Lecture Notes on Advanced Probability

Xi Geng

Abstract

These are lecture notes for the subject Advanced Probability which I taught at University of Melbourne over the past years. They appear to be substantially longer than what typically constitute a one-semester course for the reason that the subject went through a phase transition in 2021. Before that, the syllabus covered weak convergence of probability measures, sequences and sums of independent random variables, the characteristic function, central limit theorems and Stein's method for Gaussian approximations. After that, the syllabus changed to cover product measure spaces, sequences and sums of independent random variables and discrete-time martingales. These notes combine the aforementioned two versions of the subject and form a self-contained integrated text.

As a Melbourne University tradition, the first two chapters are not included in the Advanced Probability syllabus; instead they are taught in the third-year subject Probability for Inference. We have included these two chapters just for self-containedness. In addition, we have also omitted two important topics that are typical for such a subject: Markov chains and infinitely divisible distributions. The former is covered in the third-year undergraduate course Stochastic Modelling and the latter is covered in the graduate course Random Processes.

In many ways, the present notes are largely influenced by [Wil91] and also by [Bil86, Chu01, Dur19] among others. These classics have always been great sources of inspiration and enjoyment for learning modern probability theory.

Contents

1 Probability spaces

Probability theory is primarily concerned with the study of quantitative behaviours of random variables and their distributional properties. Before attempting to investigate these questions, we shall begin with the very first building block of the theory: *probability spaces*.

In this chapter, we give precise mathematical meanings to various concepts one has seen in elementary probability from a measure-theoretic perspective. The main motivating questions of this chapter are described as follows.

(i) *What is a proper mathematical way of describing events?*
(ii) *What are the natural principles of assigning probabilities to events?*
(iii) *How can one construct a "probability function" on all events given the knowledge of probabilities on "simple" events?*

The study of these three questions occupies the first three sections respectively. For one who has no prior exposure to measures on σ-algebras, this chapter may not be as entertaining as those fun combinatorial arguments in elementary probability. Nonetheless, since the development of Kolmogorov's axiomatic approach to probability in the 1930s, measure theory has become the basic language of modern probability that every probabilist speaks and uses nowadays. The essential goal of this and the next chapters is to get acquainted with this language so that one can start working on the real probabilistic problems in a precise mathematical way.

1.1 Set classes, events and Dynkin's π-λ theorem

Before introducing probabilities, one first encounters the notions of *sample spaces* and *events*. In elementary probability, a sample space is the collection of all possible outcomes of a random experiment. As a mathematical object on its own, a sample space is merely a given abstract set. The first concept that requires care is the notion of events. Heuristically, an event should clearly be considered as a subset of the sample space consisting of certain outcomes (because the relation between outcomes and events is the relation of *belongingness*: an outcome ω triggers an event A iff $\omega \in A$). In typical situations when the sample space is infinite, for subtle theoretical reasons it is inappropriate to consider *every* subset of the sample space as a legal event. How to understand events mathematically is the aim of this section. This is the first step to take before assigning probabilities to events, which will be the goal of the next section.

Let Ω be a given fixed non-empty set (the sample space). As subsets of Ω, the key axiom on events is that they should be stable / closed under natural set operations. For instance, if A, B are two legal events, then A^c, $A \cup B$ and $A \cap B$ should all be considered as legal events. In addition, for practical reasons it is necessary to form new events from *infinitely many* given ones (e.g. $\cup_{n=1}^{\infty} A_n$). It turns out that, as the minimal requirement the collection of "legal events" should be closed under any applications of the set operations $(\cdot)^c, \cup, \cap$ for countably many times. This leads to the following basic definition.

Definition 1.1. A collection $\mathcal{F}$ of subsets of Ω is called a *σ-algebra* over Ω, if it satisfies the following three properties:

(i) $\Omega \in \mathcal{F}$;
(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
(iii) if $A_n \in \mathcal{F}$ $(n \geqslant 1)$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Given a σ-algebra $\mathcal{F}$ over Ω, the pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

Members of $\mathcal{F}$ are called $\mathcal{F}$-measurable sets. We interpret them as *events*. In other words, a σ-algebra over Ω defines a family of events.

Example 1.1. There are two obvious σ-algebras over Ω: the *trivial σ-algebra* $\mathcal{F}_0 \triangleq \{\emptyset, \Omega\}$ and the *power set* $\mathcal{P}(\Omega)$ (the collection of all subsets of Ω).

Example 1.2. Let $\Omega = \mathbb{Z}$ (the set of integers). Then

$$\mathcal{F} = \{\emptyset, \Omega, \{\text{even numbers}\}, \{\text{odd numbers}\}\}$$

is a σ-algebra over Ω.

Let $\mathcal{F}$ be a σ-algebra over Ω. Any combinations of finite or countable unions, intersections, complements of events in $\mathcal{F}$ again belong to $\mathcal{F}$. For instance, $\emptyset = \Omega^c \in \mathcal{F}$ as a consequence of Property (i) and (ii). If $A, B \in \mathcal{F}$, then

$$A \cup B = A \cup B \cup B \cup B \cdots \in \mathcal{F}$$

by Property (iii). Given $\{A_n : n \geqslant 1\} \subseteq \mathcal{F}$, one has

$$\bigcap_{n=1}^{\infty} A_n = \big(\bigcup_{n=1}^{\infty} A_n^c\big)^c \in \mathcal{F}.$$

When Ω is an infinite set, σ-algebras over Ω are in general difficult to describe explicitly. Nonetheless, there are several types of set classes that are often easier

to describe and work with. A common idea in probability theory is that, once one is comfortable with properties over a simple set class, one extends them to "the σ-algebra generated by this class", the latter which being the real stuff of interest. These simpler set classes are defined by weakening the stability properties (i)–(iii) to different extents in the definition of σ-algebras.

Definition 1.2. Let $\mathcal{C}$ be a set class on Ω (i.e. a collection of subsets of Ω).

(i) $\mathcal{C}$ is called a π-*system*, if it is closed under finite intersections:

$$A, B \in \mathcal{C} \implies A \cap B \in \mathcal{C};$$

(ii) $\mathcal{C}$ is called a *semiring*, if $\mathcal{C}$ is a π-system containing $\emptyset$ and additionally for any $A, B \in \mathcal{C}$, one can write $A \backslash B$ as a disjoint union of finitely many members of $\mathcal{C}$.
(iii) $\mathcal{C}$ is called an *algebra*, if $\mathcal{C}$ is a π-system containing Ω and additionally one has

$$A \in \mathcal{C} \implies A^c \in \mathcal{C}.$$

(iv) $\mathcal{C}$ is called a λ-*system* (or a *Dynkin system*), if

$$\Omega \in \mathcal{C};\ A, B \in \mathcal{C}, A \subseteq B \implies B \backslash A \in \mathcal{C};\ A_n \in \mathcal{C}, A_n \uparrow A \implies A \in \mathcal{C}.$$

Notation. It is a common convention to use small letters ω, x, y etc. to denote elements in an event, capital letters A, C, E etc. to denote events and calligraphy letters $\mathcal{A}, \mathcal{C}, \mathcal{F}$ etc. to denote set classes.

From Definition 1.2, it is clear that

$$\sigma\text{-algebra} \implies \begin{cases} \text{algebra} \implies \text{semiring} \implies \pi\text{-system}; \\ \lambda\text{-system}, \end{cases}$$

but none of the reverse implication is true in general. π-systems, semirings and algebras are concerned with stability under finitely many steps of set operations. For instance, it is not hard to check that an algebra is closed under finite combinations of taking unions / intersections / complementations. λ-systems and σ-algebras are concerned with stability under countably many steps of set operations.

The real line is an important example to demonstrate the above concepts.

Example 1.3. Let $\Omega = \mathbb{R}$. Define the following set classes

$$\mathcal{C}_1 \triangleq \{(-\infty, b] : b \in \mathbb{R}\},$$
$$\mathcal{C}_2 \triangleq \{(a, b] : -\infty < a \leqslant b < \infty\},$$

respectively. Then $\mathcal{C}_1$ is a π-system and $\mathcal{C}_2$ is a semiring (for instance, $(0,3]\backslash(1,2]$ is the disjoint union of two members $(0,1]$ and $(2,3]$ in $\mathcal{C}_2$). Define $\mathcal{C}_3$ to be the collection of finite disjoint unions of intervals of the form $(-\infty, b]$, $(a,b]$ or (a,∞) with $a \leqslant b$. Then $\mathcal{C}_3$ is an algebra.

In Example 1.3 over $\mathbb{R}$, it is often necessary to work with subsets that are more general than intervals. In particular, we are also interested in subsets formed by countably many steps of set operations on intervals. This leads one to the important notion of generated σ-algebras as well as other types of generated classes.

Definition 1.3. Let $\mathcal{C}$ be a set class over Ω. The *σ-algebra generated by $\mathcal{C}$*, denoted as $\sigma(\mathcal{C})$, is the smallest σ-algebra that contains $\mathcal{C}$. Equivalently, it is the intersection of all σ-algebras over Ω having $\mathcal{C}$ as a subclass:

$$\sigma(\mathcal{C}) = \bigcap_{\substack{\mathcal{F}:\mathcal{F} \text{ is } \sigma\text{-alg.} \\ \mathcal{C}\subseteq\mathcal{F}}} \mathcal{F}.$$

In a similar way, one can define the notions of generated algebras, generated λ-systems etc. The *algebra* (respectively, the *λ-system*) generated by $\mathcal{C}$, denoted as $\mathcal{A}(\mathcal{C})$ (respectively, $\lambda(\mathcal{C})$), is the smallest algebra (respectively, smallest λ-system) containing $\mathcal{C}$ as a subclass.

Describing a generated algebra is easy. We illustrate this in the example of $\mathbb{R}$.

Example 1.4. Using the same notation in Example 1.3, one can show that $\mathcal{C}_3 = \mathcal{A}(\mathcal{C}_1 \cup \mathcal{C}_2)$. Indeed, it is plain to check by definition that $\mathcal{C}_3$ is an algebra containing $\mathcal{C}_1 \cup \mathcal{C}_2$. This implies that $\mathcal{C}_3 \supseteq \mathcal{A}(\mathcal{C}_1 \cup \mathcal{C}_2)$. For the reverse direction, one only needs to observe that $(a,\infty) \in \mathcal{A}(\mathcal{C}_1 \cup \mathcal{C}_2)$. But this is obvious: one has

$$(a,\infty) = (-\infty, a]^c \in \mathcal{A}(\mathcal{C}_1 \cup \mathcal{C}_2)$$

since $(-\infty, a] \in \mathcal{C}_1 \subseteq \mathcal{A}(\mathcal{C}_1 \cup \mathcal{C}_2)$.

Unfortunately, except for some special situations (e.g. $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$), it is usually impossible to describe a generated σ-algebra explicitly (cf. Appendix B for the more serious reader!). It is true though

$$\mathcal{C}_1 \subseteq \mathcal{C}_2 \implies \sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2),$$

which is seen by the fact that $\sigma(\mathcal{C}_2)$ is a σ-algebra containing $\mathcal{C}_1$. In the example of $\mathbb{R}$, one has the following property.

Proposition 1.1. *We continue using the notation $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ in Example 1.3. Let $\mathcal{C}_4$ be the collection of all open subsets of $\mathbb{R}$. Then*

$$\sigma(\mathcal{C}_1) = \sigma(\mathcal{C}_2) = \sigma(\mathcal{C}_3) = \sigma(\mathcal{C}_4).$$

Proof. We argue in the order of $\sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2) \subseteq \sigma(\mathcal{C}_3) \subseteq \sigma(\mathcal{C}_4) \subseteq \sigma(\mathcal{C}_1)$.

(i) $\sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2)$. It is sufficient to show that $\mathcal{C}_1 \subseteq \sigma(\mathcal{C}_2)$. But any set in $\mathcal{C}_1$ has the form $(-\infty, b]$ which can be written as

$$(-\infty, b] = \bigcup_{n=1}^{\infty} (-n, b]. \tag{1.1}$$

Since $(-n, b] \in \mathcal{C}_2$ for each n and $\sigma(\mathcal{C}_2)$ is a σ-algebra, one concludes that the right hand side of (1.1) and thus $(-\infty, b] \in \sigma(\mathcal{C}_2)$. Therefore, $\mathcal{C}_1 \subseteq \sigma(\mathcal{C}_2)$.

(ii) $\sigma(\mathcal{C}_2) \subseteq \sigma(\mathcal{C}_3)$. This is obvious since $\mathcal{C}_2 \subseteq \mathcal{C}_3$.

(iii) $\sigma(\mathcal{C}_3) \subseteq \sigma(\mathcal{C}_4)$. Since $\mathcal{C}_3$ consists of finite disjoint unions of intervals of the form $(-\infty, b]$, $(a, b]$ or (a, ∞), it is sufficient to see that these intervals are all contained in $\sigma(\mathcal{C}_4)$. Let us only check this for $(a, b]$. One writes

$$(a, b] = \bigcap_{n=1}^{\infty} \big(a, b + \frac{1}{n}\big).$$

Since open intervals are open subsets, one knows that the right hand side and thus $(a, b] \in \sigma(\mathcal{C}_4)$.

(iv) $\sigma(\mathcal{C}_4) \subseteq \sigma(\mathcal{C}_1)$. Let U be an open subset of $\mathbb{R}$ and we want to show that $U \in \sigma(\mathcal{C}_1)$. The key observation is that U can be written as a countable union of open intervals. Indeed, given $x \in U$, there exists $n \geqslant 1$ such that $(x-2/n, x+2/n) \subseteq U$. But x can always be approximated by rational numbers. In particular, there exists $r \in U \cap \mathbb{Q}$, such that $|x - r| < 1/n$, which implies that $x \in (r - 1/n, r + 1/n) \subseteq U$. This shows that

$$U = \bigcup_{\substack{r \in U \cap \mathbb{Q}, n \geqslant 1: \\ (r-1/n, r+1/n) \subseteq U}} \big(r - \frac{1}{n}, r + \frac{1}{n}\big).$$

Therefore, U is a countable union of open intervals. It is now enough to show that any open interval $(a, b) \in \sigma(\mathcal{C}_1)$. But this follows from the fact that

$$(a, b) = (-\infty, b) \backslash (-\infty, a] = \big(\bigcup_{n=1}^{\infty} \big(-\infty, b - \frac{1}{n}\big]\big) \backslash (-\infty, a] \in \sigma(\mathcal{C}_1).$$

□

Remark 1.1. In Proposition 1.1, by using essentially the same type of argument, one can show that $\sigma(\mathcal{C}_1)$ also coincides with the σ-algebras generated by open intervals, or closed intervals, or intervals of the form $[a, b)$, respectively.

Definition 1.4. The σ-algebra generated by open subsets of $\mathbb{R}$ is called the *Borel σ-algebra* on $\mathbb{R}$. It is denoted as $\mathcal{B}(\mathbb{R})$. Members of $\mathcal{B}(\mathbb{R})$ are called *Borel measurable sets*.

Let $\mathcal{C}$ be a given set class over Ω. The following type of questions is quite typical. Suppose one knows that certain property holds true for members of $\mathcal{C}$. How can one prove that the same property also holds *for all members of* $\sigma(\mathcal{C})$? The challenge here is that $\sigma(\mathcal{C})$ is in general hard to describe. A powerful approach to address this question (without knowing what $\sigma(\mathcal{C})$ looks like) is the so-called *Dynkin's π-λ theorem*.

Theorem 1.1. *Let $\mathcal{C}$ be a π-system. Then the λ-system generated by $\mathcal{C}$ coincides with the σ-algebra generated by $\mathcal{C}$. In other words, $\lambda(\mathcal{C}) = \sigma(\mathcal{C})$.*

Proof. We first prove a general fact: if a set class $\mathcal{L}$ is a π-system and also a λ-system, then $\mathcal{L}$ is a σ-algebra. To see that, by the definition of a λ-system, one has $\Omega \in \mathcal{L}$, and if $A \in \mathcal{L}$ then $A^c = \Omega \backslash A \in \mathcal{L}$. It remains to show that $\mathcal{L}$ is stable under countable union. Let $A_n \in \mathcal{L}$ and $A \triangleq \cup_n A_n$. For each n we set $B_n \triangleq A_1 \cup \cdots \cup A_n$. Since $B_n = (A_1^c \cap \cdots \cap A_n^c)^c$ and $\mathcal{L}$ is a π-system, one sees that $B_n \in \mathcal{L}$. In addition, since $B_n \uparrow A$ and $\mathcal{L}$ is a λ-system, one concludes that $A \in \mathcal{L}$. Therefore, $\mathcal{L}$ is a σ-algebra.

Now suppose that $\mathcal{C}$ is a given π-system. Since any σ-algebra is a λ-system, one has $\lambda(\mathcal{C}) \subseteq \sigma(\mathcal{C})$. To demonstrate the other direction, it suffices to show that $\lambda(\mathcal{C})$ is a σ-algebra. According to the previous fact, it is enough to show that $\lambda(\mathcal{C})$ is a π-system, i.e.

$$A, B \in \lambda(\mathcal{C}) \implies A \cap B \in \lambda(\mathcal{C}). \tag{1.2}$$

To this end, we introduce an intermediate step: we first prove (1.2) for $A \in \lambda(\mathcal{C})$ and $B \in \mathcal{C}$. Let $B \in \mathcal{C}$ be given fixed and define the set class

$$\mathcal{H}_B \triangleq \{A \subseteq \Omega : A \cap B \in \lambda(\mathcal{C})\}.$$

Since $\mathcal{C}$ is a π-system, one has $\mathcal{C} \subseteq \mathcal{H}_B$. We now check that $\mathcal{H}_B$ is a λ-system:

(L1) Since $\Omega \cap B = B \in \mathcal{C} \subseteq \lambda(\mathcal{C})$ by assumption, one has $\Omega \in \mathcal{H}_B$.
(L2) Let $A_1, A_2 \in \mathcal{H}_B$ with $A_1 \subseteq A_2$. Then $A_1 \cap B, A_2 \cap B \in \lambda(\mathcal{C})$. Since $\lambda(\mathcal{C})$ is a λ-system, one has

$$(A_2 \backslash A_1) \cap B = (A_2 \cap B) \backslash (A_1 \cap B) \in \lambda(\mathcal{C}).$$

Therefore, $A_2 \backslash A_1 \in \mathcal{H}_B$.
(L3) Let $A_n \in \mathcal{H}_B$ and $A_n \uparrow A$. Then $A_n \cap B \in \lambda(\mathcal{C})$ and $A_n \cap B \uparrow A \cap B$. Therefore, $A \cap B \in \lambda(\mathcal{C})$ and thus $A \in \mathcal{H}_B$.

As a consequence, $\mathcal{H}_B$ is a λ-system and one thus has $\lambda(\mathcal{C}) \subseteq \mathcal{H}_B$. In other words, for any $A \in \lambda(\mathcal{C})$ with $B \in \mathcal{C}$, one has $A \cap B \in \lambda(\mathcal{C})$. To prove the general statement (1.2), one can now fix $A \in \lambda(\mathcal{C})$ and argue in the same way to conclude that $A \cap B \in \lambda(\mathcal{C})$ for all $B \in \lambda(\mathcal{C})$. □

We conclude this section by describing the general procedure of applying Dynkin's π-λ theorem. We will see concrete examples when we have richer probabilistic contexts (cf. e.g. Proposition 1.4 in the Section 1.3 below).

Suppose that one wants to prove certain property **P** for all events in the σ-algebra $\sigma(\mathcal{C})$ generated by some π-system $\mathcal{C}$. One can then argue in the following steps:

Step 1. Define $\mathcal{H}$ to be the collection of subsets satisfying property **P**, and show that $\mathcal{C} \subseteq \mathcal{H}$ (i.e. members of $\mathcal{C}$ satisfy the desired property).
Step 2. Show that $\mathcal{H}$ is a λ-system.

Once these two steps are established, it follows that $\lambda(\mathcal{C}) \subseteq \mathcal{H}$ and Dynkin's π-λ theorem yields $\sigma(\mathcal{C}) \subseteq \mathcal{H}$. In other words, all events in $\sigma(\mathcal{C})$ satisfy property **P**.

1.2 Probability measures and their basic properties

After developing the mathematical notion of events, one can now talk about assigning probabilities to them. Essentially, this is a set function $\mathbb{P} : \mathcal{F} \to [0,1]$ $(A \mapsto \mathbb{P}(A))$ defined on a given σ-algebra $\mathcal{F}$ (collection of events) and taking values in $[0,1]$ (probabilities), which should obey some "natural axioms". Such a set function will be called a *probability measure*. At this point, the fact that $\mathbb{P}(A) \leqslant 1$ is not so important yet; it is beneficial to put aside "probability" (until we need it!) and work with the more general notion of measures.

Let $(\Omega, \mathcal{F})$ be a given measurable space. A measure should be a set function μ defined on $\mathcal{F}$, so that whenever one is given a set $A \in \mathcal{F}$, $\mu(A)$ produces a reasonable notion of its "size". In particular, μ should take values in $[0, \infty]$ (A can have infinite size in principle so that one should include ∞ as a possible value for μ). A fundamental property of a measure is *countable additivity*: the size of a countable disjoint union of sets in $\mathcal{F}$ should be the sum of their individual sizes. This is the "natural axiom" in the definition of (probability) measures.

Definition 1.5. Let $(\Omega, \mathcal{F})$ be a measurable space. A set function $\mu : \mathcal{F} \to [0, \infty]$ is called a *measure* on $(\Omega, \mathcal{F})$, if it satisfies $\mu(\emptyset) = 0$ and whenever $\{A_n : n \geqslant 1\}$ is a sequence of disjoint sets in $\mathcal{F}$ one has

$$\mu\big(\bigcup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} \mu(A_n) \qquad \text{(countably additivity)}.$$

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

As a consequence of the countable additivity axiom, one can derive the following basic properties of a measure.

Proposition 1.2. *Let μ be a measure on $(\Omega, \mathcal{F})$. Then μ satisfies:*

(i) Finite additivity: for disjoint subsets $A_1, \cdots, A_n \in \mathcal{F}$, one has

$$\mu\big(A_1 \cup \cdots \cup A_n\big) = \mu(A_1) + \cdots + \mu(A_n).$$

(ii) Monotonicity: if $A, B \in \mathcal{F}$ and $A \subseteq B$, then $\mu(A) \leqslant \mu(B)$. If in addition $\mu(A) < \infty$, then

$$\mu(B \backslash A) = \mu(B) - \mu(A).$$

(iii) Continuity from below:

$$A_n \in \mathcal{F}, A_n \uparrow A \implies \lim_{n\to\infty} \mu(A_n) = \mu(A).$$

(iv) Continuity from above:

$$A_n \in \mathcal{F}, A_n \downarrow A, \mu(A_1) < \infty \implies \lim_{n\to\infty} \mu(A_n) = \mu(A).$$

In particular, the case when $A = \emptyset$ is called continuity at $\emptyset$.
(v) Countable subadditivity: for any sequence $\{A_n : n \geqslant 1\} \subseteq \mathcal{F}$, one has

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

Proof. (i) Apply the countable additivity property to the sequence

$$A_1, A_2, \cdots, A_n, \emptyset, \emptyset, \emptyset, \cdots.$$

(ii) By applying the finite additivity property to the disjoint union $B = A \cup B \backslash A$, one gets

$$\mu(B) = \mu(A) + \mu(B \backslash A) \geqslant \mu(A).$$

If in addition $\mu(A) < \infty$, one can subtract $\mu(A)$ on both sides of the above equality to get

$$\mu(B\backslash A) = \mu(B) - \mu(A).$$

(iii) Let $\{A_n\}$ be an increasing sequence of subsets in $\mathcal{F}$ such that $A_n \uparrow A$. If $\mu(A_n) = \infty$ for some n, by monotonicity one trivially has

$$\infty = \mu(A_n) \leqslant \mu(A_{n+1}) \leqslant \cdots \leqslant \mu(A) = \infty.$$

Therefore, one may assume that $\mu(A_n) < \infty$ for each n. In this case, if one defines $B_n \triangleq A_n\backslash A_{n-1}$ ($n \geqslant 1$ and $A_0 \triangleq \emptyset$), from Part (ii) one knows that

$$\mu(B_n) = \mu(A_n) - \mu(A_{n-1}).$$

Since $\{B_n\}$ is a disjoint sequence and $\cup_{n=1}^{\infty} B_n = A$, one has

$$\mu(A) = \sum_{n=1}^{\infty} \mu(B_n) = \lim_{n\to\infty} \sum_{k=1}^{n} (\mu(A_k) - \mu(A_{k-1})) = \lim_{n\to\infty} \mu(A_n).$$

(iv) This follows from applying Part (iii) to the sequence $B_n \triangleq (A_1\backslash A_n)$ which now increases to $A_1\backslash A$.

(v) The proof of this property involves a useful technique of writing a union as a disjoint union. Let $\{A_n : n \geqslant 1\} \subseteq \mathcal{F}$. Define

$$B_n \triangleq A_n\backslash\big(A_1 \cup A_2 \cup \cdots \cup A_{n-1}\big) = A_n \cap A_1^c \cap A_2^c \cap \cdots \cap A_{n-1}^c.$$

Then $\{B_n\}$ is a disjoint sequence of subsets and

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n.$$

To see the latter relation, let $\omega \in \cup_n A_n$ so that it belongs at least one of these events. Choose m to be the smallest index n such that $\omega \in A_n$. Then $\omega \in B_m$. It is obvious that $B_n \cap B_m = \emptyset$ for $n \neq m$ and $B_n \subseteq A_n$. Therefore, by countable additivity and monotonicity one has

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

□

The main useful measures one will encounter are those with certain finiteness properties. In particular, probability measures are simply measures μ that have total mass one (i.e. $\mu(\Omega) = 1$), so that $\mu(A)$ can be interpreted as "the probability of the event A" (the monotonicity property shows that $\mu(A) \leqslant \mu(\Omega) = 1$).

Definition 1.6. A measure μ defined on a given measurable space $(\Omega, \mathcal{F})$ is called:

(i) a *finite measure* if $\mu(\Omega) < \infty$;
(ii) a *σ-finite measure* if there exists a partition $\{A_n : n \geqslant 1\}$ of Ω (i.e. a sequence of events satisfying $A_n \cap A_m = \emptyset$ whenever $n \neq m$ and $\Omega = \cup_n A_n$), such that $\mu(A_n) < \infty$ for each $n \geqslant 1$;
(iii) a *probability measure* if $\mu(\Omega) = 1$.

A probability measure is often denoted as $\mathbb{P}$. In this case, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

One can equivalently define a probability measure under *Kolmogorov's three axioms*: (i) $\mathbb{P}(A) \geqslant 0$ for all $A \in \mathcal{F}$, (ii) $\mathbb{P}(\Omega) = 1$ and (iii) $\mathbb{P}$ satisfies the countable additivity property. It should be pointed out that these axioms only provide the natural properties that every probability measure should obey. However, they do *not* construct any specific probability measure on $(\Omega, \mathcal{F})$.

Example 1.5. Consider a random experiment that produces finitely many possible outcomes equally likely (for instance, tossing a fair coin or rolling a fair die). In this case, Ω is a finite set consisting of the possible outcomes of the experiment. $\mathcal{F}$ can be taken to be the collection of all subsets of Ω (the power set). The underlying probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is defined by

$$\mathbb{P}(A) \triangleq \frac{\#(A)}{\#(\Omega)}, \quad A \in \mathcal{F},$$

where $\#(A)$ denotes the number of elements in A. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is known as a *classical probability model*. One can define other legal probability measures on $(\Omega, \mathcal{F})$ as long as they satisfy Kolmogorov's three axioms. For instance, in the example of tossing a coin, one has $\Omega = \{H, T\}$ and $\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$. Given $\alpha \in (0, 1)$, there is an associated probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ defined by

$$\mathbb{P}(\{H\}) = \alpha, \ \mathbb{P}(\{T\}) = 1 - \alpha, \ \mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\Omega) = 1.$$

This model corresponds to tossing a biased coin with "head probability" given by α.

Kolmogorov's axiomatic approach (1930s) was a milestone in the development of probability theory. Before Kolmogorov's time, it was not easy to recognise the significance of the countable additivity property. For quite a long period, probabilists were satisfied by working with the finite additivity property on an algebra (only finitely many steps of event operations were concerned). This was fine as long as the sample space Ω is finite. It was when one came across questions related to infinitely repeated experiments, sequences of independent random variables, construction of stochastic processes etc. that the notions of σ-algebras and countable additivity became essential. It is natural to ask: with finite additivity at hand *what extra condition(s) are needed to ensure countable additivity*? An answer to this question is contained in the following result.

Proposition 1.3. *Let $\mathcal{A}$ be an algebra over Ω. Let $\mu : \mathcal{A} \to [0, \infty)$ be a set function such that $\mu(\emptyset) = 0$ and it satisfies finite additivity on $\mathcal{A}$. Then the following statements are all equivalent:*

(i) μ is countably additive on $\mathcal{A}$: whenever $\{A_n : n \geqslant 1\}$ is a disjoint sequence in $\mathcal{A}$ such that $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$, one has

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

(ii) μ is continuous from below;
(iii) μ is continuous from above;
(iv) μ is continuous at $\emptyset$;
(v) whenever $A \in \mathcal{A}$ and $\{A_n : n \geqslant 1\} \subseteq \mathcal{A}$ satisfy $A \subseteq \cup_{n=1}^{\infty} A_n$, one has

$$\mu(A) \leqslant \sum_{n=1}^{\infty} \mu(A_n).$$

Proof. The argument of (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) and (i) $\Longrightarrow$ (v) is identical to the proof of Proposition 1.2. It remains to consider the following two directions. Let $\{A_n : n \geqslant 1\}$ be a sequences of disjoint sets in $\mathcal{A}$ and $A \triangleq \cup_{n=1}^{\infty} A_n \in \mathcal{A}$.

(iv) $\Longrightarrow$ (i). Since $\mathcal{A}$ is an algebra, $B_n \triangleq \cup_{k=1}^{n} A_n \in \mathcal{A}$. It is obvious that $B_n \subseteq A$ and finite additivity implies

$$\mu(A \backslash B_n) = \mu(A) - \mu(B_n) = \mu(A) - \sum_{k=1}^{n} \mu(A_k).$$

Since $(A\backslash B_n) \downarrow \emptyset$, by assumption one concludes that

$$0 = \lim_{n\to\infty} \mu(A\backslash B_n) = \mu(A) - \sum_{n=1}^{\infty} \mu(A_n),$$

yielding the desired countable additivity property.

(v) $\Longrightarrow$ (i). Define B_n as above. Observe that μ satisfies monotonicity as a consequence of finite additivity. Together with the assumption of (v), one sees that

$$\sum_{k=1}^{n} \mu(A_k) = \mu(B_n) \leqslant \mu(A) \leqslant \sum_{m=1}^{\infty} \mu(A_m)$$

for all n. By taking $n \to \infty$, one obtains the countable additivity property. □

1.3 The core of the matter: construction of probability measures

In elementary probability, one has encountered the concept of distribution function F for a random variable X (as a function on $\mathbb{R}$, $F(x) \triangleq \mathbb{P}(X \leqslant x)$). From a mathematical perspective, there is a basic question one must address: *given a distribution function F, how can one construct a probability space as well as a random variable defined on it whose distribution is F*? For instance, how can one construct a standard normal random variable mathematically?

Although we have not yet make precise the definition of a random variable, we can still outline the key idea of approaching this question. Let us take $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathbb{R})$ (cf. Definition 1.4) as the candidate measurable space. There is an obvious random variable $X : \Omega \to \mathbb{R}$ defined by $X(\omega) \triangleq \omega$ (the identity map). The missing piece is a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, under which the random variable X has distribution function given by F. To put it in another way, one needs to ensure the property $\mathbb{P}(X \leqslant x) = F(x)$, or more generally,

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a), \quad \text{for all } a \leqslant b. \tag{1.3}$$

Let us introduce the semiring

$$\mathcal{C} \triangleq \{(a, b] : a \leqslant b\}$$

on $\mathbb{R}$. The right hand side of (1.3) clearly defines a set function on $\mathcal{C}$. Now the essential point of the question becomes the following: given a "measure" on a semiring $\mathcal{C}$ (the right hand side of (1.3) in the current situation), how can one

extend it to a measure on $\sigma(\mathcal{C})$? The solution to this question is contained in a deep measure-theoretic result known as *Carathéodory's extension theorem.*

We do not want to delve too much into general measure theory. Therefore, we only state this powerful theorem and leave its complete proof to Appendix A.

Theorem 1.2 (Carathéodory's extension theorem)**.** *Let $\mathcal{C}$ be a semiring over a sample space Ω. Suppose that $\mu : \mathcal{C} \to [0, \infty]$ is a set function which satisfies $\mu(\emptyset) = 0$ and the countable additivity property. Then there exists an extension of μ to a measure on $\sigma(\mathcal{C})$ (i.e. a measure $\bar{\mu}$ on $\sigma(\mathcal{C})$ such that $\bar{\mu} = \mu$ on $\mathcal{C}$). In addition, if μ is σ-finite on $\mathcal{C}$, then the extension is unique and the extended measure on $\sigma(\mathcal{C})$ is also σ-finite.*

Here we give the proof for the uniqueness part in the context of finite measures. The argument is a good illustration on the use of Dynkin's π-λ theorem (cf. Theorem 1.1 and the paragraphs following its proof).

Proposition 1.4. *Let $\mathcal{C}$ be a π-system and $\Omega \in \mathcal{C}$. Suppose that μ_1 and μ_2 are two finite measures on $(\Omega, \sigma(\mathcal{C}))$. If $\mu_1 = \mu_2$ on $\mathcal{C}$, then $\mu_1 = \mu_2$ on $\sigma(\mathcal{C})$.*

Proof. Define

$$\mathcal{H} \triangleq \{A \subseteq \sigma(\mathcal{C}) : \mu_1(A) = \mu_2(A)\} \subseteq \sigma(\mathcal{C})$$

to be the collection of subsets satisfying the desired property. We want to show that $\mathcal{H} = \sigma(\mathcal{C})$. First of all, by assumption one knows that $\mathcal{C} \subseteq \mathcal{H}$. Next, one checks that $\mathcal{H}$ is a λ-sysetem:

(L1) By assumption $\Omega \in \mathcal{C} \subseteq \mathcal{H}$.
(L2) Let $A, B \in \mathcal{H}$ and $A \subseteq B$. From finite additivity, one has

$$\mu_1(B \backslash A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \backslash A),$$

showing that $B \backslash A \in \mathcal{H}$.
(L3) Let $\mathcal{H} \ni A_n \uparrow A$. Since probability measures are continuous from below, one has

$$\mu_1(A) = \lim_{n \to \infty} \mu_1(A_n) = \lim_{n \to \infty} \mu_2(A_n) = \mu_2(A),$$

showing that $A \in \mathcal{H}$.

Therefore, $\mathcal{H}$ is a λ-system and thus $\mathcal{H} \supseteq \lambda(\mathcal{C})$. According to Dynkin's π-λ theorem, one concludes that $\sigma(\mathcal{C}) = \lambda(\mathcal{C}) \subseteq \mathcal{H}$. □

Remark 1.2. Proposition 1.4 may fail if $\mathcal{C}$ is not a π-sytem. For example, consider $\Omega = \{1, 2, 3, 4\}$, and $\mathcal{C} = \{\emptyset, \{1, 2\}, \{1, 3\}, \Omega\}$. $\sigma(\mathcal{C})$ consists of all subsets of Ω. The following two probability measures

$$\mu_1(\{1\}) = \mu_1(\{2\}) = \mu_1(\{3\}) = \mu_1(\{4\}) = \frac{1}{4},$$

$$\mu_2(\{1\}) = \mu_2(\{4\}) = \frac{1}{3},\ \mu_2(\{2\}) = \mu_2(\{3\}) = \frac{1}{6},$$

are different but they coincide on $\mathcal{C}$.

An important application of Carathéodory's extension theorem is the construction of probability measures from distribution functions on $\mathbb{R}$. Let us relax the finiteness property for now and consider a more general situation. Throughout the rest of the notes, an *increasing* (respectively, *decreasing*) function f means $x < y \implies f(x) \leqslant f(y)$ (respectively, $\implies f(x) \geqslant y$).

Let $F : \mathbb{R} \to \mathbb{R}$ be a given increasing, right continuous function. Consider the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $\mathcal{C}$ be the semiring consisting of finite intervals $(a, b]$ $(a \leqslant b)$. Using the function F one can define a set function μ on $\mathcal{C}$ by

$$\mu((a, b]) \triangleq F(b) - F(a), \quad (a, b] \in \mathcal{C}. \tag{1.4}$$

We wish to use Carathéodory's extension theorem to obtain a unique extension of μ to $\mathcal{B}(\mathbb{R})$. Before delving into the mathematical details, there are two important examples one should keep in mind.

(i) $F(x) = x$. In this case, $\mu((a, b]) = b - a$ measures the length of the interval $(a, b]$. Therefore, the extended measure $\bar{\mu}$ gives a way of measuring the size of a general set in $\mathcal{B}(\mathbb{R})$ extending the classical notion of length. This was the original motivation of Lebesgue's measure theory in the 1900s.
(ii) $F(x)$ is the distribution function of a random variable X. In this case, the extended measure $\bar{\mu}$ is a probability measure that interprets $\bar{\mu}(A)$ as the probability of event $\{X \in A\}$ for any $A \in \mathcal{B}(\mathbb{R})$.

Let us return to verifying the conditions in Carathéodory's extension theorem for the set function μ defined by (1.4) on $\mathcal{C}$. It is clear that $\mu(\emptyset) = \mu((a, a]) = F(a) - F(a) = 0$. In addition, using the partition $\{(n, n + 1] : n \in \mathbb{Z}\}$ it is apparent that μ is σ-finite. Now it remains to verify that μ is countably additive on $\mathcal{C}$. Before proving this, one first observes that μ is finitely additive on $\mathcal{C}$; for if

$$(a, b] = (a_1, a_2] \cup (a_2, a_3] \cup \cdots \cup (a_{n-1}, a_n]$$

with $a = a_1 < \cdots < a_n = b$, then one has

$$\mu((a,b]) = F(b) - F(a) = \sum_{i=1}^{n}(F(a_i) - F(a_{i-1})) = \sum_{i=1}^{n}\mu((a_{i-1}, a_i]).$$

To show countable additivity, note that *if $\mathcal{C}$ were an algebra*, according to Proposition 1.3 (v), it is equivalent to proving that

$$A, A_n \in \mathcal{C}\ (n \geqslant 1) \text{ with } A \subseteq \bigcup_{n=1}^{\infty} A_n \implies \mu(A) \leqslant \sum_{n=1}^{\infty}\mu(A_n). \tag{1.5}$$

Here we make two comments before proceeding further.

Remark 1.3. (i) The equivalence between countable additivity and (1.5) remains true for finitely additive measures on *semirings* rather than just on algebras.
(ii) A finitely additive measure μ on a semiring $\mathcal{C}$ satisfies a finite version of the subadditivity property (1.5):

$$A, A_1, \cdots, A_n \in \mathcal{C} \text{ with } A \subseteq \bigcup_{i=1}^{n} A_i \implies \mu(A) \leqslant \sum_{i=1}^{n}\mu(A_i). \tag{1.6}$$

The proofs of these two facts in the semiring context are technical and not enlightening. We refer the reader to Lemma 1.1 in Appendix A for the details.

Now we can proceed to prove the countable additivity of μ using (1.5).

Theorem 1.3. *The set function μ is countably additive on $\mathcal{C}$. According to Carathéodory's extension theorem, there exists a unique measure $\bar{\mu}$ on $\mathcal{B}(\mathbb{R})$ such that $\bar{\mu} = \mu$ on $\mathcal{C}$, and $\bar{\mu}$ is also σ-finite.*

Proof. Since μ is finitely additive, according to Remark 1.3 (i), it is enough to check (1.5). Let $(a,b] \subseteq \cup_{n=1}^{\infty}(a_n, b_n]$. Since F is right continuous, for any $\varepsilon > 0$, one can find $a' \in (a,b]$ such that

$$\mu((a,b]) = F(b) - F(a) \leqslant F(b) - F(a') + \varepsilon = \mu((a',b]) + \varepsilon.$$

Similarly, for each n, one can find $b_n' > b_n$ such that

$$\mu((a_n, b_n]) \geqslant \mu((a_n, b_n']) - \frac{\varepsilon}{2^n}.$$

From the constructions, it is clear that

$$[a',b] \subseteq (a,b] \subseteq \bigcup_{n=1}^{\infty}(a_n, b_n] \subseteq \bigcup_{n=1}^{\infty}(a_n, b_n').$$

Since $[a', b]$ is compact, by the Heine-Borel theorem there exists $N \geqslant 1$, such that

$$(a', b] \subseteq [a', b] \subseteq \bigcup_{n=1}^{N} (a_n, b'_n) \leqslant \bigcup_{n=1}^{N} (a_n, b'_n].$$

It follows that

$$\begin{aligned} \mu((a, b]) - \varepsilon \leqslant \mu((a', b]) &\leqslant \sum_{n=1}^{N} \mu((a_n, b'_n]) \\ &\leqslant \sum_{n=1}^{N} \left(\mu((a_n, b_n]) + \frac{\varepsilon}{2^n}\right) \leqslant \sum_{n=1}^{\infty} \mu((a_n, b_n]) + \varepsilon, \end{aligned} \tag{1.7}$$

where in the second inequality we used the finite subadditivity property (1.6) that was mentioned in Remark 1.3 (ii) as a consequence of finite additivity. Since ε is arbitrary, by letting $\varepsilon \downarrow 0$ on both ends of (1.7) one obtains the desired property (1.5). □

Definition 1.7. Let $F : \mathbb{R} \to \mathbb{R}$ a right continuous and increasing function. The unique measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ obtained in Theorem 1.3 is called the *Lebesgue-Stieltjes measure* induced by the function F. It is denoted as μ_F.

The case when $F(x) = x$ is of fundamental importance in real analysis. The resulting measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called the *Lebesgue measure* and it is denoted as dx. The Lebesgue measure extends the notion of length to arbitrary Borel measurable sets.

Example 1.6. In elementary probability, one has seen the random experiment of "picking a point in the unit interval $[0, 1]$ uniformly at random". We are now able to make the construction mathematically precise: the sample space is $[0, 1]$, the family of events is $\mathcal{B}([0, 1]) \triangleq [0, 1] \cap \mathcal{B}(\mathbb{R})$, and the probability measure is the restriction of the Lebesgue measure on Borel measurable subsets of $[0, 1]$.

In the probabilistic context, another important class of examples come from the case when F is a distribution function. We first recall the following definition.

Definition 1.8. A *distribution function on* $\mathbb{R}$ is a right continuous, increasing function $F : \mathbb{R} \to \mathbb{R}$ such that

$$F(-\infty) \triangleq \lim_{x \to -\infty} F(x) = 0, \;\; F(\infty) \triangleq \lim_{x \to \infty} F(x) = 1.$$

As a consequence of Theorem 1.3, distribution functions on $\mathbb{R}$ and probability measures on $\mathcal{B}(\mathbb{R})$ are in one-to-one correspondence (hence they are essentially the same thing).

Corollary 1.1. *There is a one-to-one correspondence between distribution functions on $\mathbb{R}$ and probability measures on $\mathcal{B}(\mathbb{R})$. More precisely, given a distribution function F, the induced probability measure is the Lebesgue-Stieltjes measure μ_F. Conversely, given a probability measure μ, the corresponding distribution function is defined by $F(x) \triangleq \mu((-\infty, x])$.*

Proof. The fact that μ_F is a probability measure follows from its continuity from below as a measure:

$$\mu_F(\mathbb{R}) = \lim_{n\to\infty} \mu_F((-n, n]) = \lim_{n\to\infty} (F(n) - F(-n)) = 1,$$

since F is a distribution function. In addition, by letting $a \to -\infty$ in the relation

$$\mu_F((a, x]) = F(x) - F(a)$$

one recovers the distribution function F. Finally, given a probability measure μ, the function $F(x) \triangleq \mu((-\infty, x])$ is easily seen to be a distribution function (right continuity is a consequence of the continuity of μ applied to $(-\infty, x] = \cap_{n=1}^{\infty}(-\infty, x + 1/n]$). These properties show that the map $F \mapsto \mu_F$ defines a bijection from the space of distribution functions to the space of probability measures. □

There are natural generalisations of the aforementioned constructions to the case of $\mathbb{R}^n$, which we only outline without proofs. In multidimensions, the use of "distribution functions" becomes less natural but the idea of constructing measures is similar to the one-dimensional case.

First of all, one introduces the semiring

$$\mathcal{C} \triangleq \{(a, b] : a, b \in \mathbb{R}^n,\ a \leqslant b\}.$$

Here for $a = (a_1, \cdots, a_n)$ and $b = (b_1, \cdots, b_n)$, the notation $a \leqslant b$ means $a_i \leqslant b_i$ for each i and the "interval" $(a, b]$ refers to the rectangle

$$(a, b] \triangleq \{x = (x_1, \cdots, x_n) : a_i < x_i \leqslant b_i \text{ for each } i\}.$$

The *Borel σ-algebra* on $\mathbb{R}^n$, denoted as $\mathcal{B}(\mathbb{R}^n)$, is the σ-algebra generated by $\mathcal{C}$. It can be shown that $\mathcal{B}(\mathbb{R}^n)$ coincides with the σ-algebra generated by open subsets of $\mathbb{R}^n$.

To construct the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, one starts with the obvious notion of volume for members of $\mathcal{C}$. Precisely, one defines a set function

$$\mu((a,b]) \triangleq \prod_{i=1}^{n}(b_i - a_i)$$

on the semiring $\mathcal{C}$. After verifying the required conditions, Carathéodory's extension theorem yields a unique extension $\bar{\mu}$ of μ to $\mathcal{B}(\mathbb{R}^n)$. This measure $\bar{\mu}$, also denoted as dx, is called the *Lebesgue measure* on $\mathbb{R}^n$. It extends the notion of volume to Borel measurable sets in a natural way. The Lebesgue measure plays a fundamental role in real analysis.

The idea of constructing a probability measure (a Lebesgue-Stieltjes measure) from a given joint distribution function of an $\mathbb{R}^n$-valued random variable is similar. As usual, the first step is to write down the correct definition of μ on $\mathcal{C}$ (which is the easier part) and then to apply the extension theorem (which is the harder part). We will not provide the technical details here. We remark that, if the joint probability density function $f(x_1, \cdots, x_n)$ of the given distribution exists, the induced probability measure is simply defined by

$$\mu(A) = \int_A f(x_1, \cdots, x_n) dx_1 \cdots dx_n, \quad A \in \mathcal{B}(\mathbb{R}).$$

The above integral needs to be understood in the more general sense of Lebesgue (cf. Section 2.2).

1.4 Almost sure properties

In probability theory, one often deals with properties that hold *outside a set of zero probability*. To make this precise, we give the following definition.

Definition 1.9. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A *μ-null set* is a set $N \in \mathcal{F}$ such that $\mu(N) = 0$. A property $\mathbf{P}$ (depending on $\omega \in \Omega$) is said to hold *(μ-)almost everywhere*, denoted as *(μ-)a.e.* if it holds outside a μ-null set. In other words, there exists a μ-null set N such that

$$N^c \subseteq \{\omega : \mathbf{P} \text{ holds}\} \text{ or equivalently } \{\omega : \mathbf{P} \text{ does not hold}\} \subseteq N.$$

In the probabilistic context (i.e. when μ is a probability measure), one uses the term *almost surely* and the notation *a.s.* in replacement.

Proposition 1.5. *A countable union of μ-null sets is again a μ-null set. In addition, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if $\{\Omega_n : n \geqslant 1\}$ is a sequence of certain events (i.e. $\mathbb{P}(\Omega_n) = 1$), then $\cap_{n=1}^{\infty} \Omega_n$ is also a certain event.*

Proof. The first assertion is a simple consequence of the countable subadditivity property:

$$\mu\big(\bigcup_{n=1}^{\infty} N_n\big) \leqslant \sum_{n=1}^{\infty} \mu(N_n) = 0.$$

In the case of probability measures, the second assertion follows from taking complements. □

Consider the random experiment of choosing a point uniformly on $[0, 1]$ (cf. Example 1.6). Then for almost surely the chosen point is an irrational number.

We look at an example which to some extent reflects the need of working with properties that hold a.s. instead of deterministically. Consider the random experiment of tossing a fair coin repeatedly in a sequence. If one defines S_n to be the total number of "heads" among the first n-tosses, it is heuristically convincing that $\frac{S_n}{n} \approx \frac{1}{2}$ when n is large. However, making the phenomenon that "relative frequency eventually stabilises at the theoretical probability" mathematically precise had cost mathematicians decades of effort.

Mathematically, the underlying sample space is given by

$$\Omega = \{\omega = (\omega_1, \omega_2, \cdots) : \omega_n = H \text{ or } T \text{ for each } n\}. \tag{1.8}$$

In other words, each outcome is an infinite sequence where the n-th entry records the result of the n-th toss. A natural σ-algebra on Ω (the family of events) should be the one generated by those experiments at finite steps. More precisely, for each $n \geqslant 1$ define

$$A_n \triangleq \{\omega : \omega_n = H\},\ B_n \triangleq \{\omega : \omega_n = T\}.$$

A_n and B_n are the events corresponding a specific result ("head" or "tail") at the n-th toss (of course one has $B_n = A_n^c$). The underlying σ-algebra is defined to be the one generated by all these events A_n, B_n, namely:

$$\mathcal{F} \triangleq \sigma\left(\{A_1, B_1, A_2, B_2, A_3, B_3, \cdots\}\right). \tag{1.9}$$

In addition, by using Carathéodory's extension theorem, one can show that there is a unique probability measure $\mathbb{P}$ defined on $\mathcal{F}$ such that

$$\mathbb{P}(C_1 \cap \cdots \cap C_n) = \frac{1}{2^n} \tag{1.10}$$

for all $n \geqslant 1$ and $C_i = A_i$ or B_i $(1 \leqslant i \leqslant n)$. The construction of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is best understood using the method of product spaces (cf. Example 3.6 in Section 3.3 below).

Under such a probability model, one can talk about properties that holds a.s. but not for every ω. For instance, the event $E \triangleq \{\omega : \omega_n = H \text{ for some } n\}$ has probability one but it is not true that every outcome belongs to E (apparently $\omega = \{T, T, T, \cdots\}$ is an outcome which does not trigger E). Now let us try to rephrase the aforementioned phenomenon that "relative 'head' frequency eventually stabilises at the theoretical probability" in a more precise way. Consider the event

$$\Lambda \triangleq \Big\{\omega : \frac{\#\{i \leqslant n : \omega_i = H\}}{n} \to \frac{1}{2} \text{ as } n \to \infty\Big\},$$

where $\#(\cdot)$ counts the number of elements in a set. One can show that $\Lambda \in \mathcal{F}$ and it is a consequence of the *strong law of large numbers* that $\mathbb{P}(\Lambda) = 1$. Therefore, one can comfortably say that:

"For almost surely (With probability one), the relative 'head' frequency will eventually converge to the theoretical probability $1/2$."

Apparently, $\Lambda \neq \Omega$; indeed, the almost sure conclusion is the best one to hope for.

We make one further observation. Let $\alpha : \alpha(1) < \alpha(2) < \alpha(3) < \cdots$ be an arbitrary subsequence of positive integers. Consider the event

$$\Lambda_\alpha \triangleq \Big\{\omega : \frac{\#\{i \leqslant n : \omega_{\alpha(i)} = H\}}{n} \to \frac{1}{2} \text{ as } n \to \infty\Big\}.$$

Note that Λ_α is concerned with the same property but restricted on a given subsequence of tosses (the $\alpha(1)$-th, $\alpha(2)$-th, $\cdots$ tosses). From the probabilistic viewpoint, there is no essential difference between Λ_α and Λ. In particular, one also has $\mathbb{P}(\Lambda_\alpha) = \mathbb{P}(\Lambda) = 1$ for each fixed subsequence α. Things can go very wrong if one looks for a deterministic description of these properties. For instance, one without any prior exposure to probability theory might expect that $\cap_\alpha \Lambda_\alpha$ (intersection taken over all possible subsequences α) is a reasonable event capturing the underlying phenomenon, since the property should "apparently" be invariant when restricting on any subsequence of tosses. However, $\cap_\alpha \Lambda_\alpha = \emptyset$! Indeed, for any $\omega \in \Omega$, if there are at most finitely many "heads" in ω then ω cannot belong to any Λ_α. If there are infinitely many "heads" in ω, one can choose a subsequence α along which $\omega_{\alpha(i)} = H$ for all i. In this way ω cannot belong to Λ_α since the relative "head" frequency along this subsequence α is identically one. This shows that no outcomes ω can belong to $\cap_\alpha \Lambda_\alpha$. Note that there is no contradiction

with Proposition 1.5 since the family of all possible subsequences α is *uncountable* (why?).

Appendix A. Carathéodory's extension theorem

In this appendix, we give a complete proof of Carathéodory's extension theorem in the semiring context. In vague terms, the underlying idea of extending a set function $\mu : \mathcal{C} \to [0, \infty]$ to a measure on $\sigma(\mathcal{C})$ can be summarised as follows. First of all, one uses μ to induce a set function μ^* (the *outer measure*), which is not yet as good as a measure but has the advantage of being well-defined for *every* subsets of Ω and $\mu^* = \mu$ on $\mathcal{C}$. Next, one identifies a class $\mathcal{M}$ of subsets which happens to be a σ-algebra and the restriction of μ^* on $\mathcal{M}$ is indeed a measure (this is the most non-trivial part of the argument). Finally, one proves $\mathcal{C} \subseteq \mathcal{M}$ so that $\sigma(\mathcal{C}) \subseteq \mathcal{M}$. The restriction of μ^* on $\sigma(\mathcal{C})$ is the desired extension of μ.

Outer measures

To implement the above idea mathematically, we start with the definition of an outer measure. Let Ω be a given fixed non-empty set. Recall that $\mathcal{P}(\Omega)$ is the power set of Ω (the collection of all subsets of Ω).

Definition 1.10. An *outer measure* is a set function $\nu : \mathcal{P}(\Omega) \to [0, \infty]$ which satisfies the following properties:

(i) $\nu(\emptyset) = 0$;
(ii) whenever $A \subseteq B$, one has $\nu(A) \leqslant \nu(B)$;
(iii) for any sequence $\{A_n : n \geqslant 1\}$ of subsets, one has

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

In general, an outer measure needs not be a measure. However, there is a canonical σ-algebra associated with an outer measure, such that the outer measure restricts to an actual measure on it.

Definition 1.11. Let ν be a given outer measure. The class of *ν-measurable sets* is defined by

$$\mathcal{M} = \{A \subseteq \Omega : \forall D \subseteq \Omega,\ \nu(D) = \nu(A \cap D) + \nu(A^c \cap D)\}.$$

In other words, a subset $A \subseteq \Omega$ is ν-measurable if for any testing set D, ν is additive with respect to the decomposition $D = (A \cap D) \cup (A^c \cap D)$. Note that the defining property for ν-measurable sets is equivalent to that

$$\nu(D) \geqslant \nu(A \cap D) + \nu(A^c \cap D) \quad \text{for all } D \subseteq \Omega, \tag{1.11}$$

since the reverse inequality is obvious from Property (iii) of the definition of an outer measure.

The heart of the argument is contained in the following result.

Proposition 1.6. *The set class $\mathcal{M}$ is a σ-algebra over Ω. In addition, the restriction of ν on $\mathcal{M}$ is a measure.*

Proof. The facts that $\Omega \in \mathcal{M}$ and $A \in \mathcal{M} \implies A^c \in \mathcal{M}$ are trivial. Now let $A_n \in \mathcal{M}$ $(n \geqslant 1)$ and $A \triangleq \cup_{n=1}^{\infty} A_n$. Given a testing subset $D \subseteq \Omega$, one has

$$\begin{aligned}
\nu(D) &= \nu(A_1 \cap D) + \nu(A_1^c \cap D) \quad (\text{since } A_1 \in \mathcal{M}) \\
&= \nu(A_1 \cap D) + \nu(A_2 \cap A_1^c \cap D) + \nu(A_1^c \cap A_2^c \cap D) \quad (\text{since } A_2 \in \mathcal{M}) \\
&\cdots \\
&= \sum_{k=1}^{n} \nu\big((A_k \backslash (A_1 \cup \cdots \cup A_{k-1})) \cap D\big) + \nu(A_1^c \cap \cdots \cap A_n^c \cap D) \\
&\geqslant \sum_{k=1}^{n} \nu\big((A_k \backslash (A_1 \cup \cdots \cup A_{k-1})) \cap D\big) + \nu(A^c \cap D).
\end{aligned}$$

Since this is true for all n, by taking $n \to \infty$ (denote $B_n \triangleq A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$) one obtains that

$$\nu(D) \geqslant \sum_{n=1}^{\infty} \nu(B_n \cap D) + \nu(A^c \cap D) \geqslant \nu(A \cap D) + \nu(A^c \cap D). \tag{1.12}$$

The last inequality follows from the fact that $A \cap D = \cup_{n=1}^{\infty} (B_n \cap D)$. Therefore, A satisfies the condition (1.11) and thus $A \in \mathcal{M}$. This proves that $\mathcal{M}$ is a σ-algebra.

To see that $\nu|_{\mathcal{M}}$ is a measure, let $\{A_n\}$ be a given sequence of disjoint subsets in $\mathcal{M}$. By applying (1.12) to the case of $D = \cup_{n=1}^{\infty} A_n$, then one has $B_n = A_n$ due to disjointness and (1.12) becomes

$$\nu\Big(\bigcup_{n=1}^{\infty} A_n\Big) \geqslant \sum_{n=1}^{\infty} \nu(A_n) + 0 \geqslant \nu\Big(\bigcup_{n=1}^{\infty} A_n\Big) + 0,$$

yielding the desired countable additivity property. □

The notion of an outer measure provides a general way of obtaining a measure if one restricts the outer measure to an appropriate class of measurable sets. Another nice thing about outer measures is that they are easy (and natural) to construct, as seen from the following result.

Proposition 1.7. *Let $\mathcal{C}$ be a set class containing $\emptyset$. Suppose that $\mu : \mathcal{C} \to [0, \infty]$ satisfies $\mu(\emptyset) = 0$ and*

$$A, A_n \in \mathcal{C} \ (n \geqslant 1) \text{ with } A \subseteq \bigcup_{n=1}^{\infty} A_n \implies \mu(A) \leqslant \sum_{n=1}^{\infty} \mu(A_n). \tag{1.13}$$

Define the set function

$$\mu^*(A) \triangleq \inf \left\{ \sum_{n=1}^{\infty} \mu(A_n) : A_n \in \mathcal{C}, \ A \subseteq \bigcup_{n=1}^{\infty} A_n \right\} \tag{1.14}$$

for $A \subseteq \Omega$ (with the convention $\inf \emptyset \triangleq \infty$). Then μ^* is an outer measure. In addition, one has $\mu^* = \mu$ on $\mathcal{C}$.

Proof. We only verify Property (iii) in Definition 1.10 (the first two properties are trivial). Let $\{A_n : n \geqslant 1\}$ be a sequence of subsets of Ω. One may assume that $\mu^*(A_n) < \infty$ for each n, for otherwise the desired property is again trivial. In this case, let $\varepsilon > 0$ be given fixed. For each $n \geqslant 1$, by the definition of $\mu^*(A_n)$ there exists $\{B_{n,m} : m \geqslant 1\} \subseteq \mathcal{C}$ such that $A_n \subseteq \cup_{m=1}^{\infty} B_{n,m}$ and

$$\sum_{m=1}^{\infty} \mu(B_{n,m}) \leqslant \mu^*(A_n) + \frac{\varepsilon}{2^n}.$$

It follows that

$$\mu^*\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leqslant \sum_{n,m=1}^{\infty} \mu(B_{n,m}) \leqslant \sum_{n=1}^{\infty} \mu^*(A_n) + \varepsilon,$$

where the first inequality holds because $\cup_n A_n \subseteq \cup_{n,m} B_{n,m}$. Since ε is arbitrary, one obtains the desired countable subadditivity property.

For the last assertion, let $A \in \mathcal{C}$. Using the obvious covering $A \subseteq A \cup \emptyset \cup \emptyset \cup \cdots$ one sees that $\mu^*(A) \leqslant \mu(A)$. Conversely, for any sequence $\{A_n\} \subseteq \mathcal{C}$ with $A \subseteq \cup_{n=1}^{\infty} A_n$, by the assumption on μ one has

$$\mu(A) \leqslant \sum_{n=1}^{\infty} \mu(A_n).$$

Therefore, $\mu(A) \leqslant \mu^*(A)$. It follows that $\mu^*(A) = \mu(A)$. □

Proof of Carathéodory's extension theorem

Now we come back to the discussion of Carathéodory's extension theorem. For convenience, we recall the statement of the theorem below.

Theorem 1.4. *Let $\mathcal{C}$ be a semiring on Ω. Suppose that μ is a set function on $\mathcal{C}$ taking values in $[0,\infty]$ which satisfies $\mu(\emptyset)=0$ and the countable additivity property. Then there exists an extension of μ to a measure on $\sigma(\mathcal{C})$ (i.e. a measure $\bar{\mu}$ on $\sigma(\mathcal{C})$ such that $\bar{\mu}=\mu$ on $\mathcal{C}$). In addition, if μ is σ-finite on $\mathcal{C}$, then the extension is unique and the extended measure on $\sigma(\mathcal{C})$ is also σ-finite.*

Recall that the strategy of obtaining the extended measure consists of the following three steps: (i) obtain an outer measure μ^* from μ, (ii) restrict this outer measure to its class $\mathcal{M}$ of measurable sets, (iii) $\mathcal{M}$ contains the original semiring $\mathcal{C}$.

For the first step, in order to apply Proposition 1.7 one needs to show (1.13) from countable additivity. This is done in the following lemma. We remark that this lemma generalises Proposition 1.3 (the equivalence between (i) and (v)) to the semiring case. Such a result was implicitly used when we proved Theorem 1.3.

Lemma 1.1. *Let $\mu:\mathcal{C}\to[0,\infty]$ be a set function defined on a semiring $\mathcal{C}$ with $\mu(\emptyset)=0$. Then μ is countably additive if and only if it is finitely additive and satisfies the subadditivity property (1.13).*

Proof. Necessity. Finite additivity is obvious since $\mu(\emptyset)=0$. To prove (1.13), let $\{A_n : n\geqslant 1\}\subseteq\mathcal{C}$ and $A\in\mathcal{C}$ satisfy $A\subseteq\cup_{n=1}^{\infty}A_n$. As a standard trick one can write the union as a disjoint union

$$\bigcup_{n=1}^{\infty}A_n=\bigcup_{n=1}^{\infty}B_n$$

where $B_n\triangleq A_n\backslash(A_1\cup\cdots\cup A_{n-1})$. Next, we claim that B_n is a finite disjoint union of members of $\mathcal{C}$. Indeed, by the definition of a semiring, this is true for $A_n\backslash A_1$, and so is for $(A_n\backslash A_1)\backslash A_2$, and inductively also for $B_n=((A_n\backslash A_1)\backslash A_2)\cdots\backslash A_{n-1}$. In particular, $A\cap B_n$ is also a finite disjoint union of members of $\mathcal{C}$, say

$$A\cap B_n=\bigcup_{m=1}^{k_n}C_{n,m}$$

with some $C_{n,m}\in\mathcal{C}$ $(1\leqslant m\leqslant k_n)$ all being disjoint. It follows that $\{C_{n,m}: n\geqslant 1,\ 1\leqslant m\leqslant k_n\}$ is a countable disjoint family whose union is A. By countable

additivity, one has

$$\mu(A) = \sum_{n=1}^{\infty} \sum_{m=1}^{k_n} \mu(C_{n,m}). \tag{1.15}$$

On the other hand, since

$$\bigcup_{m=1}^{k_n} C_{n,m} \subseteq A_n,$$

similar reason shows that $A_n \backslash (\cup_{m=1}^{k_n} C_{n,m})$ is a finite disjoint union of members of $\mathcal{C}$. It then follows from finite additivity that

$$\sum_{m=1}^{k_n} \mu(C_{n,m}) \leqslant \mu(A_n). \tag{1.16}$$

The relations (1.15) and (1.16) yield the desired property

$$\mu(A) \leqslant \sum_{n=1}^{\infty} \mu(A_n).$$

Sufficiency. Let $\{A_n : n \geqslant 1\} \subseteq \mathcal{C}$ be a disjoint sequence and $A \triangleq \cup_{n=1}^{\infty} A_n \in \mathcal{C}$. By assumption, one has

$$\mu(A) \leqslant \sum_{n=1}^{\infty} \mu(A_n).$$

For the reverse inequality, since $A \backslash (A_1 \cup \cdots \cup A_n)$ is a finite disjoint union of members of $\mathcal{C}$, one knows from finite additivity that

$$\mu(A) \geqslant \sum_{k=1}^{n} \mu(A_k).$$

The desired result follows by taking $n \to \infty$. □

As a consequence, if $\mu : \mathcal{C} \to [0, \infty]$ is countably additive and $\mu(\emptyset) = 0$, it satisfies (1.13) and thus one can use equation (1.14) to induce an outer measure μ^* that agrees with μ on $\mathcal{C}$. Let $\mathcal{M}$ denote the set of μ^*-measurable sets. It follows from Proposition 1.6 that the restriction of μ^* on $\mathcal{M}$ is a measure. To further restrict μ^* to $\sigma(\mathcal{C})$, one must show that $\mathcal{C} \subseteq \mathcal{M}$. The following lemma provides a simple way of checking the measurability condition (1.11).

Lemma 1.2. *Under the same notation as above, given any subset $A \subseteq \Omega$ one has $A \in \mathcal{M}$ if and only if*

$$\mu(C) \geqslant \mu^*(A \cap C) + \mu^*(A^c \cap C), \quad \forall A \in \mathcal{C}. \tag{1.17}$$

Proof. Let $A \subseteq \Omega$ be a given subset satisfying the property (1.17). We want to verify the condition (1.11) for any $D \subseteq \Omega$. One may assume $\mu^*(D) < \infty$ for otherwise the claim is trivial. Given $\varepsilon > 0$, by the definition (1.14) of μ^*, there exists $\{C_n : n \geqslant 1\} \subseteq \mathcal{C}$ such that $D \subseteq \cup_{n=1}^{\infty} C_n$ and

$$\sum_{n=1}^{\infty} \mu(C_n) \leqslant \mu^*(D) + \varepsilon.$$

Since $C_n \in \mathcal{C}$, one knows that

$$\mu(C_n) \geqslant \mu^*(A \cap C_n) + \mu^*(A^c \cap C_n).$$

As a result,

$$\begin{aligned} \mu^*(D) + \varepsilon &\geqslant \sum_{n=1}^{\infty} \mu^*(A \cap C_n) + \sum_{n=1}^{\infty} \mu^*(A^c \cap C_n) \\ &\geqslant \mu^*(A \cap D) + \mu^*(A^c \cap D), \end{aligned}$$

where the last inequality follows from the fact that

$$A \cap D \subseteq \bigcup_{n=1}^{\infty} (A \cap C_n), \ A^c \cap D \subseteq \bigcup_{n=1}^{\infty} (A^c \cap C_n)$$

as well as the countable subadditivity property of μ^*. The desired property follows by letting $\varepsilon \downarrow 0$. □

We are now able to complete the proof of Carathéodory's extension theorem.

Completing the proof of Theorem 1.4. For the existence of extension, it remains to show that $\mathcal{C} \subseteq \mathcal{M}$. Let $A \in \mathcal{C}$. For any $C \in \mathcal{C}$, since $\mathcal{C}$ is a semiring one knows that $C \backslash A$ is a finite disjoint union of members of $\mathcal{C}$, say

$$A^c \cap C = \bigcup_{i=1}^{k} B_i,$$

where $B_i \in \mathcal{C}$ and $B_i \cap B_j = \emptyset$ $(i \neq j)$. It follows from the finite additivity of μ that

$$\mu(C) = \mu(A \cap C) + \sum_{i=1}^{k} \mu(B_i) \geqslant \mu(A \cap C) + \mu^*(A^c \cap C),$$

where the last inequality is a trivial consequence of the definition of μ^*. According to Lemma 1.2, one concludes that $A \in \mathcal{M}$. This finishes the proof of the existence of a measure extension to $\sigma(\mathcal{C})$.

Finally, we prove uniqueness under the assumption that μ is σ-finite on $\mathcal{C}$ (namely, there exists a partition $\{A_n : n \geqslant 1\} \subseteq \mathcal{C}$ of Ω such that $\mu(A_n) < \infty$ for every n). Suppose that μ_1 and μ_2 are two measures on $\sigma(\mathcal{C})$ satisfying $\mu_1 = \mu_2 = \mu$ on $\mathcal{C}$. For each fixed n, if one regards μ_1, μ_2 as measures on $(A_n, A_n \cap \sigma(\mathcal{C}))$ (one now views A_n as the sample space), then they are finite measures which coincide on the π-system $A_n \cap \mathcal{C}$ on A_n. Note that $A_n = A_n \cap A_n \in A_n \cap \mathcal{C}$. It follows from Proposition 1.4 that $\mu_1 = \mu_2$ on $A_n \cap \sigma(\mathcal{C})$. As this is true for every n, from countable additivity one concludes that

$$\mu_1(A) = \sum_{n=1}^{\infty} \mu_1(A_n \cap A) = \sum_{n=1}^{\infty} \mu_2(A_n \cap A) = \mu_2(A) \quad \forall A \in \sigma(\mathcal{C}).$$

□

Remark 1.4. The careful reader may notice that in the argument of the uniqueness part, we implicitly used the fact that the σ-algebra generated by $A_n \cap \mathcal{C}$ on A_n coincides with $A_n \cap \sigma(\mathcal{C})$. The justification of this fact is left as an exercise.

The completion of a measure space

In general, the class $\mathcal{M}$ (μ^*-measurable sets) is strictly larger than $\sigma(\mathcal{C})$. There is an important relationship between the two measure spaces $(\Omega, \mathcal{M}, \mu^*)$ and $(\Omega, \sigma(\mathcal{C}), \mu^*)$ (in the σ-finite case): the former is the *completion* of the latter. To explain this, we first introduce the notion of complete measure spaces.

Definition 1.12. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Define the collection of *μ-null sets* as

$$\mathcal{N} \triangleq \{N \in \mathcal{F} : \mu(N) = 0\}. \tag{1.18}$$

The measure space $(\Omega, \mathcal{F}, \mu)$ is said to be *complete*, if subsets of members of $\mathcal{N}$ are always $\mathcal{F}$-measurable:

$$N \in \mathcal{N}, A \subseteq N \implies A \in \mathcal{F}.$$

Remark 1.5. For the complete measure space $(\Omega, \mathcal{F}, \mu)$, it is immediate that $\mu(A) = 0$ if $A \subseteq N \in \mathcal{N}$. A simple reason of introducing the notion of completeness is that one does not want to border with subsets of null sets and one should simply make them all measurable with zero measure.

A measure space $(\Omega, \mathcal{F}_2, \mu_2)$ is called an *extension* of $(\Omega, \mathcal{F}_1, \mu_1)$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and $\mu_1 = \mu_2$ on $\mathcal{F}_1$. Among all complete extensions of a given measure space, there is always a unique smallest one.

Theorem 1.5. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. There exists a unique complete measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$, such that the following two properties hold true.*

(i) $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is an extension of $(\Omega, \mathcal{F}, \mu)$.
(ii) For any complete measure space $(\Omega, \mathcal{F}', \mu')$ that extends $(\Omega, \mathcal{F}, \mu)$, it is also an extension of $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$.

Proof. Uniqueness is obvious from Property (ii). To prove uniqueness, one constructs $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ explicitly as follows. Recall that $\mathcal{N}$ is the class of μ-null sets defined by (1.18). Set

$$\bar{\mathcal{F}} \triangleq \{A \cup E : A \in \mathcal{F}, E \subseteq N \text{ with some } N \in \mathcal{N}\},$$

and define $\bar{\mu}$ on $\bar{\mathcal{F}}$ by

$$\bar{\mu}(A \cup E) \triangleq \mu(A), \quad A \cup E \in \bar{\mathcal{F}}.$$

First of all, $\bar{\mathcal{F}}$ is a σ-algebra. Indeed, $\Omega = \Omega \cap \emptyset \in \bar{\mathcal{F}}$. In addition, $A \cup E \in \bar{\mathcal{F}}$ with $E \subseteq N \in \mathcal{N}$, then

$$(A \cup E)^c = A^c \cap E^c = \big(A^c \cap N^c\big) \cup \big(A^c \cap (E^c \backslash N^c)\big).$$

Since

$$A^c \cap (E^c \backslash N^c) = A^c \cap E^c \cap N \subseteq N,$$

one sees that $(A \cup E)^c \in \bar{\mathcal{F}}$. Finally, if $A_n \cup E_n \subseteq \bar{\mathcal{F}}$ with $E_n \subseteq N_n \in \mathcal{N}$ $(n \geqslant 1)$, then

$$\bigcup_{n=1}^{\infty}(A_n \cup E_n) = \big(\bigcup_{n=1}^{\infty} A_n\big) \cup \big(\bigcup_{n=1}^{\infty} E_n\big) \in \bar{\mathcal{F}},$$

since $\cup_{n=1}^{\infty} E_n \subseteq \cup_{n=1}^{\infty} N_n \in \mathcal{N}$.

Next, $\bar{\mu}$ is well-defined. Indeed, suppose that $A_1 \cup E_1 = A_2 \cup E_2$ for some $A_i \in \mathcal{F}$, $E_i \subseteq N_i \in \mathcal{N}$ $(i = 1, 2)$. Then $A_1 \subseteq A_2 \cup N_2$ and thus

$$\mu(A_1) \leqslant \mu(A_2) + \mu(N_2) = \mu(A_2).$$

In the same way, the reverse inequality $\mu(A_1) \geqslant \mu(A_2)$ also holds.

It is obvious that $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is an extension of $(\Omega, \mathcal{F}, \mu)$. To see its completeness, let $A \cup E \in \bar{\mathcal{F}}$ be a $\bar{\mu}$-null set. By definition one has $E \subseteq N$ for some $N \in \mathcal{N}$ and $\mu(A) = 0$. Therefore, $A \cup N \in \mathcal{N}$. Now for any subset $B \subseteq A \cup E$, one can write B as $B = \emptyset \cup (A \cup E)$ to see that $B \in \bar{\mathcal{F}}$.

It remains to check Property (ii) of the theorem. Suppose that $(\Omega, \mathcal{F}', \mu')$ is another complete measure space which extends $(\Omega, \mathcal{F}, \mu)$. Given $A \cup E \in \bar{\mathcal{F}}$ with $E \subseteq N \in \mathcal{N}$, by the completeness of $\mathcal{F}'$ one knows that $E \in \mathcal{F}'$. Since $A \in \mathcal{F} \subseteq \mathcal{F}'$, one has $A \cup E \in \mathcal{F}'$. Therefore, $\bar{\mathcal{F}} \subseteq \mathcal{F}'$. One also has $\mu' = \bar{\mu}$ on $\bar{\mathcal{F}}$ since

$$\mu(A) \leqslant \mu'(A \cup E) \leqslant \mu(A) + \mu'(E) = \mu(A)$$

whenever $A \cup E \in \bar{\mathcal{F}}$. □

We can now establish the relationship between the spaces $(\Omega, \mathcal{M}, \mu^*)$ and $(\Omega, \sigma(\mathcal{C}), \mu^*)$ arising from the proof of Carathéodory's extension theorem. Note that $(\Omega, \mathcal{M}, \mu^*)$ is a complete measure space. Indeed, if $A \subseteq \Omega$ has zero outer measure, then for any $D \subseteq \Omega$, one has

$$\mu^*(D) \geqslant \mu^*(A^c \cap D) = \mu^*(A \cap D) + \mu^*(A^c \cap D),$$

showing that $A \in \mathcal{M}$. The interesting fact is, in the σ-finite case, $(\Omega, \mathcal{M}, \mu^*)$ turns out to be the completion of $(\Omega, \sigma(\mathcal{C}), \mu^*)$.

Proposition 1.8. *Let $\mathcal{C}$ be a semiring and let $\mu : \mathcal{C} \to [0, \infty]$ be a σ-finite, countably additive set function with $\mu(\emptyset) = 0$. The measure space $(\Omega, \mathcal{M}, \mu^*)$ is the completion of $(\Omega, \sigma(\mathcal{C}), \mu^*)$.*

Proof. We only consider the case when $\mu^*(\Omega) < \infty$, leaving the σ-finite case as an exercise. It is enough to prove that $\mathcal{M} \subseteq \overline{\sigma(\mathcal{C})}$ (the completion of $\sigma(\mathcal{C})$). Let $A \in \mathcal{M}$. We first claim that, there exists $B \in \sigma(\mathcal{C})$, such that $B \supseteq A$ and $\mu^*(B) = \mu^*(A)$. Indeed, by the definition of μ^* (cf. equation (1.14)), for each $n \geqslant 1$ there exists a sequence $\{B_{n,m} : m \geqslant 1\} \subseteq \mathcal{C}$ such that $A \subseteq \cup_{m=1}^{\infty} B_{n,m}$ and

$$\sum_{m=1}^{\infty} \mu(B_{n,m}) < \mu^*(A) + \frac{1}{n}.$$

Set $B_n \triangleq \cup_{m=1}^{\infty} B_{n,m} \in \sigma(\mathcal{C})$ and $B \triangleq \cap_{n=1}^{\infty} B_n \in \sigma(\mathcal{C})$. It follows that

$$\mu^*(B) \leqslant \mu^*(B_n) \leqslant \mu^*(A) + \frac{1}{n}.$$

Letting $n \to \infty$ one obtains that $\mu^*(B) \leqslant \mu^*(A)$. One also has the reverse inequality since $A \subseteq B$. Therefore, the claim holds. By applying the claim to the set A^c, one can find another set $C \in \sigma(\mathcal{C})$ such that $C \subseteq A$ and $\mu^*(C) = \mu^*(A)$. To summarise, one has

$$B, C \in \sigma(\mathcal{C}),\ C \subseteq A \subseteq B,\ \mu^*(A) = \mu^*(C) = \mu^*(B).$$

It follows that $A \backslash C \subseteq B \backslash C$ and $B \backslash C$ is a $\sigma(\mathcal{C})$-measurable set with zero measure. Since $A = C \cup (A \backslash C)$ and $C \in \sigma(\mathcal{C})$, one concludes that $A \in \overline{\sigma(\mathcal{C})}$. □

Example 1.7. An important example is the *Lebesgue measure* on $\mathbb{R}^n$. This is obtained by extending the usual notion of volume for rectangles:

$$\mu((a, b]) = \prod_{i=1}^{n} (b_i - a_i)$$

where $a = (a_1, \cdots, a_n)$, $b = (b_1, \cdots, b_n)$ with $a_i \leqslant b_i$. We denote μ^* as the outer measure induced by μ. Members of $\mathcal{M}$ (the class of μ^*-measurable sets) are called *Lebesgue measurable sets* and the resulting measure is called the *n-dimensional Lebesgue measure*. It is interesting to point out that, there are Lebesgue measurable sets that do not belong to $\mathcal{B}(\mathbb{R}^n)$, and there are subsets of $\mathbb{R}^n$ which are not Lebesgue measurable. However, the construction of any of these examples is not a simple exercise.

Appendix B. Construction of generated σ-algebras

In this appendix, we give a "constructive description" of the σ-algebra $\sigma(\mathcal{E})$ generated by a set class $\mathcal{E}$ and discuss some of its implications. The main result is stated in Theorem 1.7 below.

Heuristically, one may expect that $\sigma(\mathcal{E})$ is obtained by performing a series of basic set operations (intersection, union, complementation) inductively to members in $\mathcal{E}$ in a countable manner. However, this description turns out to problematic if one is not careful enough about the underlying induction procedure. To give a more precise formulation, since

$$A \cap B = (A^c \cup B^c)^c,$$

one can only consider unions and complementations without loss of generality. We first make precise the definition of "performing basic set operations for countably many times". Throughout the rest, X will denote a given sample space (i.e. a non-empty set).

Definition 1.13. For any set class $\mathcal{H}$ over X containing $\emptyset$, we define $\mathcal{H}^*$ to be the collection of finite or countable unions $\cup_n H_n$, where each H_n has the form $H_n = H$ or H^c with $H \in \mathcal{H}$.

Let $\mathcal{E}$ be a set class over X containing $\emptyset$. As a natural idea, one may start with $\mathcal{E}_0 \triangleq \mathcal{E}$ and define inductively $\mathcal{E}_n \triangleq \mathcal{E}_{n-1}^*$ for $n \geqslant 1$. By definition, members of the set class $\mathcal{H} \triangleq \cup_{n \geqslant 0} \mathcal{E}_n$ are subsets that can be obtained by performing basic set operations to members in $\mathcal{E}$ in a countable manner. Apparently $\mathcal{H} \subseteq \sigma(\mathcal{E})$. One may naively expect that $\mathcal{H} = \sigma(\mathcal{E})$. However, it is a rather deep fact that this intuition is simply not true.

Proposition 1.9 (Cf. [Bil86], pp.26-28)**.** *Let $\mathcal{E}$ be the set class of intervals $(a, b]$ over $X = (0, 1]$. Define $\mathcal{E}_n$ and $\mathcal{H}$ as above. Then $\mathcal{H}$ is strictly smaller than $\mathcal{B}((0, 1])$.*

Of course, the main issue here is that $\mathcal{H}$ is not even a σ-algebra, although why it fails to be closed under countable union is far from being obvious. The idea of "inductively" constructing $\sigma(\mathcal{E})$ by countable set operations is not problematic. The crucial point is that the induction procedure here should be performed in a *transfinite* (rather than countable!) manner in order to obtain a true σ-algebra. Our main goal in this appendix is to give the transfinite construction of $\sigma(\mathcal{E})$ in a precise mathematical way. Our discussion follows the main lines of [Bil86, Hal60, HS75].

Ordinal numbers and transfinite induction

Since the construction of $\sigma(\mathcal{E})$ relies critically on the notion of ordinal numbers and transfinite induction, we begin by discussing relevant concepts in a relatively self-contained way. The reader is referred to Halmos [Hal60] and Hewitt-Stromberg [HS75] for more details.

Cardinality

The concept of cardinality is familiar to us. It is a way of measuring the size of a set in terms of the "number" of its elements. To summarise, every set A is associated with a symbol called *cardinality* (or cardinal number), which is denoted as $\text{card}(A)$. For example, a finite set with n elements has cardinality n. The set of integers has cardinality denoted as $\aleph_0$ (the *countable cardinality*). The set of real numbers has cardinality denoted as $\aleph$ (the *continuum cardinality*).

Different cardinalities can be compared in the following natural manner. Given two sets A, B, we say that

$$\mathrm{card}(A) \leqslant \mathrm{card}(B)$$

if there exists an injective map from A to B. The two sets A, B have the same cardinality if

$$\mathrm{card}(A) \leqslant \mathrm{card}(B) \text{ and } \mathrm{card}(B) \leqslant \mathrm{card}(A),$$

or equivalently, if there is a bijection between A and B. We say that

$$\mathrm{card}(A) < \mathrm{card}(B)$$

if $\mathrm{card}(A) \leqslant \mathrm{card}(B)$ but $\mathrm{card}(A) \neq \mathrm{card}(B)$.

Cardinalities can also be added and multiplied. Let α, β be the cardinality of A, B respectively. Then $\alpha + \beta$ (respectively, $\alpha \times \beta$) is the cardinality of the disjoint union of A and B (respectively, the Cartesian product of A and B). It is easy to see that this definition is independent of the representatives A, B for the cardinalities α, β.

We mention some basic properties of cardinality. For instance, one has

$$n < \aleph_0 < \aleph \quad \forall n \in \mathbb{N}.$$

In addition,

$$\aleph_0 + \aleph_0 = \aleph_0 \times \aleph_0 = \aleph_0, \ \aleph + \aleph = \aleph \times \aleph = \aleph.$$

For any set A, one has

$$\mathrm{card}(A) < \mathrm{card}(\mathcal{P}(A)),$$

where $\mathcal{P}(A)$ is the *power set* of A, which by definition is the set of all subsets of A. The set of infinite $\{0,1\}$-sequences has cardinality continuum. This can be shown by representing real numbers as binary sequences. Another useful fact is that a continuum union of continuum is again continuum. More precisely, let $\mathcal{I}$ be a set with cardinality $\aleph$ and for each $i \in \mathcal{I}$ let A_i be a set with cardinality $\aleph$. Then one has

$$\mathrm{card}\big(\bigcup_{i\in\mathcal{I}} A_i\big) = \aleph.$$

This can be seen by defining an injective map T from $\cup_{i\in\mathcal{I}} A_i$ into $\mathbb{R} \times \mathbb{R}$ in the following way:

$$\bigcup_{i\in\mathcal{I}} A_i \ni z \mapsto (i, z) \in \mathcal{I} \times A_i,$$

where one picks an i such that $z \in A_i$ and identifies $\mathcal{I} \times A_i$ with $\mathbb{R} \times \mathbb{R}$.

The *continuum hypothesis* (CH) asserts that there does not exist a set whose cardinality is strictly between $\aleph_0$ and $\aleph$. In other words, the "next" cardinality after countability is continuum. In Zermelo–Fraenkel set theory with the axiom of choice (ZFC), it was proved by Paul Cohen in 1963 that CH is independent of ZFC, i.e. CH cannot be proven or disproved by the ZFC axioms. Cohen was awarded the Fields Medal in 1966 for his proof.

Ordinal numbers

The notion of cardinality has nothing to do with ordering. If one takes into account orders, one is led to an important concept of ordinal numbers as well as a powerful technique of transfinite induction which generalises the classical mathematical induction.

Recall that, a *partially ordered set* is a set equipped with a partial order. A *totally ordered set* is a partially ordered set in which any two elements are comparable. A crucial concept is a *well ordered set*, which by definition is a totally ordered set such that every non-empty subset has a least element in the given ordering (a *least element* of a subset S is an element of S which is smaller than every other element in S). For example, $\mathbb{R}$ with the usual ordering is totally ordered but not well ordered, while $\mathbb{N}$ is a well ordered set.

Every totally ordered set A is associated with a symbol called the *order type* of A. This is the content of the following definition.

Definition 1.14. Let A, B be totally ordered sets. An *order isomorphism* from A to B is a bijection $f : A \to B$ such that $x \leqslant y$ implies that $f(x) \leqslant f(y)$. If such an isomorphism exists, we say that A and B have the *same order type*. We use the symbol $\mathrm{ord}(A)$ to denote the order type of A. If in addition A is well ordered, we call $\mathrm{ord}(A)$ an *ordinal number*.

Under the usual ordering of real numbers, one has $\mathrm{ord}(\{1, 2, \cdots, n\}) = n$. Symbolically, one writes $\mathrm{ord}(\mathbb{N}) = \omega$, $\mathrm{ord}(\mathbb{Q}) = \eta$. Note that ω is an ordinal number since $\mathbb{N}$ is well ordered, while η is not an ordinal number since $\mathbb{Q}$ is not well ordered.

