag{15.10.6}$$

as the mesh size tends to zero. Now, by Taylor expansion,

$$\begin{aligned}&\Bigg|f(Y(t) - Y(s)) - f(0) - \sum_{i=0}^{n-1} f'(Y(s_i) - Y(s))(Y(s_{i+1}) - Y(s_i)) \\ &\qquad - \frac{1}{2}\sum_{i=0}^{n-1} f''(Y(s_i) - Y(s))(Y(s_{i+1}) - Y(s_i))^2\Bigg| \\ &\le \frac{1}{2}\omega(\delta)\sum_{i=0}^{n-1}(Y(s_{i+1}) - Y(s_i))^2.\end{aligned}$$

Using (15.10.3), (15.10.4), (15.10.6), and the above bound, we get

$$\begin{aligned}\mathbb{E}(f(Y(t) - Y(s)); A) &= \mathbb{P}(A) + \frac{1}{2}\mathbb{E}\left(\int_s^{t\wedge T} f''(Y(u) - Y(s))du; A\right) \\ &= \mathbb{P}(A) + \frac{1}{2}\mathbb{E}\left(\int_s^{t\wedge T} f''(X(u) - X(s))du; A\right),\end{aligned}$$

where the last step holds because $Y(u) = X(u)$ when $u \le T$. Now let $M \to \infty$, so that $T \to \infty$ (because X is a continuous process). By the boundedness of f and f'' and the dominated convergence theorem, we can take the limits inside the expectations in the above identity and get

$$\mathbb{E}(f(X(t) - X(s)); A) = \mathbb{P}(A) + \frac{1}{2}\mathbb{E}\left(\int_s^t f''(X(u) - X(s))du; A\right).$$

Let $\phi(t) := \mathbb{E}(f(X(t) - X(s)); A)$ for $t \ge s$. Since $f''(x) = -\theta^2 f(x)$, the above identity yields the following integral equation for ϕ:

$$\phi(t) = \mathbb{P}(A) - \frac{\theta^2}{2}\int_s^t \phi(u)du.$$

The unique solution of this equation is $\phi(t) = \mathbb{P}(A)e^{-\theta^2(t-s)/2}$. (Uniqueness is easily established, for example, by Picard iteration.) This proves (15.10.1), and hence completes the proof of the theorem. □