Just like cardinality, two ordinal numbers can be compared. More precisely, if A is well ordered set and $x \in A$, we define $A_x \triangleq \{y \in A : y < x\}$ to be the *initial segment* of A determined by x. Given two ordinal numbers $\alpha = \mathrm{ord}(A)$ and $\beta = \mathrm{ord}(B)$, we say that $\alpha < \beta$ if A is order isomorphic to B_x for some $x \in B$. We say that $\alpha \leqslant \beta$ if either $\alpha < \beta$ or $\alpha = \beta$. It is easy to see that this definition is independent of the representatives A and B. Furthermore, it can be shown

that for any two ordinal numbers α and β, precisely one of the following three alternatives occurs: $\alpha < \beta$, $\alpha = \beta$ or $\alpha > \beta$. It follows that any set of ordinal numbers is totally ordered.

Below is a fundamental result about ordinal numbers. For each ordinal number α, let us define

$$P_\alpha \triangleq \{\beta : \beta \text{ is an ordinal number and } 0 \leqslant \beta < \alpha\}$$

to be the set of ordinal numbers that are strictly smaller than α.

Theorem 1.6. *There is a smallest ordinal number, denoted as Ω, such that P_Ω is uncountable (in other words, if α is any ordinal number with P_α being uncountable, then $\Omega \leqslant \alpha$). In addition, the set P_Ω satisfies the following properties:*

(i) P_Ω is well ordered;
(ii) P_Ω is uncountable and under the continuum hypothesis one has $\mathrm{card}(P_\Omega) = \aleph$*;*
(iii) for each $\alpha \in P_\Omega$, the set P_α is countable;
(iv) for any countable subset $C \subseteq P_\Omega$, there exists $\beta \in P_\Omega$ such that $\alpha < \beta$ for all $\alpha \in C$.

Transfinite induction

Finally, we discuss the method of transfinite induction for well ordered sets, which is a natural generalisation of the classical mathematical induction.

The Principle of Transfinite Induction. Let W be a well ordered set. Suppose that one wants to prove certain property $\mathbf{P}$ that depends on $w \in W$. The principle consists of the following two steps:

(i) *Initial Step.* Show that $\mathbf{P}$ holds when w is the least element of W.
(ii) *Induction Step.* Let $\alpha \in W$ and assume that $\mathbf{P}$ holds for all $\beta \in W$ that are strictly smaller than α. Show that $\mathbf{P}$ holds for α.

Then one is able to conclude that the property $\mathbf{P}$ holds for all $\alpha \in W$.

Proof of the principle of transfinite induction. The proof is indeed rather straight forward. Let $S \subseteq W$ be the subset of elements for which the property $\mathbf{P}$ holds. If $S \neq W$, by the well-orderedness assumption one knows that $W \backslash S$ has a least element, say α. This means that every $\beta \in W \backslash S$ is greater than or equal to α. Equivalently, if $\beta < \alpha$ then $\beta \in S$. By the induction step, one concludes that $\alpha \in S$, which is a contradiction. Therefore, $S = W$ and thus $\mathbf{P}$ holds for all $w \in W$.

□

Construction of $\sigma(\mathcal{E})$

Let X be the underlying sample space. Let $\mathcal{E}$ be a given set class over X containing $\emptyset$. Our goal is to give an "explicit" construction of $\sigma(\mathcal{E})$. Heuristically, the main idea is that $\sigma(\mathcal{E})$ should be the class of subsets that can be obtained from $\mathcal{E}$ by performing series of basic set operations in a suitably inductive manner.

Let Ω denote the smallest uncountable ordinal number given by Theorem 1.6. For each ordinal number $0 \leqslant \alpha < \Omega$, we define a set class $\mathcal{E}_\alpha$ over X as follows. $\mathcal{E}_0$ is defined to be $\mathcal{E}$. Given $0 < \alpha < \Omega$, if $\mathcal{E}_\beta$ is already defined for all ordinal numbers $0 \leqslant \beta < \alpha$, we then set (cf. Definition 1.13 for the notation $(\cdot)^*$)

$$\mathcal{E}_\alpha \triangleq \big(\bigcup_{0 \leqslant \beta < \alpha} \mathcal{E}_\beta \big)^*,$$

Finally, we define

$$\mathcal{F} \triangleq \bigcup_{0 \leqslant \alpha < \Omega} \mathcal{E}_\alpha. \tag{1.19}$$

The main theorem about the construction of $\sigma(\mathcal{E})$ is stated as follows.

Theorem 1.7. *Let $\mathcal{E}$ be a class of subsets of X containing $\emptyset$. Define $\mathcal{F}$ as above. Then $\mathcal{F} = \sigma(\mathcal{E})$.*

Proof. Firstly, we apply transfinite induction to the well ordered set

$$P_\Omega \triangleq \{0 \leqslant \alpha < \Omega : \alpha \text{ is an ordinal number}\} \tag{1.20}$$

to show that $\mathcal{F} \subseteq \sigma(\mathcal{E})$ (equivalently, $\mathcal{E}_\alpha \subseteq \sigma(\mathcal{E})$ for all $\alpha \in P_\Omega$). As the initial step, it is obvious that $\mathcal{E}_0 \subseteq \sigma(\mathcal{E})$. Now let $\alpha \in P_\Omega$ and assume that $\mathcal{E}_\beta \subseteq \sigma(\mathcal{E})$ for all $\beta < \alpha$. Note that any element in $\mathcal{E}_\alpha$ has the form $\cup_n A_n$, where $A_n = A$ or A^c with $A \in \mathcal{E}_\beta$ for some $\beta < \alpha$. By the transfinite induction hypothesis, one knows that $A_n \in \sigma(\mathcal{E})$. Therefore, $\cup_n A_n \in \sigma(\mathcal{E})$. This shows that $\mathcal{E}_\alpha \subseteq \sigma(\mathcal{E})$. By transfinite induction, one concludes that $\mathcal{F} \subseteq \sigma(\mathcal{E})$.

For the other inclusion $\sigma(\mathcal{E}) \subseteq \mathcal{F}$, since $\mathcal{E} = \mathcal{E}_0 \subseteq \mathcal{F}$ it suffices to show that $\mathcal{F}$ is a σ-algebra. To this end, one first observes that

$$X = \emptyset^c = \emptyset^c \cup \emptyset \cup \emptyset \cup \cdots \in \mathcal{E}_1 \subseteq \mathcal{F}.$$

In addition, if $A \in \mathcal{F}$, then $A \in \mathcal{E}_\alpha$ for some $\alpha < \Omega$. Pick some β with $\alpha < \beta < \Omega$. Then one has

$$A^c = A^c \cup A^c \cup A^c \cup \cdots \in \mathcal{E}_\alpha^* \subseteq \mathcal{E}_\beta \subseteq \mathcal{F}.$$

Finally, if $A_n \in \mathcal{F}$, then $A_n \in \mathcal{E}_{\alpha_n}$ for some $\alpha_n < \Omega$. According to Theorem 1.6 (iv), there exists $\beta < \Omega$ such that $\alpha_n < \beta$ for all n. It follows that

$$\bigcup_{n=1}^{\infty} A_n \in \big(\bigcup_{n=1}^{\infty} \mathcal{E}_{\alpha_n}\big)^* \subseteq \mathcal{E}_\beta \subseteq \mathcal{F}.$$

Therefore, $\mathcal{F}$ is a σ-algebra. □

The following corollary of Theorem 1.7 provides some information about the "size" of $\sigma(\mathcal{E})$.

Corollary 1.2. *Let $\mathcal{E}$ be a class of subsets of X containing $\emptyset$. Suppose that $\mathcal{E}$ has cardinality at most continuum. Then $\sigma(\mathcal{E})$ also has cardinality at most continuum.*

Proof. According to Theorem 1.7, it suffices to show that the σ-algebra $\mathcal{F}$ defined by (1.19) has cardinality at most continuum. To this end, one first notes that the set P_Ω defined by (1.20) has continuum cardinality (cf. Theorem 1.6 (ii)). Since a continuum union of continuum is a continuum (cf. the discussion on cardinality below), it remains to show that $\text{card}(\mathcal{E}_\alpha) \leqslant \aleph$ for each ordinal number $\alpha < \Omega$. We again use transfinite induction to prove this. By assumption, one knows that $\mathcal{E}_0 = \mathcal{E}$ has cardinality at most continuum. Now suppose that $0 < \alpha < \Omega$ is an ordinal number and the claim is true for $\mathcal{E}_\beta$ with any $\beta < \alpha$. Since

$$P_\alpha \triangleq \{0 \leqslant \beta < \alpha : \beta \text{ is an ordinal number}\}$$

is a subset of P_Ω, one knows that $\text{card}(P_\alpha) \leqslant \aleph$. In particular, by induction hypothesis one has

$$\text{card}\big(\bigcup_{0 \leqslant \beta < \alpha} \mathcal{E}_\beta\big) \leqslant \aleph,$$

since the union is viewed as a continuum union of continuum. Since $\mathcal{E}_\alpha = (\cup_{\beta<\alpha}\mathcal{E}_\beta)^*$, the induction step will be completed by using the following general property.

Claim. If a class $\mathcal{H}$ has cardinality at most continuum, so does $\mathcal{H}^*$.

To prove the above claim, let $\tilde{\mathcal{H}} \triangleq \mathcal{H} \cup \mathcal{H}^c$ where $\mathcal{H}^c$ is the class obtained by taking complements of members in $\mathcal{H}$. Observe that $\tilde{\mathcal{H}}$ has the same cardinality as $\mathcal{H}$ (i.e. at most continuum). In addition, $\mathcal{H}^*$ is easily identified as a subset of $\tilde{\mathcal{H}}^{\mathbb{N}}$ through

$$\mathcal{H}^* \ni H = \bigcup_{n=1}^{\infty} H_n \mapsto \{H_n : n \geqslant 1\} \in \tilde{\mathcal{H}}^{\mathbb{N}},$$

where the notation A^B means the set of maps from B to A. The claim then follows from the standard fact that $\mathbb{R}^{\mathbb{N}}$ has continuum cardinality; indeed, one has

$$\mathbb{R}^{\mathbb{N}} \approx (\{0,1\}^{\mathbb{N}})^{\mathbb{N}} \approx \{0,1\}^{\mathbb{N}\times\mathbb{N}} \approx \{0,1\}^{\mathbb{N}} \approx \mathbb{R},$$

where $\approx$ means equal in cardinality. □

Remark 1.6. If $\mathcal{E}$ is finite, it is obvious that $\sigma(\mathcal{E})$ is also finite. If $\mathcal{E}$ is countably infinite, then $\sigma(\mathcal{E})$ must be a continuum. Indeed, Theorem 1.7 shows that the cardinality of $\sigma(\mathcal{E})$ is at most continuum. One thus only needs to show that it is at least continuum. This is a consequence of the more general fact that *any infinite σ-algebra $\mathcal{F}$ must have cardinality at least continuum.* To see this, assume on the contrary that $\mathcal{F} = \{A_0, A_1, A_2, \cdots\}$, where all the A_n's are distinct subsets. For each $\{\pm 1\}$-sequence $\alpha = (\alpha_0, \alpha_1, \alpha_2, \cdots) \in \{0,1\}^{\mathbb{N}}$, we define the subset

$$B^\alpha = \bigcup_{n\in\mathbb{N}} A_n^{\alpha_n} \in \mathcal{F},$$

where $A_n^{\alpha_n} = A_n$ or A_n^c depending on whether $\omega_n = 1$ or 0. It is apparent that

$$\alpha \neq \beta \in \{0,1\}^{\mathbb{N}} \implies B^\alpha \cap B^\beta = \emptyset.$$

The next observation is that $\mathcal{B} \triangleq \{B^\alpha : \alpha \in \{0,1\}^{\mathbb{N}}\}$ contains at least countably many elements. Indeed, if there were only finitely many elements in $\mathcal{B}$, according to the simple relation

$$A_n = \bigcup_{\alpha:\alpha_n=1} B^\alpha$$

it is impossible that the A_n's are all distinct. Now pick a sequence $B_0, B_1, B_2, \cdots$ of distinct non-empty members in $\mathcal{B}$. Define a map T from $\{0,1\}^{\mathbb{N}}$ to $\mathcal{F}$ by

$$\alpha = (\alpha_0, \alpha_1, \cdots) \mapsto T(\alpha) \triangleq \bigcup_{n:\alpha_n=1} B_n.$$

One checks that T is injective. Since $\{0,1\}^{\mathbb{N}} = \aleph$, it follows that $\mathcal{F}$ has cardinality at least continuum.

Example 1.8. One knows from the previous remark that $\mathcal{B}(\mathbb{R})$ has continuum cardinality. As a result, in terms of cardinality $\mathcal{B}(\mathbb{R})$ is much smaller that $\mathcal{P}(\mathbb{R})$ and thus there are plenty of subsets of $\mathbb{R}$ that are not Borel measurable.

An application: product σ-algebras over topological spaces

As an application of Theorem 1.7 (indeed of Corollary 1.2), we discuss another peculiar fact about generated σ-algebras: the product σ-algebra may not be equal to the topological σ-algebra for the product of topological spaces.

Definition 1.15. Let X is a topological space. The *Borel σ-algebra* over X, denoted as $\mathcal{B}(X)$, is the σ-algebra generated by open subsets of X.

Suppose that X, Y are topological spaces with Borel σ-algebras $\mathcal{B}(X), \mathcal{B}(Y)$ respectively. There are two apparent notions of σ-algebras over the product space $X \times Y$. On the one hand, from a measure-theoretic perspective one can form the *product σ-algebra* $\mathcal{B}(X) \otimes \mathcal{B}(Y)$. This is the σ-algebra generated by measurable rectangles:

$$\mathcal{B}(X) \otimes \mathcal{B}(Y) = \sigma(\{A \times B : A \in \mathcal{B}(X), B \in \mathcal{B}(Y)\}).$$

On the other hand, since $X \times Y$ is a topological space under the product topology (i.e. the coarsest topology under which the projections $X \times Y \to X$, $X \times Y \to Y$ are both continuous), it also carries a Borel σ-algebra $\mathcal{B}(X \times Y)$ generated by the open subsets of $X \times Y$.

In general, it is obvious that

$$\mathcal{B}(X) \otimes \mathcal{B}(Y) \subseteq \mathcal{B}(X \times Y);$$

for if $A \subseteq X$ and $B \subseteq Y$ are open subsets respectively, then $A \times B$ is open in $X \times Y$. On the other hand, if X and Y are both *separable* metric spaces equipped with the metric topology, the reversed inclusion is also valid. This is because any open subset G of $X \times Y$ is a countable union of open rectangles, i.e.

$$G = \bigcup_{n=1}^{\infty} U_n \times V_n$$

for suitable open subsets $U_n \subseteq X$, $V_n \subseteq Y$. The same conclusion holds for topological spaces with countable base.

The peculiar situation is that, even in the context of metric spaces, the reversed inclusion may fail in general if the spaces are *not separable*. Indeed, one has the following result.

Theorem 1.8. *Let X be a metric space whose cardinality is strictly larger than continuum. Then $\mathcal{B}(X) \otimes \mathcal{B}(X)$ is a proper subset of $\mathcal{B}(X \times X)$. In particular, X is not separable.*

Remark 1.7. An interesting corollary of Theorem 1.8 is that every separable metric space has cardinality at most continuum. This fact can also be proved directly as follows. Let X be a separable metric space with a countable dense subset $\{x_n : n \geqslant 1\}$. The trick is to view the continuum as the power set of $\mathbb{N}$ (the collection of all subsets of natural numbers, denoted as $\mathcal{P}(\mathbb{N})$). Since X is separable, for each $y \in X$, one can select a subsequence $\{x_{k_n}\}$ such that $x_{k_n} \to y$. Viewing $\{k_n\}$ as a subset of natural numbers, one then constructs a map from X to $\mathcal{P}(\mathbb{N})$ by sending y to $\{k_n\}$. It is apparent that this map is injective, for if $y \neq z \in X$, the two associated subsequences of $\{x_n\}$ must be different.

The proof of Theorem 1.8 is an application of Corollary 1.2 along with the aid of the following lemma.

Lemma 1.3. *Let $\mathcal{E}$ be a set class over a sample space X. Then for any $E \in \sigma(\mathcal{E})$, there exists a countable subclass $\mathcal{E}_0$ of $\mathcal{E}$ (possibly depending on E), such that $E \in \sigma(\mathcal{E}_0)$.*

Proof. Let $\mathcal{H}$ be the family of subsets $E \in \sigma(\mathcal{E})$ satisfying the desired property. If $E \in \mathcal{E}$, one has $E \in \sigma(\mathcal{E}_0)$ with $\mathcal{E}_0 \triangleq \{E\}$. Therefore, $\mathcal{E} \subseteq \mathcal{H}$. It remains to show that $\mathcal{H}$ is a σ-algebra. Apparently, $X \in \mathcal{H}$ and $\mathcal{H}$ is closed under complementation. In addition, let $E_n \in \mathcal{H}$ $(n \geqslant 1)$ with associated countable subclass $\mathcal{E}_n \subseteq \mathcal{E}$ so that $E_n \in \sigma(\mathcal{E}_n)$. Then $\cup_n E_n \in \mathcal{H}$ with associated countable subclass $\mathcal{E}_0 \triangleq \cup_n \mathcal{E}_n$. □

Now we give the proof of Theorem 1.8.

Proof of Theorem 1.8. Consider the diagonal subset

$$D = \{(x, y) : x = y\} \subseteq X \times X.$$

Since X is Hausdorff, one knows that D is closed and hence it belongs to $\mathcal{B}(X) \times \mathcal{B}(X)$. We claim that D does not belong to $\mathcal{B}(X) \otimes \mathcal{B}(X)$. Suppose on the contrary that $D \in \mathcal{B}(X) \otimes \mathcal{B}(X)$. According to Lemma 1.3, there exists a countable collection of measurable rectangles

$$\mathcal{E} = \{A_n \times B_n : A_n, B_n \in \mathcal{B}(X)\},$$

such that $D \in \sigma(\mathcal{E})$. Let us define

$$\mathcal{A} \triangleq \big\{A_i, B_j : i, j \geqslant 1\big\}.$$

Then one has

$$D \in \sigma(\mathcal{E}) \subseteq \sigma(\mathcal{A}) \otimes \sigma(\mathcal{A}).$$

It is standard that for each $y \in X$, the section

$$D_y \triangleq \{x \in X : (x, y) \in D\} = \{y\} \in \sigma(\mathcal{A}).$$

As a result, X has cardinality at most of $\sigma(\mathcal{A})$.

On the other hand, since $\mathcal{A}$ is countable, one knows from Corollary 1.2 that $\sigma(\mathcal{A})$ has cardinality at most continuum. Therefore, X has cardinality at most continuum, which contradicts the assumption on the cardinality of X.

□

Example 1.9. A "trivial" example of a metric space whose cardinality is strictly larger than continuum is the following: take X to be the power set of $\mathbb{R}$ and define a metric on X by

$$d(A, B) \triangleq \begin{cases} 1, & \text{if } A \neq B; \\ 0, & \text{if } A = B. \end{cases}$$

2 The mathematical expectation

In this chapter, we study random variables and construct their mathematical expectations (integration). We begin with the more general notion of integration with respect to measures and then specialise in the probabilistic context. We also study the important concept of conditional expectation. This chapter provides the necessary analytic tools for the study of various topics in the subject.

2.1 Measurable functions and random variables

In elementary probability, one has seen the notion of random variables in a semi-rigorous way. Roughly speaking, a random variable is a function X defined on the sample space Ω for which one can talk about probabilities of events like $\{\omega : X(\omega) \in A\}$ whenever $A \subseteq \mathbb{R}$. However, it is unreasonable to allow $\{X \in A\}$ to be legal events for *all subsets* A of $\mathbb{R}$. Measurability properties need to be introduced and the legal events are precisely those $\{X \in A\}$'s with A being Borel measurable subsets of $\mathbb{R}$.

For many purposes, it will be convenient to allow random variables to take $\pm\infty$-values. For instance, let us consider the random experiment of tossing a fair coin repeatedly until forever. The random variable defined by the first time a "head" appears (cf. Example 2.1 below) takes value $+\infty$ at the outcome given by all tails, even though with probability one a "head" appears in finite time. There are also natural random variables which admit $+\infty$-value with positive probability. For instance, suppose that a frog is jumping among integer points on the real line. It starts at the origin $x = 0$, and at each step it jumps one unit to the left with probability $2/3$ and one unit to the right with probability $1/3$. Let X be the first time the frog reaches $x = 1$. It can be shown that $\mathbb{P}(X = +\infty) > 0$. From an analytic viewpoint, it is often needed to consider the supremum or infimum of a sequence of finitely valued functions, which could fail to be finite in general.

We make some standard conventions when working with $\pm\infty$-values. Let $\bar{\mathbb{R}} \triangleq \mathbb{R} \cup \{\pm\infty\}$ denote the extended real line. Define $\mathcal{B}(\bar{\mathbb{R}})$ to be the σ-algebra generated by $\mathcal{B}(\mathbb{R})$ along with $\{-\infty\}$ and $\{+\infty\}$. Proposition 1.1 generalises to $\bar{\mathbb{R}}$ in the obvious way. For instance, $\mathcal{B}(\bar{\mathbb{R}})$ is the σ-algebra generated by the class of subsets of the form $\{x : -\infty \leqslant x \leqslant a\}$ $(a \in \mathbb{R})$. Through the rest, we always adopt obvious conventions like

$$\infty - (-\infty) = \infty,\ (-2) \times \infty = -\infty,\ 0 \cdot \pm\infty = 0,\ \frac{3}{-\infty} = 0,\ 4 - (-\infty) = +\infty,$$

etc. when one performs arithmetic operations on $\bar{\mathbb{R}}$. Expressions like

$$+\infty - (+\infty),\ \frac{0}{0},\ \frac{2}{0},\ \frac{\pm\infty}{\pm\infty}$$

will be considered *not* well-defined.

For conventional reason, we save the term *random variables* for $\mathbb{R}$-valued functions and use *measurable functions* for the $\bar{\mathbb{R}}$-valued case.

Definition 2.1. Let $(\Omega, \mathcal{F})$ be a measurable space. A *measurable function* on Ω is a map $X : \Omega \to \bar{\mathbb{R}}$ such that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}(\bar{\mathbb{R}})$. A *random variable* on Ω is a $\mathbb{R}$-valued measurable function.

By definition, for a measurable function X, subsets like $\{\omega \in \Omega : a < X(\omega) < b\}$ or $\{\omega \in \Omega : X(\omega) = c\}$ are events in $\mathcal{F}$. The result below gives a simple way of checking the required measurability property. To ease notation, for any $A \subseteq \bar{\mathbb{R}}$ we will use $\{X \in A\}$ to denote the subset $\{\omega \in \Omega : X(\omega) \in A\}$.

Lemma 2.1. *The following statements are equivalent:*

(i) $X : \Omega \to \bar{\mathbb{R}}$ is a measurable function;
(ii) for any $a \in \mathbb{R}$, $\{X < a\} \in \mathcal{F}$;
(iii) for any $a \in \mathbb{R}$, $\{X \leqslant a\} \in \mathcal{F}$;
(iv) for any $a \in \mathbb{R}$, $\{X > a\} \in \mathcal{F}$;
(v) for any $a \in \mathbb{R}$, $\{X \geqslant a\} \in \mathcal{F}$.

Proof. Denote $\bar{\mathcal{C}} \triangleq \{[-\infty, a) : a \in \mathbb{R}\}$. Then $\mathcal{B}(\bar{\mathbb{R}}) = \sigma(\bar{\mathcal{C}})$. The equivalence between (i) and (ii) is a consequence of the fact that

$$X^{-1}(\mathcal{B}(\bar{\mathbb{R}})) = X^{-1}(\sigma(\bar{\mathcal{C}})) = \sigma\left(X^{-1}(\bar{\mathcal{C}})\right) \subseteq \mathcal{F}.$$

Their equivalence to (iii)–(v) follows from a similar reason. □

It is useful to express a measurable function as the difference of two non-negative measurable functions. In this way, one can often reduce the study to the non-negative case.

Definition 2.2. Let X be a measurable function on Ω. The functions

$$X^+ \triangleq \max\{X, 0\},\ X^- \triangleq \max\{-X, 0\}$$

are called the *positive* and *negative parts* of X respectively.

From the definition, it is clear that

$$X = X^+ - X^-,\ |X| = X^+ + X^-.$$

The relation $\{X^+ > a\} = \{X > a\} \in \mathcal{F}$ $(a \geqslant 0)$ implies that X^+ is a measurable function. Similarly, X^- is also a measurable function.

One can form new measurable functions from given ones by elementary operations.

Proposition 2.1. *Let X, Y, X_n $(n \geqslant 1)$ be measurable functions on Ω. Whenever the following objects are well-defined, they are all measurable functions:*

$$X \pm Y,\ XY,\ \frac{X}{Y},\ \inf_{n\geqslant 1} X_n,\ \sup_{n\geqslant 1} X_n,\ \varliminf_{n\to\infty} X_n,\ \varlimsup_{n\to\infty} X_n.$$

Proof. For simplicity we only consider the case when X, Y are random variables (i.e. $\mathbb{R}$-valued).

(i) $X + Y$ is a random variable:

$$\{X + Y < a\} = \bigcup_{r\in\mathbb{Q}} \big(\{X < r\} \cap \{Y < a - r\}\big) \in \mathcal{F}.$$

A similar argument shows that $X - Y$ is also a random variable.

(ii) For the XY case, we first assume that X, Y are both non-negative. In this case,

$$\begin{aligned}\{XY < a\} = \{X = 0\} \cup \{Y = 0\} \\ \cup \Big(\bigcup_{r\in\mathbb{Q}_+} \big(\{0 < X < r\} \cap \{0 < Y < \frac{a}{r}\}\big)\Big) \in \mathcal{F}.\end{aligned}$$

The general case follows from the decomposition

$$XY = (X^+ - X^-)(Y^+ - Y^-) = X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-,$$

which is a measurable function as a consequence of Part (i) and the non-negative case.

(iii) For the $\frac{X}{Y}$ case, from Part (ii) it is enough to consider $\frac{1}{Y}$. If $a > 0$, one has

$$\Big\{\frac{1}{Y} < a\Big\} = \{Y < 0\} \cup \Big\{Y > \frac{1}{a}\Big\} \in \mathcal{F}.$$

The case when $a < 0$ is treated similarly.

(iv) The following relations

$$\{\inf_{n\geqslant 1} X_n < a\} = \bigcup_{n=1}^{\infty}\{X_n < a\},\ \{\sup_{n\geqslant 1} X_n \leqslant a\} = \bigcap_{n=1}^{\infty}\{X_n \leqslant a\}$$

imply that $\inf_n X_n$ and $\sup_n X_n$ are both measurable functions. It follows that

$$\underline{\lim}_{n\to\infty} X_n = \sup_{n\geqslant 1}\inf_{m\geqslant n} X_n,\ \overline{\lim}_{n\to\infty} X_n = \inf_{n\geqslant 1}\sup_{m\geqslant n} X_n$$

are also measurable functions. □

Example 2.1. Consider the example of tossing a fair coin repeatedly in a sequence. Recall that the sample space Ω and the σ-algebra $\mathcal{F}$ are defined by (1.8) and (1.9) respectively. Let X be the first time that a "head" appears. Mathematically,

$$X(\omega) \triangleq \inf\{n \geqslant 1 : \omega_n = H\}, \quad \omega = (\omega_1, \omega_2, \cdots) \in \Omega.$$

Then X is a measurable function on Ω. Indeed, note that X takes values in $\mathbb{Z} \cup \{+\infty\}$. Moreover, for any $n \geqslant 1$ one has

$$\{\omega : X(\omega) = n\} = \{\omega : \omega_1 = \cdots = \omega_{n-1} = T, \omega_n = H\} \in \mathcal{F}.$$

From this fact it is not hard to verify any of the conditions in Lemma 2.1.

2.2 Integration with respect to measures

A basic numerical feature of a random variable is its expectation with respect to a probability measure. This is essentially the notion of integration. At this point, the probabilistic structure is of no significance yet and it is advantageous to begin with general measures as well as measurable functions.

Let $(\Omega, \mathcal{F}, \mu)$ be a given measure space. Let X be a measurable function on Ω. We want to define the notion of "the integral of X with respect to the measure μ". The idea of constructing this integral $\int_\Omega X d\mu$ is very natural. In the situation when X is the *indicator function* associated with some event $A \in \mathcal{F}$, i.e. if

$$X(\omega) = \mathbf{1}_A(\omega) \triangleq \begin{cases} 1, & \omega \in A; \\ 0, & \omega \notin A, \end{cases}$$

the integral $\int_\Omega X d\mu$ should just be defined as $\mu(A)$. In the probabilistic context when μ is a probability measure, this X is a Bernoulli random variable and its

expectation is just the probability of A. Since the integration map $X \mapsto \int_\Omega X d\mu$ should be *linear*, one knows immediately how the integral of a linear combination of indicator functions (*simple* functions) should be defined. To extend the construction to general measurable functions, one needs a key lemma for approximating measurable functions by simple functions as well as a procedure of passing to the limit. It is convenient to first treat the case of non-negative functions, since the general case can be dealt with using the decomposition $X = X^+ - X^-$.

We start with the following definition.

Definition 2.3. A *non-negative, simple, measurable function* on Ω is a linear combination of indicator functions, i.e.

$$X(\omega) = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}(\omega) \tag{2.1}$$

for some $n \geqslant 1$, $a_i \in [0, +\infty)$ and $A_i \in \mathcal{F}$ $(1 \leqslant i \leqslant n)$. The space of non-negative, simple, measurable functions is denoted as $\mathcal{S}^+$.

It is not hard to see that a non-negative measurable function X is simple if and only if it takes finitely many values in $[0, +\infty)$. This observation immediately tells us that

$$X, Y \in \mathcal{S}^+ \implies X + Y,\ XY,\ \max(X, Y),\ \min(X, Y),\ \alpha X\ (\alpha > 0) \in \mathcal{S}^+.$$

In addition, if $X \geqslant Y$ then $X - Y \in \mathcal{S}^+$.

Note that the representation (1.6) of $X \in \mathcal{S}^+$ may not be unique. For instance, on $\Omega = [0, 1]$,

$$X = \mathbf{1}_{[0,2/3]} = \mathbf{1}_{[0,1/3]} + \mathbf{1}_{(1/3,2/3]}$$

are two different representations of the same function. Among all representations, there is one in which the events A_i's form a finite *partition* of Ω. Such kind of representations is more convenient to work with.

Lemma 2.2. *Let $X \in \mathcal{S}^+$ be a non-negative simple measurable function. Then X can be written as $X = \sum_{j=1}^m b_j \mathbf{1}_{B_j}$ where $B_j \in \mathcal{F}$ $(1 \leqslant j \leqslant m)$ form a partition of Ω.*

Proof. Suppose that $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i} \in \mathcal{S}^+$. For each $0 \leqslant i \leqslant n$, we set $C_0^{(i)} \triangleq A_i^c$ and $C_1^{(i)} \triangleq A_i$. Let

$$\mathcal{J} = \{(j_1, \cdots, j_n) : j_i = 0 \text{ or } 1\}$$

denote the set of 0-1 sequences of length n. Given $J = (j_1, \cdots, j_n) \in \mathcal{J}$, define

$$B_J = C_{j_1}^{(1)} \cap C_{j_2}^{(2)} \cap \cdots \cap C_{j_n}^{(n)}$$

and

$$b_J \triangleq \sum_{1 \leqslant i \leqslant n : j_i = 1} a_i \quad (b_J \triangleq 0 \text{ if } J = \{0, 0, \cdots, 0\}).$$

Apparently, $\{B_J : J \in \mathcal{J}\}$ is a (finite) partition of Ω and one has

$$X = \sum_{J \in \mathcal{J}} b_J B_J. \tag{2.2}$$

□

Remark 2.1. Despite of the above seemingly complicated notation, the underlying idea is quite simple. In the case when $n = 2$, one has

$$\begin{aligned} X &= a_1 \mathbf{1}_{A_1} + a_2 \mathbf{1}_{A_2} \\ &= 0 \cdot \mathbf{1}_{A_1^c \cap A_2^c} + a_1 \cdot \mathbf{1}_{A_1 \cap A_2^c} + a_2 \cdot \mathbf{1}_{A_1^c \cap A_2} + (a_1 + a_2) \cdot \mathbf{1}_{A_1 \cap A_2}. \end{aligned}$$

The definition of integration for non-negative simple functions is obvious.

Definition 2.4. For $X \in \mathcal{S}^+$ with given representation (2.1), the *integral* of X with respect to the measure μ is defined by

$$\int_\Omega X d\mu \triangleq \sum_{i=1}^n a_i \mu(A_i) \in [0, +\infty].$$

Before studying basic properties of this integral, one must first show that it is well-defined, namely it is independent of the representation of X. The proof is quite technical and one should not bother with the not-so-inspiring details if one is convinced by drawing simple pictures.

Proposition 2.2. *The integral $\int_\Omega X d\mu$ is well-defined for $X \in \mathcal{S}^+$.*

Proof. Suppose that $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$. We first claim that, the quantity $\int_\Omega X d\mu$ remains the same if one uses the representation of X given by (2.2) over a finite partition of Ω. Indeed, under the same notation as in the proof of Lemma 2.2, for each $1 \leqslant i \leqslant n$, A_i admits the following disjoint decomposition:

$$A_i = \bigcup_{J = (j_1, \cdots, j_n) \in \mathcal{J} : j_i = 1} B_J.$$

According to the finite additivity of μ, one has

$$\begin{aligned}\sum_{i=1}^{n} a_i\mu(A_i) &= \sum_{i=1}^{n} a_i \sum_{J=(j_1,\cdots,j_n)\in\mathcal{J}:j_i=1} \mu(B_J)\\ &= \sum_{J=(j_1,\cdots,j_n)\in\mathcal{J}} \sum_{1\leqslant i\leqslant n:j_i=1} a_i\mu(B_J)\\ &= \sum_{J\in\mathcal{J}} b_J\mu(B_J).\end{aligned}$$

Therefore, the claim holds.

Since any representation of X can always be reduced to one over a finite partition, it remains to show that the quantity $\int_\Omega X d\mu$ is invariant when X admits two representations

$$X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i} = \sum_{j=1}^{m} b_j \mathbf{1}_{B_j}$$

with $\{A_1, \cdots, A_n\}$ and $\{B_1, \cdots, B_m\}$ being two partitions of Ω. But this fact is straight forward to see:

$$\begin{aligned}\sum_{i=1}^{n} a_i\mu(A_i) &= \sum_{i=1}^{n} a_i\mu\big(A_i \cap \big(\bigcup_{j=1}^{m} B_j\big)\big) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i\mu(A_i\cap B_j)\\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} b_j\mu(A_i\cap B_j) = \sum_{j=1}^{m} b_j\mu(B_j),\end{aligned}$$

since $a_i = b_j$ on $A_i \cap B_j$. □

The definition of $\int_\Omega X d\mu$ ($X \in \mathcal{S}^+$) automatically ensures its *linearity*: if $X, Y \in \mathcal{S}^+$ and $\alpha \geqslant 0$, then

$$\int_\Omega (X+Y)d\mu = \int_\Omega X d\mu + \int_\Omega Y d\mu,\ \int_\Omega \alpha X d\mu = \alpha\int_\Omega X d\mu. \tag{2.3}$$

Another basic property which is less obvious is *monotonicity*: if $X, Y \in \mathcal{S}^+$ and $X \leqslant Y$, then

$$\int_\Omega X d\mu \leqslant \int_\Omega Y d\mu. \tag{2.4}$$

Indeed, if one expresses X, Y by

$$X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i},\ Y = \sum_{j=1}^{m} b_j \mathbf{1}_{B_j}$$

where $\{A_1, \cdots, A_n\}$ and $\{B_1, \cdots, B_m\}$ are two partitions of Ω respectively, then

$$X = \sum_{i,j} a_i \mathbf{1}_{A_i \cap B_j}, \ Y = \sum_{i,j} b_j \mathbf{1}_{A_i \cap B_j}$$

with $a_i \leqslant b_j$ on $A_i \cap B_j$ (since $X \leqslant Y$). Therefore,

$$\int_\Omega X d\mu = \sum_{i,j} a_i \mu(A_i \cap B_j) \leqslant \sum_{i,j} b_i \mu(A_i \cap B_j) \leqslant \int_\Omega Y d\mu.$$

Next, we establish a property of the integral $\int_\Omega X d\mu$ ($X \in \mathcal{S}^+$) which is crucial for extending the construction to general measurable functions. This is also the preliminary version of the more general *monotone convergence theorem* (cf. Theorem 2.2 below). We use the notation $X_n \uparrow X$ to mean that $X_n(\omega)$ increases to $X(\omega)$ as $n \to \infty$ for every $\omega \in \Omega$.

Proposition 2.3. *Let $X_n, X \in \mathcal{S}^+$ ($n \geqslant 1$) and $X_n \uparrow X$. Then $\int_\Omega X_n d\mu \uparrow \int_\Omega X d\mu$.*

Proof. We first consider the case when $X = \mathbf{1}_A$. Suppose that $X_n \uparrow \mathbf{1}_A$. Since

$$\int_\Omega X_n d\mu \leqslant \int_\Omega \mathbf{1}_A d\mu = \mu(A),$$

one has

$$\overline{\lim_{n\to\infty}} \int_\Omega X_n d\mu \leqslant \mu(A). \tag{2.5}$$

To obtain the matching lower estimate, let $\varepsilon > 0$ and define

$$A_n \triangleq \{\omega \in A : X_n \geqslant 1 - \varepsilon\}.$$

Since $X_n \uparrow \mathbf{1}_A$, on has $A_n \uparrow A$ and thus $\mu(A_n) \uparrow \mu(A)$. But from the definition of A_n one also knows that

$$X_n \geqslant (1 - \varepsilon) \cdot \mathbf{1}_{A_n}.$$

Therefore,

$$\int_\Omega X_n d\mu \geqslant (1 - \varepsilon) \cdot \mu(A_n),$$

which implies that

$$\underline{\lim_{n\to\infty}} \int_\Omega X_n d\mu \geqslant (1 - \varepsilon) \cdot \mu(A). \tag{2.6}$$

Since ε is arbitrary, by letting $\varepsilon \downarrow 0$ in (2.6) and together with (2.5) one concludes that the limit of $\int_\Omega X_n d\mu$ is equal to $\mu(A)$.

For the general case, suppose that $X_n \uparrow X$ where $X = \sum_{i=1}^m a_i \mathbf{1}_{A_i}$ with $a_i \geqslant 0$ and the A_i's form a partition of Ω. For those i's such that $a_i > 0$, by the convergence assumption $X_n \uparrow X$ one has

$$\mathcal{S}^+ \ni Y_n^{(i)} \triangleq \frac{1}{a_i} \mathbf{1}_{A_i} \cdot X_n \uparrow \mathbf{1}_{A_i}.$$

It follows from the case we have just treated that

$$\lim_{n\to\infty} \int_\Omega Y_n^{(i)} d\mu = \mu(A_i).$$

As a result,

$$\lim_{n\to\infty} \int_\Omega X_n \mathbf{1}_{A_i} d\mu = \lim_{n\to\infty} \int_\Omega a_i Y_n^{(i)} d\mu = a_i \mu(A_i).$$

Note that the above relation remains valid if $a_i = 0$ (since $0 \leqslant X_n \leqslant X$, if $a_i = 0$ then $X = 0$ on A_i and so is X_n, which shows that $X_n \mathbf{1}_{A_i} = 0$ in this case). Since $\{A_1, \cdots, A_m\}$ is a partition of Ω, one knows that

$$\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_m} = 1.$$

Therefore, one has

$$\int_\Omega X_n d\mu = \int_\Omega X_n \big(\sum_{i=1}^m \mathbf{1}_{A_i} \big) d\mu \overset{n\to\infty}{\longrightarrow} \sum_{i=1}^m a_i \mu(A_i) = \int_\Omega X d\mu.$$

□

Finally, we move to the real stuff: extending the construction to general measurable functions. We first consider the case of non-negative measurable functions; the general case follows easily from the decomposition $X = X^+ - X^-$. The idea is to approximate a non-negative measurable function by a sequence of non-negative simple functions and show that the result sequence of integrals converges accordingly. The following result is the key lemma to implement this idea. It is also useful in many other situations.

Lemma 2.3. *Let X be a non-negative measurable function on Ω. Then there exists a sequence $\{X_n : n \geqslant 1\}$ of non-negative simple measurable functions such that $X_n \uparrow X$.*

Proof. Given $n \geqslant 1$, define $X_n : \Omega \to [0, \infty)$ by

$$X_n(\omega) \triangleq \begin{cases} n, & \text{if } X(\omega) \geqslant n; \\ \frac{k}{2^n}, & \text{if } \frac{k}{2^n} \leqslant X(\omega) < \frac{k+1}{2^n} \text{ for some } 0 \leqslant k \leqslant n2^n - 1, \end{cases}$$

or in more compact form,

$$X_n(\omega) \triangleq \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\{k/2^n \leqslant X < (k+1)/2^n\}} + n\mathbf{1}_{\{X \geqslant n\}}.$$

It is obvious that $X_n \in \mathcal{S}^+$ and we let the reader check that $X_n \uparrow X$. □

Remark 2.2. If X is bounded, then X_n converges to X uniformly. Indeed, suppose that $0 \leqslant X \leqslant M$ for some constant $M > 0$. From the above proof, one sees that

$$0 \leqslant X(\omega) - X_n(\omega) \leqslant \frac{1}{2^n} \quad \forall \omega,$$

provided $n > M$ since in this case one always has $X(\omega) < n$.

With the aid of Lemma 2.3, one can now define the integral of a non-negative measurable function.

Definition 2.5. Let $X : \Omega \to [0, \infty]$ be a non-negative measurable function. The *integral* of X with respect to the measure μ is defined by

$$\int_\Omega X d\mu \triangleq \lim_{n\to\infty} \int_\Omega X_n d\mu,$$

where $X_n \in \mathcal{S}^+$ is any sequence such that $X_n \uparrow X$.

Just like the case for simple functions, one must show that $\int_\Omega X d\mu$ is well-defined.

Proposition 2.4. *For any non-negative measurable function X, the integral $\int_\Omega X d\mu$ is well-defined and it is non-negative.*

Proof. If $\mathcal{S}^+ \ni X_n \uparrow X$, one knows from the monotonicity property (2.4) that $\int_\Omega X_n d\mu$ is increasing. Therefore, its limit exists in $[0, \infty]$. Suppose that $X_n, Y_n \in \mathcal{S}^+$ are two sequences both satisfying $X_n \uparrow X$ and $Y_n \uparrow X$. We want to show that

$$\lim_{n\to\infty} \int_\Omega X_n d\mu = \lim_{n\to\infty} \int_\Omega Y_n d\mu. \tag{2.7}$$

To this end, let us fix $m \geqslant 1$ for now and consider the sequence $Z_n \triangleq \min\{X_n, Y_m\}$ $(n \geqslant 1)$. It is clear that $Z_n \in \mathcal{S}^+$ and $Z_n \uparrow \min\{X, Y_m\} = Y_m$. According to Proposition 2.3 (the monotone convergence theorem for simple functions), one knows that

$$\int_\Omega Y_m d\mu = \lim_{n\to\infty} \int_\Omega Z_n d\mu \leqslant \lim_{n\to\infty} \int_\Omega X_n d\mu,$$

where the inequality follows from $Z_n \leqslant X_n$. By taking $m \to \infty$ on the left hand side, one concludes that

$$\lim_{m\to\infty} \int_\Omega Y_m d\mu \leqslant \lim_{n\to\infty} \int_\Omega X_n d\mu.$$

By symmetry the reverse inequality also holds. Therefore, the relation (2.7) follows. □

Finally, we are able to give the construction of integration in full generality. Let $X : \Omega \to \bar{\mathbb{R}}$ be a measurable function on Ω. Recall that one can always express X as the difference between its positive and negative parts:

$$X = X^+ - X^-, \quad X^+ \triangleq \max\{X, 0\}, \ X^- \triangleq \max\{-X, 0\}.$$

Definition 2.6. We say that the integral of X with respect to μ *exists*, if at least one of $\int_\Omega X^+ d\mu$ and $\int_\Omega X^- d\mu$ is finite. In this case, one defines

$$\int_\Omega X d\mu \triangleq \int_\Omega X^+ d\mu - \int_\Omega X^- d\mu.$$

If both of $\int_\Omega X^+ d\mu$ and $\int_\Omega X^- d\mu$ are finite, we say that X is *integrable* and simply write $X \in L^1$.

Remark 2.3. Since $|X| = X^+ + X^-$, from definition X is integrable if and only if $|X|$ is integrable. Regardless of integrability it is always true that $\int_\Omega |X| d\mu = \int_\Omega X^+ d\mu + \int_\Omega X^- d\mu$.

Remark 2.4. It is also common to write the integral as $\int_\Omega X(\omega)\mu(d\omega)$ to indicate the hidden variable ω.

An important property of integration is *linearity*. Indeed, it has guided us in the construction of the integral.

Theorem 2.1. *Let $X, Y \in L^1$ and $\alpha \in \mathbb{R}$. Then $X + Y, \alpha X \in L^1$ and*

$$\int_\Omega (X + Y) d\mu = \int_\Omega X d\mu + \int_\Omega Y d\mu, \quad \int_\Omega \alpha X d\mu = \alpha \int_\Omega X d\mu.$$

Proof. First of all, note that

$$(X^+ - X^-) + (Y^+ - Y^-) = X + Y = (X+Y)^+ - (X+Y)^-,$$

or equivalently

$$X^- + Y^- + (X+Y)^+ = X^+ + Y^+ + (X+Y)^-. \tag{2.8}$$

Next, in the non-negative case, the linearity property of the integral is a consequence of approximation by functions in $\mathcal{S}^+$ as well as the linearity property (2.3) in the case of simple functions. Therefore, by integrating (2.8) one obtains that

$$\begin{aligned}&\int_\Omega X^- d\mu + \int_\Omega Y^- d\mu + \int_\Omega (X+Y)^+ d\mu \\ &= \int_\Omega X^+ d\mu + \int_\Omega Y^+ d\mu + \int_\Omega (X+Y)^- d\mu.\end{aligned}$$

Now the first assertion follows from rearrangement of the terms and Definition 2.6 of the integral. The second assertion is proved in a similar way by decomposition into positive and negative parts. □

Sometimes one needs to consider the integral of X over a measurable subset $A \in \mathcal{F}$ instead of the whole space Ω. Let X be an integrable function. Then for any $A \in \mathcal{F}$, the function $X\mathbf{1}_A$ is integrable (since $|X\mathbf{1}_A| \leqslant |X|$). The integral $\int_\Omega X\mathbf{1}_A d\mu$ is called the *integral of X over A* and it is denoted as $\int_A X d\mu$. By linearity, it is easily seen that

$$\int_{A\cup B} X d\mu = \int_\Omega X\mathbf{1}_{A\cup B} d\mu = \int_\Omega X(\mathbf{1}_A + \mathbf{1}_B) d\mu = \int_A X d\mu + \int_B X d\mu \tag{2.9}$$

for any $A, B \in \mathcal{F}$ with $A \cap B = \emptyset$. This property is known as the *additivity* of integration.

Before stating other properties of the integral, we make a note that the integral $\int_\Omega X d\mu$ is invariant under changing the values of X on a μ-null set.

Proposition 2.5. *Let X be a given integrable function. Then $\int_N X d\mu = 0$ for any μ-null set N. As a consequence, suppose that Y is another measurable function such that $X = Y$ a.e. Then Y is also integrable and $\int_\Omega X d\mu = \int_\Omega Y d\mu$.*

Proof. For the first assertion, it is enough to consider the case when $X \geqslant 0$. Let N be a given μ-null set. Choose a sequence of simple functions $X_n \in \mathcal{S}^+$

such that $X_n \uparrow X$. Then $\mathcal{S}^+ \ni X_n \mathbf{1}_N \uparrow X\mathbf{1}_N$. Since X_n has a general form of $X_n = \sum_{i=1}^m a_i \mathbf{1}_{A_i}$ and $\mu(N) = 0$, it is clear that

$$\int_\Omega X_n \mathbf{1}_N d\mu = \sum_{i=1}^m a_i \mu(A_i \cap N) = 0.$$

By the definition of the integral, one has

$$\int_\Omega X\mathbf{1}_N d\mu = \lim_{n\to\infty} \int_\Omega X_n \mathbf{1}_N d\mu = 0.$$

For the second assertion, let N be a μ-null set such that $X(\omega) = Y(\omega)$ when $\omega \in N^c$. Then one has $X\mathbf{1}_{N^c} = Y\mathbf{1}_{N^c}$ on Ω . It follows that

$$\int_{N^c} X d\mu = \int_\Omega X\mathbf{1}_{N^c} d\mu = \int_\Omega Y\mathbf{1}_{N^c} d\mu = \int_{N^c} Y d\mu.$$

According to the additivity property (2.9) and the first assertion, one has

$$\int_\Omega X d\mu = \int_N X d\mu + \int_{N^c} X d\mu = \int_\Omega X\mathbf{1}_{N^c} d\mu = \int_\Omega Y\mathbf{1}_{N^c} d\mu = \int_\Omega Y d\mu.$$

□

Proposition 2.6. *Let X, Y be integrable functions. Then one has the following properties.*

(i) Triangle inequality: $\left|\int_\Omega X d\mu\right| \leqslant \int_\Omega |X| d\mu$.
(ii) Monotonicity: if $X \leqslant Y$ a.e. then $\int_\Omega X d\mu \leqslant \int_\Omega Y d\mu$.

Proof. (i) This follows from the trivial inequality

$$-\int_\Omega X^+ d\mu - \int_\Omega X^- d\mu \leqslant \int_\Omega X^+ d\mu - \int_\Omega X^- d\mu \leqslant \int_\Omega X^+ d\mu + \int_\Omega X^- d\mu.$$

(ii) Because of Proposition 2.5, one may assume that $X \leqslant Y$ on Ω. By the definition of the positive and negative parts, it is not hard to see that

$$X^+ \leqslant Y^+, \; Y^- \leqslant X^-. \tag{2.10}$$

If one can prove monotonicity of the integral for non-negative functions, the result will follow from integrating (2.10) as well as the definition of the integral. To treat

the non-negative case, suppose that $0 \leqslant X \leqslant Y$ and choose $X_n, Y_n \in \mathcal{S}^+$ so that $X_n \uparrow X, Y_n \uparrow Y$. Then

$$\mathcal{S}^+ \ni Z_n \triangleq \max\{X_n, Y_n\} \uparrow \max\{X, Y\} = Y.$$

It follows that

$$\int_\Omega X d\mu = \lim_{n\to\infty} \int_\Omega X_n d\mu \leqslant \lim_{n\to\infty} \int_\Omega Z_n d\mu = \int_\Omega Y d\mu.$$

□

In some cases, one can deduce properties of the integrand from the knowledge of the integral.

Proposition 2.7. *(i) Let $X \geqslant 0$ a.e. Then*

$$\int_\Omega X d\mu = 0 \iff X = 0 \quad \text{a.e.} \tag{2.11}$$

and

$$\int_\Omega X d\mu < \infty \implies X < \infty \quad \text{a.e.}$$

(ii) Let $X, Y \in L^1$. If

$$\int_A X d\mu \leqslant \int_A Y d\mu \quad \textit{for all } A \in \mathcal{F},$$

then $X \leqslant Y$ a.e.

Proof. (i) For the first assertion, the sufficiency part is trivial. For the necessity part, for each $n \geqslant 1$ consider $X_n \triangleq n^{-1} \cdot \mathbf{1}_{\{X \geqslant n^{-1}\}}$. It is obvious that $X_n \leqslant X$. As a result,

$$n^{-1}\mu(X \geqslant n^{-1}) = \int_\Omega X_n d\mu \leqslant \int_\Omega X d\mu = 0.$$

It follows that

$$\mu(X > 0) = \lim_{n\to\infty} \mu(X \geqslant n^{-1}) = 0.$$

The second assertion is proved by a similar argument. For each $n \geqslant 1$, one has $\mathbf{1}_{\{X \geqslant n\}} \leqslant X/n$ and thus

$$\mu(X \geqslant n) = \int_\Omega \mathbf{1}_{\{X \geqslant n\}} d\mu \leqslant \frac{1}{n} \int_\Omega X d\mu.$$

Therefore,

$$\mu(X = \infty) = \lim_{n\to\infty} \mu(X \geqslant n) = 0.$$

(ii) Choose $A \triangleq \{X > Y\}$. From the assumption and linearity, one knows that $\int_\Omega (X-Y)\mathbf{1}_{\{X>Y\}} d\mu \leqslant 0$. On the other hand, it is apparent that $(X-Y)\mathbf{1}_{\{X>Y\}} \geqslant 0$. Hence $\int_\Omega (X-Y)\mathbf{1}_{\{X>Y\}} d\mu \geqslant 0$. It follows that

$$\int_\Omega (X-Y)\mathbf{1}_{\{X>Y\}} d\mu = 0.$$

According to (2.11), one concludes that

$$Z \triangleq (X-Y)\mathbf{1}_{\{X>Y\}} = 0 \quad \text{a.e.}$$

or equivalently, $\mu(Z \neq 0) = 0$. But $\{X > Y\} \subseteq \{Z \neq 0\}$. Therefore, $\mu(X > Y) = 0$. □

Remark 2.5. As a consequence of Proposition 2.7 (ii), if $\int_A X d\mu = \int_A Y d\mu$ for all $A \in \mathcal{F}$, then $X = Y$ a.e. This is useful e.g. when establishing uniqueness properties for densities or conditional expectations as we will see later on.

2.3 Taking limit under the integral sign

It is often useful to know under what conditions one can interchange the limit and integral signs:

$$X_n \to X \overset{?}{\Longrightarrow} \int_\Omega X_n d\mu \to \int_\Omega X d\mu.$$

There are three basic results of this kind, which may apply in different situations. The first one is known as the *monotone convergence theorem*. It is concerned with non-negative sequences.

Theorem 2.2. *Let $X_n, X \geqslant 0$ a.e. and $X_n \uparrow X$ a.e. Then one has $\int_\Omega X_n d\mu \uparrow \int_\Omega X d\mu$.*

Proof. First of all, since the integrals will not change under modification of X_n and X on μ-null sets, one may assume without loss of generality that the given a.e. properties hold for every $\omega \in \Omega$. Next, for each fixed n, let $Y_{n,m}$ $(m \geqslant 1)$ be a sequence in $\mathcal{S}^+$ such that $Y_{n,m} \uparrow X_n$ as $m \to \infty$. Define

$$Z_m \triangleq \max\{Y_{1,m}, Y_{2,m}, \cdots, Y_{m,m}\} \in \mathcal{S}^+, \quad m \geqslant 1.$$

It is obvious that $Z_m \leqslant Z_{m+1}$. We claim that $Z_m \uparrow X$. Indeed, for fixed n, when $m > n$ one has $Z_m \geqslant Y_{n,m}$. Letting $m \to \infty$ yields

$$\lim_{m\to\infty} Z_m \geqslant \lim_{m\to\infty} Y_{n,m} = X_n.$$

Since this is true for all n, by taking $n \to \infty$ one obtains that

$$\lim_{m\to\infty} Z_m \geqslant \lim_{n\to\infty} X_n = X.$$

The reverse inequality is trivial as $Z_m \leqslant X$ for all m. Therefore, $Z_m \uparrow X$. It follows from the definition of $\int_\Omega X d\mu$ in the non-negative case that

$$\lim_{m\to\infty} \int_\Omega Z_m d\mu = \int_\Omega X d\mu.$$

On the other hand, since $Y_{n,m} \leqslant X_n \leqslant X_m$ whenever $n \leqslant m$, it is clear that $Z_m \leqslant X_m$. Consequently, one has

$$\lim_{m\to\infty} \int_\Omega X_m d\mu \geqslant \lim_{m\to\infty} \int_\Omega Z_m d\mu = \int_\Omega X d\mu.$$

The reverse inequality is trivial, thus finishing the proof of the theorem. □

Corollary 2.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of measurable functions such that $X_n \geqslant 0$ a.e. for each n. Then one has*

$$\int_\Omega \big(\sum_{n=1}^{\infty} X_n\big) d\mu = \sum_{n=1}^{\infty} \int_\Omega X_n d\mu.$$

In particular, if $\sum_{n=1}^{\infty} \int_\Omega X_n d\mu < \infty$, then $\sum_{n=1}^{\infty} X_n < \infty$ a.e. and $X_n \to 0$ a.e.

Proof. Define $S_n \triangleq \sum_{k=1}^{n} X_k$. Then $S_n \geqslant 0$ and $S_n \uparrow \sum_{n=1}^{\infty} X_n$. The result follows from the monotone convergence theorem. □

The second result is known as *Fatou's lemma*. It is also concerned with non-negative sequences and is more flexible to use.

Theorem 2.3. *Let $\{X_n : n \geqslant 1\}$ be a sequence of measurable functions such that $X_n \geqslant 0$ a.e. Then*

$$\int_\Omega \underline{\lim}_{n\to\infty} X_n d\mu \leqslant \underline{\lim}_{n\to\infty} \int_\Omega X_n d\mu.$$

Proof. Define $Y_n \triangleq \inf_{m \geqslant n} X_m$. It is clear that $Y_n \geqslant 0$ a.e. and from the definition of "$\underline{\lim}$" one also knows that

$$Y_n \uparrow X \triangleq \varliminf_{n\to\infty} X_n.$$

According to the monotone convergence theorem, one concludes that

$$\int_\Omega X d\mu = \lim_{n\to\infty} \int_\Omega Y_n d\mu \leqslant \varliminf_{n\to\infty} \int_\Omega X_n d\mu,$$

where the last inequality follows from the fact that $Y_n \leqslant X_n$ for each n. □

Corollary 2.2. *Let X_n, X, Y be measurable functions. Suppose that $0 \leqslant X_n \leqslant Y$, $X_n \to X$ a.e. and $Y \in L^1$. Then $X \in L^1$ and one has*

$$\lim_{n\to\infty} \int_\Omega X_n d\mu = \int_\Omega X d\mu. \tag{2.12}$$

Proof. Since $X_n \geqslant 0$ a.e. for all n, so is X. According to Fatou's lemma, one has

$$0 \leqslant \int_\Omega X d\mu = \int_\Omega \varliminf_{n\to\infty} X_n d\mu \leqslant \varliminf_{n\to\infty} \int_\Omega X_n d\mu \leqslant \int_\Omega Y d\mu < \infty. \tag{2.13}$$

In particular, X is integrable. To prove the convergence result, one applies Fatou's lemma to the a.e. non-negative sequence $Y - X_n$:

$$\int_\Omega (Y - X) d\mu = \int_\Omega \varliminf_{n\to\infty} (Y - X_n) d\mu \leqslant \varliminf_{n\to\infty} \int_\Omega (Y - X_n) d\mu.$$

Equivalently, one has

$$\int_\Omega X d\mu \geqslant \varlimsup_{n\to\infty} \int_\Omega X_n d\mu.$$

Together with the middle inequality in (2.13), one obtains the convergence property (2.12). □

The last result is known as the dominated convergence theorem. It does not require $X_n \geqslant 0$ but one needs a uniform control on their magnitudes.

Theorem 2.4. *Let X_n, X, Y be measurable functions. Suppose that $|X_n| \leqslant Y$ a.e. for all n, $X_n \to X$ a.e. and $Y \in L^1$. Then $X \in L^1$ and*

$$\lim_{n\to\infty} \int_\Omega X_n d\mu = \int_\Omega X d\mu.$$

Proof. By assumption, one has

$$X_n^{\pm} \leqslant |X_n| = X_n^+ + X_n^- \leqslant Y \quad \text{a.e.}$$

for all n. In addition, since the functions $x \mapsto \max\{x, 0\}$ and $x \mapsto \{-x, 0\}$ are continuous, one knows that $X_n^{\pm} \to X^{\pm}$ a.e. According to Corollary 2.2, one has $\int_\Omega X^{\pm} d\mu < \infty$ (thus $X \in L^1$) and

$$\lim_{n\to\infty} \int_\Omega X_n^{\pm} d\mu = \int_\Omega X^{\pm} d\mu.$$

Now the result follows from linearity of integration. □

Example 2.2. An important example is the case when $(\Omega, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $\mu = dx$ is the Lebesgue measure. The integral $\int_{\mathbb{R}^n} X(x)dx$ is called the *Lebesgue integral* of X. When $\Omega = [a, b] \subseteq \mathbb{R}^1$ and $X : [a, b] \to \mathbb{R}$ is a continuous function, the Lebesgue integral $\int_{[a,b]} X(x)dx$ coincides with the Riemann integral.

Before specialising in the probabilistic case, we present a useful continuity property of integration.

Proposition 2.8. *Let X be an integrable function on $(\Omega, \mathcal{F}, \mu)$. Then for any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$F \in \mathcal{F}, \mu(F) < \delta \implies \int_F |X| d\mu < \varepsilon.$$

Proof. Let $X_n \triangleq |X|\mathbf{1}_{\{|X|>n\}}$. According to the dominated convergence theorem, one has

$$\int_\Omega X_n d\mu = \int_{\{|X|>n\}} |X| d\mu \to 0$$

as $n \to \infty$. As a result, given any $\varepsilon > 0$, there exists $n > 0$ such that

$$\int_{\{|X|>n\}} |X| d\mu < \frac{\varepsilon}{2}.$$

For the above n and any $F \in \mathcal{F}$, one has

$$\begin{aligned} \int_F |X| d\mu &= \int_{F\cap\{|X|\leqslant n\}} |X| d\mu + \int_{F\cap\{|X|>n\}} |X| d\mu \\ &\leqslant n\mu(F) + \int_{\{|X|>n\}} |X| d\mu < n\mu(F) + \frac{\varepsilon}{2}. \end{aligned}$$

Take $\delta \triangleq \varepsilon/2n$. It follows that $\int_F |X| d\mu < \varepsilon$ whenever $\mu(F) < \delta$. This gives the desired continuity property of the integral. □

2.4 The mathematical expectation of a random variable

The probabilistic case (i.e. when μ is a probability measure) is of central importance to us. All the previous discussions on integration carry through and we give the integral a specific name in this case: the *mathematical expectation*.

Definition 2.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a random variable on Ω. If the integral $\int_\Omega X d\mathbb{P}$ exists, we call it the (mathematical) expectation of X and denote it as $\mathbb{E}[X]$. As usual, we say that X is *integrable* and write $X \in L^1$ if $\mathbb{E}[X]$ exists finitely.

There is a special feature of the expectation that does not have its counterpart for general measures. Namely, one has $\mathbb{E}[1] = 1$ and thus $\mathbb{E}[c] = c$ for any constant $c \in \mathbb{R}$. This is a trivial consequence of that fact that $\mathbb{P}(\Omega) = 1$ but it has several nice implications (cf. Corollary 2.3 as one such example).

Notation. Given $A \in \mathcal{F}$ and random variable X, we sometimes write $\mathbb{E}[X; A] \triangleq \int_A X d\mathbb{P}$ (integral of X over A).

2.4.1 The law of a random variable and the change of variable formula

In elementary probability, one defines the expectation of a random variable using the formulae:

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in S_X} x\mathbb{P}(X = x), & X \text{ discrete;} \\ \int_{-\infty}^{\infty} x f_X(x) dx, & X \text{ continuous with density function } f_X \end{cases} \tag{2.14}$$

We now illustrate how these two formulae are unified and connected to the general Definition 2.7.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable on Ω. X induces a probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in a natural way through

$$\mu_X(A) \triangleq \mathbb{P}(X \in A), \ A \in \mathcal{B}(\mathbb{R}).$$

Definition 2.8. The above probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induced by X is called the *law* of the random variable X.

The relation between the law μ_X and the distribution function of X is straight forward:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mu_X((-\infty, x]).$$

From this relation, one sees that μ_X is the Lebesgue-Stieltjes measure induced by F_X.

The connection between the two viewpoints of the expectation (Definition 2.7 and the equation (2.14)) is contained in the following change of variable formula.

Theorem 2.5. *Let X be a random variable on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $g : \mathbb{R} \to \mathbb{R}$ be an $\mathbb{R}$-valued $\mathcal{B}(\mathbb{R})$-measurable function. Suppose that $g(X)$ is integrable with respect to $\mathbb{P}$. Then g is integrable with respect to μ_X (the law of X) and one has*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\mu_X(dx).$$

Proof. The argument is nothing deeper than the definition of μ_X and construction of the integrals. By writing $g = g^+ - g^-$ it is enough to consider the case when $g \geqslant 0$, which by approximation further reduces to the case when g is non-negative simple, and eventually to the case when $g = \mathbf{1}_A$ ($A \in \mathcal{B}(\mathbb{R})$). But in this case, by the definition of μ_X one trivially has

$$\mathbb{E}[\mathbf{1}_A(X)] = \mathbb{P}(X \in A) = \mu_X(A) = \int_{\mathbb{R}} \mathbf{1}_A d\mu_X.$$

□

In the context of Theorem 2.5, by choosing $g(x) = x$ one immediately recovers (2.14). In fact, Theorem 2.5 shows that

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\mu(dx).$$

If X is a discrete random variable whose set of possible values is $S_X = \{x_1, x_2, \cdots\}$, then

$$\int_{\mathbb{R}} x\mathbb{P}^X(dx) = \sum_{n=1}^{\infty} x_n \mu_X(\{x_n\}) = \sum_{n=1}^{\infty} x_n \mathbb{P}(X = x_n). \tag{2.15}$$

If X is continuous with density function f_X, then

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

and thus

$$\int_{\mathbb{R}} x\mu_X(dx) = \int_{\mathbb{R}} x f_X(x)dx. \tag{2.16}$$

Therefore, the relation (2.14) holds. As a good exercise we let the reader think about why (2.15) and (2.16) hold.

2.4.2 Some basic inequalities for the expectation

We conclude by discussing a few basic integral inequalities that will be used frequently in probability theory. Since we are motivated from the probabilistic side, we will always work on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and integration thus becomes expectation. But we should point out that most of the inequalities (only except for Corollary 2.3) have their obvious extensions to general measures.

The first inequality is *Markov's inequality*. The proof is nearly trivial but it has a broad range of applications in tail probability estimates, convergence of random variables, large deviation principles, error estimates etc.

Theorem 2.6. *Let X be a random variable on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then for any $\alpha, \lambda > 0$, one has*

$$\mathbb{P}(|X| \geqslant \lambda) \leqslant \frac{\mathbb{E}[|X|^\alpha]}{\lambda^\alpha}.$$

Proof. The result follows from taking expectation on both sides of the following obvious inequality:

$$\mathbf{1}_{\{|X| \geqslant \lambda\}} \leqslant \frac{|X|^\alpha}{\lambda^\alpha}.$$

□

Remark 2.6. The case when $\alpha = 2$ (sometimes with X replaced by $X - \mathbb{E}[X]$ provided that $X \in L^1$), i.e.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant \lambda) \leqslant \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\lambda^2},$$

is known as *Chebyshev's inequality*.

Example 2.3. Let X be a standard normal random variable. According to Markov's inequality, for any $x > 0$ one has

$$\mathbb{P}(X \geqslant x) = \mathbb{P}(e^X \geqslant e^x) \leqslant e^{-tx}\mathbb{E}[e^{tX}] = e^{-tx+t^2/2} \quad \forall t > 0. \tag{2.17}$$

Here we used the formula $\mathbb{E}[e^{tX}] = e^{t^2/2}$ for the moment generating function of X. By optimising the right hand side of (2.17) over $t > 0$, one has

$$\mathbb{P}(X \geqslant x) \leqslant \inf_{t>0} e^{-tx+t^2/2} = e^{-x^2/2}.$$

In other words, one obtains the useful analytic inequality

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du \leqslant e^{-x^2/2} \quad \forall x > 0.$$

The next two inequalities are called *Hölder's inequality* and *Minkowski's inequality*. They have significant applications in several areas of mathematics apart from probability theory. We first recall two elementary inequalities for real numbers.

Lemma 2.4. *Let $a, b \in \mathbb{R}$, $r > 0$ and $1 < p, q < \infty$ with $1/p + 1/q = 1$. Then the following inequalities hold true:*

(i) $|a+b|^r \leqslant \max\{1, 2^{r-1}\} \cdot (|a|^r + |b|^r)$.
(ii) [Young's inequality] $|ab| \leqslant |a|^p/p + |b|^q/q$.

Proof. (i) Without loss of generality, let us assume that $a, b > 0$ and $r \neq 1$ (for otherwise the inequality is trivial). Consider the function

$$f(t) \triangleq t^r + (1-t)^r, \quad t \in [0,1].$$

Then

$$f'(t) = r(t^{r-1} - (1-t)^{r-1}).$$

If $r > 1$, $f(t)$ attains minimum at $t = 1/2$, yielding the inequality

$$t^r + (1-t)^r \geqslant \frac{1}{2^{r-1}}.$$

If $0 < r < 1$, $f(t)$ attains minimum at the end points $t = 0, 1$, yielding the inequality

$$t^r + (1-t)^r \geqslant 1.$$

The result follows by replacing t with $\frac{a}{a+b}$.

(ii) Since $\log x$ is a concave function, for any $x, y > 0$ and $\alpha, \beta \geqslant 1$ with $\alpha + \beta = 1$, one has

$$\alpha \log x + \beta \log y \leqslant \log(\alpha x + \beta y),$$

or equivalently,

$$x^\alpha y^\beta \leqslant \alpha x + \beta y.$$

The result follows by substituting $|a| = x^\alpha$, $|b| = y^\beta, p = 1/\alpha, q = 1/\beta$. □

Before stating Hölder's inequality and Minkowski's inequality, it is convenient to introduce the following notation. Let $p \geqslant 1$. Given a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$, we set

$$\|X\|_p \triangleq \left(\mathbb{E}[|X|^p]\right)^{1/p},$$

and we say that $X \in L^p$ if $\|X\|_p$ is finite (i.e. X has finite p-th moment).

Remark 2.7. The functional $\|\cdot\|_p$ defines a notion of length on the space of random variables with finite p-th moments. For one who is familiar with the language of functional analysis, this space becomes a Banach space when it is equipped with the norm $\|\cdot\|_p$ and two random variables are identified whenever they are equal a.s.

Theorem 2.7. *(i) [Hölder's inequality] Let $1 < p, q < \infty$ be such that $1/p + 1/q = 1$. Suppose that $X \in L^p$ and $Y \in L^q$. Then $XY \in L^1$ and one has*

$$|\mathbb{E}[XY]| \leqslant \mathbb{E}[|XY|] \leqslant \|X\|_p \cdot \|Y\|_q. \tag{2.18}$$

(ii) [Minkowski's inequality] Let $1 \leqslant p < \infty$. Then for any $X, Y \in L^p$, one has $X + Y \in L^p$ and

$$\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p.$$

Proof. (i) We only need to prove the second part of (2.18). Define $U \triangleq \frac{X}{\|X\|_p}$ and $V \triangleq \frac{Y}{\|Y\|_q}$. By Young's inequality (cf. Lemma 2.4 (ii)), one has

$$|UV| \leqslant \frac{U^p}{p} + \frac{V^q}{q}.$$

The result follows by taking expectation on both sides and expressing U, V in terms of X, Y.

(ii) The case when $p = 1$ is a simple consequence of the usual triangle inequality. We therefore assume that $p > 1$. Firstly, note from Lemma 2.4 (i) that $|X+Y|^p \leqslant 2^{p-1}(|X|^p + |Y|^p)$, which implies $X + Y \in L^p$. Next, one has

$$\begin{aligned} \mathbb{E}[|X + Y|^p] &= \mathbb{E}[|X + Y| \cdot |X + Y|^{p-1}] \\ &\leqslant \mathbb{E}[|X| \cdot |X + Y|^{p-1}] + \mathbb{E}[|Y| \cdot |X + Y|^{p-1}]. \end{aligned} \tag{2.19}$$

By using Hölder's inequality (with $q \triangleq \frac{p}{p-1}$ so that $1/p + 1/q = 1$), one sees that

$$\begin{aligned} &\mathbb{E}[|X| \cdot |X + Y|^{p-1}] \\ &\leqslant \big(\mathbb{E}[|X|^p]\big)^{1/p} \cdot \big(\mathbb{E}[|X + Y|^{(p-1)q}]\big)^{1/q} \\ &= \|X\|_p \cdot \big(\mathbb{E}[|X + Y|^p]\big)^{1/q} \quad \text{(note that } p = (p-1)q\text{)}, \end{aligned}$$

and similarly for the second term on the right hand side of (2.19). By substituting them into (2.19), one arrives at

$$\mathbb{E}[|X + Y|^p] \leqslant \big(\|X\|_p + \|Y\|_p\big) \cdot \big(\mathbb{E}[|X + Y|^p]\big)^{1/q}.$$

Now the result follows by dividing $\big(\mathbb{E}[|X+Y|^p]\big)^{1/q}$ to the left hand side and using the relation $1/p + 1/q = 1$. □

Remark 2.8. In Hölder's inequality, the case when $p = q = 2$ is of special importance and is known as the *Cauchy-Schwarz inequality*. The Minkowski inequality is essentially a triangle inequality if one interprets $\|\cdot\|_p$ as a notion of length on the space of random variables with finite p-th moment.

The power of these two inequalities (in particular of Hölder's inequality) will be seen in the study of martingales, diffusion processes, Gaussian analysis etc. We give one simple application here.

Corollary 2.3. *Suppose that* $1 \leqslant p < q < \infty$ *and* $X \in L^q$. *Then* $X \in L^p$ *and* $\|X\|_p \leqslant \|X\|_q$.

Proof. Let $r \triangleq q/p > 1$ and choose the corresponding exponent $s > 1$ so that $1/r + 1/s = 1$. According to Hölder's inequality, one has

$$\mathbb{E}[|X|^p] = \mathbb{E}[|X|^p \cdot 1] \leqslant \big(\mathbb{E}[|X|^{pr}]\big)^{1/r} \cdot \big(\mathbb{E}[1^s]\big)^{1/s} = \big(\mathbb{E}[|X|^q]\big)^{p/q}.$$

The result follows from taking p-th root on both sides. □

The last inequality we shall present is *Jensen's inequality*. It is related to the notion of convexity. Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be *convex*, if

$$\varphi(\alpha x + \beta y) \leqslant \alpha\varphi(x) + \beta\varphi(y)$$

for any $x, y \in \mathbb{R}$ and $\alpha, \beta \geqslant 0$ with $\alpha + \beta = 1$. A convex function on $\mathbb{R}$ is necessarily continuous. In addition, given any $x \in \mathbb{R}$ the right derivative

$$\varphi'_+(x) \triangleq \lim_{h \downarrow 0} \frac{\varphi(x+h) - \varphi(x)}{h}$$

of φ at x exists finitely and it satisfies

$$\varphi'_+(x) \cdot (y - x) \leqslant \varphi(y) - \varphi(x) \quad \forall x, y \in \mathbb{R}. \tag{2.20}$$

Theorem 2.8. *[Jensen's inequality] Suppose that* X *is an integrable random variable on* $(\Omega, \mathcal{F}, \mathbb{P})$. *Let* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a convex function. Then* $\mathbb{E}[\varphi(X)]$ *exists (which may not necessarily be finite) and*

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)].$$

Proof. In (2.20), by taking $x = \mathbb{E}[X]$ and $y = X$ one has

$$\varphi'_+(\mathbb{E}[X]) \cdot (X - \mathbb{E}[X]) \leqslant \varphi(X) - \varphi(\mathbb{E}[X]).$$

The result follows from taking expectation on both sides. □

2.5 The conditional expectation

Another essential technique in probability theory is *conditioning*. This is a very natural concept as one often wants to know how a priori knowledge of partial information affects the original distribution. Since different σ-algebras represent different amounts of information, it is natural to consider conditional probabilities / distributions *given a σ-algebra*. This leads one to the general notion of conditional expectation.

2.5.1 The general idea

Let X be a given integrable random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. Heuristically, $\mathcal{G}$ contains a subcollection of information from $\mathcal{F}$. We want to define the conditional expectation of X given $\mathcal{G}$ (denoted as $\mathbb{E}[X|\mathcal{G}]$).

Let us first make two extreme observations. If $\mathcal{G} = \{\emptyset, \Omega\}$ (the trivial σ-algebra), the information contained in $\mathcal{G}$ is trivial. In this case, the most effective prediction of X given the information in $\mathcal{G}$ should merely be its mean value, i.e. $\mathbb{E}[X|\{\emptyset, \Omega\}] = \mathbb{E}[X]$. Next, suppose that $\mathcal{G} = \mathcal{F}$ (the total information). Since X is $\mathcal{F}$-measurable, heuristically the information in $\mathcal{F}$ allows one to determine the value of X at each random experiment. As a result, the prediction of X given $\mathcal{F}$ should just be the random variable X itself, i.e. $\mathbb{E}[X|\mathcal{F}] = X$. For those intermediate situations where $\mathcal{G}$ is a non-trivial proper sub-σ-algebra of $\mathcal{F}$, it is thus reasonable to expect that $\mathbb{E}[X|\mathcal{G}]$ *should be defined as a suitable random variable*.

To motivate its definition, we consider an elementary situation. Suppose that A is a given event. The *conditional probability* of an arbitrary event B given A is defined as

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

When viewed as a set function, $\mathbb{P}(\cdot|A)$ is the *conditional probability measure* given A. The integral of X with respect to this conditional probability measure gives the *average value of X given the occurrence of A*:

$$\mathbb{E}[X|A] = \int_\Omega X d\mathbb{P}(\cdot|A) = \frac{\mathbb{E}[X\mathbf{1}_A]}{\mathbb{P}(A)}. \tag{2.21}$$

Now suppose that the given sub-σ-algebra $\mathcal{G}$ is generated by a partition of Ω, say

$$\mathcal{G} = \sigma(A_1, A_2, \cdots, A_n)$$

where $A_i \cap A_j = \emptyset$ and $\Omega = \cup_{i=1}^n A_i$. To define the random variable $\mathbb{E}[X|\mathcal{G}]$, the main idea is that on each event A_i the value of $\mathbb{E}[X|\mathcal{G}]$ should simply be the average value of X given that A_i occurs. Mathematically, one has

$$\mathbb{E}[X|\mathcal{G}](\omega) = \sum_{i=1}^n c_i \mathbf{1}_{A_i}(\omega),$$

where

$$c_i \triangleq \mathbb{E}[X|A_i] = \frac{\mathbb{E}[X\mathbf{1}_{A_i}]}{\mathbb{P}(A_i)}, \quad i = 1, 2, \cdots, n.$$

From this definition, it is clear that $\mathbb{E}[X|\mathcal{G}]$ is a $\mathcal{G}$-measurable random variable. In addition, a key observation is that the integral of $\mathbb{E}[X|\mathcal{G}]$ on each event A_i coincides with the integral of X on the same event:

$$\int_{A_i} \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_{A_i} \sum_{j=1}^n c_j \mathbf{1}_{A_j}(\omega)d\mathbb{P} = c_i\mathbb{P}(A_i) = \mathbb{E}[X\mathbf{1}_{A_i}] = \int_{A_i} X d\mathbb{P}.$$

The above integral property motivates the following general definition of the conditional expectation.

Definition 2.9. Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. The *conditional expectation of X given $\mathcal{G}$* is an integrable, $\mathcal{G}$-measurable random variable Y such that

$$\int_A Y d\mathbb{P} = \int_A X d\mathbb{P} \quad \forall A \in \mathcal{G}. \tag{2.22}$$

This random variable is denoted as $\mathbb{E}[X|\mathcal{G}]$.

It is not clear at all whether the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ exists uniquely. A standard measure-theoretic way of proving its existence and uniqueness is through the so-called *Radon-Nikodym theorem.* Instead of elaborating this more general method, we will give an alternative geometric construction of the conditional expectation which appears to be more enlightening.

2.5.2 Geometric construction of the conditional expectation

Our construction relies on some basic notions from Hilbert space which we shall first recall. The reader is referred to [Lan93] for a systematic introduction.

In Euclidean geometry, elements in $\mathbb{R}^2$ or $\mathbb{R}^3$ are viewed as vectors. There is a notion of inner product $\langle v, w\rangle$ between two vectors v, w which satisfies the following basic properties:

(i) *symmetry*: $\langle v, w\rangle = \langle w, v\rangle$;
(ii) *bilinearity*: $\langle cv_1 + v_2, w\rangle = c\langle v_1, w\rangle + \langle v_2, w\rangle$ where $c \in \mathbb{R}$ is a scalar;
(iii) *positive definiteness*: $\langle v, v\rangle \geqslant 0$ and equality holds iff $v = 0$.

The inner product can be used to measure all sorts of geometric properties e.g. length ($|v| = \sqrt{\langle v, v\rangle}$), angle ($\angle_{v,w} = \frac{\langle v,w\rangle}{|v|\cdot|w|}$), orthogonality ($v \perp w \iff \langle v, w\rangle = 0$) etc. Given a vector v and a subspace $E \subseteq \mathbb{R}^3$, one can naturally talk about the orthogonal projection of v onto E.

The concept of Hilbert space generalises the above considerations to an abstract setting.

Definition 2.10. Let H be a vector space over $\mathbb{R}$. An *inner product* over H is a function $\langle \cdot, \cdot\rangle : H \times H \to \mathbb{R}$ which satisfies the above Properties (i)–(iii). A vector space equipped with an inner product is called an *inner product space*.

Two elements v, w are said to be *orthogonal* (denoted as $v \perp w$) if $\langle v, w\rangle = 0$. By using the inner product, one can define the notion of length (more commonly known as a *norm*) by

$$\|v\| \triangleq \sqrt{\langle v, v\rangle}, \quad v \in H.$$

With this norm structure one can talk about convergence just like in Euclidean spaces: we say that v_n *converges to* v if $\|v_n - v\| \to 0$ as $n \to \infty$. A sequence $\{v_n : n \geqslant 1\}$ in H is said to be a *Cauchy sequence* in H if for any $\varepsilon > 0$, there exists $N \geqslant 1$ such that

$$m, n > N \implies \|v_m - v_n\| < \varepsilon.$$

Definition 2.11. A *Hilbert space* is a complete inner product space, i.e. an inner product space in which every Cauchy sequence converges.

Example 2.4. $\mathbb{R}^d$ is a Hilbert space when equipped with the Euclidean inner product:

$$\langle x, y\rangle \triangleq x_1 y_1 + \cdots + x_d y_d.$$

The following example is of our main interest.

Example 2.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $H \triangleq L^2(\Omega, \mathcal{F}, \mathbb{P})$ be the space of square integrable random variables. To be more precise, we shall identify

two random variables X, Y if $X = Y$ a.s. As a result, H is indeed the space of equivalence classes of square integrable random variables. Define an inner product over H by

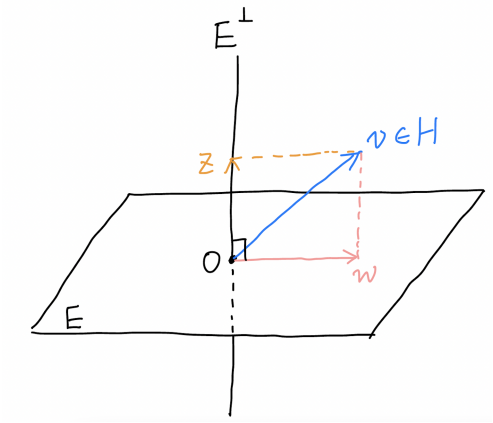
$$\langle X, Y \rangle_{L^2} \triangleq \mathbb{E}[XY], \quad X, Y \in H.$$

Then $(H, \langle \cdot, \cdot \rangle_{L^2})$ is a Hilbert space.

In finite dimensions there is no need to emphasise completeness as every finite dimensional inner product space is complete. The completeness property is essential when one considers infinite dimensional spaces such as a space of functions. In infinite dimensions it is also important to emphasise closedness when one comes to the study of subspaces: a subspace E is *closed* if $v_n \in E, v_n \to v \implies v \in E$. This notion is again not needed in finite dimensions as every subspace is closed in that case.

Example 2.6. Under the same notation as in Example 2.5, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let $E \triangleq L^2(\Omega, \mathcal{G}, \mathbb{P})$ denote the subspace of H consisting of square integrable random variables that are $\mathcal{G}$-measurable. More precisely, an equivalence class $[X] \in H$ belongs to E iff $[X]$ admits a $\mathcal{G}$-measurable representative. Then E is a closed subspace of H.

A basic property of closed subspaces in Hilbert spaces that is relevant to us is the existence of orthogonal projections. Given any subset $E \subseteq H$, we use $E^\perp$ to denote the collection of elements $v \in H$ such that $v \perp w$ for all $w \in E$.



Proposition 2.9. *Let E be a closed subspace of a Hilbert space H. For any $v \in H$, there exists a unique decomposition $v = w + z$ where $w \in E$ and $z \in E^\perp$.*

The vector w is the unique element in E that has minimal distance to v:

$$\|v - w\| = \min_{w' \in E} \|v - w'\|.$$

Definition 2.12. The element w in Proposition 2.9 is called the *orthogonal projection* of v onto the closed subspace E.

Having the above preparations, we can now give the geometric construction of the conditional expectation.

Theorem 2.9. *Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. There exists an integrable, $\mathcal{G}$-measurable random variable Y, such that*

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P} \quad \forall A \in \mathcal{G}. \tag{2.23}$$

Such Y is unique in the sense that if Y_1, Y_2 are two integrable, $\mathcal{G}$-measurable random variables satisfying (2.23), then $Y_1 = Y_2$ a.s.

Proof. Uniqueness follows from Remark 2.5. To prove existence, we first consider the case when X is square integrable. Recall from Examples 2.5 and 2.6 that $E \triangleq L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of the Hilbert space $H \triangleq L^2(\Omega, \mathcal{F}, \mathbb{P})$. Let Y be the orthogonal projection of X onto H_0 whose existence is ensured by Proposition 2.9. According to the same proposition, one knows that

$$X - Y \perp Z \quad \forall Z \in E.$$

Taking $Z = \mathbf{1}_A$ with $A \in \mathcal{G}$, one finds that

$$\langle X - Y, \mathbf{1}_A \rangle_{L^2} = \mathbb{E}\big[(X - Y)\mathbf{1}_A\big] = 0,$$

which is exactly the property (2.23).

We now consider the case when X is integrable. We first assume that $X \geqslant 0$. For each $n \geqslant 1$, define $X_n \triangleq \min\{X, n\}$. Then $X_n \in H$ and thus there exists $Y_n \in H_0$ such that

$$\int_A X_n d\mathbb{P} = \int_A Y_n d\mathbb{P} \quad \forall A \in \mathcal{G}. \tag{2.24}$$

Since $X_n \geqslant 0$, from the relation (2.24) one sees that

$$\int_A Y_n d\mathbb{P} \geqslant 0 \quad \forall A \in \mathcal{G},$$

which implies by Proposition 2.7 (ii) that $Y_n \geqslant 0$ a.s. In addition, since $X_{n+1} \geqslant X_n$, the same relation (2.24) yields that

$$\int_A Y_{n+1} d\mathbb{P} \geqslant \int_A Y_n d\mathbb{P} \ (\iff \int_A (Y_{n+1} - Y_n) d\mathbb{P} \geqslant 0) \quad \forall A \in \mathcal{G}.$$

As a result, Y_n is increasing a.s. Set $Y \triangleq \lim_{n\to\infty} Y_n$. It is then clear that $Y \geqslant 0$ a.s. and Y is $\mathcal{G}$-measurable. By using the monotone convergence theorem, after sending $n \to \infty$ in (2.24) one obtains the desired relation (2.23). The integrability of Y follows by taking $A = \Omega$. Finally, for the general case, one considers $X = X^+ - X^-$ and use linearity. □

2.5.3 Basic properties of the conditional expectation

By taking $A = \Omega$ in (2.22), it is clear that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]$. This is known as the *law of total expectation*. We list a few basic properties of the conditional expectation that are useful later on.

Theorem 2.10. *The conditional expectation satisfies the following properties. We always assume that the underlying random variables are integrable.*

(i) The map $X \mapsto \mathbb{E}[X|\mathcal{G}]$ is linear.
(ii) If $X \leqslant Y$, then $\mathbb{E}[X|\mathcal{G}] \leqslant \mathbb{E}[Y|\mathcal{G}]$. In particular,

$$|\mathbb{E}[X|\mathcal{G}]| \leqslant \mathbb{E}[|X||\mathcal{G}].$$

(iii) If Z is $\mathcal{G}$-measurable, then

$$\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}].$$

(iv) [The tower rule] If $\mathcal{G}_1 \subseteq \mathcal{G}_2$ are sub-σ-algebras of $\mathcal{F}$, then

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[X|\mathcal{G}_1].$$

(v) If X and $\mathcal{G}$ are independent, then

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X].$$

(vi) [Jensen's inequality] Let φ be a convex function on $\mathbb{R}$. Then

$$\varphi(\mathbb{E}[X|\mathcal{G}]) \leqslant \mathbb{E}[\varphi(X)|\mathcal{G}]. \tag{2.25}$$

Proof. (i) Let X_1, X_2 be integrable random variables and $c \in \mathbb{R}$. Define $Y_i \triangleq \mathbb{E}[X_i|\mathcal{G}]$ $(i = 1, 2)$. By using (2.23) and linearity of integration, one has

$$\int_A (cX_1 + X_2)d\mathbb{P} = \int_A (cY_1 + Y_2)d\mathbb{P} \quad \forall A \in \mathcal{G}.$$

Since $cY_1 + Y_2$ is $\mathcal{G}$-measurable, it follows from Theorem 2.9 (which gives a characterisation of the conditional expectation) that

$$cY_1 + Y_2 = \mathbb{E}[cX_1 + X_2|\mathcal{G}].$$

(ii) The result follows from the fact that

$$\int_A \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_A X d\mathbb{P} \leqslant \int_A Y d\mathbb{P} = \int_A \mathbb{E}[Y|\mathcal{G}]d\mathbb{P} \quad \forall A \in \mathcal{G}.$$

(iii) Suppose that $Z = \mathbf{1}_G$ for some $G \in \mathcal{G}$. Then for any $A \in \mathcal{G}$, one has

$$\int_A ZX d\mathbb{P} = \int_{A\cap G} X d\mathbb{P} = \int_{A\cap G} \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_A Z\mathbb{E}[X|\mathcal{G}]d\mathbb{P}.$$

Since $Z\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, one concludes by the characterising property (2.23) that $Z\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[ZX|\mathcal{G}]$. The general case follows by the standard argument (simple function approximation).

(iv) Since $\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1]$ is $\mathcal{G}_1$-measurable, one only needs to check the characterising property; indeed,

$$\int_A \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1]d\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{G}_2]d\mathbb{P} = \int_A X d\mathbb{P} \quad \forall A \in \mathcal{G}_1,$$

where the last equality follows from the assumption that $\mathcal{G}_1 \subseteq \mathcal{G}_2$.

(v) Left as an exercise (consider simple function approximation).

(vi) The proof is similar to the unconditional case (cf. Theorem 2.8). Setting $x = \mathbb{E}[X|\mathcal{G}]$ and $y = X$ in (2.20), one finds that

$$\varphi'_+(\mathbb{E}[X|\mathcal{G}])(X - \mathbb{E}[X|\mathcal{G}]) \leqslant \varphi(X) - \varphi(\mathbb{E}[X|\mathcal{G}]).$$

By taking conditional expectation on both sides and using Property (iii), it follows that

$$\mathbb{E}\big[\varphi(X)|\mathcal{G}\big] - \varphi(\mathbb{E}[X|\mathcal{G}]) \geqslant \varphi'_+(\mathbb{E}[X|\mathcal{G}]) \cdot \mathbb{E}\big[(X - \mathbb{E}[X|\mathcal{G}])\big|\mathcal{G}\big] = 0.$$

This proves Jensen's inequality (2.25) for the conditional expectation.

□

The intuition behind some of these properties is clear. For instance, for Property (iii), given the information in $\mathcal{G}$ the value of Z is known. In other words, conditional on $\mathcal{G}$ the random variable is frozen (treated as a constant) and can thus be moved outside the conditional expectation. In Property (v), by independence the knowledge of $\mathcal{G}$ provides no meaningful information for predicting X. As a result, the most effective prediction of X is its unconditional mean.

One can easily write down parallel versions of monotone convergence, Fatou's lemma and dominated convergence for the conditional expectation. We will not give the details here.

3 Product measure spaces

In Chapter 1, one knows how to construct a random variable with given distribution (construction of the Lebesgue-Stieltjes measure). The next question is: *how can one construct independent random variables with given marginal distributions*? This is an important question since a substantial part of classical probability theory is related to understanding the asymptotic behaviour of independent sequences. The path from the construction of a single to independent random variables is through the notion of product spaces.

In this chapter, we introduce the notion of product measure spaces and use them to construct independent random variables. We begin by discussing the product measurable structure in Section 3.1. This is a technical prerequisite for proving Fubini's theorem for bounded, measurable functions, the latter of which easily yields the construction of the product measure and thus a canonical way of constructing (finitely many) independent random variables. The general Fubini's theorem follows from the bounded case by a standard argument. These results are discussed in Section 3.2. The case of countable products and independent sequences is more delicate and requires deeper considerations (Kolmogorov's extension theorem). We deal with it in Section 3.3.

3.1 Product measurable structure

Our first goal is to define the "product" of two measure spaces as a new measure space. Before working with measures, we first try to understand the underlying measurable structure. Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be given measurable spaces. To form their product as a new measurable space, it is clear that the sample space should just be the Cartesian product $\Omega_1 \times \Omega_2$. However, for the σ-algebra one cannot simply take the Cartesian product

$$\mathcal{F}_1 \times \mathcal{F}_2 \triangleq \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\},$$

due to the obvious reason that this is not a σ-algebra! Instead, one should consider the σ-algebra generated by $\mathcal{F}_1 \times \mathcal{F}_1$. This leads to the following definition.

Definition 3.1. The *product measurable space* $(\Omega, \mathcal{F})$ of $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ is defined by

$$\Omega = \Omega_1 \times \Omega_2, \ \mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \triangleq \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$$

Next, we discuss a basic property of bounded, measurable functions on $(\Omega, \mathcal{F})$. This is a necessary (technical) ingredient for proving Fubini's theorem and the

construction of product measures later on. For convenience, we use the notation $b\mathcal{F}$ to denote the space of bounded, measurable functions on $(\Omega, \mathcal{F})$.

Lemma 3.1. *Let $f \in b\mathcal{F}$. Then the following statements hold true:*

(i) for each fixed $\omega_1 \in \Omega_1$, the function $\omega_2 \mapsto f(\omega_1, \omega_2)$ is bounded, $\mathcal{F}_2$-measurable;
(ii) for each fixed $\omega_2 \in \Omega_2$, the function $\omega_1 \mapsto f(\omega_1, \omega_2)$ is bounded, $\mathcal{F}_1$-measurable.

Proof. We again use the standard argument, i.e. using Dynkin's π-λ theorem and approximating general measurable functions by simple ones. First of all, f can be uniformly approximated by a sequence f_n of simple functions (cf. Lemma 2.3 and Remark 2.2). Since measurability is preserved under pointwise limit (cf. Proposition 2.1), it suffices to prove the desired properties when $f = \mathbf{1}_A$ ($A \in \mathcal{F}$).

To this end, we define

$$\mathcal{H} \triangleq \{A \in \mathcal{F} : \mathbf{1}_A \text{ satisfies properties (i) and (ii)}\}.$$

It is obvious that the Cartesian product $\mathcal{F}_1 \times \mathcal{F}_2$ is a π-system and $\mathcal{F}_1 \times \mathcal{F}_2 \subseteq \mathcal{H}$. Next, we check that $\mathcal{H}$ is a λ-system:

(L1) $\Omega \in \mathcal{H}$ is obvious.
(L2) Suppose that $A, B \in \mathcal{H}$ with $A \subseteq B$. Then $\mathbf{1}_{B\setminus A} = \mathbf{1}_B - \mathbf{1}_A$. In particular, the measurability properties (i), (ii) satisfied by $\mathbf{1}_A$ and $\mathbf{1}_B$ are inherited by $\mathbf{1}_{B\setminus A}$ due to linearity. Therefore, $B\setminus A \in \mathcal{H}$.
(L3) Suppose that $A_n \in \mathcal{H}$, $A_n \uparrow A$. Then $\mathbf{1}_{A_n} \uparrow \mathbf{1}_A$ for every $(\omega_1, \omega_2) \in \Omega$. Since the measurability properties (i), (ii) satisfied by $\mathbf{1}_{A_n}$ are preserved under pointwise limit, one concludes that $A \in \mathcal{H}$.

It follows from Dynkin's π-λ theorem that $\lambda(\mathcal{F}_1 \times \mathcal{F}_2) = \mathcal{F} = \mathcal{H}$. In other words, all members of $\mathcal{F}$ satisfy the desired properties. □

Remark 3.1. If S is a topological space, one can define its *Borel σ-algebra* $\mathcal{B}(S)$ to be the σ-algebra generated by all open subsets of S. Given topological spaces S, T, it is apparent that

$$\mathcal{B}(S) \otimes \mathcal{B}(T) \subseteq \mathcal{B}(S \times T),$$

where $\mathcal{B}(S \times T)$ denotes the Borel σ-algebra with respect to the product topology. However, it is a deep fact that $\mathcal{B}(S) \otimes \mathcal{B}(T)$ may be strictly smaller than $\mathcal{B}(S \times T)$ in general (cf. Chapter 1, Appendix B, Theorem 1.8).

3.2 Product measures and Fubini's theorem

Suppose that μ_i is a finite measure on $(\Omega_i, \mathcal{F}_i)$ $(i = 1, 2)$. Our goal is to construct the "product measure" $\mu_1 \otimes \mu_2$ on $(\Omega, \mathcal{F}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$. We will take the route of first proving Fubini's theorem (for bounded, measurable functions) and then using it to construct $\mu_1 \otimes \mu_2$.

3.2.1 Fubini's theorem for the bounded case and construction of the product measure

Let $f \in \mathrm{b}\mathcal{F}$ be given fixed. Fubini's theorem asserts that when performing the double integral of f, the order of integration does not matter. We now make this mathematically precise. First of all, according to Lemma 3.1, one knows that $f(\omega_1, \cdot) \in \mathrm{b}\mathcal{F}_2$ for each fixed $\omega_1 \in \Omega_1$. In particular, one can define the marginal integral (as a function of ω_1)

$$I_1^f : \Omega_1 \to \mathbb{R},\ \omega_1 \mapsto I_1^f(\omega_1) \triangleq \int_{\Omega_2} f(\omega_1, \omega_2)\mu_2(d\omega_2).$$

Similarly, one also defines

$$I_2^f : \Omega_2 \to \mathbb{R},\ \omega_2 \mapsto I_2^f(\omega_2) \triangleq \int_{\Omega_1} f(\omega_1, \omega_2)\mu_1(d\omega_1).$$

Proposition 3.1 (Fubini's theorem for bounded, measurable functions)**.** *For each $f \in b\Sigma$, one has $I_1^f \in b\mathcal{F}_1$, $I_2^f \in b\mathcal{F}_2$ and*

$$\int_{\Omega_1} I_1^f(\omega_1)\mu_1(d\omega_1) = \int_{\Omega_2} I_2^f(\omega_2)\mu_2(d\omega_2).$$

Proof. The argument is almost identical to the proof of Lemma 3.1 (Dynkin's π-λ theorem and simple function approximation). We leave it as an exercise. □

One can now use Proposition 3.1 to define the product measure. More precisely, for each given $F \in \mathcal{F}$, with $f = \mathbf{1}_F$ one defines

$$\mu(F) \triangleq \int_{\Omega_1} I_1^f(\omega_1)\mu_1(d\omega_1) = \int_{\Omega_2} I_2^f(\omega_2)\mu_2(d\omega_2).$$

It is clear that $\mu \geqslant 0$ and $\mu(\emptyset) = 0$. In addition, since $\mathbf{1}_F \leqslant 1$ one has

$$\mu(F) \leqslant \int_{\Omega_1} \Big(\int_{\Omega_2} 1 \cdot \mu_2(d\omega_2)\Big)\mu_1(d\omega_1) = \mu_1(\Omega_1)\mu_2(\Omega_2) < \infty.$$

Theorem 3.1. *The set function μ is a well-defined, finite measure on the product measurable space $(\Omega, \mathcal{F})$. It is the unique measure on $(\Omega, \mathcal{F})$ such that*

$$\mu(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2) \quad \forall A_i \in \mathcal{F}_i \ (i = 1, 2). \tag{3.1}$$

Proof. For the first part, it remains to verify countable additivity. Let F_n be a disjoint sequence in $\mathcal{F}$. Define

$$F \triangleq \bigcup_{n=1}^{\infty} F_n, \ G_n \triangleq \bigcup_{k=1}^{n} F_k.$$

Since $\mathbf{1}_{G_n} = \mathbf{1}_{F_1} + \cdots + \mathbf{1}_{F_n}$, by the linearity of integration one has

$$\mu(G_n) = \int_{\Omega_1} I_1^{\mathbf{1}_{G_n}}(\omega_1)\mu_1(d\omega_1) = \sum_{k=1}^{n} \int_{\Omega_1} I_1^{\mathbf{1}_{F_k}}(\omega_1)\mu_1(d\omega_1) = \sum_{k=1}^{n} \mu(F_k). \tag{3.2}$$

In addition, since $\mathbf{1}_{G_n} \uparrow \mathbf{1}_F$, by applying monotone convergence (twice!) one obtains that

$$\mu(F) = \int_{\Omega_1} I_1^{\mathbf{1}_F}(\omega_1)\mu_1(d\omega_1) = \lim_{n\to\infty} \int_{\Omega_1} I_1^{\mathbf{1}_{G_n}}(\omega_1)\mu_1(d\omega_1) = \lim_{n\to\infty} \mu(G_n).$$

By taking $n \to \infty$ in (3.2), it follows that

$$\mu(F) = \sum_{k=1}^{\infty} \mu(F_k)$$

which gives the countable additivity.

For the second part of the theorem, the relation (3.1) is obvious:

$$\begin{aligned}\mu(A_1 \times A_2) &= \int_{\Omega_1} \Big(\int_{\Omega_2} \mathbf{1}_{A_1 \times A_2}(\omega_1, \omega_2)\mu_2(d\omega_2)\Big)\mu_1(d\omega_1) \\ &= \int_{\Omega_1} \mathbf{1}_{A_1}(\omega_1)\Big(\int_{\Omega_2} \mathbf{1}_{A_2}(\omega_2)\mu_2(d\omega_2)\Big)\mu_1(d\omega_1) = \mu_1(A_1)\mu_2(A_2).\end{aligned}$$

Uniqueness of μ is a direct consequence of Proposition 1.4 (note that $\mathcal{F}_1 \times \mathcal{F}_2$ is a π-system and $\Omega \in \mathcal{F}_1 \times \mathcal{F}_2$). □

Definition 3.2. The measure μ, also denoted as $\mu_1 \otimes \mu_2$, is called the *product measure* of μ_1 and μ_2 on $(\Omega, \mathcal{F})$. The measure space $(\Omega, \mathcal{F}, \mu)$ is called the *product measure space* of $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$. One often writes

$$(\Omega, \mathcal{F}, \mu) = (\Omega_1, \mathcal{F}_1, \mu_1) \otimes (\Omega_2, \mathcal{F}_2, \mu_2).$$

3.2.2 Fubini's theorem for the general case

Once the product measure μ is constructed, extension of Fubini's theorem from the bounded to general case is routine. Recall that $(\Omega, \mathcal{F}, \mu)$ is the product measure space constructed before.

Theorem 3.2 (Fubini's theorem for general measurable functions)**.** *Suppose that f is a non-negative, measurable function on $(\Omega, \mathcal{F})$. Then one has*

$$\int_\Omega f d\mu = \int_{\Omega_1} I_1^f(\omega_1)\mu_1(d\omega_1) = \int_{\Omega_2} I_2^f(\omega_2)\mu_2(d\omega_2) \in [0, +\infty]. \tag{3.3}$$

If f is a general measurable function which is integrable with respect to μ, then (3.3) remains valid with value in $\mathbb{R}$.

Proof. For the first part, take a sequence of non-negative, simple functions $0 \leqslant f_n \uparrow f$. The relation (3.3) holds for each f_n as a consequence of Proposition 3.1. The claim follows from the monotone convergence theorem. For the second part, write $f = f^+ - f^-$. The relation (3.3) is valid for $f^\pm$ with values in $\mathbb{R}$ (since $\int_\Omega f^\pm d\mu < \infty$ by assumption). The claim follows from linearity of integration. □

Example 3.1. Consider the following bivariate function defined on $[0,1] \times [0,1]$:

$$f(x,y) = \begin{cases} \frac{x^2-y^2}{(x^2+y^2)^2}, & x^2 + y^2 \neq 0; \\ 0, & x = y = 0. \end{cases}$$

By explicit integration, one finds that

$$\int_0^1 \Big(\int_0^1 f(x,y)dy \Big) dx = \frac{\pi}{4}, \ \int_0^1 \Big(\int_0^1 f(x,y)dx \Big) dy = -\frac{\pi}{4}.$$

In particular, Fubini's theorem fails for this example. The issue here is that f is not integrable with respect to the Lebesgue measure on $[0,1] \times [0,1]$. Indeed, one has

$$f^+(x,y) = \begin{cases} \frac{x^2-y^2}{(x^2+y^2)^2}, & 0 \leqslant y < x \leqslant 1; \\ 0, & \text{otherwise}. \end{cases}$$

Direct calculation shows that

$$\int_0^1 f^+(x,y)dy = \frac{1}{x} \int_0^1 \frac{1-t^2}{(1+t^2)^2} dt.$$

Since $1/x$ is not integrable on $[0,1]$, one has

$$\int_0^1 \Big(\int_0^1 f^+(x,y)dy \Big) dx = \infty.$$

Similarly, one also has

$$\int_0^1 \Big(\int_0^1 f^-(x,y)dy \Big) dx = \infty.$$

As a result, the two-dimensional integral of f over $[0,1] \times [0,1]$ does not exist.

Remark 3.2. The construction of product measures and Fubini's theorem extend naturally to the σ-finite case. Suppose that μ_i is a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$ $(i = 1, 2)$. By definition, there is a partition $\{A_n^{(i)}\} \subseteq \mathcal{F}_i$ of Ω_i such that $\mu_i(A_n^{(i)}) < \infty$ for all n $(i = 1, 2)$. Define

$$\mu(F) \triangleq \sum_{m,n=1}^{\infty} \mu\big(F \cap (A_m^{(1)} \times A_n^{(2)})\big), \quad \mathcal{F} \in \mathcal{F}_1 \otimes \mathcal{F}_2.$$

Note that $\mu(\cdot \cap (A_m^{(1)} \times A_n^{(2)}))$ is just the product of the finite measures $\mu_1\big|_{A_m^{(1)}}$ and $\mu_2\big|_{A_n^{(2)}}$. The σ-finite measure μ is the *product measure* of μ_1 and μ_2. Theorem 3.2 remains valid in this case.

Example 3.2. The σ-finite assumption in Remark 3.2 cannot be removed. Indeed, consider $\Omega_1 = \Omega_2 = [0,1]$, $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{B}([0,1])$ (the σ-algebra generated by $\{[a,b] : a \leqslant b \in [0,1]\}$). Let μ_1 be the Lebesgue measure on $[0,1]$ and let μ_2 be the counting measure, i.e.

$$\mu_2(A) \triangleq \begin{cases} \#\text{ of elements in } A, & \text{if } A \in \mathcal{B}([0,1]) \text{is a finite set;} \\ \infty, & \text{otherwise.} \end{cases}$$

Define $f \triangleq \mathbf{1}_F$ with $F = \{(x,y) \in [0,1]^2 : x = y\}$ (why does $F \in \mathcal{B}([0,1]) \otimes \mathcal{B}([0,1])$?) Then one has

$$I_1^f(\cdot) \equiv 1, \ I_2^f(\cdot) \equiv 0.$$

In particular,

$$\int_{\Omega_1} I_1^f(\omega_1)\mu_1(d\omega_1) \neq \int_{\Omega_2} I_2^f(\omega_2)\mu_2(d\omega_2).$$

Example 3.3 (The Dirichlet integral)**.** We use Fubini's theorem to derive an important identity in analysis:

$$\int_0^\infty \frac{\sin x}{x} dx \triangleq \lim_{t\to\infty} \int_0^t \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

This identity will be used in Section 6.2 for proving the inversion formula for the characteristic function. below The key observation is that

$$\frac{1}{x} = \int_0^\infty e^{-ux} du \quad \forall x > 0.$$

By using this formula, one can write

$$\int_0^t \frac{\sin x}{x} dx = \int_0^t \sin x \big(\int_0^\infty e^{-ux} du \big) dx.$$

According to Fubini's theorem, the last expression is equal to $\int_0^\infty \big(\int_0^t e^{-ux} \sin x dx \big) du$. The calculation of the inner integral is straight forward. Indeed, using integration by parts twice one has

$$\begin{aligned}
\int_0^t e^{-ux} \sin x dx &= \big(- e^{-ux} \cos x \big)_0^t - \int_0^t u e^{-ux} \cos x dx \\
&= -e^{-ut} \cos t + 1 - \big(u e^{-ux} \sin x \big|_0^t + \int_0^t u^2 e^{-ux} \sin x dx \big) \\
&= 1 - e^{-ut} \cos t - u e^{-ut} \sin t - u^2 \int_0^t e^{-ux} \sin x dx.
\end{aligned}$$

By rearrangement, one finds that

$$\int_0^t e^{-ux} \sin x dx = \frac{1}{1+u^2} \big(1 - e^{-ut} \cos t - u e^{-ut} \sin t \big).$$

It follows that

$$\int_0^t \frac{\sin x}{x} dx = \int_0^\infty \frac{1}{1+u^2} \big(1 - e^{-ut} \cos t - u e^{-ut} \sin t \big) du.$$

Letting $t \to \infty$ and applying dominated convergence, one arrives at

$$\int_0^\infty \frac{\sin x}{x} dx = \int_0^\infty \frac{1}{1+u^2} du = \frac{\pi}{2}.$$

We let the reader justify the uses of Fubini's theorem and dominated convergence here.

Example 3.4 (Volume of unit n-ball)**.** Let $\omega_n(r)$ denote the volume of a ball in $\mathbb{R}^n$ with radius r. We simply write $\omega_n \triangleq \omega_n(1)$. It is clear that

$$\omega_n(r) = r^n \omega_n. \tag{3.4}$$

We are going to use Fubini's theorem to compute ω_n. Let

$$A \triangleq \{(x_1, \cdots, x_n) : x_1^2 + \cdots + x_n^2 \leqslant 1\}$$

denote the unit ball in $\mathbb{R}^n$. Then

$$\omega_n = \int_A dx_1 \cdots dx_n.$$

The trick is to divide the variables $(x_1, \cdots, x_n)$ into two parts (x_1, x_2), $(x_3, \cdots, x_n)$ and integrate out the latter part first (Fubini's theorem). One gets that

$$\begin{aligned} \omega_n &= \int_{\{x_1^2+x_2^2 \leqslant 1\}} dx_1 dx_2 \int_{\{x_3^2+\cdots+x_n^2 \leqslant 1-x_1^2-x_2^2\}} dx_3 \cdots dx_n \\ &= \int_{\{x_1^2+x_2^2 \leqslant 1\}} \sqrt{1-x_1^2-x_2^2}^{n-2} dx_1 dx_2 \times \omega_{n-2}, \end{aligned} \tag{3.5}$$

where the last equality follows from the scaling property (3.4) applied to dimension $n-2$ with $r = \sqrt{1-x_1^2-x_2^2}$. The integral in (3.5) can be easily evaluated by a change of variables into polar coordinates. This gives the recursive relation

$$\omega_n = \frac{2\pi}{n} \omega_{n-2}. \tag{3.6}$$

By taking $\omega_0 \triangleq 1$ and noting that $\omega_1 = 2$, one can solve the relation (3.6) to conclude that

$$\omega_{2n-1} = \frac{2(2\pi)^{n-1}}{1 \cdot 3 \cdot 5 \cdots (2n-1)}, \ \omega_{2n} = \frac{(2\pi)^n}{2 \cdot 4 \cdots (2n)}, \quad n \geqslant 1. \tag{3.7}$$

What if one first integrates over the $(x_2, \cdots, x_n)$ part? This gives the relation

$$\begin{aligned} \omega_n &= \int_{-1}^{1} dx_1 \int_{\{x_2^2+\cdots+x_n^2 \leqslant 1-x_1^2\}} dx_2 \cdots dx_n \\ &= \int_{-1}^{1} \sqrt{1-x_1^2}^{n-1} dx_1 \times \omega_{n-1} \\ &= 2 \int_0^{\pi/2} (\cos t)^n dt \times \omega_{n-1}. \end{aligned} \tag{3.8}$$

Let us denote $I_n \triangleq \int_0^{\pi/2} (\cos t)^n dt$. By substituting the formulae (3.7) into (3.8), one easily finds that

$$I_{2n} = \frac{\pi}{2} \times \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots (2n)}, \; I_{2n+1} = \frac{2 \cdot 4 \cdot 6 \cdots (2n)}{1 \cdot 3 \cdot 5 \cdots (2n+1)}, \quad n \geqslant 1$$

and of course $I_0 = \pi/2$, $I_1 = 1$. Explicit calculation shows that the relation $I_{2n-1} > I_{2n} > I_{2n+1}$ (I_n is obviously decreasing in n) is equivalent to the relation that

$$\frac{\pi}{2} \cdot \frac{2n}{2n+1} < \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdots \frac{2n}{2n-1} \cdot \frac{2n}{2n+1} < \frac{\pi}{2}.$$

After taking $n \to \infty$, one arrives at the so-called *Wallis formula* for π:

$$\pi = 2 \times \left(\frac{2}{1} \cdot \frac{2}{3}\right) \cdot \left(\frac{4}{3} \cdot \frac{4}{5}\right) \cdot \left(\frac{6}{5} \cdot \frac{6}{7}\right) \cdots .$$

The following formula is a useful application of Fubini's theorem. It provides a way of computing expectation through integrating tail probabilities.

Proposition 3.2. *Let X be a non-negative random variable defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define μ to be the product measure of $\mathbb{P}$ and the Lebesgue measure on $\mathcal{F} \otimes \mathcal{B}([0,\infty))$. Let*

$$F \triangleq \{(\omega, x) : 0 \leqslant x < X(\omega)\}.$$

Then

$$\mu(F) = \mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x) dx. \tag{3.9}$$

Proof. The measurability of F follows from the fact that

$$F = \bigcup_{r \in \mathbb{Q} \cap [0,\infty)} \big(\{\omega : X(\omega) > r\} \times [0, r)\big).$$

The relation (3.9) is obtained by taking $f = \mathbf{1}_F$ in (3.3). □

Remark 3.3. This formula also indicates that $\mathbb{E}[X]$ is the "μ-area under the graph of X".

Another useful application of Fubini's theorem is the integration by parts formula for distribution functions.

Proposition 3.3. *Let F, G be two distribution functions on $\mathbb{R}$ and let μ_F, μ_G be the associated Lebesgue-Stieltjes measures respectively. For any $a < b$, one has*

$$\int_{(a,b]} F(x)\mu_G(dx) = F(b)G(b) - F(a)G(a) - \int_{(a,b]} G(x-)\mu_F(dx), \tag{3.10}$$

where $G(x-) \triangleq \lim_{y \uparrow x} G(y)$.

Proof. One first writes

$$\begin{aligned}\int_{(a,b]} F(x)\mu_G(dx) &= \int_{(a,b]} \big(F(x) - F(a) + F(a)\big)\mu_G(dx) \\ &= F(a)\big(G(b) - G(a)\big) + \int_{(a,b]} \big(\int_{(a,x]} \mu_F(dy)\big)\mu_G(dx). \end{aligned} \tag{3.11}$$

By using Fubini's theorem, one has

$$\begin{aligned}&\int_{(a,b]} \big(\int_{(a,x]} \mu_F(dy)\big)\mu_G(dx) \\ &= \int_{\{(x,y):a<y\leqslant x\leqslant b\}} \mu_G \otimes \mu_F(dx \otimes dy) = \int_{(a,b]} \big(\int_{[y,b]} \mu_G(dx)\big)\mu_F(dy) \\ &= \int_{(a,b]} \big(G(b) - G(y-)\big)\mu_F(dy) = G(b)\big(F(b) - F(a)\big) - \int_{(a,b]} G(y-)\mu_F(dy).\end{aligned}$$

The result follows by substituting the last expression into (3.11). □

The extension from two-fold to n-fold products is straight forward. As an immediate application, one obtains another (equivalent) way of defining the n-dimensional Lebesgue measure on

$$\mathcal{B}(\mathbb{R}^n) = \underbrace{\mathcal{B}(\mathbb{R}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R})}_{n}$$

as the n-fold product of the Lebesgue measure on $\mathcal{B}(\mathbb{R})$.

3.2.3 Construction of pairs of independent random variables

With the notion of product measure spaces, we can now give the mathematical construction of (pairs of) independent random variables. We first recall some basic definitions.

Definition 3.3. Let X, Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Their *joint distribution function* is the function $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F_{X,Y}(x, y) \triangleq \mathbb{P}(X \leqslant x, Y \leqslant y), \quad (x, y) \in \mathbb{R}^2.$$

Their *joint law* is the probability measure on $\mathcal{B}(\mathbb{R}^2)$ defined by

$$\mu_{X,Y}(F) \triangleq \mathbb{P}((X, Y) \in F), \quad F \in \mathcal{B}(\mathbb{R}^2).$$

We say that X, Y are *independent*, if

$$F_{X,Y}(x, y) = F_X(x) F_Y(y)$$

for all $(x, y) \in \mathbb{R}^2$.

The following equivalent characterisations of independence are particularly useful.

Proposition 3.4. *Let X, Y be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following three statements are equivalent.*

(i) X, Y are independent.
(ii) For any $A, B \in \mathcal{B}(\mathbb{R})$, one has

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B). \tag{3.12}$$

(iii) For any bounded, Borel measurable functions $f, g : \mathbb{R} \to \mathbb{R}$, one has

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]. \tag{3.13}$$

Proof. (i) $\Longrightarrow$ (ii). We first consider the case when $B = (-\infty, y]$ is given fixed while A is arbitrary. Recall that

$$\mathcal{C} \triangleq \{(-\infty, x] : x \in \mathbb{R}\}$$

s a π-system over $\mathbb{R}$. Define $\mathcal{H}$ to be the collection of $A \in \mathcal{B}(\mathbb{R})$ that satisfies

$$\mathbb{P}(X \in A, Y \leqslant y) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \leqslant y). \tag{3.14}$$

It is immediate from the definition of independence that $\mathcal{C} \subseteq \mathcal{H}$. Next, we check that $\mathcal{H}$ is a λ-system. It is clear that $\mathbb{R} \in \mathcal{H}$. In addition, let $A, B \in \mathcal{H}$ with $A \subseteq B$. Then one has

$$\begin{aligned}
\mathbb{P}(X \in A, Y \leqslant y) &= \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \leqslant y) \\
\mathbb{P}(X \in B, Y \leqslant y) &= \mathbb{P}(X \in B) \cdot \mathbb{P}(Y \leqslant y).
\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}(X \in B\backslash A, Y \leqslant y) &= \mathbb{P}(X \in B, Y \leqslant y) - \mathbb{P}(X \in A, Y \leqslant y) \\ &= \mathbb{P}(X \in B) \cdot \mathbb{P}(Y \leqslant y) - \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \leqslant y) \\ &= \mathbb{P}(X \in B\backslash A) \cdot \mathbb{P}(Y \leqslant y).\end{aligned}$$

Therefore, $B\backslash A \in \mathcal{H}$. The third property for being a λ-system is also easy to check. According to Dynkin's π-λ theorem, one concludes that $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C}) \subseteq \mathcal{H}$. In other words, all members of $\mathcal{B}(\mathbb{R})$ satisfy the property (3.14).

To prove the full relation (3.12), we now fix $A \in \mathcal{B}(\mathbb{R})$ and define $\mathcal{E}$ to be the collection of $B \in \mathcal{B}(\mathbb{R})$ such that (3.12) holds. The same argument as before shows that $\mathcal{E} = \mathcal{B}(\mathbb{R})$, i.e. all members of $\mathcal{B}(\mathbb{R})$ satisfy (3.12).

(ii) $\Longrightarrow$ (iii). By the linearity of expectation, the property (3.12) implies that (3.13) is true whenever f, g are simple functions. To prove the claim for general bounded, $\mathcal{B}(\mathbb{R})$-measurable functions, recall that such functions can be approximated uniformly by simple functions (cf. Lemma 2.3 and Remark 2.2). The result then follows from dominated convergence.

(iii) $\Longrightarrow$ (i). Take $f = \mathbf{1}_{(-\infty,x]}$ and $g = \mathbf{1}_{(-\infty,y]}$. □

Note that the relation (3.14) means that $\mu_{X,Y} = \mu_X \otimes \mu_Y$ (i.e. joint law equals the product of marginal laws). This fact motivates the following canonical way of constructing independent random variables.

Suppose that F, G are given distribution functions on $\mathbb{R}$. From Chapter 1, one knows how to construct random variables with distribution functions F and G respectively. To ensure the additional independence property, let us consider the product space

$$\Omega = \mathbb{R}^2, \ \mathcal{F} = \mathcal{B}(\mathbb{R}^2), \ \mathbb{P} \triangleq \mu_F \otimes \mu_G,$$

where μ_F, μ_G are the Lebesgue-Stieltjes measures induced by F, G respectively. Define two random variables X, Y on $(\Omega, \mathcal{F}, \mathbb{P})$ simply by taking coordinate projections:

$$X(x, y) \triangleq x, \ Y(x, y) = y.$$

It is clear that the law of X (respectively, of Y) under $\mathbb{P}$ is μ_F (respectively, μ_G); indeed,

$$\mathbb{P}(X \in A) = (\mu_F \otimes \mu_G)(A \times \mathbb{R}) = \mu_F(A)\mu_G(\mathbb{R}) = \mu_F(A) \quad \forall A \in \mathcal{B}(\mathbb{R}).$$

In addition, denoting their joint distribution function as $F_{X,Y}$ one has

$$F_{X,Y}(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y) = (\mu_F \otimes \mu_G)((-\infty, x] \times (-\infty, y]) = F(x)G(y).$$

As a result, X and Y are independent. The construction apparently extends to the case of finitely many random variables.

3.3 Countable product spaces and Kolmogorov's extension theorem

In order to construct sequences of independent random variables, one needs to extend the notion of finite product measures to the countable case. Such an extension requires extra effort, as Fubini's theorem does not work in the first place (one cannot perform iterated integrals for infinitely many times). The main idea is to use Carathéodory's extension theorem.

We begin with the construction of probability measures on $\mathbb{R}^\infty$ (Kolmogorov's extension theorem), which yields a canonical construction of random sequences with given finite dimensional distributions. Then we discuss the special case of countable product measures (sequences of independent random variables). For simplicity without losing the essential picture, we restrict ourselves to the countable product of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ instead of general measurable spaces. We give some comments on possible generalisations at the end of this section.

3.3.1 The measurable space $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$

To fix notation, we set $\mathbb{R}^\infty \triangleq \prod_{n=1}^\infty \mathbb{R}$ (countably many copies of $\mathbb{R}$). Elements in $\mathbb{R}^\infty$ are of the form

$$x = (x_1, x_2, x_3, \cdots), \quad x_i \in \mathbb{R}.$$

For each $n \geqslant 1$, one can view $\mathbb{R}^\infty = \mathbb{R}^n \times \mathbb{R}^{>n}$ where $\mathbb{R}^n$ represents the first n-components $(x_1, \cdots, x_n)$ and $\mathbb{R}^{>n}$ represents the remainder $(x_{n+1}, x_{n+2}, \cdots)$. To define the product σ-algebra, we first introduce the algebra $\mathcal{A}$ of cylindrical subsets defined by

$$\mathcal{A} \triangleq \bigcup_{n=1}^\infty \mathcal{F}_n, \tag{3.15}$$

where

$$\mathcal{F}_n \triangleq \{G_n \times \mathbb{R}^{>n} : G_n \in \mathcal{B}(\mathbb{R}^n)\}, \quad n \geqslant 1.$$

Note that elements in $\mathcal{A}$ are of the form $\Gamma = G_n \times \mathbb{R}^{>n}$ for some $n \geqslant 1$ and $G_n \in \mathcal{B}(\mathbb{R}^n)$. This cylindrical representation may not be unique, e.g. $G_1 \times \mathbb{R}^{>1} = (G_1 \times \mathbb{R}) \times \mathbb{R}^{>2}$.

Lemma 3.2. *$\mathcal{A}$ is an algebra over $\mathbb{R}^\infty$.*

Proof. Obviously, $\mathbb{R}^\infty \in \mathcal{A}$. Next, suppose that $\Gamma \in \mathcal{A}$, say $\Gamma = G_n \times \mathbb{R}^{>n}$. Then $\Gamma^c = G_n^c \times \mathbb{R}^{>n} \in \mathcal{A}$. Finally, let $\Gamma, \Lambda \in \mathcal{A}$ with representations

$$\Gamma = G_m \times \mathbb{R}^{>m}, \ \Lambda = G_n \times \mathbb{R}^{>n}$$

and assume without loss of generality that $m \geqslant n$. Then

$$\Gamma \cap \Lambda = \big(G_m \cap (G_n \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{m-n})\big) \times \mathbb{R}^{>m} \in \mathcal{A}.$$

□

Definition 3.4. The *product σ-algebra over* $\mathbb{R}^\infty$ is defined by $\mathcal{B}(\mathbb{R}^\infty) \triangleq \sigma(\mathcal{A})$.

Remark 3.4. The σ-algebra $\mathcal{B}(\mathbb{R}^\infty)$ is the smallest σ-algebra $\mathcal{F}$ such that the projection

$$\pi_n : \mathbb{R}^\infty \to \mathbb{R}, \ (x_1, x_2, \cdots, x_n, \cdots) \mapsto x_n$$

is $\mathcal{F}$-measurable for all n (why?).

3.3.2 Kolmogorov's extension theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space. Let $X = \{X_n : n \geqslant 1\}$ be a sequence of random variables defined on it. One can equivalently view X as a "random variable" taking values in $\mathbb{R}^\infty$ by

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty)), \quad X(\omega) = (X_1(\omega), X_2(\omega), \cdots).$$

Note that X is a measurable map.

Definition 3.5. The *law* of X is the probability measure on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ defined by

$$\mu_X(\Gamma) \triangleq \mathbb{P}(X \in \Gamma), \quad \Gamma \in \mathcal{B}(\mathbb{R}^\infty).$$

For each $n \geqslant 1$, the joint law of the first n components $(X_1, \cdots, X_n)$ is the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ defined by

$$\nu_X^{(n)}(G_n) \triangleq \mathbb{P}\big((X_1, \cdots, X_n) \in G_n\big) = \mu_X(G_n \times \mathbb{R}^{>n}), \quad G_n \in \mathcal{B}(\mathbb{R}^n). \tag{3.16}$$

The family $\{\nu_X^{(n)} : n \geqslant 1\}$ satisfies the following obvious consistency relation:

$$\nu_X^{(n+m)}(G_n \times \mathbb{R}^m) = \nu_X^{(n)}(G_n) \tag{3.17}$$

for all $m, n \geqslant 1$ and $G_n \in \mathcal{B}(\mathbb{R}^n)$. The law of X is uniquely determined by this family of finite dimensional laws; indeed, according to Proposition 1.4 there is at most one probability measure μ_X on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that (3.16) holds for all n and all $G_n \in \mathcal{B}(\mathbb{R}^n)$.

Here comes a key question. Suppose that $\{\nu^{(n)} : n \geqslant 1\}$ is a given sequence of probability measures which satisfy the consistency relation (3.17). *How can one construct a sequence* $X = \{X_n : n \geqslant 1\}$ *of random variables on some probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that the joint law of* $(X_1, \cdots, X_n)$ *is* $\nu^{(n)}$ *for all* n*?* The answer is essentially contained in the following so-called *Kolmogorov's extension theorem* (in the countable case).

Theorem 3.3. *For each* $n \geqslant 1$*, let* $\nu^{(n)}$ *be a given probability measure on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$*. Suppose that they satisfy the consistency relation (3.17). Then there exists a unique probability measure* μ *on* $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$*, such that*

$$\mu(G_n \times \mathbb{R}^{>n}) = \nu^{(n)}(G_n) \quad \forall n \geqslant 1, G_n \in \mathcal{B}(\mathbb{R}^n).$$

To see how Theorem 3.3 addresses the above question, one simply takes

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mu)$$

where μ is the probability measure given by the theorem, and define

$$X_n : \Omega \to \mathbb{R}, \ \omega = (x_1, x_2, \cdots, x_n, \cdots) \mapsto X_n(\omega) \triangleq x_n$$

to be the canonical projection (equivalently, $X : \Omega \to \mathbb{R}^\infty$ is the identity map). It is clear from the construction as well as the theorem that the joint law of $(X_1, \cdots, X_n)$ is $\nu^{(n)}$ for all n (equivalently, the law of X is μ).

Example 3.5. For a Markov chain $\{X_n\}$, the one-step transition probabilities together with the initial distribution determine the joint law $\nu^{(n)}$ of $(X_1, \cdots, X_n)$ for each n. These probability measures $\{\nu^{(n)}\}$ satisfy the consistency condition (3.17). Therefore, the above discussion provides a mathematical construction of Markov chains with given transition matrix and initial distribution.

The rest of this section is devoted to the proof of Theorem 3.3. The uniqueness part has already been discussed. We now prove existence.

In the first place, the family $\{\nu^{(n)}\}$ induces an obvious "probability measure" on the algebra $\mathcal{A}$ (cf. (3.15)) defined by

$$\hat{\mu}(\Gamma) \triangleq \nu^{(n)}(G_n), \quad \Gamma = G_n \times \mathbb{R}^{>n}.$$

Note that $\hat{\mu}$ is well-defined (i.e. independent of representations of Γ) as a result of the consistency relation (3.17). We shall use Carathéodory's extension theorem to obtain a probability measure μ on $\mathcal{B}(\mathbb{R}^\infty)$ as an extension of $\hat{\mu}$. To this end,

since $\hat{\mu}$ is finitely additive on $\mathcal{A}$ (why?), the key ingredient is proving its countably additivity. We do so by using the criterion given by Proposition 1.3 (iv), i.e. we want to show that

$$\Gamma_n \in \mathcal{A},\ \Gamma_n \downarrow \emptyset \implies \lim_{n\to\infty} \hat{\mu}(\Gamma_n) = 0. \tag{3.18}$$

We prove (3.18) by contradiction. Let $\Gamma_n \in \mathcal{A}$ be such that $\Gamma_n \downarrow \emptyset$. Suppose on the contrary that

$$\lim_{n\to\infty} \hat{\mu}(\Gamma_n) = \varepsilon > 0. \tag{3.19}$$

Our eventual goal is to produce an element $x^* \in \cap_n \Gamma_n$ which then leads to a contradiction. The argument for this purpose contains the following steps.

(i) One may assume without loss of generality that $\Gamma_n = G_n \times \mathbb{R}^{>n}$ where $G_n \in \mathcal{B}(\mathbb{R}^n)$. Indeed, since $\Gamma_n \in \mathcal{A}$ there exists $m_n \geqslant 1$ such that $\Gamma_n = G_{m_n} \times \mathbb{R}^{>m_n}$. By adding sufficiently $\mathbb{R}$'s following G_{m_n} if necessary, one may assume that

$$1 \leqslant m_1 < m_2 < m_3 < \cdots \uparrow \infty.$$

In this case, one defines

$$\begin{aligned}
\hat{\Gamma}_1 &\triangleq \mathbb{R} \times \mathbb{R}^{>1},\ \hat{\Gamma}_2 \triangleq \mathbb{R}^2 \times \mathbb{R}^{>2},\ \cdots, \hat{\Gamma}_{m_1-1} \triangleq \mathbb{R}^{m_1-1} \times \mathbb{R}^{>m_1-1}, \\
\hat{\Gamma}_{m_1} &\triangleq \Gamma_1 = G_{m_1} \times \mathbb{R}^{>m_1}, \\
\hat{\Gamma}_{m_1+1} &\triangleq (G_{m_1} \times \mathbb{R}) \times \mathbb{R}^{>m_1+1}, \cdots, \hat{\Gamma}_{m_2-1} \triangleq (G_{m_1} \times \mathbb{R}^{m_2-m_1-1}) \times \mathbb{R}^{>m_2-1}, \\
\hat{\Gamma}_{m_2} &\triangleq \Gamma_2 = G_{m_2} \times \mathbb{R}^{>m_2}, \cdots.
\end{aligned}$$

It is clear from the definition that $\hat{\Gamma}_n$ has the form $\hat{G}_n \times \mathbb{R}^{>n}$ for all n. In addition, since the $\hat{\Gamma}_n$'s are obtained from the Γ_n's by trivially inserting $\mathbb{R}$'s, one also has $\hat{\Gamma}_{n+1} \subseteq \hat{\Gamma}_n$ and

$$\bigcap_{n=1}^{\infty} \hat{\Gamma}_n = \bigcap_{n=1}^{\infty} \Gamma_n = \emptyset.$$

We rename $\hat{\Gamma}_n$ as Γ_n to ease notation.

(ii) Here comes an essential point where topological properties of $\mathbb{R}^n$ become relevant. We quote the following deep result about probability measures on complete, separable metric spaces (in particular, on $\mathbb{R}^n$). Its proof can be found in [Fol99].

Proposition 3.5. *Let ν be a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. For any $G \in \mathcal{B}(\mathbb{R}^n)$ and $\delta > 0$, there exists a compact subset K of G, such that $\nu(G \backslash K) < \delta$.*

By the assumption (3.19), one has

$$\nu^{(n)}(G_n) = \hat{\mu}(\Gamma_n) \geqslant \varepsilon \quad \forall n.$$

According to Proposition 3.5, there exists a compact subset $K_n \subseteq G_n$ (also setting $\Lambda_n \triangleq K_n \times \mathbb{R}^{>n}$), such that

$$\hat{\mu}(\Gamma_n \backslash \Lambda_n) = \nu^{(n)}(G_n \backslash K_n) < \frac{\varepsilon}{2^{n+1}}.$$

Let us set

$$\tilde{K}_n \triangleq (K_1 \times \mathbb{R}^{n-1}) \cap (K_2 \times \mathbb{R}^{n-2}) \cap \cdots \cap (K_{n-1} \times \mathbb{R}) \cap K_n$$

and $\tilde{\Lambda}_n \triangleq \tilde{K}_n \times \mathbb{R}^{>n}$. Since $\tilde{\Lambda}_n = \Lambda_1 \cap \cdots \cap \Lambda_n$, it follows that

$$\begin{aligned}
\nu^{(n)}(\tilde{K}_n) = \hat{\mu}(\tilde{\Lambda}_n) &= \hat{\mu}(\Gamma_n) - \hat{\mu}(\Gamma_n \backslash \tilde{\Lambda}_n) = \hat{\mu}(\Gamma_n) - \hat{\mu}\big(\bigcup_{k=1}^{n} \Gamma_n \backslash \Lambda_k\big) \\
&\geqslant \hat{\mu}(\Gamma_n) - \sum_{k=1}^{n} \hat{\mu}(\Gamma_n \backslash \Lambda_k) \\
&\geqslant \hat{\mu}(\Gamma_n) - \sum_{k=1}^{n} \hat{\mu}(\Gamma_k \backslash \Lambda_k) \quad (\text{since } \Gamma_n \subseteq \Gamma_k) \\
&> \varepsilon - \sum_{k=1}^{n} \frac{\varepsilon}{2^{k+1}} > \varepsilon - \sum_{k=1}^{\infty} \frac{\varepsilon}{2^{k+1}} = \frac{\varepsilon}{2} > 0.
\end{aligned}$$

In particular, $\tilde{K}_n \neq \emptyset$ for all n.

(iii) According to Step (ii), one can choose a point $(x_1^{(n)}, \cdots, x_n^{(n)}) \in \tilde{K}_n$ for each n. By the definition of $\tilde{K}_n$, the sequence $x_1^{(n)}$ is contained in K_1. Since K_1 is compact, there exists a subsequence $x_1^{m_1(n)}$ which converges to some point $x_1 \in K_1$. Next, consider the sequence $(x_1^{m_1(n)}, x_2^{m_1(n)}) \in K_2$. Again by compactness, it contains a further subsequence

$$(x_1^{m_2(n)}, x_2^{m_2(n)}) \to \text{some } (x_1', x_2) \in K_2.$$

Indeed $x_1' = x_1$ since $x_1^{m_2(n)}$ is a subsequence of $x_1^{m_1(n)}$. Arguing inductively, at the n-th step one finds a point $(x_1, \cdots, x_n) \in K_n$ ($(x_1, \cdots, x_{n-1})$ coincides with the point coming from the $(n-1)$-th step). Finally, one defines $x^* \triangleq (x_1, x_2, x_3, \cdots)$. Since $(x_1, \cdots, x_n) \in K_n$ for all n, it follows that

$$x^* \in \bigcap_{n=1}^{\infty} \Lambda_n \subseteq \bigcap_{n=1}^{\infty} \Gamma_n.$$

This contradicts the assumption that $\cap_n \Gamma_n = \emptyset$. As a consequence, $\hat{\mu}$ is continuous at $\emptyset$. According to Proposition 1.3 (iv), $\hat{\mu}$ is countably additive on $\mathcal{A}$. The existence part of Theorem 3.3 now follows from Carathéodory's extension theorem.

3.3.3 Construction of independent sequences

Suppose that $\{\nu_n : n \geqslant 1\}$ is a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define the product measure $\nu^{(n)} \triangleq \nu_1 \otimes \cdots \otimes \nu_n$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ for each n. Then $\{\nu^{(n)}\}$ satisfy the consistency relation (3.17). According to Theorem 3.3, there exists a unique probability measure μ on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that

$$\mu(B_1 \times \cdots \times B_n \times \mathbb{R}^n) = \nu_1(B_1) \cdots \nu_n(B_n) \quad \forall n \geqslant 1,\ B_1, \cdots, B_n \in \mathcal{B}(\mathbb{R}).$$

Definition 3.6. The probability space $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mu)$ is called the *infinite product space* of $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \nu_n)$ $(n \geqslant 1)$.

To construct an independent sequence with marginal laws $\{\nu_n\}$, one takes

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mu)$$

and define

$$X_n(\omega) \triangleq x_n, \quad \omega = (x_1, x_2, x_3, \cdots) \in \Omega$$

as before. It is readily checked that the sequence of random variables $\{X_n : n \geqslant 1\}$ are independent with $X_n \overset{\text{law}}{=} \nu_n$ for all n. This mathematically justifies the existence of independent sequences with given marginal laws (e.g. an i.i.d. sequence of Bernoulli random variables).

Remark 3.5. We have not yet define the independence in general. This will done in Section 5.1.1 below; for now a random sequence $\{X_n\}$ is independent means that $X_1, \cdots, X_n$ are independent for each n.

3.3.4 Some generalisations

In the first place, Kolmogorov's extension theorem has a natural generalisation to the case of *uncountable* products (the argument is essentially the same the one given above). Such a generalisation is particularly useful in the study of stochastic processes (e.g. construction of Brownian motion).

Secondly, there is a version of infinite products of general measurable spaces instead of just $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The construction of probability measures on such product spaces requires the notion of *transition probability kernels*. The argument is

measure-theoretic and does not require topological considerations. However, the general existence of such kernels relies on topological properties of the underlying measurable spaces (typically, one requires them to be complete, separable metric spaces). In any case, some sort of "compactness" property is needed for the extension theorem to hold.

There is one exceptional situation where the construction is entirely measure-theoretic and no topological assumptions are needed: the infinite product of probability spaces $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ $(n \geqslant 1)$. Similar to the case of $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \nu_n)$, let us set

$$\Omega \triangleq \prod_{k=1}^{\infty} \Omega_k, \ \mathcal{F} \triangleq \sigma(\mathcal{A}),$$

where $\mathcal{A}$ is the algebra of cylindrical subsets defined in a similar way by

$$\mathcal{A} \triangleq \Big\{ \Gamma_n = G_n \times \prod_{k>n} \Omega_k : n \geqslant 1, \ G_n \in \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n \Big\}.$$

Theorem 3.4. *There exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, such that*

$$\mathbb{P}\big(A_1 \times \cdots \times A_n \times \prod_{k>n} \Omega_k\big) = \mathbb{P}_1(A_1) \cdots \mathbb{P}_n(A_n)$$

for all $n \geqslant 1$ and $A_i \in \mathcal{F}_i$ $(1 \leqslant i \leqslant n)$.

Proof. The argument is parallel to the proof of Theorem 3.3 with one exceptional difference (replacing the compactness argument by a measure-theoretic one). As before, we first define a set function $\hat{\mathbb{P}} : \mathcal{A} \to [0,1]$ by

$$\hat{\mathbb{P}}(G_n \times \prod_{k>n} \Omega_k) \triangleq (\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)(G_n).$$

Note that $\hat{\mathbb{P}}$ is well-defined and finitely additive on $\mathcal{A}$. Following the proof of Theorem 3.3, the core step is to show that

$$\Gamma_n \in \mathcal{A}, \ \Gamma_n \downarrow \emptyset \implies \lim_{n\to\infty} \hat{\mathbb{P}}(\Gamma_n) = 0,$$

where one assumes without loss of generality that Γ_n has the form $\Gamma_n = G_n \times \prod_{k>n} \Omega_k$. Suppose on the contrary that

$$\hat{\mathbb{P}}(\Gamma_n) \geqslant \varepsilon > 0 \quad \forall n.$$

We want to find a point $\omega^* \in \cap_n \Gamma_n$, which then leads to a contradiction.

For $n \geqslant 2$, define the function $\phi_{1,n} : \Omega_1 \to \mathbb{R}$ by

$$\phi_{1,n}(\omega_1) \triangleq \int_{\Omega_2 \times \cdots \times \Omega_n} \mathbf{1}_{G_n}(\omega_1, \omega_2, \cdots, \omega_n) \mathbb{P}_2(d\omega_2) \cdots \mathbb{P}_n(d\omega_n).$$

One can easily check that

$$0 \leqslant \phi_{1,n+1}(\omega_1) \leqslant \phi_{1,n}(\omega_1) \leqslant 1 \quad \forall n \geqslant 1, \omega_1 \in \Omega_1.$$

Set $\phi_1 \triangleq \lim_{n\to\infty} \phi_{1,n}$. By the dominated convergence theorem, one has

$$\begin{aligned} \int_{\Omega_1} \phi_1 d\mathbb{P}_1 &= \lim_{n\to\infty} \int_{\Omega_1} \Big(\int_{\Omega_2 \times \cdots \times \Omega_n} \mathbf{1}_{G_n}(\omega_1, \omega_2, \cdots, \omega_n) \mathbb{P}_2(d\omega_2) \cdots \mathbb{P}_n(d\omega_n) \Big) \mathbb{P}_1(d\omega_1) \\ &= \lim_{n\to\infty} (\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n)(G_n) = \lim_{n\to\infty} \hat{\mathbb{P}}(\Gamma_n) \geqslant \varepsilon > 0. \end{aligned} \tag{3.20}$$

As a result, there exists $\omega_1^* \in \Omega_1$, such that $\phi_1(\omega_1^*) > 0$. We claim that $\omega_1^* \in G_1$. Indeed, if $\omega_1^* \notin G_1$, then $(\omega_1^*, \omega_2, \cdots, \omega_n) \notin G_n$ for all $n \geqslant 2$ and $\omega_i \in \Omega_i$ (since $G_n \subseteq G_1 \times \prod_{k=2}^n \Omega_k$ by assumption), which implies that $\phi_{1,n}(\omega_1^*) = 0$ for all n, contradicting the property (3.20).

Next, for $n \geqslant 3$ one defines $\phi_{2,n} : \Omega_2 \to \mathbb{R}$ by

$$\phi_{2,n}(\omega_2) \triangleq \int_{\Omega_3 \times \cdots \times \Omega_n} \mathbf{1}_{G_n}(\omega_1^*, \omega_2, \cdots, \omega_n) \mathbb{P}_3(d\omega_3) \cdots \mathbb{P}_n(d\omega_n).$$

For the same reason as before, with $\phi_2 \triangleq \lim_{n\to\infty} \phi_{2,n}$ one has

$$\int_{\Omega_2} \phi_2 d\mathbb{P}_2 = \lim_{n\to\infty} \int_{\Omega_2 \times \cdots \times \Omega_n} \mathbf{1}_{G_n}(\omega_1^*, \omega_2, \cdots, \omega_n) \mathbb{P}_2(d\omega_2) \cdots \mathbb{P}_n(d\omega_n) = \phi_1(\omega_1^*) > 0.$$

As a result, there exists $\omega_2^* \in \Omega_2$ such that $\phi_2(\omega_2^*)$. One also sees in the same way as before that $(\omega_1^*, \omega_2^*) \in G_2$. It is now a routine matter of induction to see that one can choose $\omega_n^* \in \Omega_n$, such that $(\omega_1^*, \cdots, \omega_n^*) \in G_n$ ($\omega_1^*, \cdots, \omega_{n-1}^*$ are chosen in the first $(n-1)$ steps). Consequently, one finds that

$$\omega^* = (\omega_1^*, \omega_2^*, \omega_3^*, \cdots) \in \bigcap_{n=1}^{\infty} \Gamma_n,$$

which gives a contradiction. □

Example 3.6. Consider the random experiment of tossing a fair coin independently in a sequence. For each $n \geqslant 1$, we define $\Omega_n = \{H, T\}$, $\mathcal{F}_n = \{\emptyset, \{H\}, \{T\}, \Omega_n\}$ and $\mathbb{P}_n(\{H\}) = \mathbb{P}_n(\{T\}) = 1/2$. The classical probability model $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ represents the n-th toss. The countable product space

$$(\Omega, \mathcal{F}, \mathbb{P}) \triangleq \bigotimes_{n=1}^{\infty} (\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$$

defines a canonical probability space which models the underlying random experiment.

4 Modes of convergence

A central theme of study in this subject is related to understanding asymptotic behaviours of sequences of random variables. Unlike sequences of real numbers where there is no ambiguity of talking about convergence, there are various natural ways of defining convergence for random sequences. For instance, the law of large numbers is concerned with almost sure convergence as well as convergence in probability (strong and weak laws respectively), while the central limit theorem is concerned with weak convergence (or convergence in distribution). On the other hand, L^p-convergence (primarily $p = 1, 2$) plays an essential role in martingale theory and stochastic calculus.

Before moving to concrete probabilistic settings and examples, in this chapter we develop some general tools for studying convergence of random variables and probability measures. In Section 4.1, we introduce the four types of convergence we shall encounter in the sequel and discuss some of their basic relations. In Section 4.2, we discuss the connection between L^1-convergence and convergence in probability through the important concept of uniform integrability. Section 4.3 is the core of this chapter where we study weak convergence of probability measures in depth.

4.1 Basic convergence concepts

Let X_n, X be random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. There are various ways of interpreting the convergence of X_n towards X. Among others, we are going to discuss four basic types of convergence: almost sure convergence, convergence in probability, L^1-convergence and weak convergence. The basic relations are that

$$\begin{cases} \text{a.s. convergence} \implies \text{convergence in probability} \implies \text{weak convergence}, \\ \text{convergence in probability} + \text{"uniform integrability"} \iff L^1\text{-convergence}. \end{cases}$$

Different modes of convergence have different applications depending on the nature of the underlying problem.

We begin with the strongest one: almost sure convergence.

Definition 4.1. We say that X_n converges to X *almost surely* or *with probability one*, if there exists a null event N such that for any $\omega \notin N$ one has

$$\lim_{n\to\infty} X_n(\omega) = X(\omega).$$

Equivalently,

$$\mathbb{P}\big(\big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}\big) = 1.$$

We often use the shorthanded notation "$X_n \to X$ a.s." to denote almost sure convergence.

In contrast to almost sure convergence, one has the weaker notion of convergence in probability.

Definition 4.2. We say that X_n converges to X *in probability*, if for any $\varepsilon > 0$ one has

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

We often use the shorthanded notation "$X_n \to X$ in prob." to denote convergence in probability.

The third type of convergence we shall consider is L^1-convergence.

Definition 4.3. We say that X_n *converges to* X *in* L^1 if

$$\lim_{n\to\infty} \mathbb{E}\big[|X_n - X|\big] = 0.$$

The basic relation between the above three types of convergence is summarised as follows.

Proposition 4.1. *Either almost sure convergence or L^1-convergence implies convergence in probability.*

Proof. Suppose that X_n converges to X a.s. Let $\varepsilon > 0$ be given fixed. Since

$$\big\{\lim_{n\to\infty} X_n = X\big\} \subseteq \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \big\{|X_m - X| \leqslant \varepsilon\big\},$$

one finds that

$$\begin{aligned} 1 = \mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) &\leqslant \lim_{n\to\infty} \mathbb{P}\big(\bigcap_{m=n}^{\infty} \big\{|X_m - X| \leqslant \varepsilon\big\}\big) \\ &\leqslant \lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \leqslant \varepsilon\big). \end{aligned}$$

As a result, $\mathbb{P}(|X_n - X| \leqslant \varepsilon)] \to 1$ or equivalently $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. This gives convergence in probability.

Now suppose that X_n converges to X in L^1. By Markov's inequality, one has

$$\mathbb{P}\big(|X_n - X| > \varepsilon\big) \leqslant \frac{1}{\varepsilon}\mathbb{E}\big[|X_n - X|\big],$$

which gives convergence in probability immediately. □

As the example below suggests, the converse of Proposition 4.1 is not true in general.

Example 4.1. (i) *Convergence in probability does not imply almost sure convergence.* Consider the random experiment of choosing a point $\omega \in \Omega = [0,1]$ uniformly at random. We construct a sequence $\{Y_n : n \geqslant 1\}$ of random variables as follows. Firstly, divide $[0,1]$ into two sub-intervals, and define $Y_1 \triangleq \mathbf{1}_{[0,1/2]}$ and $Y_2 \triangleq \mathbf{1}_{[1/2,1]}$. Next, divide $[0,1]$ into three sub-intervals, and define $Y_3 \triangleq \mathbf{1}_{[0,1/3]}$, $Y_4 \triangleq \mathbf{1}_{[1/3,2/3]}$ and $Y_5 \triangleq \mathbf{1}_{[2/3,1]}$. Now the procedure continues in the obvious way to define the whole sequence $\{Y_n\}$. Since the event

$$\{\omega \in [0,1] : |Y_n(\omega)| > \varepsilon\} = \{\omega : Y_n(\omega) = 1\}$$

is given by a particular sub-interval whose length tends to zero, we conclude that Y_n converges to zero in probability. However, $Y_n(\omega)$ does not converge to zero at any $\omega \in [0,1]$. Indeed, for each ω, by the construction there must exist a subsequence n_k such that $Y_{n_k}(\omega) = 1$ for all k.
(ii) *Convergence in probability does not imply L^1-convergence.* Take $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), dx)$. Define $X_n(\omega) \triangleq n\mathbf{1}_{[0,1/n]}(\omega)$. Since $\mathbb{P}(|X_n| > \varepsilon) = 1/n$, one knows that $X_n \to 0$ in probability. However, X_n does not converges to 0 in L^1 since $\mathbb{E}[|X_n|] = 1$ for all n.

The above three types of convergence rely on the realisations of X_n, X on a common probability space (coupling between X_n and X). In particular, one cannot talk about such convergence by only looking at the distributions of X_n and X separately. There is the weakest notion of convergence which has such a "distributional" nature: weak convergence.

Definition 4.4. Let $F_n(x), F(x)$ be the distribution functions of X_n, X respectively. We say that X_n *converges weakly to* X, if $F_n(x)$ converges to $F(x)$ at every continuity point x of F. Weak convergence is also known as *convergence in distribution*.

The concept of weak convergence is only distributional; it depends only on the distribution functions $F_n(x)$ and $F(x)$ but has nothing to do with how one chooses the probability space and constructs the random variables on it. The fact that weak convergence is the weakest among the four requires more tools from weak convergence theory. We will prove it in Proposition 4.6 below. The following example shows that weak convergence does not imply convergence in probability in general.

Example 4.2. Let W be a Bernoulli random variable with parameter $1/2$. Define $Z_n \triangleq W$ for all n and $Z \triangleq 1 - W$. Since Z_n and Z are both Bernoulli random variables with parameter $1/2$, it is trivial that Z_n converges weakly to Z. However, for any $\varepsilon \in (0, 1)$ one has

$$\mathbb{P}\big(|Z_n - Z| > \varepsilon\big) = \mathbb{P}(|2W - 1| > \varepsilon) = 1.$$

Therefore, Z_n does not converge to Z in probability.

There is a special situation where the two concepts are equivalent, i.e. when the limiting random variable is a deterministic constant.

Proposition 4.2. *Suppose that X_n converges weakly to a deterministic constant c. Then $X_n \to c$ in probability.*

Proof. The distribution function of the constant random variable $X \equiv c$ is given by

$$F(x) \triangleq \begin{cases} 0, & x < c; \\ 1, & x \geqslant c. \end{cases}$$

By assumption, one knows that

$$\mathbb{P}(X_n \leqslant x) \to F(x)$$

for all $x \neq c$ (any $x \neq c$ is a continuity point of F). Therefore, given $\varepsilon > 0$, one has

$$\begin{aligned} \mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n > c + \varepsilon) + \mathbb{P}(X_n < c - \varepsilon) \\ &\leqslant 1 - \mathbb{P}(X_n \leqslant c + \varepsilon) + \mathbb{P}(X_n \leqslant c - \varepsilon) \\ &\to 1 - F(c + \varepsilon) + F(c - \varepsilon) \\ &\to 1 - 1 + 0 = 0. \end{aligned}$$

This shows that $X_n \to c$ in probability. □

4.2 Uniform integrability and L^1-convergence

Sometimes it is useful to have L^1-convergence (e.g. if one wants to have $\mathbb{E}[X_n] \to \mathbb{E}[X]$). We have seen that L^1-convergence implies convergence in probability but the converse is not true in general (cf. Example 4.1 (ii)). The bridge connecting these two concepts is the so-called *uniform integrability*.

To motivate its definition, we first consider a single integrable random variable X. An application of dominated convergence shows that

$$\lim_{\lambda\to\infty} \mathbb{E}[|X|\mathbf{1}_{\{|X|>\lambda\}}] = 0. \tag{4.1}$$

More generally, the integral $\mathbb{E}[|X|\mathbf{1}_F]$ can be made arbitrarily small whenever $\mathbb{P}(F)$ is small enough (cf. Proposition 2.8). The idea of uniform integrability for a family $\{X_t\}$ of random variables is that the convergence (4.1) should be required to hold uniformly with respect to the entire family $\{X_t\}$.

Definition 4.5. Let $\{X_t : t \in \mathcal{T}\}$ be a family of integrable random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the family $\{X_t\}$ is *uniformly integrable*, if

$$\lim_{\lambda\to\infty} \sup_{t\in\mathcal{T}} \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\lambda\}}\big] = 0.$$

Similar to the case of a single random variable, uniform integrability essentially suggests that the integral $\mathbb{E}[|X_t|\mathbf{1}_F]$ can be made arbitrarily small *in a uniform manner* as long as $\mathbb{P}(F)$ is small enough. This is made precise by the following characterisation.

Theorem 4.1. *A family $\{X_t : t \in \mathcal{T}\}$ of integrable random variables is uniformly integrable if and only if the following two statements hold true.*

(i) The family $\{X_t\}$ is bounded in L^1, i.e. there exists $M > 0$ such that

$$\mathbb{E}[|X_t|] \leqslant M \quad \forall t \in \mathcal{T}.$$

(ii) For any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\forall F \in \mathcal{F}, \mathbb{P}(F) < \delta \implies \mathbb{E}[|X_t|\mathbf{1}_F] < \varepsilon \quad \forall t \in \mathcal{T}. \tag{4.2}$$

Proof. Necessity. Suppose that the family $\{X_t : t \in \mathcal{T}\}$ is uniformly integrable. By definition, given any $\varepsilon > 0$, there exists $\Lambda > 0$ such that

$$\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\Lambda\}}\big] \leqslant \varepsilon \quad \forall t \in \mathcal{T}.$$

It follows that

$$\mathbb{E}[|X_t|] = \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|\leqslant\Lambda\}}\big] + \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\Lambda\}}\big] \leqslant \Lambda + \varepsilon,$$

which gives the L^1-boundedness. In addition, for the above Λ and any $F \in \mathcal{F}$, one has

$$\begin{aligned}\mathbb{E}[|X_t|\mathbf{1}_F] &= \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|\leqslant\Lambda\}\cap F}\big] + \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\Lambda\}\cap F}\big] \\ &\leqslant \Lambda\mathbb{P}(F) + \mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\Lambda\}}\big] < \Lambda\mathbb{P}(F) + \varepsilon.\end{aligned}$$

Take $\delta \triangleq \varepsilon/\Lambda$. Whenever $\mathbb{P}(F) < \delta$, one has

$$\mathbb{E}\big[|X_t|\mathbf{1}_F\big] < 2\varepsilon \quad \forall t \in \mathcal{T}.$$

This gives the uniform continuity property (4.2).

Sufficiency. Since $\{X_t\}$ is bounded in L^1, by Markov's inequality one has

$$\mathbb{P}\big(|X_t| > \lambda\big) \leqslant \frac{1}{\lambda}\mathbb{E}[|X_t|] \leqslant \frac{1}{\lambda}M \quad \forall t \in \mathcal{T}.$$

Therefore, given $\varepsilon > 0$ there exists $\Lambda > 0$ such that

$$\lambda > \Lambda \implies \mathbb{P}\big(|X_t| > \lambda\big) < \delta,$$

where δ is the number appearing in (4.2). By using that assumed property, one obtains that

$$\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\lambda\}}\big] < \varepsilon \quad \forall t \in \mathcal{T}.$$

This gives the uniform integrability. □

Verifying either the definition or the conditions in Theorem 4.1 is not always easy. We present two useful sufficient conditions for uniform integrability.

Proposition 4.3. *Let $\{X_t : t \in \mathcal{T}\}$ be a family of random variables. Suppose that one of the following two conditions holds true.*

(i) There exists $p > 1$ and $M > 0$ such that

$$\mathbb{E}[|X_t|^p] \leqslant M \quad \forall t \in \mathcal{T}.$$

(ii) There exists an integrable random variable Y such that

$$|X_t| \leqslant Y \ \text{ a.s. } \forall t \in \mathcal{T}.$$

Then the family $\{X_t\}$ is uniformly integrable.

Proof. We only consider the first case and leave the second one as an exercise. By the assumption, one has

$$\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\lambda\}}\big] \leqslant \mathbb{E}\big[(|X_t|/\lambda)^{p-1}|X_t|\big] = \frac{1}{\lambda^{p-1}}\mathbb{E}[|X_t|^p] \leqslant \frac{M}{\lambda^{p-1}} \quad \forall t \in \mathcal{T}.$$

Therefore,

$$\lim_{\lambda\to\infty}\sup_{t\in\mathcal{T}}\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|>\lambda\}}\big] = 0.$$

□

Below is a particularly useful way of constructing uniformly integrable families. It plays an important role in the study of martingales.

Proposition 4.4. *Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\{\mathcal{G}_t : t \in \mathcal{T}\}$ be a family of sub-σ-algebras of $\mathcal{F}$. Then the family $\{\mathbb{E}[X|\mathcal{G}_t] : t \in \mathcal{T}\}$ is uniformly integrable.*

Proof. Denote $X_t \triangleq \mathbb{E}[X|\mathcal{G}_t]$. By using properties of the conditional expectation, one has

$$\begin{aligned}\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|\}>\lambda}\big] &\leqslant \mathbb{E}\big[\mathbb{E}[|X|\big|\mathcal{G}_t]\mathbf{1}_{\{|X_t|>\lambda\}}\big] \\ &= \mathbb{E}\big[\mathbb{E}[|X|\mathbf{1}_{\{|X_t|>\lambda\}}\big|\mathcal{G}_t]\big] = \mathbb{E}\big[|X|\mathbf{1}_{\{|X_t|>\lambda\}}\big]. \end{aligned} \tag{4.3}$$

On the other hand, since X is integrable, by Proposition 2.8 for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\forall F \in \mathcal{F}, \mathbb{P}(F) < \delta \implies \mathbb{E}\big[|X|\mathbf{1}_F\big] < \varepsilon. \tag{4.4}$$

By Markov's inequality, one has

$$\mathbb{P}(|X_t| > \lambda) \leqslant \frac{1}{\lambda}\mathbb{E}[|X_t|] \leqslant \frac{1}{\lambda}\mathbb{E}\big[\mathbb{E}[|X|\big|\mathcal{G}_t]\big] = \frac{1}{\lambda}\mathbb{E}[|X|].$$

In particular, there exists $\Lambda > 0$ such that

$$\lambda > \Lambda \implies \mathbb{P}(|X_t| > \lambda) < \delta \quad \forall t \in \mathcal{T},$$

which further implies by (4.3) and (4.4) that

$$\mathbb{E}\big[|X_t|\mathbf{1}_{\{|X_t|\}>\lambda}\big] < \varepsilon \quad \forall t \in \mathcal{T}.$$

This gives the uniform integrability of $\{X_t\}$. □

In the sequential context, uniform integrability plays an essential role in the connection between convergence in probability and convergence in L^1.

Theorem 4.2. *Let X_n, X $(n \geqslant 1)$ be integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The following two statements are equivalent.*

(i) X_n converges to X in L^1.
(ii) X_n convergence to X in probability and the family $\{X_n : n \geqslant 1\}$ is uniformly integrable.

Proof. Necessity. Suppose that X_n converges to X in L^1. It is immediate from Markov's inequality that X_n converges to X in probability. We now use Theorem 4.1 to prove uniform integrability. Boundedness in L^1 is obvious. To prove the uniform continuity property (4.2), one first notes that

$$\mathbb{E}\big[|X_n|\mathbf{1}_F\big] \leqslant \mathbb{E}\big[|X_n - X|\mathbf{1}_F\big] + \mathbb{E}\big[|X|\mathbf{1}_F\big] \leqslant \mathbb{E}\big[|X_n - X|\big] + \mathbb{E}\big[|X|\mathbf{1}_F\big]$$

for all $F \in \mathcal{F}$. By the L^1-convergence and integrability of X, for any $\varepsilon > 0$ there exist $N \geqslant 1$ and $\delta > 0$, such that

$$n > N,\ \mathbb{P}(F) < \delta \implies \mathbb{E}\big[|X_n - X|\big] < \varepsilon,\ \mathbb{E}\big[|X|\mathbf{1}_F\big] < \varepsilon.$$

As a result, one has

$$\sup_{n>N} \mathbb{E}\big[|X_n|\mathbf{1}_F\big] \leqslant 2\varepsilon \tag{4.5}$$

for any $F \in \mathcal{F}$ with $\mathbb{P}(F) < \delta$. By further reducing δ if necessary, one can include the first N terms into the relation (4.5). This proves the uniform integrability.

Sufficiency. Suppose that X_n converges to X in probability and $\{X_n\}$ is uniformly integrable. For any $\varepsilon > 0$ and $n \geqslant 1$, one can write

$$\begin{aligned}\mathbb{E}[|X_n - X|] &\leqslant \mathbb{E}\big[|X_n - X|\mathbf{1}_{\{|X_n - X| \leqslant \varepsilon\}}\big] + \mathbb{E}\big[|X_n - X|\mathbf{1}_{\{|X_n - X| > \varepsilon\}}\big] \\ &\leqslant \varepsilon + \mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n - X| > \varepsilon\}}\big] + \mathbb{E}\big[|X|\mathbf{1}_{\{|X_n - X| > \varepsilon\}}\big]. \end{aligned} \tag{4.6}$$

According to the property (4.2) and integrability of X, there exists $\delta > 0$ such that

$$F \in \mathcal{F},\ \mathbb{P}(F) < \delta \implies \sup_{n \geqslant 1} \mathbb{E}\big[|X_n|\mathbf{1}_F\big] < \varepsilon,\ \mathbb{E}\big[|X|\mathbf{1}_F\big] < \varepsilon.$$

In addition, by the convergence in probability there exists $N \geqslant 1$ such that

$$n > N \implies \mathbb{P}\big(|X_n - X| > \varepsilon\big) < \delta.$$

It follows from (4.6) that for all $n > N$, one has

$$\mathbb{E}\big[|X_n - X|\big] < 3\varepsilon.$$

This proves L^1-convergence. □

4.3 Weak convergence of probability measures

Weak convergence, being the weakest type of convergence among the four, is essentially a distributional property (it is concerned with probability laws of the random variables). It does not reflect the correlations among the random variables and it does not rely on the probability space where the random variables are defined. As a consequence, it applies to a broader range of problems and has larger flexibility to support finer quantitative estimates. In this section, we develop the basic tools for the study of weak convergence in reasonable generality.

4.3.1 Recapturing weak convergence for distribution functions

Let F_n, F be given distribution functions on $\mathbb{R}$. Recall from Definition 4.4 that F_n converges weakly to F if $F_n(x) \to F(x)$ at every continuity point of F (the definition has nothing to do with the actual random variables; it is essentially about convergence of the distribution functions). The reason why one cannot replace the definition with the "seemingly more natural" condition

$$\text{“}F_n(x) \to F(x) \quad \text{for every } x \in \mathbb{R}\text{”}$$

is best illustrated by the following simple example. Let $X_n = 1/n$ be the deterministic random variable taking value $1/n$. Obviously, any useful and reasonable notion of convergence should ensure that X_n "converges" to the zero random variable $X = 0$ as $n \to \infty$. On the other hand, the distribution functions of X_n and X are given by

$$F_n(x) = \begin{cases} 0, & x < 1/n; \\ 1, & x \geqslant 1/n, \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0; \\ 1, & x \geqslant 0, \end{cases}$$

respectively. It is apparent that

$$F_n(0) = 0 \nrightarrow F(0) = 1$$

as $n \to \infty$. This simple example shows that it is generally too restrictive to require $F_n(x)$ converging to $F(x)$ *for all* $x \in \mathbb{R}$. In this example, the issue occurs precisely at $x = 0$, which is a discontinuity point of F. It is easily seen that at every continuity point of F (i.e. whenever $x \neq 0$) one has $F_n(x) \to F(x)$. In other words, X_n converges weakly to X in the sense of Definition 4.4.

Example 4.3. Let X_n be a discrete uniform random variable over $\{1, 2, \cdots, n\}$, i.e.

$$\mathbb{P}(X_n = k) = \frac{1}{n}, \quad k = 1, 2, \cdots n.$$

Let X be a continuous uniform random variable over $[0,1]$. Then $X_n/n \to X$ weakly as $n \to \infty$. Indeed, the distribution function of X_n is given by

$$F_n(x) = \mathbb{P}\big(\frac{X_n}{n} \leqslant x\big) = \mathbb{P}(X_n \leqslant nx) = \begin{cases} 0, & x < 0; \\ \frac{[nx]}{n}, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1, \end{cases}$$

where $[nx]$ denotes the integer part of nx. From the simple inequality

$$\frac{[nx]}{n} \leqslant \frac{nx}{n} = x \leqslant \frac{[nx]}{n} + \frac{1}{n},$$

we know that $\frac{[nx]}{n} \to x$ as $n \to \infty$. It follows that

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0, & x < 0; \\ x, & 0 \leqslant x < 1; \\ 1, & x \geqslant 1, \end{cases}$$

which is precisely the distribution function $F(x)$ of X. Therefore, by definition one concludes that $X_n/n \to X$ weakly. Note that $F(x)$ is continuous at every $x \in \mathbb{R}$.

Example 4.4. Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables with finite mean and variance. Define $S_n \triangleq X_1 + \cdots + X_n$. Then the sample average S_n/n converges to $\mathbb{E}[X_1]$ a.s. In addition, the normalised fluctuation $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ converges weakly to the standard normal distribution. These are the contents of the strong law of large numbers and the central limit theorem. We will make precise the definition of independence and prove these facts later on.

When the limiting random variable X is continuous (i.e. the distribution function of X being continuous), weak convergence does become pointwise convergence at every $x \in \mathbb{R}$. A surprising fact is that one can obtain the stronger property of *uniform convergence* in this context. Such a result was due to G. Pólya.

Theorem 4.3. *Let X_n, X be random variables with distribution functions F_n, F respectively. Suppose that F is continuous on $\mathbb{R}$. Then X_n converges weakly to X if and only if F_n converges to F uniformly on $\mathbb{R}$.*

Proof. We only need to prove necessity as the other direction is trivial. Suppose that F_n converges to F at every $x \in \mathbb{R}$. Let $k \geqslant 1$ be an arbitrary given integer. We choose a partition

$$-\infty = x_0 < x_1 < x_2 < \cdots < x_{k-1} < x_k = \infty$$

such that

$$F(x_i) = \frac{i}{k}, \quad i = 0, 1, \cdots, k.$$

This is possible since F is continuous on $\mathbb{R}$. For any $x \in [x_{i-1}, x_i)$, according to the monotonicity of distribution functions, one has

$$F_n(x) - F(x) \leqslant F_n(x_i) - F(x_{i-1}) = F_n(x_i) - F(x_i) + \frac{1}{k}.$$

Similarly, one also has

$$F_n(x) - F(x) \geqslant F_n(x_{i-1}) - F(x_i) = F_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k}.$$

Combining the two inequalities, one obtains that

$$|F_n(x) - F(x)| \leqslant \max\left\{\left|F_n(x_{i-1}) - F(x_{i-1})\right|, \left|F_n(x_i) - F(x_i)\right|\right\} + \frac{1}{k},$$

for all i and $x \in [x_{i-1}, x_i)$. As a result,

$$\sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right| \leqslant \max_{0\leqslant i\leqslant k} \left|F_n(x_i) - F(x_i)\right| + \frac{1}{k}$$

Since $F_n(x_i) \to F(x_i)$ at each x_i, by letting $n \to \infty$ on both sides one finds that

$$\overline{\lim_{n\to\infty}} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leqslant \frac{1}{k}.$$

Since k is arbitrary, it follows that

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0.$$

□

In many situations, working with distribution functions may not be as convenient as working with probability measures (this is particularly the case in higher dimensions). Recall from Corollary 1.1 that distribution functions on $\mathbb{R}$ and probability measures on $\mathcal{B}(\mathbb{R})$ are essentially the same thing. As a result, it is reasonable to expect that Definition 4.4 has a counterpart for probability measures on $\mathcal{B}(\mathbb{R})$. We first give the following definition.

Definition 4.6. Let μ be a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A real number $a \in \mathbb{R}$ is called a *continuity point* of μ if $\mu(\{a\}) = 0$. The set of continuity points of μ is denoted as $\mathcal{C}(\mu)$.

Remark 4.1. The complement of $\mathcal{C}(\mu)$ is at most countable. To see this, one first observes that for each fixed $\varepsilon > 0$, the set $E_\varepsilon \triangleq \{a \in \mathbb{R} : \mu(\{a\}) \geqslant \varepsilon\}$ is at most finite. For otherwise, say E_ε contains an infinite sequence $\{a_1, a_2, \cdots\}$, one then has

$$\mu(\{a_1, a_2, \cdots\}) = \sum_{i=1}^{\infty} \mu(\{a_i\}) \geqslant \sum_{i=1}^{\infty} \varepsilon = \infty,$$

contradicting the finiteness of μ. It follows that

$$\mathcal{C}(\mu)^c = \big\{a \in \mathbb{R} : \mu(\{a\}) > 0\big\} = \bigcup_{n=1}^{\infty} \big\{a \in \mathbb{R} : \mu(\{a\}) \geqslant \frac{1}{n}\big\}$$

is at most countable. As a consequence, one also knows that $\mathcal{C}(\mu)$ is dense in $\mathbb{R}$.

The following result provides the counterpart of Definition 4.4 in the context of probability measures.

Proposition 4.5. *Let F_n, F be distribution functions and let μ_n, μ be the induced probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then F_n converges weakly to F if and only if*

$$\mu_n((a, b]) \to \mu((a, b])$$

for all continuity points $a < b$ of μ.

The necessity part of Proposition 4.5 is trivial. Indeed, suppose that F_n converges weakly to F. Note that a is a continuity point of μ if and only if it is a continuity point of F. Therefore, for any continuity points $a < b$ of μ, one has

$$\mu_n((a, b]) = F_n(b) - F_n(a) \to F(b) - F(a) = \mu((a, b]).$$

The sufficiency part is not as obvious. The crucial point is a *tightness* property for the sequence $\{\mu_n\}$, which is in turn based on the fact that μ_n and μ are *probability* measures. We take this result as granted in order to motivate the next definitions. We will come back to its proof when we are acquainted with more tools on weak convergence.

4.3.2 Vague convergence and Helly's theorem

Inspired by Proposition 4.5, we introduce the following definition.

Definition 4.7. Let μ_n $(n \geqslant 1)$ and μ be finite measures on $\mathcal{B}(\mathbb{R})$. We say that μ_n *converges vaguely* to μ, if $\mu_n((a,b]) \to \mu((a,b])$ for all continuity points $a < b$ of μ. If additionally $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$, we say that μ_n *converges weakly* to μ.

When μ_n and μ are probability measures, vague and weak convergence are the same thing since $\mu_n(\mathbb{R}) = \mu(\mathbb{R}) = 1$. In general, these two notions of convergence are different as seen from the following example.

Example 4.5. Consider $\mu_n = \delta_n$ (the Dirac mass at the point $x = n$) and $\mu = 0$ (the zero measure). Every real number is a continuity point of μ. For any fixed $a < b$, when n is large (precisely when $n > b$) one has $\mu_n((a,b]) = 0$. In particular, μ_n converges vaguely to μ. But

$$\mu_n(\mathbb{R}) = 1 \nrightarrow 0 = \mu(\mathbb{R}).$$

In other words, μ_n does not converge weakly to μ.

Intervals are too special for many purposes. One needs to identify more robust characterisations of vague and weak convergence in order to generalise these concepts to higher (random vectors) and even infinite dimensions (stochastic processes). The next two results provide particularly useful characterisations in terms of integration against suitable test functions.

Recall that a continuous function on $\mathbb{R}^d$ is said to *have compact support* if it vanishes identically outside some bounded subset of $\mathbb{R}^d$. The space of continuous functions on $\mathbb{R}^d$ with compact support is denoted as $\mathcal{C}_c(\mathbb{R}^d)$. Respectively, the space of bounded continuous functions on $\mathbb{R}^d$ is denoted as $\mathcal{C}_b(\mathbb{R}^d)$. Apparently, $\mathcal{C}_c(\mathbb{R}^d) \subseteq \mathcal{C}_b(\mathbb{R}^d)$.

First of all, one has the following characterisation of vague convergence.

Theorem 4.4. *Let μ_n $(n \geqslant 1)$ and μ be finite measures on $\mathbb{R}$. Then μ_n converges vaguely to μ if and only if*

$$\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx) \quad \textit{for all } f \in \mathcal{C}_c(\mathbb{R}).$$

Proof. Necessity. Let $f \in \mathcal{C}_c(\mathbb{R})$. Firstly, we choose $a < b$ in $\mathcal{C}(\mu)$ (continuity points of μ) such that $f(x) = 0$ outside $[a,b]$. Since f is continuous, it is uniformly continuous on $[a,b]$. In particular, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x, y \in [a,b], |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

For such δ, we choose a partition

$$a = x_0 < x_1 < \cdots < x_{k-1} < x_k = b$$

such that $x_i \in \mathcal{C}(\mu)$ and $|x_i - x_{i-1}| < \delta$. This is possible since $\mathcal{C}(\mu)$ is dense in $\mathbb{R}$ (cf. Remark 4.1). If we define the step function

$$g(x) \triangleq \sum_{i=1}^{k} f(x_{i-1}) \mathbf{1}_{(x_{i-1}, x_i]}(x),$$

then $f(x) = g(x) = 0$ when $x \notin (a, b]$ and $|f(x) - g(x)| < \varepsilon$ when $x \in [a, b]$. It follows that

$$\begin{aligned}
&\Big| \int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx) \Big| \\
&\leqslant \Big| \int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} g(x)\mu_n(dx) \Big| + \Big| \int_{\mathbb{R}} g(x)\mu_n(dx) - \int_{\mathbb{R}} g(x)\mu(dx) \Big| \\
&\quad + \Big| \int_{\mathbb{R}} g(x)\mu(dx) - \int_{\mathbb{R}} f(x)\mu(dx) \Big| \\
&\leqslant \varepsilon \cdot \mu_n((a, b]) + \sum_{i=1}^{k} |f(x_{i-1})| \cdot |\mu_n((x_{i-1}, x_i]) - \mu((x_{i-1}, x_i])| + \varepsilon \cdot \mu((a, b]).
\end{aligned} \tag{4.7}$$

Since $a, b, x_i \in \mathcal{C}(\mu)$, one has

$$\mu_n((a, b]) \to \mu((a, b]), \ \ \mu_n((x_{i-1}, x_i]) \to \mu((x_{i-1}, x_i])$$

as $n \to \infty$. As a consequence, by taking $n \to \infty$ in (4.7) one arrives at

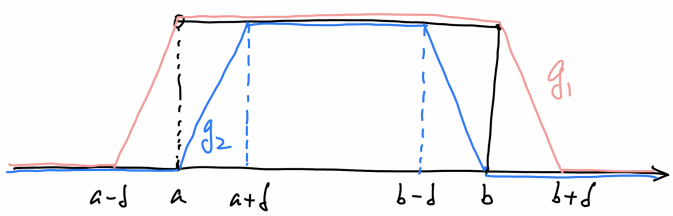
$$\overline{\lim_{n\to\infty}} \Big| \int_{\mathbb{R}} f(x)\mu_n(dx) - \int_{\mathbb{R}} f(x)\mu(dx) \Big| \leqslant 2\varepsilon \cdot \mu((a, b]).$$

Since ε is arbitrary, it follows that

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx).$$

Sufficiency. Let $a < b$ be two continuity points of μ and define $g(x) \triangleq \mathbf{1}_{(a,b]}(x)$. Given fixed $\delta > 0$, we are going to define two "tent-shaped" functions $g_1, g_2 \in \mathcal{C}_c(\mathbb{R})$ that approximate g from above and below respectively. Precisely, $g_1(x) \triangleq 1$ when

$x \in [a,b]$, $g_1(x) \triangleq 0$ when $x \notin [a-\delta, b+\delta]$, and $g_1(x)$ is linear when $x \in [a-\delta, a]$ and $x \in [b, b+\delta]$. Similarly, $g_2(x) \triangleq 1$ when $x \in [a+\delta, b-\delta]$, $g_1(x) \triangleq 0$ when $x \notin [a,b]$, and $g_2(x)$ is linear when $x \in [a, a+\delta]$ and $x \in [b-\delta, b]$.



By the construction, it is not hard to see that

$$g_2(x) \leqslant g(x) \leqslant g_1(x) \quad \forall x \in \mathbb{R}, \tag{4.8}$$

and

$$g_1 = g_2 \text{ on } U^c, \ 0 \leqslant g_1 - g_2 \leqslant 1 \text{ on } U \tag{4.9}$$

with $U \triangleq (a-\delta, a+\delta) \cup (b-\delta, b+\delta)$.

By integrating (4.8) against μ_n and μ respectively, one obtains that

$$\int g_2 d\mu_n \leqslant \int g d\mu_n = \mu_n((a,b]) \leqslant \int g_1 d\mu_n, \ \int g_2 d\mu \leqslant \mu((a,b]) \leqslant \int g_1 d\mu.$$

Therefore,

$$\int g_2 d\mu_n - \int g_1 d\mu \leqslant \mu_n((a,b]) - \mu((a,b]) \leqslant \int g_1 d\mu_n - \int g_2 d\mu. \tag{4.10}$$

By taking $n \to \infty$ in the first inequality and using (4.9), one finds that

$$\begin{aligned} \varliminf_{n\to\infty} \big(\mu_n((a,b]) - \mu((a,b])\big) &\geqslant \int g_2 d\mu - \int g_1 d\mu \geqslant -\mu(U) \\ &= -\big(\mu((a-\delta, a+\delta)) + \mu((b-\delta, b+\delta))\big). \end{aligned}$$

Since a, b are continuity points of μ and δ is arbitrary, by letting $\delta \to 0$ the last term goes to zero and thus

$$\varliminf_{n\to\infty} \big(\mu_n((a,b]) - \mu((a,b])\big) \geqslant 0.$$

Exactly the same argument applied to the second inequality in (4.10) yields

$$\overline{\lim_{n\to\infty}}\big(\mu_n((a,b])-\mu((a,b])\big)\leqslant 0.$$

Therefore, one arrives at

$$\lim_{n\to\infty}\mu_n((a,b])=\mu((a,b]).$$

□

Respectively, one has the following characterisation of weak convergence.

Theorem 4.5. *Let μ_n $(n\geqslant 1)$ and μ be finite measures on $\mathbb{R}$. Then μ_n converges weakly to μ if and only if*

$$\int_{\mathbb{R}} f(x)\mu_n(dx)\to\int_{\mathbb{R}} f(x)\mu(dx)\quad \textit{for all } f\in\mathcal{C}_b(\mathbb{R}). \tag{4.11}$$

Proof. Sufficiency. The condition already implies vague convergence as a consequence of Theorem 4.4 since $\mathcal{C}_c(\mathbb{R})\subseteq\mathcal{C}_b(\mathbb{R})$. In addition, by taking $f=1$ one also has $\mu_n(\mathbb{R})\to\mu(\mathbb{R})$. Therefore, μ_n converges weakly to μ.

Necessity. Let $f\in\mathcal{C}_b(\mathbb{R})$ and suppose that $|f(x)|\leqslant M$ for all x. Given $\varepsilon>0$, we pick two continuity points $a<b$ of μ so that $\mu((a,b]^c)<\varepsilon$. By the weak convergence assumption, one has

$$\mu_n((a,b]^c)=\mu_n(\mathbb{R})-\mu_n((a,b])\to\mu(\mathbb{R})-\mu((a,b])=\mu((a,b]^c).$$

In particular, $\mu_n((a,b]^c)<\varepsilon$ when n is large. It follows that

$$\begin{aligned}
&\Big|\int_{\mathbb{R}} f(x)\mu_n(dx)-\int_{\mathbb{R}} f(x)\mu(dx)\Big|\\
&\leqslant\Big|\int_{(a,b]} f(x)\mu_n(dx)-\int_{(a,b]} f(x)\mu(dx)\Big|+\Big|\int_{(a,b]^c} f(x)\mu_n(dx)-\int_{(a,b]^c} f(x)\mu(dx)\Big|\\
&\leqslant\Big|\int_{(a,b]} f(x)\mu_n(dx)-\int_{(a,b]} f(x)\mu(dx)\Big|+2M\varepsilon.
\end{aligned}\tag{4.12}$$

By using the same approximation argument as in the necessity part of Theorem 4.4, one can show that the first term on the right hand side of (4.12) vanishes as $n\to\infty$. As a result,

$$\overline{\lim_{n\to\infty}}\Big|\int_{\mathbb{R}} f(x)\mu_n(dx)-\int_{\mathbb{R}} f(x)\mu(dx)\Big|\leqslant 2M\varepsilon,$$

which further implies

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx)$$

since ε is arbitrary. □

The characterisations given by Theorem 4.4 and Theorem 4.5 allow one to generalise the concepts of vague and weak convergence to higher dimensions naturally.

Definition 4.8. Let μ_n $(n \geqslant 1)$ and μ be finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

(i) We say that μ_n *converges vaguely to* μ if

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \to \int_{\mathbb{R}^d} f(x)\mu(dx)$$

for every $f \in \mathcal{C}_c(\mathbb{R}^d)$.

(ii) We say that μ_n *converges weakly to* μ if

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \to \int_{\mathbb{R}^d} f(x)\mu(dx)$$

for every $f \in \mathcal{C}_b(\mathbb{R}^d)$.

It can be shown that weak convergence is equivalent to vague convergence plus the additional property that $\mu_n(\mathbb{R}^d) \to \mu(\mathbb{R}^d)$. In particular, when μ_n, μ are probability measures, the two notions of convergence are again the same thing. In the context of $\mathbb{R}^d$-valued random variables X_n and X, we say that X_n *converges weakly to* X if the law of X_n converges weakly to the law of X. According to Theorem 2.5, this is also equivalent to saying that

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \text{for all } f \in \mathcal{C}_b(\mathbb{R}^d).$$

Recall from real analysis that a bounded sequence in $\mathbb{R}^d$ admits a convergent subsequence. The extension of this result to probability measures is the content of *Helly's theorem*. This theorem is important because it is often the first step towards proving weak convergence of probability measures. Before stating the theorem, we first introduce the following definition.

Definition 4.9. A *sub-probability measure* μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a finite measure such that $\mu(\mathbb{R}^d) \leqslant 1$.

Helly's theorem for vague convergence of sub-probability measures is stated as follows.

Theorem 4.6. *Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then there exists a subsequence μ_{n_k} and a sub-probability measure μ, such that μ_{n_k} converges vaguely to μ as $k \to \infty$.*

Proof. We only prove the result in one dimension and let the reader to adapt the argument to higher dimensions. We break down the proof into several steps.

Step One. Consider the corresponding sequence of distribution functions $F_n(x) \triangleq \mu_n((-\infty, x])$. Let $D = \{x_j : j \geqslant 1\}$ be a countable dense subset of $\mathbb{R}$ (e.g. the rational numbers). We claim that there exists a subsequence $\{F_{n_k}\}$ of $\{F_n\}$, such that $\lim_{k\to\infty} F_{n_k}(x_j)$ exists for every $x_j \in D$. To prove this, let us start with the sequence $\{F_n(x_1)\}$ of real numbers. Since this is a bounded sequence, there exists a subsequence $\{n_1(k) : k \geqslant 1\}$ of $\mathbb{N}$ and some real number denoted as $G(x_1)$, such that $F_{n_1(k)}(x_1) \to G(x_1)$. Next, for the bounded sequence $\{F_{n_1(k)}(x_2) : k \geqslant 1\}$, there exists a further subsequence $\{n_2(k)\}$ of $\{n_1(k)\}$ and some real number denoted as $G(x_2)$, such that $F_{n_2(k)}(x_2) \to G(x_2)$. If one continues this procedure, at the j-th step one finds a subsequence $\{n_j(k)\}$ of the previous sequence $\{n_{j-1}(k)\}$ as well as $G(x_j) \in \mathbb{R}$, such that $F_{n_j(k)}(x_j) \to G(x_j)$. Now we consider the "diagonal" subsequence $\{n_k(k) : k \geqslant 1\}$. For each fixed j, from the construction $\{n_k(k) : k \geqslant j\}$ is a subsequence of $\{n_j(k) : k \geqslant 1\}$. In particular, one has

$$\lim_{k\to\infty} F_{n_k(k)}(x_j) = G(x_j),$$

which proves the desired claim.

Step Two. Using the previous limit points $\{G(x_j) : j \geqslant 1\}$, we define the function

$$F(x) \triangleq \inf\{G(x_j) : x_j > x\}.$$

It is obvious that $0 \leqslant F(x) \leqslant 1$ and $F(x)$ is increasing. Moreover, $F(x)$ is right continuous. Indeed, let $x \in \mathbb{R}$ and $\varepsilon > 0$. By the definition of F, there exists $x_j > x$ such that $G(x_j) < F(x) + \varepsilon$. It follows that whenever $0 < h < x_j - x$ one has $x + h < x_j$ and thus

$$F(x + h) \leqslant G(x_j) < F(x) + \varepsilon.$$

This shows that F is right continuous at x.

Step Three. At every continuity point x of F, one has $F_{n_k(k)}(x) \to F(x)$. For simplicity we write $n_k \triangleq n_k(k)$. Given $\varepsilon > 0$, there exists $x_p > x$ such that

$$G(x_p) < F(x) + \varepsilon. \tag{4.13}$$

In addition, since x is a continuity point of F, there exists $y < x$ such that $F(x) - F(y) < \varepsilon$. Pick any $x_q \in D \cap (y, x)$. It follows that $F(y) \leqslant G(x_q)$ and thus

$$F(x) - G(x_q) \leqslant F(x) - F(y) < \varepsilon. \tag{4.14}$$

Adding (4.13) and (4.14) gives $G(x_p) - G(x_q) < 2\varepsilon$. As a consequence, one finds that

$$\begin{aligned} |F_{n_k}(x) - F(x)| &\leqslant |F_{n_k}(x) - F_{n_k}(x_p)| + |F_{n_k}(x_p) - G(x_p)| + |G(x_p) - F(x)| \\ &\leqslant \big(F_{n_k}(x_p) - F_{n_k}(x_q)\big) + |F_{n_k}(x_p) - G(x_p)| + \varepsilon. \end{aligned}$$

By taking $k \to \infty$, one obtains that

$$\overline{\lim_{k\to\infty}} \big|F_{n_k}(x) - F(x)\big| \leqslant G(x_p) - G(x_q) + \varepsilon \leqslant 3\varepsilon.$$

Since ε is arbitrary, it follows that $F_{n_k}(x) \to F(x)$.

Step Four. According to Corollary 1.1, the function F induces a unique sub-probability μ on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$ such that $\mu((a, b]) = F(b) - F(a)$ for any $a < b$. Step three shows that μ_{n_k} converges vaguely to μ. This completes the proof of the theorem. □

It is important to point out that one cannot strengthen the conclusion of Helly's theorem to weak convergence, as the limit point μ may fail to be a probability measure.

Example 4.6. Let μ_n be the uniform distribution over $[-n, n]$. Then μ_n converges vaguely to the zero measure (and so does any of its subsequence). Indeed, for any fixed $a < b$, when n is large one has

$$\mu_n((a, b]) = \frac{b - a}{2n},$$

which converges to zero as $n \to \infty$.

The next natural question is to investigate when a vague limit point is indeed a probability measure. The answer to this question is intimately related to the concept of *tightness* which will be discussed in Section 4.3.4 below. Here we look at a simple motivating example.

Example 4.7. Let $M > 0$ be a fixed number. Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mu_n([-M, M]) = 1$ for each n. Then

every vague convergent subsequence of μ_n must converge weakly to a probability measure. Indeed, suppose that μ_{n_k} converges vaguely to some sub-probability measure μ. Pick two continuity points a, b of μ such that $a < -M$ and $b > M$. Then

$$1 = \mu_n((a, b]) \to \mu((a, b]),$$

showing that μ has to be a probability measure and thus μ_{n_k} converges weakly to μ. The key point in this example is that masses for the sequence $\{\mu_n\}$ are uniformly concentrated on a large interval. This is essentially the tightness property which is of fundamental importance in the study of weak convergence.

4.3.3 Weak convergence on metric spaces and the Portmanteau theorem

Working with probability measures over $\mathbb{R}^d$ (i.e. in finite dimensions) is unfortunately not always sufficient. For instance, when one studies distributions of stochastic processes, one is led to the consideration of probability measures over infinite dimensional spaces (space of paths). It is essential to extend the notion of weak convergence to the more general context of metric spaces.

Metric spaces

We first recall the relevant concepts. Heuristically, a metric space is a set equipped with a distance function.

Definition 4.10. Let S be a non-empty set. A *metric* on S is a non-negative function $\rho : S \times S \to [0, \infty)$ which satisfies the following three properties.

(i) Positive definiteness: $\rho(x, y) = 0$ if and only if $x = y$;
(ii) Symmetry: $\rho(x, y) = \rho(y, x)$;
(iii) Triangle inequality: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$.

When a set S is equipped with a metric ρ, we call (S, ρ) a *metric space*.

Example 4.8. An obvious metric on $\mathbb{R}^d$ is the Euclidean metric:

$$\rho(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}.$$

There are other choices of metrics, such as

$$\rho'(x, y) = |x_1 - y_1| + \cdots + |x_d - y_d| \quad (\text{the } l^1 \text{ metric})$$

and

$$\rho''(x, y) = \max_{1 \leqslant i \leqslant d} |x_i - y_i| \quad (\text{the } l^\infty \text{ metric}).$$

Example 4.9. An important infinite dimensional example of a metric space is the space of paths. More precisely, let $W = C[0,1]$ be the set of all continuous functions $w : [0,1] \to \mathbb{R}$. Define a function $\rho : W \times W \to [0,\infty)$ by

$$\rho(w_1, w_2) \triangleq \sup_{0 \leqslant t \leqslant 1} |w_1(t) - w_2(t)|, \quad w_1, w_2 \in W.$$

It is a simple exercise to check that ρ is a metric on W (it is called the *uniform metric*). One will frequently encounter this metric space (W, ρ) in the study of continuous-time stochastic processes such as Brownian motion, Gaussian processes etc.

Let (S, ρ) be a given metric space. We introduce a few basic set classes over S. Unlike the usual $\mathbb{R}^d$, there are no analogues of intervals on S. However, one has the natural notion of *open balls*

$$B(x, r) \triangleq \{y \in S : d(y, x) < r\}$$

and similarly of closed balls.

Definition 4.11. Given a subset $A \subseteq S$, a point $x \in A$ is called an *interior point* of A if there exists $r > 0$ such that $B(x, r) \subseteq A$. A subset $G \subseteq S$ is said to be *open* if every point in G is an interior point. A subset $F \subseteq S$ is said to be *closed* if its complement F^c is open. A subset $K \subseteq S$ is said to be *compact* if any open cover of K contains a finite subcover, namely whenever K is contained in the union of a family of open sets, one can always choose finitely many members in that family whose union still contains K.

These concepts are better illustrated along with the notion of convergence.

Definition 4.12. Let x_n $(n \geqslant 1)$ and x be points in S. We say that x_n *converges to* x, denoted as $x_n \to x$, if $\rho(x_n, x) \to 0$ as $n \to \infty$.

It can be shown that a subset F is closed if and only if

$$x_n \in F, x_n \to x \implies x \in F.$$

In addition, a subset K is compact if and only if it is closed and any sequence in K admits a convergent subsequence.

Definition 4.13. Let A be a subset of S. The *closure* of A, denoted as $\bar{A}$, is the smallest closed subset containing A. Equivalently, $\bar{A}$ consists of all limit points of A. The *interior* of A, denoted as $\mathring{A}$, is the largest open subset contained in A. Equivalently, $\mathring{A}$ is the set of interior points of A. The *boundary* of A is defined to be $\partial A \triangleq \bar{A} \backslash \mathring{A}$.

Continuous functions and uniformly continuous functions are defined in the usual way.

Definition 4.14. A function $f : S \to \mathbb{R}$ is *continuous* at x, if

$$x_n \in S,\ x_n \to x \implies f(x_n) \to f(x).$$

A *continuous function* on S is a function that is continuous at every point in S. A function is *uniformly continuous*, if for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x, y \in S,\ d(x, y) < \delta \implies |f(y) - f(x)| < \varepsilon.$$

If $f : S \to \mathbb{R}$ is continuous, then

$$\begin{aligned} U \subseteq \mathbb{R},\ U \text{ is open} &\implies f^{-1}U \text{ is open in } S; \\ C \subseteq \mathbb{R},\ C \text{ is closed} &\implies f^{-1}C \text{ is closed in } S. \end{aligned}$$

The space of bounded, continuous functions on S is denoted as $\mathcal{C}_b(S)$.

Remark 4.2. All the above concepts are natural generalisations of the $\mathbb{R}^d$ case. They are best visualised when one looks at the case of $\mathbb{R}^d$. A major difference from the $\mathbb{R}^d$ case is notion of compact sets. In $\mathbb{R}^d$, one knows that a subset is compact if and only if it is bounded and closed. This is *not* true for general metric spaces (compactness is much harder to characterise in infinite dimensions). In the appendix, we give the description of compact sets in the path space $C[0, 1]$ of Example 4.9.

Weak convergence and the Portmanteau theorem

Before studying probability measures and weak convergence on a metric space, one needs to introduce a natural σ-algebra over it.

Definition 4.15. The *Borel σ-algebra* over S, denoted as $\mathcal{B}(S)$, is σ-algebra generated by all open subsets of S.

Example 4.10. Open balls, closed balls, open sets, closed sets, compact sets and countable unions / intersections of these sets are members of $\mathcal{B}(S)$.

Remark 4.3. In the case of $\mathbb{R}$, the Borel σ-algebra is generated by the class of open intervals (a, b). For general metric spaces, the Borel σ-algebra may not necessarily be generated by open balls. Nonetheless, this will be the case if the metric space (S, ρ) is *separable*, namely if there exists a countable subset $D \subseteq S$ such that $\bar{D} = S$.

We will always work with the Borel σ-algebra $\mathcal{B}(S)$ and all probability measures are assumed to be defined on $\mathcal{B}(S)$. A natural way of generalising weak convergence to metric spaces is through the characterisation given by Theorem 4.5 in the $\mathbb{R}^d$ case.

Definition 4.16. Let μ_n $(n \geqslant 1)$ and μ be probability measures on a metric space $(S, \mathcal{B}(S), \rho)$. We say that μ_n *converges weakly to* μ, if

$$\int_S f(x)\mu_n(dx) \to \int_S f(x)\mu(dx)$$

for all $f \in \mathcal{C}_b(S)$.

One may wonder if weak convergence can be seen through testing against "sets". As in the case of $\mathbb{R}$, one cannot expect that $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}(S)$ and some sort of continuity for the set A with respect to μ is needed.

Definition 4.17. Let μ be a probability measure on $(S, \mathcal{B}(S), \rho)$. A subset $A \in \mathcal{B}(S)$ is said to be μ-*continuous*, if $\mu(\partial A) = 0$.

Example 4.11. In the case of $S = \mathbb{R}$, an interval $(a, b]$ is μ-continuous if and only if a, b are both continuity points of μ (cf. Definition 4.6).

Example 4.12. Let $S = \{(x, y) : 0 \leqslant x, y \leqslant 1\}$ be the unit square in $\mathbb{R}^2$ and let μ be the Lebesgue measure on $(S, \mathcal{B}(S))$. Then any region in S enclosed by a smooth curve is μ-continuous, since its boundary curve has zero measure.

The following basic result, known as the *Portmanteau theorem*, provides a set of equivalent characterisations for weak convergence.

Theorem 4.7. *Let μ_n $(n \geqslant 1)$ and μ be probability measures on a metric space $(S, \mathcal{B}(S), \rho)$. The following statements are equivalent:*

(i) μ_n converges weakly to μ;
(ii) for any bounded, uniformly continuous function f on S, one has

$$\int_S f(x)\mu_n(dx) \to \int_S f(x)\mu(dx);$$

(iii) for any closed subset $F \subseteq S$, one has

$$\overline{\lim_{n\to\infty}} \mu_n(F) \leqslant \mu(F);$$

(iv) for any open subsets $G \subseteq S$, one has

$$\varliminf_{n\to\infty} \mu_n(G) \geqslant \mu(G);$$

(v) for any $A \in \mathcal{B}(S)$ that is μ-continuous, one has

$$\lim_{n\to\infty} \mu_n(A) = \mu(A).$$

Proof. (i) $\Longrightarrow$ (ii) is trivial.

(ii) $\Longrightarrow$ (iii). Let F be a closed subset of S. For $k \geqslant 1$, we define

$$f_k(x) = \left(\frac{1}{1+\rho(x,F)}\right)^k, \ x \in S,$$

where $\rho(x,F)$ is the distance between x and F. Then f_k is bounded and uniformly continuous. In addition,

$$\mathbf{1}_F(x) \leqslant f_k(x) \leqslant 1, \tag{4.15}$$

and $f_k(x) \downarrow \mathbf{1}_F(x)$ as $k \to \infty$. It follows from (4.15) and the assumption that

$$\varlimsup_{n\to\infty} \mu_n(F) \leqslant \lim_{n\to\infty} \int_S f_k(x)\mu_n(dx) = \int_S f_k(x)\mu(dx)$$

for every $k \geqslant 1$. By taking $k \to \infty$ and using the dominated convergence theorem, one finds that

$$\varlimsup_{n\to\infty} \mu_n(F) \leqslant \mu(F).$$

(iii)$\Longleftrightarrow$(iv) is obvious as they are complementary to each other.

(iii)+(iv) $\Longrightarrow$ (v). Let $A \in \mathcal{B}(S)$ be such that $\mu(\partial A) = 0$. Then

$$\mu(\mathring{A}) = \mu(A) = \mu(\bar{A}).$$

By the assumptions of (iii) and (iv), one has

$$\varlimsup_{n\to\infty} \mu_n(A) \leqslant \varlimsup_{n\to\infty} \mu_n(\bar{A}) \leqslant \mu(\bar{A}) = \mu(A) = \mu(\mathring{A}) \leqslant \varliminf_{n\to\infty} \mu_n(\mathring{A}) \leqslant \varliminf_{n\to\infty} \mu_n(A).$$

As a result, $\mu_n(A) \to \mu(A)$.

(v) $\Longrightarrow$ (i). Let $f \in \mathcal{C}_b(S)$ be given fixed. One may assume that $0 < f < 1$; for otherwise, say $a < f(x) < b$, one can consider the normalised function

$$0 < \bar{f}(x) \triangleq \frac{f(x)-a}{b-a} < 1.$$

The idea is to approximate f by linear combinations of indicator functions of μ-continuous sets. Since μ is a probability measure, for each $n \geqslant 1$ the set $\{a \in \mathbb{R}^1 : \mu(f = a) \geqslant 1/n\}$ must be finite, and thus the set $\{a \in \mathbb{R}^1 : \mu(f = a) > 0\}$ is at most countable. Given $k \geqslant 1$, for each $1 \leqslant i \leqslant k$ one can then choose some $a_i \in ((i-1)/k, i/k)$ such that $\mu(f = a_i) = 0$. Set $a_0 \triangleq 0$ and $a_{k+1} \triangleq 1$. Note that $|a_i - a_{i-1}| < 2/k$ for all i. Next, define the subsets

$$B_i \triangleq \{x \in S : a_{i-1} \leqslant f(x) < a_i\}, \quad 1 \leqslant i \leqslant k+1.$$

The B_i's are disjoint and $S = \cup_{i=1}^{k+1} B_i$ since $0 < f < 1$. In addition, it is seen from the continuity of f that

$$\overline{B_i} \subseteq \{a_{i-1} \leqslant f \leqslant a_i\}, \ \{a_{i-1} < f < a_i\} \subseteq \mathring{B}_i.$$

Therefore, one has

$$\partial B_i \subseteq \{f = a_{i-1}\} \cup \{f = a_i\},$$

showing that $\mu(\partial B_i) = 0$. We now consider the step function

$$g(x) \triangleq \sum_{i=1}^{k+1} a_{i-1} \mathbf{1}_{B_i}(x).$$

The function g approximates f in the sense that

$$|f(x) - g(x)| \leqslant \frac{2}{k} \quad \forall x \in S,$$

which is easily seen from the construction of the B_i's and g. It follows that

$$\begin{aligned}
&\left| \int_S f d\mu_n - \int_S f d\mu \right| \\
&\leqslant \int_S |f(x) - g(x)| d\mu_n + \int_S |f(x) - g(x)| d\mu + \Big| \int_S g d\mu_n - \int_S g d\mu \Big| \\
&\leqslant \frac{4}{k} + \sum_{i=1}^{k+1} a_{i-1} \cdot \big| \mu_n(B_i) - \mu(B_i) \big|.
\end{aligned}$$

Since $\mu(\partial B_i) = 0$, by taking $n \to \infty$ one has

$$\overline{\lim_{n \to \infty}} \left| \int_S f d\mu_n - \int_S f d\mu \right| \leqslant \frac{4}{k}.$$

Since k is arbitrary, one concludes that $\int_S f d\mu_n \to \int_S f d\mu$. $\square$

As an application of the Portmanteau theorem, we return to prove an earlier claim that weak convergence is the weakest among the four types of convergence we defined before.

Proposition 4.6. *Convergence in probability implies weak convergence.*

Proof. Suppose that X_n converges to X in probability. We use the second charaterisation in the Portmanteau theorem to show that X_n converges weakly to X. To this end, let f be a bounded and uniformly continuous function on $\mathbb{R}$. Given $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|x - y| \leqslant \delta \implies |f(x) - f(y)| \leqslant \varepsilon.$$

It follows that

$$\begin{aligned}
&\big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\big| \\
&\leqslant \mathbb{E}[|f(X_n) - f(X)|] \\
&= \mathbb{E}[|f(X_n) - f(X)|; |X_n - X| \leqslant \delta] + \mathbb{E}[|f(X_n) - f(X)|; |X_n - X| > \delta] \\
&\leqslant \varepsilon + 2\|f\|_\infty \mathbb{P}(|X_n - X| > \delta),
\end{aligned}$$

where $\|f\|_\infty \triangleq \sup_{x\in\mathbb{R}} |f(x)|$. Since $X_n \to X$ in probability, by letting $n \to \infty$ one finds that

$$\overline{\lim_{n\to\infty}} \big|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\big| \leqslant \varepsilon.$$

As ε is arbitrary, one concludes that $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$, yielding the desired weak convergence. □

4.3.4 Tightness and Prokhorov's theorem

Helly's theorem ensures the existence of a vaguely convergent subsequence for a given family of probability measures (though it is only true in finite dimensions!). To understand whether vague limit points are always probability measures, one is led to the concept of tightness.

Definition 4.18. A family $\{\mu : \mu \in \Lambda\}$ of probability measures on a metric space $(S, \mathcal{B}(S), \rho)$ is said to be *tight*, if for any $\varepsilon > 0$ there exists a compact subset $K \subseteq S$, such that

$$\mu(K) \geqslant 1 - \varepsilon \quad \text{for every } \mu \in \Lambda. \tag{4.16}$$

We say that a family of random variables is *tight* if the induced family of laws on $S = \mathbb{R}$ is tight.

Note that when $S = \mathbb{R}$, the condition (4.16) means that for any $\varepsilon > 0$ there exists $M > 0$, such that

$$\mu([-M, M]) \geqslant 1 - \varepsilon \quad \text{for every } \mu.$$

The following result, known as *Prokhorov's theorem*, is of fundamental importance in the study of weak convergence. We only prove the finite dimensional version, which gives a precise condition under which vague limit points are always probability measures, thus enhancing Helly's theorem to the level of weak convergence.

Theorem 4.8 (Prokhorov's theorem in $\mathbb{R}^d$)**.** *Let $\{\mu : \mu \in \Lambda\}$ be a family of probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The the following two statements are equivalent:*

(i) The family $\{\mu : \mu \in \Lambda\}$ is tight;
(ii) Every sequence in the family $\{\mu : \mu \in \Lambda\}$ admits a weakly convergent subsequence.

Proof. For simplicity we only consider the one dimensional case (i.e. $d = 1$).

(i) $\Longrightarrow$ (ii). Let $\mu_n \in \Lambda$ be a given sequence in the family. According to Helly's theorem (cf. Theorem 4.6), there exists a subsequence μ_{n_k} and a sub-probability measure μ, such that μ_{n_k} converges vaguely to μ. We need to show that μ is a probability measure. Since the family is tight by assumption, for given $m \geqslant 1$ there exists a closed interval K_m such that

$$\mu_{n_k}(K_m) \geqslant 1 - \frac{1}{m} \quad \text{for all } k.$$

One may assume that K_m is contained in $(a_m, b_m]$, where $a_m < b_m$ are continuity points of μ such that $a_m \downarrow -\infty$ and $b_m \uparrow \infty$ (as $m \to \infty$). It follows that

$$\mu_{n_k}((a_m, b_m]) \geqslant 1 - \frac{1}{m} \quad \text{for all } k.$$

Letting $k \to \infty$, one obtains that $\mu((a_m, b_m]) \geqslant 1 - 1/m$. By further sending $m \to \infty$ one concludes that $\mu(\mathbb{R}^1) \geqslant 1$. Therefore, μ must be a probability measure.

(ii) $\Longrightarrow$ (i). Suppose on the contrary that the family is not tight. Then there exists $\varepsilon > 0$, such that for each closed interval $[-n, n]$ one can find $\mu_n \in \Lambda$ with

$$\mu_n([-n, n]) < 1 - \varepsilon. \tag{4.17}$$

On the other hand, by the assumption of (ii) μ_n has a weakly convergent subsequence, say μ_{n_k} converging weakly to some probability measure μ. The property (4.17) implies that, for each fixed n, when k is large one has

$$\mu_{n_k}([-n,n]) \leqslant \mu_{n_k}([-n_k,n_k]) < 1-\varepsilon.$$

It follows from the Portmanteau theorem (cf. Theorem 4.7 (iv)) that

$$\mu((-n,n)) \leqslant \varliminf_{k\to\infty} \mu_{n_k}((-n,n)) \leqslant 1-\varepsilon$$

for every fixed n. Letting $n\to\infty$, one obtains that $\mu(\mathbb{R}) \leqslant 1-\varepsilon$ which contradicts the fact that μ is a probability measure. Therefore, the family $\{\mu : \mu\in\Lambda\}$ is tight. □

Example 4.13. Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables such that

$$L \triangleq \sup_n \mathbb{E}[|X_n|] < \infty.$$

Then this family is tight. Indeed, let μ_n be the law of X_n. Then for each $M>0$, one has

$$\mu_n([-M,M]^c) = \mathbb{P}(|X_n|>M) \leqslant \frac{\mathbb{E}[|X_n|]}{M} \leqslant \frac{L}{M}$$

for all $n\geqslant 1$. When M is large enough, the right hand side is arbitrarily small uniformly in n. This gives the tightness property.

We must point out the remarkable fact that Prokhorov's theorem holds in the general context of metric spaces. We only state the result as its proof is beyond the scope of the subject. A metric space (S,ρ) is said to be *complete*, if every Cauchy sequence is convergent. Examples 4.8 and 4.9 are both complete (and separable) metric spaces. The general Prokhorov's theorem is stated as follows.

Theorem 4.9 (Prokhorov's theorem in metric spaces)**.** *Let $\{\mu : \mu\in\Lambda\}$ be a family of probability measures defined on a separable metric space $(S,\mathcal{B}(S),\rho)$.*

(i) If the family $\{\mu : \mu\in\Lambda\}$ is tight, then every sequence in the family admits a weakly convergent subsequence.
(ii) Suppose further that S is complete. If every sequence in the family $\{\mu : \mu\in\Lambda\}$ admits a weakly convergent subsequence, then the family is tight.

4.3.5 An important example: $C[0,1]$

We conclude this topic by discussing a useful tightness criterion for the Example 4.9 of the path space $C[0,1]$. This is an important result for studying convergence of stochastic processes such as functional central limit theorems.

Definition 4.19. A *stochastic process* on $[0,1]$ is a family of random variables $\{X(t) : t \in [0,1]\}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Since the $X(t)$'s are random variables, there is a hidden dependence on sample points $\omega \in \Omega$. It is therefore more precise to write $X(t,\omega)$ to indicate such dependence. A useful way of looking at a stochastic process is that for each fixed $\omega \in \Omega$, the function $[0,1] \ni t \mapsto X(t,\omega)$ defines a real valued path on $[0,1]$ (called a *sample path*). In this way, a stochastic process on $[0,1]$ can be equivalently viewed as a mapping from Ω to "the space of paths".

Recall that $W = C[0,1]$ is the space of continuous functions (paths) $w : [0,1] \to \mathbb{R}$ equipped with the uniform metric (cf. Example 4.9). Let $\{X(t) : t \in [0,1]\}$ be a stochastic process defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The process is said to be *continuous*, if every sample path is continuous, i.e. for every $\omega \in \Omega$, the function $[0,1] \ni t \mapsto X(t,\omega)$ is continuous. Using the sample path viewpoint, a continuous stochastic process can be regarded as a mapping from Ω to W.

Definition 4.20. Let $X = \{X(t) : t \in [0,1]\}$ be a continuous stochastic process defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, viewed as a measurable mapping $X : (\Omega, \mathcal{F}) \to (W, \mathcal{B}(W))$. The *law* of X is the probability measure μ_X on $(W, \mathcal{B}(W))$ defined by

$$\mu_X(\Gamma) \triangleq \mathbb{P}(X \in \Gamma), \quad \Gamma \in \mathcal{B}(W).$$

We are often interested in the weak convergence of a sequence of stochastic processes $X_n(t)$. The following result provides a useful criterion for tightness in this context, which is an important ingredient in the study of weak convergence. Its proof, which is enlightening but also quite involved, is deferred to the appendix.

Theorem 4.10. *Let $X_n = \{X_n(t) : t \in [0,1]\}$ $(n \geqslant 1)$ be a sequence of continuous stochastic processes defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that the following two conditions hold true.*

(i) There exists $r > 0$ such that

$$\sup_{n \geqslant 1} \mathbb{E}[|X_n(0)|^r] < \infty.$$

(ii) There exists $\alpha, \beta, C > 0$ such that

$$\mathbb{E}[|X_n(t) - X_n(s)|^\alpha] \leqslant C|t-s|^{1+\beta}$$

for all $s, t \in [0,1]$ and $n \geqslant 1$.

Let μ_n be the law of X_n on $(W, \mathcal{B}(W))$. Then the sequence of probability measures $\{\mu_n : n \geqslant 1\}$ is tight.

Appendix. Compactness and tightness in $C[0,1]$

The characterisation of compact sets in $W = C[0,1]$ is given by the well known *Arzelà-Ascoli theorem* in functional analysis.

Theorem 4.11. *A subset $F \subseteq W$ is precompact (i.e. the closure of F is compact) if and only if the following two conditions hold true.*

(i) F is bounded at $t = 0$, in the sense that there exists $M > 0$ such that

$$|w(0)| \leqslant M \quad \text{for all } w \in F.$$

(ii) F is uniformly equicontinuous, in the sense that for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|w(t) - w(s)| < \varepsilon$$

for all $w \in F$ and $s, t \in [0,1]$ with $|t-s| < \delta$.

In particular, F is compact if and only if it is closed and Conditions (i) & (ii) hold.

Remark 4.4. Conditions (i) and (ii) can be written in the following more concise forms:

$$\text{(i)} \ \sup_{w \in F} |w(0)| < \infty \quad \text{and} \quad \text{(ii)} \ \lim_{\delta \downarrow 0} \sup_{w \in F} \Delta(\delta; w) = 0$$

respectively, where $\Delta(\delta; w)$ is the *modulus of continuity* for w defined by

$$\Delta(\delta; w) \triangleq \sup_{|t-s|<\delta} |w(t) - w(s)|. \tag{4.18}$$

Remark 4.5. In its more common form, Condition (i) is often replaced by the following uniform boundedness condition: there exists $M > 0$ such that

$$|w(t)| \leqslant M \quad \text{for all } w \in F \text{ and } t \in [0,1].$$

With the presence of Condition (ii), this uniform boundedness condition is equivalent to Condition (i) (why?).

Example 4.14. Let $L > 0$ be a fixed number. Define F to be the set of paths $w \in W$ such that $w(0) = 0$ and

$$|w(t) - w(s)| \leqslant L|t - s| \quad \forall s, t \in [0, 1].$$

Then F is compact.

Since the notion of tightness is closely related to compact sets, it is reasonable to expect that tightness over W can be characterised in terms of suitable probabilistic versions of Conditions (i) and (ii) in the Arzelà-Ascoli theorem. This is the content of the following result.

Theorem 4.12. *Let $\{\mu : \mu \in \Lambda\}$ be a family of probability measures on $(W, \mathcal{B}(W))$. Suppose that the following two conditions hold true.*

(i) One has

$$\lim_{M\to\infty} \sup_{\mu\in\Lambda} \mu\big(\{w : |w(0)| > M\}\big) = 0;$$

(ii) for any $\varepsilon > 0$, one has

$$\lim_{\delta\downarrow 0} \sup_{\mu\in\Lambda} \mu\big(\{w : |\Delta(\delta; w)| > \varepsilon\}\big) = 0. \tag{4.19}$$

Then the family $\{\mu : \mu \in \Lambda\}$ is tight.

Proof. Let $\varepsilon > 0$. Our goal is to find a compact subset $K \subseteq W$ such that $\mu(K^c) < \varepsilon$ for all $\mu \in \Lambda$. To this end, by Assumption (i) there exists $M > 0$, such that

$$\mu(\{w : |w(0)| > M\}) < \frac{\varepsilon}{2} \quad \forall \mu \in \Lambda.$$

In addition, by Assumption (ii), for each $n > 0$ there exists $\delta_n > 0$ such that

$$\mu\big(\big\{w : |\Delta(\delta_n; w)| > \frac{1}{n}\big\}\big) < \frac{\varepsilon}{2^{n+1}} \quad \forall \mu \in \Lambda.$$

Now let us define

$$\Gamma_\varepsilon \triangleq \big\{w : |w(0)| \leqslant M\big\} \cap \big(\bigcap_{n=1}^{\infty} \big\{w : |\Delta(\delta_n; w)| \leqslant \frac{1}{n}\big\}\big).$$

It is easy to check that Γ_ε satisfies the two conditions in the Arzelà-Ascoli theorem and is thus precompact (i.e. $\overline{\Gamma_\varepsilon}$ is compact). On the other hand, one also has

$$\overline{\Gamma_\varepsilon}^c \subseteq \Gamma_\varepsilon^c = \big\{w : |w(0)| > M\big\} \cup \big(\bigcup_{n=1}^{\infty} \big\{w : |\Delta(\delta_n; w)| > \frac{1}{n}\big\}\big).$$

It follows that

$$\mu(\overline{\Gamma_\varepsilon}^c) \leqslant \mu(\{w : |w(0)| > M) + \sum_{n=1}^{\infty} \mu\big(\{w : |\Delta(\delta_n; w)| > \frac{1}{n}\}\big)$$
$$< \frac{\varepsilon}{2} + \sum_{n=1}^{\infty} \frac{\varepsilon}{2^{n+1}} < \varepsilon$$

for all $\mu \in \Lambda$. This gives the tightness property. □

We now use the tightness criterion given by Theorem 4.12 to prove Theorem 4.10. We first recall an elementary fact about real numbers that will be needed for the proof. We shall make use of *dyadic partitions* of $[0, 1]$. For $m \geqslant 0$, define

$$D_m = \{k/2^m : 0 \leqslant k \leqslant 2^m\}$$

to be the m-th dyadic partition of $[0, 1]$. Let $D \triangleq \cup_{m=0}^{\infty} D_m$. D is the collection of dyadic points on $[0, 1]$. Every real number $t \in [0, 1]$ admits a unique dyadic expansion

$$t = \sum_{i=0}^{\infty} a_i(t) 2^{-i}$$

where $a_i(t) = 0$ or 1 for each i. If $t \in D$, then the expansion is a finite sum (i.e. there are at most finitely many 1's among the $a_i(t)$'s). For instance,

$$D \ni \frac{11}{16} = 0 \cdot 2^{-0} + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} + 0 + 0 + \cdots .$$

Proof of Theorem 4.10. One needs to check the two conditions in Theorem 4.12 for the laws of the sequence $\{X_n : n \geqslant 1\}$. Condition (i) is a simple consequence of Chebyshev's inequality:

$$\mathbb{P}(|X_n(0)| > M) \leqslant \frac{\mathbb{E}[|X_n(0)|^r]}{M^r} \leqslant \frac{L}{M^r}$$

where

$$L \triangleq \sup_{n \geqslant 1} \mathbb{E}[|X_n(0)|^r] < \infty.$$

It follows that

$$\lim_{M \to \infty} \sup_{n \geqslant 1} \mathbb{P}(|X_n(0)| > M) = 0$$

and thus Condition (i) holds. Checking Condition (ii) requires deeper probabilistic reasoning. Since the following argument is uniform in n, to ease notation we will write $Y(t) = X_n(t)$.

Let $\gamma \in (0, \beta/\alpha)$ be a fixed number (α, β are the exponents appearing in the second assumption of the theorem). According to Chebyshev's inequality, one has

$$\begin{aligned}
&\mathbb{P}\big(|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant 2^{\alpha\gamma m} \cdot \mathbb{E}\big[|Y(k/2^m) - Y((k-1)/2^m)|^\alpha\big] \\
&\leqslant C \cdot 2^{-m(1+\beta-\alpha\gamma)},
\end{aligned}$$

for all $1 \leqslant k \leqslant 2^m$. It follows that,

$$\begin{aligned}
&\mathbb{P}\big(\max_{1\leqslant k\leqslant 2^m} |Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant \mathbb{P}\big(\bigcup_{k=1}^{2^m} \big\{|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big\}\big) \\
&\leqslant \sum_{k=1}^{2^m} \mathbb{P}\big(|Y(k/2^m) - Y((k-1)/2^m)| > 2^{-\gamma m}\big) \\
&\leqslant C \cdot 2^{-m(\beta-\alpha\gamma)}.
\end{aligned}$$

Since $\gamma < \beta/\alpha$, the right hand side is summable in m. In particular, given $\eta > 0$ there exists $p \geqslant 1$, such that if one defines

$$\Omega_p \triangleq \bigcup_{m=p}^{\infty} \big\{ \max_{1\leqslant k\leqslant 2^m} \big|Y(k/2^m) - Y((k-1)/2^m)\big| > 2^{-\gamma m}\big\},$$

then

$$\mathbb{P}(\Omega_p) \leqslant C \cdot \sum_{m=p}^{\infty} 2^{-m(\beta-\alpha\gamma)} < \eta.$$

Now let $\varepsilon > 0$ be given. We claim that $\{\Delta(\delta; Y) > \varepsilon\} \subseteq \Omega_p$, or equivalently,

$$\Omega_p^c \subseteq \big\{\Delta(\delta; Y) \leqslant \varepsilon\big\}$$

for δ small enough, where $\Delta(\delta; w)$ is the modulus of continuity for Y defined by (4.18). To this end, suppose that Ω_p^c occurs. Then one has

$$|Y(k/2^m) - Y((k-1)/2^m)| \leqslant 2^{-\gamma m} \quad \text{for all } m \geqslant p \text{ and } 1 \leqslant k \leqslant 2^m.$$

Let $s, t \in D$ (the set of dyadic points on $[0,1]$) be such that

$$0 < |t-s| < \delta \triangleq 2^{-p}.$$

For each l we use the notation s_l (respectively, t_l) to denote the largest l-th dyadic point in D_l such that $s_l \leqslant s$ (respectively, $t_l \leqslant t$). Let $m \geqslant p$ be the unique integer such that

$$2^{-(m+1)} < |t-s| < 2^{-m}.$$

Note that either $s_m = t_m$ or $t_m - s_m = 2^{-m}$. It follows that

$$\begin{aligned} |Y(t)-Y(s)| &\leqslant |Y(t_m)-Y(s_m)| + \sum_{l=m}^{\infty} |Y(t_{l+1})-Y(t_l)| + \sum_{l=m}^{\infty} |Y(s_{l+1})-Y(s_l)| \\ &\leqslant 2^{-\gamma m} + 2\sum_{l=m}^{\infty} 2^{-\gamma(l+1)} = \big(1+\frac{2}{2^\gamma-1}\big)\cdot 2^{-\gamma m} \\ &\leqslant 2^\gamma\big(1+\frac{2}{2^\gamma-1}\big)|t-s|^\gamma < 2^\gamma\big(1+\frac{2}{2^\gamma-1}\big)\cdot 2^{-p\gamma}. \end{aligned} \tag{4.20}$$

If it is further assumed that p satisfies

$$2^\gamma\big(1+\frac{2}{2^\gamma-1}\big)\cdot 2^{-p\gamma} < \varepsilon$$

at the beginning, then (4.20) gives that $|Y(t)-Y(s)| < \varepsilon$. Since this holds for all $s, t \in D$ with $|t-s| < \delta$ and D is dense in $[0,1]$, by continuity one concludes that $\Delta(\delta; Y) < \varepsilon$.

Now the proof of Theorem 4.10 is complete.

□

5 Sequences and sums of independent random variables

This chapter is an introduction to the realm of probabilistic limit theory. The study of asymptotic behaviours of stochastic processes, random dynamical systems, complex random structures etc. has been (and will continue to be) a central theme of modern research in probability theory. Basic limit theorems such as law of large numbers, central limit theorem etc. also have enormous applications in a variety of mathematical and scientific areas.

To introduce some of the essential probabilistic ideas, in this chapter we will focus on the most basic and classical situation: sequence of *independent* random variables. In particular, we are going to establish the following three major theorems:

(i) Kolmogorov's two-series theorem for random series;
(ii) The strong law of large numbers;
(iii) Cramér's theorem of large deviations.

Among others, these three theorems are of fundamental importance in probability theory; apart from their broad applications, the proofs of these results also contain deep mathematical ideas and techniques that have far-reaching implications.

In Section 5.1, we begin by introducing Kolmogorov's zero-one law and the Borel-Cantelli lemma. These are powerful tools which are particularly useful for proving various probabilistic limit theorems. In Section 5.3, we establish Kolmogorov's two-series theorem which gives a useful criterion for the convergence of random series. It also provides a basic tool for proving the strong law of large numbers. In Sections 5.4 and 5.6, we establish the strong law of large numbers and the large deviation principle respectively in the context of i.i.d. sequences. The former result mathematically justifies the phenomenon that the sample average asymptotically stabilises at its theoretical mean. The latter result quantifies the concentration of measure associated with the law of large numbers.

5.1 Kolmogorov's zero-one law and the Borel-Cantelli lemma

Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables. We are interested in deriving conditions under which the random series $\sum_{n=1}^{\infty} X_n$ is convergent. When the X_n's are independent, it is a remarkable theorem (Kolmogorov's zero-one law) that the

event

$$\Big\{\sum_{n=1}^{\infty} X_n \text{ is convergent}\Big\}$$

has probability either zero or one. Before discussing this result, we first spend some time recalling the general definition of independence.

5.1.1 Definition of independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space. Let $\{\mathcal{G}_n : n \geqslant 1\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$.

Definition 5.1. We say that $\mathcal{G}_1, \mathcal{G}_2, \cdots$ are *independent* if whenever $G_i \in \mathcal{G}_i$ and $i_1, \cdots, i_n$ are distinct, one has

$$\mathbb{P}\big(G_{i_1} \cap \cdots \cap G_{i_n}\big) = \mathbb{P}(G_{i_1}) \cdots \mathbb{P}(G_{i_n}).$$

Remark 5.1. One can of course just talk about the independence between two (or finitely many) σ-algebras: $\mathcal{G}_1, \mathcal{G}_2$ are independent if

$$\mathbb{P}(G_1 \cap G_2) = \mathbb{P}(G_1)\mathbb{P}(G_2) \quad \forall G_i \in \mathcal{G}_i \ (i = 1, 2).$$

Definition 5.1 is the most general one; it covers all notions of independence we have seen before. For instance, two random variables X, Y are independent if and only if $\sigma(X), \sigma(Y)$ are independent. Here

$$\sigma(X) \triangleq \sigma(\{X \leqslant x\} : x \in \mathbb{R}) = X^{-1}\mathcal{B}(\mathbb{R})$$

denotes the *σ-algebra generated by X*. A collection of events $A_1, \cdots, A_n$ are independent if and only if the σ-algebras $\sigma(A_1), \cdots, \sigma(A_n)$ are independent (recall that $\sigma(A_i) \triangleq \{\emptyset, A_i, A_i^c, \Omega\}$). This is also equivalent to saying that the random variables $\mathbf{1}_{A_1}, \cdots, \mathbf{1}_{A_n}$ are independent.

Definition 5.2. A sequence $\{X_n : n \geqslant 1\}$ of random variables are said to be *independent*, if the sequence $\sigma(X_1), \sigma(X_2), \cdots$ of σ-algebras are independent in the sense of Definition 5.1.

5.1.2 Tail σ-algebras and Kolmogorov's zero-one law

Let $\{\mathcal{G}_n : n \geqslant 1\}$ be a given sequence of sub-σ-algebras over $(\Omega, \mathcal{F}, \mathbb{P})$. For each $n \geqslant 1$, let us define

$$\mathcal{T}_n \triangleq \sigma\Big(\bigcup_{k=n+1}^{\infty} \mathcal{G}_k\Big)$$

to be the σ-algebra generated by the tail sequence $\mathcal{G}_{n+1}, \mathcal{G}_{n+2}, \cdots$. We then set

$$\mathcal{T} \triangleq \bigcap_{n=1}^{\infty} \mathcal{T}_n.$$

It is routine to check that $\mathcal{T}$ is a sub-σ-algebra of $\mathcal{F}$. Heuristically, $\mathcal{T}$ is generated by the information encoded in the "infinitely far tail" of the sequence $\{\mathcal{G}_n\}$. This leads to the following definition.

Definition 5.3. The σ-algebra $\mathcal{T}$ is called the *tail σ-algebra* of the sequence $\{\mathcal{G}_n\}$. Members of $\mathcal{T}$ are called *tail events*.

The most important situation to have in mind (which will always be the case in our study unless otherwise stated) is when $\{\mathcal{G}_n\}$ come from a sequence of random variables $\{X_n\}$ (i.e. $\mathcal{G}_n = \sigma(X_n)$). Below are some important examples of tail events in this case.

Example 5.1. The following events are tail events of the sequence $\{X_n\}$:

$$F_1 = \big\{ \lim_{n\to\infty} X_n \text{ exists} \big\} \triangleq \big\{ \omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists} \big\},$$
$$F_2 = \big\{ \sum_{n=1}^{\infty} X_n \text{ is convergent} \big\},$$
$$F_3 = \big\{ \lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} \text{ exists} \big\}.$$

We only look at F_2 and leave the other two as an exercise. To show that $F_2 \in \mathcal{T}$, by definition one needs to show that $F_2 \in \mathcal{T}_n$ for each fixed n. But this is obvious, since F_2 can also be written as

$$F_2 = \big\{ \sum_{m=n+1}^{\infty} X_m \text{ is convergent} \big\},$$

which is clearly measurable with respect to $X_{n+1}, X_{n+2}, \cdots$.

Example 5.2. The event $\{X_n > 0 \ \forall n\}$ is not a tail event, since its occurrence relies on the values of all the X_n's and cannot be determined by the information encoded in $\{X_{n+1}, X_{n+2}, \cdots\}$ for arbitrarily large n.

Kolmogorov's zero-one law asserts that tail events of an *independent* sequence always have probabilities either zero or one.

Theorem 5.1. *Let $\{\mathcal{G}_n : n \geqslant 1\}$ be an independent sequence of sub-σ-algebras over $(\Omega, \mathcal{F}, \mathbb{P})$. For any $A \in \mathcal{T}$, one has*

$$\mathbb{P}(A) = 0 \text{ or } 1.$$

Proof. Define

$$\mathcal{F}_n \triangleq \sigma(\mathcal{G}_1, \cdots, \mathcal{G}_n).$$

By using Dynkin's π-λ theorem, one sees that $\mathcal{F}_n$ and $\mathcal{T}_m$ are independent for all $m \geqslant n$. In particular, $\mathcal{F}_n$ and $\mathcal{T}$ are independent (for every n). Let us define

$$\mathcal{F}_\infty \triangleq \sigma\big(\bigcup_{n=1}^{\infty} \mathcal{F}_n\big) = \sigma(\mathcal{G}_1, \mathcal{G}_2, \cdots).$$

It follows from a similar reason that $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent. Given $A \in \mathcal{T}$, one can trivially write $A = A \cap A$. By viewing the first A as a member of $\mathcal{F}_\infty$ and the second A as a member of $\mathcal{T}$, one concludes by their independence that

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2 \iff \mathbb{P}(A)\big(1 - \mathbb{P}(A)\big) = 0.$$

As a consequence, $\mathbb{P}(A) = 0$ or 1. □

The following result is an immediate consequence of Theorem 5.1. We leave its proof as an exercise.

Corollary 5.1. *Under the assumption of Theorem 5.1, let Y be a random variable that is measurable with respect to $\mathcal{T}$. Then Y is degenerate, i.e. $Y =$ constant a.s.*

5.1.3 The Borel-Cantelli lemma

According to Example 5.1 and Theorem 5.1, for an independent sequence $\{X_n\}$ the event

$$\big\{\sum_{n=1}^{\infty} X_n \text{ is convergent}\big\}$$

is a tail event and thus has probability either zero or one. However, it is a priori not clear which case occurs. In probability theory, there is a particularly useful tool (the Borel-Cantelli lemma) which can often be applied to answer such questions.

We first recall the following definitions. Let $\{A_n : n \geqslant 1\}$ be a given sequence of events. We use the notation

$$\overline{\lim_{n\to\infty}} A_n \triangleq \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$$

to denote the event that "A_n happens for infinitely many n's (or "A_n happens infinitely often"). Sometimes we simply write "A_n i.o." Respectively, the notation

$$\varliminf_{n\to\infty} A_n \triangleq \bigcup_{n=1}\bigcap_{m=n} A_m$$

denotes the event that "from some point on every A_n happens" (or "A_n happens for all but finitely many n's"). Sometimes we simply write "A_n eventually." It is obvious that

$$\big(\varlimsup_{n\to\infty} A_n\big)^c = \varliminf_{n\to\infty} A_n^c,\ \big(\varliminf_{n\to\infty} A_n\big)^c = \varlimsup_{n\to\infty} A_n^c.$$

The Borel-Cantelli lemma is stated as follows.

Theorem 5.2. *Let $\{A_n : n \geqslant 1\}$ be a sequence of events defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

(i) [1st Borel-Cantelli lemma] If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}\big(\varlimsup_{n\to\infty} A_n\big) = 0$.

(ii) [2nd Borel-Cantelli lemma] Suppose further that the sequence $\{A_n\}$ are independent. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}\big(\varlimsup_{n\to\infty} A_n\big) = 1$.

Proof. (i) By the definition of $\varlimsup_{n\to\infty} A_n$ and the assumption, one has

$$\mathbb{P}\big(\varlimsup_{n\to\infty} A_n\big) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{m=n}^{\infty} A_m\Big) \leqslant \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0.$$

(ii) By considering the complement, one has

$$\mathbb{P}\big(\big(\varlimsup_{n\to\infty} A_n\big)^c\big) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m^c\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{\infty} A_m^c\Big) = \lim_{n\to\infty}\lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{m=n}^{N} A_m^c\Big).$$

Now we study the above limit. First of all, by independence one knows that

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{m=n}^{N} A_m^c\Big) &= \big(1-\mathbb{P}(A_n)\big)\cdots\big(1-\mathbb{P}(A_N)\big)\\ &= \exp\Big(\sum_{m=n}^{N}\log(1-\mathbb{P}(A_m)\Big) \leqslant \exp\Big(-\sum_{m=n}^{N}\mathbb{P}(A_m))\Big),\end{aligned}$$

where we used the elementary inequality that $\log(1-x) \leqslant -x$. Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, by letting $N \to \infty$ one finds that

$$\lim_{N\to\infty} \mathbb{P}\big(\bigcap_{m=n}^{N} A_m^c\big) \leqslant \exp\big(-\sum_{m=n}^{\infty} \mathbb{P}(A_m)\big) = \exp(-\infty) = 0.$$

This is true for every n. As a consequence,

$$\mathbb{P}\big(\big(\overline{\lim_{n\to\infty}} A_n\big)^c\big) = \lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\big(\bigcap_{m=n}^{N} A_m^c\big) = 0$$

and the result thus follows. □

As illustrated by the following example, the independence assumption is essential for the second Borel-Cantelli's lemma to hold.

Example 5.3. Let X be a uniform random variable over $[0, 1]$. Define $A_n \triangleq \{X \leqslant 1/n\}$ $(n \geqslant 1)$. Then $\mathbb{P}(A_n) = 1/n$ and thus $\sum_n \mathbb{P}(A_n) = \infty$. However, one has

$$\overline{\lim_{n\to\infty}} A_n = \{X = 0\},$$

which is an event of zero probability. The issue here is that the A_n's are not independent.

Remark 5.2. The second Borel-Cantelli lemma remains valid under the assumption of *pairwise independence*, i.e. only assuming that A_n and A_m are independent for each pair of $n \neq m$.

The following example is a trivial application of the second Borel-Cantelli lemma, although it may still look surprising to non-probabilists.

Example 5.4. Suppose that one tosses a fair coin independently in a sequence. Let A_1 be the event that the first 10^{10} consecutive tosses all result in "heads". Let A_2 be the event that the next 10^{10} consecutive tosses all result in "heads", and so forth. It is apparent that the A_n's are independent and each of them has a rather small probability:

$$\mathbb{P}(A_n) = 0.5^{10^{10}} > 0.$$

Nonetheless, one still has $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. It follows from the second Borel-Cantelli lemma that

$$\mathbb{P}\big(\overline{\lim_{n\to\infty}} A_n\big) = 1.$$

In other words, with probability one, there will be infinitely many intervals of length 10^{10} that contain only "heads"! There is another interesting way of describing this phenomenon. If a monkey randomly types one letter at each time, then with probability one it will eventually produce an exact copy of Shakespeare's "Hamlet" (in fact infinitely many copies!). The next question is: *how long on average does the monkey need to produce such a copy for the first time*? We will answer this question in Section 8.7.1 below by using martingale theory.

5.1.4 An application to random walks: recurrence / transience

We discuss an interesting application of Kolmogorov's zero-one law and the Borel-Cantelli lemma to the recurrence / transience of simple random walks. Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence with distribution

$$\mathbb{P}(X_1 = 1) = p, \ \mathbb{P}(X_1 = -1) = q \triangleq 1 - p,$$

where $p \in (0, 1)$ is given fixed. For each $n \geqslant 1$, we set $S_n \triangleq X_1 + \cdots + X_n$ ($S_0 \triangleq 0$). The sequence $\{S_n\}$ defines a *random walk on the integer lattice* $\mathbb{Z}$. Suppose that its current location is $S_n = x \in \mathbb{Z}$. In the next move, it will jump to $x + 1$ with probability p and to $x - 1$ with probability q respectively. We are interested in the probability that the random walk will return to the origin infinitely often.

Proposition 5.1. *(i) Suppose that $p \neq 1/2$. Then $\mathbb{P}(S_n = 0 \ i.o.) = 0$.*
(ii) Suppose that $p = 1/2$. Then $\mathbb{P}(S_n = 0 \ i.o.) = 1$.

Proof. (i) Note that S_n can only return to the origin in even number of steps. It is elementary to see that

$$\mathbb{P}(S_{2n} = 0) = \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2} (pq)^n.$$

In addition, recall from Stirling's formula (cf. Proposition 7.1 below) that

$$n! \sim \sqrt{2\pi n} \big(\frac{n}{e}\big)^n \quad \text{as } n \to \infty,$$

where $a_n \sim b_n$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. As a consequence,

$$\mathbb{P}(S_{2n} = 0) \sim \frac{\sqrt{2\pi \cdot 2n}(2n/e)^{2n}}{\big(\sqrt{2\pi n}(n/e)^n\big)^2} (pq)^n = \frac{1}{\sqrt{\pi n}} (4pq)^n \quad \text{as } n \to \infty. \tag{5.1}$$

Since $p \neq 1/2$, one has

$$4pq = 4p(1-p) = 4p - 4p^2 = 1 - (1-2p)^2 < 1.$$

It follows from (5.1) that

$$\sum_{n=1}^{\infty} \mathbb{P}(S_{2n} = 0) < \infty.$$

According to the first Borel-Cantelli lemma, one concludes that

$$\mathbb{P}(S_n = 0 \text{ i.o.}) = \mathbb{P}(S_{2n} = 0 \text{ i.o.}) = 0.$$

(ii) If $p = 1/2$, the relation (5.1) yields that

$$\sum_{n=1}^{\infty} \mathbb{P}(S_{2n} = 0) = \infty.$$

However, one cannot apply the second Borel-Cantelli lemma directly since the events $\{S_{2n} = 0\}$ $(n \geqslant 1)$ are not independent. To solve the problem, we claim that

$$\mathbb{P}\big(\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{n}} = \infty \text{ and } \underline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{n}} = -\infty\big) = 1. \tag{5.2}$$

If this is true, it implies that with probability one, there are subsequences of times along which S_n explodes to ∞ and $-\infty$ respectively. In particular, with probability one it has to return to the origin infinitely often (in order to "fluctuate between $\pm\infty$").

To prove (5.2), let us set

$$A \triangleq \big\{\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{n}} = \infty\big\}.$$

For each $M > 0$ we define

$$A_M \triangleq \big\{\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{n}} \geqslant M\big\}.$$

Then $A_M \downarrow A$ and thus $\mathbb{P}(A_M) \downarrow \mathbb{P}(A)$ as $M \to \infty$. In order to prove $\mathbb{P}(A) = 1$, it suffices to show that $\mathbb{P}(A_M) = 1$ for all M. To this end, a key observation is that A_M is a tail event of the sequence $\{X_n\}$ (why?). As a result of Theorem 5.1, it is enough to show that $\mathbb{P}(A_M) > 0$. Since

$$\overline{\lim_{n\to\infty}} \big\{\frac{S_n}{\sqrt{n}} > M\big\} \subseteq A_M,$$

one finds that

$$\mathbb{P}(A_M) \geqslant \mathbb{P}\big(\overline{\lim_{M\to\infty}}\{\frac{S_n}{\sqrt{n}} > M\}\big) = \mathbb{P}\big(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\{\frac{S_m}{\sqrt{m}} > M\}\big)$$
$$= \lim_{n\to\infty}\mathbb{P}\big(\bigcup_{m=n}^{\infty}\{\frac{S_m}{\sqrt{m}} > M\}\big) \geqslant \lim_{n\to\infty}\mathbb{P}\big(\frac{S_n}{\sqrt{n}} > M\big).$$

According to the classical central limit theorem (cf. Theorem 7.1 below), one has

$$\lim_{n\to\infty}\mathbb{P}\big(\frac{S_n}{\sqrt{n}} > M\big) = \frac{1}{\sqrt{2\pi}}\int_M^{\infty} e^{-x^2/2}dx > 0.$$

Therefore $\mathbb{P}(A_M) > 0$, which implies that $\mathbb{P}(A_M) = 1$ (for all M) and thus $\mathbb{P}(A) = 1$. The second event in (5.2) is treated in the same way (or simply use $\{S_n\} \overset{\text{law}}{=} \{-S_n\}$). □

5.2 The weak law of large numbers

We demonstrate another important application of the Borel-Cantelli lemma: the *weak law of large numbers*. We first prove a simple property for the expectation that will be used later on.

Lemma 5.1. *Let X be non-negative random variable. Then one has*

$$\mathbb{E}[X] < \infty \iff \sum_{n=1}^{\infty}\mathbb{P}(X > n) < \infty. \tag{5.3}$$

Proof. We only prove necessity part and leave the other direction as an exercise. According to Proposition 3.2, one has

$$\mathbb{E}[X] = \int_0^{\infty}\mathbb{P}(X > x)dx = \sum_{n=1}^{\infty}\int_{n-1}^{n}\mathbb{P}(X > x)dx$$
$$\geqslant \sum_{n=1}^{\infty}\int_{n-1}^{n}\mathbb{P}(X > n)dx = \sum_{n=1}^{\infty}\mathbb{P}(X > n).$$

The implication " $\Longrightarrow$ " in (5.3) thus follows. □

The weak law of large numbers (LLN) is stated as follows.

Theorem 5.3. *Let $\{X_n : n \geqslant 1\}$ be a sequence of pairwise independent, identically distributed random variables with finite mean m. Define $S_n \triangleq X_1 + \cdots + X_n$. Then*

$$\lim_{n\to\infty} \frac{S_n}{n} = m \quad \textit{in prob.} \tag{5.4}$$

Before developing its proof, we first examine a particularly simple but enlightening situation. For the moment, let us further assume that all the X_n's have finite variance σ^2. By Chebyshev's inequality, in this case one has

$$\mathbb{P}\big(\big|\frac{S_n}{n} - m\big| > \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\mathrm{Var}\big[\frac{S_n}{n}\big] = \frac{1}{\varepsilon^2 n^2}\mathrm{Var}[S_n] = \frac{\sigma^2}{n\varepsilon^2}. \tag{5.5}$$

This trivially gives the convergence (5.4). The key point here is that $\mathrm{Var}[S_n] = o(n^2)$ as $n \to \infty$ ($o(n^2)$ means a real sequence a_n such that $a_n/n^2 \to 0$).

The main idea of proving the weak LLN in the general case is to truncate X_n to a bounded random variable. This is a basic technique that will be used again in the study of the strong LLN.

Proof of Theorem 5.3. We divide the proof into several steps. Let $F(x)$ be the distribution function of X_1 (equivalently, of any X_n).

Step one: truncation. We define

$$Y_n \triangleq \begin{cases} X_n, & \text{if } |X_n| \leqslant n; \\ 0, & \text{otherwise.} \end{cases}$$

Observe that $\{X_n \neq Y_n\} = \{|X_n| > n\}$. Since all the X_n's are identically distributed, one has

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty,$$

where the last summability follows from Lemma 5.1. According to the first Borel-Cantelli lemma,

$$\mathbb{P}\big(X_n \neq Y_n \ \text{ for infinitely many } n\big) = 0.$$

In other words, with probability one $X_n = Y_n$ for all n sufficiently large.

Step two: the weak LLN for $\{Y_n\}$. Define $T_n \triangleq Y_1 + \cdots + Y_n$. Inspired by the earlier argument for (5.5), let us estimate $\mathrm{Var}[T_n]$. Since $Y_1, \cdots, Y_n$ are independent, one has

$$\mathrm{Var}[T_n] = \sum_{j=1}^{n} \mathrm{Var}[Y_j] \leqslant \sum_{j=1}^{n} \mathbb{E}[Y_j^2].$$

Our goal is to show that the above quantity is of order $o(n^2)$. By the construction of Y_n, one has

$$\sum_{j=1}^{n} \mathbb{E}[Y_j^2] = \sum_{j=1}^{n} \mathbb{E}[X_j^2 \mathbf{1}_{\{|X_j| \leqslant j\}}] = \sum_{j=1}^{n} \int_{\{|x| \leqslant j\}} x^2 dF(x)$$
$$= \sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} x^2 dF(x) + \sum_{\sqrt{n} < j \leqslant n} \int_{\{|x| \leqslant j\}} x^2 dF(x). \tag{5.6}$$

We estimate the above two sums separately. For the first one,

$$\sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} x^2 dF(x) \leqslant \sum_{j \leqslant \sqrt{n}} \int_{\{|x| \leqslant j\}} \sqrt{n} \cdot |x| dF(x)$$
$$\leqslant \sum_{j \leqslant \sqrt{n}} \sqrt{n} \int_{-\infty}^{\infty} |x| dF(x) = n \cdot \mathbb{E}[|X_1|].$$

For the second one,

$$\sum_{\sqrt{n} < j \leqslant n} \int_{\{|x| \leqslant j\}} x^2 dF(x) = \sum_{\sqrt{n} < j \leqslant n} \Big(\int_{\{|x| \leqslant \sqrt{n}\}} x^2 dF(x) + \int_{\{\sqrt{n} < |x| \leqslant j\}} x^2 dF(x) \Big)$$
$$\leqslant n\sqrt{n} \cdot \int_{-\infty}^{\infty} |x| dF(x) + n^2 \int_{\{|x| > \sqrt{n}\}} |x| dF(x)$$
$$= n\sqrt{n} \cdot \mathbb{E}[|X_1|] + n^2 \mathbb{E}[|X_1| \cdot \mathbf{1}_{\{|X_1| > \sqrt{n}\}}].$$

Note that (cf. Proposition 2.8)

$$\lim_{n \to \infty} \mathbb{E}[|X_1| \cdot \mathbf{1}_{\{|X_1| > \sqrt{n}\}}] = 0.$$

As a result, both sums on the right hand side of (5.6) is of order $o(n^2)$. Therefore, $\mathrm{Var}[T_n] = o(n^2)$. By the same argument leading to (5.5), one obtains tha

$$\lim_{n \to \infty} \frac{T_n - \mathbb{E}[T_n]}{n} = 0 \quad \text{in prob.}$$

Step three: relating back to the sequence $\{X_n\}$. To complete the proof, let us compare $\frac{S_n}{n} - m$ with $\frac{T_n - \mathbb{E}[T_n]}{n}$. Firstly, observe that

$$\Big| \Big(\frac{S_n}{n} - m \Big) - \Big(\frac{T_n - \mathbb{E}[T_n]}{n} \Big) \Big| \leqslant \frac{|S_n - T_n|}{n} + \Big| \frac{\mathbb{E}[T_n]}{n} - m \Big|.$$

In Step One, we have seen that with probability one $X_n = Y_n$ for all sufficiently large n. This implies that with probability one,

$$S_n - T_n = (X_1 - Y_1) + \cdots + (X_n - Y_n)$$

stops depending on n after some point and thus

$$\frac{|S_n - T_n|}{n} \to 0 \quad \text{as } n \to \infty.$$

In addition, it is apparent that

$$\mathbb{E}[Y_n] = \int_{\{|x| \leqslant n\}} x dF(x) \to \int_{-\infty}^{\infty} x dF(x) = m$$

as $n \to \infty$. It follows that

$$\frac{\mathbb{E}[T_n]}{n} = \frac{\mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_n]}{n} \to m,$$

where we used the elementary analytic fact that

$$a_n \to a \in \mathbb{R} \implies \frac{a_1 + \cdots + a_n}{n} \to a. \tag{5.7}$$

To summarise, one concludes that with probability one,

$$\lim_{n \to \infty} \left| \left(\frac{S_n}{n} - m \right) - \left(\frac{T_n - \mathbb{E}[T_n]}{n} \right) \right| = 0.$$

Combining with Step Two, the result thus follows.

□

Remark 5.3. It is a remarkable fact that the conclusion of Theorem 5.3 can be strengthened to almost sure convergence under exactly the same assumption, hence yielding a strong LLN. In Section 5.4, we will prove such a result under the stronger assumption of total independence.

5.3 Kolmogorov's two-series theorem

Our study of the strong LLN relies heavily on techniques from random series. We develop a few relevant tools in this section. Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 5.4. We say that the random series $\sum_{n=1}^{\infty} X_n$ is *convergent almost surely* (a.s.), if

$$\mathbb{P}\big(\{\omega : \sum_{n=1}^{\infty} X_n(\omega) \text{ is convergent}\}\big) = 1.$$

Our first task is to derive a characterisation of the a.s. convergence of a random series. In real analysis, the convergence of a real series $\sum_{n=1}^{\infty} x_n$ is characterised by the *Cauchy criterion*: the series $\sum_{n=1}^{\infty} x_n$ is convergent if and only if for any $\varepsilon > 0$, there exists $n \geqslant 1$ such that for any $l \geqslant n$ one has

$$\big|s_l - s_n\big| < \varepsilon,$$

where $s_n \triangleq x_1 + \cdots + x_n$. Equivalently,

$$\sum_{n=1}^{\infty} x_n \text{ not convergent } \iff \exists \varepsilon, \forall n, \exists l \geqslant n \text{ s.t. } |s_l - s_n| \geqslant \varepsilon. \tag{5.8}$$

The statement "$\exists l \geqslant n$ s.t. $|s_l - s_n| \geqslant \varepsilon$" can clearly be replaced by

$$\exists N \geqslant n \text{ s.t. } \max_{n \leqslant l \leqslant N} |s_l - s_n| \geqslant \varepsilon.$$

In the context of random series, it is thus reasonable to expect that $\sum_n X_n$ is convergent a.s. if and only if

$$\mathbb{P}\big(\exists \varepsilon, \forall n, \exists N \geqslant n \text{ s.t. } \max_{n \leqslant l \leqslant N} |S_l - S_n| \geqslant \varepsilon\big) = 0,$$

where $S_n \triangleq X_1 + \cdots + X_n$. This is precisely the following probabilistic version of the Cauchy criterion.

Proposition 5.2. *Let $\{X_n : n \geqslant 1\}$ be a sequence of random variables and set $S_n \triangleq X_1 + \cdots + X_n$. The random series $\sum_{n=1}^{\infty} X_n$ is convergent a.s. if and only if*

$$\lim_{n \to \infty} \lim_{N \to \infty} \mathbb{P}\big(\max_{n \leqslant l \leqslant N} |S_l - S_n| \geqslant \varepsilon\big) = 0 \tag{5.9}$$

for all $\varepsilon > 0$.

Proof. One can clearly pretend that $\varepsilon \in \mathbb{Q}$ (why?). Based on the above reasoning, it is seen that

$$\mathbb{P}\big(\sum_{n=1}^{\infty} X_n \text{ is not convergent}\big) = 0 \iff \mathbb{P}\big(\bigcup_{\varepsilon > 0, \varepsilon \in \mathbb{Q}} \bigcap_{n \geqslant 1} \bigcup_{N \geqslant n} \big\{ \max_{n \leqslant l \leqslant N} |S_l - S_n| \geqslant \varepsilon\big\}\big) = 0.$$

The last property is equivalent to that

$$\mathbb{P}\Big(\bigcap_{n\geqslant 1}\bigcup_{N\geqslant n}\big\{\max_{n\leqslant l\leqslant N}|S_l-S_n|\geqslant\varepsilon\big\}\Big)=0\quad\forall\varepsilon\in\mathbb{Q}\cap(0,\infty).$$

By the continuity of probability measures, one also has

$$\mathbb{P}\Big(\bigcap_{n\geqslant 1}\bigcup_{N\geqslant n}\big\{\max_{n\leqslant l\leqslant N}|S_l-S_n|\geqslant\varepsilon\big\}\Big)=\lim_{n\to\infty}\lim_{N\to\infty}\mathbb{P}\big(\max_{n\leqslant l\leqslant N}|S_l-S_n|\geqslant\varepsilon\big).$$

The result thus follows. □

The probabilistic Cauchy criterion (5.9) is difficult to verify directly in general. Nonetheless, there is a particularly simple and useful criterion in the independent context. This is known as *Kolmogorov's two-series theorem.*

Theorem 5.4. *Let $\{X_n : n\geqslant 1\}$ be a sequence of independent random variables. Suppose that each X_n has finite mean and variance. If both of the real series $\sum_n\mathbb{E}[X_n]$ and $\sum_n\mathrm{Var}[X_n]$ are convergent, then the random series $\sum_n X_n$ is convergent a.s.*

The above theorem is almost an immediate consequence of the following inequality of Kolmogorov, whose proof is truly ingenious.

Lemma 5.2 (Kolmogorov's maximal inequality)**.** *Let $X_1,\cdots,X_n$ be independent random variables. Suppose that $\mathbb{E}[X_k]=0$ and $\mathrm{Var}[X_k]<\infty$ for each k. Then for any $\varepsilon>0$, one has*

$$\mathbb{P}\big(\max_{1\leqslant k\leqslant n}|S_k|\geqslant\varepsilon\big)\leqslant\frac{1}{\varepsilon^2}\sum_{k=1}^{n}\mathrm{Var}[X_k],\tag{5.10}$$

where $S_k\triangleq X_1+\cdots+X_k$.

Proof. We decompose the event

$$A\triangleq\big\{\max_{1\leqslant k\leqslant n}|S_k|\geqslant\varepsilon\big\}$$

according to the first k such that $|S_k|\geqslant\varepsilon$. More precisely, for each $1\leqslant k\leqslant n$ we introduce the event

$$A_k\triangleq\big\{|S_1|<\varepsilon,\cdots,|S_{k-1}|<\varepsilon,\ |S_k|\geqslant\varepsilon\big\}.$$

It is obvious that $A_1, \cdots, A_n$ are disjoint and $A = \cup_{k=1}^n A_k$. Therefore, one has

$$\mathbb{P}\big(A\big) = \sum_{k=1}^n \mathbb{P}(A_k) \leqslant \frac{1}{\varepsilon^2} \sum_{k=1}^n \mathbb{E}[S_k^2 \mathbf{1}_{A_k}], \tag{5.11}$$

where the last inequality follows from the fact that $|S_k| \geqslant \varepsilon$ on A_k.

Here is the crucial point. We claim that

$$\mathbb{E}[S_k^2 \mathbf{1}_{A_k}] \leqslant \mathbb{E}[S_n^2 \mathbf{1}_{A_k}] \tag{5.12}$$

for every $k \leqslant n$. Coming up with such an observation is much harder than its proof; the intuition behind (5.12) is better understood with insight from martingale theory. Here we just prove this inequality directly. Note that

$$\begin{aligned} \mathbb{E}[S_n^2 \mathbf{1}_{A_k}] &= \mathbb{E}[(S_n - S_k + S_k)^2 \mathbf{1}_{A_k}] \\ &= \mathbb{E}[(S_n - S_k)^2 \mathbf{1}_{A_k}] + 2\mathbb{E}[(S_n - S_k) S_k \mathbf{1}_{A_k}] + \mathbb{E}[S_k^2 \mathbf{1}_{A_k}]. \end{aligned} \tag{5.13}$$

Since $X_1, \cdots, X_n$ are independent, one has

$$\begin{aligned} \mathbb{E}[(S_n - S_k) S_k \mathbf{1}_{A_k}] &= \mathbb{E}[(X_{k+1} + \cdots + X_n) S_k \mathbf{1}_{A_k}] \\ &= \mathbb{E}[X_{k+1} + \cdots + X_n] \cdot \mathbb{E}[S_k \mathbf{1}_{A_k}] \\ &= 0. \end{aligned}$$

In addition, the first term on the right hand side of (5.13) is non-negative. As a result, the inequality (5.12) holds.

It follows from (5.11) and (5.12) that

$$\mathbb{P}\big(A\big) \leqslant \frac{1}{\varepsilon^2} \sum_{k=1}^n \mathbb{E}[S_n^2 \mathbf{1}_{A_k}] = \frac{1}{\varepsilon^2} \mathbb{E}[S_n^2 \mathbf{1}_A] \leqslant \frac{1}{\varepsilon^2} \mathbb{E}[S_n^2].$$

On the other hand, due to the independence and mean zero assumptions one also has

$$\begin{aligned} \mathbb{E}[S_n^2] = \mathbb{E}[(X_1 + \cdots + X_n)^2] &= \sum_{k=1}^n \mathbb{E}[X_k^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] \\ &= \sum_{k=1}^n \mathbb{E}[X_k^2] + \sum_{i \neq j} \mathbb{E}[X_i] \mathbb{E}[X_j] = \sum_{k=1}^n \mathrm{Var}[X_k]. \end{aligned}$$

Therefore, the desired inequality (5.10) follows. □

Using Kolmogorov's maximal inequality and the probabilistic Cauchy criterion, the two-series theorem can be proved quite easily.

Proof of Theorem 5.4. We shall verify the criterion (5.9). Without loss of generality, one may assume that $\mathbb{E}[X_n] = 0$; for otherwise one can consider the sequence $X_n - \mathbb{E}[X_n]$ instead. In this case, according to the maximal inequality (5.10), for any $\varepsilon > 0$ one has

$$\mathbb{P}\big(\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2}\big(\mathrm{Var}[X_{n+1}] + \cdots + \mathrm{Var}[X_N]\big),$$

where $S_n \triangleq X_1 + \cdots + X_n$. It follows that

$$\lim_{N\to\infty} \mathbb{P}\big(\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2} \sum_{k=n+1}^{\infty} \mathrm{Var}[X_k].$$

Since $\sum_n \mathrm{Var}[X_n] < \infty$, one further has

$$\lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\big(\max_{n\leqslant l\leqslant N} |S_l - S_n| \geqslant \varepsilon\big) \leqslant \frac{1}{\varepsilon^2} \lim_{n\to\infty} \sum_{k=n+1}^{\infty} \mathrm{Var}[X_k] = 0.$$

Therefore, the property (5.9) holds and one concludes from Proposition 5.2 that the random series $\sum_n X_n$ is convergent a.s.

□

Example 5.5. From real analysis, the harmonic series $\sum_n 1/n$ diverges while the alternating harmonic series $\sum_n \frac{(-1)^{n-1}}{n}$ is convergent. It is interesting to investigate the convergence of the series if one puts a "random $\pm$-sign" in front of each $1/n$. A natural mathematical formulation is to consider the random series $\sum_n \frac{X_n}{n}$, where $\{X_n : n \geqslant 1\}$ is an i.i.d. symmetric Bernoulli sequence:

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}.$$

It follows easily from Kolmogorov's two-series theorem that $\sum_n \frac{X_n}{n}$ is convergent a.s.

Remark 5.4. There is a more general characterisation of the a.s. convergence of $\sum_n X_n$ which covers the case when the X_n's fail to have finite variance (*Kolmogorov's three-series theorem*). Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables. Then the random series $\sum_n X_n$ converges a.s. if and only if

there exists $C > 0$ (equivalently, for every $C > 0$), such that the following three properties hold true:

(i) $\sum_n \mathbb{P}(|X_n| > C) < \infty$;
(ii) $\sum_n \mathbb{E}[X_n \mathbf{1}_{\{|X_n| \leqslant C\}}]$ is convergent;
(iii) $\sum_n \mathrm{Var}[X_n \mathbf{1}_{\{|X_n| \leqslant C\}}] < \infty$.

The reader is referred to [Shi96] for a proof of this result.

5.4 The strong law of large numbers

In this section, we use tools from random series (in particular, the two-series theorem) to establish the strong LLN in the i.i.d. context. Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As usual, we denote $S_n \triangleq X_1 + \cdots + X_n$ as the partial sum sequence. The strong LLN is stated as follows.

Theorem 5.5. *(i) Suppose that $\mathbb{E}[|X_1|] < \infty$. Then one has*

$$\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}[X_1] \quad \text{a.s.} \tag{5.14}$$

(ii) Suppose that $\mathbb{E}[|X_1|] = \infty$. Then one has

$$\overline{\lim_{n\to\infty}} \frac{|S_n|}{n} = \infty \quad \text{a.s.} \tag{5.15}$$

We are going to take an analytic perspective to prove Theorem 5.5. Conceptually, the LLN is essentially related to the following type of convergence properties:

$$\frac{1}{a_n} \sum_{j=1}^{n} x_j \to 0 \tag{5.16}$$

where $0 < a_n \uparrow \infty$ and $x_n \in \mathbb{R}$. It is typical that $a_n = n$ and $x_n = X_n(\omega) - \mathbb{E}[X_n]$. The property (5.16) is often shown by means of the following analytic fact known as *Kronecker's lemma.*

Lemma 5.3. *Let $\{x_n\}$ be a real sequence and $\{a_n\}$ be a positive sequence increasing to infinity. Suppose that the series $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$ is convergent. Then the property (5.16) holds.*

The proof of this lemma is deferred to the appendix for not distracting the reader from the main probabilistic picture. An inspiration from this lemma is that one can try to prove the strong LLN through the convergence of suitable random series. We now give the precise argument for this.

Proof of Theorem 5.5. The argument is quite involved and we divide it into several steps. The idea is similar to the proof of the weak law (truncation). We first prove (5.14) which is the main case of interest.

Step one: truncation. We again introduce the following truncated sequence:

$$Y_n \triangleq \begin{cases} X_n, & \text{if } |X_n| \leqslant n; \\ 0, & \text{otherwise.} \end{cases}$$

In the same way as in Step One for the proof of the weak law, one concludes that

$$\mathbb{P}\big(X_n = Y_n \text{ for all sufficiently large } n\big) = 1.$$

As a consequence,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (X_j - Y_j) = 0 \quad \text{a.s.} \tag{5.17}$$

Step two: *convergence of* $\sum_n \frac{Y_n - \mathbb{E}[Y_n]}{n}$. Our next step is to apply Kolmogorov's two-series theorem to the random series $\sum_n Z_n$ where $Z_n \triangleq \frac{Y_n - \mathbb{E}[Y_n]}{n}$. Since Z_n has mean zero, one only needs to check that $\sum_n \mathrm{Var}[Z_n] < \infty$. To this end, let μ denote the law of X_1. By the definition of Y_n, one has

$$\mathrm{Var}[Z_n] \leqslant \frac{1}{n^2}\mathbb{E}[Y_n^2] = \frac{1}{n^2} \int_{\{|x|\leqslant n\}} x^2 \mu(dx).$$

It follows that

$$\begin{aligned} \sum_{n=1}^{\infty} \mathrm{Var}[Z_n] &\leqslant \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{\{|x|\leqslant n\}} x^2 \mu(dx) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{j=1}^{n} \int_{\{j-1<|x|\leqslant j\}} x^2 \mu(dx) \\ &= \sum_{j=1}^{\infty} \Big(\int_{\{j-1<|x|\leqslant j\}} x^2 \mu(dx) \Big) \sum_{n=j}^{\infty} \frac{1}{n^2} \quad \text{(exchange of summation)}. \end{aligned}$$

To analyse the last expression, one first observes that

$$\int_{\{j-1<|x|\leqslant j\}} x^2 \mu(dx) \leqslant j \cdot \int_{\{j-1<|x|\leqslant j\}} |x| \mu(dx).$$

In addition, one also has

$$\sum_{n=j}^{\infty} \frac{1}{n^2} \leqslant \sum_{n=j}^{\infty} \frac{1}{(n-1)n} = \sum_{n=j}^{\infty} \big(\frac{1}{n-1} - \frac{1}{n}\big) = \frac{1}{j-1} \leqslant \frac{2}{j}$$

for all $j \geqslant 2$. Note that the above inequality is also valid when $j = 1$ since

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < 2.$$

As a result, one obtains that

$$\begin{aligned} \sum_{n=1}^{\infty} \mathrm{Var}[Z_n] &\leqslant \sum_{j=1}^{\infty} j \cdot \Big(\int_{\{j-1<|x|\leqslant j\}} |x| \mu(dx) \Big) \cdot \frac{2}{j} \\ &= 2 \sum_{j=1}^{\infty} \int_{\{j-1<|x|\leqslant j\}} |x| \mu(dx) = 2\mathbb{E}[|X_1|] < \infty. \end{aligned}$$

According to Kolmogorov's two-series theorem, one concludes that the random series $\sum_n Z_n$ converges a.s. It then follows from Kronecker's lemma (cf. Lemma 5.3) with $a_n = n$ and $x_n = Y_n - \mathbb{E}[Y_n]$ that

$$\frac{1}{n} \sum_{j=1}^{n} (Y_j - \mathbb{E}[Y_j]) \to 0 \quad \text{a.s.} \tag{5.18}$$

as $n \to \infty$.

Step three: *convergence of* $\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[Y_j]$. By definition and the dominated convergence theorem, one has

$$\mathbb{E}[Y_n] = \int_{\{|x|\leqslant n\}} x \mu(dx) \to \mathbb{E}[X_1]. \tag{5.19}$$

It follows from (5.18) and the elementary fact (5.7) that

$$\frac{1}{n} \sum_{j=1}^{n} Y_j \to \mathbb{E}[X_1] \quad \text{a.s.}$$

The strong LLN (5.14) is now a consequence of (5.17) obtained in the first step.

Finally, we consider the divergent case. Suppose that $\mathbb{E}[|X_1|] = \infty$. A simple adaptation of the proof of Lemma 5.1 implies that

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_1| > An) = \infty$$

for any given $A > 0$. Since the X_n's are identically distributed, it follows that

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > An) = \infty.$$

By independence and the second Borel-Cantelli lemma, one thus finds that

$$\mathbb{P}(|X_n| > An \text{ for infinitely many } n) = 1.$$

Observe that

$$\{X_n > An\} \subseteq \big\{|S_n| > \frac{An}{2}\big\} \cup \big\{|S_{n-1}| > \frac{A(n-1)}{2}\big\}.$$

As a result, one has

$$\mathbb{P}\big(|S_n| > \frac{An}{2} \text{ for infinitely many } n\big) = 1.$$

Note that this is true for all $A > 0$. To conclude, for each $m \geqslant 1$ we define

$$\Omega_m \triangleq \big\{|S_n| > mn \text{ for infinitely many } n\big\}$$

and set $\Omega \triangleq \cap_{m=1}^{\infty} \Omega_m$. Then $\mathbb{P}(\Omega) = 1$ and one also has

$$\Omega \subseteq \big\{\overline{\lim_{n\to\infty}} \frac{|S_n|}{n} \geqslant m \ \forall m\big\} = \big\{\overline{\lim_{n\to\infty}} \frac{|S_n|}{n} = \infty\big\}.$$

Consequently, the divergence property (5.15) follows.

□

Remark 5.5. It was a remarkable result of N. Etemadi [Ete81] that Theorem 5.5 remains true when the assumption of total independence is weakened to pairwise independence.

5.5 Some applications of the law of large numbers

In this section, we present a few applications of (weak and strong) LLN.

5.5.1 Bernstein's polynomial approximation theorem

As the first example, we discuss an application of the weak LLN to polynomial approximations of continuous functions. From calculus, one can easily approximate a smooth function by polynomials through the Taylor expansion. If the function is only assumed to be *continuous*, how can one construct its polynomial approximation in some natural way? This question is of practical importance since polynomials are much easier to analyse both theoretically and computationally. Among various different approaches, the following elegant one was originally due to S. Bernstein.

Theorem 5.6. *Let $f(x)$ be a continuous function on $[0,1]$. For each $n \geqslant 1$, define the polynomial*

$$p_n(x) \triangleq \sum_{k=0}^{n} f\big(\frac{k}{n}\big) \binom{n}{k} x^k (1-x)^{n-k}, \quad x \in [0,1].$$

Then p_n converges to f uniformly on $[0,1]$ as $n \to \infty$.

Proof. Fix $x \in [0,1]$. Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence each following the Bernoulli distribution with parameter x, i.e.

$$\mathbb{P}(X_n = 1) = x, \ \mathbb{P}(X_n = 0) = 1 - x.$$

Define $S_n \triangleq X_1 + \cdots + X_n$. It is straight forward to check that $p_n(x) = \mathbb{E}\big[f\big(\frac{S_n}{n}\big)\big]$. According to the weak LLN, $S_n/n \to \mathbb{E}[X_1] = x$ in probability. In particular, $S_n/n \to x$ weakly. Since f is bounded and continuous, this already implies that

$$p_n(x) = \mathbb{E}\big[f\big(\frac{S_n}{n}\big)\big] \to \mathbb{E}[f(x)] = f(x)$$

for every given $x \in [0,1]$.

Proving uniform convergence requires extra effort. First of all, since f is uniformly continuous on $[0,1]$, for any given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$x, y \in [0,1], \ |x - y| \leqslant \delta \implies |f(x) - f(y)| \leqslant \varepsilon.$$

Exactly the same argument as the proof of Proposition 4.6 yields that

$$\big|p_n(x) - f(x)\big| \leqslant \varepsilon + 2\|f\|_\infty \cdot \mathbb{P}\big(\big|\frac{S_n}{n} - x\big| > \delta\big),$$

where $\|f\|_\infty \triangleq \sup_{x\in[0,1]} |f(x)|$. In addition, from Chebyshev's inequality one has

$$\mathbb{P}\big(\big|\frac{S_n}{n} - x\big| > \delta\big) \leqslant \frac{1}{\delta^2}\mathrm{Var}\big[\frac{S_n}{n}\big] = \frac{x(1-x)}{n\delta^2} \leqslant \frac{1}{4n\delta^2},$$

where we used the elementary inequality that $x(1-x) \leqslant 1/4$. Therefore, one arrives at

$$\big|p_n(x) - f(x)\big| \leqslant \varepsilon + \frac{\|f\|_\infty}{2n\delta^2}.$$

When n is large, the right hand side can be made smaller than 2ε uniformly in $x \in [0,1]$. This proves the desired uniform convergence. □

5.5.2 Borel's theorem on normal numbers

As the second example, we discuss an interesting application of the strong LLN to number theory. Recall that every real number $x \in (0,1)$ admits a decimal expansion

$$x = 0.x_1x_2\cdots x_n\cdots$$

where $x_n = 0, 1, \cdots, 9$. Except for countably many points in $(0,1)$ (precisely, points of the form $x = m/10^n$ with m, n being positive integers) whose expansions terminate in finitely many steps, such a representation is infinite and unique.

Given $x \in (0,1)$ and $0 \leqslant k \leqslant 9$, let $\nu_n^{(k)}(x)$ be the number of digits among the first n positions in the expansion of x that are equal to k. Apparently, $\nu_n^{(k)}(x)/n$ is the relative frequency of the digit k in the first n places. It is reasonable to expect that for "generic" points in $(0,1)$, this frequency should become close to $\frac{1}{10}$ when n gets large. Probabilistically, all the ten digits should occur equally likely in the decimal expansion of x if x is chosen in a suitably random manner.

Definition 5.5. A real number $x \in (0,1)$ is said to be *simply normal* (in base 10) if

$$\lim_{n\to\infty} \frac{\nu_n^{(k)}(x)}{n} = \frac{1}{10} \quad \text{for every } k = 0, 1, \cdots, 9.$$

The following result, which was originally due to Borel, asserts that almost every real number in $(0,1)$ is simply normal.

Theorem 5.7. *Let X be a point in $(0,1)$ chosen uniformly at random (i.e. $X \overset{d}{=} U(0,1)$). Then with probability one, X is a simply normal number.*

Proof. We express X in terms of its decimal expansion: $X = 0.X_1X_2\cdots X_n\cdots$. The crucial observation is that the sequence $\{X_n : n \geqslant 1\}$ of digits are i.i.d. with discrete uniform distribution

$$\mathbb{P}(X_n = k) = \frac{1}{10}, \quad k = 0, 1, \cdots, 9. \tag{5.20}$$

We first show that (5.20) holds. To understand the event $\{X_n = k\}$, let $A_1, \cdots, A_m$ ($m = 10^{n-1}$) be the partition of $(0, 1)$ into 10^{n-1} subintervals of equal length. For each j, we further divide A_j into 10 equal subintervals and let $B_{j,k}$ be the k-th one. It is not hard to see that

$$\{X_n = k\} = \bigcup_{j=1}^{m} \{X \in B_{j,k}\}.$$

Therefore, one has

$$\mathbb{P}(X_n = k) = \sum_{j=1}^{m} \mathbb{P}\big(X \in B_{j,k}\big) = 10^{n-1} \cdot \frac{1}{10^n} = \frac{1}{10}.$$

This shows that the X_n's are identically distributed. The geometric intuition behind the above argument is best seen when one considers binary instead of decimal expansions and draw a picture for the cases $n = 1, 2, 3$. In addition, for any $n \geqslant 1$ and $0 \leqslant k_1, \cdots k_n \leqslant 9$, the event

$$\{X_1 = k_1, X_2 = k_2, \cdots, X_n = k_n\}$$

simply means that X falls into one particular subinterval (depending on $k_1, \cdots, k_n$) in the partition of $(0, 1)$ into 10^n equal subintervals. In particular, one has

$$\mathbb{P}\big(X_1 = k_1, \cdots, X_n = k_n\big) = \frac{1}{10^n} = \mathbb{P}(X_1 = k_1)\cdots\mathbb{P}(X_n = k_n).$$

This gives the independence among $X_1, \cdots, X_n$.

To prove the theorem, let $0 \leqslant k \leqslant 9$ be given fixed and consider the i.i.d. Bernoulli sequence

$$Y_n = \begin{cases} 1, & X_n = k; \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that $\nu_n^{(k)}(X) = Y_1 + \cdots + Y_n$. According to the strong LLN (5.14), one has

$$\lim_{n\to\infty} \frac{\nu_n^{(k)}(X)}{n} = \mathbb{E}[Y_1] = \mathbb{P}(X_1 = k) = \frac{1}{10} \quad \text{a.s.} \tag{5.21}$$

In other words,

$$\mathbb{P}(\Omega_k) = 1 \text{ where } \Omega_k \triangleq \big\{\frac{\nu_n^{(k)}(X)}{n} \to \frac{1}{10}\big\}.$$

The conclusion of the theorem follows by observing that

$$\mathbb{P}\big(\frac{\nu_n^{(k)}(X)}{n} \to \frac{1}{10} \text{ for every } k\big) = \mathbb{P}\big(\bigcap_{k=0}^{9} \Omega_k\big) = 1.$$

□

Remark 5.6. Although Theorem 5.7 asserts that almost every real number in $(0, 1)$ is simply normal, it does not provide an explicit example of a single one! In fact, one can easily come up with numbers that are not simply normal e.g. $x = 1/3 = 0.333\cdots$. Explicitly constructing simply normal numbers appears to be a bit more challenging. It is typical that probabilistic methods provide simpler ways of proving existence theorems but they often have a non-constructive nature as a price to pay. Another famous example is the existence of continuous but nowhere differentiable functions. While it is quite non-trivial to explicitly write down one such function (e.g. the *Weierstrass function*), one can use the notion of Brownian motion to produce a rich class of examples (with probability one, Brownian trajectories are continuous but nowhere differentiable!).

5.5.3 Poincaré's lemma on Gaussian measures

As the third example, we prove a striking fact that the standard Gaussian measure on $\mathbb{R}^n$ can be realised through projections of spherical measures from "infinitely" high dimensions. Such a result, which was originally due to H. Poincaré, plays a fundamental role in Gaussian analysis. For instance, based on such a property one can derive isoperimetric inequalities for the Gaussian measure from classical isometric inequalities on spheres.

We first introduce some basic concepts. The standard Gaussian measure (law of standard Gaussian vector) on $\mathbb{R}^n$ is defined by

$$\gamma_n(d\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{-|\mathbf{x}|^2/2} d\mathbf{x},$$

where $|\cdot|$ denotes the Euclidean norm and $d\mathbf{x}$ is the Lebesgue measure on $\mathbb{R}^n$. In what follows, n is fixed and $N > n$ is varying (we shall send $N \to \infty$ eventually).

The map $\Pi_{N+1,n} : \mathbb{R}^{N+1} \to \mathbb{R}^n$ denotes the project onto the first n coordinates. The N-sphere in $\mathbb{R}^{N+1}$ with radius ρ is denoted as

$$S_\rho^N \triangleq \{\mathbf{x} = (x_1, \cdots, x_{N+1})^T : \sqrt{x_1^2 + \cdots + x_{N+1}^2} = \rho\} \subseteq \mathbb{R}^{N+1},$$

where $(\cdot)^T$ means matrix transpose (we adopt the convention that elements in $\mathbb{R}^{N+1}$ are column vectors). The normalised surface measure on S_ρ^N is denoted as σ_ρ^N. In other words, for any Borel measurable subset $B \subseteq S_\rho^N$, one has

$$\sigma_\rho^N(B) = \frac{\text{Area of } B}{\text{Total area of } S_\rho^N}.$$

It is classical that the total area of S_ρ^N is equal to $\frac{2\pi^{(N+1)/2}}{\Gamma((N+1)/2)}\rho^N$. Note that σ_ρ^N is a probability measure and it is rotationally invariant in the sense that $\sigma_\rho^N(OB) = \sigma_\rho^N(B)$ for any $B \in \mathcal{B}(S_\rho^N)$ and $(N+1) \times (N+1)$ orthogonal matrix O ($OB \triangleq \{O \cdot \mathbf{x} : \mathbf{x} \in B\}$). Indeed, such a property uniquely characterises σ_ρ^N (cf. [Lan93]).

Proposition 5.3. *σ_ρ^N is the unique rotationally invariant probability measure on S_ρ^N.*

Poincaré's lemma for the Gaussian measure γ_n is stated as follows.

Theorem 5.8. *For every Borel set A in $\mathbb{R}^n$, one has*

$$\lim_{N\to\infty} \sigma_{\sqrt{N}}^N \left(\Pi_{N+1,n}^{-1}(A) \cap S_{\sqrt{N}}^N\right) = \gamma_n(A). \tag{5.22}$$

Proof. If one is satisfied with weak convergence, the argument is conceptually simpler by using the strong LLN. To see this, let $\{Z_1, \cdots, Z_{N+1}\}$ be an i.i.d. family of standard normals and set

$$R_{N+1} \triangleq \sqrt{Z_1^2 + \cdots + Z_{N+1}^2}.$$

A key property of the standard Gaussian measure is its rotational invariance. In other words, letting $\mathbf{Z} \triangleq (Z_1, \cdots, Z_{N+1})^T$ one has $O \cdot \mathbf{Z} \overset{\text{law}}{=} \mathbf{Z}$ for any $(N+1) \times (N+1)$ orthogonal matrix O (why?). In particular, the law of the normalised random vector

$$\frac{\sqrt{N}}{R_{N+1}}(Z_1, \cdots, Z_{N+1}) \in S_{\sqrt{N}}^N$$

is rotationally invariant on the sphere $S^N_{\sqrt{N}}$, thus being equal to $\sigma^N_{\sqrt{N}}$ as a consequence of Proposition 5.3. It follows that the law of the random vector

$$\frac{\sqrt{N}}{R_{N+1}}(Z_1, \cdots, Z_n) \in \mathbb{R}^n$$

is $\sigma^N_{\sqrt{N}}\big(\Pi^{-1}_{N+1,n}(\cdot) \cap S^N_{\sqrt{N}}\big)$. On the other hand, according to the strong LLN,

$$\frac{\sqrt{N}}{R_{N+1}}(Z_1, \cdots, Z_n) \to (Z_1, \cdots, Z_n) \quad \text{a.s. as } N \to \infty.$$

Since a.s. convergence implies weak convergence, one thus concludes that

$$\sigma^N_{\sqrt{N}}\big(\Pi^{-1}_{N+1,n}(\cdot) \cap S^N_{\sqrt{N}}\big) \to \gamma_n \quad \text{weakly as } N \to \infty.$$

Proving convergence for every $A \in \mathcal{B}(\mathbb{R}^n)$ requires some extra work. Let $\{Z_k : k \geqslant 1\}$ be an i.i.d. sequence of standard normals. For each N we set $R_N \triangleq \sqrt{Z_1^2 + \cdots + Z_N^2}$. We have already seen that

$$\begin{aligned}\sigma^N_{\sqrt{N}}\big(\Pi^{-1}_{N+1,n}(A) \cap S^N_{\sqrt{N}}\big) &= \mathbb{P}\big(\frac{\sqrt{N}}{R_{N+1}}(Z_1, \cdots, Z_n) \in A\big) \\ &= \mathbb{P}\Big(\sqrt{\frac{NR_n^2}{R_{N+1}^2}} \cdot \frac{(Z_1, \cdots, Z_n)}{R_n} \in A\Big) \qquad (5.23)\end{aligned}$$

□

To proceed further, the key observation is that R_n^2/R_{N+1}^2 is independent of $(Z_1, \cdots, Z_n)/R_n$. To see this, it suffices to show that $(Z_1, \cdots, Z_n)/R_n$ is independent of R_n (why?). Recall from the rotational invariance of γ_n that $O \cdot \mathbf{Z} \overset{\text{law}}{=} \mathbf{Z}$, where $\mathbf{Z} \triangleq (Z_1, \cdots, Z_n)^T$. It follows that

$$\frac{O \cdot \mathbf{Z}}{r}\big|_{|\mathbf{Z}|=r} = \frac{O \cdot \mathbf{Z}}{|O \cdot \mathbf{Z}|}\big|_{|O\cdot\mathbf{Z}|=r} \overset{\text{law}}{=} \frac{\mathbf{Z}}{|\mathbf{Z}|}\big|_{|\mathbf{Z}|=r} = \frac{\mathbf{Z}}{r}\big|_{|\mathbf{Z}|=r},$$

where $|\cdot|$ denotes the Euclidean norm. In particular, this shows that conditional on $|\mathbf{Z}| = r$, the law of $\mathbf{Z}/r$ on the unit sphere S_1^{n-1} is rotational invariant. As a result, this conditional distribution must be the normalised surface measure on S_1^{n-1}. But the unconditional law of $\mathbf{Z}/|\mathbf{Z}|$ is also the normalised surface measure. As a consequence, $\mathbf{Z}/|\mathbf{Z}|$ and $|\mathbf{Z}|$ are independent.

Note that R_n^2/R_{N+1}^2 is β-distributed with parameters $(n/2,(N+1-n)/2)$. It follows from (5.23) and the above independence property that

$$\begin{aligned}
&\sigma_{\sqrt{N}}^N\big(\Pi_{N+1,n}^{-1}(A)\cap S_{\sqrt{N}}^N\big)\\
&=\frac{\Gamma(n/2)}{2\pi^{n/2}}\beta\big(\frac{n}{2},\frac{N+1-n}{2}\big)^{-1}\int_{S_1^{n-1}}\int_0^1 \mathbf{1}_A(\sqrt{Nt}x)t^{\frac{n}{2}-1}(1-t)^{\frac{N+1-n}{2}-1}dtd\sigma_1^{n-1}(x)\\
&=\frac{\Gamma(\frac{N+1}{2})}{\pi^{n/2}N^{n/2}\Gamma(\frac{N+1-n}{2})}\int_{S_1^{n-1}}\int_0^{\sqrt{N}}\mathbf{1}_A(ux)u^{n-1}\big(1-\frac{u^2}{N}\big)^{\frac{N+1-n}{2}-1}dud\sigma_1^{n-1}(x),
\end{aligned}$$

where we made the change of variable $u=\sqrt{Nt}$ to reach the last line. By applying the dominated convergence theorem on the inner integral, after sending $N\to\infty$ the right hand side converges to

$$\frac{1}{(2\pi)^{n/2}}\int_0^\infty\int_{S_1^{n-1}}\mathbf{1}_A(ux)u^{n-1}e^{-u^2/2}d\sigma_1^{n-1}(x)du,$$

which is exactly $\gamma_n(A)$ computed under polar coordinates.

5.6 Introduction to the large deviation principle

Let $\{X_n : n\geqslant 1\}$ be an i.i.d. sequence of random variables with finite mean. We define the sample average sequence

$$\bar{S}_n\triangleq\frac{1}{n}(X_1+\cdots+X_n),\quad n\geqslant 1.$$

By the strong LLN, one knows that $\bar{S}_n\to\bar{x}$ a.s. However, this result does not contain quantitative information about the underlying convergence. In this section, we discuss a particularly typical phenomenon associated with a LLN: the *large deviation principle*. As we will see, the central limit theorem (cf. Chapter 7 below) and the large deviation principle quantify the strong LLN at different levels. The central limit theorem indicates that the error $\bar{S}_n-\bar{x}$ from the strong LLN, at the level of random variables, is roughly of order $1/\sqrt{n}$. On the other hand, the large deviation principle is more like a concentration of measure property; it suggests that the law of $\bar{S}_n$ concentrates at the Dirac delta measure $\delta_{\bar{x}}$ exponentially fast.

5.6.1 Motivation and formulation of Cramér's theorem

To motivate the underlying phenomenon, let B be an arbitrary subset of $\mathbb{R}$ that has positive distance (say ε) away from $\bar{x}$ (in particular, $\bar{x} \notin B$). In other words, $(\bar{x} - \varepsilon, \bar{x} + \varepsilon) \subseteq B^c$ which further implies that

$$\mathbb{P}(\bar{S}_n \in B) \leqslant \mathbb{P}(|\bar{S}_n - \bar{x}| \geqslant \varepsilon).$$

Since a.s. convergence implies convergence in probability, it follows that

$$\lim_{n\to\infty} \mathbb{P}(\bar{S}_n \in B) = 0.$$

The *large deviation principle* quantifies the precise decay rate of the probability $\mathbb{P}(\bar{S}_n \in B)$ as $n \to \infty$. It turns out that this probability decays to zero exponentially fast:

$$\mathbb{P}\big(\bar{S}_n \in B\big) \approx e^{-nI_B} \quad \text{as } n \to \infty. \tag{5.24}$$

Here I_B is an exponent that depends on B, which can be determined explicitly from the distribution μ. Heuristically, (5.24) describes a *concentration of measure* phenomenon. It suggests that masses of $\bar{S}_n$ over any set that "deviates" from $\bar{x}$ vanish exponentially fast. As a result, masses of $\bar{S}_n$ concentrate around $\bar{x}$ with exponential speed as $n \to \infty$.

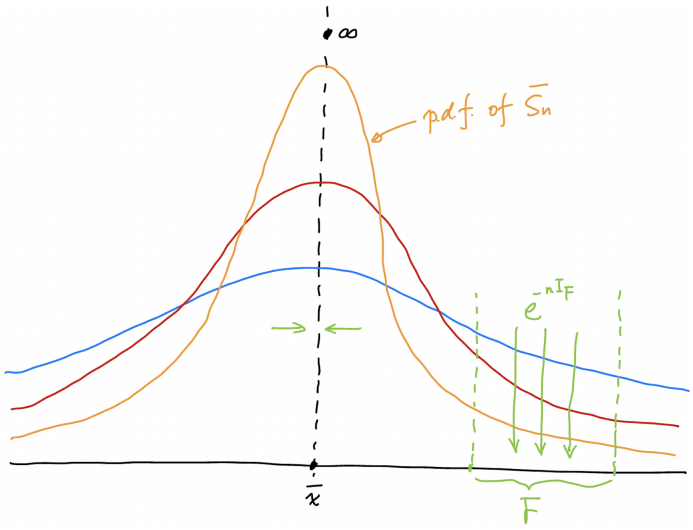


Figure 5.1: Concentration of measure in LDP

The statement (5.24) is rather vague at this stage (and in fact, it is not exactly true!). Before introducing the precise formulaltion of the result, let us take a little bit extra effort to motivate the expression of the exponent I_B.

For simplicity, we consider $B = [x, \infty)$ where $x > \bar{x}$. By using Markov's inequality, for any $\lambda \geqslant 0$ one has

$$\mathbb{P}\big(\bar{S}_n \in B\big) \leqslant \mathbb{P}\big(e^{n\lambda \bar{S}_n} \geqslant e^{n\lambda x}\big) \leqslant e^{-n\lambda x}\mathbb{E}[e^{n\lambda \bar{S}_n}] = e^{-n\lambda x}\mathbb{E}[e^{\lambda X_1}]^n. \tag{5.25}$$

Recall that the *cumulant generating function* of X_1 is defined by

$$\Lambda(\lambda) \triangleq \log \mathbb{E}[e^{\lambda X_1}], \quad \lambda \in \mathbb{R}. \tag{5.26}$$

By using $\Lambda(\lambda)$, one can rewrite (5.25) as

$$\mathbb{P}\big(\bar{S}_n \in B\big) \leqslant e^{-n(\lambda x - \Lambda(\lambda))}. \tag{5.27}$$

Since (5.27) is true for all $\lambda \geqslant 0$, by optimising it over λ one finds that

$$\mathbb{P}\big(\bar{S}_n \in B\big) \leqslant \exp\big(-n \sup_{\lambda \geqslant 0}\{\lambda x - \Lambda(\lambda)\}\big). \tag{5.28}$$

Now it becomes natural to introduce the following function:

$$\Lambda^*(x) \triangleq \sup_{\lambda \geqslant 0}\{\lambda x - \Lambda(\lambda)\}, \quad x > \bar{x}. \tag{5.29}$$

It will be shown in Lemma 5.4 below that $\Lambda^*(x)$ (as a function of x) is increasing on $(\bar{x}, \infty)$. As a consequence, the inequality (5.28) can also be expressed as

$$\mathbb{P}\big(\bar{S}_n \in B\big) \leqslant \exp\big(-n \inf_{y \in B} \Lambda^*(y)\big) \iff \frac{1}{n}\log \mathbb{P}\big(\bar{S}_n \in B\big) \leqslant -\inf_{y \in B} \Lambda^*(B). \tag{5.30}$$

It turns out that as $n \to \infty$, the above estimate becomes asymptotically sharp and the exponent I_B appearing in (5.24) is given by the right hand side of (5.30)

We now proceed to introduce the precise mathematical formulation of the large deviation principle. From the above discussion, the function $\Lambda^*(x)$ shall play a central role in this problem and we first define it in a more careful way. Recall that $\Lambda(\lambda)$ is the cumulant generating function of X_1 defined by (5.26). Note that $\Lambda(0) = 0$ and $\Lambda(\lambda) \in (-\infty, \infty]$ for all $\lambda \in \mathbb{R}$ (why?).

Definition 5.6. The *Legendre transform* of $\Lambda(\lambda)$ is the function defined by

$$\Lambda^*(x) \triangleq \sup_{\lambda \in \mathbb{R}}\{\lambda x - \Lambda(\lambda)\}, \quad x \in \mathbb{R}.$$

By definition, $\Lambda^*(x)$ measures the maximal excess of the straight line $\lambda \mapsto x \cdot \lambda$ over the function $\Lambda(\lambda)$. Since $\Lambda(0) = 0$, it is clear that $\Lambda^* \geqslant 0$. It is possible that $\Lambda^*(x) = \infty$. Some basic properties of $\Lambda^*(x)$ are given in Lemma 5.4 below (see also Figure 5.2 for the geometric intuition).

Remark 5.7. In Definition 5.6, the supremum is taken over $\mathbb{R}$ while it is over $\lambda \geqslant 0$ in (5.29). These two representations are identical when $x \in (\bar{x}, \infty)$ (cf. (5.34) below).

Example 5.6. Suppose that $X_1 \sim N(0, \sigma^2)$. Its cumulant generating function is given by

$$\Lambda(\lambda) = \log e^{\lambda^2\sigma^2/2} = \frac{1}{2}\sigma^2\lambda^2.$$

It follows that

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}} \left\{\lambda x - \frac{1}{2}\sigma^2\lambda^2\right\} = \frac{x^2}{2\sigma^2}.$$

Note that $\bar{x} = 0$ is the unique global minimum of Λ^* and $I_B \in (0, \infty)$ for all B having positive distance to 0.

Example 5.7. Suppose that X_1 follows the symmetric Bernoulli distribution:

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}.$$

Then one has

$$\Lambda(\lambda) = \log\left(\frac{e^\lambda + e^{-\lambda}}{2}\right).$$

For each $x \in \mathbb{R}$, define

$$\varphi_x(\lambda) \triangleq \lambda x - \Lambda(\lambda) = \lambda x - \log\left(\frac{e^\lambda + e^{-\lambda}}{2}\right), \quad \lambda \in \mathbb{R}.$$

Simple calculus shows that when $x \in (-1, 1)$, $\varphi_x(\lambda)$ attains its maximum at $\lambda_x \triangleq \frac{1}{2}\log\left(\frac{1+x}{1-x}\right)$ with value

$$\Lambda^*(x) = \varphi_x(\lambda_x) = \frac{1+x}{2}\log(1+x) + \frac{1-x}{2}\log(1-x).$$

If $x = 1$ (respectively, $x = -1$), the supremum $\Lambda^*(x) = \log 2$ is asymptotically attained at $\lambda \to \infty$ (respectively, $\lambda \to -\infty$). If $x \notin [-1, 1]$, one has $\Lambda^*(x) = \infty$. To summarise,

$$\Lambda^*(x) = \begin{cases} \frac{1+x}{2}\log(1+x) + \frac{1-x}{2}\log(1-x), & -1 \leqslant x \leqslant 1; \\ \infty, & \text{otherwise.} \end{cases}$$

This example also shows that $\Lambda^*(x)$ needs not be finite for all x.

The large deviation principle (LDP) for the sequence $\{\bar{S}_n : n \geqslant 1\}$, which is also known as *Cramér's theorem* on $\mathbb{R}$, is stated as follows. It contains both an upper and a lower estimate.

Theorem 5.9. *Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence and define $\bar{S}_n \triangleq \frac{X_1+\cdots+X_n}{n}$.*
(i) [Upper bound] For any closed subset $F \subseteq \mathbb{R}$, one has

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in F\big) \leqslant -\inf_{x\in F} \Lambda^*(x). \tag{5.31}$$

(ii) [Lower bound] for any open subset $G \subseteq \mathbb{R}$, one has

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in G\big) \geqslant -\inf_{x\in G} \Lambda^*(x). \tag{5.32}$$

Heuristically, Cramér's theorem suggests that for "suitably reasonable" subsets B, the probability $\mathbb{P}(\bar{S}_n \in B)$ decays like e^{-nI_B} as $n \to \infty$ with rate exponent

$$I_B = \inf_{x\in B} \Lambda^*(x).$$

Indeed, for any Borel subset $B \in \mathcal{B}(\mathbb{R})$, according to (5.31) and (5.32) one has

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in B\big) \leqslant \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in \bar{B}\big) \leqslant -\inf_{x\in \bar{B}} \Lambda^*(x)$$

and

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in B\big) \geqslant \underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in \mathring{B}\big) \geqslant -\inf_{x\in \mathring{B}} \Lambda^*(x)$$

respectively. Here $\bar{B}$ denotes the closure of B and $\mathring{B}$ is its interior. The above two inequalities hold simply because $\mathring{B} \subseteq B \subseteq \bar{B}$. As a consequence, all possible limit points of the sequence $\{n^{-1} \log \mathbb{P}(\bar{S}_n \in B)\}$ are contained in the interval

$$\big[-\inf_{x\in \mathring{B}} \Lambda^*(x), -\inf_{x\in \bar{B}} \Lambda^*(x)\big].$$

If the subset $B \in \mathcal{B}(\mathbb{R})$ satisfies

$$\inf_{x\in \bar{B}} \Lambda^*(x) = \inf_{x\in \mathring{B}} \Lambda^*(x),$$

then one has

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in B\big) = -\inf_{x\in B} \Lambda^*(x) \tag{5.33}$$

On the other hand, it should be pointed out that the sequence $\{n^{-1}\log\mathbb{P}(\bar{S}_n \in B)\}$ itself may fail to converge and (5.33) may not hold in general. For instance, consider $X_1 \sim N(0,1)$ and $B = \mathbb{Q}$. The left hand side is trivially equal to $-\infty$ since $\mathbb{P}(\bar{S}_n \in \mathbb{Q}) = 0$, while the right hand side is zero as seen from Example 5.6. Nonetheless, (5.33) is true for $B = [y, \infty)$ (cf. Corollary 5.2 below).

Remark 5.8. Later on it will be clear that $\Lambda^*(\bar{x}) = 0$. In particular, $I_B = 0$ if $\bar{x} \in B$, in which case the LDP does not contain much useful information. The main interesting regime is when B has a positive distance to $\bar{x}$, in which case one typically observes an exponential decay with a meaningful exponent $I_B \in (0, \infty)$ (not always true though!). Therefore, the theorem really measures the (un)likelihood of "large deviations" of $\bar{S}_n$ from its mean $\bar{x}$. The function $\Lambda^*(x)$ is commonly known as the *rate function* for the LDP of $\{\bar{S}_n\}$.

Example 5.8. Cramér's theorem is easily appreciated in the simple example of Gaussian distributions. Let $\{X_n\}$ be i.i.d. standard normals. Then $\bar{S}_n \overset{d}{=} N(0, 1/n)$. We also recall from Example 5.6 that $\Lambda^*(x) = x^2/2$. For simplicity, let us take $B = [a, b]$ with $a > 0$ (whether one includes any of the endpoints is of no significance). Then

$$\mathbb{P}(\bar{S}_n \in B) = \int_a^b \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-nx^2/2} dx \leqslant \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-na^2/2}(b-a),$$

from which it follows that

$$\overline{\lim_{n\to\infty}} \frac{1}{n}\log\mathbb{P}(\bar{S}_n \in B) \leqslant -\frac{a^2}{2} = -\inf_{x\in B}\Lambda^*(x).$$

On the other hand, for any small $\varepsilon > 0$ one also has

$$\mathbb{P}(\bar{S}_n \in B) \geqslant \int_a^{a+\varepsilon} \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-nx^2/2} dx \geqslant \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-n(a+\varepsilon)^2/2}\varepsilon,$$

which implies that

$$\underline{\lim_{n\to\infty}} \frac{1}{n}\log\mathbb{P}(\bar{S}_n \in B) \geqslant -\frac{(a+\varepsilon)^2}{2}.$$

Since ε is arbitrary, one easily obtains the matching lower bound. The density function

$$\rho_n(x) = \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-nx^2/2}$$

for $\bar{S}_n$ in this example also illustrates the aforementioned exponential concentration phenomenon around $\bar{x} = 0$ (cf. Figure 5.1).

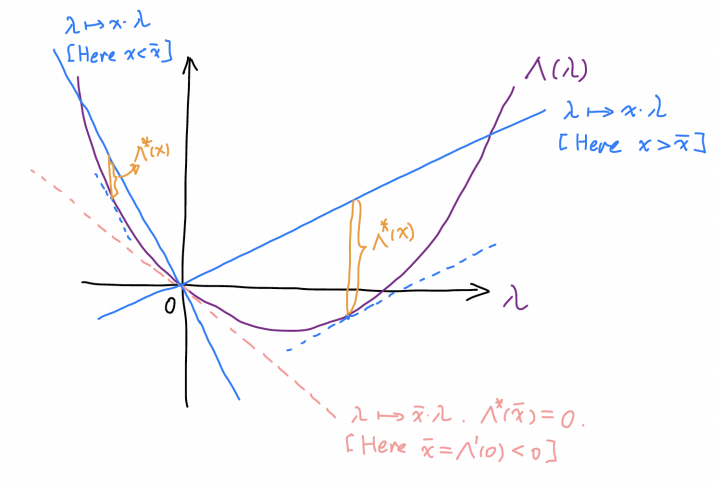


Figure 5.2: Geometric interpretation of $\Lambda^*(x)$

5.6.2 Basic properties of the Legendre transform

Before proving Theorem 5.9, we summarise a few basic properties of the function Λ^* in the lemma below. We only discuss the ones that are directly related to the proof of the theorem. Examples 5.6, 5.7 and Figure 5.2 should provide some geometric intuition behind the function Λ^*.

Lemma 5.4. *Suppose that $\mathbb{E}[X_1]$ has finite mean. Then the following properties of the Legendre transform Λ^* hold true.*

(i) $\Lambda^*(\bar{x}) = \inf_{x\in\mathbb{R}} \Lambda^*(x) = 0$.

(ii) For any $x > \bar{x}$, one has

$$\Lambda^*(x) = \sup_{\lambda\geqslant 0}\{\lambda x - \Lambda(\lambda)\}. \tag{5.34}$$

The function $\Lambda^*(x)$ is increasing on $(\bar{x}, \infty)$.

(iii) For any $x < \bar{x}$, one has

$$\Lambda^*(x) = \sup_{\lambda\leqslant 0}\{\lambda x - \Lambda(\lambda)\}.$$

The function $\Lambda^*(x)$ is decreasing on $(-\infty, \bar{x})$.

Proof. (i) We have seen that $\Lambda^* \geqslant 0$. In addition, since $\log x$ is a concave function, by Jensen's inequality one has

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X_1}] \geqslant \mathbb{E}\big[\log e^{\lambda X_1}\big] = \lambda \bar{x} \quad \forall \lambda \in \mathbb{R}. \tag{5.35}$$

As a result,

$$\lambda \bar{x} - \Lambda(\lambda) \leqslant 0 \ \forall \lambda \in \mathbb{R} \implies \Lambda^*(\bar{x}) = \sup_{\lambda \in \mathbb{R}} \{\lambda \bar{x} - \Lambda(\lambda)\} \leqslant 0.$$

Therefore, $\Lambda^*(\bar{x}) = 0$.

(ii) Let $x > \bar{x}$. According to (5.35), for any $\lambda < 0$ one has

$$\lambda x - \Lambda(\lambda) \leqslant \lambda \bar{x} - \Lambda(\lambda) \leqslant 0 = 0 \cdot x - \Lambda(0).$$

In particular, the values of $\lambda x - \Lambda(\lambda)$ for all $\lambda < 0$ cannot exceed the supremum over $\lambda \geqslant 0$. As a result, one has the representation (5.34). Under such a representation (for $x > \bar{x}$), since the function $x \mapsto \lambda x - \Lambda(\lambda)$ is increasing for each $\lambda \geqslant 0$, it follows that $\Lambda^*(x)$ is increasing on $(\bar{x}, \infty)$.

(iii) The argument is parallel to Part (ii) and is thus omitted. □

In what follows, we develop the proof of Cramér's theorem. For simplicity, we assume exclusively that X_1 *has finite mean*. We remark that the theorem remains true even if $\mathbb{E}[X_1]$ does not exist.

5.6.3 Proof of Cramér's theorem: upper bound

We have already obtained the following lemma in (5.30) before. This is the key ingredient for proving the LDP upper bound (5.31).

Lemma 5.5. *For any $x > \bar{x}$, one has*

$$\mathbb{P}\big(\bar{S}_n \geqslant x\big) \leqslant e^{-n\Lambda^*(x)}. \tag{5.36}$$

Similarly, for any $x < \bar{x}$, one has

$$\mathbb{P}(\bar{S}_n \leqslant x) \leqslant e^{-n\Lambda^*(x)}.$$

Proof. In view of (5.30), one only needs to apply (5.34) which holds since $x > \bar{x}$. The other case is treated in a similar way. □

Proof of LDP upper bound (5.31). Let $F \subseteq \mathbb{R}$ be a closed subset and define

$$I_F \triangleq \inf_{x \in F} \Lambda^*(x).$$

One may assume that $\bar{x} \notin F$; for otherwise $I_F = 0$ (since $\Lambda^*(\bar{x}) = 0$ by Lemma 5.4) and the result is trivial. Define

$$x^- \triangleq \sup\{r < \bar{x} : r \in F\}, \ x^+ \triangleq \inf\{r > \bar{x} : r \in F\}.$$

Note that $x^- < \bar{x} < x^+$ and at lease one of $x^\pm$ is finite (since $F \neq \emptyset$). In addition, whenever $x^\pm$ is finite one has $x^\pm \in F$ (since F is closed). As a result,

$$F \subseteq (-\infty, x^-] \cup [x^+, \infty),$$

which implies by Lemma 5.5 that

$$\mathbb{P}\big(\bar{S}_n \in F\big) \leqslant \mathbb{P}\big(\bar{S}_n \leqslant x^-\big) + \mathbb{P}\big(\bar{S}_n \geqslant x^+\big) \leqslant e^{-n\Lambda^*(x^-)} + e^{-n\Lambda^*(x^+)} \leqslant 2e^{-nI_F}.$$

The desired estimate (5.31) follows from this inequality.

□

5.6.4 Proof of Cramér's theorem: lower bound

Next, we turn to the proof of the LDP lower bound (5.32). We first make the following key observation.

Lemma 5.6. *In order to establish (5.32), it is sufficient to show that for any $\delta > 0$ and any marginal distribution μ (the distribution of X_1), one has*

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in (-\delta, \delta)\big) \geqslant \inf_{\lambda \in \mathbb{R}} \Lambda_\mu(\lambda), \tag{5.37}$$

where $\Lambda_\mu(\lambda)$ denotes the cumulant generating function of μ.

Proof. Suppose that (5.37) is true for any marginal distribution μ. Note that the right hand side of (5.37) is also equal to $-\Lambda^*(0)$. As a result, given any $x \in \mathbb{R}$ one has

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (x-\delta, x+\delta)\big) = \varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^Y \in (-\delta, \delta)\big) \geqslant -\Lambda_Y^*(0). \tag{5.38}$$

Here $\bar{S}_n^X$ refers to the sample average associated with the i.i.d. sequence $X = \{X_n\}$ whose marginal distribution is the given fixed μ in the LDP. $\bar{S}_n^Y$ denotes the sample

average of the sequence $Y \triangleq \{Y_n \triangleq X_n - x\}$ and Λ_Y^* is defined with respect to the law of Y_1. It is easy to figure out the relation between Λ_X^* and Λ_Y^*; indeed, since

$$\Lambda_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y_1}] = \log \mathbb{E}[e^{\lambda(X_1 - x)}] = \Lambda_X(\lambda) - \lambda x,$$

one has

$$\Lambda_Y^*(y) = \sup_{\lambda \in \mathbb{R}} (\lambda y - \Lambda_Y(\lambda)) = \sup_{\lambda \in \mathbb{R}} (\lambda(x+y) - \Lambda_X(\lambda)) = \Lambda_X^*(y+x).$$

As a result, the inequality (5.38) becomes

$$\underline{\lim}_{n \to \infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (x - \delta, x + \delta)\big) \geqslant -\Lambda_X^*(x). \tag{5.39}$$

To establish the LDP lower bound (5.32), let $G \subseteq \mathbb{R}$ be a given open subset. For any $x \in G$, choose $\delta > 0$ such that $(x - \delta, x + \delta) \subseteq G$. It follows from (5.39) that

$$\underline{\lim}_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\bar{S}_n^X \in G) \geqslant \underline{\lim}_{n \to \infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (x - \delta, x + \delta)\big) \geqslant -\Lambda_X^*(x).$$

Since this is true for all $x \in G$, the desired estimate (5.32) follows. □

To complete the proof, it remains to establish the key estimate (5.37). The argument for this part contains an essential technique of *change of measure*. Such a technique has deep extensions and rich applications in probability theory. Let μ denote the law of X_1 and $\Lambda(\lambda)$ is its cumulant generating function. We divide the discussion into three cases.

Case I: μ has positive measure on both $(-\infty, 0)$ and $(0, \infty)$, and it is supported in a bounded interval, say $[-M, M]$.

Equivalently, it is assumed that

$$\mathbb{P}(X_1 > 0) > 0, \ \mathbb{P}(X_1 < 0) > 0, \ |X_1| \leqslant M \text{ a.s.}$$

The heart of the proof is contained in this case. The main benefit here is that the function $\Lambda(\lambda)$ is continuously differentiable on $\mathbb{R}$, and one also has

$$\lim_{\lambda \to \pm\infty} \Lambda(\lambda) = \infty. \quad \text{(why?)} \tag{5.40}$$

As a consequence, there exists $\eta \in \mathbb{R}$ such that $\Lambda'(\eta) = 0$.

To proceed further, the key idea is to change μ to a new probability measure $\tilde{\mu}$ that has mean zero. To motivate its construction, one first notes that

$$0 = \Lambda'(\eta) = \frac{M'(\eta)}{M(\eta)} = \int_{\mathbb{R}} x e^{\eta x - \Lambda(\eta)} \mu(dx). \tag{5.41}$$

As a result, one defines $\tilde{\mu}$ by

$$\tilde{\mu}(dx) \triangleq e^{\eta x - \Lambda(\eta)} \mu(dx). \quad (\text{Equivalently, } \tilde{\mu}(A) = \int_A e^{\eta x - \Lambda(\eta)} \mu(dx) \ \forall A \in \mathcal{B}(\mathbb{R}).)$$

It follows from (5.41) that $\int_{\mathbb{R}} x \tilde{\mu}(dx) = 0$.

Recall that $X = \{X_n\}$ is an i.i.d. sequence with marginal distribution μ. Let $\tilde{X} = \{\tilde{X}_n\}$ denote an i.i.d. sequence with marginal distribution $\tilde{\mu}$. We use $\bar{S}_n^X$, $\bar{S}_n^{\tilde{X}}$ to denote their sample average sequences respectively. For any $\varepsilon \in (0, \delta)$ (recall that δ is given fixed in (5.37)), one has

$$\begin{aligned} \mathbb{P}\big(\bar{S}_n^X \in (-\varepsilon, \varepsilon)\big) &= \int_{\left\{(x_1, \cdots, x_n): \left|\frac{x_1 + \cdots + x_n}{n}\right| < \varepsilon\right\}} \mu(dx_1) \cdots \mu(dx_n) \\ &= \int_{\left\{(x_1, \cdots, x_n): \left|\frac{x_1 + \cdots + x_n}{n}\right| < \varepsilon\right\}} e^{-\eta(x_1 + \cdots + x_n) + n\Lambda(\eta)} \tilde{\mu}(dx_1) \cdots \tilde{\mu}(dx_n) \end{aligned} \tag{5.42}$$

Since $(-\varepsilon, \varepsilon) \subseteq (-\delta, \delta)$ and

$$\Big|\frac{x_1 + \cdots + x_n}{n}\Big| < \varepsilon \implies |\eta(x_1 + \cdots + x_n)| < n\varepsilon|\eta| \implies e^{-n\varepsilon|\eta|} < e^{-\eta(x_1 + \cdots + x_n)},$$

it follows from (5.42) that

$$\begin{aligned} \mathbb{P}\big(\bar{S}_n^X \in (-\delta, \delta)\big) &\geqslant e^{-n\varepsilon|\eta| + n\Lambda(\eta)} \int_{\left\{(x_1, \cdots, x_n): \left|\frac{x_1 + \cdots + x_n}{n}\right| < \varepsilon\right\}} \tilde{\mu}(dx_1) \cdots \tilde{\mu}(dx_n) \\ &= e^{-n\varepsilon|\eta| + n\Lambda(\eta)} \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \in (-\varepsilon, \varepsilon)\big). \end{aligned}$$

After taking logarithm, one arrives at

$$\frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (-\delta, \delta)\big) \geqslant -\varepsilon|\eta| + \Lambda(\eta) + \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \in (-\varepsilon, \varepsilon)\big). \tag{5.43}$$

Here is the main benefit of the change of measure: $\{\tilde{X}_n\}$ is an i.i.d. sequence with *mean zero*. One can therefore apply the strong LLN to conclude that $\bar{S}_n^{\tilde{X}} \to 0$ a.s. In particular,

$$\lim_{n \to \infty} \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \in (-\varepsilon, \varepsilon)\big) = 1.$$

By taking $n \to \infty$ in (5.43), it follows that

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (-\delta, \delta)\big) \geqslant -\varepsilon|\eta| + \Lambda(\eta) \geqslant -\varepsilon|\eta| + \inf_{\lambda\in\mathbb{R}} \Lambda(\lambda).$$

Since $\varepsilon \in (0, \delta)$ is arbitrary, one concludes that

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^X \in (-\delta, \delta)\big) \geqslant \inf_{\lambda\in\mathbb{R}} \Lambda(\lambda),$$

which gives the desired estimate (5.37).

Case II: μ has positive measure on both $(-\infty, 0)$ and $(0, \infty)$, but it needs not be supported in a bounded interval. This case is a rather technical extension of Case I (it can be omitted on first reading).

Choose $M_0 > 0$ such that

$$\mu([-M, 0)) > 0, \ \mu((0, M]) > 0 \quad \forall M \geqslant M_0.$$

For each fixed $M \geqslant M_0$, define ν to be the conditional law of X_1 given $|X_1| \leqslant M$ and ν_n to be the conditional law of $\bar{S}_n$ given $\{|X_i| \leqslant M, \forall i = 1, \cdots, n\}$. Note that ν_n is just the law of the sample average of an i.i.d. sequence whose marginal distribution is ν (why?). We also denote μ_n as the (unconditional) law of $\bar{S}_n$. By the definition of ν_n, one has

$$\mu_n((-\delta, \delta)) \geqslant \nu_n((-\delta, \delta)) \cdot \mu([-M, M])^n.$$

It follows from the result of Case I that

$$\begin{aligned} \varliminf_{n\to\infty} \frac{1}{n} \log \mu_n((-\delta, \delta)) &\geqslant \varliminf_{n\to\infty} \frac{1}{n} \log \nu_n((-\delta, \delta)) + \log \mu([-M, M]) \\ &\geqslant \inf_{\lambda\in\mathbb{R}} \Lambda_\nu(\lambda) + \log \mu([-M, M]). \end{aligned} \tag{5.44}$$

To compute $\Lambda_\nu(\lambda)$, by definition one has

$$M_\nu(\lambda) \triangleq \mathbb{E}\big[e^{\lambda X_1} \big| |X_1| \leqslant M\big] = \frac{\int_{[-M,M]} e^{\lambda x} \mu(dx)}{\mu([-M, M])}.$$

Hence

$$\Lambda_\nu(\lambda) \triangleq \log M_\nu(\lambda) = \Lambda_M(\lambda) - \log \mu([-M, M]), \tag{5.45}$$

where $\Lambda_M(\lambda) \triangleq \log \int_{[-M,M]} e^{\lambda x} \mu(dx)$. By substituting (5.45) into (5.44), one finds that

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mu_n((-\delta, \delta)) \geqslant \inf_{\lambda\in\mathbb{R}} \Lambda_M(\lambda) =: J_M.$$

It is apparent that J_M is increasing in M. Denoting $J^* \triangleq \lim_{M\to\infty} J_M$, it follows that

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mu_n((-\delta, \delta)) \geqslant J^*.$$

We claim that there exists $\lambda_0 \in \mathbb{R}$, such that $\Lambda(\lambda_0) \leqslant J^*$. The desired inequality (5.37) follows from this claim trivially. To prove the claim, for each $M \geqslant M_0$ we introduce the set

$$K_M \triangleq \{\lambda \in \mathbb{R} : \Lambda_M(\lambda) \leqslant J^*\}.$$

We summarise the key properties of K_M in the following two lemmas, which then lead to the conclusion of the claim easily.

Lemma 5.7. *K_M is a non-empty compact subset of $\mathbb{R}$.*

Proof. The key observation is that Λ_M is continuous on $\mathbb{R}$ and

$$\lim_{\lambda\to\pm\infty} \Lambda_M(\lambda) = \infty, \tag{5.46}$$

which follows from the same reason leading to (5.40). As a result, the global infimum of $\Lambda_M(\lambda)$ is attained at some $\lambda_M \in \mathbb{R}$, i.e.

$$\Lambda_M(\lambda_M) = J_M \leqslant J^*.$$

This shows that $K_M \neq \emptyset$. The compactness of K_M follows from the fact that it is a bounded, closed subset, which is again a consequence of (5.46). □

Lemma 5.8. *Let $\{L_n : n \geqslant 1\}$ be a decreasing sequence of non-empty compact subsets of $\mathbb{R}$. Then*

$$\bigcap_{n=1}^{\infty} L_n \neq \emptyset.$$

Proof. This is well-known in real analysis, but we still give its proof for completeness. Suppose on the contrary that $\cap_n L_n = \emptyset$. Then $\cup_n L_n^c = \mathbb{R} \supseteq L_1$. Since L_n^c is open for every n, by compactness there exists $N \geqslant 1$ such that

$$L_1 \subseteq \bigcup_{n=1}^{N} L_n^c = L_N^c.$$

This is absurd since $L_N \subseteq L_1$. Therefore, $\cap_n L_n \neq \emptyset$. □

According to Lemma 5.7, K_M ($M \geqslant M_0$) is a decreasing sequence of non-empty compact subsets of $\mathbb{R}$ (decreasingness follows from monotonicity of $M \mapsto \Lambda_M(\lambda)$). It then follows from Lemma 5.8 that

$$\bigcap_{M=M_0}^{\infty} K_M \neq \emptyset.$$

In particular, there exists $\lambda_0 \in \mathbb{R}$ such that

$$\Lambda_M(\lambda_0) \leqslant J^* \quad \forall M \geqslant M_0.$$

Letting $M \uparrow \infty$, one concludes that $\Lambda(\lambda_0) \leqslant J^*$. This proves the desired claim.

Case III: Either $\mu((-\infty, 0))$ or $\mu((0, \infty))$ is zero, say $\mu((-\infty, 0)) = 0$. This case can be dealt with independently in a relatively easy way.

In this case, one has

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X_1}] = \log \big(\mathbb{P}(X_1 = 0) + \mathbb{E}[e^{\lambda X_1}; X_1 > 0]\big).$$

In particular, $\Lambda(\lambda)$ is increasing on $\mathbb{R}$, and thus

$$\inf_{\lambda \in \mathbb{R}} \Lambda(\lambda) = \lim_{\lambda \to -\infty} \Lambda(\lambda) = \log \mathbb{P}(X_1 = 0). \tag{5.47}$$

To prove (5.37), one simply notes that

$$\mathbb{P}\big(\bar{S}_n \in (-\delta, \delta)\big) \geqslant \mathbb{P}(\bar{S}_n = 0) \geqslant \mathbb{P}(X_1 = 0, \cdots, X_n = 0) = \mathbb{P}(X_1 = 0)^n.$$

It follows from (5.47) that

$$\frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in (-\delta, \delta)\big) \geqslant \log \mathbb{P}(X_1 = 0) = \inf_{\lambda \in \mathbb{R}} \Lambda(\lambda).$$

Since this is true for all n, by taking $n \to \infty$ one obtains the desired inequality (5.37).

The following corollary gives the exact convergence (5.33) for special subsets. Its proof is only a small adaptation of what we have obtained so far.

Corollary 5.2. *Under the setting of Cramér's theorem, one has*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in [y, \infty)\big) = - \inf_{x \in [y, \infty)} \Lambda^*(x)$$

for all $y \in \mathbb{R}$.

Proof. Since $[y, \infty)$ is closed, one already has the upper bound (5.31) for the "limsup". To establish a matching lower bound for the "liminf", the point is to strengthen the key inequality (5.37) to the following one:

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\bar{S}_n \in [0, \delta)\big) \geqslant \inf_{\lambda\in\mathbb{R}} \Lambda_\mu(\lambda).$$

The entire argument developed in Section 5.6.4 carries through in exactly the same way with $(x-\delta, x+\delta)$, $(-\delta, \delta)$, $(-\varepsilon, \varepsilon)$ replaced by $[x, x+\delta)$, $[0, \delta)$, $[0, \varepsilon)$ respectively. There is only one exception though: in order to show that (cf. (5.43))

$$\frac{1}{n} \log \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \in [0, \varepsilon)\big) \to 0,$$

one now observes

$$\mathbb{P}\big(\bar{S}_n^{\tilde{X}} \in [0, \varepsilon)\big) = \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \geqslant 0\big) - \mathbb{P}\big(\bar{S}_n^{\tilde{X}} \geqslant \varepsilon\big) \to \frac{1}{2} - 0 = \frac{1}{2},$$

where the first limit $1/2$ follows from the central limit theorem and the second limit 0 is a consequence of LLN. □

Remark 5.9. Cramér's theorem has a natural extension to multidimensions. LDP for infinite dimensional distributions (e.g. laws of stochastic processes) is a significant research topic in modern probability theory. The general definition is given as follows. Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures over a topological space E equipped with its Borel σ-algebra (the σ-algebra generated by open subsets). We say that $\{\mu_n\}$ satisfies the *large deviation principle* with a rate function $I : E \to [0, \infty]$, if the following estimates hold true.

(i) [Upper bound] For any closed subset $F \subseteq E$, one has

$$\varlimsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leqslant -\inf_{x\in F} I(x).$$

(ii) [Lower bound] For any open subset $G \subseteq E$, one has

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geqslant -\inf_{x\in G} I(x).$$

We briefly mention one fundamental example which has profound implications in modern probability and PDE theory. Consider a *stochastic differential equation* (parametrised by n)

$$\begin{cases} dX_t^{(n)} = b(X_t^{(n)})dt + \frac{1}{\sqrt{n}}\sigma(X_t^{(n)})dB_t, \quad 0 \leqslant t \leqslant 1; \\ X_0^{(n)} = x, \end{cases}$$

where B_t is the so-called *Bownian motion* which resembles a continuous-time version of the simple random walk. Since n is large, one can think of the above stochastic dynamics as a "small random perturbation" of the deterministic dynamics (ordinary differential equation)

$$\begin{cases} dx_t = b(x_t)dt, \quad 0 \leqslant t \leqslant 1; \\ x_0 = x. \end{cases}$$

There is an obvious "law of large numbers" in this situation: the stochastic process $X_t^{(n)}$ converges to the deterministic function x_t as $n \to \infty$, simply because the random perturbation $\frac{1}{\sqrt{n}}\sigma(X_t^{(n)})dB_t$ vanishes in the limit. The renowned *Freidlin–Wentzell theorem* establishes a large deviation principle for the law of $X^{(n)}$ (as a stochastic process) on the infinite dimensional space of paths with an explicit rate function induced by the stochastic dynamics (b, σ, B). Although the setting here is quite involved, the essential idea is to some extent inspired by (and is not too much deeper than) the proof of Cramér's theorem we developed in this section.

Appendix. Proof of Kronecker's lemma

In this appendix, we give a proof of Kronecker's lemma (cf. Lemma 5.3).

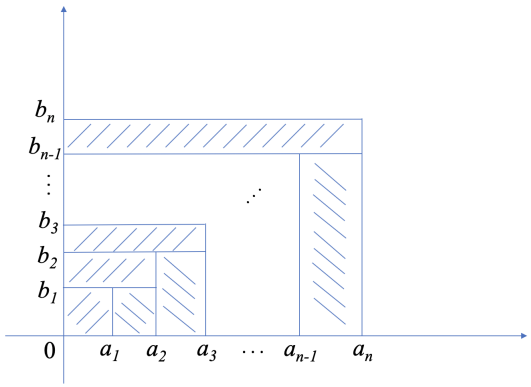
Proof of Lemma 5.3. Define $b_n \triangleq \sum_{j=1}^n \frac{x_j}{a_j}$ and set $a_0 = b_0 \triangleq 0$. Then $x_n = a_n(b_n - b_{n-1})$ and thus

$$\frac{1}{a_n}\sum_{j=1}^n x_j = \frac{1}{a_n}\sum_{j=1}^n a_j(b_j - b_{j-1}).$$

The crucial step is to write

$$\sum_{j=1}^n a_j(b_j - b_{j-1}) = a_n b_n - \sum_{j=0}^{n-1} b_j(a_{j+1} - a_j). \tag{5.48}$$

This is a discrete version of integration by parts. The intuition behind (5.48) is best illustrated by the figure below.



As a result of (5.48), one obtains that

$$\frac{1}{a_n}\sum_{j=1}^{n} x_j = b_n - \sum_{j=0}^{n-1} \frac{a_{j+1}-a_j}{a_n}\cdot b_j.$$

By the assumption of the lemma, say $b_n \to b \in \mathbb{R}$. We claim that

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} \frac{a_{j+1}-a_j}{a_n}\cdot b_j = b.$$

Indeed, given $\varepsilon > 0$, there exists $N \geqslant 1$ such that for all $n > N$, one has $|b_n - b| < \varepsilon$. It follows that for $n > N$,

$$\begin{aligned}
&\Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)b_j}{a_n} - b\Big| \\
&= \Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)(b_j-b)}{a_n}\Big| = \Big|\Big(\sum_{j\leqslant N} + \sum_{N<j\leqslant n-1}\Big)\frac{(a_{j+1}-a_j)(b_j-b)}{a_n}\Big| \\
&\leqslant \frac{a_{N+1}}{a_0}\cdot 2M + \varepsilon\cdot \sum_{N<j\leqslant n-1} \frac{a_{j+1}-a_j}{a_n} \leqslant \frac{2Ma_{N+1}}{a_n} + \varepsilon,
\end{aligned}$$

where $M > 0$ is a constant such that $|b_n| \leqslant M$ for all n. By letting $n \to \infty$, one obtains that

$$\overline{\lim_{n\to\infty}}\Big|\sum_{j=0}^{n-1} \frac{(a_{j+1}-a_j)b_j}{a_n} - b\Big| \leqslant \varepsilon.$$

The result follows as ε is arbitrary.

□

6 The characteristic function

In this chapter, we develop a fundamental tool for the study of distributional properties of random variables: the *characteristic function*.

In elementary probability, we have seen the notion of moment generating functions. There are many important reasons for introducing the moment generating function. For instance, it uniquely determines the law of a random variable. It can be used to compute moments effectively and to study convergence in distribution. One disadvantage of the moment generating function is that it is not always well-defined (the Cauchy distribution is such an example). Even if it is defined, it comes with its intrinsic domain of definition making the analysis cumbersome.

On the other hand, the characteristic function is always well-defined for any random variable. It has better analytic properties making it more convenient to work with, although a price to pay is that one needs to work with complex numbers (mostly in the obvious manner). The method of characteristic functions is rather powerful in the study of limiting behaviours of random variables, in particular in questions related to weak convergence (e.g the central limit theorem as we will see in the next chapter).

In Section 6.1, we give the definition of the characteristic function and discuss some of its basic properties. In Section 6.2, we discuss how one can recover the original distribution from the characteristic function (the inversion formula). In Section 6.3, we characterise weak convergence in terms of convergence of characteristic functions (the Lévy-Cramér continuity theorem). In Section 6.4, we discuss a few simple applications of the characteristic function. In Section 6.5, we discuss an elegant and useful result of G. Pólya which provides a sufficient condition for being characteristic functions.

6.1 Definition of the characteristic function and its basic properties

The characteristic function is a complexified version of the moment generating function. In particular, it takes complex values in general. We first recall a basic formula for complex exponentials. For $z = x + iy \in \mathbb{C}$, e^z is the complex number given by

$$e^z = e^x(\cos y + i \sin y).$$

Setting $x = 0$, one obtains *Euler's formula*:

$$e^{iy} = \cos y + i \sin y, \quad y \in \mathbb{R}. \tag{6.1}$$

Definition 6.1. Let X be a random variable. The *characteristic function* of X is the $\mathbb{C}$-valued function defined by

$$f_X(t) \triangleq \mathbb{E}[e^{itX}], \quad t \in \mathbb{R}. \tag{6.2}$$

Remark 6.1. Using Euler's formula (6.1), equation (6.2) is interpreted as

$$f_X(t) = \mathbb{E}[\cos tX] + i\mathbb{E}[\sin tX].$$

In most circumstances, there is no need to treat the real and imaginary parts separately; it is more effective to work over $\mathbb{C}$.

Remark 6.2. The characteristic function is defined in terms of the distribution of X and the underlying probability space plays no role. In fact, it is more intrinsic to write

$$f_X(t) = \int_{-\infty}^{\infty} e^{itx} \mu_X(dx)$$

where μ_X is the law of X. Equivalently, one can directly define the *characteristic function of a probability measure* μ on $\mathbb{R}$ as

$$f_\mu(t) \triangleq \int_{\mathbb{R}} e^{itx} \mu(dx)$$

without referring to any random variables. When X (or μ) admits a density function $\rho(x)$, the characteristic function is given by

$$f(t) = \int_{-\infty}^{\infty} e^{itx} \rho(x) dx.$$

This is also known as the *Fourier transform* of the function $\rho(x)$.

The first benefit of working with the characteristic function is that it is well-defined for all $t \in \mathbb{R}$. Indeed, by the triangle inequality one has

$$|f_X(t)| \leqslant \mathbb{E}[|e^{itX}|] = \mathbb{E}[1] = 1.$$

In addition, it is obvious that $f_X(0) = 1$ and

$$\overline{f_X(t)} = \overline{\mathbb{E}[e^{itX}]} = \mathbb{E}[e^{-itX}] = f_X(-t).$$

At a formal level, the characteristic function is related to the moment generating function $M_X(t)$ by the simple relation

$$f_X(t) = M_X(it).$$

In particular, one has the following parallel properties for the characteristic function.

Proposition 6.1. *Let $a, b \in \mathbb{R}$ and X, Y be random variables.*
(i) $f_{aX+b}(t) = e^{itb} f_X(at)$ and $f_{-X}(t) = \overline{f_X(t)}$.
(ii) If X and Y are independent, then $f_{X+Y}(t) = f_X(t) \cdot f_Y(t)$.

Proof. (i) By definition,

$$f_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{itaX} \cdot e^{itb}] = e^{itb} \cdot f_X(at),$$

and

$$f_{-X}(t) = \mathbb{E}[e^{it\cdot(-X)}] = f_X(-t) = \overline{f_X(t)}.$$

(ii) Recall from Proposition 3.4 that X and Y are independent if and only if

$$\mathbb{E}[\varphi(X)\psi(Y)] = \mathbb{E}[\varphi(X)] \cdot \mathbb{E}[\psi(Y)]$$

for any bounded Borel-measurable functions φ, ψ. Applying this to the function $\varphi(x) = \psi(x) = e^{itx}$ (for fixed t), one gets that

$$f_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX} \cdot e^{itY}] = \mathbb{E}[e^{itX}] \cdot \mathbb{E}[e^{itY}] = f_X(t) f_Y(t).$$

□

On the other hand, below is a nice analytic property which does not have its counterpart for moment generating functions.

Proposition 6.2. *The characteristic function $f_X(t)$ is uniformly continuous on $\mathbb{R}$.*

Proof. By definition, for any $t, h \in \mathbb{R}$ one has

$$\begin{aligned} f_X(t+h) - f_X(t) &= \int_{-\infty}^{\infty} (e^{i(t+h)x} - e^{itx}) \mu_X(dx) \\ &= \int_{-\infty}^{\infty} e^{itx} (e^{ihx} - 1) \mu_X(dx). \end{aligned}$$

According to the triangle inequality,

$$|f_X(t+h) - f_X(t)| \leqslant \int_{-\infty}^{\infty} |e^{ihx} - 1| \mu_X(dx). \tag{6.3}$$

Note that the right hand side is independent of t and the integrand $|e^{ihx} - 1| \to 0$ as $h \to 0$ (for each fixed x). By dominated convergence, the right hand side of (6.3) converges to zero as $h \to 0$. This gives the uniform continuity of $f_X(t)$. □

We do not list the explicit formulae for the characteristic functions of those special distributions one encounters in elementary probability. Some of them are straight forward while some could be tricky to derive. Here we look at one example: the standard normal distribution.

Example 6.1. The characteristic function of $X \stackrel{d}{=} \mathcal{N}(0,1)$ is given by $f(t) = e^{-t^2/2}$. We start with the definition

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx.$$

By differentiation and integration by parts,

$$\begin{aligned} f'(t) &= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{itx} e^{-x^2/2} dx = -\frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} d\big(e^{-x^2/2}\big) \\ &= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} d(e^{itx}) = -\frac{t}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = -t f(t). \end{aligned}$$

This is a first order ODE that can be solved uniquely with the initial condition $f(0) = 1$. Its solution is $f(t) = e^{-t^2/2}$.

We conclude this section with an elementary inequality for the complex exponential that will be used frequently later on.

Lemma 6.1. *For any $a, b \in \mathbb{R}$, one has*

$$|e^{ib} - e^{ia}| \leqslant |b - a|. \tag{6.4}$$

Proof. Assume that $a < b$. A simple application of the triangle inequality yields

$$|e^{ib} - e^{ia}| = \Big| \int_a^b i e^{it} dt \Big| \leqslant \int_a^b \big| i e^{it} \big| dt = \int_a^b 1 dt = b - a.$$

The desired inequality thus follows. □

6.2 Uniqueness theorem and inversion formula

One basic reason of working with the characteristic function is that it uniquely determines the distribution of the random variable. In addition, one can recover the distribution as well as many of its properties from the characteristic function in a fairly explicit way.

The main result here is the following *inversion formula*, which easily implies the uniqueness property.

Theorem 6.1. *Let μ be a probability measure on $\mathbb{R}$ and let $f(t)$ be its characteristic function. Then for any real numbers $x_1 < x_2$, one has*

$$\mu((x_1, x_2)) + \frac{1}{2}\mu(\{x_1\}) + \frac{1}{2}\mu(\{x_2\}) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt. \quad (6.5)$$

Remark 6.3. The function $\frac{e^{-itx_1}-e^{-itx_2}}{it}$ at $t = 0$ is defined in the limiting sense as $x_2 - x_1$. It should be pointed out that the right hand side of (6.5) cannot simply be understood as the integral

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt.$$

Indeed, such an integral over $(-\infty, \infty)$ may not be well-defined unless $f(t)$ is integrable over $\mathbb{R}$.

We postpone the proof of Theorem 6.1 to the end of this section and first discuss some of its implications. First of all, it implies the following uniqueness result, which asserts that a probability measure is uniquely determined by its characteristic function.

Corollary 6.1. *Let μ_1 and μ_2 be two probability measures. Suppose that they have the same characteristic function. Then $\mu_1 = \mu_2$.*

Proof. Let $D_i \triangleq \{x \in \mathbb{R}^1 : \mu_i(\{x\}) > 0\}$ denote the set of atoms (discontinuity points) for μ_i $(i = 1, 2)$ and $D \triangleq D_1 \cup D_2$. Since μ_1 and μ_2 have the same characteristic function, by the inversion formula (6.5) one has

$$\mu_1((x_1, x_2)) = \mu_2((x_1, x_2)), \quad \text{for all } x_1 < x_2 \text{ in } D^c. \quad (6.6)$$

On the other hand, D_1, D_2 are both countable and so is D. In particular, D^c is dense in $\mathbb{R}$. By a simple approximation argument, the relation (6.6) is enough to conclude that $\mu_1((a, b]) = \mu_2((a, b])$ for all real numbers $a < b$. This in turns implies $\mu_1 = \mu_2$ by Proposition 1.4. □

Due to the uniqueness theorem, many properties of the original distribution can be detected from its characteristic function. We give two examples of this kind. The first one only uses the uniqueness property while the second one requires an application of the inversion formula.

Proposition 6.3. *Let X be a random variable with characteristic function $f_X(t)$. Then $f_X(t)$ is real-valued if and only if X and $-X$ have the same distribution.*

Proof. Note that $f_{-X}(t) = f_X(-t) = \overline{f_X(t)}$. Therefore, $f_X(t)$ is real-valued if and only if $f_{-X}(t) = f_X(t)$, which according to the uniqueness theorem is equivalent to saying that $X \overset{\text{law}}{=} -X$. □

Proposition 6.4. *Let X be a random variable with distribution function $F(x)$ and characteristic function $f(t)$ respectively. Suppose that $f(t)$ is integrable over $(-\infty, \infty)$. Then $F(x)$ is continuously differentiable on $\mathbb{R}$ and its derivative (the probability density function) is given by the formula*

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt. \tag{6.7}$$

Proof. In this case, the inversion formula (6.5) is equivalently expressed as

$$\mathbb{P}(x_1 < X < x_2) + \frac{1}{2}\mathbb{P}(X = x_1) + \frac{1}{2}\mathbb{P}(X = x_2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t) dt. \tag{6.8}$$

Indeed, according to (6.4) one has

$$\left| \frac{e^{-itx_1} - e^{-itx_2}}{it} \right| \leqslant |x_1 - x_2|.$$

It follows from assumption that the function $t \mapsto \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)$ is integrable over $(-\infty, \infty)$.

We first show that F is left continuous (and thus continuous). Let $x \in \mathbb{R}$ and $h > 0$. Using the relations

$$\begin{cases} \mathbb{P}(x - h < X < x) = F(x-) - F(x - h), \\ \mathbb{P}(X = x - h) = F(x - h) - F((x - h)-), \\ \mathbb{P}(X = x) = F(x) - F(x-), \end{cases}$$

the inversion formula (6.8) applied to $x_1 = x - h$ and $x_2 = x$ yields that

$$\frac{1}{2}(F(x) - F(x - h)) + \frac{1}{2}(F(x-) - F((x - h)-)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-it(x-h)} - e^{-itx}}{it} f(t) dt. \tag{6.9}$$

Note that one always has

$$\lim_{h \downarrow 0} F((x - h)-) = F(x-) \quad \text{(why?)}.$$

In addition, since

$$\lim_{h\downarrow 0}\frac{e^{-it(x-h)}-e^{-itx}}{it}=0$$

for every fixed t, by dominated convergence the right hand side of (6.9) tends to zero as $h\downarrow 0$. Therefore, $F(x-h)\to F(x)$ as $h\downarrow 0$ which shows that F is left continuous at x.

Since F is continuous, by applying the inversion formula to $x_1=x$, $x_2=x+h$ and dividing it by h, one obtains that

$$\frac{F(x+h)-F(x)}{h}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-itx}-e^{-it(x+h)}}{ith}f(t)dt.$$

By dominated convergence, the right hand side tends to $\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}f(t)dt$ as $h\to 0$. Therefore, F is differentiable at x with derivative

$$F'(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}f(t)dt.$$

The continuity of $F'(x)$ follows from the continuity of $x\mapsto\int_{-\infty}^{\infty}e^{-itx}f(t)dt$, which is again a simple consequence of dominated convergence. □

Proof of the inversion formula (6.5)

The proof of (6.5) relies on the following Dirichlet integral

$$\int_0^{\infty}\frac{\sin u}{u}du=\frac{\pi}{2} \tag{6.10}$$

which was derived in Example 3.3 before. Note that this integral needs to be understood as an improper integral $\lim_{R\to\infty}\int_0^R\frac{\sin u}{u}du$ and $\frac{\sin 0}{0}\triangleq 1$.

To prove the inversion formula we begin with its right hand side. We fix $x_1<x_2$ throughout the discussion. By the definition of the characteristic function, for each $T>0$ one has

$$\begin{aligned}\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}f(t)dt&=\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}\Big(\int_{-\infty}^{\infty}e^{itx}\mu(dx)\Big)dt\\&=\int_{-\infty}^{\infty}\Big(\int_{-T}^{T}\frac{e^{-it(x_1-x)}-e^{-it(x_2-x)}}{it}dt\Big)\mu(dx).\end{aligned} \tag{6.11}$$

Here we used Fubini's theorem to change the order of integration. This is legal since

$$\left|\frac{e^{-itx_1}-e^{-itx_2}}{it}\cdot e^{itx}\right| \leqslant |x_2-x_1|$$

which is integrable over $[-T,T]\times\mathbb{R}$ with respect to the product measure $dt\otimes\mu$.

Next, we set

$$I_T(x;x_1,x_2) \triangleq \int_{-T}^{T}\frac{e^{-it(x_1-x)}-e^{-it(x_2-x)}}{it}dt.$$

By writing out the real and imaginary parts one obtains that

$$\begin{aligned}
&I_T(x;x_1,x_2)\\
&=\int_{-T}^{T}\frac{\big(\cos t(x_1-x)-\cos t(x_2-x)\big)+i\big(\sin t(x_2-x)-\sin t(x_1-x)\big)}{it}dt\\
&=2\Big(\int_0^T\frac{\sin t(x_2-x)}{t}dt-\int_0^T\frac{\sin t(x_1-x)}{t}dt\Big),
\end{aligned}$$

where the cosine part vanishes since $\cos x$ is an even function. By applying a change of variables and discussing according to different scenarios of x, it is routine to see that

$$I_T(x;x_1,x_2)=\begin{cases}
2\big(\int_0^{T(x_2-x)}\sin u/u du-\int_0^{T(x_1-x)}\sin u/u du\big), & x<x_1;\\
2\int_0^{T(x_2-x)}\sin u/u du, & x=x_1;\\
2\big(\int_0^{T(x_2-x)}\sin u/u du+\int_0^{T(x-x_1)}\sin u/u du\big), & x_1<x<x_2;\\
2\int_0^{T(x-x_1)}\sin u/u du, & x=x_2;\\
2\big(-\int_0^{T(x-x_2)}\sin u/u du+\int_0^{T(x-x_1)}\sin u/u du\big), & x>x_2.
\end{cases}\tag{6.12}$$

Sending $T\to\infty$ and using the Dirichlet integral (6.10), one obtains that

$$\lim_{T\to\infty}I_T(x;x_1,x_2)=\begin{cases}
0, & x<x_1;\\
\pi, & x=x_1;\\
2\pi, & x_1<x<x_2;\\
\pi, & x=x_2;\\
0, & x>x_2.
\end{cases}\tag{6.13}$$

Note that we have expressed the right hand side of the inversion formula (6.5) as

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-\infty}^{\infty}I_T(x;x_1,x_2)\mu(dx).$$

The equation (6.13) urges us to take limit under the integral sign. This is indeed legal as a result of the following elementary fact.

Lemma 6.2. *One has*

$$0 \leqslant \int_0^y \frac{\sin u}{u} du \leqslant \int_0^\pi \frac{\sin u}{u} du \quad \text{for all } y \geqslant 0.$$

In view of the expression (6.12) of $I_T(x; x_1, x_2)$, Lemma 6.2 shows that

$$|I_T(x; x_1, x_2)| \leqslant 4 \int_0^\pi \frac{\sin u}{u} du < \infty \quad \text{for all } x \text{ and } T.$$

According to the dominated convergence theorem and (6.13), one obtains that

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} I_T(x; x_1, x_2) \mu(dx) = \frac{1}{2}\mu(\{x_1\}) + \mu((x_1, x_2)) + \frac{1}{2}\mu(\{x_2\})$$

which concludes the desired inversion formula.

As the last piece of the puzzle, it remains to prove Lemma 6.2.

Proof of Lemma 6.2. We first show that $D(y) \triangleq \int_0^y \frac{\sin u}{u} du$ is non-negative for all $y \geqslant 0$. This is obvious when $y \in [0, \pi]$. When $y \in [(2k-1)\pi, (2k+1)\pi]$ with any $k \geqslant 1$, one has

$$\begin{aligned}
\int_0^y \frac{\sin u}{u} du &\geqslant \int_0^{2k\pi} \frac{\sin u}{u} du = \sum_{l=1}^{k} \int_{2(l-1)\pi}^{2l\pi} \frac{\sin u}{u} du \\
&= \sum_{l=1}^{k} \Big(\int_{(2l-2)\pi}^{(2l-1)\pi} \frac{\sin u}{u} du + \int_{(2l-1)\pi}^{2l\pi} \frac{\sin u}{u} du \Big) \\
&= \sum_{l=1}^{k} \int_{(2l-2)\pi}^{(2l-1)} (\sin u) \cdot \Big(\frac{1}{u} - \frac{1}{u+\pi} \Big) du \\
&\geqslant 0,
\end{aligned}$$

where we have applied a change of variables to the integral $\int_{(2l-1)\pi}^{2l\pi} \sin u/u du$ to reach the last equality.

Next, we show that $D(y)$ is maximised at $y = \pi$. Due to the sign pattern of $\sin u$, it is enough to show that

$$\int_{\pi}^{(2k+1)\pi} \frac{\sin u}{u} du \leqslant 0 \quad \forall k \geqslant 0. \text{ (why?)}$$

This can be proved in a similar way to the positivity part:

$$\begin{aligned}\int_{\pi}^{(2k+1)\pi} \frac{\sin u}{u} du &= \sum_{l=1}^{k} \Big(\int_{(2l-1)\pi}^{2l\pi} \frac{\sin u}{u} du + \int_{2l\pi}^{(2l+1)\pi} \frac{\sin u}{u} du \Big) \\ &= \sum_{l=1}^{k} \int_{(2l-1)\pi}^{2l\pi} (\sin u) \cdot \big(\frac{1}{u} - \frac{1}{u+\pi}\big) du \\ &\leqslant 0.\end{aligned}$$

□

Remark 6.4. The proof of the inversion formula (6.5) we gave here is not entire satisfactory, since we started from the right hand side of the formula pretending that it was known in advance. In the context of Fourier transform, it took mathematicians quite a while to understand why the simple inversion formula

$$\rho(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt$$

recovers the original function $\rho(x)$ from its Fourier transform $f(t)$. One needs to delve deeper into Fourier analysis to understand how the inversion formula arises naturally (cf. Appendix B for a discussion).

6.3 The Lévy-Cramér continuity theorem

One of the most important properties of the characteristic function is that weak convergence of random variables is equivalent to pointwise convergence of their characteristic functions. This is the content of *the Lévy-Cramér continuity theorem*. As we will see, it provides a useful tool for proving central limit theorems.

We start with the easier part of the theorem.

Theorem 6.2. *Let μ_n $(n \geqslant 1)$ and μ be probability measures on $\mathbb{R}$ with characteristic functions f_n $(n \geqslant 1)$ and f respectively. Suppose that μ_n converges weakly to μ. Then f_n converges to f uniformly on every finite interval of $\mathbb{R}$.*

Proof. For each fixed t, the function $x \mapsto e^{itx}$ is bounded and continuous. The convergence of $f_n(t)$ to $f(t)$ (for fixed t) is thus a trivial consequence of the weak convergence of μ_n to μ.

The uniformity assertion requires more effort than pointwise convergence. We first claim that, under the current assumption the family of functions $\{f_n : n \geqslant 1\}$

is *uniformly equicontinuous* on $\mathbb{R}$, in the sense that for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f_n(t) - f_n(s)| < \varepsilon$$

for all $n \geqslant 1$ and all s, t with $|t - s| < \delta$. To prove the uniform equicontinuity of $\{f_n\}$, first note that the family of probability measures $\{\mu_n : n \geqslant 1\}$ is tight as a consequence of weak convergence. In particular, given $\varepsilon > 0$, there exists $A = A(\varepsilon) > 0$ such that

$$\mu_n([-A, A]^c) < \varepsilon \quad \text{for all } n.$$

Next, for any real numbers t and h and $n \geqslant 1$, one has

$$\begin{aligned}
|f_n(t+h) - f_n(t)| &= \Big| \int_{-\infty}^{\infty} e^{i(t+h)x} \mu_n(dx) - \int_{-\infty}^{\infty} e^{itx} \mu_n(dx) \Big| \\
&\leqslant \int_{-\infty}^{\infty} |e^{ihx} - 1| \mu_n(dx) \\
&= \int_{\{x:|x|\leqslant A\}} |hx| \mu_n(dx) + \int_{\{x:|x|>A\}} 2\mu_n(dx) \\
&\leqslant |h|A + 2\mu_n([-A, A]^c) \\
&< |h|A + 2\varepsilon.
\end{aligned}$$

When $|h|$ is small enough (in a way independent of t), the right hand side can be made less than 3ε. This proves the uniform equicontinuity property.

Now we establish the desired uniform convergence. Let $I = [a, b]$ be an arbitrary finite interval. First of all, given $\varepsilon > 0$, by uniform equicontinuity there exists $\delta > 0$, such that whenever $|t - s| < \delta$ one has

$$|f_n(t) - f_n(s)| < \varepsilon \quad \forall n \geqslant 1.$$

We may also assume that for the same δ one has $|f(t) - f(s)| < \varepsilon$, since f is uniformly continuous (cf. Proposition 6.2). Next, we fix a finite partition

$$\mathcal{P} : a = t_0 < t_1 < \cdots < t_{r-1} < t_r = b$$

of $[a, b]$ such that $|t_i - t_{i-1}| < \delta$ for all $1 \leqslant i \leqslant r$. Since at each partition point t_i one has the pointwise convergence $f_n(t_i) \to f(t_i)$ and there are finitely many of them, one can find $N \geqslant 1$ such that

$$|f_n(t_i) - f(t_i)| < \varepsilon \text{ for all } n > N \text{ and } 0 \leqslant i \leqslant r.$$

It follows that for each $n > N$ and $t \in [a, b]$, with $t_i \in \mathcal{P}$ being the partition point such that $t \in [t_i, t_{i+1}]$, one has

$$\begin{aligned} |f_n(t) - f(t)| &\leqslant |f_n(t) - f_n(t_i)| + |f_n(t_i) - f(t_i)| + |f(t_i) - f(t)| \\ &< \varepsilon + \varepsilon + \varepsilon = 3\varepsilon. \end{aligned}$$

This gives the uniform convergence of f_n to f on $[a, b]$. □

The harder (and more useful) part of the theorem is the other direction which asserts that weak convergence can be established through pointwise convergence of the characteristic functions.

Theorem 6.3. *Let $\{\mu_n : n \geqslant 1\}$ be a sequence of probability measures on $\mathbb{R}$ with characteristic functions $\{f_n : n \geqslant 1\}$ respectively. Suppose that the following two conditions hold:*

(i) $f_n(t)$ converges pointwisely to some limiting function $f(t)$;
(ii) $f(t)$ is continuous at $t = 0$.

Then there exists a probability measure μ, such that μ_n converges weakly to μ. In addition, f is the characteristic function of μ.

We postpone its proof to the end of this section. There are two important remarks concerning the assumptions in the above two theorems. On the one hand, in Theorem 6.2 it is crucial to assume weak convergence of μ_n. As illustrated by the following example, f_n may fail to converge if only vague convergence is imposed.

Example 6.2. Let $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n$ be the two-point distribution at 0 and n with equal probabilities. It is a simple exercise that μ_n converges vaguely to the zero measure on $\mathbb{R}$. The characteristic function of μ_n is given by $f_n(t) = \frac{1}{2} + \frac{1}{2}e^{int}$, which fails to converge at any $t \notin 2\pi\mathbb{Z}$.

On the other hand, the following example illustrates that in Theorem 6.3, the continuity assumption of the limiting function at $t = 0$ cannot be removed. As we will also see in the proof of that theorem, such an assumption ensures tightness which is essential to expect weak convergence.

Example 6.3. Let μ_n be the normal distribution with mean zero and variance n. Then

$$f_n(t) = e^{-\frac{1}{2}nt^2} \overset{n\to\infty}{\longrightarrow} f(t) = \begin{cases} 0, & t \neq 0; \\ 1, & t = 0. \end{cases}$$

Note that although f_n converges pointwisely, the limiting function is not continuous at $t = 0$. The sequence μ_n converges vaguely to the zero measure and thus fails to be weakly convergent.

Combining the two theorems, one obtains the following neater but slightly weaker formulation.

Corollary 6.2. *Let μ_n $(n \geqslant 1)$ and μ be probability measures on $\mathbb{R}$, with characteristic functions f_n $(n \geqslant 1)$ and f respectively. Then μ_n converges weakly to μ if and only if f_n converges pointwisely to f.*

Proof. Necessity is trivial. For sufficiency, since f is a characteristic function it must be continuous at $t = 0$. In particular, the two conditions of Theorem 6.3 are both verified. As a result, there exists a probability measure ν such that μ_n converges weakly to ν and f is the characteristic function of ν. Since f is assumed to be the characteristic function of μ, by the uniqueness theorem one has $\nu = \mu$, hence showing that μ_n converges weakly to μ. □

Proof of Theorem 6.3

Before proving Theorem 6.3, we first derive a general estimate for the characteristic function which is also of independent interest.

Lemma 6.3. *Let μ be a probability measure on $\mathbb{R}$ with characteristic function f. Then for any $\delta > 0$, one has*

$$\mu([-2\delta^{-1}, 2\delta^{-1}]) \geqslant \frac{1}{\delta}\Big|\int_{-\delta}^{\delta} f(t)dt\Big| - 1. \tag{6.14}$$

Proof. By definition, one has

$$\begin{aligned}\int_{-\delta}^{\delta} f(t)dt &= \int_{-\delta}^{\delta}\int_{-\infty}^{\infty} e^{itx}\mu(dx)dt \\ &= \int_{-\infty}^{\infty}\mu(dx)\int_{-\delta}^{\delta}(\cos tx + i\sin tx)dt \\ &= \int_{-\infty}^{\infty}\frac{2\sin\delta x}{x}\mu(dx).\end{aligned}$$

Since $|\frac{\sin x}{x}| \leqslant 1$, it follows that

$$\begin{aligned}\frac{1}{2\delta}\Big|\int_{-\delta}^{\delta} f(t)dt\Big| &= \Big|\int_{-\infty}^{\infty} \frac{\sin \delta x}{\delta x}\mu(dx)\Big| \\ &\leqslant \int_{\{x:|\delta x|\leqslant 2\}} \mu(dx) + \int_{\{x:|\delta x|>2\}} \frac{1}{|\delta x|}\mu(dx) \\ &\leqslant \mu([-2\delta^{-1}, 2\delta^{-1}]) + \frac{1}{2}\mu([-2\delta^{-1}, 2\delta^{-1}]^c) \\ &= \frac{1}{2} + \frac{1}{2}\mu([-2\delta^{-1}, 2\delta^{-1}]).\end{aligned}$$

Rearranging the terms gives the desired inequality. □

The significance of Lemma 6.3 is that the continuity of $f(t)$ at $t = 0$ controls the speed that μ loses its mass at infinity. Indeed, a further rearrangement of (6.14) yields

$$\begin{aligned}\mu([-2\delta^{-1}, 2\delta^{-1}]^c) &\leqslant 2 - \frac{1}{\delta}\Big|\int_{-\delta}^{\delta} f(t)dt\Big| \\ &= \frac{\big|\int_{-\delta}^{\delta} f(0)dt\big| - \big|\int_{-\delta}^{\delta} f(t)dt\big|}{\delta} \quad (\text{since } f(0) = 1) \\ &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta} \big|f(t) - f(0)\big|dt. \qquad (6.15)\end{aligned}$$

This inequality shows that the speed that $\mu([-2\delta^{-1}, 2\delta^{-1}]^c) \to 0$ as $\delta \downarrow 0$ is controlled by the speed of convergence to zero for the right hand side, which is in turn controlled by the (modulus of) continuity of $f(t)$ at $t = 0$.

Remark 6.5. In the language of analysis, understanding the precise relationship between the tail behaviour of a function and the behaviour near the origin of its Fourier transform falls into the scope of *Tauberian theory*.

The key step for proving Theorem 6.3 is to show that the family $\{\mu_n : n \geqslant 1\}$ is tight, which in turn ensures the existence of a weakly convergent subsequence. The assumptions in the theorem play essential roles for establishing tightness through the estimate (6.15).

Proof of Theorem 6.3. Step one: tightness of $\{\mu_n\}$. According to (6.15), for every

$\delta > 0$ one has

$$\begin{aligned}\mu_n([-2\delta^{-1}, 2\delta^{-1}]^c) &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta} \big|f_n(t) - f_n(0)\big| dt \\ &\leqslant \frac{1}{\delta}\int_{-\delta}^{\delta} \big|f_n(t) - f(t)\big| dt + \frac{1}{\delta}\int_{-\delta}^{\delta} \big|f(t) - f(0)\big| dt\end{aligned}$$

where we have also used $f_n(0) = f(0) = 1$. Now given $\varepsilon > 0$, by the continuity assumption for $f(t)$ at $t = 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\frac{1}{\delta}\int_{-\delta}^{\delta} \big|f(t) - f(0)\big| dt < \varepsilon.$$

Since $f_n(t) \to f(t)$ for every t and $|f_n(t) - f(t)| \leqslant 2$, by the dominated convergence theorem (for such fixed δ) one sees that

$$\lim_{n\to\infty} \int_{-\delta}^{\delta} |f_n(t) - f(t)| dt = 0.$$

In particular, there exists $N = N(\varepsilon) \geqslant 1$ such that

$$\frac{1}{\delta}\int_{-\delta}^{\delta} \big|f_n(t) - f(t)\big| dt < \varepsilon \quad \text{for all } n > N.$$

It follows that

$$\mu_n([-2\delta^{-1}, 2\delta^{-1}]^c) < 2\varepsilon \quad \text{for all } n > N. \tag{6.16}$$

By further shrinking δ, one can ensure that (6.16) holds for $\mu_1, \cdots, \mu_N$ as well and thus for all n. This gives the tightness property.

Step two: there is precisely one weak limit point of μ_n. Since the family $\{\mu_n\}$ is tight, there exists a subsequence μ_{n_k} converging weakly to some probability measure μ. Let μ_{m_j} another subsequence which converges weakly to another probability measure ν. According to Theorem 6.2, one has

$$f_{n_k}(t) \to f_\mu(t), \ f_{m_j}(t) \to f_\nu(t)$$

for the corresponding characteristic functions. But from assumption, $f_n(t)$ converges pointwisely. As a result, one concludes that $f_\nu(t) = f_\mu(t)$, which implies $\nu = \mu$ by uniqueness. Consequently, the sequence has one and only one weak limit point μ.

Step three: μ_n converges weakly to μ. Let $f \in C_b(\mathbb{R})$ and denote $c_n \triangleq \int_{-\infty}^{\infty} f(x)\mu_n(dx)$. Suppose that c is a limit point of c_n, say along a subsequence c_{m_j}. By tightness, there is a further weakly convergent subsequence $\mu_{m_{j_l}}$, whose weak limit has to be μ by Step Two. As a result, one has

$$c_{m_{j_l}} = \int_{-\infty}^{\infty} f(x)\mu_{m_{j_l}}(dx) \to \int_{-\infty}^{\infty} f(x)\mu(dx)$$

as $l \to \infty$. This shows that $c = \int_{-\infty}^{\infty} f(x)\mu(dx)$. In other words, c_n has precisely one limit point c. It follows that

$$c_n = \int_{-\infty}^{\infty} f(x)\mu_n(dx) \to c = \int_{-\infty}^{\infty} f(x)\mu(dx)$$

as $n \to \infty$. This proves the weak convergence of μ_n to μ.

□

6.4 Some applications of the characteristic function

We discuss a few simple applications of the characteristic function. Its more powerful applications to central limit theorems will be discussed in the Chapter 7.

In the first place, by formally differentiating the expression $f(t) = \mathbb{E}[e^{itX}]$ at $t = 0$, one obtains that $f^{(k)}(0) = i^k\mathbb{E}[X^k]$. This suggests that the characteristic function can be used to compute moments. The following result makes this point precise.

Theorem 6.4. *Suppose that the random variable X has finite absolute moments up to order n. Then its characteristic function $f(t)$ has bounded, continuous derivatives up to order n, and they are given by*

$$f^{(k)}(t) = i^k\mathbb{E}[X^k e^{itX}], \quad 1 \leqslant k \leqslant n.$$

In particular, $\mathbb{E}[X^k] = \frac{f^{(k)}(0)}{i^k}$ for each $1 \leqslant k \leqslant n$.

Proof. We only consider the case when $n = 1$ as the general case follows by induction. First of all, for any real numbers t and h one has

$$\frac{f(t+h) - f(t)}{h} = \mathbb{E}\Big[\frac{e^{i(t+h)X} - e^{itX}}{h}\Big].$$

Note that

$$\frac{e^{i(t+h)X} - e^{itX}}{h} \to iXe^{itX} \quad \text{as } h \to 0,$$

and $\left|\frac{e^{i(t+h)X}-e^{itX}}{h}\right| \leqslant |X|$ which is integrable by assumption. It follows from the dominated convergence theorem that

$$\frac{f(t+h) - f(t)}{h} \to \mathbb{E}[iXe^{itX}] \quad \text{as } h \to 0,$$

which is also the derivative of $f(t)$. Its continuity is another consequence of dominated convergence. □

The following result is a direct corollary of Theorem 6.4 and the Taylor approximation theorem in real analysis.

Corollary 6.3. *Under the same assumption as in Theorem 6.4, one has*

$$f(t) = \sum_{k=0}^{n} \frac{i^k \mathbb{E}[X^k]}{k!} t^k + o(|t|^n),$$

where $o(|t|^n)$ denotes a function such that $o(|t|^n)/|t|^n \to 0$ as $t \to 0$.

As another application, we reproduce the weak LLN in the i.i.d. case by using the characteristic function.

Theorem 6.5. *Let $\{X_n : n \geqslant 1\}$ be a sequence of i.i.d. random variables with finite mean $m \triangleq \mathbb{E}[X_1]$. Then*

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = m \quad \text{in prob.}$$

Proof. Since the asserted limit is a deterministic constant, it is equivalent to proving weak convergence (cf. Proposition 4.2). Let $f(t)$ be the characteristic function of X_1 (thus of X_n for every n). Then, with $S_n \triangleq X_1 + \cdots + X_n$ one has

$$f_{S_n/n}(t) = \mathbb{E}\big[e^{it(X_1+\cdots+X_n)/n}\big] = \big(f\big(\frac{t}{n}\big)\big)^n.$$

Since X_1 has finite mean, by Corollary 6.3 one can write

$$f_{S_n/n}(t) = \big(1 + \frac{imt}{n} + o(1/n)\big)^n = (1+q_n)^{\frac{1}{q_n}\cdot nq_n},$$

where $q_n \triangleq imt/n + o(1/n)$. Note that $q_n \to 0$ and $nq_n \to imt$ as $n \to \infty$. Therefore, $(1+q_n)^{1/q_n} \to e$ and $f_{S_n/n}(t) \to e^{imt}$ as $n \to \infty$. Since e^{imt} is the characteristic function of the constant random variable $X = m$, one concludes from the Lévy-Cramér continuity theorem that S_n/n converges weakly to m. □

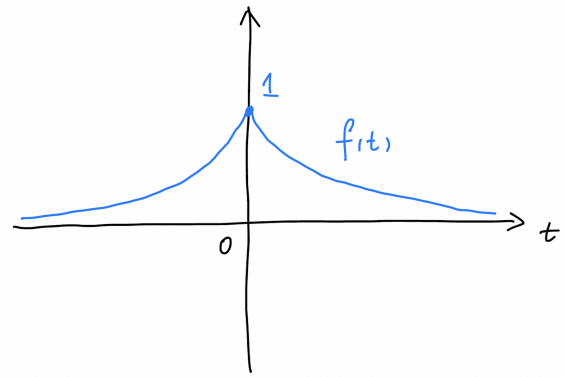
6.5 Pólya's criterion for characteristic functions

Let us consider the following question. Suppose that $f(t)$ is a given function. *How can one know if it is the characteristic function of some random variable / probability measure*? There is a general theorem due to S. Bochner, which provides a necessary and sufficient condition for $f(t)$ to be a characteristic function. Bochner's criterion is not easy to verify in practice. On the other hand, there is a simple sufficient condition discovered by G. Pólya which is more useful in many situations. Pólya's criterion can often be checked explicitly and be used to construct a rich class of characteristic functions. The main theorem is stated as folllows.

Theorem 6.6. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function which satisfies the following properties:*

(i) $f(0) = 1$ and $f(t) = f(-t)$ for all t;
(ii) $f(t)$ is decreasing and convex on $(0, \infty)$;
(iii) $f(t)$ is continuous at the origin and $\lim_{t\to\infty} f(t) = 0$.
Then $f(t)$ is the characteristic function of some random variable / probability measure.

The generic shape of functions that satisfy Pólya's criterion is sketched in the figure below.



Remark 6.6. The conditions in the theorem imply that $f(t)$ is non-negative. The condition that $f(t)$ is continuous at $t = 0$ is important. Indeed, the function

$$f(t) \triangleq \begin{cases} 1, & t = 0; \\ 0, & t \neq 0 \end{cases}$$

satisfies all conditions of the theorem except for continuity at the origin. This function is clearly not a characteristic function. The condition that $\lim_{t\to\infty} f(t) = 0$ is not essential and can be replaced by $\lim_{t\to\infty} f(t) = c > 0$ for some $c \in (0,1)$. Indeed, in the latter case one considers

$$g(t) \triangleq \frac{f(t) - c}{1 - c}.$$

Then $g(t)$ satisfies the conditions of the theorem and is thus a characteristic function. But one can then write

$$f(t) = (1 - c) \cdot g(t) + c \cdot 1,$$

which is a convex combination of two characteristic functions ($g(t)$ and 1). As a result, $f(t)$ is also a characteristic function (cf. Lemma 6.4 below).

Before proving Theorem 6.6, we first look at a simple but enlightening example.

Example 6.4. A basic example that satisfies Pólya's criterion is given as follows:

$$f(t) = (1 - |t|)^+ \triangleq \begin{cases} 1 - |t|, & |t| \leqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

For this example, there is no need to use Theorem 6.6 to see that it is a characteristic function. By evaluating the inversion formula (6.7) explicitly, one finds that $f(t)$ is the characteristic function of the distribution whose probability density function is given by

$$\rho(x) = \frac{1 - \cos x}{\pi x^2}, \quad x \in \mathbb{R}.$$

Example 6.5. Another example that satisfies Pólya's criterion is the function $f_\alpha(t) \triangleq e^{-|t|^\alpha}$ ($\alpha \in (0,1]$). In particular, this covers the case of the Cauchy distribution (when $\alpha = 1$). When $\alpha \in (1,2)$, f_α is still a characteristic function, however, Theorem 6.6 does not apply since f_α is no longer convex. The treatment of this case will be given in Section 7.3 below by using a different approach when we study the central limit theorem.

The starting point for proving Theorem 6.6 is the observation that convex combinations of characteristic functions are again characteristic functions.

Lemma 6.4. *Suppose that f_1, f_2 are characteristic functions and $\lambda \in (0,1)$. Then $\lambda f_1 + (1 - \lambda) f_2$ is also a characteristic function.*

Proof. Let μ_1, μ_2 be the probability measures corresponding to f_1, f_2 respectively. By definition, one has

$$f_i(t) = \int_{\mathbb{R}} e^{itx} \mu_i(dx), \quad i = 1, 2.$$

It follows that

$$f(t) \triangleq \lambda f_1(t) + (1-\lambda) f_2(t) = \int_{\mathbb{R}} e^{itx} \big(\lambda \mu_1 + (1-\lambda)\mu_2\big)(dx).$$

In other words, $f(t)$ is the characteristic function of the probability measure $\lambda\mu_1 + (1-\lambda)\mu_2$. □

Lemma 6.4 naturally extends to the case of more than two members: if $f_1, \cdots, f_n$ are characteristic functions and $\lambda_1, \cdots, \lambda_n$ are positive numbers such that $\lambda_1 + \cdots + \lambda_n = 1$, then

$$\lambda_1 f_1 + \cdots + \lambda_n f_n$$

is also a characteristic function. Without surprise, this fact can further be generalised to the case of the convex combination of a continuous family of characteristic functions. To be specific, let ν be a probability measure on $(0, \infty)$ and for each $r \in (0, \infty)$ let $t \mapsto f_r(t)$ be a characteristic function. Under suitable measurability condition on $r \mapsto f_r$, it can be shown that the function

$$t \mapsto \int_{(0,\infty)} f_r(t) \nu(dr)$$

is also a characteristic function. The assumption that ν is a probability measure on $(0, \infty)$ ensures that this is a "convex combination" of the family $\{f_r : r \in (0, \infty)\}$ of characteristic functions weighted by the measure ν.

As a result, the key idea of proving Theorem 6.6 is to express $f(t)$ as a convex combination of a (continuous) family of characteristic functions, more precisely, as

$$f(t) = \int_{(0,\infty)} f_r(t) \nu(dr) \tag{6.17}$$

where f_r is some classical characteristic function (for each $r > 0$) and ν is a probability measure on $(0, \infty)$. The above discussion then shows that f must also be a characteristic function. We now implement this idea mathematically.

Proof of Theorem 6.6. Step one. We first collect some standard properties arising from the convexity of $f(t)$ as well as the other assumptions in the theorem (we will not prove them here). Recall that the right derivative of $f(t)$ is given by

$$f'_+(t) \triangleq \lim_{h\downarrow 0} \frac{f(t+h)-f(t)}{h}.$$

(i) f'_+ is well defined and one has $-\infty < f'_+(t) \leqslant 0$ for every $t > 0$.
(ii) f'_+ is increasing and right continuous on $(0,\infty)$.
(iii) For each given $t > 0$, f is Lipschitz (and absolutely continuous) on $[t,\infty)$.
(iv) Since $\lim_{t\to\infty} f(t) = 0$, one has

$$\lim_{t\to\infty} f'_+(t) = 0.$$

Step two. Since f'_+ is increasing and right continuous, it induces a Lebesgue-Stieltjes measure μ on $\mathcal{B}(\mathbb{R})$ which satisfies

$$\mu((a,b]) \triangleq f'_+(b) - f'_+(a), \quad 0 < a < b.$$

Using μ and the density function $\rho(r) = r$, we introduce another measure ν on $(0,\infty)$ by

$$\nu(dr) \triangleq r\mu(dr).$$

The definition of ν is understood as

$$\nu(A) \triangleq \int_A r\nu(dr), \ \ A \in \mathcal{B}((0,\infty)).$$

Step three. We shall express $f(t)$ as an integral with respect to ν in the form (6.17). To this end, one first observes from the definition of μ that

$$-f'_+(s) = 0 - f'_+(s) = f'_+(\infty) - f'_+(s) = \int_s^\infty \mu(dr) = \int_s^\infty r^{-1}\nu(dr)$$

for every $s > 0$. In addition, by the fundamental theorem of calculus one has

$$f(t) = -(f(\infty) - f(t)) = -\int_t^\infty f'_+(s)ds = \int_t^\infty \int_s^\infty r^{-1}\nu(dr)ds$$

for every $t > 0$. Using Fubini's theorem, one obtains that

$$\begin{aligned} f(t) &= \int_t^\infty \Big(\int_t^r ds\Big) r^{-1}\nu(dr) = \int_t^\infty \big(1 - \frac{t}{r}\big)\nu(dr) \\ &= \int_{(0,\infty)} \big(1 - \frac{t}{r}\big)^+ \nu(dr), \quad \text{for all } t > 0. \end{aligned}$$

Since $f(t)$ is an even function, one arrives at

$$f(t) = \int_{(0,\infty)} \big(1 - \frac{|t|}{r}\big)^+ \nu(dr), \quad \text{for all } t \in \mathbb{R}\backslash\{0\}. \tag{6.18}$$

Step four. For each given $r > 0$, the function

$$f_r(t) \triangleq \big(1 - \frac{|t|}{r}\big)^+, \quad t \in \mathbb{R}$$

is a characteristic function. This is a direct consequence of Example 6.4 and the scaling property of the characteristic function.

Step five. It remains to show that ν is a probability measure on $(0, \infty)$ which then recognises (6.18) as a convex combination of the family $\{f_r : r > 0\}$. To this end, one sends $t \downarrow 0$ in the equation (6.18). By the assumption, the left hand side converges to $f(0) = 1$. For the right hand side, note that for each fixed r one has

$$\big(1 - \frac{|t|}{r}\big)^+ \uparrow 1 \quad \text{as } t \downarrow 0.$$

According to the monotone convergence theorem,

$$1 = \lim_{t\downarrow 0} \int_{(0,\infty)} \big(1 - \frac{|t|}{r}\big)^+ \nu(dr) = \int_{(0,\infty)} 1\nu(dr) = \nu(0,\infty).$$

Therefore, ν is a probability measure on $(0, \infty)$, hence finishing the proof of Theorem 6.6.

□

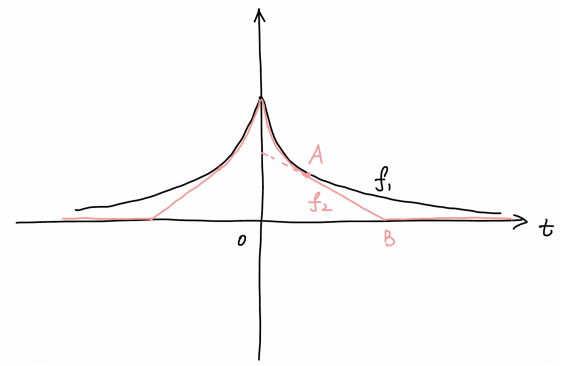
We conclude this chapter with two applications of Pólya's theorem.

Corollary 6.4. *Let $c > 0$. There exist two different characteristic functions f_1, f_2 such that*

$$f_1(t) = f_2(t) \quad \text{for } t \in (-c, c).$$

Proof. Let $f_1(t) = e^{-|t|}$ be the characteristic function of the Cauchy distribution. We draw the tangent line of $f_1(t)$ at the point $A = (c, f_1(c))$ and let this line intersect the positive t-aixs at the point B. We construct f_2 whose graph on $(0, \infty)$ consists of the following parts:

(i) the graph of f_1 on the part of $(0, c)$;
(ii) the line segment $\overline{AB}$ on the part from A to B;
(iii) the zero function from B to infinity. □



We also extend $f_2(t)$ to the negative t-axis by symmetry. It is readily checked that f_2 satisfies Pólya's criterion and is thus a characteristic function. The functions f_1, f_2 satisfy the desired property.

Corollary 6.5. *There exist three characteristic functions f_1, f_2, f_3 such that $f_1 \neq f_2$ but $f_1 f_3 = f_2 f_3$.*

Proof. Let f_1, f_2 be given as in Corollary 6.4 and set $f_3(t) \triangleq (1 - |t|/c')^+$ where $c' \in (0, c)$ is a fixed constant. Then f_1, f_2, f_3 are desired. □

Remark 6.7. Corollary 6.5 tells us that the cancellation law does not hold for characteristic functions:

$$f_1 f_3 = f_2 f_3 \nRightarrow f_1 = f_2.$$

Appendix A. The uniqueness theorem without inversion

Using the inversion formula to prove the uniqueness result (as we did before) is quite involved and unnatural. There is another argument which provides more insight into the uniqueness property. Suppose that μ_1 and μ_2 have the same characteristic function, i.e.

$$\int_{-\infty}^{\infty} e^{itx} \mu_1(dx) = \int_{-\infty}^{\infty} e^{itx} \mu_2(dx) \quad \text{for all } t \in \mathbb{R}.$$

We want to show that $\mu_1 = \mu_2$. A natural idea is described as follows.

(i) It is enough to show that

$$\int_{-\infty}^{\infty} f(x) \mu_1(dx) = \int_{-\infty}^{\infty} f(x) \mu_2(dx) \tag{6.19}$$

for a sufficiently rich class of functions f.
(ii) Such a class of functions can be approximated by linear combinations of functions from the family $\{e^{itx} : t \in \mathbb{R}\}$.

The first point is reasonable to expect. The fact that the family $\{e^{itx} : t \in \mathbb{R}\}$ generates a rich class of functions is also natural from the view of *Fourier series*: any continuous periodic function $f(x)$ with period $T = 1$ (i.e. $f(x+1) = f(x)$) admits a Fourier series expansion

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n x}, \quad x \in [0,1],$$

where $c_n \triangleq \int_0^1 f(x) e^{2\pi i n x} dx$ is the n-th Fourier coefficient.

Instead of using Fourier series, we shall take a different approach to implement the above idea mathematically. The key ingredient is the so-called *Stone-Weierstrass theorem*, which is stated in the context of periodic functions as follows. Its proof can be found in [Lan93].

Theorem 6.7 (The Stone-Weierstrass Theorem for Periodic Functions)**.** *Let $T > 0$. Define $\mathcal{C}_T$ to be the space of periodic functions $f : \mathbb{R} \to \mathbb{C}$ with period T. Let $\mathcal{A}$ be a subset of $\mathcal{C}_T$ satisfying the following three properties.*

(i) $\mathcal{A}$ is an algebra: $f, g \in \mathcal{A}, a, b \in \mathbb{R} \implies af + bg, f \cdot g \in \mathcal{A}$.
(ii) $\mathcal{A}$ vanishes at no point: for any $x \in [0,T)$, there exists $f \in \mathcal{A}$ such that $f(x) \neq 0$.
(iii) $\mathcal{A}$ separates points: for any $x \neq y \in [0,T)$, there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

Then $\mathcal{A}$ is dense in $\mathcal{C}_T$ with respect to uniform convergence on $[0,T]$. More precisely, for any periodic function $f \in \mathcal{C}_T$ and $\varepsilon > 0$, there exists $g \in \mathcal{A}$ such that

$$\sup_{t \in [0,T]} |f(t) - g(t)| < \varepsilon.$$

Now we proceed to prove the uniqueness result for the characteristic function by using the Stone-Weierstrass theorem.

Another proof of Corollary 6.1. Let μ_1, μ_2 be two probability measures with the same characteristic function.

We first claim that (6.19) holds for any continuous periodic function f. Indeed, let $T > 0$ be an arbitrary positive number and define $\mathcal{C}_T$ to be the space of periodic functions $f : \mathbb{R} \to \mathbb{C}$ with period T. Let $\mathcal{A}_T \subseteq \mathcal{C}_T$ be the vector space spanned by

the family $\{e^{2\pi inx/T} : n \in \mathbb{Z}\}$ of functions. It is routine to check that $\mathcal{A}_T$ satisfies all the assumptions in Theorem 6.7. As a result, $\mathcal{A}_T$ is dense in $\mathcal{C}_T$ with respect to uniform convergence on $[0, T]$. On the other hand, by assumption one knows that (6.19) holds for all $f \in \mathcal{A}_T$. It follows from approximation that (6.19) holds for all $f \in \mathcal{C}_T$.

Next, we claim that (6.19) holds for all bounded, continuous function f. The idea is to replace f by a periodic function with large period. Given an arbitrary $\varepsilon > 0$, there exists $M > 0$ such that

$$\mu_i([-M, M]^c) < \varepsilon \quad \text{for } i = 1, 2.$$

Let $g : [-M - 1, M + 1] \to \mathbb{R}$ be the continuous function given by

$$g(x) \triangleq \begin{cases} f(x), & x \in [-M, M]; \\ 0, & x \in (-\infty, -M - 1) \cup (M + 1, \infty); \\ \text{linear}, & x \in [-M - 1, -M] \text{ or } x \in [M, M + 1]. \end{cases}$$

By definition, one has $g(-M - 1) = g(M + 1)$ and

$$|g(x)| \leqslant \|f\|_\infty \quad \forall x \in [-M - 1, M + 1].$$

Let $\bar{g} : \mathbb{R} \to \mathbb{R}$ be the periodic extension of g to $\mathbb{R}$ with period $T = 2M + 2$. From the previous step one knows that (6.19) holds for $\bar{g}$. Since $f = \bar{g}$ on $[-M, M]$, it follows that

$$\begin{aligned} &\Big| \int f d\mu_1 - \int f d\mu_2 \Big| \\ &\leqslant \Big| \int f d\mu_1 - \int \bar{g} d\mu_1 \Big| + \Big| \int \bar{g} d\mu_1 - \int \bar{g} d\mu_2 \Big| + \Big| \int \bar{g} d\mu_2 - \int f d\mu_2 \Big| \\ &= \Big| \int f d\mu_1 - \int \bar{g} d\mu_1 \Big| + \Big| \int \bar{g} d\mu_2 - \int f d\mu_2 \Big| \\ &\leqslant 2\|f\|_\infty \cdot \big(\mu_1([-M, M]^c) + \mu_2([-M, M]^c)\big) < 4\|f\|_\infty \varepsilon. \end{aligned}$$

As ε is arbitrary, one obtains (6.19) for f.

Finally, if (6.19) holds for all bounded, continuous functions, one must have $\mu_1 = \mu_2$. The proof of this claim is left as an exercise.

□

Appendix B. Formal derivation of the inversion formula

Let μ be a probability measure on $\mathbb{R}$ and let $f(t)$ be its characteristic function. Recall that the inversion formula is given by

$$\mu((x_1, x_2)) + \frac{1}{2}\mu(\{x_1\}) + \frac{1}{2}\mu(\{x_2\}) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)dt. \quad (6.20)$$

In Section 6.2, we proved the above formula by starting from the right hand side of the equation (as if we know the formula in advance). This perspective is thus not very natural and satisfactory. In this appendix, we give a more constructive derivation of the inversion formula (indeed of (6.21) below). Our discussion here aims at conveying the essential idea and is thus only semi-rigorous.

First of all, recall from Proposition 6.4 that if $f(t)$ is integrable on $\mathbb{R}$, then μ admits a continuous density function given by

$$\rho(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} f(t)dt. \quad (6.21)$$

At a formal level, it is not hard to see how the two formulae (6.20) and (6.21) are related with each other. On the one hand, presuming that μ has no discontinuity points, by taking $x_1 = x$, $x_2 = x + h$ and sending $h \to 0$ after dividing both sides by h, one obtains (6.21). On the other hand, integrating (6.21) with respect to x over $[x_1, x_2]$ produces (6.20).

For simplicity, we shall work at the level of functions instead of measures. In particular, our main goal is to give a constructive derivation of the inversion formula (6.21). In the context of functions, the characteristic function is more commonly known as the *Fourier transform*. To be precise, the Fourier transform of a function $g : \mathbb{R} \to \mathbb{C}$ is the function defined by

$$\hat{g}(t) \triangleq \int_{\mathbb{R}} e^{itx} g(x)dx, \quad t \in \mathbb{R}.$$

The question thus becomes: *how can one recover g from $\hat{g}$?* There are two key observations before answering this question.

(i) Suppose that X is a random variable. For $\varepsilon > 0$, let N_ε be a Gaussian random variable with mean zero and variance ε, and we assume that X, N_ε are independent. It is reasonable to expect that $X + N_\varepsilon$ converges to X as $\varepsilon \to 0$ in

a reasonable sense. Now suppose that $g(x)$ is the probability density function of X. One knows that the convolution

$$(g * \rho_\varepsilon)(x) \triangleq \int_{\mathbb{R}} g(y)\rho_\varepsilon(x-y)dy$$

is the density of $X + N_\varepsilon$, where $\rho_\varepsilon(x) \triangleq \frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{x^2}{2\varepsilon}}$ is the density of N_ε. As a consequence, it is natural to expect that

$$\int_{\mathbb{R}} g(y)\rho_\varepsilon(x-y)dy \to g(x)$$

as $\varepsilon \to 0$.

(ii) The following property of the Fourier transform is fairly straight forward:

$$\int_{\mathbb{R}} \hat{g}(t)h(t)dt = \int_{\mathbb{R}} g(t)\hat{h}(t)dt. \tag{6.22}$$

Indeed, one has

$$\begin{aligned}
&\int_{\mathbb{R}} \hat{g}(t)h(t)dt = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} e^{itx} g(x)dx \Big) h(t)dt \\
&= \int_{\mathbb{R}} g(x)dx \int_{\mathbb{R}} e^{itx} h(t)dt \quad \text{(Fubini's theorem)} \\
&= \int_{\mathbb{R}} g(x)\hat{h}(x)dx.
\end{aligned}$$

Accepting the above two points, here is a natural idea of recovering the function g from $\hat{g}$. Firstly, we recall from Point (i) that

$$g(x) = \lim_{\varepsilon \to 0} \int_{\mathbb{R}} g(y)\rho_\varepsilon(x-y)dy.$$

To proceed further, let us ask the following question: *given fixed x and ε, which function $h_{x,\varepsilon}$ has Fourier transform given by $\hat{h}_{x,\varepsilon}(y) = \rho_\varepsilon(x-y)$?*

If one knows the answer, by applying (6.22) from Point (ii) one would have

$$g(x) = \lim_{\varepsilon \to 0} \int_{\mathbb{R}} g(y)\hat{h}_{x,\varepsilon}(y)dy = \lim_{\varepsilon \to 0} \int_{\mathbb{R}} \hat{g}(t)h_{x,\varepsilon}(t)dt.$$

One then expects that the last expression yields an explicit inversion formula.

It is not too difficult to figure out the answer to the above question. In fact, one notes that $y \mapsto \rho_\varepsilon(x-y)$ is the density of the Gaussian distribution with mean x and variance ε. In particular, its Fourier transform (or characteristic function) is given by

$$t \mapsto e^{ixt-\frac{1}{2}\varepsilon^2 t^2}.$$

Based on this observation, a moment's though reveals that

$$h_{x,\varepsilon}(t) \triangleq \frac{1}{2\pi} e^{-ixt-\frac{1}{2}\varepsilon^2 t^2}$$

is the desired function. One can of course directly verify that $\hat{h}_{x,\varepsilon}(\cdot) = \rho_\varepsilon(x-\cdot)$ again by using the formula for the Fourier transform of a Gaussian density.

To summarise the above formal discussion, one concludes that

$$g(x) = \lim_{\varepsilon\to 0} \frac{1}{2\pi} \int_{\mathbb{R}} \hat{g}(t) e^{-ixt-\frac{1}{2}\varepsilon^2 t^2} dt = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \hat{g}(t) dt.$$

This is exact the inversion formula (6.21) one looks for!

Appendix C. A geometric proof of Pólya's theorem

The proof of Pólya's theorem (cf. Theorem 6.6) we gave in Section 6.5 is based on expressing $f(t)$ as a convex combination of a continuous family of characteristic functions. However, discovering such an ingenious representation (6.18) requires deep insight and is not obvious at all. The purpose of this appendix is to develop an alternative (and more constructive) proof of Pólya's theorem. The argument here relies heavily on Euclidean geometric considerations.

Since the theorem is concerned with even functions, from now on we will only work on the positive t-axis and assume that all functions are extended to the negative t-axis by symmetry.

The building block: the simplest example

Our starting point is the following classical fact: the function

$$f_1(t) = (1-|t|)^+, \quad t \in \mathbb{R}.$$

is a characteristic function. Indeed, using the inversion formula one checks that f_1 is the characteristic function of the distribution whose density function is given by

$$\rho(x) = \frac{1-\cos x}{\pi x^2}, \quad x \in \mathbb{R}.$$

By the scaling property, for each $r > 0$ the function

$$f_r(t) \triangleq \big(1 - \frac{|t|}{r}\big)^+, \quad t \in \mathbb{R}, \tag{6.23}$$

is also a characteristic function. The shape of f_r is sketched in Figure 6.1 below.

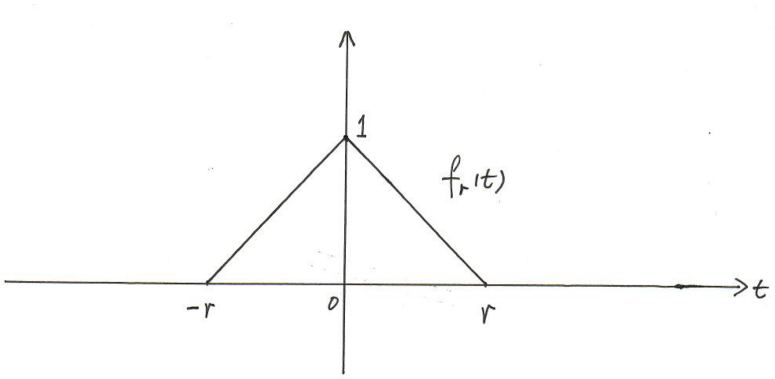


Figure 6.1: The characteristic function f_r

To illustrate the intuition better, in what follows we will avoid the use of equations and describe the constructions in geometric terms. For instance, one can rephrase f_r in the following way. Let A be the point $(0, 1)$ on the vertical axis (we always call this point A). For each given point B on the positive axis, the function whose graph is given by the polygon $\overline{AB\infty}$ (i.e. the segment $\overline{AB}$ plus the segment from B to ∞ along the positive axis) is the characteristic function f_r with $r = B$.

We denote $\mathcal{C} = \{f_r : r > 0\}$ to be the class of characteristic functions provided by this example.

A slightly more complicated example

As the next step, let us consider a slightly more complicated situation. The graph of the function $f(t)$ we shall consider in this example is given by the polygon $\overline{ABC\infty}$ as illustrated in Figure 6.2.

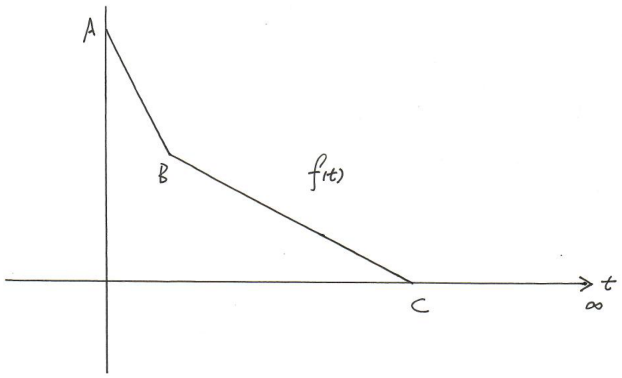


Figure 6.2: The function $f(t)$

We claim that $f(t)$ is a characteristic function.The idea is to show that f is the convex combination of two characteristic functions from the above class $\mathcal{C}$. This is illustrated by Figure 6.3 below.

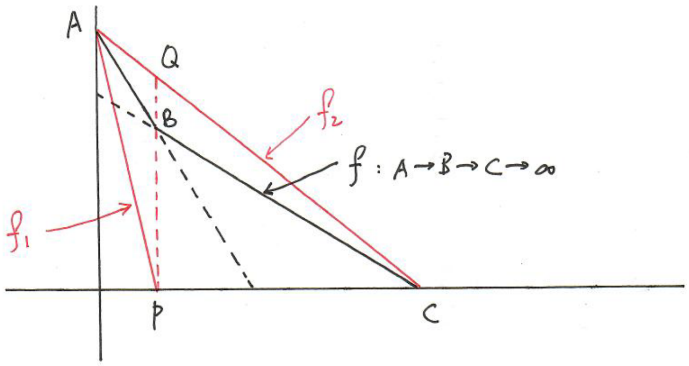


Figure 6.3: $f(t)$ as a convex combination of f_1, f_2

The two red segments $\overline{AP}$ and $\overline{AC}$ correspond to two characteristic functions $f_1, f_2 \in \mathcal{C}$ respectively. It is useful to note that the points B and C can be regarded as the "turning points" for the function f. The functions f_1, f_2 are associated with these two "turning points" in a natural way. More precisely, they are defined uniquely in terms of the t-coordinates of B and C respectively.

We look at the triangles ΔAQP and ΔCQP separately, which corresponds to the first two pieces of $f(t)$ respectively. In the triangle ΔAQP, the segment $\overline{AB}$

correspond to the function f, while the segments $\overline{AP}$ and $\overline{AQ}$ correspond to f_1 and f_2 respectively. From the relation among these three segments, it is apparent that

$$f = \lambda f_1 + (1-\lambda) f_2$$

on this part, where $\lambda \triangleq \frac{|QB|}{|QP|}$.This is merely a consequence of the relation that

$$B = \lambda \cdot P + (1-\lambda) \cdot Q \tag{6.24}$$

in terms of coordinates.

In the triangle ΔCQP, the segment $\overline{BC}$ corresponds to f, while the segments $\overline{PC}$ and $\overline{QC}$ correspond to f_1 and f_2 respectively. In view of the relation (6.24) on the segment $\overline{QP}$, one has exactly the same relation

$$f = \lambda f_1 + (1-\lambda) f_2$$

for this part!

Note that this relation holds trivially on the part $\overline{C\infty}$. Therefore, one concludes that f is a characteristic function (as the convex combination of two characteristic functions).

An even more complicated construction

To prove Pólya's theorem, one has to consider an even more complicated situation. The analysis developed in this example will provide the core ingredient of the entire proof. The function $f(t)$ we shall consider here is constructed by adding one extra piece to the last example. In view of Figure 6.4 below, the graph of $f(t)$ is given by the polygon $\overline{ABDE\infty}$. To compare $f(t)$ with the previous example, one has two extra points $D \in \overline{BC}$ and $E \in \overline{C\infty}$. The new function $f(t)$ is identical with the earlier one on the parts $\overline{AB}$ and $\overline{BD}$. On the part from D to E, the previous function (given by $\overline{DCE}$) is changed to the new function $f(t)$ (given by $\overline{DE}$). We claim that $f(t)$ is a characteristic function.

To this end, the main idea is to show that $f(t)$ is now a convex combination of three functions from the class $\mathcal{C}$. These three functions are marked by red segments and are denoted as f_1, f_2, f_3 respectively (cf. Figure 6.5 below). In a similar way as before, B, D, E are the three "turning points" of f, and the three functions $f_1, f_2, f_3 \in \mathcal{C}$ are associated with these three points B, D, E in the sense that they are determined by the t-coordinates of these points respectively. However, directly showing that f is a convex combination of f_1, f_2, f_3 is too complicated and not inspiring. We shall make use of what we have already obtained in the previous

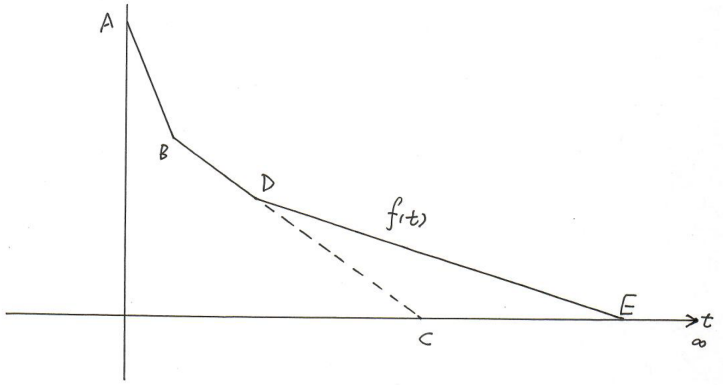


Figure 6.4: The function $f(t)$

example and apply an induction argument. This will also lead to a complete proof of Pólya's theorem without much extra effort!

To make the connection clear, let us denote the current function $f(t)$ by $f^{(3)}(t)$ and let $f^{(2)}(t)$ be the function defined in the previous example (in Figure 6.5 below, $f^{(2)}(t) = \overline{ABC\infty}$). In the previous example, we have seen that

$$f^{(2)} = \lambda_1 \cdot f_1 + \lambda_2 \cdot \tilde{f}_2 \tag{6.25}$$

with some $\lambda_1, \lambda_2 > 0$ and $\lambda_1 + \lambda_2 = 1$. Here the function $\tilde{f}_2 \in \mathcal{C}$ in Figure 6.5 is the function f_2 in Figure 6.3 of the previous example. Note that $f^{(3)} = f^{(2)}$ on $\overline{ABD}$. Therefore, one has

$$f^{(3)} = \lambda_1 \cdot f_1 + \lambda_2 \cdot \tilde{f}_2$$

on the part $\overline{ABD}$.

The next observation is that, in the triangle ΔAPQ one easily expresses $\tilde{f}_2$ as a convex combination of f_2 and f_3. To be more precise, let $\alpha \triangleq \frac{|RP|}{|QP|}$. Then one has

$$\tilde{f}_2 = (1-\alpha) \cdot f_2 + \alpha \cdot f_3$$

on the part up to the point D (in the sense of t-coordinate). As a consequence, for the same part one has

$$f^{(3)} = \lambda_1 \cdot f_1 + \lambda_2(1-\alpha) \cdot f_2 + \lambda_2 \alpha \cdot f_3. \tag{6.26}$$

Note that this is a convex combination of f_1, f_2, f_3 (though for the moment this is only true up to the point D).

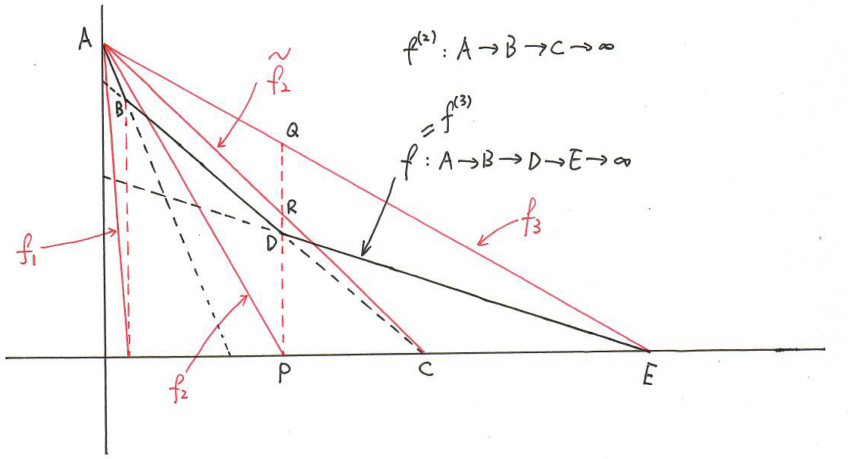


Figure 6.5: $f(t)$ as a convex combination of f_1, f_2, f_3

It remains to show that the relation (6.26) is also valid on the part $\overline{DE}$. Since $f_1 = f_2 = 0$ on this part, it reduces to showing that

$$f^{(3)} = \lambda_2 \alpha \cdot f_3 \tag{6.27}$$

on $\overline{DE}$. To this end, one first looks at the segment $\overline{DC}$. This segment is from the graph of $f^{(2)}$. Since $f_1 = 0$ on this part, the relation (6.25) for $f^{(2)}$ becomes

$$f^{(2)} = \lambda_2 \cdot \tilde{f}_2.$$

In view of the triangle ΔCRP, this relation yields $\lambda_2 = \frac{|DP|}{|RP|}$. Now if one looks at the triangle ΔEQP, on the part $\overline{DE}$ one has

$$f^{(3)} = \frac{|DP|}{|QP|} \cdot f_3 = \frac{|DP|}{|RP|} \cdot \frac{|RP|}{|QP|} \cdot f_3 = \lambda_2 \alpha \cdot f_3.$$

This gives the desired relation (6.27).

Note that (6.26) is trivial on $\overline{E\infty}$. Therefore, one concludes that $f(t) = f^{(3)}(t)$ is a convex combination of the characteristic functions f_1, f_2, f_3. It follows that $f(t)$ is a characteristic function.

Generalising the construction to arbitrarily many pieces

One can continue the previous construction by adding in more pieces. To be precise, if one picks two arbitrary points $F \in \overline{DE}$, $G \in \overline{E\infty}$ and create a new

segment $\overline{FG}$, one obtains a new function $f^{(4)}$ from $f^{(3)}$ whose graph is given by $\overline{ABDFG\infty}$. One can keep adding new pieces to obtain $f^{(5)}$ from $f^{(4)}$, and $f^{(6)}$ from $f^{(5)}$ etc. in a similar way. Note that these functions are not arbitrary piecewise linear functions. The fact that the slope gets larger when passing through each "turning point" is an important feature of the construction. Let $\mathcal{D}$ denote the family of functions that can be constructed from this manner.

We claim that these functions $f^{(m)}$ are all characteristic functions. The crucial observation is that the previous argument has an inductive nature and therefore it requires almost no extra effort to treat this general $f^{(m)}$. To convey the idea clearly, let $f^{(m)} \in \mathcal{D}$ be a function whose graph is given by $\overline{A_0A_1\cdots A_m\infty}$ where $A_0 = (0,1)$, $A_1\cdots, A_{m-1}$ all lie above the t-axis and A_m lies on the t-axis. The points $A_1, \cdots, A_m$ are "turning points" through which the slopes of the segments $\overline{A_{i-1}A_i}$ get increased. For each $1 \leqslant i \leqslant m$, let $f_i \in \mathcal{C}$ be the characteristic function associated with the "turning point" A_i. More precisely, letting t_i be the t-coordinate of A_i, one has $f_i \triangleq f_{t_i}$ where f_{t_i} is the function defined by (6.27) and is thus in class $\mathcal{C}$. We propose the following induction hypothesis.

Induction hypothesis: $f^{(m)}$ is a convex combination of $f_1, \cdots, f_m$.

To develop the inductive step, let $f^{(m+1)}$ be a function obtained from $f^{(m)}$ in the following way. Let $A'_m \in \overline{A_{m-1}A_m}$ and $A'_{m+1} \in \overline{A_m\infty}$. Then $f^{(m+1)} = f^{(m)}$ from A_0 up to the point A'_m. On the part after A'_m, the graph of $f^{(m+1)}$ is given by $\overline{A'_mA'_{m+1}\infty}$. Let $\tilde{f}_m, \tilde{f}_{m+1} \in \mathcal{C}$ be the characteristic functions associated with the "turning points" A'_m, A'_{m+1} respectively. We need to show that $f^{(m+1)}$ is a convex combination of $f_1, \cdots f_{m-1}, \tilde{f}_m, \tilde{f}_{m+1}$.

By the induction hypothesis, one knows that

$$f^{(m)} = \lambda_1 f_1 + \cdots + \lambda_{m-1} f_{m-1} + \lambda_m f_m$$

where $\lambda_1, \cdots, \lambda_m$ are positive numbers such that $\lambda_1 + \cdots + \lambda_m = 1$. One is now in exactly the same situation as in the last example. Figure 6.6 describes the main geometric intuition. In the figure, one views B as A_{m-1}, D as A'_m, C as A_m, E as A'_{m+1}, and replaces the line segment $\overline{AB}$ by $\overline{A_0\cdots A_{m-1}}$. This does not change the argument at all. In the same way as in the last example, we set $\alpha \triangleq \frac{|RP|}{|QP|}$. Then

$$f_m = (1-\alpha)\tilde{f}_m + \alpha\tilde{f}_{m+1}$$

on the part from A_0 to A'_m. In particular, one has

$$f^{(m+1)} = f^{(m)} = (\lambda_1 f_1 + \cdots + \lambda_{m-1} f_{m-1}) + \lambda_m(1-\alpha)\tilde{f}_m + \lambda_m\alpha\tilde{f}_{m+1} \qquad (6.28)$$

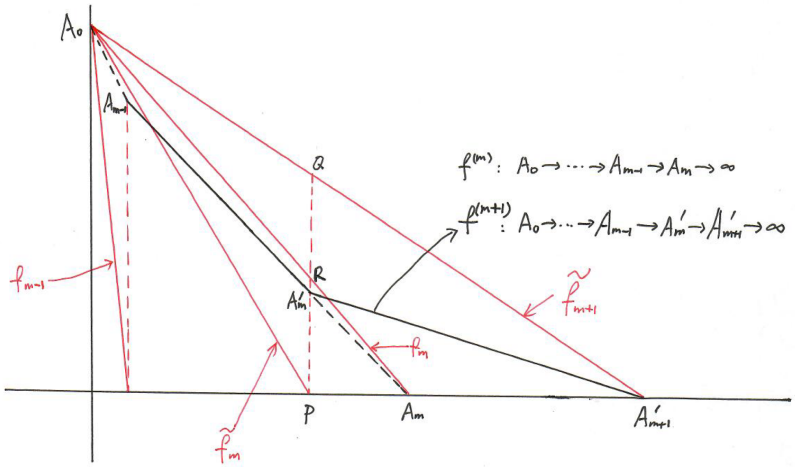


Figure 6.6: The induction step

for this part. Note that on the part $\overline{A_{m-1}A_m}$ one has $f_1 = \cdots = f_{m-1} = 0$ and thus $f^{(m)} = \lambda_m f_m$ on this part. Using the triangle $\Delta A_m RP$, one sees that

$$\lambda_m = \frac{|A'_m P|}{|RP|}.$$

By exactly the same argument as in the last example, one finds that

$$f^{(m+1)} = \lambda_m \alpha \tilde{f}_{m+1}$$

on the part $\overline{A'_m A'_{m+1}}$. This relation is equivalent to (6.28) since

$$f_1 = \cdots = f_{m-1} = \tilde{f}_m = 0$$

on this part. Therefore, the convex linear relation (6.28) holds for $f^{(m+1)}$ for all t. This concludes the induction step.

Proof of Theorem 6.6

Now we are able to give another independent proof of Pólya's theorem. The main idea is to approximate $f(t)$ by using the functions $f^{(m)} \in \mathcal{D}$ constructed previously. Intuitively, these functions will be piecewise linear interpolations of $f(t)$.

Let $m \geqslant 1$ and let

$$\mathcal{P}_m : 0 = t_0 < t_1 < \cdots < t_{k_m} = m$$

be a finite partition of $[0, m]$ such that

$$\text{mesh}(\mathcal{P}_m) \triangleq \max_{1 \leqslant i \leqslant k_m} |t_i - t_{i-1}| \to 0$$

as $m \to \infty$. We define an approximating function $\varphi^{(m)}(t)$ in the following way. On the interval $[0, t_{k_m}]$, $\varphi^{(m)}(t)$ is the linear interpolation of $f(t)$ over the partition $\mathcal{P}_m$, i.e.

$$\varphi^{(m)}(t_i) = f(t_i), \quad t_i \in \mathcal{P}_m,$$

and $\varphi^{(m)}(t)$ is linear on each sub-interval $[t_{i-1}, t_i]$. It is clear that this part of $\varphi^{(m)}$ is given by a polygon $\overline{A_0 A_1 \cdots A_{k_m}}$ where $A_i = (t_i, f(t_i))$. To define the remaining part of $\varphi^{(m)}$, we extend the last piece $\overrightarrow{A_{k_{m-1}} A_{k_m}}$ until it meets the t-axis at a point A'_{k_m}. The part of $\varphi^{(m)}$ on the interval $[t_{k_m}, \infty)$ will be given by the graph $\overline{A_{k_m} A'_{k_m} \infty}$. Since $f(t)$ is decreasing and convex, one sees that $\varphi^{(m)} \in \mathcal{D}$ and they are thus characteristic functions. It is routine to check that

$$\lim_{m \to \infty} \varphi^{(m)}(t) = f(t), \quad \text{for every } t \in \mathbb{R}.$$

It follows from the Lévy-Cramér theorem that $f(t)$ is a characteristic function.

The proof of Theorem 6.6 is now complete.

Comparison with the short / ingenious proof

To relate the current proof with the one given in Section 6.5, recall from (6.17) that $f(t)$ admits the following integral representation:

$$f(t) = \int_{(0,\infty)} f_r(t) \nu(dr), \tag{6.29}$$

where $d\nu \triangleq r df'_+(r)$. Using this ingenious formula, one can easily express the earlier $\varphi^{(m)}$ (more generally, $f^{(m)} \in \mathcal{D}$) as a convex combination of members in $\mathcal{C}$. Let us again consider $f^{(m)}$ given by $\overline{A_0 A_1 \cdots A_m \infty}$ as in Figure 6.6. Let t_i be the t-coordinate of A_i and let $a_i > 0$ be the change of slope from $\overline{A_{i-1} A_i}$ to $\overline{A_i A_{i+1}}$ $(1 \leqslant i \leqslant m)$ where $A_{m+1} \triangleq \infty$. By explicit calculation, one finds that

$$\nu = \sum_{i=1}^{m} a_i t_i \delta_{t_i}.$$

It follows from the general formula (6.29) that

$$f^{(m)}(t) = \sum_{i=1}^{m} a_i t_i f_{t_i}(t),$$

which is of no surprise a convex combination of $f_{t_1}, \cdots, f_{t_m}$.

7 The central limit theorem

The LLN asserts that for i.i.d. random variables, the sample average $\bar{S}_n$ stabilises at its theoretical mean $\bar{x}$ as $n \to \infty$. The LDP (Cramér's theorem) describes an associated concentration of measure phenomenon: the law of $\bar{S}_n$ concentrates at the Dirac delta mass $\delta_{\bar{x}}$ exponentially fast. The *central limit theorem* (CLT), on the other hand, quantifies the rate of convergence for the LLN. Roughly speaking, it describes the behaviour that

$$\bar{S}_n - \bar{x} \approx \frac{1}{\sqrt{n}} \sigma Z$$

where $\sigma^2 = \mathrm{Var}[X_1]$ and Z is standard normal (one needs to be very careful about the interpretation of such a statement though!). Another way of interpreting the CLT is that the fluctuation of the partial sum of an i.i.d. sequence around its mean is asymptotically Gaussian. It is a rather striking fact that such a behaviour is *universal* for a wide class of models; it arises as along as the contribution of each random individual is "small" and the dependence among different individuals is "weak". In addition, the particular distribution of each individual is of little relevance and one ends up with a canonical Gaussian limit. Of course such statements are quite vague and of no mathematical precision at this point. The mathematics behind such a phenomenon as well as the appearance of the Gaussian nature is rather deep.

In this chapter, we develop some insights into the hidden mechanism behind the CLT from several perspectives. In Section 7.1, we recapture the classical CLT for i.i.d. sequences from the viewpoints of characteristic functions and moments. Such a theorem is qualitative and only gives weak convergence without any quantitative rate of convergence. In Section 7.2, we prove Lindeberg's CLT which extend the classical CLT to a more general context (still independent but not necessarily identically distributed). In particular, we introduce a different method that describes the weak convergence in a more quantitative form. In Section 7.3, we use an explicit example to illustrate the possibility of having non-Gaussian limit even in the i.i.d. case if the random variables have heavy tails. In Section 7.4, we introduce a powerful modern technique of establishing rates of convergence for distributional approximations: Stein's method. To illustrate the essential ideas, we only consider Gaussian approximations in the context of independent sequences.

7.1 The classical central limit theorem

We start by recapturing the classical central limit theorem (CLT) in the i.i.d. context. This fundamental result was due to J. Lindeberg and P. Lévy. Its proof is a typical application of characteristic functions.

Theorem 7.1. *Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence of random variables which has finite mean and variance. Then $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ converges weakly to the standard normal distribution as $n \to \infty$, where $S_n \triangleq X_1 + \cdots + X_n$.*

Proof. One may assume that $\mathbb{E}[X_1] = 0$, for otherwise one can consider the sequence $X_n - \mathbb{E}[X_n]$ instead. Let $f(t)$ be the characteristic function of X_1. Since X_1 has finite second moment, according to Corollary 6.3 one has

$$f(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

where $\sigma^2 \triangleq \mathrm{Var}[X_1]$. In addition, since $\{X_n\}$ is an i.i.d. sequence, it is easily seen that the characteristic function of $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ is given by

$$f_n(t) = \big(f\big(\frac{t}{\sigma\sqrt{n}}\big)\big)^n = \big(1 - \frac{t^2}{2n} + o\big(\frac{t^2}{n\sigma^2}\big)\big)^n.$$

Note that t is fixed here and the infinitesimal term $o(t^2/n\sigma^2)$ is understood in the limit as $n \to \infty$. By writing

$$c_n \triangleq -\frac{t^2}{2n} + o\big(\frac{t^2}{n\sigma^2}\big),$$

one finds that

$$f_n(t) = (1 + c_n)^{\frac{1}{c_n} \cdot n c_n} \to e^{-t^2/2}$$

as $n \to \infty$. The above limit is precisely the characteristic function of the standard normal distribution. According to the Lévy-Cramér continuity theorem, one concludes that

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} \to N(0, 1)$$

weakly as $n \to \infty$. This proves the classical CLT. □

The above proof, as the most standard one, is so simple that it has unfortunately concealed many of the deeper insights into this fundamental theorem.

The use of characteristic functions is somehow like a piece of magic, leaving the audience in shock after the play is over without telling the deeper truth of why.

The following alternative argument perhaps provides a bit more clues towards the hidden mechanism. In the first place, it is reasonable to believe that most of the common distributions one encounters in elementary probability are uniquely determined by the sequence of moments. The normal distribution is such an example (we will not prove this fact here). Recall that the moments of $Z \stackrel{d}{=} N(0,1)$ are given by

$$\mathbb{E}[Z^{2m-1}] = 0,\ \mathbb{E}[Z^{2m}] = (2m-1)\cdot(2m-3)\cdot\cdots\cdot 3\cdot 1$$

for each $m \geqslant 1$.

Let us compute moments of the quantity $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ correspondingly. For simplicity, we assume that $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}[X_1] = 1$ ($\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ becomes $S_n/\sqrt{n}$). To make use of the method of moments, we further assume that X_1 has finite moments of all orders. The lemma below provides the key reason for the weak convergence of $S_n/\sqrt{n}$ towards the standard normal distribution.

Lemma 7.1. *For each $m \geqslant 1$, one has*

$$\lim_{n\to\infty} \mathbb{E}\big[\big(\frac{S_n}{\sqrt{n}}\big)^m\big] = L_m,$$

where $L_m \triangleq \mathbb{E}[Z^m]$ is the m-th moment of the standard normal distribution.

Proof. We prove the claim by induction on m. The case when $m = 1$ is trivial. When $m = 2$, one has

$$\mathbb{E}[\big(\frac{S_n}{\sqrt{n}}\big)^2\big] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[X_j^2] = 1 = L_2.$$

Now suppose that the claim is true for general m. To examine the $(m+1)$-case, one first observes that

$$\begin{aligned}
\mathbb{E}[S_n^{m+1}] &= \mathbb{E}[(X_1+\cdots+X_n)S_n^m] = n\cdot\mathbb{E}[X_nS_n^m] \quad \text{(since } \{X_n\} \text{ are i.i.d.)} \\
&= n\cdot\mathbb{E}[X_n(X_n+S_{n-1})^m] = n\cdot\sum_{j=0}^{m}\binom{m}{j}\mathbb{E}[X_n^{j+1}]\mathbb{E}[S_{n-1}^{m-j}] \\
&= nm\cdot\mathbb{E}[S_{n-1}^{m-1}] + n\cdot\sum_{j=2}^{m}\binom{m}{j}\mathbb{E}[X_n^{j+1}]\mathbb{E}[S_{n-1}^{m-j}], \qquad (7.1)
\end{aligned}$$

where we used $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^2] = 1$ to reach the last equality.

We now take into account the $\sqrt{n}$-normalisation. To simplify the notation, we set

$$L_m(n) \triangleq \mathbb{E}\big[\big(\frac{S_n}{\sqrt{n}}\big)^m\big],\ C_j \triangleq \mathbb{E}[X_n^{j+1}].$$

It follows from (7.1) that

$$\begin{aligned} L_{m+1}(n) = mL_{m-1}(n-1)\cdot\big(\frac{n-1}{n}\big)^{\frac{m-1}{2}} \\ + \sum_{j=2}^{m}\binom{m}{j} C_j L_{m-j}(n-1)\cdot\frac{(n-1)^{(m-j)/2}}{n^{(m-1)/2}}. \end{aligned}$$

According to the induction hypothesis and the simple observation that (for $j \geqslant 2$)

$$\lim_{n\to\infty}\frac{(n-1)^{(m-j)/2}}{n^{(m-1)/2}} = 0,$$

one finds that $L_{m+1}(n) \to mL_{m-1}$ as $n \to \infty$. This not only shows the convergence of $L_{m+1}(n)$, but more importantly its convergence to the correct limit

$$mL_{m-1} = L_m,$$

which is precisely the relation satisfied by the moments of $N(0,1)$. This completes the proof of the lemma. □

Due to Lemma 7.1, it becomes reasonable to expect that the CLT holds true (i.e. $S_n/\sqrt{n}$ converges weakly to $N(0,1)$). Technically, there is still a missing step in the above proof, i.e. *why the convergence of moments implies the weak convergence of distributions*? This question falls into the scope of the so-called *moment problem* which typically studies the relation between moments and distributions. We will not delve deeper into this direction and refer the interested reader to [Bil86] for a discussion.

An application: Stirling's formula

We discuss an enlightening application of the classical CLT: *Stirling's formula*. We used such a formula when we studied the random walk in Proposition 5.1

Proposition 7.1. *One has the following asymptotic equivalence:*

$$\lim_{n\to\infty}\frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1. \tag{7.2}$$

We fist provide a heuristic argument which explains the key point of the proof. Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and Poisson random variables with parameter 1. Define $S_n \triangleq X_1 + \cdots + X_n$. Then one has

$$\mathbb{P}(S_n = n) = \mathbb{P}(n-1 < S_n \leqslant n) = \mathbb{P}\big(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leqslant 0\big).$$

By the CLT, $\frac{S_n - n}{\sqrt{n}} \to N(0,1)$ weakly. In particular,

$$\mathbb{P}(S_n = n) \approx \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx$$

when n is large. Note that

$$\mathbb{P}(S_n = n) = \frac{n^n e^{-n}}{n!}$$

since $S_n \stackrel{d}{=} \text{Poisson}(n)$, and one also has

$$\int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx \approx \frac{1}{\sqrt{n}}.$$

It follows that

$$\frac{n^n e^{-n}}{n!} \approx \frac{1}{\sqrt{2\pi n}}$$

which is precisely the Stirling approximation (7.2). However, this argument is not entirely rigorous; indeed, the step

$$\mathbb{P}\big(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leqslant 0\big) \approx \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx$$

is by no mean a simple consequence of the CLT as one also varies the end point of the interval here.

To give a rigorous treatment, we consider a sequence $\{X_n : n \geqslant 1\}$ of independent and exponential random variables with parameter 1. Note that $S_n \triangleq X_1 + \cdots + X_n$ follows a Gamma distribution with parameters n and 1. Using the explicit formula for the Gamma density, it is plain to check that

$$\begin{aligned} &\mathbb{P}\big(0 \leqslant \frac{S_{n+1} - (n+1)}{\sqrt{n+1}} \leqslant 1\big) \\ &= \frac{\sqrt{n+1}}{n!} \int_0^1 (\sqrt{n+1} \cdot (x + \sqrt{n+1}))^n e^{-\sqrt{n+1}(x+\sqrt{n+1})} dx. \end{aligned} \tag{7.3}$$

In the first place, according to the CLT one has

$$\lim_{n\to\infty} \mathbb{P}\big(0 \leqslant \frac{S_{n+1}-(n+1)}{\sqrt{n+1}} \leqslant 1\big) = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2} dx. \tag{7.4}$$

On the other hand, by applying two steps of change of variables

$$y = \sqrt{n+1}(x+\sqrt{n+1}),\ z = \frac{y-n}{\sqrt{n}}$$

to the right hand side of (7.3), one is led to

$$\begin{aligned}\mathbb{P}\big(0 \leqslant \frac{S_{n+1}-(n+1)}{\sqrt{n+1}} \leqslant 1\big) &= \frac{1}{n!} \int_{1+n}^{1+n+\sqrt{n+1}} y^n e^{-y} dy \\ &= \frac{\sqrt{n} n^n e^{-n}}{n!} \int_{\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}+\sqrt{1+\frac{1}{n}}} \big(1+\frac{z}{\sqrt{n}}\big)^n e^{-\sqrt{n}z} dz.\end{aligned}$$

Note that

$$\big(1+\frac{z}{\sqrt{n}}\big)^n = \exp\big(n \log\big(1+\frac{z}{\sqrt{n}}\big)\big) = \exp\big(n \cdot \big(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + o(\frac{1}{n})\big)\big),$$

which implies

$$\lim_{n\to\infty} \big(1+\frac{z}{\sqrt{n}}\big)^n e^{-\sqrt{n}z} = e^{-z^2/2}.$$

It follows that

$$\mathbb{P}\big(0 \leqslant \frac{S_{n+1}-(n+1)}{\sqrt{n+1}} \leqslant 1\big) \sim \frac{\sqrt{n} n^n e^{-n}}{n!} \int_0^1 e^{-z^2/2} dz. \tag{7.5}$$

Stirling's formula (7.2) thus follows by comparing (7.4) and (7.5).

7.2 Lindeberg's central limit theorem

There are at least two reasons for push our understanding of the CLT further. The first reason is that in the classical CLT, we have made the restrictive assumption that the sequence $\{X_n\}$ is i.i.d . The two proofs given in the last section make use of this condition in a crucial way. However, the i.i.d. assumption is not strictly essential for general CLTs. One needs to understand the deeper mechanism leading to such a phenomenon. The second reason is that the previous proofs are only *qualitative*; it does not contain any information about how close the distribution

of $\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}}$ is to the standard normal for each given n. For practical purposes it is necessary and important to develop robust tools for studying the rate of convergence in the CLT.

Lindeberg's CLT provides some deeper insights into the above two aspects. For the first aspect, it suggests that some sort of "uniform negligibility of each summand X_m $(1 \leqslant m \leqslant n)$ with respect to S_n" is essential for the CLT to hold. For the second aspect, recall that a sequence of probability measures μ_n on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ converges weakly to μ if and only if

$$\int_{\mathbb{R}} f(x)\mu_n(dx) \to \int_{\mathbb{R}} f(x)\mu(dx) \quad \forall f \in \mathcal{C}_b(\mathbb{R}).$$

In this spirit, a natural way of comparing the "distance" between μ_n and μ is to quantitatively estimate the distance $\left| \int_{\mathbb{R}} f d\mu_n - \int_{\mathbb{R}} f d\mu \right|$ for each f within a suitable class of functions. In the CLT context, one is thus led to estimating the distance

$$\left| \mathbb{E}\big[f\big(\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} \big) \big] - \mathbb{E}[f(Z)] \right| \quad \text{for suitable class of functions } f,$$

where $Z \overset{d}{=} N(0,1)$. Lindeberg's CLT gives an answer to questions of such kind.

Before stating the theorem, we first present the basic set-up. We again consider a sequence $\{X_n : n \geqslant 1\}$ of *independent* (but not necessarily identically distributed!) random variables with finite mean and variance. We assume that $\mathbb{E}[X_n] = 0$, for otherwise one can always centralise the sequence to have mean zero. For each $n \geqslant 1$, let us set

$$\sigma_n \triangleq \sqrt{\mathrm{Var}[X_n]}, \ \Sigma_n \triangleq \sqrt{\mathrm{Var}[S_n]}, \ \hat{S}_n \triangleq \frac{S_n}{\Sigma_n}.$$

We introduce two key quantities that will appear in the rate of convergence estimate:

$$r_n \triangleq \max_{1 \leqslant m \leqslant n} \frac{\sigma_m}{\Sigma_n} \tag{7.6}$$

and

$$g_n(\varepsilon) \triangleq \frac{1}{\Sigma_n^2} \sum_{m=1}^{n} \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon \Sigma_n], \quad \varepsilon > 0.$$

Vaguely speaking, these two quantities reflect the relative magnitude of each summand X_m $(1 \leqslant m \leqslant n)$ with respect to S_n. We also recall the notation $\|f\|_\infty \triangleq \sup_{x \in \mathbb{R}} |f(x)|$ for a given function $f : \mathbb{R} \to \mathbb{R}$.

Now we are able to state Lindeberg's CLT. In many ways it is deeper than the classical CLT in the last section. We denote $\mathcal{C}_b^3(\mathbb{R})$ as the space of functions $f : \mathbb{R} \to \mathbb{R}$ that are continuously differentiable with bounded derivatives up to order three.

Theorem 7.2. *Under the aforementioned set-up, let $f \in \mathcal{C}_b^3(\mathbb{R})$. Then for each $\varepsilon > 0$ and $n \geqslant 1$, one has*

$$\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(Z)]\big| \leqslant \big(\frac{\varepsilon}{6} + \frac{\gamma \cdot r_n}{6}\big)\|f'''\|_\infty + g_n(\varepsilon) \cdot \|f''\|_\infty, \tag{7.7}$$

where $Z \overset{d}{=} N(0,1)$ and $\gamma \triangleq \mathbb{E}[|Z|^3] = \sqrt{8/\pi}$ is the third absolute moment of Z. In addition, if

$$\lim_{n\to\infty} g_n(\varepsilon) = 0 \quad \text{for every } \varepsilon > 0, \tag{7.8}$$

then one has

$$\hat{S}_n \to Z \quad \text{weakly}$$

as $n \to \infty$, hence giving the CLT for $\{X_n\}$.

The condition (7.8) is known as *Lindeberg's condition*. Theorem 7.2 therefore indicates that Lindeberg's condition implies a CLT in the context of independent random variables with finite mean and variance. As a direct corollary, one recovers the CLT. Indeed, for an i.i.d. sequence $\{X_n : n \geqslant 1\}$ one has $\Sigma_n = \sqrt{n}\sigma$ ($\sigma^2 \triangleq \mathrm{Var}[X_1]$) and thus

$$\begin{aligned} g_n(\varepsilon) &= \frac{1}{n\sigma^2}\sum_{m=1}^{n} \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon\sqrt{n}\sigma] \\ &= \frac{1}{\sigma^2}\mathbb{E}[X_1^2; |X_1| \geqslant \varepsilon\sqrt{n}\sigma], \end{aligned}$$

which vanishes as $n \to \infty$. In particular, Lindeberg's condition holds. Another interesting corollary of Lindeberg's theorem is the following *Liapunov's CLT*.

Corollary 7.1. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables with mean zero and finite third moment. Define S_n, Σ_n as before and we also set*

$$\Gamma_n \triangleq \sum_{m=1}^{n} \mathbb{E}[|X_m|^3].$$

Suppose that $\Gamma_n/\Sigma_n^3 \to 0$. Then S_n/Σ_n converges weakly to $N(0,1)$.

Proof. One verifies Lindeberg's condition by using Chebyshev's inequality:

$$g_n(\varepsilon) = \frac{1}{\Sigma_n^2} \sum_{m=1}^{n} \mathbb{E}[X_m^2; |X_m| \geqslant \varepsilon \Sigma_n] \leqslant \frac{1}{\varepsilon \Sigma_n^3} \sum_{m=1}^{n} \mathbb{E}[|X_m|^3] = \frac{\Gamma_n}{\varepsilon \Sigma_n^3} \to 0.$$

□

Remark 7.1. Liapunov's CLT can also be derived using the method of characteristic functions (third-order Taylor expansion for the characteristic function).

The rest of this section is devoted to the proof of Lindeberg's CLT.

Proof of Theorem 7.2

We first establish the quantitative estimate (7.7) and then show how it leads to the weak convergence property for the CLT.

The quantitative estimate.

Let $n \geqslant 1$ be given fixed. For each $1 \leqslant m \leqslant n$, we define $\hat{X}_m \triangleq X_m/\Sigma_n$ so that

$$\hat{S}_n = \hat{X}_1 + \cdots + \hat{X}_n.$$

The main idea of the proof is to swap each $\hat{X}_m$ to a reference normal random variable $\hat{Y}_m$ (one flip at each step) in a way that after n swaps the accumulated error is controllable.

Step one: Introducing the reference normal random variables. To implement the swapping idea mathematically, we first assume that there are n standard normal random variables $Y_1, \cdots, Y_n$ along with the X_i's being defined on the same probability space and the random variables

$$X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_n$$

are all independent. This is always possible by using product spaces (why?). We then set

$$\hat{Y}_m \triangleq \frac{\sigma_m Y_m}{\Sigma_n}, \quad 1 \leqslant m \leqslant n,$$

and

$$\hat{T}_n \triangleq \hat{Y}_1 + \cdots + \hat{Y}_n.$$

Observe that $\hat{Y}_m$ is Gaussian with mean zero and the same variance as $\hat{X}_m$'s. In addition, $\hat{T}_n \overset{d}{=} N(0,1)$. The problem is essentially reduced to estimating the quantity

$$\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)]\big|,$$

where $f \in \mathcal{C}_b^3(\mathbb{R})$ is a given fixed test function.

Step two: Forming the telescoping sum. We estimate the above quantity by consecutively flipping $\hat{X}_m$ to $\hat{Y}_m$ (one at each time). More precisely, one writes

$$\begin{aligned}
&\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)] \\
&= \mathbb{E}[f(\hat{X}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \hat{X}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{X}_3 + \cdots + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \hat{Y}_2 + \hat{Y}_3 + \cdots + \hat{X}_n)] \\
&\quad \cdots \\
&\quad + \mathbb{E}[f(\hat{Y}_1 + \cdots \hat{Y}_{n-1} + \hat{X}_n)] - \mathbb{E}[f(\hat{Y}_1 + \cdots + \hat{Y}_{n-1} + \hat{Y}_n)]. \qquad (7.9)
\end{aligned}$$

To rewrite the expression in a more enlightening form, let us introduce for $1 \leqslant m \leqslant n$,

$$U_m \triangleq \hat{Y}_1 + \cdots + \hat{Y}_{m-1} + \hat{X}_{m+1} + \cdots + \hat{X}_n.$$

Then (7.9) can be expressed as

$$\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)] = \sum_{m=1}^{n} \big(\mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)]\big).$$

Step three: Introducing the Taylor approximation. We now use Taylor's approximation for the function f to estimate the quantity

$$\big|\mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)]\big|.$$

For this purpose, let us define

$$R_m(\xi) \triangleq f(U_m + \xi) - f(U_m) - f'(U_m)\xi - \frac{f''(U_m)}{2}\xi^2, \quad \xi \in \mathbb{R}.$$

This is the remainder for the second-order Taylor expansion of f around U_m. Since $U_m, \hat{X}_m, \hat{Y}_m$ are independent and $\hat{X}_m, \hat{Y}_m$ have the same mean and variance, one sees that

$$\mathbb{E}[f(U_m + \hat{X}_m)] - \mathbb{E}[f(U_m + \hat{Y}_m)] = \mathbb{E}[R_m(\hat{X}_m)] - \mathbb{E}[R_m(\hat{Y}_m)].$$

It follows that

$$\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(\hat{T}_n)]\big| \leqslant \sum_{m=1}^{n} \big|\mathbb{E}[R_m(\hat{X}_m)]\big| + \sum_{m=1}^{n} \big|\mathbb{E}[R_m(\hat{Y}_m)]\big|. \qquad (7.10)$$

Step four: Estimating the $\mathbb{E}[R_m(\hat{X}_m)]$- and $\mathbb{E}[R_m(\hat{Y}_m)]$-sums separately. We estimate the right hand side of (7.10). First of all, by using a third-order Taylor expansion of f one has

$$|R_m(\xi)| \leqslant \frac{1}{3!}\|f'''\|_\infty |\xi|^3, \tag{7.11}$$

In addition, the second-order Taylor expansion gives

$$|f(U_m+\xi) - f(U_m) - f'(U_m)\xi| \leqslant \frac{1}{2}\|f''\|_\infty |\xi|^2.$$

As a result, one also has

$$|R_m(\xi)| \leqslant \frac{1}{2}\|f''\|_\infty |\xi|^2 + \frac{1}{2}|f''(U_m)|\cdot|\xi|^2 \leqslant \|f''\|_\infty |\xi|^2. \tag{7.12}$$

We use (7.11) to estimate the $\mathbb{E}[R_m(\hat{Y}_m)]$-sum as follows:

$$\begin{aligned}\sum_{m=1}^n \big|\mathbb{E}[R_m(\hat{Y}_m)]\big| &\leqslant \frac{1}{6}\|f'''\|_\infty \sum_{m=1}^n \mathbb{E}[|\hat{Y}_m|^3] = \frac{\gamma}{6}\|f'''\|_\infty \sum_{m=1}^n \frac{\sigma_m^3}{\Sigma_n^3} \\ &\leqslant \frac{\gamma}{6}\|f'''\|_\infty \cdot \frac{\max_{1\leqslant m\leqslant n}\sigma_m}{\Sigma_n}\cdot \sum_{m=1}^n \frac{\sigma_m^2}{\Sigma_n^2} = \frac{\gamma}{6}\|f'''\|_\infty \cdot r_n, \end{aligned} \tag{7.13}$$

where r_n is defined in (7.6) and $\gamma \triangleq \mathbb{E}[|Y_1|^3] = \sqrt{8/\pi}$ is the third absolute moment of $N(0,1)$.

The estimation of the $\mathbb{E}[R_m(\hat{X}_m)]$-sum is a bit more involved. One needs to decompose the region of integration into two parts:

$$\begin{aligned}&\sum_{m=1}^n \big|\mathbb{E}[R_m(\hat{X}_m)]\big| \\ &= \sum_{m=1}^n \big|\mathbb{E}[R_m(\hat{X}_m); |\hat{X}_m| < \varepsilon]\big| + \sum_{m=1}^n \big|\mathbb{E}[R_m(\hat{X}_m); |\hat{X}_m| \geqslant \varepsilon]\big| \\ &\leqslant \frac{\|f'''\|_\infty}{6}\sum_{m=1}^n \mathbb{E}[|\hat{X}_m|^3; |\hat{X}_m| < \varepsilon] + \|f''\|_\infty \sum_{m=1}^n \mathbb{E}[|\hat{X}_m|^2; |\hat{X}_m| \geqslant \varepsilon] \\ &\leqslant \frac{\|f'''\|_\infty \varepsilon}{6}\sum_{m=1}^n \frac{\sigma_m^2}{\Sigma_n^2} + \|f''\|_\infty g_n(\varepsilon) \\ &= \frac{\varepsilon}{6}\|f'''\|_\infty + g_n(\varepsilon)\|f''\|_\infty, \end{aligned} \tag{7.14}$$

where we used (7.11) and (7.12) to estimate the two parts respectively.

The desired estimate (7.7) is now a consequence of (7.13) and (7.14).

Proving weak convergence in the CLT.

Finally, we show that Lindeberg's condition (7.8) implies that

$$\hat{S}_n \to N(0,1) \quad \text{weakly.}$$

To this end, let m be the integer at which the maximum in (7.6) is attained, i.e. $r_n = \sigma_m/\Sigma_n$. It follows that

$$\begin{aligned} r_n^2 = \frac{\sigma_m^2}{\Sigma_n^2} &= \mathbb{E}[\hat{X}_m^2] \\ &= \mathbb{E}[\hat{X}_m^2; |\hat{X}_m| < \varepsilon] + \mathbb{E}[\hat{X}_m^2 : |\hat{X}_m| \geqslant \varepsilon] \\ &\leqslant \varepsilon^2 + g_n(\varepsilon), \end{aligned}$$

for every $\varepsilon > 0$. In particular, if Lindeberg's condition (7.8) holds, then $r_n \to 0$. According to (7.7), one has

$$\mathbb{E}[f(\hat{S}_n)] \to \mathbb{E}[f(Z)] \quad \text{for every } f \in \mathcal{C}_b^3(\mathbb{R}),$$

where $Z \overset{d}{=} N(0,1)$.

In order to establish weak convergence, using the second characterisation in the Portmanteau theorem, one has to strengthen the class $\mathcal{C}^3(\mathbb{R})$ of test functions to the class of bounded, uniformly continuous functions. This is possible due to a standard mollification technique in analysis (which is particularly useful in PDE theory).

Lemma 7.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a bounded, uniformly continuous function. Then there exists a sequence $f_n \in \mathcal{C}_b^3(\mathbb{R})$ such that f_n converges uniformly to f.*

Proof. The main idea is to convolute f with a "nice" function. One possible choice is the following. For each $\eta > 0$, we define

$$\rho_\eta(x) = \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{x^2}{2\eta}}, \quad x \in \mathbb{R}$$

to be the density function of $N(0,\eta)$. Let

$$f_\eta(x) \triangleq (\rho_\eta * f)(x) \triangleq \int_{\mathbb{R}} \rho_\eta(x-y) f(y) dy.$$

Since $\int_{\mathbb{R}} \rho_\eta(x)dx = 1$ and f is bounded, one knows that f_η is well-defined. Indeed, f_η is smooth and its k-th derivative is given by

$$f_\eta^{(k)}(x) = \int_{\mathbb{R}} \rho_\eta^{(k)}(x-y)f(y)dy$$

which is easily seen to be bounded on $\mathbb{R}$.

We now show that f_η converges uniformly to f as $\eta \to 0$. First of all, since f is uniformly continuous, given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|y-x| < \delta \implies |f(y) - f(x)| < \varepsilon.$$

It follows that

$$\begin{aligned}
\big|f_\eta(x) - f(x)\big| &= \Big| \int_{\mathbb{R}} \rho_\eta(x-y)(f(y)-f(x))dy\Big| \\
&\leqslant \Big| \int_{\{y:|y-x|<\delta\}} \rho_\eta(x-y)(f(y)-f(x))dy\Big| \\
&\quad + \Big| \int_{\{y:|y-x|\geqslant\delta\}} \rho_\eta(x-y)(f(y)-f(x))dy\Big| \\
&\leqslant \varepsilon + 2\|f\|_\infty \cdot \int_{\{y:|y-x|\geqslant\delta\}} \rho_\eta(x-y)dy \\
&= \varepsilon + 2\|f\|_\infty \cdot \mathbb{P}(|X_\eta| \geqslant \delta)
\end{aligned}$$

where $X_\eta \overset{d}{=} N(0,\eta)$. Note that

$$\mathbb{P}(|X_\eta| \geqslant \delta) = \mathbb{P}\big(Z \geqslant \frac{\delta}{\sqrt{\eta}}\big) \to 0 \quad \text{as } \eta \to 0,$$

where $Z \overset{d}{=} N(0,1)$. Therefore, one obtains that

$$\overline{\lim_{\eta\to 0}} \|f_\eta - f\|_\infty \leqslant \varepsilon.$$

The result follows as ε is arbitrary. □

To complete the proof of the CLT, let f be a bounded and uniformly continuous function on $\mathbb{R}$. Given $\varepsilon > 0$, let $g \in \mathcal{C}_b^3(\mathbb{R})$ be such that

$$\|g - f\|_\infty \triangleq \sup_{x\in\mathbb{R}} |g(x) - f(x)| < \varepsilon.$$

The existence of g is guaranteed by Lemma 7.2. It follows that

$$\begin{aligned}
&\big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[f(Z)]\big| \\
&\leqslant \big|\mathbb{E}[f(\hat{S}_n)] - \mathbb{E}[g(\hat{S}_n)]\big| + \big|\mathbb{E}[g(\hat{S}_n)] - \mathbb{E}[g(Z)]\big| \\
&\quad + \big|\mathbb{E}[g(Z)] - \mathbb{E}[f(Z)]\big| \\
&\leqslant 2\varepsilon + \big|\mathbb{E}[g(\hat{S}_n)] - \mathbb{E}[g(Z)]\big|.
\end{aligned}$$

According to (7.7), the second term tends to zero as $n \to \infty$. Since ε is arbitrary, one concludes that

$$\mathbb{E}[f(\hat{S}_n)] \to \mathbb{E}[f(Z)].$$

This yields the desired weak convergence.

Remark 7.2. As we have seen, Lindeberg's condition (7.8) implies that (i) $S_n/\Sigma_n \to N(0,1)$ weakly and (ii) $r_n \to 0$. Later on, W. Feller proved that Lindeberg's condition is also necessary for (i) and (ii) to hold. This result together with Theorem 7.2 is known as the *Lindeberg-Feller theorem*. We refer the reader to [Chu01] for its proof.

7.3 Non-Gaussian central limit theorems: an example

In the i.i.d. context, if the random variables have finite mean and variance, the limiting distribution of the normalised partial sum is Gaussian. However, if the random variables have *heavy tails* (thus having less integrability), the limiting distribution (if it exists) could fail to be Gaussian in general. In this section, we use one example to demonstrate such a phenomenon. Note that non-Gaussian type limit theorems already appear for instance in the Poisson approximation of binomial distributions:

$$\text{Binomial}(n, p_n) \overset{\text{weakly}}{\longrightarrow} \text{Poisson}(\lambda) \quad \text{if } np_n \to \lambda > 0.$$

Let $0 < \alpha < 2$ be a given fixed number. Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence with probability density function

$$p_\alpha(x) \triangleq \begin{cases} \frac{\alpha}{2|x|^{1+\alpha}}, & |x| \geqslant 1; \\ 0, & \text{otherwise.} \end{cases}$$

We are interested in the asymptotic behaviour of $\frac{X_1+\cdots+X_n}{a_n}$ with a suitable normalising sequence a_n. Note that X_1 does not have finite variance and thus the classical CLT does not apply.

Let $f_\alpha(t)$ be the characteristic function of X_1. The crucial point for understanding this situation is to figure out the behaviour of $f_\alpha(t)$ near $t = 0$. Since $f_\alpha(0) = 1$, let us write

$$\begin{aligned} 1 - f_\alpha(t) &= \int_{-\infty}^{\infty} \big(1 - e^{itx}\big) p_\alpha(x) dx \\ &= \alpha \int_1^{\infty} \frac{1 - \cos tx}{x^{1+\alpha}} dx \\ &= \alpha |t|^\alpha \int_{|t|}^{\infty} \frac{1 - \cos u}{u^{1+\alpha}} du \\ &= \alpha |t|^\alpha \Big(\int_0^{\infty} \frac{1 - \cos u}{u^{1+\alpha}} du - \int_0^{|t|} \frac{1 - \cos u}{u^{1+\alpha}} du \Big). \end{aligned}$$

Since $1 - \cos u = \frac{1}{2}u^2 + o(u^2)$, the first integral on the right hand side is finite, In addition, for the second integral one has

$$\int_0^{|t|} \frac{1 - \cos u}{u^{1+\alpha}} du = \int_0^{|t|} \frac{\frac{1}{2}u^2 + o(u^2)}{u^{1+\alpha}} du = O(|t|^{2-\alpha}).$$

As a result, one finds that

$$1 - f_\alpha(t) = C_\alpha |t|^\alpha + O(|t|^2) \quad \text{as } t \to 0, \tag{7.15}$$

where $C_\alpha > 0$ is a constant depending only on α.

The relation (7.15) naturally leads to the correct normalisation in the corresponding CLT. In fact, the characteristic function of $S_n/n^{1/\alpha}$ ($S_n \triangleq X_1 + \cdots + X_n$) is given by

$$f_{S_n/n^{1/\alpha}}(t) = \big(f_\alpha\big(\frac{t}{n^{1/\alpha}}\big)\big)^n = \big(1 - \frac{C_\alpha |t|^\alpha}{n} + O\big(\frac{t^2}{n^{2/\alpha}}\big)\big)^n.$$

In the above equation, t is fixed and the term $O(\frac{t^2}{n^{2/\alpha}})$ is understood in the limit $n \to \infty$. It follows that

$$\lim_{n\to\infty} f_{S_n/n^{1/\alpha}}(t) = e^{-C_\alpha |t|^\alpha}.$$

According to the Lévy-Cramér theorem, the function $g_\alpha(t) \triangleq e^{-C_\alpha |t|^\alpha}$ must be a characteristic function (of some distribution G_α) and one has

$$\frac{S_n}{n^{1/\alpha}} \to G_\alpha \quad \text{weakly}$$

as $n \to \infty$.

Remark 7.3. When $\alpha = 1$, G_α is a Cauchy distribution. When $\alpha > 2$, one is back to the setting of the classical CLT and thus $S_n/\sqrt{n}$ converges weakly to a normal distribution. *What if* $\alpha = 2$?

7.4 Introduction to Stein's method

To gain deeper understanding on the CLT, it is necessary to develop effective methods of analysing the associated error (rate of convergence) at various quantitative levels. A powerful modern technique for this purpose is known as *Stein's method*. In this section, we develop the basic ingredients behind this method and use it to derive quantitative error estimates in the CLT. Although the scope of Stein's method is rather broad, to illustrate the essential ideas we confine ourselves to the context of *independent random variables*.

7.4.1 The general picture and basic ingredients

Recall that the CLT asserts that $\hat{S}_n \to Z$ weakly, where $\hat{S}_n$ is a suitably normalised random variable and $Z \stackrel{d}{=} N(0,1)$. To understand the rate of convergence in the CLT, one first needs a natural notion of "distance" between two distribution functions (or equivalently, between two probability measures).

Different notions of distance for distributions

To get the main idea, suppose that W and Z are two random variables with distribution functions F and G respectively. Among others, there are at least two apparent notions of "distance" between F and G :

(i) *The uniform distance*:

$$\|F - G\|_\infty \triangleq \sup_{x \in \mathbb{R}} |F(x) - G(x)|. \tag{7.16}$$

(ii) *The* L^1*-distance*:

$$\|F - G\|_{L^1} \triangleq \int_{\mathbb{R}} |F(x) - G(x)| dx. \tag{7.17}$$

There is a unified viewpoint to look at these two distances. Let μ, ν be the probability laws of W, Z respectively. We have seen in the definition of weak convergence and the proof of the CLT that the quantity

$$\big|\mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)]\big| = \big| \int_{\mathbb{R}} \varphi d\mu - \int_{\mathbb{R}} \varphi d\nu \big|,$$

when φ ranges over certain class of test functions, gives a natural sense of "closeness" between the two distributions. In fact, if one fixes a suitable class $\mathcal{H}$ of test functions on $\mathbb{R}$, there is an associated notion of distance defined by

$$d_{\mathcal{H}}(\mu,\nu) \triangleq \sup\big\{\big|\int_{\mathbb{R}}\varphi d\mu - \int_{\mathbb{R}}\varphi d\nu\big| : \varphi\in\mathcal{H}\big\}. \tag{7.18}$$

Apparently, this notion of distance depends crucially on the underlying class $\mathcal{H}$ of test functions.

(i) Suppose that $\mathcal{H}$ is the class of *indicator functions of semi-infinite intervals*, i.e.

$$\mathcal{H} \triangleq \{\mathbf{1}_{(-\infty,a]}(x) : a\in\mathbb{R}\}.$$

Then $d_{\mathcal{H}}(\mu,\nu)$ recovers the uniform distance between F and G defined in (7.16). The uniform distance is also known as the *Kolmogorov distance*.
(ii) Assume further that W and Z both have finite mean. If one takes $\mathcal{H}$ to be the class of 1-*Lipschitz functions*, i.e. the class of functions $\varphi:\mathbb{R}\to\mathbb{R}$ such that

$$|\varphi(x)-\varphi(y)| \leqslant |x-y| \quad \text{for all } x,y\in\mathbb{R},$$

then it can be shown that $d_{\mathcal{H}}(\mu,\nu)$ recovers the L^1-distance between F and G defined in (7.17). This fact, which is not entirely obvious at the moment, will be shown in the appendix. The distance in this case is known as the 1-*Wasserstein distance*.
(iii) There is another natural distance associated with the class of test functions taken to be *indicator functions of Borel subsets*, i.e. $\mathcal{H} \triangleq \{\mathbf{1}_A(x) : A\in\mathcal{B}(\mathbb{R})\}$. The associated distance, given by

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{A\in\mathcal{B}(\mathbb{R})}\big|\int_{\mathbb{R}}\mathbf{1}_A d\mu - \int_{\mathbb{R}}\mathbf{1}_A d\nu\big| = \sup_{A\in\mathcal{B}(\mathbb{R})}\big|\mu(A)-\nu(A)\big|,$$

is known as the *total variation distance*. This distance is commonly used in the context of discrete random variables e.g. in the study of Poisson approximations.

From the above discussion, in order to estimate the "distance" between the distributions of W and Z, an essential ingredient is to find an effective way to estimate the quantity

$$|\mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)]| \tag{7.19}$$

in terms of suitable "norms" of the test function φ. For instance, we have seen such type of estimate in terms of the third derivative of φ in Lindeberg's CLT. However,

this is not sufficient for many applications and it is necessary to strengthen the estimate to weaker norms of φ (e.g. in terms of the first derivative of φ).

In the 1960s, C. Stein developed a powerful method to estimate distributional distances defined through quantities like (7.19). The scope of Stein's method goes way beyond the CLT and Gaussian approximations. Here we only discuss the Gaussian case in the classical setting. Nonetheless, our analysis contains the essential ideas behind this general method (at least in the classical sense). Our main goal is to estimate the L^1-*distance* between the distributions of $W = \hat{S}_n$ and $Z \stackrel{d}{=} N(0,1)$ in the context of independent random variables. Such an estimate is known as the L^1-*Berry-Esseen estimate*. The uniform Berry-Esseen estimate (i.e. the corresponding estimate for the uniform distance) is technically much harder to obtain, but it is also achievable within Stein's method.

Basic ingredients of Stein's method for Gaussian approximation

Recall that in the CLT context, Z is a standard normal random variable and $W = \hat{S}_n$. The starting point of Stein's method is the following simple calculation. Let f be a suitably regular test function. By applying integration by parts (assuming the boundary term goes away), one has

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(z) e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} z f(z) e^{-z^2/2} dz \\ &= \mathbb{E}[Z f(Z)].\end{aligned}$$

A key observation is that the above property indeed characterises the standard normal distribution. Namely, *a random variable Z is $N(0,1)$-distributed if and only if*

$$\mathbb{E}[f'(Z)] - \mathbb{E}[Z f(Z)] = 0 \tag{7.20}$$

for a wide class of test functions f. This will be the content of Stein's lemma in Section 7.4.2 below. Based on this fact, one naturally expects that if the distribution of W is "close to" $N(0,1)$, the quantity $\mathbb{E}[f'(W)] - \mathbb{E}[W f(W)]$ should be "small".

To quantify such a property, recall that we wish to estimate (7.19) for given test function φ, where W is a general random variable and $Z \stackrel{d}{=} N(0,1)$. The next key step is to write down a so-called *Stein's equation* associated with the given function φ:

$$f'(x) - x f(x) = \varphi(x) - c_\varphi, \tag{7.21}$$

where

$$c_\varphi \triangleq \mathbb{E}[\varphi(Z)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \varphi(z) e^{-z^2/2} dz$$

is the mean of φ with respect to the standard normal distribution. The form of this equation is naturally motivated from the characterisation (7.20). Stein's equation (7.21) is a first order linear ODE, whose solution f can be written down easily. It follows that

$$f'(W) - Wf(W) = \varphi(W) - c_\varphi.$$

If one takes expectation on both sides, one arrives at

$$\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)] = \mathbb{E}[\varphi(W)] - \mathbb{E}[\varphi(Z)].$$

In particular, the original task of estimating (7.19) is magically transferred to the estimation of the quantity

$$\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)]. \tag{7.22}$$

Note that if $W = Z$, this quantity is zero which is consistent with the characterisation (7.20). In general, this quantity can be estimated in terms of certain derivatives of the function f (the development of this part is the last step of Stein's method). Since the original goal is to estimate (7.19) in terms of φ, one must find a way to estimate derivatives of the solution f in terms of suitable norms of φ. This part corresponds to the *analysis of Stein's equation*, which is the second step of Stein's method and will be developed in Section 7.4.3 below.

The last step is to *estimate the quantity (7.22)*. There is no universal approach to this step and the analysis for this part depends heavily on the nature of the underlying problem (i.e. the specific assumption on the random variable W). To illustrate the essential idea, we will only develop this step in the context of independent random variables, i.e. when $W = \hat{S}_n$ with $\{X_n : n \geqslant 1\}$ being an independent sequence (cf. Section 7.4.4 below). Nonetheless, we must point out that this step can be established in a much wider context (e.g. for random variables with local dependence), which makes Stein's method robust and powerful.

The three main ingredients of Stein's method are summarised as follows.

Step one. Establish the characterising property for the standard normal distribution. In abstract terms, this characterising property takes the form

$$\mathbb{E}[\mathcal{A}f(Z)] = 0 \quad \text{for all suitable test functions } f.$$

For the standard normal distribution, one has $(\mathcal{A}f)(x) = f'(x) - xf(x)$.

Step two. Write down Stein's equation associated with a given test function φ. This equation takes the form

$$\mathcal{A}f = \varphi - c_\varphi.$$

For the standard normal distribution, this equation is given by (7.21). One also needs to estimate the solution f in terms of the given function φ.

Step three. Using the specific structure of the random variable W to estimate the quantity $\mathbb{E}[\mathcal{A}f(W)]$ in terms of f. In the Gaussian context, this quantity is given by (7.22).

Remark 7.4. Although we only consider Gaussian approximations here, the formulation of these three steps is robust and applies to other types of distributional approximations (e.g. when the limiting distribution is Poisson, the "generator" $\mathcal{A}$ takes a different form).

In the following sections, we develop the above three ingredients mathematically with our ultimate goal towards the L^1-Berry-Esseen estimate in the independent context.

7.4.2 Step one: Stein's lemma

We start by establishing the characterising property (7.20) of $N(0,1)$. This is known as *Stein's lemma* for the normal distribution.

Lemma 7.3. *Let Z be a random variable. Then the following two statements are equivalent.*

(i) $Z \overset{d}{=} N(0,1)$.
(ii) For any piecewise differentiable function $f : \mathbb{R} \to \mathbb{R}$ that is integrable with respect to the standard Gaussian measure, both of $\mathbb{E}[f'(Z)]$ and $\mathbb{E}[Zf(Z)]$ are finite and one has

$$\mathbb{E}[f'(Z)] = \mathbb{E}[Zf(Z)].$$

Proof. (i) $\Longrightarrow$ (ii). Suppose that $Z \overset{d}{=} N(0,1)$. Given f satisfying the assumptions,

one has

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(z)e^{-z^2/2}dz \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} f'(z)\Big(\int_{-\infty}^{z}(-x)e^{-x^2/2}dx\Big)dz \\ &\quad + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} f'(z)\Big(\int_{z}^{\infty} xe^{-x^2/2}dx\Big)dz,\end{aligned}$$

where we used the relation

$$e^{-z^2/2} = \int_{-\infty}^{z}(-x)e^{-x^2/2}dx = \int_{z}^{\infty} xe^{-x^2/2}dx.$$

By using Fubini's theorem, one has

$$\begin{aligned}\int_{-\infty}^{0} f'(z)\Big(\int_{-\infty}^{z}(-x)e^{-x^2/2}dx\Big)dz &= \int_{-\infty}^{0}(-x)e^{-x^2/2}dx\int_{x}^{0} f'(z)dz \\ &= \int_{-\infty}^{0}(-x)e^{-x^2/2}\big(f(0)-f(x)\big)dx \\ &= \int_{-\infty}^{0} x\big(f(x)-f(0)\big)e^{-x^2/2}dx.\end{aligned}$$

Similarly,

$$\int_{0}^{\infty} f'(z)\Big(\int_{z}^{\infty} xe^{-x^2/2}dx\Big)dz = \int_{0}^{\infty} x\big(f(x)-f(0)\big)e^{-x^2/2}dx.$$

It follows that

$$\begin{aligned}\mathbb{E}[f'(Z)] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} x\big(f(x)-f(0)\big)e^{-x^2/2}dx \\ &\quad + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} x\big(f(x)-f(0)\big)e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xf(x)e^{-x^2/2}dx \quad \Big(\text{since } \int_{-\infty}^{\infty} xe^{-x^2/2}dx = 0\Big) \\ &= \mathbb{E}[Zf(Z)].\end{aligned}$$

(ii) $\Longrightarrow$ (i). Let $\varphi(t) \triangleq \mathbb{E}[e^{itZ}]$ be the characteristic function of Z. Taking $f = 1$ in the assumption, one sees that $\mathbb{E}[Z]$ is finite. According to Theorem 6.4, $\varphi(t)$ is differentiable and

$$\varphi'(t) = i\mathbb{E}[Ze^{itZ}].$$

On the other hand, by choosing $f(x) = e^{itx}$ (with t fixed) one has

$$\mathbb{E}[f'(Z)] = it\mathbb{E}[e^{itZ}] = it\varphi(t),$$

and

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[Ze^{itZ}] = -i\varphi'(t).$$

The assumption implies that $it\varphi(t) = -i\varphi'(t)$, or equivalently

$$\varphi'(t) = -t\varphi(t).$$

Since $\varphi(0) = 1$, the above ODE has the unique solution $\varphi(t) = e^{-t^2/2}$ which is precisely the characteristic function of $N(0,1)$. Therefore, one concludes that $Z \overset{d}{=} N(0,1)$. □

7.4.3 Step two: Analysing Stein's equation

For a given function φ, we wish to estimate the solution f to Stein's equation

$$f'(x) - xf(x) = \varphi(x) - c_\varphi$$

in terms of φ. To do so, one needs the following lemma regarding Gaussian tail estimates.

Lemma 7.4. *For any $x \in \mathbb{R}$, we have*

$$|x|e^{x^2/2}\int_{|x|}^{\infty} e^{-t^2/2}dt \leqslant 1, \quad e^{x^2/2}\int_{|x|}^{\infty} e^{-t^2/2}dt \leqslant \sqrt{\frac{\pi}{2}}.$$

Proof. Apparently we can assume that $x \geqslant 0$. The first claim follows from

$$xe^{x^2/2}\int_x^{\infty} e^{-t^2/2}dt \leqslant e^{x^2/2}\int_x^{\infty} te^{-t^2/2}dt = 1.$$

For the second claim, one considers the function

$$q(x) \triangleq e^{x^2/2}\int_x^{\infty} e^{-t^2/2}dt, \quad x \geqslant 0.$$

Using the first part, one sees that

$$q'(x) = xe^{x^2/2}\int_x^{\infty} e^{-t^2/2}dt - 1 \leqslant 0$$

and thus

$$q(x) \leqslant q(0) = \int_0^{\infty} e^{-t^2/2}dt = \sqrt{\frac{\pi}{2}}.$$

□

Remark 7.5. Such type of Gaussian tail estimate was already obtained in Example 2.3 based on Chebyshev's inequality. Here we reproduced a similar result by using an analytic argument.

The main result for the analysis of Stein's equation is stated as follows.

Proposition 7.2. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable with uniformly bounded derivative. Set*

$$\tilde{\varphi}(x) \triangleq \varphi(x) - c_{\varphi},$$

where we recall that $c_{\varphi} \triangleq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \varphi(x) e^{-x^2/2} dx$ is the mean of φ with respect to $N(0,1)$. Then

$$f(x) \triangleq e^{x^2/2} \int_{-\infty}^{x} \tilde{\varphi}(t) e^{-t^2/2} dt, \quad x \in \mathbb{R} \tag{7.23}$$

is the unique bounded solution to Stein's equation

$$f'(x) - x f(x) = \tilde{\varphi}(x). \tag{7.24}$$

In addition, f has bounded, continuous derivatives up to order two, and the following estimates hold:

$$\|f\|_{\infty} \leqslant 2\|\varphi'\|_{\infty}, \ \ \|f'\|_{\infty} \leqslant 3\sqrt{\frac{\pi}{2}}\|\varphi'\|_{\infty}, \ \ \|f''\|_{\infty} \leqslant 6\|\varphi'\|_{\infty}.$$

Proof. From standard ODE theory, the general solution to the linear ODE (7.24) is found to be

$$f_c(x) = c e^{x^2/2} + f(x),$$

where $f(x)$ is the function defined by (7.23) and c is an arbitrary constant. In what follows, we prove that $f(x)$ is bounded. It is then clear that $f(x)$ is the unique bounded solution, since any choice of $c \neq 0$ would lead to an unbounded solution due to the unboundedness of the factor $e^{x^2/2}$.

(i) *Estimating f.* Let us assume that $\varphi(0) = 0$, as subtracting a constant to φ does not change $\tilde{\varphi}$ or the ODE (7.24). In this case, one has

$$|\varphi(t)| = |\varphi(t) - \varphi(0)| \leqslant \|\varphi'\|_{\infty} \cdot |t| \tag{7.25}$$

and

$$|c_{\varphi}| \leqslant \|\varphi'\|_{\infty} \cdot \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |t| e^{-t^2/2} dt = \|\varphi'\|_{\infty} \cdot \sqrt{\frac{2}{\pi}}, \tag{7.26}$$

where we used the explicit expression for the first absolute moment of $N(0,1)$.

To estimate $f(x)$, one first considers the case when $x \leqslant 0$. By using (7.25) and (7.26), one has

$$\begin{aligned}
|f(x)| &\leqslant e^{x^2/2} \int_{|x|}^{\infty} \big(\|\varphi'\|_\infty \cdot t + \|\varphi'\|_\infty \cdot \sqrt{\frac{2}{\pi}}\big) e^{-t^2/2} dt \\
&= \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^{\infty} t e^{-t^2/2} dt + \sqrt{\frac{2}{\pi}} \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^{\infty} e^{-t^2/2} dt \\
&= \|\varphi'\|_\infty + \sqrt{\frac{2}{\pi}} \|\varphi'\|_\infty \cdot e^{x^2/2} \int_{|x|}^{\infty} e^{-t^2/2} dt.
\end{aligned}$$

According to Lemma 7.4, one sees that

$$|f(x)| \leqslant 2\|\varphi'\|_\infty. \tag{7.27}$$

If $x \geqslant 0$, one use the alternative expression for f given by

$$f(x) = -e^{x^2/2} \int_{x}^{\infty} \tilde{\varphi}(t) e^{-t^2/2} dt, \tag{7.28}$$

which follows from the observation that

$$\int_{-\infty}^{\infty} \tilde{\varphi}(t) e^{-t^2/2} dt = 0.$$

In this case, the same argument applied to (7.28) gives the same estimate (7.27). Therefore, one obtains that

$$\|f\|_\infty \leqslant 2\|\varphi'\|_\infty.$$

(ii) *Estimating* f'. Since φ is differentiable, by differentiating the ODE (7.24) one has

$$f''(x) - x f'(x) = f(x) + \varphi'(x). \tag{7.29}$$

Inspired by the previous argument, in order to estimate f', just like the case for f one may wish to express f' as the product of $e^{x^2/2}$ and another function (an $\int_{-\infty}^{x}$-integral). For this purpose, one computes

$$\begin{aligned}
\frac{d}{dx}\big(e^{-x^2/2} f'(x)\big) &= e^{-x^2/2} f''(x) - x e^{-x^2/2} f'(x) \\
&= e^{-x^2/2} \big(f(x) + \varphi'(x)\big).
\end{aligned}$$

It follows that

$$f'(x) = e^{x^2/2} \cdot \int_{-\infty}^{x} \big(f(t) + \varphi'(t)\big) e^{-t^2/2} dt. \tag{7.30}$$

Similar to Part (i), one first considers $x \leqslant 0$. In this case, using the estimate (7.27) on f as well as Lemma 7.4, one finds that

$$|f'(x)| \leqslant 3\|\varphi'\|_\infty e^{x^2/2} \int_{|x|}^{\infty} e^{-t^2/2} dt \leqslant 3\sqrt{\frac{\pi}{2}} \|\varphi'\|_\infty. \tag{7.31}$$

If $x \geqslant 0$, one turns to the alternative expression

$$f'(x) = -e^{x^2/2} \int_{x}^{\infty} \big(f(t) + \varphi'(t)\big) e^{-t^2/2} dt. \tag{7.32}$$

This is legal since

$$\int_{-\infty}^{\infty} \big(f(t) + \varphi'(t)\big) e^{-t^2/2} dt = \int_{-\infty}^{\infty} \big(f''(t) + t f'(t)\big) e^{-t^2/2} dt = 0,$$

where the second equality follows from integration by parts. The same argument applied to (7.32) again gives (7.31) in this case. Therefore, one arrives at

$$\|f'\|_\infty \leqslant 3\sqrt{\frac{\pi}{2}} \|\varphi'\|_\infty.$$

(iii) *Estimating* f''. According to the equation (7.29) for f'' and the expression (7.30) for f', one has

$$f''(x) = x e^{x^2/2} \int_{-\infty}^{x} \big(f(t) + \varphi'(t)\big) e^{-t^2/2} dt + \big(f(x) + \varphi'(x)\big).$$

One has already got all the needed ingredients to estimate the above terms. To be precise, again by considering the cases $x \leqslant 0$ and $x \geqslant 0$ separately one finds that

$$\begin{aligned} |f''(x)| &\leqslant \big(\|f\|_\infty + \|\varphi'\|_\infty\big) \cdot |x| e^{x^2/2} \int_{|x|}^{\infty} e^{-t^2/2} dt + \big(\|f\|_\infty + \|\varphi'\|_\infty\big) \\ &\leqslant 2\big(\|f\|_\infty + \|\varphi'\|_\infty\big) \\ &\leqslant 6\|\varphi'\|_\infty, \end{aligned}$$

where we used Lemma 7.4 and the estimate on f obtained in Part (i).

The proof of the proposition is now complete. □

7.4.4 Step three: Establishing the L^1-Berry-Esseen estimate

The first two steps are general. To develop the last step, we restrict ourselves to the independent case. In its precise form, the main theorem is stated as follows.

Theorem 7.3. *Let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables, each having mean zero and finite third moment. For each n, we set*

$$\Sigma_n \triangleq \sqrt{\mathrm{Var}[S_n]},\ \tau_n \triangleq \big(\mathbb{E}[|X_n|^3]\big)^{1/3},\ \hat{S}_n \triangleq \frac{S_n}{\Sigma_n},$$

where $S_n \triangleq X_1 + \cdots + X_n$. Then for any continuously differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ with bounded derivative, one has

$$\big|\mathbb{E}[\varphi\big(\hat{S}_n\big)] - \mathbb{E}[\varphi(Z)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3} \quad \forall n \geqslant 1, \tag{7.33}$$

where $Z \overset{d}{=} N(0,1)$. In particular, if $\{X_n : n \geqslant 1\}$ is an i.i.d. sequence with mean zero, unit variance and $\tau \triangleq \big(\mathbb{E}[|X_1|^3]\big)^{1/3} < \infty$, then one has

$$\big|\mathbb{E}[\varphi\big(\hat{S}_n\big)] - \mathbb{E}[\varphi(Z)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\tau}{\sqrt{n}}.$$

Proof. Let f be the unique bounded solution to Stein's equation (7.24) corresponding to φ. Then

$$\mathbb{E}[\varphi\big(\hat{S}_n\big)] - \mathbb{E}[\varphi(Z)] = \mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)].$$

As the last step in Stein's method, our goal is to estimate the right hand side of the above equation. For this purpose, we first introduce the following notation:

$$\hat{X}_m \triangleq \frac{X_m}{\Sigma_n},\ \hat{\sigma}_m \triangleq \frac{\sigma_m}{\Sigma_n}, \quad 1 \leqslant m \leqslant n.$$

Note that $\hat{\sigma}_m^2 = \mathbb{E}[\hat{X}_m^2]$ and

$$\sum_{m=1}^n \hat{X}_m = \hat{S}_n,\ \sum_{m=1}^n \hat{\sigma}_m^2 = 1.$$

One can now rewrite

$$\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)] = \sum_{m=1}^n \mathbb{E}[\hat{\sigma}_m^2 f'(\hat{S}_n)] - \sum_{m=1}^n \mathbb{E}[\hat{X}_m f(\hat{S}_n)].$$

The next crucial point is to relate $f(\hat{S}_n)$ with $f(\hat{S}_n - \hat{X}_m)$ through f' (this is beneficial since $\hat{S}_n - \hat{X}_m$ and $\hat{X}_m$ are independent). To this end, recall from calculus that

$$f(y) = f(x) + \int_0^1 f'\big((1-t)x + ty\big) \cdot (y-x)dt.$$

Taking $x = \hat{S}_n - \hat{X}_m$ and $y = \hat{S}_n$, one can write

$$f(\hat{S}_n) = f(\hat{S}_n - \hat{X}_m) + \int_0^1 f'(T_{n,m}(t))\hat{X}_m dt,$$

where we set $T_{n,m}(t) \triangleq (1-t)(\hat{S}_n - \hat{X}_m) + t\hat{S}_n$ to simplify notation. It follows that

$$\begin{aligned}\mathbb{E}[\hat{X}_m f(\hat{S}_n)] &= \mathbb{E}[\hat{X}_m f(\hat{S}_n - \hat{X}_m)] + \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big] \\ &= \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big].\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)] \\ &= \sum_{m=1}^n \mathbb{E}[\hat{\sigma}_m^2 f'(\hat{S}_n)] - \sum_{m=1}^n \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 f'(T_{n,m}(t))dt\big] \\ &= \sum_{m=1}^n \mathbb{E}\big[\hat{\sigma}_m^2 \cdot \big(f'(\hat{S}_n) - f'(T_{n,m}(0))\big)\big] \\ &\quad + \sum_{m=1}^n \mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 \big(f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big)dt\big], \qquad (7.34)\end{aligned}$$

where we used the observation $T_{n,m}(0) = \hat{S}_n - \hat{X}_m$ to reach the last equality and thus

$$\mathbb{E}[\hat{\sigma}_m^2 f'(T_{n,m}(0))] = \mathbb{E}[\hat{X}_m^2 f'(T_{n,m}(0))].$$

To estimate the first summation on the right hand side of (7.34), we shall use the inequality

$$\big|f'(\hat{S}_n) - f'(T_{n,m}(0))\big| \leqslant \|f''\|_\infty \cdot |\hat{X}_m|.$$

This gives

$$\big|\mathbb{E}\big[\hat{\sigma}_m^2 \cdot \big(f'(\hat{S}_n) - f'(T_{n,m}(0))\big)\big]\big| \leqslant \|f''\|_\infty \hat{\sigma}_m^2 \cdot \mathbb{E}[|\hat{X}_m|] \leqslant \|f''\|_\infty \cdot \frac{\tau_m^3}{\Sigma_n^3},$$

where we used the fact that $p \mapsto \|X\|_p$ is increasing in $p \geqslant 1$ (cf. Corollary 2.3); in particular, $\mathbb{E}[|X_m|] \leqslant \tau_m$ and $\sigma_m \leqslant \tau_m$. To estimate the second summation on the right hand side of (7.34), note that

$$\big|f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big)\big| \leqslant t\|f''\|_\infty \cdot |\hat{X}_m|.$$

This gives

$$\big|\mathbb{E}\big[\hat{X}_m^2 \cdot \int_0^1 \big(f'(T_{n,m}(t)) - f'(T_{n,m}(0))\big)dt\big]\big| \leqslant \frac{1}{2}\|f''\|_\infty \cdot \frac{\tau_m^3}{\Sigma_n^3}.$$

Finally, using the estimate $\|f''\|_\infty \leqslant 6\|\varphi'\|_\infty$ given by Proposition 7.2, one arrives at

$$\big|\mathbb{E}[f'(\hat{S}_n)] - \mathbb{E}[\hat{S}_n f(\hat{S}_n)]\big| \leqslant 9\|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3},$$

which is the desired estimate. □

The estimate (7.33) is not yet in the shape of an L^1-distance estimate. To complete the discussion, we now proceed to see how Theorem 7.3 gives rise to an L^1-estimate for the distribution functions. Here for simplicity we only sketch the argument. Making the analysis fully rigorous requires more technical effort; this unpleasant task will be deferred to the appendix.

Let F_n be the distribution function of $\hat{S}_n$ and let Φ be the distribution function of $N(0,1)$. First of all, a naive integration by parts gives

$$\int_{\mathbb{R}} \varphi(x)dF_n(x) - \int_{\mathbb{R}} \varphi(x)d\Phi(x) = \int_{\mathbb{R}} \varphi'(x)(\Phi(x) - F_n(x))dx.$$

Therefore, Theorem 7.3 yields that

$$\big|\int_{\mathbb{R}} \varphi'(x)\big(\Phi(x) - F_n(x)\big)\big| \leqslant C_n \cdot \|\varphi'\|_\infty \tag{7.35}$$

for any φ with $\varphi' \in \mathcal{C}_b(\mathbb{R})$, where

$$C_n \triangleq 9 \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

is the constant that provides the rate of convergence. Symbolically, one can just put $\psi \triangleq \varphi'$ to see that

$$\Big| \int_{\mathbb{R}} \psi(x)(F_n(x) - \Phi(x))dx \Big| \leqslant C_n \|\psi\|_\infty$$

for any bounded continuous function ψ. Based on this point, it is not too surprising to expect that

$$\|F_n - \Phi\|_{L^1} = \int_{\mathbb{R}} |F_n(x) - \Phi(x)| \leqslant C_n.$$

In fact, if one were allowed to even choose $\psi(x) = \operatorname{sgn}(F_n(x) - G(x))$ where $\operatorname{sgn}(x)$ is the function defined by

$$\operatorname{sgn}(x) \triangleq \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0, \end{cases}$$

then one has $\|\psi\|_\infty \leqslant 1$ and

$$\int_{\mathbb{R}} \psi(x)(F_n(x) - \Phi(x))dx = \int_{\mathbb{R}} |F_n(x) - \Phi(x)|dx,$$

yielding the desired L^1-estimate. The main difficulty here is that $\psi(x)$ is not a continuous function. Getting around this difficulty requires more analysis and this is done in the appendix.

To summarise, the *L^1-Berry-Esseen estimate* is the content of the following result.

Corollary 7.2. *Under the same set-up as in Theorem 7.3, one has*

$$\int_{\mathbb{R}} |F_n(x) - \Phi(x)|dx \leqslant 9 \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}. \tag{7.36}$$

In particular, in the i.i.d. context one has

$$\int_{\mathbb{R}} |F_n(x) - \Phi(x)|dx \leqslant 9 \frac{\tau}{\sqrt{n}}.$$

7.4.5 Further remarks and scopes

We conclude this chapter with a few further discussions on Stein's method.

(i) Let us take a second look at our previous heuristic argument about obtaining the L^1-Berry-Esseen estimate. The fact that the right hand side of (7.35) is the uniform norm of φ' leads one to the L^1-estimate for the distribution functions. Through a naive duality viewpoint, if one were able to replace the right hand side of (7.35) by the L^1-norm of φ' ($\|\varphi'\|_{L^1} \triangleq \int_{\mathbb{R}} |\varphi'(x)|dx$), one should then be able to deduce the *uniform Berry-Esseen estimate*. This requires strengthening the analysis of Stein's equation (cf. Proposition 7.2) to estimating the uniform norms of f, f', f'' in terms of $\|\varphi'\|_{L^1}$. It can be done for f, f' but not for f''! This is why the uniform Berry-Esseen estimate is much harder than the L^1-estimate to obtain. The result is stated as follows. A complete proof along the above lines of argument can be found in [Str11].

Theorem 7.4 (The Uniform Berry-Esseen Estimate)**.** *Under the same notation as in Theorem 7.3 and Corollary 7.2, one has*

$$\|F_n - \Phi\|_\infty \leqslant 10 \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}.$$

In particular, in the i.i.d. context one has

$$\|F_n - \Phi\|_\infty \leqslant 10 \cdot \frac{\tau}{\sqrt{n}}.$$

(ii) As we mentioned earlier, Step 3 in Stein's method can be developed in the more general context of dependent random variables. In addition, the underlying principles are robust enough to be applied to other types of distributional approximations. One significant application is Poisson approximations. The monograph [BC05] contains a excellent exposition on this topic.

(iii) There are extensions of the one-dimensional theory we developed here to multivariate Gaussian approximations. An effective way of performing the analysis in the multidimensional context is to make use of modern tools from Gaussian analysis such as the Malliavin calculus. The monograph [NP12] contains a nice introduction of relevant techniques.

(iv) There is a modern viewpoint of Stein's method, known as the *generator approach*, that leads to deeper applications such as distributional approximations for stochastic processes. Suppose that μ is the target distribution that one wishes

to approximate. Here μ can be defined on $\mathbb{R}$, $\mathbb{R}^n$ or even an infinite dimensional space (e.g. the space of continuous paths in the context of stochastic process approximations). As the first step in Stein's method, one needs to identify the Stein operator, say $\mathcal{A}$, which is an operator acting on a space of functions on S, so that

$$\mathbb{E}[\mathcal{A}f(Z)] = 0 \quad \forall f$$

uniquely characterises the distribution μ. The key point behind the generator approach is to realise μ as the invariant measure of some S-valued Markov process. The Stein operator $\mathcal{A}$ will then be given by the generator of this Markov process. It turns out that the associated Stein's equation can be studied effectively through the analysis of this Markov process. The monograph [BC05] also contains an introduction to this approach.

Appendix. A functional-analytic lemma for the L^1-Berry-Esseen estimate

We now provide the precise details that lead to the L^1-Berry-Esseen estimate (7.36) from Theorem 7.3. The key technical ingredient is a functional-analytic lemma which gives a representation of the L^1-norm from a duality perspective.

Lemma 7.5. *Let $F, G : \mathbb{R} \to \mathbb{R}$ be distribution functions with finite first moment, i.e. $\int_{\mathbb{R}} |x| dF(x)$ and $\int_{\mathbb{R}} |x| dG(x)$ are both finite. Let ψ be a bounded Borel-measurable function and define $\varphi(x) \triangleq \int_0^x \psi(u)du$. Then*

$$\int_{\mathbb{R}} \varphi(x)dF(x) - \int_{\mathbb{R}} \varphi(x)dG(x) = \int_{\mathbb{R}} \psi(x)\big(G(x) - F(x)\big)dx. \tag{7.37}$$

Proof. One can write

$$\begin{aligned}
&\int_{\mathbb{R}} \varphi(x)dF(x) \\
&= \int_{\mathbb{R}} \Big(\int_0^x \psi(u)du \Big) dF(x) \\
&= -\int_{-\infty}^0 \int_x^0 \psi(u)du dF(x) + \int_0^\infty \int_0^x \psi(u)du dF(x) \\
&= -\int_{-\infty}^0 \psi(u)F(u)du + \int_0^\infty \psi(u)(1 - F(u))du,
\end{aligned}$$

where we used Fubini's theorem to reach the last equality . Similarly, one has

$$\int_{\mathbb{R}} \varphi(x) dG(x) = -\int_{-\infty}^{0} \psi(u) G(u) du + \int_{0}^{\infty} \psi(u)(1 - G(u)) du.$$

By subtracting the two results, one obtains (7.37). □

Lemma 7.6. *Let $Q : \mathbb{R} \to \mathbb{R}$ be a function which contains at most countably many discontinuity points and suppose that $\int_{\mathbb{R}} |Q(x)| dx < \infty$. Then*

$$\int_{\mathbb{R}} |Q(x)| dx = \sup \big\{ \big| \int_{\mathbb{R}} \varphi(x) Q(x) dx \big| : \varphi \in \mathcal{C}_b(\mathbb{R}), \|\varphi\|_\infty \leqslant 1 \big\}. \tag{7.38}$$

Proof. For any φ with $\|\varphi\|_\infty \leqslant 1$, one has

$$\big| \int_{\mathbb{R}} \varphi(x) Q(x) dx \big| \leqslant \|\varphi\|_\infty \cdot \int_{\mathbb{R}} |Q(x)| dx \leqslant \int_{\mathbb{R}} |Q(x)| dx.$$

Therefore, the right hand side of (7.38) is not greater than the left hand side. To prove the other direction, first note that

$$\int_{\mathbb{R}} |Q(x)| dx = \int_{\mathbb{R}} \operatorname{sgn}(Q(x)) \cdot Q(x) dx,$$

where $\operatorname{sgn}(x)$ is the function defined by

$$\operatorname{sgn}(x) \triangleq \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ 0, & x = 0. \end{cases}$$

We set $\psi(x) \triangleq \operatorname{sgn}(Q(x))$. The main difficulty here is that $\psi(x)$ is not a continuous function. One thus needs to construct $\mathcal{C}_b(\mathbb{R})$-approximations.

For this purpose, for each $\varepsilon > 0$, let us choose a continuous function $\rho_\varepsilon : \mathbb{R} \to \mathbb{R}$ such that

$$\rho_\varepsilon \geqslant 0, \ \int_{\mathbb{R}} \rho_\varepsilon(x) dx = 1$$

and $\rho_\varepsilon(x) = 0$ for any $|x| > \varepsilon$. Define ψ_ε to be the convolution of ψ and ρ_ε, i.e.

$$\psi_\varepsilon(x) \triangleq \int_{\mathbb{R}} \psi(x - y) \rho_\varepsilon(y) dy = \int_{\mathbb{R}} \rho_\varepsilon(x - y) \psi(y) dy. \tag{7.39}$$

Using the latter expression, one can check that ψ_ε is continuous. Since $|\psi| \leqslant 1$, one also knows that

$$|\psi_\varepsilon(x)| \leqslant \int_{\mathbb{R}} |\psi(x-y)| \cdot \rho_\varepsilon(y)dy \leqslant \int_{\mathbb{R}} \rho_\varepsilon(y)dy = 1.$$

Therefore, $\psi_\varepsilon \in \mathcal{C}_b(\mathbb{R})$. It may not be true that $\psi_\varepsilon(x) \to \psi(x)$ for every $x \in \mathbb{R}$ as $\varepsilon \to 0$. However, we claim that

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}} \psi_\varepsilon(x)Q(x)dx = \int_{\mathbb{R}} \psi(x)Q(x)dx. \tag{7.40}$$

If one can prove this, it is then immediate that the left hand side of (7.38) is not greater than the right hand side, and the proof of (7.38) will be finished.

To prove (7.40), let $\mathcal{C}_Q$ be the set of continuity points of Q. The crucial observation is that

$$\psi_\varepsilon(x)\mathbf{1}_{\{x:Q(x)\neq 0\}\cap \mathcal{C}_Q}(x) \to \psi(x)\mathbf{1}_{\{x:Q(x)\neq 0\}\cap \mathcal{C}_Q}(x) \tag{7.41}$$

as $\varepsilon \to 0$. Indeed, if x is a continuity point of Q and $Q(x) \neq 0$, one knows by continuity that $Q(x)$ does not change sign in a small neighbourhood of x. Suppose that $Q(x) > 0$ (so that $\psi(x) = 1$). Then there exists $\delta > 0$ such that $Q(x-y) > 0$ for any $y \in (-\delta, \delta)$. In particular,

$$\psi(x-y) = \operatorname{sgn}(Q(x-y)) = 1, \quad y \in (-\delta, \delta).$$

According to the constructions of ρ_ε and ψ_ε, for any $\varepsilon < \delta$ one has

$$\psi_\varepsilon(x) = \int_{\mathbb{R}} \psi(x-y)\rho_\varepsilon(y)dy = \int_{(-\varepsilon,\varepsilon)} 1 \cdot \rho_\varepsilon(y)dy = 1 = \psi(x),$$

which trivially implies that $\psi_\varepsilon(x) \to \psi(x)$ as $\varepsilon \to 0$. Therefore, (7.41) holds. The dominated convergence theorem then implies that

$$\int_{\{x:Q(x)\neq 0\}\cap \mathcal{C}_Q} \psi_\varepsilon(x)Q(x)dx \to \int_{\{x:Q(x)\neq 0\}\cap \mathcal{C}_Q} \psi(x)Q(x)dx.$$

On the other hand, since $\mathcal{C}_Q^c$ is at most countable (and thus has zero Lebesgue measure), one knows that

$$\int_{\mathbb{R}} \psi_\varepsilon(x)Q(x)dx = \int_{\{x:Q(x)\neq 0\}} \psi_\varepsilon(x)Q(x)dx = \int_{\{x:Q(x)\neq 0\}\cap \mathcal{C}_Q} \psi_\varepsilon(x)Q(x)dx.$$

The same property holds for $\psi(x)$. Therefore, one concludes that (7.40) holds. □

Finally, we are in a position to complete the proof of the L^1-Berry-Esseen estimate.

Proof of Corollary 7.2. In Theorem 7.3, we have shown that

$$\Big| \int_{\mathbb{R}} \varphi(x) dF_n(x) - \int_{\mathbb{R}} \varphi(x) dG(x) \Big| \leqslant 9 \|\varphi'\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

for any φ with $\varphi' \in \mathcal{C}_b(\mathbb{R})$. Using Lemma 7.5 and setting $\psi \triangleq \varphi'$, one concludes that

$$\Big| \int_{\mathbb{R}} \psi(x) \big(F(x) - G(x)\big) dx \Big| \leqslant 9 \|\psi\|_\infty \cdot \frac{\sum_{m=1}^n \tau_m^3}{\Sigma_n^3}$$

for any $\psi \in \mathcal{C}_b(\mathbb{R})$. The L^1-estimate (7.36) then follows from Lemma 7.6.

□

8 Discrete-time martingales

In previous chapters, we have been mostly dealing with sequences of independent random variables. Another type of random sequences that exhibit rich and interesting properties are *martingale sequences*. Martingale theory is of fundamental importance for several reasons. Apart from its wide applications in applied areas (e.g. statistics, finance, physics, biology etc.), martingale methods have also been proven to be powerful in modern mathematics. Within the realm of probability theory, one of its most important applications is the study of stochastic calculus (stochastic integration and differential equations). Outside probability theory, a notable example is its use in obtaining weak solutions to parabolic PDEs. Applications of martingale theory to other mathematical areas such as number theory, group theory, complex analysis, harmonic analysis, differential geometry have been explored in depth since the second half of the last century. It will continue to be a rich area of study in modern probability theory.

In this chapter, we study the basic theory of discrete-time martingales in depth. There are three fundamental results we shall establish:

(i) The optional sampling theorem;
(ii) The maximal and L^p-inequalities;
(iii) The martingale convergence theorem.

The development of martingale theory was largely due to J.B. Doob around the 1950s. Following [Wil91], we will take a unified (and rather enlightening) approach to study these basic results (the martingale transform).

In Section 8.1, we begin by introducing the definition of (sub/super)martingales and related concepts. In Section 8.2, we discuss the martingale transform which is the core technique for proving the basic martingale theorems. In Sections 8.3–8.5, we develop the aforementioned three results respectively. In Section 8.6, we study uniformly integrable martingales which exhibit better convergence properties. In Section 8.7, we discuss a few enlightening applications of martingale methods.

8.1 Martingales, submartingales and supermartingales

Heuristically, a martingale models the wealth sequence of a fair game. The description of "fairness" relies on the notion of "information growth" in the evolution of time: *filtration*. We first define it mathematically before discussing the conpect of a martingale.

8.1.1 Filtration and adaptedness

Let $(\Omega, \mathcal{F})$ be a given measurable space. Let $X = \{X_n : n \geqslant 0\}$ be a sequence of random variables on it.

Definition 8.1. A *filtration* over $(\Omega, \mathcal{F})$ is a sequence $\{\mathcal{F}_n : n \geqslant 0\}$ of sub-σ-algebras of $\mathcal{F}$ such that

$$\mathcal{F}_n \subseteq \mathcal{F}_{n+1} \quad \forall n \geqslant 0.$$

The sequence X is said to be *adapted to the filtration* $\{\mathcal{F}_n\}$ (or simply $\{\mathcal{F}_n\}$-*adapted*), if X_n is $\mathcal{F}_n$-measurable (i.e. $X_n^{-1}B \in \mathcal{F}_n$ for all $B \in \mathcal{B}(\mathbb{R})$) for every n.

Heuristically, $\mathcal{F}_n$ represents the accumulative information up to time n. Adaptedness means that given the information up to time n, one can determine the exact value of X_n (indeed, the values of $X_0, \cdots, X_n$). Given a random sequence, there is a canonical filtration with respect to which X is adapted.

Definition 8.2. Let $X = \{X_n \geqslant 0\}$ be a sequence of random variables on $(\Omega, \mathcal{F})$. The *natural filtration* of X is defined by

$$\mathcal{F}_n^X \triangleq \sigma(X_0, X_1, \cdots, X_n) \triangleq \sigma\big(\{X_k^{-1}B : B \in \mathcal{B}(\mathbb{R}), 0 \leqslant k \leqslant n\}\big)$$

for each $n \geqslant 0$.

By definition, it is obvious that X is adapted to its natural filtration.

8.1.2 Definition of (sub/super)martingale sequences

We now give the precise definition of a martingale. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a given filtration $\{\mathcal{F}_n : n \geqslant 0\}$.

Definition 8.3. A sequence $X = \{X_n : n \geqslant 0\}$ of random variables is called an $\{\mathcal{F}_n\}$-*martingale* (respectively, a *submartingale* / *supermartingale*) if the following properties hold true:

(i) X is $\{\mathcal{F}_n\}$-adapted;
(ii) X_n is integrable for every $n \geqslant 0$;
(iii) for every $n \geqslant 0$, one has

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n \quad \text{(respectively "}\geqslant\text{" / "}\leqslant\text{")}. \tag{8.1}$$

The martingale property (8.1) is equivalent to $\mathbb{E}[X_m|\mathcal{F}_n] = X_n$ for all $m > n$ (why?). Heuristically, this property suggests that given the information up to the present time, the rational expectation of future wealth should simply be the current wealth. In particular, one will not gain or lose money under such a situation. Therefore, a martingale models a "fair game" in a mathematical way.

Remark 8.1. Definition 8.3 relies crucially on the underlying filtration. As a shorthanded convention, we sometimes say that $\{X_n, \mathcal{F}_n\}$ is a (sub/super)martingale. Note that a martingale with respect to one filtration may fail to be a martingale with respect to another.

Example 8.1. Let $X = \{X_n : n = 1, 2, \cdots\}$ denote an i.i.d. sequence of random variables with distribution

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}.$$

We consider the natural filtration associated with the sequence X:

$$\mathcal{F}_0 \triangleq \{\emptyset, \Omega\};\ \mathcal{F}_n \triangleq \sigma(X_1, \cdots, X_n), \quad n \geqslant 1.$$

Define $S_n \triangleq X_1 + \cdots + X_n$ ($S_0 \triangleq 0$). Then $\{S_n, \mathcal{F}_n : n \geqslant 0\}$ is a martingale. Indeed, it is clear that S_n is integrable and $\mathcal{F}_n$-measurable. To obtain the martingale property (8.1), for any $n \geqslant 0$ one has

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = \mathbb{E}[S_n + X_{n+1}|\mathcal{F}_n] = \mathbb{E}[S_n|\mathcal{F}_n] + \mathbb{E}[X_{n+1}|\mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] = S_n.$$

A simple way of constructing submartingales from a given martingale is to compose with convex functions.

Proposition 8.1. *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a martingale (respectively, a submartingale). Suppose that $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function (respectively, a convex and increasing function). If $\varphi(X_n)$ is integrable for every n, then $\{\varphi(X_n), \mathcal{F}_n : n \geqslant 0\}$ is a submartingale.*

Proof. The adaptedness and integrability conditions are clearly satisfied. To see the submartingale property, one applies Jensen's inequality (2.25) to find that

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] \geqslant \varphi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) \geqslant \varphi(X_n)$$

for all n. □

Example 8.2. The functions

$$\varphi_1(x) = x^+ \triangleq \max\{x, 0\},\ \varphi_2(x) = |x|^p\ (p \geqslant 1)$$

are convex on $\mathbb{R}$. As a result, if $\{X_n, \mathcal{F}_n\}$ is a martingale, then $\{X_n^+\}$ and $\{|X_n|^p\}$ ($p \geqslant 1$) are both $\{\mathcal{F}_n\}$-submartingales, provided that $\mathbb{E}[|X_n|^p] < \infty$ for every n.

8.2 A fundamental technique: the martingale transform

We discuss a particularly useful construction of martingales. This also provides an essential tool for proving several basic theorems in martingale theory. We begin by introducing a few definitions.

Definition 8.4. Let $\{\mathcal{F}_n : n \geqslant 0\}$ be a filtration. A random sequence $\{C_n : n \geqslant 1\}$ is said to be $\{\mathcal{F}_n\}$-*predictable* if C_n is $\mathcal{F}_{n-1}$-measurable for every $n \geqslant 1$.

Heuristically, predictability means that the future value C_{n+1} can be determined by the information up to the present time n.

Let $\{X_n : n \geqslant 0\}$ and $\{C_n : n \geqslant 1\}$ be two random sequences. We define another sequence $\{Y_n : n \geqslant 0\}$ by $Y_0 \triangleq 0$ and

$$Y_n \triangleq \sum_{k=1}^{n} C_k(X_k - X_{k-1}), \quad n \geqslant 1.$$

Definition 8.5. The sequence $\{Y_n : \ n \geqslant 0\}$ is called the *martingale transform* of $\{X_n\}$ by $\{C_n\}$. We often write Y_n as $(C \bullet X)_n$.

Remark 8.2. The martingale transform is a discrete-time version of stochastic integration as seen from the following continuous / discrete comparison:

$$\int C_t dX_t \approx \sum_k C_k(X_{t_k} - X_{t_{k-1}}).$$

The result below justifies the name of martingale transform. It plays a fundamental role in our later study.

Theorem 8.1. *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a martingale (respectively, submartingale / supermartingale) and let $\{C_n : n \geqslant 1\}$ be an $\{\mathcal{F}_n\}$-predictable random sequence which is uniformly bounded (respectively, uniformly bounded and non-negative). Then the martingale transform $\{(C \bullet X)_n, \mathcal{F}_n : n \geqslant 0\}$ is a martingale (respectively, submartingale / supermartingale).*

Proof. We only consider the martingale case. Adaptedness and integrability are clear. To check the martingale property, one computes

$$\begin{aligned}
\mathbb{E}[(C \bullet X)_{n+1}|\mathcal{F}_n] &= \mathbb{E}[(C \bullet X)_n + C_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] \\
&= (C \bullet X)_n + C_{n+1} \cdot \big(\mathbb{E}[X_{n+1}|\mathcal{F}_n] - X_n\big) \\
&= (C \bullet X)_n,
\end{aligned}$$

where we used the predictability of $\{C_n\}$ to reach the second last identity. □

Remark 8.3. The boundedness of $\{C_n\}$ is not an essential assumption. It is imposed to ensure the integrability of Y_n.

The following intuition behind the martingale transform is particularly useful. Suppose that you are gambling over the time horizon $\{1, 2, \cdots\}$.The quantity C_n represents your stake at game n. Predictability means that you are making your next decision on the stake amount C_{n+1} based on the information $\mathcal{F}_n$ observed up to the present round n. The quantity $X_n - X_{n-1}$ represents your winning at game n per unit stake. As a result, Y_n is your total winning up to time n. Theorem 8.1 asserts that if the game is fair (i.e. $\{X_n, \mathcal{F}_n\}$ is a martingale) and if you are playing the game based on the intrinsic information carried by the game itself (i.e. predictability), then you cannot beat fairness (i.e. your wealth process $\{Y_n, \mathcal{F}_n\}$ is also a martingale).

As we will see, the martingale transform can be used as a unified approach to establish several fundamental results in martingale theory.

8.3 Doob's optional sampling theorem

Apart from deterministic times, in practice it is often useful to consider random times as well. For instance, when predicting the behaviour of a volcano one relies on the dynamical data / information up to the next time of its eruption. However, the next eruption day is itself a random variable. In this case, one is talking about the accumulative information up to a random time. It is natural to expect that the martingale property remains valid even when one evaluates the martingale sequence at a suitable random time and looks at information up to such a time. The precise mathematical formulation of this result is content of the *optional sampling theorem*. Before discussing it, we first introduce the concept of *stopping time*.

8.3.1 Stopping times

Let $(\Omega, \mathcal{F})$ be a measurable space equipped with a filtration $\{\mathcal{F}_n : n \geqslant 0\}$. By a random time we shall mean a function $\tau : \Omega \to \mathbb{N} \cup \{\infty\} = \{0, 1, 2, \cdots, \infty\}$ (one can of course consider time continuously but in our current study time is always assumed to be discrete). Allowing τ to achieve ∞-value is convenient since it may not always be the case that τ is finite. In the volcano example, it is theoretically possible that the volcano never erupts any more in the future ($\tau(\omega) = +\infty$). Similar to the notion of random variables, in order to study its distributional properties one needs to impose suitable measurability condition on

a random time. Such a condition should respect the information flow given by the filtration $\{\mathcal{F}_n\}$.

Definition 8.6. A random time $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ is said to be an $\{\mathcal{F}_n\}$-*stopping time*, if

$$\{\omega \in \Omega : \tau(\omega) \leqslant n\} \in \mathcal{F}_n \quad \forall n \geqslant 0. \tag{8.2}$$

The idea behind the measurability condition (8.2) can be described as follows. Suppose that one is given the accumulative information up to time n. Then one is able to determine whether the event $\{\tau \leqslant n\}$ occurs or not. If it is the case that $\tau \leqslant n$, one can actually further determine the exact value of τ. Indeed, since one can decide whether $\{\tau \leqslant k\}$ happens for every $k \leqslant n$ (the information up to k is also known as part of $\mathcal{F}_n$), the value of τ can then be extracted from the first $k \leqslant n$ such that $\{\tau \leqslant k\}$ happens while $\{\tau \leqslant k-1\}$ fails. On the other hand, if it is the case that $\tau > n$, no further implication on the value of τ can be made. In the volcano example, if one has observed its activity continuously for 100 days, one certainly knows whether the volcano has erupted within this period of 100 days (i.e. whether $\tau \leqslant 100$ or not). If it does, the given data should further indicate the exact day of eruption, while if does not one cannot determine its next eruption day by using the given information.

Proposition 8.2. *A random time τ is a stopping time if and only if $\{\tau = n\} \in \mathcal{F}_n$ for all n.*

Proof. Sufficiency. Suppose that $\{\tau = n\} \in \mathcal{F}_n$ for all n. Then

$$\{\tau \leqslant n\} = \bigcup_{k=0}^{n} \{\tau = k\} \in \mathcal{F}_n,$$

since $\{\tau = k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$ for all $k \leqslant n$.

Necessity. Suppose that τ is a stopping time. Then

$$\{\tau = n\} = \{\tau \leqslant n\} \backslash \{\tau \leqslant n-1\} \in \mathcal{F}_n,$$

since $\{\tau \leqslant n-1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$. □

Apparently, every deterministic time is an $\{\mathcal{F}_n\}$-stopping time. Moreover, one can construct new stopping times from given ones. We use $a \wedge b$ (respectively, $a \vee b$) to denote the minimum (respectively, the maximum) between two numbers a, b.

Proposition 8.3. *Suppose that σ, τ, τ_m $(m \geqslant 1)$ are $\{\mathcal{F}_n\}$-stopping times. Then*

$$\sigma + \tau,\ \sigma \wedge \tau,\ \sigma \vee \tau,\ \sup_m \tau_m$$

are all $\{\mathcal{F}_n\}$-stopping times.

Proof. We only consider the first and the last cases (the other two are left as an exercise). For $\sigma + \tau$, one uses Proposition 8.2:

$$\{\sigma + \tau = n\} = \bigcup_{k=0}^{n} \big(\{\sigma = k\} \cap \{\tau = n - k\}\big) \in \mathcal{F}_n.$$

For $\sup_m \tau_m$, one observes that

$$\big\{\sup_m \tau_m \leqslant n\big\} = \bigcap_{m=1}^{\infty} \{\tau_m \leqslant n\} \in \mathcal{F}_n.$$

□

It is helpful to re-examine the above property from the heuristic perspective. Suppose that one know the accumulative information up to time n. The criterion of being a stopping time is to see if one can decide whether $\{\sigma+\tau \leqslant n\}$ happens or not. Since σ, τ are both stopping times, the following scenarios are all decidable:

$$\sigma \leqslant n,\ \sigma > n,\ \tau \leqslant n,\ \tau > n.$$

If it is determined that either $\{\sigma > n\}$ or $\{\tau > n\}$ happens, then one decides that $\sigma + \tau > n$ since both σ, τ are non-negative. If it is determined that $\sigma \leqslant n$ and $\tau \leqslant n$, from earlier discussion the exact values of σ and τ are both decidable. As a result, the value of $\sigma + \tau$ can then be determined, which certainly allows one to further decide if $\{\sigma + \tau \leqslant n\}$ happens or not.

One can use this kind of heuristic argument to discuss the other cases in Proposition 8.3. It also allows one to see e.g. why $\sigma - \tau$ may not necessarily be a stopping time (assuming $\sigma \geqslant \tau$). Indeed, suppose again that the information up to n is presented. If one finds that $\sigma \leqslant n$, then $\{\sigma - \tau \leqslant n\}$ happens. However, if one finds that $\sigma > n$, no further implication on the value of σ can be obtained. In this case, $\sigma - \tau$ can either be smaller or larger than n and the occurrence of $\{\sigma - \tau \leqslant n\}$ is thus not decidable.

Example 8.3. Consider the random experiment of tossing a coin repeatedly. Let $\mathcal{F}_n$ denote the σ-algebra generated by results of the first n outcomes. Define τ to be the first time that a "Head" appears. Then τ is an $\{\mathcal{F}_n\}$-stopping time. Indeed,

$$\{\tau \leqslant n\} = \{\text{at leat one H appears among the first } n \text{ tosses}\} \in \mathcal{F}_n.$$

8.3.2 The σ-algebra at a stopping time

With the notion of a stopping time τ, it becomes natural to talk about the accumulative information up to τ. Similar to $\mathcal{F}_n$, this should be a suitable sub-σ-algebra of $\mathcal{F}$. The essential idea behind defining this σ-algebra is described as follows. First of all, an event $A \in \mathcal{F}_\tau$ means that knowing the information up to τ allows one to determine whether A happens or not. To rephrase this point properly in terms of the filtration $\{\mathcal{F}_n\}$, let n be a given fixed deterministic time. Suppose that one knows the information up to time n. Since τ is a stopping time, one can determine whether $\{\tau \leqslant n\}$ has occurred or not. If it is the first case, the information up to τ is then known since one is given the information up to n and has already determined that $\tau \leqslant n$. In this scenario, one can decide whether A occurs or not. If it happens to be the second case ($\tau > n$), since one only has the information up to n, the information over the period from $n+1$ to τ is missing and one should not be able to decide whether A occurs in this scenario. To summarise, given the information up to time n, it is only in the scenario $\{\tau \leqslant n\}$ that one can determine whether A occurs or not. This heuristic argument leads to the following precise mathematical definition.

Definition 8.7. Let τ be a given $\{\mathcal{F}_n\}$-stopping time. The *σ-algebra at the stopping time* τ is defined by

$$\mathcal{F}_\tau \triangleq \{A \in \mathcal{F} : A \cap \{\tau \leqslant n\} \in \mathcal{F}_n \quad \forall n \geqslant 0\}.$$

The following fact justifies the definition of $\mathcal{F}_\tau$.

Proposition 8.4. *The set class $\mathcal{F}_\tau$ is a sub-σ-algebra of $\mathcal{F}$.*

Proof. (i) Since τ is a stopping time, for each $n \geqslant 0$ one has

$$\Omega \cap \{\tau \leqslant n\} = \{\tau \leqslant n\} \in \mathcal{F}_n.$$

Therefore, $\Omega \in \mathcal{F}_\tau$.

(ii) Suppose $A \in \mathcal{F}_\tau$. Given an arbitrary $n \geqslant 0$, note that both of $\{\tau \leqslant n\}$ and $A \cap \{\tau \leqslant n\}$ belong to $\mathcal{F}_n$. As a result,

$$A^c \cap \{\tau \leqslant n\} = \{\tau \leqslant n\} \backslash \big(A \cap \{\tau \leqslant n\}\big) \in \mathcal{F}_n.$$

In particular, $A^c \in \mathcal{F}_\tau$.

(iii) Let $A_m \in \mathcal{F}_\tau$ ($m \geqslant 1$). For each $n \geqslant 0$, one has

$$\big(\bigcup_{m=1}^{\infty} A_m\big) \cap \{\tau \leqslant n\} = \bigcup_{m=1}^{\infty} \big(A_m \cap \{\tau \leqslant n\}\big) \in \mathcal{F}_n,$$

since each $A_m \cap \{\tau \leqslant n\} \in \mathcal{F}_n$. Therefore, $\cup_m A_m \in \mathcal{F}_\tau$. □

8.3.3 The optional sampling theorem

The optional sampling theorem asserts that the martingale property (8.1) remains valid when one samples along suitable stopping times. To elaborate this fact, let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a (sub/super)martingale and let τ be an $\{\mathcal{F}_n\}$-stopping time. We introduce the *stopped process*

$$X_n^\tau \triangleq X_{\tau\wedge n} = \begin{cases} X_n, & n \leqslant \tau, \\ X_\tau, & n > \tau. \end{cases}$$

Note that $X_{\tau\wedge n}$ means the random variable $\omega \mapsto X_{\tau(\omega)\wedge n}(\omega)$.

Theorem 8.2. *The stopped process $\{X_n^\tau\}$ is an $\{\mathcal{F}_n\}$-(sub/super)martingale.*

Proof. The main idea is to represent X_n^τ as a martingale transform (and such a method will appear for many times later on!). To this end, we consider a gambling model where $X_n - X_{n-1}$ represents the winning at game n per unit stake. The gambling strategy is constructed as follows:

Keep playing unit stake from the beginning and quit immediately after the time τ.

Mathematically, the strategy is given by

$$C_n \triangleq \mathbf{1}_{\{n\leqslant\tau\}}, \quad n \geqslant 1.$$

Since

$$\{C_n = 1\} = \{n \leqslant \tau\} = \{\tau \leqslant n-1\}^c \in \mathcal{F}_{n-1},$$

the sequence $\{C_n : n \geqslant 1\}$ is $\{\mathcal{F}_n\}$-predictable. The total winning up to time n is given by $(C \bullet X)_n = X_{\tau\wedge n} - X_0$. According to Theorem 8.1, one concludes that $\{X_n^\tau, \mathcal{F}_n\}$ is a martingale. □

Next, we consider the situation when one also stops the filtration at a stopping time. For simplicity, we only consider the case of bounded stopping times which is sufficient for most applications. Situations involving unbounded stopping times are often dealt with by truncation to the bounded case and then passing to the limit (see the application to random walks below for such a situation and also Propositions 8.6, 8.8 for suitable extensions to integrable stopping times).

Theorem 8.3 (The Optional Sampling Theorem)**.** *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a martingale. Suppose that σ, τ are two bounded $\{\mathcal{F}_n\}$-stopping times such that $\sigma \leqslant \tau$. Then X_σ (respectively, X_τ) is integrable, $\mathcal{F}_\sigma$-measurable (respectively, $\mathcal{F}_\tau$-measurable) and one has*

$$\mathbb{E}[X_\tau|\mathcal{F}_\sigma] = X_\sigma. \tag{8.3}$$

Proof. Assume that $\sigma \leqslant \tau \leqslant N$ for some constant $N \geqslant 0$. Integrability is seen by the following simple estimate:

$$\mathbb{E}[|X_\sigma|] = \sum_{n=0}^{N} \mathbb{E}\big[|X_\sigma| \mathbf{1}_{\{\sigma=n\}}\big] = \sum_{n=0}^{N} \mathbb{E}\big[|X_n| \mathbf{1}_{\{\sigma=n\}}\big] \leqslant \sum_{n=0}^{N} \mathbb{E}[|X_n|] < \infty.$$

To prove $\mathcal{F}_\sigma$-measurability, given $B \in \mathcal{B}(\mathbb{R})$ and any $n \geqslant 0$, one has

$$\begin{aligned} \{X_\sigma \in B\} \cap \{\sigma \leqslant n\} &= \bigcup_{k=0}^{n} \big(\{X_\sigma \in B\} \cap \{\sigma = k\}\big) \\ &= \bigcup_{k=0}^{n} \big(\{X_k \in B\} \cap \{\sigma = k\}\big) \in \mathcal{F}_n. \end{aligned}$$

By the definition of $\mathcal{F}_\sigma$, one concludes that $X_\sigma^{-1} B \in \mathcal{F}_\sigma$.

To obtain the martingale property (8.3), since X_σ is $\mathcal{F}_\sigma$-measurable, by the definition of conditional expectation one needs to show that

$$\int_F X_\tau d\mathbb{P} = \int_F X_\sigma d\mathbb{P} \quad \forall F \in \mathcal{F}_\sigma. \tag{8.4}$$

Let $F \in \mathcal{F}_\sigma$ be given fixed. Consider the gambling strategy of *playing unit stake at each time step from* $\sigma + 1$ *until* τ *under the occurrence of* F:

$$C_n \triangleq \mathbf{1}_F \mathbf{1}_{\{\sigma < n \leqslant \tau\}}, \quad n \geqslant 1.$$

The strategy sequence $\{C_n\}$ is $\{\mathcal{F}_n\}$-predictable as seen by

$$F \cap \{\sigma < n \leqslant \tau\} = F \cap \{\sigma \leqslant n-1\} \cap (\tau \leqslant n-1)^c \in \mathcal{F}_{n-1}.$$

The total winning by time N is $(C \bullet X)_N = (X_\tau - X_\sigma)\mathbf{1}_F$. According to Theorem 8.1, one concludes that $\{(C \bullet X)_n, \mathcal{F}_n\}$ is a martingale. In particular,

$$\mathbb{E}[(C \bullet X)_N] = \mathbb{E}[(X_\tau - X_\sigma)\mathbf{1}_F] = \mathbb{E}[(C \bullet X)_0] = 0,$$

which gives the desired property (8.4). □

Remark 8.4. Theorem 8.3 and its proof clearly extends to the case of sub/super martingales as well.

Remark 8.5. Sometimes a more useful form of (8.3) is obtained after taking expectation:

$$\mathbb{E}[X_\tau] = \mathbb{E}[X_\sigma] (= \mathbb{E}[X_0]). \tag{8.5}$$

Example 8.4. Consider the martingale $\{S_n\}$ defined by the simple random walk on $\mathbb{Z}$ (cf. Example 8.1 for the precise definition). Consider the stopping time

$$\tau \triangleq \inf\{n \geqslant 0 : S_n = 1\}.$$

One knows from (5.2) that $\tau < \infty$ a.s. As a result, S_τ is a.s. well-defined; indeed, by definition $S_\tau = 1$ trivially. In particular,

$$1 = \mathbb{E}[S_\tau] \neq \mathbb{E}[S_0] = 0.$$

In other words, the optional sampling theorem does not hold for this τ. The main issue here is that τ is not bounded (it is not even integrable).

An application: gambler's ruin

Imagine there are two gamblers A and B. Their initial capitals at time $n = 0$ are a and b respectively (a, b are given fixed positive integers). At each round $n \geqslant 1$, either A wins \$1 from B or the otherwise. Suppose that A's winning probability at each round is given by $p \in (0, 1)$. Different rounds are assumed to be independent. The game is finished if either one of them goes bankrupt. We are interested in computing *the average length of the game* and *the probability that Player A first goes bankrupt*.

We now introduce a suitable mathematical model to describe the underlying problem. Consider a random walk $S_n \triangleq X_1 + \cdots + X_n$ with initial position $S_0 \triangleq a$. Here $\{X_n : n \geqslant 1\}$ is an i.i.d. sequence with distribution

$$\mathbb{P}(X_1 = 1) = p, \ \mathbb{P}(X_1 = -1) = q \triangleq 1 - p.$$

Define

$$\tau = \inf\{n : S_n = 0 \text{ or } S_n = a + b\}.$$

The problem is thus about computing $\mathbb{E}[\tau]$ and $\gamma \triangleq \mathbb{P}(S_\tau = 0)$.

(i) *The case when* $p \neq 1/2$.

We first introduce some martingales associated with such a model. More specifically, we shall look for two martingales of the forms

$$M_n \triangleq \alpha^{S_n}, \ N_n \triangleq S_n - \beta n$$

respectively, where α, β are parameters to be determined. For M_n to be a martingale, one requires that

$$\mathbb{E}[\alpha^{S_{n+1}} | \mathcal{F}_n] = \mathbb{E}[\alpha^{S_n + X_{n+1}} | \mathcal{F}_n] = p\alpha^{S_n + 1} + q\alpha^{S_n - 1} = \alpha^{S_n},$$

where $\mathcal{F}_n \triangleq \sigma(X_1, \cdots, X_n)$. This implies that

$$\alpha p + \frac{q}{\alpha} = 1 \iff \alpha = 1 \text{ or } \alpha = \frac{q}{p}.$$

Of course the meaningful choice is the latter. Similarly, for N_n to be a martingale one needs

$$\mathbb{E}[S_{n+1} - \beta(n+1)|\mathcal{F}_n] = p(S_n + 1) + q(S_n - 1) - \beta n - \beta = S_n - \beta n.$$

This implies that $\beta = p - q$. To summarise, we have proved the following fact.

Lemma 8.1. *$M_n \triangleq (q/p)^{S_n}$ and $N_n \triangleq S_n - (p-q)n$ are both $\{\mathcal{F}_n\}$-martingales.*

For each $n \geqslant 1$, since $\tau \wedge n$ is a bounded stopping time, by applying the optional sampling theorem to the martingale N_n one finds that

$$\mathbb{E}[N_{\tau\wedge n}] = \mathbb{E}[N_0] \iff \mathbb{E}[S_{\tau\wedge n}] - (p-q)\mathbb{E}[\tau \wedge n] = a. \tag{8.6}$$

Since $0 \leqslant S_{\tau\wedge n} \leqslant a + b$, one has

$$\mathbb{E}[\tau] = \lim_{n\to\infty} \mathbb{E}[\tau \wedge n] \leqslant \frac{2a+b}{|p-q|} < \infty.$$

In particular, τ is finite a.s. (with probability one, one of the two players will go bankrupt). By taking $n \to \infty$ in (8.6), one obtains that

$$(p-q)\mathbb{E}[\tau] = \gamma \cdot 0 + (1-\gamma)\cdot(a+b) - a = b - (a+b)\gamma \tag{8.7}$$

Next, by applying optional sampling to the martingale $M_n = \alpha^{S_n}$ (with $\alpha \triangleq q/p$) one also has

$$\mathbb{E}[\alpha^{S_{\tau\wedge n}}] = \mathbb{E}[\alpha^{S_0}] = \alpha^a \quad \forall n.$$

Taking $n \to \infty$ yields

$$\gamma \cdot \alpha^0 + (1-\gamma)\cdot\alpha^{a+b} = \alpha^a.$$

Therefore,

$$\gamma = \frac{\alpha^a - \alpha^{a+b}}{1 - \alpha^{a+b}}.$$

By substituting this into (8.7), one finds that

$$\mathbb{E}[\tau] = \frac{1}{p-q}\Big(b - (a+b)\frac{\alpha^a - \alpha^{a+b}}{1-\alpha^{a+b}}\Big).$$

(ii) *The case when* $p = 1/2$.

In this case, we consider the martingales S_n and $S_n^2 - n$. By optional sampling, one has

$$\mathbb{E}[S_{\tau\wedge n}] = \mathbb{E}[S_0] = a$$

and

$$\mathbb{E}[S_{\tau\wedge n}^2 - \tau \wedge n] = \mathbb{E}[S_0^2] = a^2$$

for all n. By taking $n \to \infty$, in a similar way as before one finds that

$$\gamma = \frac{b}{a+b},\ \mathbb{E}[\tau] = ab.$$

Remark 8.6. With no surprise, one can check that

$$\frac{b}{a+b} = \lim_{p\to 1/2} \frac{\alpha^a - \alpha^{a+b}}{1-\alpha^{a+b}},\ ab = \lim_{p\to 1/2} \frac{1}{p-q}\big(b - (a+b)\frac{\alpha^a - \alpha^{a+b}}{1-\alpha^{a+b}}\big).$$

8.4 Doob's maximal inequality

In probability theory (in particular, in the study of stochastic processes), it is often useful to know how to control the supremum of a random sequence. This can be done in a neat and simple way in the (sub)martingale context with the aid of the optional sampling theorem. As a submartingale exhibits an increasing trend, it is not surprising that its running maximum can be controlled by the terminal value in a reasonable sense.

Theorem 8.4. *Let* $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ *be a submartingale. For all* $N \geqslant 0$ *and* $\lambda > 0$, *one has*

$$\mathbb{P}\big(\max_{0\leqslant n\leqslant N} X_n \geqslant \lambda\big) \leqslant \frac{\mathbb{E}[X_N^+]}{\lambda}. \tag{8.8}$$

Proof. Let $\sigma \triangleq \inf\{n \leqslant N : X_n \geqslant \lambda\}$ denote the first time (up to N) that X_n exceeds the level λ. We set $\sigma = N$ if no such $n \leqslant N$ exists. Clearly σ is an $\{\mathcal{F}_n\}$-stopping time bounded by N. By taking expectation on both sides of (8.3) (in the submartingale case), one obtains that

$$\mathbb{E}[X_N] \geqslant \mathbb{E}[X_\sigma]. \tag{8.9}$$

On the other hand, one can write

$$X_\sigma = X_\sigma \mathbf{1}_{\{X_N^* \geqslant \lambda\}} + X_\sigma \mathbf{1}_{\{X_N^* < \lambda\}}$$

where

$$X_N^* \triangleq \max_{0 \leqslant n \leqslant N} X_n.$$

On the event $\{X_N^* \geqslant \lambda\}$, the process X_n does exceed λ at some $n \leqslant N$ and thus $X_\sigma \geqslant \lambda$. On the event $\{X_N^* < \lambda\}$ no such exceeding occurs and thus $X_\sigma = X_N$ ($\sigma = N$ in this case). As a result, one has

$$\begin{aligned} \mathbb{E}[X_N] \geqslant \mathbb{E}[X_\sigma] &= \mathbb{E}\big[X_\sigma \mathbf{1}_{\{X_N^* \geqslant \lambda\}}\big] + \mathbb{E}\big[X_\sigma \mathbf{1}_{\{X_N^* < \lambda\}}\big] \\ &\geqslant \lambda \mathbb{P}\big(X_N^* \geqslant \lambda\big) + \mathbb{E}\big[X_N \mathbf{1}_{\{X_N^* < \lambda\}}\big]. \end{aligned}$$

It follows that

$$\lambda \mathbb{P}\big(X_N^* \geqslant \lambda\big) \leqslant \mathbb{E}[X_N] - \mathbb{E}[X_N \mathbf{1}_{\{X_N^* < \lambda\}}] = \mathbb{E}[X_N \mathbf{1}_{\{X_N^* \geqslant \lambda\}}] \leqslant \mathbb{E}[X_N^+], \tag{8.10}$$

which yields the desired inequality. □

Remark 8.7. Kolmogorov's maximal inequality (cf. Lemma 5.2) can be seen as a special case of Theorem 8.4. Indeed, let $\{X_n : n \geqslant 1\}$ be a sequence of independent random variables with mean zero and finite variance. Define $S_n \triangleq X_1 + \cdots + X_n$. Then $\{S_n\}$ is a martingale with respect to its natural filtration. Since $x \mapsto x^2$ is a convex function on $\mathbb{R}$, it follows from Proposition 8.1 that $\{S_n^2\}$ is a submartingale. According to Doob's maximal inequality (8.8), one has

$$\mathbb{P}\big(\max_{1 \leqslant k \leqslant n} |S_k| > \varepsilon\big) = \mathbb{P}\big(\max_{1 \leqslant k \leqslant n} S_k^2 > \varepsilon^2\big) \leqslant \frac{1}{\varepsilon^2}\mathbb{E}[S_n^2] = \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \mathrm{Var}[X_k].$$

An important corollary of the maximal inequality is the *L^p-inequality* for the running maximum. Due to the former result, it is not too surprising that the integrability of the running maximum can also be controlled by the integrability of the terminal value. Recall that $\|X\|_p \triangleq \mathbb{E}[|X|^p]^{1/p}$ denotes the *L^p-norm* ($p \geqslant 1$) of a random variable X and we say that $X \in L^p$ if $\|X\|_p < \infty$.

Corollary 8.1. *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a non-negative submartingale. Let $p > 1$ and suppose that $X_n \in L^p$ for all n. Then for every $N \geqslant 0$, one has*

$$\|\max_{0 \leqslant n \leqslant N} X_n\|_p \leqslant \frac{p}{p-1}\|X_N\|_p,$$

In particular, $\max_{0 \leqslant n \leqslant N} X_n \in L^p$.

In order to prove this result, we first need the following lemma.

Lemma 8.2. *Suppose that X, Y are two non-negative random variables such that*

$$\mathbb{P}(X \geqslant \lambda) \leqslant \frac{\mathbb{E}[Y\mathbf{1}_{\{X\geqslant\lambda\}}]}{\lambda} \quad \forall \lambda > 0. \tag{8.11}$$

Then for any $p > 1$, one has

$$\|X\|_p \leqslant q\|Y\|_p, \tag{8.12}$$

where $q \triangleq p/(p-1)$ (so that $1/p + 1/q = 1$).

Proof. Suppose $\|Y\|_p < \infty$ for otherwise the result is trivial. We write

$$\mathbb{E}[X^p] = \mathbb{E}\Big[\int_0^X p\lambda^{p-1}d\lambda\Big] = \mathbb{E}\left[\int_0^\infty p\lambda^{p-1}\mathbf{1}_{\{X\geqslant\lambda\}}d\lambda\right].$$

By using Fubini's theorem, one finds that

$$\begin{aligned}
\mathbb{E}[X^p] &= \int_0^\infty p\lambda^{p-1}\mathbb{P}(X \geqslant \lambda)d\lambda \\
&\leqslant \int_0^\infty p\lambda^{p-2}\mathbb{E}\big[Y\mathbf{1}_{\{X\geqslant\lambda\}}\big]d\lambda \\
&= \mathbb{E}\Big[Y\int_0^X p\lambda^{p-2}d\lambda\Big] \\
&= \frac{p}{p-1}\mathbb{E}[YX^{p-1}].
\end{aligned} \tag{8.13}$$

To proceed further, we assume for the moment that $X \in L^p$. According to Hölder's inequality (cf. (2.18)), one has

$$\mathbb{E}[YX^{p-1}] \leqslant \|Y\|_p\|X^{p-1}\|_q = \|Y\|_p\|X\|_p^{p-1}.$$

The inequality (8.12) thus follows by diving $\|X\|_p^{p-1}$ to the left hand side of (8.13). If $\|X\|_p = \infty$, we let $X^N \triangleq X \wedge N$ ($N \geqslant 1$). By considering the cases $\lambda > N$ and $\lambda \leqslant N$ separately, it is not hard to see that the condition (8.11) holds for the pair (X^N, Y). The desired inequality (8.12) follows by first considering X^N and then applying the monotone convergence theorem. □

Proof of Corollary 8.1. Let us write $X_N^* \triangleq \max\limits_{0\leqslant n\leqslant N} X_n$. We have shown in (8.10) that

$$\mathbb{P}(X_N^* \geqslant \lambda) \leqslant \frac{\mathbb{E}[X_N\mathbf{1}_{\{X_N^*\geqslant\lambda\}}]}{\lambda}.$$

In particular, the condition (8.11) holds with $(X, Y) = (X_N^*, X_N)$. The result follows immediate from Lemma 8.2.

□

8.5 The martingale convergence theorem

The (sub/super)martingale property (8.1) exhibits certain kind of monotone behaviour. It is thus reasonable to expect that a (sub/super)martingale converges in a suitable sense under some boundedness condition. To make this idea mathematically precise, one is led to the *martingale convergence theorem.*

8.5.1 A general strategy for proving almost sure convergence

Before establishing such a theorem, we first explain a general strategy of proving almost sure convergence. Let $X = \{X_n : n \geqslant 0\}$ be a random sequence on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given fixed $\omega \in \Omega$, the real sequence $X_n(\omega)$ is convergent if and only if

$$\underline{\lim}_{n\to\infty} X_n(\omega) = \overline{\lim}_{n\to\infty} X_n(\omega).$$

Here we take the convention that $X_n \to \infty$ or $X_n \to -\infty$ is also considered as being convergent. Therefore,

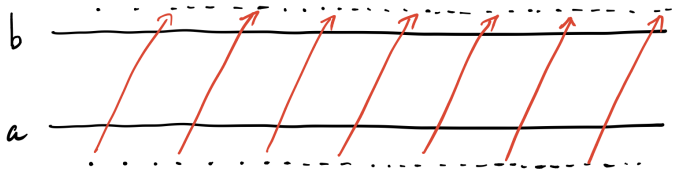
$$\begin{aligned}\{X_n \text{ is not convergent}\} &\subseteq \big\{ \underline{\lim}_{n\to\infty} X_n < \overline{\lim}_{n\to\infty} X_n \big\} \\ &\subseteq \bigcup_{\substack{a<b \\ a,b\in\mathbb{Q}}} \big\{ \underline{\lim}_{n\to\infty} X_n < a < b < \overline{\lim}_{n\to\infty} X_n \big\}.\end{aligned}$$

In order to prove that X_n converges a.s., it suffices to show that

$$\mathbb{P}\big(\underline{\lim}_{n\to\infty} X_n < a < b < \overline{\lim}_{n\to\infty} X_n \big) = 0 \tag{8.14}$$

for every pair of given numbers $a < b$.

Here is the key observation. Due to the definitions of the liminf and limsup, the event in (8.14) implies that there is a subsequence of X_n lying below a and in the meanwhile there is another subsequence of X_n lying above b. This further implies that, as n increases there must be infinitely many *upcrossings* by the sequence X_n from below the level a to above the level b.

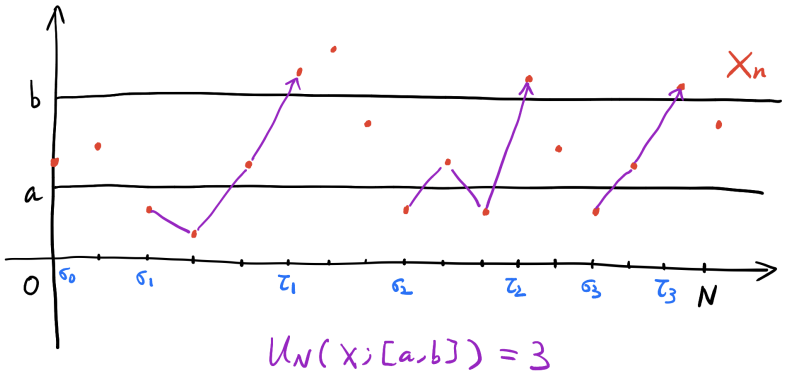


From this reasoning, the key step for proving the a.s. convergence of $\{X_n\}$ is to control its total upcrossing number with respect to the interval $[a, b]$. More specifically, one is led to showing that with probability one, there are at most finitely many upcrossings with respect to $[a, b]$.

8.5.2 The upcrossing inequality

We now define the upcrossing number mathematically. Consider the following two sequences of random times: $\sigma_0 \triangleq 0$,

$$\begin{aligned}
\sigma_1 &\triangleq \inf\{n \geqslant 0 : X_n < a\}, & \tau_1 &\triangleq \inf\{n > \sigma_1 : X_n > b\},\\
\sigma_2 &\triangleq \inf\{n > \tau_1 : X_n < a\}, & \tau_2 &\triangleq \inf\{n > \sigma_2 : X_n > b\},\\
&\cdots & &\\
\sigma_k &\triangleq \inf\{n > \tau_{k-1} : X_n < a\}, & \tau_k &\triangleq \inf\{n > \sigma_k : X_n > b\},\\
&\cdots. & &
\end{aligned}$$



Definition 8.8. Given $N \geqslant 0$, the *upcrossing number* $U_N(X; [a, b])$ with respect to the interval $[a, b]$ by the sequence $\{X_n\}$ up to time N is define by the random number

$$U_N(X; [a, b]) \triangleq \sum_{k=1}^{\infty} \mathbf{1}_{\{\tau_k \leqslant N\}}.$$

Note that $U_N(X; [a, b]) \leqslant N/2$. Moreover, if $\{\mathcal{F}_n\}$ is a given filtration and X is $\{\mathcal{F}_n\}$-adapted, then σ_k, τ_k are $\{\mathcal{F}_n\}$-stopping times. In particular, $U_N(X; [a, b])$ is $\mathcal{F}_N$-measurable. The main result for controlling the quantity $U_N(X; [a, b])$ in the martingale context is stated as follows.

Proposition 8.5 (The Upcrossing Inequality). *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a supermartingale. Then the upcrossing number $U_N(X; [a,b])$ satisfies the following inequality:*

$$\mathbb{E}[U_N(X;[a,b])] \leqslant \frac{\mathbb{E}[(X_N - a)^-]}{b-a}, \tag{8.15}$$

where $x^- \triangleq \max\{-x, 0\}$.

Proof. We again use the method of martingale transform. This time we construct a gambling strategy as follows: repeat the following two steps forever.

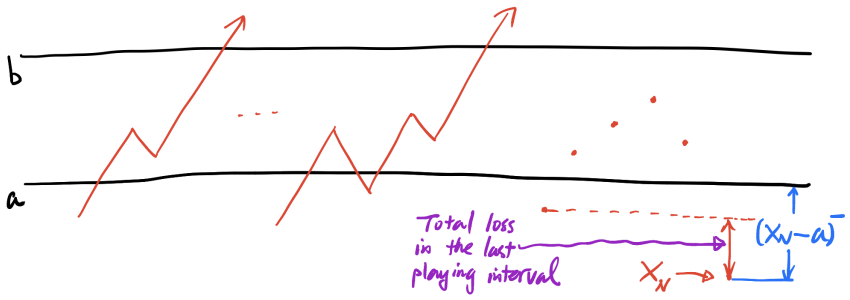
(i) Wait for X_n to drop below a;
(ii) play unit stakes onwards until X_n gets above b and then stop playing.

Mathematically, the strategy $\{C_n : n \geqslant 1\}$ is defined by the following equations

$$C_1 \triangleq \mathbf{1}_{\{X_0 < a\}};\ C_n \triangleq \mathbf{1}_{\{C_{n-1}=0\}}\mathbf{1}_{\{X_{n-1}<a\}} + \mathbf{1}_{\{C_{n-1}=1\}}\mathbf{1}_{\{X_{n-1}\leqslant b\}}, \quad n \geqslant 2.$$

Let $\{Y_n\}$ be the martingale transform of X_n by C_n. Then Y_N represents the total winning up to time N. Observe that Y_N comes from two parts:

(i) the playing intervals corresponding to complete upcrossings;
(ii) the last playing interval corresponding to the last incomplete upcrossing (which may possibly not exist).



The total winning Y_N from the first part is clearly bounded from below by $(b-a)U_N(X;[a,b])$. The total winning in the last playing interval (if it exists) is bounded from below by $-(X_N - a)^-$ (the worst scenario when a loss occurs). Consequently, one finds that

$$Y_N \geqslant (b-a)U_N(X;[a,b]) - (X_N - a)^-.$$

On the other hand, one checks by definition that $\{C_n\}$ is bounded, non-negative and $\{\mathcal{F}_n\}$-predictable. According to Theorem 8.1, $\{Y_n, \mathcal{F}_n\}$ is a supermartingale. Therefore,

$$\mathbb{E}[Y_N] \leqslant \mathbb{E}[Y_0] = 0.$$

Rearranging this relation gives the upcrossing inequality (8.15). □

Remark 8.8. There is also a version of the upcrossing inequality for the submartingale case. However, the proof of that case is quite different from what we gave here. Since they both lead to the same convergence theorem, we only consider the supermartingale case.

8.5.3 The convergence theorem

Since $U_N(X; [a, b])$ is increasing in N, one can define the total upcrossing number for all time to be

$$U_\infty(X; [a, b]) \triangleq \lim_{N\to\infty} U_N(X; [a, b]).$$

Now let us further assume that the supermartingale $\{X_n, \mathcal{F}_n\}$ is *bounded in* L^1, namely there exists $M > 0$ such that

$$\mathbb{E}[|X_n|] \leqslant M \quad \forall n \geqslant 0. \tag{8.16}$$

According to the upcrossing inequality, one finds that

$$\mathbb{E}[U_\infty(X; [a, b])] = \lim_{N\to\infty} \mathbb{E}[U_N(X; [a, b])] \leqslant \frac{M + |a|}{b - a} < \infty.$$

In particular, $U_\infty(X; [a, b]) < \infty$ a.s. It then follows from the relation

$$\Big\{ \varliminf_{n\to\infty} X_n < a < b < \varlimsup_{n\to\infty} X_n \Big\} \subseteq \{U_\infty(X; [a, b]) = \infty\}$$

that (8.14) holds. As a consequence, the sequence X_n is convergent a.s. Let us denote the limiting random variable as X_∞. From Fatou's lemma, under the L^1-boundedness assumption (8.16) one also finds that

$$\mathbb{E}[|X_\infty|] = \mathbb{E}\Big[\lim_{n\to\infty} |X_n| \Big] \leqslant \varliminf_{n\to\infty} \mathbb{E}[|X_n|] \leqslant M < \infty.$$

To summarise, we have thus established the following convergence result.

Theorem 8.5 (The Supermartingale Convergence Theorem)**.** *Let* $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ *be a supermartingale which is bounded in* L^1*. Then* X_n *converges a.s. to an integrable random variable* X_∞*.*

Remark 8.9. Since martingales are supermartingales and a submartingale is the negative of a supermartingale, it is immediate that the above convergence theorem is also valid for (sub)martingales.

8.6 Uniformly integrable martingales

Theorem 8.5 is an assertion about a.s. convergence of a (super)martingale sequence $\{X_n\}$. On the other hand, it is natural to ask if one also has $\mathbb{E}[X_n] \to \mathbb{E}[X_\infty]$ (in the martingale context, this is asking whether $\mathbb{E}[X_\infty] = \mathbb{E}[X_0]$)? Though it seems reasonable to expect so, the example below shows that this is surprisingly not true in general.

Example 8.5. Consider a sequence $\{X_n : n \geqslant 1\}$ of i.i.d. standard normal random variables. Let $S_n \triangleq X_1 + \cdots + X_n$ ($S_0 \triangleq 0$). It is routine to check that $M_n \triangleq e^{S_n - n/2}$ is a martingale with respect to its natural filtration. Since $\mathbb{E}[M_n] = 1$ and $M_n > 0$, it is trivially bounded in L^1. According to the martingale convergence theorem, M_n converges a.s. to some integrable random variable M_∞. We claim that $M_\infty = 0$ a.s.! Indeed, one knows from Proposition 5.1 (ii) (cf. (5.2)) that with probability one, $S_n < 0$ along some subsequence of times, say $n_k \uparrow \infty$ (n_k depends on ω). It follows that

$$0 < M_{n_k} \leqslant e^{-n_k/2} \to 0$$

as $k \to \infty$. As a result, M_∞ has to be zero (since the limit does not depend on the choice of subsequences). This example shows that

$$1 = \mathbb{E}[M_n] \nrightarrow \mathbb{E}[M_\infty] = 0,$$

even though $\{M_n\}$ is bounded in L^1 and $M_n \to M_\infty$ a.s. In particular, M_n does not converge to M_∞ in L^1.

The main issue in the above example is that $\{M_n\}$ fails to be *uniformly integrable*. As we have seen in Theorem 4.2, uniform integrability is the bridge connecting convergence in probability and L^1-convergence. Uniformly integrable martingales thus possess richer convergence properties and are better behaved objects to work with.

8.6.1 Uniformly integrable martingales and L^1-convergence

The following result provides the L^1-convergence counterpart of the martingale convergence theorem.

Theorem 8.6. *Let $\{X_n, \mathcal{F}_n : n \geqslant 0\}$ be a supermartingale which is bounded in L^1, so that X_n converges a.s. to some integrable random variable X_∞. Then the following two statements are equivalent:*

(i) The family $\{X_n\}$ is uniformly integrable.
(ii) X_n converges to X_∞ in L^1.
In this case, one also has

$$\mathbb{E}[X_\infty|\mathcal{F}_n] \leqslant X_n \quad a.s. \tag{8.17}$$

If $\{X_n, \mathcal{F}_n\}$ is a martingale, then (i) / (ii) is also equivalent to (8.17) with "$\leqslant$" replaced by "$=$".

Proof. Since a.s. convergence implies convergence in probability, the equivalence between (i) and (ii) is a direct consequence of Theorem 4.2. To obtain (8.17), it suffices to show that

$$\int_A X_\infty d\mathbb{P} \leqslant \int_A X_n d\mathbb{P}, \quad \forall A \in \mathcal{F}_n. \tag{8.18}$$

To this end, by the supermartingale property one has

$$\int_A X_m d\mathbb{P} \leqslant \int_A X_n d\mathbb{P} \quad \forall m \geqslant n, \ A \in \mathcal{F}_n.$$

The relation (8.18) follows by letting $m \to \infty$, which is legal due to L^1-convergence. The last part of the theorem in the martingale case is a direct consequence of Proposition 4.4. □

The following convergence result due to P. Lévy is a particularly useful situation.

Corollary 8.2 (Lévy's Forward Theorem)**.** *Let Y be an integrable random variable and let $\{\mathcal{F}_n : n \geqslant 0\}$ be a filtration. Then $X_n = \mathbb{E}[Y|\mathcal{F}_n]$ is a uniformly integrable $\{\mathcal{F}_n\}$-martingale and*

$$X_n \to \mathbb{E}[Y|\mathcal{F}_\infty] \text{ both a.s. and in } L^1,$$

where $\mathcal{F}_\infty \triangleq \sigma(\cup_n \mathcal{F}_n)$.

Proof. The martingale property follows from

$$\mathbb{E}[X_m|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[Y|\mathcal{F}_m]|\mathcal{F}_n] = \mathbb{E}[Y|\mathcal{F}_n] = X_n \quad \forall m > n.$$

Uniform integrability follows from Proposition 4.4. In particular, one knows from Theorem 4.1 that $\{X_n\}$ is bounded in L^1. According to Theorem 8.6, X_n converges to some X_∞ a.s. and in L^1.

It remains to show that $X_\infty = \mathbb{E}[Y|\mathcal{F}_\infty]$. Since $\cup_n \mathcal{F}_n$ is a π-system, one only needs to verify

$$\int_A X_\infty d\mathbb{P} = \int_A Y d\mathbb{P}$$

for all $A \in \mathcal{F}_n$ and $n \geqslant 0$. But this follows from taking $m \to \infty$ in the relation

$$\int_A X_m d\mathbb{P} = \int_A Y d\mathbb{P} \quad \forall m \geqslant n,\ A \in \mathcal{F}_n.$$

□

Example 8.6 (Pólya's urn)**.** Initially, an urn contains b black balls and w white balls. At each time, a ball is drawn uniformly at random and it is returned to the urn along with c additional balls of the same colour. Let X_n denote the fraction of black balls after the n-th draw. We claim that X_n is a martingale with respect to its natural filtration $\{\mathcal{F}_n\}$. Indeed, let B_n (respectively, W_n) denote the total number of black (respectively, white) balls in the urn after the n-th draw. Then given $\mathcal{F}_n$, one has

$$X_{n+1} = \begin{cases} \dfrac{B_n + c}{B_n + W_n + c}, & \text{with prob. } \dfrac{B_n}{B_n + W_n}; \\ \dfrac{B_n}{B_n + W_n + c}, & \text{with prob. } \dfrac{W_n}{B_n + W_n}. \end{cases}$$

As a result,

$$\begin{aligned} \mathbb{E}[X_{n+1}|\mathcal{F}_n] &= \frac{B_n + c}{B_n + W_n + c} \times \frac{B_n}{B_n + W_n} + \frac{B_n}{B_n + W_n + c} \times \frac{W_n}{B_n + W_n} \\ &= \frac{B_n}{B_n + W_n} = X_n. \end{aligned}$$

Note that X_n is also uniformly bounded (since $X_n \in [0, 1]$) and is thus uniformly integrable by Proposition 4.3. According to Theorem 8.6, X_n converges to some integrable random variable X_∞ a.s. and in L^1.

There is an interesting observation of exchangeability in this example. For each $n \geqslant 1$, let us define

$$Z_n \triangleq \begin{cases} 1, & \text{if the } n\text{-th draw is black;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{z} = (z_1, \cdots, z_n)$ be a given word where $z_i \in \{0, 1\}$ and there are precisely k of 1's in $\mathbf{z}$. We claim that

$$\begin{aligned} &\mathbb{P}\big((Z_1, \cdots, Z_n) = (z_1, \cdots, z_n)\big) \\ &= \frac{b}{b+w} \cdot \frac{b+c}{b+w+c} \cdots \frac{b+(k-1)c}{b+w+(k-1)c} \\ &\quad \times \frac{w}{b+w+kc} \cdot \frac{w+c}{b+w+(k+1)c} \cdots \frac{w+(n-k-1)c}{b+w+(n-1)c}. \end{aligned} \tag{8.19}$$

This formula shows that the outcome distribution of the first n draws depends only on k (the number of black draws) but not on the specific times that the black balls are drawn. To prove (8.19), one observes that the formula is obvious when

$$(z_1, \cdots, z_n) = \big(\underbrace{1, \cdots, 1}_{k}, \underbrace{0, \cdots, 0}_{n-k}\big).$$

For a different word $\mathbf{z}$ (with k of 1's), the denominator in (8.19) remains unchanged while the numerator is suitably permuated from (8.19). The resulting product is thus the same.

To get some feeling about the distribution of X_∞, let us first consider the special situation when $b = w = c = 1$. According to the formula (8.19), for any $k = 0, 1, \cdots, n$ one has

$$\begin{aligned} \mathbb{P}(B_n = k+1) &= \mathbb{P}(\text{precisely } k \text{ black among first } n \text{ draws}) \\ &= \binom{n}{k} \times \frac{(1 \cdot 2 \cdots k)(1 \cdot 2 \cdots (n-k))}{2 \cdot 3 \cdots (n+1)} = \frac{1}{n+1}. \end{aligned}$$

In other words, B_n is a discrete uniform random variable over $\{1, 2, \cdots, n+1\}$. As a consequence, $X_\infty \sim U[0, 1]$. Next, let us assume that $b = c = 1$ but $w = 2$. Similar calculation shows that

$$\mathbb{P}(B_n = k+1) = \frac{2(n-k+1)}{n+2} \times \frac{1}{n+1}.$$

If $n \to \infty$ and $k/n \to x$, the fraction $\frac{2(n-k+1)}{n+2}$ converges to $2(1-x)$. This suggest that the probability density of the random variable X_∞ should be given by

$$f(x) = 2(1-x), \ x \in (0,1).$$

Question. Can you derive the distribution of X_∞ for general b, w, c?

8.6.2 Lévy's backward theorem and a martingale proof of strong LLN

There is a "backward" counterpart of Corollary 8.2 (running time backward towards $-\infty$). As we will see, a benefit of running backward in time is that one has a.s. convergence for free! We first introduce the following definition.

Definition 8.9. A *backward (sub/super)martingale* is a martingale indexed by negative integers. More precisely, it is a sequence $\{X_n, \mathcal{G}_n : n \leqslant -1\}$ satisfying the following properties:

(i) $\mathcal{G}_n$ is a sub-σ-algebra of $\mathcal{F}$ such that

$$\cdots \subseteq \mathcal{G}_n \subseteq \cdots \subseteq \mathcal{G}_{-2} \subseteq \mathcal{G}_{-1};$$

(ii) X_n is integrable and $\mathcal{G}_n$-measurable;
(iii) for any $n < m \leqslant -1$, one has

$$\mathbb{E}[X_m | \mathcal{G}_n](\geqslant / \leqslant) = X_n \quad \text{a.s.}$$

Below is the backward martingale convergence theorem. As usual, we only state the version for supermartingales while the (sub)martinagle counterpart of the theorem should be transparent.

Theorem 8.7. *Let $\{X_n, \mathcal{G}_n : n \leqslant -1\}$ be a backward supermartingale. Then the limit*

$$X_{-\infty} \triangleq \lim_{n \to -\infty} X_n \in [-\infty, \infty] \tag{8.20}$$

exists a.s. Suppose further that

$$\sup_{n \leqslant -1} \mathbb{E}[X_n] < \infty. \tag{8.21}$$

Then $\{X_n\}$ is uniformly integrable, $X_{-\infty}$ is integrable and the limit (8.20) holds in L^1 as well. Moreover, in this case one has

$$\mathbb{E}[X_n | \mathcal{G}_{-\infty}] \leqslant X_{-\infty} \quad \text{a.s.}$$

where $\mathcal{G}_{-\infty} \triangleq \cap_{m \leqslant -1} \mathcal{G}_m$.

Remark 8.10. If $\{X_n, \mathcal{G}_n : n \leqslant -1\}$ is a martingale, the condition (8.21) is automatically satisfied. As a result, *every backward martingale converges a.s. and in* L^1.

Proof. To prove the a.s. convergence of X_n, we use the same technique as in the proof of Theorem 8.5. Recall that the essential point is to estimate the expected upcrossing number. The notable difference here is that the upcrossing inequality (8.15), when applied to the martingale $\{X_n, \mathcal{G}_n : n = -N, \cdots, -1\}$, becomes

$$\mathbb{E}[U_{-N}(X; [a,b])] \leqslant \frac{\mathbb{E}[(X_{-1} - a)^-]}{b-a}.$$

By taking $N \uparrow \infty$, it follows that

$$\mathbb{E}[U_{-\infty}(X; [a,b])] \leqslant \frac{\mathbb{E}[(X_{-1} - a)^-]}{b-a}$$

Here $U_{-\infty}(X; [a,b])$ denotes the upcrossing number with respect to the interval $[a,b]$ by the entire sequence $\{X_n\}$. Since X_{-1} is integrable, one obtains the a.s. finiteness of $U_{-\infty}(X; [a,b])$ without the L^1-boundedness assumption! From this point on, the argument leading to the a.s. convergence of X_n is identical to the forward case. Note that the limit $X_{-\infty}$ is not necessarily finite.

Now suppose further that (8.21) holds. We first observe that $\{X_n\}$ is bounded in L^1. Indeed, since $\{X_n^-\}$ is a submartingale (it is the composition of the convex function $\max\{x, 0\}$ and the submartingale $-X_n$), one has

$$\mathbb{E}[|X_n|] = \mathbb{E}[X_n] + 2\mathbb{E}[X_n^-] \leqslant \mathbb{E}[X_n] + 2\mathbb{E}[X_{-1}^-] \quad \forall n \leqslant -1.$$

As a result,

$$M \triangleq \sup_{n \leqslant -1} \mathbb{E}[|X_n|] \leqslant 2\mathbb{E}[X_{-1}^-] + \sup_{n \leqslant -1} \mathbb{E}[X_n] < \infty. \tag{8.22}$$

This already implies by Fatou's lemma that $X_{-\infty}$ is finite a.s.

To prove the uniform integrability of $\{X_n\}$, let $\lambda > 0$ and $n \leqslant k \leqslant -1$. According to the supermartingale property, one has

$$\begin{aligned}
\mathbb{E}\big[|X_n| \mathbf{1}_{\{|X_n| > \lambda\}}\big] &= \mathbb{E}\big[X_n \mathbf{1}_{\{X_n > \lambda\}}\big] - \mathbb{E}\big[X_n \mathbf{1}_{\{X_n < -\lambda\}}\big] \\
&= \mathbb{E}[X_n] - \mathbb{E}\big[X_n \mathbf{1}_{\{X_n \leqslant \lambda\}}\big] - \mathbb{E}\big[X_n \mathbf{1}_{\{X_n < -\lambda\}}\big] \\
&\leqslant \mathbb{E}[X_n] - \mathbb{E}\big[X_k \mathbf{1}_{\{X_n \leqslant \lambda\}}\big] - \mathbb{E}\big[X_k \mathbf{1}_{\{X_n < -\lambda\}}\big] \\
&= \mathbb{E}[X_n] - \mathbb{E}[X_k] + \mathbb{E}\big[X_k \mathbf{1}_{\{X_n > \lambda\}}\big] - \mathbb{E}\big[X_k \mathbf{1}_{\{X_n < -\lambda\}}\big] \\
&\leqslant \mathbb{E}[X_n] - \mathbb{E}[X_k] + \mathbb{E}\big[|X_k| \mathbf{1}_{\{|X_n| > \lambda\}}\big].
\end{aligned} \tag{8.23}$$

Given $\varepsilon > 0$, by the assumption (8.21) there exists $k \leqslant -1$, such that

$$0 \leqslant \mathbb{E}[X_n] - \mathbb{E}[X_k] \leqslant \frac{\varepsilon}{2} \quad \forall n \leqslant k.$$

We fix such a k. Since X_k is integrable, there exists $\delta > 0$ such that

$$A \in \mathcal{F}, \mathbb{P}(A) < \delta \implies \mathbb{E}[|X_k|\mathbf{1}_A] < \frac{\varepsilon}{2}.$$

In view of the L^1-boundedness property (8.22), one can find $\Lambda > 0$ such that for all $\lambda > \Lambda$,

$$\mathbb{P}(|X_n| > \lambda) \leqslant \frac{\mathbb{E}[|X_n|]}{\lambda} \leqslant \frac{M}{\lambda} < \delta \quad \forall n \leqslant -1 \text{ and } \lambda > 0.$$

It follows from (8.23) that

$$\mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>\lambda\}}\big] < \varepsilon$$

for all $n \leqslant k$ and $\lambda > \Lambda$. This yields the uniform integrability (why?). It is now a consequence of Theorem 4.2 that $X_n \to X_{-\infty}$ in L^1 as well as $n \to -\infty$.

The last part of the theorem follows by taking $m \to -\infty$ in the following relation:

$$\int_A X_n d\mathbb{P} \leqslant \int_A X_m d\mathbb{P} \quad \forall A \in \mathcal{G}_{-\infty} \text{ and } m \leqslant n \leqslant -1.$$

□

The following result, which is the backward version of Corollary 8.2, is an immediate consequence of Theorem 8.7.

Corollary 8.3 (Lévy's Backward Theorem)**.** *Let $\{\mathcal{G}_n : n \leqslant -1\}$ be a given filtration over $(\Omega, \mathcal{F}, \mathbb{P})$ as before. Let Y be an integrable random variable and define*

$$X_n \triangleq \mathbb{E}[Y|\mathcal{G}_n], \quad n \leqslant -1.$$

Then

$$X_{-\infty} \triangleq \lim_{n\to-\infty} X_n \text{ exists a.s. and in } L^1,$$

and one has

$$X_{-\infty} = \mathbb{E}[Y|\mathcal{G}_{-\infty}] \quad a.s. \tag{8.24}$$

where $\mathcal{G}_{-\infty} \triangleq \cap_{n\leqslant -1}\mathcal{G}_n$.

As an application of Lévy's backward theorem, we give an alternative (and much shorter) proof of the strong LLN (Theorem 5.5 (i)) with an extra gain of L^1-convergence.

Theorem 8.8. *Let $\{X_n\}$ be an i.i.d. sequence with mean μ. Define $S_n \triangleq X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \quad \textit{a.s. and in } L^1$$

as $n \to \infty$.

Proof. For each $n \geqslant 1$, define

$$\mathcal{G}_{-n} \triangleq \sigma(S_n, S_{n+1}, S_{n+2}, \cdots) = \sigma(S_n, X_{n+1}, X_{n+2}, \cdots).$$

Since the X_n's are i.i.d., one has □

$$\mathbb{E}[X_1|\mathcal{G}_{-n}] = \mathbb{E}[X_1|S_n] = \frac{S_n}{n} \quad \text{a.s. (why?)}$$

According to Corollary 8.3, S_n/n converges to some integrable random variable $X_{-\infty}$ a.s. and in L^1 as $n \to \infty$. Since the random variable $\lim\limits_{n\to\infty} S_n/n$ is measurable with respect to the tail σ-algebra of the sequence $\{X_n\}$, by Corollary 5.1 it is constant a.s. In particular,

$$X_{-\infty} = \mathbb{E}[X_{-\infty}] = \lim_{n\to\infty} \mathbb{E}\big[\frac{S_n}{n}\big] = \mu \quad \text{a.s.}$$

The result thus follows.

8.6.3 An application: Wald's identity

Consider a random walk $S_n \triangleq X_1 + \cdots + X_n$ ($S_0 \triangleq 0$) where $X = \{X_n\}$ is an i.i.d. sequence with finite mean $\mu \triangleq \mathbb{E}[X_1]$. Let $\mathcal{F}_n \triangleq \sigma(X_1, \cdots, X_n)$. The following result, which was due to A. Wald, partially generalises the optional sampling theorem to the case of integrable stopping times in the current context.

Proposition 8.6. *Let τ be an integrable, $\{\mathcal{F}_n\}$-stopping time. Then $\mathbb{E}[S_\tau] = \mu\mathbb{E}[\tau]$.*

Proof. It is easily checked that $S_n - \mu n$ is an $\{\mathcal{F}_n\}$-martingale. According to the optional sampling theorem, one has

$$\mathbb{E}[S_{\tau\wedge n}] = \mu\mathbb{E}[\tau \wedge n] \quad \forall n \geqslant 1.$$

Letting $n \to \infty$, the right hand side converges to $\mu\mathbb{E}[\tau]$ by the monotone convergence theorem. For the left hand side, one has $S_{\tau\wedge n} \to S_\tau$ a.s. (note that $\tau < \infty$ a.s.) and additionally

$$|S_{\tau\wedge n}| \leqslant |X_1| + \cdots + |X_{\tau\wedge n}| \leqslant |X_1| + \cdots + |X_\tau| = \sum_{k=1}^{\infty} |X_k| \mathbf{1}_{\{\tau \geqslant k\}}$$

for all n. We claim that the last random variable is integrable. Indeed, since

$$\{\tau \geqslant k\} = \{\tau \leqslant k-1\}^c \in \mathcal{F}_{k-1},$$

one knows that X_k and $\{\tau \geqslant k\}$ are independent. As a result,

$$\begin{aligned}\mathbb{E}\big[\sum_{k=1}^{\infty} |X_k| \mathbf{1}_{\{\tau \geqslant k\}}\big] &= \sum_{k=1}^{\infty} \mathbb{E}\big[|X_k| \mathbf{1}_{\{\tau \geqslant k\}}\big] = \sum_{k=1}^{\infty} \mathbb{E}[|X_k|] \cdot \mathbb{P}(\tau \geqslant k) \\ &= \mathbb{E}[|X_1|] \cdot \sum_{k=1}^{\infty} \mathbb{P}(\tau \geqslant k) < \infty,\end{aligned}$$

where the last property follows from Lemma 5.1. It follows from the dominated convergence theorem that $S_\tau \in L^1$ and $\mathbb{E}[S_{\tau\wedge n}] \to \mathbb{E}[S_\tau]$. The result thus follows. □

Proposition 8.6 is useful in determining the exact value of $\mathbb{E}[\tau]$ when $\mu \neq 0$. If $\mu = 0$, one needs to move to the quadratic level.

Proposition 8.7. *Suppose that $\mu = 0$ and $\sigma^2 \triangleq \mathbb{E}[X_1^2] < \infty$. Let τ be an integrable, $\{\mathcal{F}_n\}$-stopping time. Then $\mathbb{E}[S_\tau^2] = \sigma^2\mathbb{E}[\tau]$.*

Proof. First of all, one observes that $S_n^2 - \sigma^2 n$ is a martinagle. By optional sampling, one has

$$\mathbb{E}[S_{\tau\wedge n}^2] = \sigma^2\mathbb{E}[\tau \wedge n] \leqslant \sigma^2\mathbb{E}[\tau] \quad \forall n.$$

According to Proposition 4.3 (i), the family $\{S_{\tau\wedge n}\}$ is uniformly integrable. Since $S_{\tau\wedge n} \to S_\tau$ a.s. trivially, it follows that the same convergence also holds in L^1. Next, we claim that

$$S_{\tau\wedge n} = \mathbb{E}[S_\tau | \mathcal{F}_{\tau\wedge n}]. \tag{8.25}$$

Indeed, given $A \in \mathcal{F}_{\tau\wedge n}$, by optional sampling one has

$$\int_A S_{\tau\wedge n} d\mathbb{P} = \int_A S_{\tau\wedge m} d\mathbb{P} \quad \forall m > n.$$

Taking $m \to \infty$, one obtains from L^1-convergence that the right hand side goes to $\int_A S_\tau d\mathbb{P}$. This justifies the claim (8.25).

Now one can use the conditional Jensen's inequality to see that

$$S_{\tau\wedge n}^2 = \big(\mathbb{E}[S_\tau|\mathcal{F}_{\tau\wedge n}]\big)^2 \leqslant \mathbb{E}[S_\tau^2|\mathcal{F}_{\tau\wedge n}].$$

As a result,

$$\sigma^2\mathbb{E}[\tau] = \lim_{n\to\infty} \sigma^2\mathbb{E}[\tau\wedge n] = \lim_{n\to\infty} \mathbb{E}[S_{\tau\wedge n}^2] \leqslant \mathbb{E}[S_\tau^2].$$

On the other hand, by Fatou's lemma one also has

$$\mathbb{E}[S_\tau^2] \leqslant \varliminf_{n\to\infty} \mathbb{E}[S_{\tau\wedge n}^2] = \varliminf_{n\to\infty} \sigma^2\mathbb{E}[\tau\wedge n] = \sigma^2\mathbb{E}[\tau].$$

Therefore, one concludes that $\mathbb{E}[S_\tau^2] = \sigma^2\mathbb{E}[\tau]$. □

To conclude, we look at the example of Bernoulli random walks. Suppose that X_1 follows the distribution

$$\mathbb{P}(X_1 = 1) = p,\ \mathbb{P}(X_1 = -1) = q \triangleq 1 - p$$

where $p \in (0, 1)$. Note that $\mu = p - q$. Let x be a fixed positive integer and define

$$\tau \triangleq \inf\{n \geqslant 1 : S_n = x\}.$$

The case when $p > q$. Since

$$\mu\mathbb{E}[\tau\wedge n] = \mathbb{E}[S_{\tau\wedge n}] \leqslant x,$$

by monotone convergence one sees that

$$\mu\mathbb{E}[\tau] \leqslant x < \infty \implies \mathbb{E}[\tau] < \infty.$$

According to Proposition 8.6,

$$\mathbb{E}[\tau] = \frac{\mathbb{E}[S_\tau]}{\mu} = \frac{x}{\mu}.$$

The case when $p \leqslant q$. In this case, one must have $\mathbb{E}[\tau] = \infty$; for otherwise, one knows from Proposition 8.6 that

$$0 < x = \mathbb{E}[S_\tau] = \mu\mathbb{E}[\tau] \leqslant 0,$$

which is absurd. We now compute $\mathbb{P}(\tau < \infty)$. Let

$$\varphi(t) \triangleq \mathbb{E}[e^{tX_1}] = pe^t + qe^{-t}$$

denote the moment generating function of X_1. The key observation is that for each fixed t, the sequence $e^{tS_n}/\varphi(t)^n$ is a martingale. Let us first assume that $p < q$. In this case, since $\mu = \varphi'(0) < 0$ there exists an $r > 0$ such that $\varphi(r) = \varphi(0) = 1$ (why?). In particular, e^{rS_n} is a martingale. By the optional sampling theorem, one has

$$\mathbb{E}[e^{rS_{\tau\wedge n}}] = 1 \quad \forall n \geqslant 1. \tag{8.26}$$

We now write

$$\mathbb{E}[e^{rS_{\tau\wedge n}}] = \mathbb{E}\big[e^{rS_\tau}; \tau \leqslant n\big] + \mathbb{E}\big[e^{rS_n}; \tau > n\big] = e^{rx}\mathbb{P}(\tau \leqslant n) + \mathbb{E}\big[e^{rS_n}; \tau > n\big].$$

The first term converges to $e^{rx}\mathbb{P}(\tau < \infty)$ as $n \to \infty$. For the second term, due to the strong LLN one has

$$S_n/n \to \mu < 0 \text{ a.s.} \implies e^{rS_n} \to 0 \text{ a.s.}$$

In addition, on the event $\{\tau > n\}$ it is obvious that $e^{rS_n} \leqslant e^{rx}$. By dominated convergence, one finds that

$$\lim_{n\to\infty} \mathbb{E}\big[e^{rS_n}; \tau > n\big] = 0.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{E}[e^{rS_{\tau\wedge n}}] = e^{rx}\mathbb{P}(\tau < \infty).$$

It follows from (8.26) that $\mathbb{P}(\tau < \infty) = e^{-rx}$.

Next, we look at the symmetric case ($p = q = 1/2$). In this case, $\varphi(t) = \cosh t$. It is clear from (5.2) that $\mathbb{P}(\tau < \infty) = 1$. But this fact can be easily reproduced by the current martingale method. Recall that for each fixed $t > 0$, one has

$$\mathbb{E}\big[(\mathrm{sech}t)^{\tau\wedge n} e^{tS_{\tau\wedge n}}\big] = 1 \quad \forall n. \tag{8.27}$$

In addition, note that

$$\lim_{n\to\infty} (\mathrm{sech}t)^{\tau\wedge n} e^{tS_{\tau\wedge n}} = \begin{cases} (\mathrm{sech}t)^\tau e^{tx}, & \text{on } \{\tau < \infty\}; \\ 0, & \text{on } \{\tau = \infty\}. \end{cases}$$

The second scenario follows from the fact that $\mathrm{sech}t < 1$. By dominated convergence, one can take $n \to \infty$ in (8.27) to obtain that

$$\mathbb{E}[(\mathrm{sech}t)^\tau e^{tx}] = 1.$$

Here $(\mathrm{sech}t)^\tau \triangleq 0$ if $\tau = \infty$. In other words,

$$\mathbb{E}[(\mathrm{sech}t)^\tau \mathbf{1}_{\{\tau<\infty\}}] = e^{-tx} \quad \forall t > 0. \tag{8.28}$$

Since$(\mathrm{sech}t)^\tau \mathbf{1}_{\{\tau<\infty\}} \leqslant 1$ for all $t > 0$, by taking $t \searrow 0$ and using dominated convergence again, one concludes that

$$\mathbb{P}(\tau < \infty) = \lim_{t\downarrow 0} \mathbb{E}[(\mathrm{sech}t)^\tau \mathbf{1}_{\{\tau<\infty\}}] = \lim_{t\downarrow 0} e^{-tx} = 1.$$

It is actually possible to compute the distribution of τ explicitly. We only consider the symmetric case and $x = 1$ for simplicity. By letting $\alpha \triangleq \mathrm{sech}t$ in (8.28), one finds that

$$\mathbb{E}[\alpha^\tau] = e^{-t} = \frac{1 - \sqrt{1-\alpha^2}}{\alpha}.$$

We now expand the right hand side into a power series of α. First recall that

$$\sqrt{1-\alpha^2} = \sum_{n=0}^{\infty} \binom{1/2}{n} (-1)^n \alpha^{2n}.$$

As a result,

$$\mathbb{E}[\alpha^\tau] = \frac{1}{\alpha}\big(1 - \sum_{n=0}^{\infty} \binom{1/2}{n} (-1)^n \alpha^{2n}\big) = \sum_{n=1}^{\infty} \binom{1/2}{n} (-1)^{n+1} \alpha^{2n-1}.$$

On the other hand, one also has

$$\mathbb{E}[\alpha^\tau] = \sum_{n=1}^{\infty} \mathbb{P}(\tau = 2n-1) \cdot \alpha^{2n-1}$$

since τ cannot achieve even values. By comparing the two expansions, it is easily seen that

$$\mathbb{P}(\tau = 2n-1) = (-1)^{n+1} \binom{1/2}{n}.$$

8.7 Some applications of martingale methods

We conclude by discussing three enlightening applications of martingale methods.

8.7.1 Monkey typing Shakespeare

The first application is an artificial one (and just for fun!). Nonetheless, it contains a very insightful method and a non-trivial application of optional sampling.

Imagine that a monkey is typing letters on the keyboard. At each time, it selects one of the 26 letters uniformly at random and different selections are assumed to be independent from each other. We have seen in Example 5.4 that (by using the second Borel-Cantelli lemma) with probability one, the monkey will eventually produce an exact copy (in fact, infinitely many!) of Shakespeares' "Hamlet". Our next question is: *how long on average does it take to produce one for the first time?*

Let us formulate the mathematics precisely. Our random experiment is tossing a die with 26 faces (each letter per face) repeatedly and independently in a sequence. We consider the string ALPHABETALPHA as a toy model of the "Hamlet". Let τ denote the first time this string appears. We wish to compute $\mathbb{E}[\tau]$.

To solve this problem, we introduce the following imaginary model. Suppose that a casino is proposing a new game called ALPHABETALPHA. The dealer rolls a die with 26 faces repeatedly. At every round, precisely one gambler enters the game and she bets in the following way. She bets \$1 on the first letter A in the string ALPHABETALPHA. She quits if she loses, while if she wins the dealer pays her \$26 dollars and she bets all this amount on the second letter L at the next round. If she loses she quits, while if she wins again the dealer pays her $\$26^2$ and she further bets all the money on the third letter P at the next round. The strategy continues and she quits the game either when she loses at some point or when she wins the entire string.

We now introduce several basic mathematical objects associated with this model. The underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as follows. The sample space is given by

$$\Omega = \{\omega = (\omega_1, \omega_2, \cdots) : \omega_n = \text{A,B,C}, \cdots\}.$$

Let $\xi_n(\omega) \triangleq \omega_n$ be the random variable giving the n-th outcome of the game. $\mathcal{F}$ is the σ-algebra generated by all the ξ_n's. $\mathbb{P}$ is the unique probability measure on $\mathcal{F}$ such that

$$\mathbb{P}\big(\xi_1 = a_1, \cdots, \xi_n = a_n\big) = \frac{1}{26^n}$$

for all n and $a_i \in \{A, \cdots, Z\}$. One can also construct $(\Omega, \mathcal{F}, \mathbb{P})$ as the countable product space of the probability model for each single toss. We also introduce

the filtration given by $\mathcal{F}_n \triangleq \sigma(\xi_1, \cdots, \xi_n)$ (the σ-algebra generated by the first n outcomes).

Let M_n ($n \geqslant 1$) denote the net gain of the casino up to the n-th round. The first key observation is the following lemma.

Lemma 8.3. *The sequence $\{M_n, \mathcal{F}_n\}$ is a martingale with uniformly bounded increments, i.e. there exists $C > 0$ such that $|M_n(\omega) - M_{n-1}(\omega)| \leqslant C$ for all ω, n.*

Proof. Let S denote the string ALPHABETALPHA and set $N \triangleq 13$ (length of S). The key point is to find a suitable way of representing M_n. Given $j, n \geqslant 1$, we introduce Y_n^j to be the net gain of Player j at the end of Game n. We first show that $Y^j = \{Y_n^j\}$ is a martingale by realising it as a martingale transform. According to the rules of the game, the strategy for this player at Game n is given by

$$C_n^j = 0 \quad \text{if } n < j \text{ or } n > j + N - 1$$

and

$$C_n^j = 26^{n-j} \mathbf{1}_{A_n^j} \quad \text{if } j \leqslant n \leqslant j + N - 1,$$

where A_n^j denotes the event that "Game j results in the first letter A in S, Game $j+1$ results in the second letter L in S, $\cdots$, Game $n-1$ results in the $(n-j)$-th letter in S". Clearly, $C_n^j \in \mathcal{F}_{n-1}$ so that the strategy sequence $C^j = \{C_n^j\}$ is $\{\mathcal{F}_n\}$-predictable. For $j \leqslant n \leqslant j + N - 1$, let ξ_n^j denote the net gain if one bets \$1 on seeing the $(n - j + 1)$-th letter in S in the outcome of Game n. It is not hard to see that

$$\xi_n^j = \begin{cases} 25, & \text{if Game } n \text{ produces the } (n-j+1)\text{-th letter in } S; \\ -1, & \text{otherwise.} \end{cases}$$

We also set $\xi_n^j \triangleq 0$ if $n < j$ or $n > j + N - 1$. One checks by definition that

$$G_n^j \triangleq \sum_{k=1}^{n} \xi_k^j, \quad n \geqslant 1$$

is an $\{\mathcal{F}_n\}$-martingale. Since

$$Y_n^j = \sum_{k=1}^{n} C_k^j \xi_k^j = \sum_{k=1}^{n} C_k^j (G_k^j - G_{k-1}^j),$$

by Theorem 8.1 one knows that $\{Y_n^j, \mathcal{F}_n\}$ is a martingale.

On the other hand, since the net gain of the casino is equal to the net loss of all players, one sees that

$$M_n = -\sum_{j=1}^{\infty} Y_n^j = -\sum_{j=1}^{n} Y_n^j.$$

It follows that

$$\begin{aligned}
\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= -\mathbb{E}\big[\sum_{j=1}^{n+1} Y_{n+1}^j|\mathcal{F}_n\big] = -\sum_{j=1}^{n+1}\mathbb{E}\big[Y_{n+1}^j|\mathcal{F}_n\big] \\
&= -\sum_{j=1}^{n+1} Y_n^j \quad (\text{since } Y^j \text{ is a martingale}) \\
&= -\sum_{j=1}^{n} Y_n^j = M_n. \quad (\text{since } Y_n^{n+1} = 0)
\end{aligned}$$

In other words, $\{M_n, \mathcal{F}_n\}$ is a martingale.

It remains to see that $\{M_n\}$ has uniformly bounded increments. To this end, one first observes that

$$\begin{aligned}
|M_{n+1} - M_n| &= \big|\sum_{j=1}^{n}\big(Y_{n+1}^j - Y_n^j\big) + Y_{n+1}^{n+1}\big| = \big|\sum_{j=1}^{n} C_{n+1}^j \xi_{n+1}^j + Y_{n+1}^{n+1}\big| \\
&\leqslant \sum_{j=1}^{n} |C_{n+1}^j| \cdot |\xi_{n+1}^j| + |Y_{n+1}^{n+1}|. \qquad (8.29)
\end{aligned}$$

In addition, one has $|Y_{n+1}^{n+1}| \leqslant 26$ and

$$|C_{n+1}^j| \cdot |\xi_{n+1}^j| \leqslant 26^{N-1} \cdot 26 = 26^N.$$

Note that the above term is non-zero only when

$$j \leqslant n+1 \leqslant j+N-1 (\implies j \geqslant n-N+2).$$

In particular, there are at most $N-1$ non-zero terms in the summation in (8.29). As a result, one finds that

$$|M_{n+1} - M_n| \leqslant (N-1)26^N + 26.$$

□

Recall that τ is the first time that the string S appears. The next key observation is the integrability of τ.

Lemma 8.4. *The stopping time τ is integrable.*

Proof. Recall that $N = 13$ is the length of the string S. Let B_n denote the event that the string S appears over the Games $n+1$, $n+2$, $\cdots$, $n+N$. Then $B_n \subseteq \{\tau \leqslant n+N\}$. As a result, for any $A \in \mathcal{F}_n$ one has

$$\begin{aligned}
&\mathbb{P}(\{\tau \leqslant n+N\} \cap A) \\
&\geqslant \mathbb{P}(B_n \cap A) = \mathbb{P}(B_n)\mathbb{P}(A) \quad (B_n \text{ and } A \text{ are independent}) \\
&= 26^{-N}\mathbb{P}(A).
\end{aligned}$$

By the definition of the conditional expectation, one obtains that

$$\mathbb{P}(\tau \leqslant n+N|\mathcal{F}_n) \geqslant 26^{-N} =: \varepsilon \quad \forall n. \tag{8.30}$$

It follows from (8.30) that (for all $k, r \geqslant 1$)

$$\begin{aligned}
\mathbb{P}(\tau > kN+r) &= \mathbb{P}(\tau > kN+r,\ \tau > (k-1)N+r) \\
&= \mathbb{E}\big[\mathbf{1}_{\{\tau>(k-1)N+r\}}\mathbb{P}(\tau > kN+r|\mathcal{F}_{(k-1)N+r})\big] \\
&\leqslant (1-\varepsilon)\cdot\mathbb{P}(\tau > (k-1)N+r).
\end{aligned}$$

Iterating the same inequality, one arrives at

$$\mathbb{P}(\tau > kN+r) \leqslant (1-\varepsilon)^{k-1}\mathbb{P}(\tau > N+r).$$

It follows that

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{j=1}^{\infty}\mathbb{P}(\tau \geqslant j) = \sum_{j=1}^{N+1}\mathbb{P}(\tau \geqslant j) + \sum_{n=1}^{\infty}\mathbb{P}(\tau > N+n) \\
&= \sum_{j=1}^{N+1}\mathbb{P}(\tau \geqslant j) + \sum_{r=1}^{N}\sum_{k=1}^{\infty}\mathbb{P}(\tau > kN+r) \\
&\leqslant (N+1) + \Big(\sum_{r=1}^{N}\mathbb{P}(\tau > N+r)\Big)\cdot\Big(\sum_{k=1}^{\infty}(1-\varepsilon)^{k-1}\Big) < \infty.
\end{aligned}$$

This gives the integrability of τ. □

The result below is another extension of the optional sampling theorem to the case of integrable stopping times. Note that Proposition 8.6 does not apply here since M_n is not a random walk.

Proposition 8.8. *Suppose that X is a martingale with uniformly bounded increments and σ is an integrable stopping time. Then $\mathbb{E}[X_\sigma] = \mathbb{E}[X_0]$.*

Proof. One begins by writing

$$\begin{aligned}
\mathbb{E}[X_\sigma] &= \sum_{n=0}^{\infty} \mathbb{E}[X_\sigma \mathbf{1}_{\{\sigma=n\}}] = \sum_{n=0}^{\infty} \mathbb{E}[X_n \mathbf{1}_{\{\sigma=n\}}] \\
&= \mathbb{E}[X_0 \mathbf{1}_{\{\sigma=0\}}] + \sum_{n=1}^{\infty} \mathbb{E}\big[\big(\sum_{k=1}^{n} (X_k - X_{k-1}) + X_0\big)\mathbf{1}_{\{\sigma=n\}}\big] \\
&= \mathbb{E}[X_0 \mathbf{1}_{\{\sigma=0\}}] + \sum_{n=1}^{\infty}\sum_{k=1}^{n} \mathbb{E}[(X_k - X_{k-1})\mathbf{1}_{\{\sigma=n\}}] + \sum_{n=1}^{\infty} \mathbb{E}[X_0 \mathbf{1}_{\{\sigma=n\}}] \\
&= \mathbb{E}[X_0] + \sum_{k=1}^{\infty} \mathbb{E}\big[(X_k - X_{k-1}) \sum_{n=k}^{\infty} \mathbf{1}_{\{\sigma=n\}}\big] \qquad (8.31) \\
&= \mathbb{E}[X_0] + \sum_{k=1}^{\infty} \mathbb{E}\big[(X_k - X_{k-1})\mathbf{1}_{\{\sigma\geqslant k\}}\big].
\end{aligned}$$

To reach the equality (8.31), we used Fubini's theorem to change the order of summation which is legal due to the boundedness of $X_k - X_{k-1}$ and the integrability of σ; indeed

$$\begin{aligned}
&\sum_{k=1}^{\infty}\sum_{n=k}^{\infty} \mathbb{E}\big[|X_k - X_{k-1}|\mathbf{1}_{\{\sigma=n\}}\big] \\
&\leqslant C \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} \mathbb{P}(\sigma = n) = C \sum_{k=1}^{\infty} \mathbb{P}(\sigma \geqslant k) = C\mathbb{E}[\sigma] < \infty.
\end{aligned}$$

Now the main observation is that

$$\{\sigma \geqslant k\} = \{\sigma \leqslant k-1\}^c \in \mathcal{F}_{k-1}.$$

By the martingale property, one has

$$\mathbb{E}\big[(X_k - X_{k-1})\mathbf{1}_{\{\sigma\geqslant k\}}\big] = 0,$$

which implies that $\mathbb{E}[X_\sigma] = \mathbb{E}[X_0]$. □

We now have all the required preliminaries to compute $\mathbb{E}[\tau]$ explicitly.

Let L_n denote the total loss of the casino up to the n-th round. Then $M_n = n - L_n$ and by Proposition 8.8 one has

$$0 = \mathbb{E}[M_0] = \mathbb{E}[M_\tau] = \mathbb{E}[\tau] - \mathbb{E}[L_\tau].$$

On the other hand, from the explicit shape of the string S it is easily found that

$$L_\tau = 26^{13} + 26^5 + 26.$$

Therefore, one obtains that

$$\mathbb{E}[\tau] = \mathbb{E}[L_\tau] = 26^{13} + 26^5 + 26.$$

Remark 8.11. It is not hard to obtain the correct value of $\mathbb{E}[\tau]$ by formally applying the optional sampling theorem to the martingale $\{M_n\}$ at the stopping time τ (as in the above calculation). The main delicate point here is to justify such a procedure mathematically, which is not a direct consequence of Theorem 8.3.

Remark 8.12. Although one knows for sure it will appear in finite time, one probably needs to wait until the end of the universe to see it (even for a string as short as S)!

8.7.2 Kolmogorov's law of the iterated logarithm

The second application is Kolmogorov's law of the iterated logarithm which provides fine information about the large time behaviour of random walks.

Let $\{X_n : n \geqslant 1\}$ be an i.i.d. sequence of random variables with mean zero and unit variance. Define $S_n \triangleq X_1 + \cdots + X_n$. The central limit theorem $S_n/\sqrt{n} \xrightarrow{d} N(0,1)$ suggests that the magnitude of S_n is roughly of order $\sqrt{n}$ when n is large. The *law of the iterated logarithm*, which was originally due to Kolmogorov in 1929, provides a precise (pathwise) growth rate of S_n. It asserts that along a subsequence of time S_n grows like $\sqrt{2n \log\log n}$ as $n \to \infty$.

Theorem 8.9. *With probability one,*

$$\overline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2n\log\log n}} = 1, \quad \underline{\lim_{n\to\infty}} \frac{S_n}{\sqrt{2n\log\log n}} = -1. \tag{8.32}$$

For simplicity, we only consider the special situation where $X_n \sim N(0,1)$. The general case requires more delicate analysis on the moment generating function,

but the essential idea is similar. The proof is based on Doob's maximal inequality and the Borel-Cantelli lemma. We also need the following lower estimate of the Gaussian tail, which complements the upper bound obtained in Example 2.3 before.

Lemma 8.5. *Let Z be a standard normal random variable. Then*

$$\mathbb{P}(X_1 > x) \geqslant \frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} e^{-x^2/2} \quad \forall x > 0. \tag{8.33}$$

Proof. Consider the function $f(x) \triangleq x^{-1} e^{-x^2/2}$ $(x > 0)$. One has

$$f'(x) = -(1 + x^{-2}) e^{-x^2/2}.$$

By integration from x to ∞, one obtains that

$$x^{-1} e^{-x^2/2} = \int_x^\infty (1 + y^{-2}) e^{-y^2/2} dy \leqslant (1 + x^{-2}) \int_x^\infty e^{-y^2/2} dy.$$

The lower estimate in (8.33) follows after rearrangement. □

We now proceed to prove Theorem 8.9 in the Gaussian context. Suppose that $\{X_n\}$ is an i.i.d. sequence of standard normal random variables.

Proof of Theorem 8.9. We only prove the first part of (8.32). The other part follows by considering $-S_n$. In what follows, we denote $h(m) \triangleq \sqrt{2m \log\log m}$. For each $\theta \in \mathbb{R}$, since the function $x \mapsto e^{\theta x}$ is convex, one knows from Proposition 8.1 that $e^{\theta S_m}$ is a submartingale (with respect to the natural filtration of $\{X_m\}$). According to Doob's maximal inequality (8.9), for all $c > 0$ one has

$$\mathbb{P}\big(\max_{k \leqslant m} S_k > c\big) = \mathbb{P}\big(\max_{k \leqslant m} e^{\theta S_k} > e^{\theta c}\big) \leqslant e^{-\theta c} \mathbb{E}[e^{\theta S_m}] = e^{-\theta c + \theta^2 m/2}, \tag{8.34}$$

where we used the fact that $S_m \sim N(0, m)$ to reach the last equality. Since (8.34) is true for all θ, by optimising the exponent $-\theta c + \theta^2 m/2$ (i.e. taking $\theta = c/m$), one obtains that

$$\mathbb{P}\big(\max_{k \leqslant m} S_k > c\big) \leqslant \exp\big(\inf_{\theta \in \mathbb{R}} \{-\theta c + \theta^2 m/2\}\big) = e^{-\frac{1}{2m} c^2} \quad \forall m \geqslant 1. \tag{8.35}$$

The proof of (8.32) consists of two steps: establishing an upper and a matching lower estimate. We first derive the upper bound. Let $K > 1$ be a fixed number. By applying (8.35) to $m = K^n$ and $c = K h(K^{n-1})$, one finds that

$$\mathbb{P}\big(\max_{k \leqslant K^n} S_k > c_n\big) \leqslant (n-1)^{-K} (\log K)^{-K}.$$

Since the right hand side yields a convergent series, by the first Borel-Cantelli lemma one has

$$\mathbb{P}\big(\max_{k\leqslant K^n} S_k > c_n \text{ i.o.}\big) = 1.$$

In other words, with probability one

$$\max_{k\leqslant K^n} S_k \leqslant c_n = Kh(K^{n-1}) \quad \forall n \text{ sufficiently large.}$$

Since $m \mapsto h(m)$ is increasing, it follows that with probability one

$$S_k \leqslant Kh(k) \quad \forall \text{ large } n \text{ and } K^{n-1} \leqslant k \leqslant K^n. \tag{8.36}$$

As a result,

$$\overline{\lim_{k\to\infty}} \frac{S_k}{h(k)} \leqslant K \quad \text{a.s.}$$

Since this is true for each $K > 1$, by taking $K \downarrow 1$ (along a countable sequence) one obtains that

$$\overline{\lim_{k\to\infty}} \frac{S_k}{h(k)} \leqslant 1 \quad \text{a.s.} \tag{8.37}$$

We now turn to establishing a matching lower bound. Let $N > 1$ and $\varepsilon \in (0,1)$ be given fixed. Consider the event

$$A_n \triangleq \big\{S_{N^{n+1}} - S_{N^n} > (1-\varepsilon)h(N^{n+1} - N^n)\big\}.$$

Note that $S_{N^{n+1}} - S_{N^n} \sim N(0, N^{n+1} - N^n)$. By using the lower bound in (8.33), one finds that

$$\mathbb{P}(A_n) = \mathbb{P}\big(Z > y\big) \geqslant \frac{1}{\sqrt{2\pi}} \frac{y}{y^2+1} e^{-y^2/2}, \tag{8.38}$$

where $Z \sim N(0,1)$ and $y \triangleq (1-\varepsilon)\sqrt{2\log\log(N^{n+1} - N^n)}$. Explicit calculation shows that

$$e^{-y^2/2} = \frac{1}{(n\log N + \log(N-1))^{(1-\varepsilon)^2}}.$$

In particular, the right hand side of (8.38) yields a divergent series. Since the A_n's are independent, by the second Borel-Cantelli lemma one has $\mathbb{P}(A_n \text{ i.o.}) = 1$. In other words, with probability one

$$S_{N^{n+1}} > (1-\varepsilon)h(N^{n+1} - N^n) + S_{N^n} \quad \text{for infinitely many } n.$$

On the other hand, recall from (8.36) that (with $K = 2$) $S_k \geqslant -2h(k)$ for all large k a.s. (why?) Therefore, one finds that

$$S_{N^{n+1}} > (1-\varepsilon)h(N^{n+1} - N^n) - 2h(N^n) \quad \text{for infinitely many } n \quad \text{a.s.} \tag{8.39}$$

By explicit calculation,

$$\lim_{n\to\infty} \frac{(1-\varepsilon)h(N^{n+1} - N^n) - 2h(N^n)}{h(N^{n+1})} = (1-\varepsilon)\sqrt{\frac{N-1}{N}} - \frac{2}{\sqrt{N}}.$$

It follows from (8.39) that

$$\varlimsup_{k\to\infty} \frac{S_k}{h(k)} \geqslant \varlimsup_{n\to\infty} \frac{S_{N^{n+1}}}{h(N^{n+1})} \geqslant (1-\varepsilon)\sqrt{\frac{N-1}{N}} - \frac{2}{\sqrt{N}} \quad \text{a.s.}$$

Taking $N \to \infty$ and $\varepsilon \downarrow 0$, one arrives at

$$\varlimsup_{k\to\infty} \frac{S_k}{h(k)} \geqslant 1 \quad \text{a.s.} \tag{8.40}$$

By putting (8.37) and (8.40) together, one obtains the first part of (8.32). The proof of Theorem 8.9 is thus complete.

□

8.7.3 Recurrence / Transience of Markov chains

The last application we shall discuss is the study of recurrence / transience properties of Markov chains. We assume that the reader is familiar with basic concepts and results from Markov chain theory (cf. [Str05] for an excellent introduction).

Let $X = \{X_n : n \geqslant 0\}$ be a Markov chain on a countable state space S with one-step transition probabilities $P = (P_{ij})_{i,j\in S}$. Given a function $f : S \to \mathbb{R}$, we define

$$(Pf)(i) \triangleq \sum_{j\in S} P_{ij} f(j), \quad i \in S$$

provided that the above summation is convergent. We begin with a particularly useful martingale property associated with X. Such a property plays a fundamental role in the study of Markov processes (in particular, in diffusion and SDE theory).

Proposition 8.9. *Let $f : S \to \mathbb{R}$ be a given function. Define*

$$Y_n^f \triangleq f(X_n) - f(X_0) - \sum_{k=0}^{n-1} (Pf - f)(X_k),\ Y_0^f \triangleq 0. \tag{8.41}$$

Suppose that Y_n^f is well-defined and integrable for each n. Then $\{Y_n^f\}$ is a martingale with respect to the natural filtration of X.

Proof. By the definition of Y_n^f, one has

$$\begin{aligned}
\mathbb{E}[Y_{n+1}^f|\mathcal{F}_n] &= \mathbb{E}\big[f(X_{n+1}) - f(X_0) - \sum_{k=1}^{n} (Pf - f)(X_k)|\mathcal{F}_n\big] \\
&= \mathbb{E}[f(X_{n+1})|\mathcal{F}_n] - f(X_0) - \sum_{k=1}^{n} (Pf - f)(X_k) \\
&= Pf(X_n) - f(X_0) - \sum_{k=1}^{n} (Pf - f)(X_k) \quad \text{(by Markov property)} \\
&= f(X_n) - f(X_0) - \sum_{k=1}^{n-1} (Pf - f)(X_k) \\
&= Y_n^f.
\end{aligned}$$

The martingale property thus follows. □

The essential idea here is that recurrence / transience properties of X is closely related to the existence of certain superharmonic functions on the state space S. We first introduce the following key definition.

Definition 8.10. A function $f : S \to \mathbb{R}$ is said to be *P-superharmonic* on $F \subseteq S$ if

$$(Pf)(i) \triangleq \sum_{j \in S} P_{ij} f(j) \leqslant f(i) \quad \forall i \in F,$$

provided that Pf is well-defined. We simply say that f is P-superharmonic if $F = S$.

The proposition below suggests that superharmonicity is naturally connected with a supermartingale property.

Proposition 8.10. *Let $f : S \to [0, \infty)$ be a given function such that $f(X_0)$ is integrable.*

(i) Suppose that f is P-superharmonic. Then $\{f(X_n)\}$ is a supermartingale.
(ii) Let $F \subseteq S$. Define $\tau \triangleq \inf\{n \geqslant 1 : X_n \in F\}$. Suppose that f is P-superharmonic outside F. Then one has

$$\mathbb{E}\big[f(X_{\tau\wedge n})|X_0 = i\big] \leqslant f(i) \quad \forall i \in F^c, n \geqslant 0. \tag{8.42}$$

Proof. (i) Since f is P-superharmonic, one has

$$\mathbb{E}[f(X_{n+1})|\mathcal{F}_n] = Pf(X_n) \leqslant f(X_n).$$

Therefore, $\{f(X_n)\}$ is a supermartingale.

(ii) Let Y_n be defined by (8.41) associated with f. Since $\{Y_n\}$ is a martingale by Proposition 8.9, so is $\{Y_{\tau\wedge n}\}$ by Theorem 8.2. As a result, one has

$$0 = \mathbb{E}[Y_0] = \mathbb{E}[Y_{\tau\wedge n}] = \mathbb{E}\big[f(X_{\tau\wedge n}) - f(X_0) - \sum_{k=0}^{\tau\wedge n-1} (Pf - f)(X_k)\big].$$

It follows that

$$\mathbb{E}[f(X_{\tau\wedge n})|X_0 = i] = f(i) + \mathbb{E}\big[\sum_{k=0}^{\tau\wedge n-1} (Pf - f)(X_k)|X_0 = i\big].$$

Since f is P-superharmonic outside F, one knows that

$$\sum_{k=0}^{\tau\wedge n-1} (Pf - f)(X_k) \leqslant 0.$$

The desired inequality (8.42) follows immediately. □

The following result provides a simple criterion for the recurrence of X.

Theorem 8.10. *Suppose that X is irreducible. Then X is recurrent if and only if all non-negative P-superharmonic functions are constant.*

Proof. Suppose that X is recurrent. Let $f \geqslant 0$ be a P-superharmonic function on S. Given $i \neq j \in S$, define

$$\rho_j \triangleq \inf\{n \geqslant 1 : X_n = j\}.$$

Since X is recurrent, one has

$$\mathbb{P}(\rho_j < \infty | X_0 = i) = 1.$$

On the other hand, Proposition 8.10 (ii) implies that (with $F = \{j\}$)

$$\mathbb{E}[f(X_{\rho_j \wedge n})|X_0 = i] \leqslant f(i) \quad \forall n.$$

By using Fatou's lemma, one obtains that

$$\begin{aligned} f(j) = \mathbb{E}[f(X_{\rho_j})|X_0 = i] &= \mathbb{E}\big[\varliminf_{n\to\infty} f(X_{\rho_j \wedge n})|X_0 = i\big] \\ &\leqslant \varliminf_{n\to\infty} \mathbb{E}[f(X_{\rho_j \wedge n})|X_0 = i] \leqslant f(i). \end{aligned}$$

Since i, j are arbitrary, one concludes that f is a constant function.

Conversely, suppose that X is transient. We define the function $G(i,j)$ on $S \times S$ by

$$G(i,j) \triangleq \sum_{n=0}^{\infty} P_{ij}^n,$$

where $P_{ij}^n \triangleq \mathbb{P}(X_n = j | X_0 = i)$ is the n-step transition probability from i to j. By definition, one has

$$\begin{aligned} (PG(\cdot, j))(i) &= \sum_{k\in S} P_{ik} G(k,j) = \sum_{k\in S} P_{ik} \sum_{n=0}^{\infty} P_{kj}^n = \sum_{n=0}^{\infty} \sum_{k\in S} P_{ik} P_{kj}^n \\ &= \sum_{n=0}^{\infty} P_{ij}^{n+1} = G(i,j) - \delta_{ij} \leqslant G(i,j). \end{aligned} \tag{8.43}$$

In particular, for each fixed $j \in S$ the function $G(\cdot, j)$ is a non-negative P-superharmonic function. We claim that $G(\cdot, j)$ is non-constant. Indeed, from (8.43) one has

$$(PG(\cdot, j))(j) = G(j,j) - 1 \neq G(j,j).$$

In particular, $G(\cdot, j)$ is not P-harmonic (a function f is *P-harmonic* if $Pf = f$). Since any constant function is obviously P-harmonic, one concludes that $G(\cdot, j)$ must be non-constant. □

The following result is a simple application of Theorem 8.10.

Proposition 8.11. *Let X be an irreducible Markov chain on a finite state space S. Then X is recurrent.*

Proof. Let f be a non-negative, P-superharmonic function on S. Then one has

$$\sum_{j\in S} P_{ij}^n f(j) \leqslant f(i) \quad \forall n \geqslant 1, j \in S.$$

We choose n such that $P_{ij}^n > 0$ for all $i, j \in S$ (this is possible due to irreducibility). Let $i_0 \in S$ be such that $f(i_0) = \min_S f$. Then

$$f(i_0) \geqslant \sum_{j\in S} P_{i_0 j}^n f(j) \geqslant \sum_{j\in S} P_{ij}^n f(i_0) = f(i_0).$$

As a result, one must have equality in the above estimate, which implies that $f(j) = f(i_0)$ for all j (since $P_{i_0 j}^n > 0$). In particular, f is a constant function. According to Theorem 8.10, one concludes that X is recurrent. □

Theorem 8.10 indicates that transience can be detected from the existence of a non-negative, non-constant P-superharmonic function on S. The following result suggests that recurrence can also be detected from the existence of certain unbounded, (partially) P-superharmonic functions.

Theorem 8.11. *Let $j \in S$ be a fixed state. Let $\{B_m : m \geqslant 1\}$ be an increasing family of non-empty subsets of S such that $j \in B_0$ and for each m, with probability one X (starting at j) exits B_m in finite time. Suppose that there exists $f : S \to [0, \infty)$ which is P-superharmonic on $\{j\}^c$ and*

$$a_m \triangleq \inf_{i\notin B_m} f(i) \to \infty \quad \text{as } m \to \infty.$$

Then the state j is recurrent.

Proof. Let us define

$$\tau_m \triangleq \inf\{n \geqslant 1 : X_n \notin B_m\}, \ \rho_j \triangleq \inf\{n \geqslant 1 : X_n = j\}$$

and set

$$\zeta_{j,m} \triangleq \rho_j \wedge \tau_m = \inf\{n \geqslant 1 : X_n \in \{j\} \cup B_m^c\}.$$

According to the Proposition 8.10 (ii), one has

$$f(j) \geqslant \mathbb{E}_j[f(X_{\zeta_{j,m}\wedge n})] \geqslant \mathbb{E}_j[f(X_{\zeta_{j,m}\wedge n})\mathbf{1}_{\{\tau_m \leqslant n\wedge\rho_j\}}] \geqslant a_m \mathbb{P}_j(\tau_m \leqslant n \wedge \rho_j), \quad (8.44)$$

where the subscript j in $\mathbb{E}, \mathbb{P}$ means that the initial position of X is j. Since $\mathbb{P}_j(\tau_m < \infty) = 1$ by assumption, one has

$$\{\tau_m \leqslant \rho_j\} = \bigcup_{n=1}^{\infty} \{\tau_m \leqslant n \wedge \rho_j\} \quad \mathbb{P}_j\text{-a.s.}$$

As a result, by taking $n \to \infty$ in (8.44) one finds that

$$f(j) \geqslant a_m \cdot \mathbb{P}_j(\tau_m \leqslant \rho_j).$$

Since $a_m \to \infty$, it must be the case that

$$\lim_{m \to \infty} \mathbb{P}_j(\tau_m \leqslant \rho_j) = 0.$$

As a consequence,

$$\mathbb{P}_j(\rho_j < \infty) \geqslant \mathbb{P}_j(\rho_j < \tau_m) = 1 - \mathbb{P}_j(\tau_m \leqslant \rho_j) \to 1$$

as $m \to \infty$. This gives the recurrence of the state j. □

Remark 8.13. One cannot expect that the function f in the theorem is superharmonic on the entire S, for otherwise it would also give transience (assuming X is irreducible) which is absurd.

Remark 8.14. It is possible to study positive / null recurrence by using superharmonic functions. But we will not discuss it here.

As an application, we discuss the recurrence / transience of simple random walks on $\mathbb{Z}^d$. Let $\{\mathbf{e}_1, \cdots, \mathbf{e}_d\}$ be the canonical basis of $\mathbb{Z}^d$.

Definition 8.11. The simple random walk on $\mathbb{Z}^d$ is the Markov chain on $\mathbb{Z}^d$ with one-step transition probabilities $(P_{\mathbf{xy}})_{\mathbf{x},\mathbf{y} \in \mathbb{Z}^d}$ given by

$$P_{\mathbf{xy}} \triangleq \begin{cases} \frac{1}{2d}, & \text{if } \mathbf{y} = \mathbf{x} \pm \mathbf{e}_i \text{ for some } i; \\ 0, & \text{otherwise.} \end{cases}$$

We first consider the one-dimensional case.

Proposition 8.12. *The simple random walk on $\mathbb{Z}$ is recurrent.*

Proof. Consider the function $f(k) \triangleq k$. By definition,

$$Pf(k) = \frac{1}{2}(|k+1| + |k-1|).$$

In particular, when $k \neq 0$ one has

$$Pf(k) = |k| = f(k).$$

As a result, f is superharmonic outside $\{0\}$. Since $f(k) \to \infty$ as $|k| \to \infty$, one concludes from Theorem 8.11 that the origin is recurrent. By irreducibility, the entire chain X is recurrent. □

Remark 8.15. The function f in the above proof is not superharmonic at the origin, as seen from

$$Pf(0) = 1 > 0 = f(0).$$

Next, we consider the two-dimensional case.

Proposition 8.13. *The simple random walk on $\mathbb{Z}^2$ is recurrent.*

Proof. We define

$$f(\mathbf{k}) \triangleq \begin{cases} \log(k_1^2 + k_2^2 - 1/2), & \mathbf{k} = (k_1, k_2) \neq (0,0); \\ \kappa, & \mathbf{k} = (0,0), \end{cases}$$

where κ is a suitable number to be chosen later on (the motivation of this construction comes from the continuous situation; cf. Remark 8.16 for a discussion). By explicit calculation, one finds that

$$\begin{aligned}(Pf - f)(\mathbf{k}) = \frac{1}{4}\log\Big[&\Big((k_1+1)^2 + k_2^2 - \frac{1}{2}\Big)\Big((k_1-1)^2 + k_2^2 - \frac{1}{2}\Big) \\ &\times \Big(k_1^2 + (k_2+1)^2 - \frac{1}{2}\Big)\Big(k_1^2 + (k_2-1)^2 - \frac{1}{2}\Big)/\Big(k_1^2 + k_2^2 - \frac{1}{2}\Big)^4\Big]\end{aligned}$$

for any $\mathbf{k} \notin \{\mathbf{0}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ are the four neighbouring points of $\mathbf{0}$. The reason for imposing this constraint on $\mathbf{k}$ is that if $\mathbf{k} = \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$, the term $f(\mathbf{0})$ will appear in the computation of Pf which is clearly not $\log(-1/2)$! To show that f is P-superharmonic on $\{\mathbf{0}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}^c$, one only needs to check that the expression inside the above logarithm is not greater than 1. But this follows from explicit calculation:

$$\frac{((k_1 \pm 1)^2 + k_2^2 - 1/2) \cdot (k_1^2 + (k_2 \pm 1)^2 - 1/2)}{(k_1^2 + k_2^2 - \frac{1}{2})^4} - 1 = -\frac{4(k_1^2 - k_2^2)^2}{(k_1^2 + k_2^2 - 1/2)^4} \leqslant 0.$$

Next, we choose $\kappa = f(\mathbf{0})$ properly so that f is also superharmonic at each of the four points $\{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$. By symmetry, one only needs to consider the value of f at $\mathbf{a}$. The requirement is that

$$\frac{1}{4}\big(\kappa + f(1,1) + f(2,0) + f(1,-1)\big) \leqslant f(1,0).$$

Clearly, a choice of κ satisfying the above inequality can be made. The resulting function f on $\mathbb{Z}^2$ is thus superharmonic on $\mathbb{Z}^2\backslash\{\mathbf{0}\}$. □

Remark 8.16. The above construction of f is motivated from its continuous counterpart. Indeed, for the function $F(x,y) \triangleq \log(x^2+y^2+a)$ $(a \in \mathbb{R})$ one finds that

$$\Delta F \triangleq \frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} = \frac{4a}{(x^2+y^2+a)^2},$$

which is negative when $a < 0$. The differential operator $\frac{1}{2}\Delta$ is the generator of the Brownian motion on $\mathbb{R}^2$ (the continuous analogue of the simple random walk). As an analogue in the discrete situation, one naturally considers functions of this type, e.g. $f(k_1,k_2) \triangleq \log(k_1^2 + k_2^2 - 1/2)$.

Finally, we consider the situation when dimension is three or higher.

Proposition 8.14. *The simple random walk on $\mathbb{Z}^d$ $(d \geqslant 3)$ is transient.*

Proof. The key step is the case of dimension three. The higher dimensional case follows easily from the result in dimension three. We define

$$f_\alpha(k) \triangleq (\alpha^2 + |\mathbf{k}|^2)^{-1/2}, \quad \mathbf{k} = (k_1,k_2,k_3) \in \mathbb{Z}^3$$

where $\alpha \geqslant 1$ is a parameter to be specified later on (in a way such that f is superharmonic). The construction of f_α is again motivated from the continuous situation (cf. Remark 8.17 for a discussion). By definition, f is superharmonic if and only if

$$\frac{1}{6}\sum_{i=1}^{3}\big((\alpha^2 + |\mathbf{k}+\mathbf{e}_i|^2)^{-1/2} + (\alpha^2 + |\mathbf{k}-\mathbf{e}_i|^2)^{-1/2}\big) \leqslant (\alpha^2 + |\mathbf{k}|^2)^{-1/2}. \tag{8.45}$$

Setting $M \triangleq 1 + \alpha^2 + |\mathbf{k}|^2$ and $x_i \triangleq k_i/M$ $(i = 1,2,3)$, it is a simple matter of

algebra that the inequality (8.45) is equivalent to

$$\begin{aligned}
&\frac{1}{6}\sum_{i=1}^{3}\big(\big(\frac{M+2k_i}{M}\big)^{-1/2}+\big(\frac{M-2k_i}{M}\big)^{-1/2}\big)\leqslant\big(\frac{M-1}{M}\big)^{-1/2}\\
\Longleftrightarrow\ &\frac{1}{6}\sum_{i=1}^{3}\big((1+2x_i)^{-1/2}+(1-2x_i)^{-1/2}\big)\leqslant\big(1-\frac{1}{M}\big)^{-1/2}\\
\Longleftrightarrow\ &\frac{1}{6}\sum_{i=1}^{3}\frac{(1+2x_i)^{1/2}+(1-2x_i)^{1/2}}{(1-4x_i^2)^{1/2}}\leqslant\big(1-\frac{1}{M}\big)^{-1/2}.
\end{aligned}\tag{8.46}$$

To proceed further, we shall make use of the following elementary estimate:

$$\frac{1}{2}\big((1+\xi)^{1/2}+(1-\xi)^{1/2}\big)\leqslant 1-\frac{1}{8}\xi^2\quad\forall\xi\in[-1,1].\tag{8.47}$$

To prove (8.47), by symmetry it is sufficient to consider $\xi\in[0,1]$. Define

$$h(\xi)\triangleq\frac{1}{2}\big((1+\xi)^{1/2}+(1-\xi)^{1/2}\big)-1+\frac{1}{8}\xi^2,\quad[0,1].$$

Then one has

$$\begin{aligned}
h'(\xi)&=\frac{1}{4}\big((1+\xi)^{-1/2}-(1-\xi)^{-1/2}\big)+\frac{1}{4}\xi,\\
h''(\xi)&=-\frac{1}{8}\big((1+\xi)^{-3/2}+(1-\xi)^{-3/2}\big)+\frac{1}{4}.
\end{aligned}$$

Since

$$(1+\xi)^{-3/2}+(1-\xi)^{-3/2}\geqslant 2\cdot(1-\xi^2)^{-3/4}\geqslant 2,$$

one obtains that

$$h''(\xi)\leqslant-\frac{1}{8}\times 2+\frac{1}{4}=0.$$

Therefore,

$$h'(\xi)\leqslant h'(0)=0\implies h(\xi)\leqslant h(0)=0.$$

The estimate (8.47) thus follows. It follows from (8.46) and (8.47) that

$$\frac{1}{6}\sum_{i=1}^{3}\frac{(1+2x_i)^{1/2}+(1-2x_i)^{1/2}}{(1-4x_i^2)^{1/2}}\leqslant\frac{1}{3}\sum_{i=1}^{3}\frac{1}{(1-4x_i^2)^{1/2}}-\frac{1}{6}|\mathbf{x}|^2.\tag{8.48}$$

Our next step is to further upper bound the summation on the right hand side of (8.48). To this end, one first observes that

$$\begin{aligned} 4x_i^2 = \frac{4k_i^2}{(1+\alpha^2+|\mathbf{k}|^2)^2} &\leqslant \frac{4k_i^2}{(2+|\mathbf{k}|^2)^2} \quad (\text{since } \alpha \geqslant 1) \\ &\leqslant \frac{4|\mathbf{k}|^2}{(2+|\mathbf{k}|^2)^2} \leqslant \sup_{r\geqslant 1} \frac{4r}{(2+r)^2} =: \beta^2 < 1. \end{aligned}$$

Let us consider the function

$$f(y) \triangleq (1-y)^{-1/2}, \quad y \in [0, \beta^2]$$

and set

$$g(y) \triangleq f(y) - 1 - \frac{1}{2}y - Ky^2$$

where K is some universal constant to be specified later on. It follows that

$$g''(y) = \frac{3}{4}(1-y)^{-5/2} - 2K \leqslant \frac{3}{4}(1-\beta^2)^{-5/2} - 2K.$$

If one chooses

$$K \triangleq \frac{3}{8}(1-\beta^2)^{-5/2},$$

then $g''(y) \leqslant 0$ and simple calculus shows that $g(y) \leqslant 0$. As a result,

$$(1-4x_i^2)^{-1/2} \leqslant 1 + 2x_i^2 + 16Kx_i^4,$$

and one thus obtains that

$$\frac{1}{3}\sum_{i=1}^{3} \frac{1}{(1-4x_i^2)^{1/2}} \leqslant 1 + \frac{2}{3}|\mathbf{x}|^2 + C|\mathbf{x}|^4$$

where $C \triangleq \frac{16K}{1-4\beta^2}$. It follows from (8.48) that

$$\frac{1}{6}\sum_{i=1}^{3} \frac{(1+2x_i)^{1/2} + (1-2x_i)^{1/2}}{(1-4x_i^2)^{1/2}} \leqslant 1 + \frac{2}{3}|\mathbf{x}|^2 + C|\mathbf{x}|^4 - \frac{1}{6}|\mathbf{x}|^2. \tag{8.49}$$

On the other hand, by simple calculus the right hand side of (8.46) admits the following lower bound:

$$\big(1 - \frac{1}{M}\big)^{-1/2} \geqslant 1 + \frac{1}{2M}. \tag{8.50}$$

In view of (8.46), (8.49) and (8.50), in order that f_α is superharmonic it is sufficient to choose $\alpha \geqslant 1$ such that

$$1 + \frac{1}{2M} \geqslant 1 + \frac{2}{3}|\mathbf{x}|^2 + C|\mathbf{x}|^4 - \frac{|\mathbf{x}|^2}{6}.$$

Recalling that $M = 1 + \alpha^2 + |\mathbf{k}|^2$, the above inequality is also equivalent to

$$\frac{1}{2M} \geqslant \frac{|\mathbf{k}|^2}{2M^2} + \frac{C|\mathbf{k}|^4}{M^4} \iff 2C|\mathbf{k}|^4 \leqslant (1+\alpha^2)M^2. \tag{8.51}$$

Since $M^2 \geqslant |\mathbf{k}|^4$, by choosing α satisfying $1 + \alpha^2 \geqslant 2C$ one can ensure that (8.51) holds. As a consequence, with such a choice of α one concludes that f_α is a non-constant, non-negative superharmonic function on $\mathbb{Z}^3$. According to Theorem 8.10, the simple random walk on $\mathbb{Z}^3$ is transient.

Finally, we consider higher dimensions. Let $X_n = (X_n^1, X_n^2, X_n^3, \cdots, X_n^d)$ be the simple random walk on $\mathbb{Z}^d$ $(d \geqslant 4)$ and define $Y_n \triangleq (X_n^1, X_n^2, X_n^3)$. Note that $\{Y_n\}$ is a random walk on $\mathbb{Z}^3$ with step distribution

$$\frac{1}{2d}\sum_{i=1}^{3}\delta_{\mathbf{e}_i} + \big(1 - \frac{3}{d}\big)\delta_{\mathbf{0}}.$$

Let $\{Z_n\}$ be the simple random walk on $\mathbb{Z}^3$. It is simple algebra that $f : \mathbb{Z}^3 \to [0, \infty)$ is superharmonic for $\{Y_n\}$ if and only if it is superharmonic for $\{Z_n\}$. According to Theorem 8.10, $\{Y_n\}$ and $\{Z_n\}$ are both recurrent or transient at the same time. It then follows from the three dimensional case that $\{Y_n\}$ is transient. This implies that $\{X_n\}$ is also transient (since $\{X_n\}$ recurrent $\Longrightarrow$ $\{Y_n\}$ recurrent). □

Remark 8.17. The construction of f_α is also motivated from the its continuous counterpart. On $\mathbb{R}^3$, one checks that the function $F_\alpha(\mathbf{x}) \triangleq (\alpha^2 + |\mathbf{x}|^2)^{-1/2}$ satisfies

$$\Delta F_\alpha \triangleq \frac{\partial^2 F_\alpha}{\partial x_1^2} + \frac{\partial^2 F_\alpha}{\partial x_2^2} + \frac{\partial^2 F_\alpha}{\partial x_3^2} = -3\alpha^2(\alpha^2 + |\mathbf{x}|^2)^{-1/2} \leqslant 0.$$

This motivates the construction of f_α as the discrete analogue of F_α. The main extra difficulty here is that the choice of α making f_α superharmonic is not entirely obvious, due to the more complicated shape of the discrete Laplacian.

References

[BC05] A.D. Barbour and L.H. Chen. *An introduction to Stein's method*. World Scientific, 2005.

[Bil68] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1968.

[Bil86] P. Billingsley. *Probability and measure*. John Wiley & Sons, 1986.

[Chu01] K.L. Chung. *A course in probability theory*, 3rd Edition. Academic Press, 2001.

[DZ09] A. Dembo and O. Zeitouni. *Large deviations: techniques and applications*, 2nd Edition. Springer, 2009.

[Dur19] R. Durrett. *Probability: theory and examples*. Fifth Edition. Cambridge University Press, 2019.

[Ete81] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrscheinlichkeitstheorie* 55 (1981): 119–122.

[Fel68] W. Feller. *An introduction to probability theory and its applications*, Volumes I & II. John Wiley & Sons, 1968 & 1971.

[Fol99] G.B. Folland. *Real analysis: modern techniques and their applications*, 2nd Edition. John Wiley & Sons, 1999.

[Hal60] P.R. Halmos. *Naive set theory*. Van Nostrand Reinhold Company, 1960.

[HS75] E. Hewitt and K. Stromberg. *Real and abstract analysis*. Springer-Verlag, 1975.

[Kel97] O. Kallenberg. *Foundations of modern probability*. Springer, 1997.

[Lan93] S. Lang. *Real and functional analysis*, 3rd Edition. Graduate texts in Mathematics. Springer, 1993.

[NP12] I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus*. Cambridge University Press, 2012.

[Shi96] A.N. Shiryaev. *Probability*, 2nd Edition. Graduate Texts in Mathematics. Springer, 1996.

[Str05] D.W. Stroock. *An introduction to Markov processes*. Graduate Texts in Mathematics. Springer, 2005.

[Str11] D.W. Stroock. *Probability theory: an analytic view*, 2nd Edition. Cambridge University Press, 2011.

[Wil91] D. Willliams. *Probability with martingales*. Cambridge University Press, 1991.