Vladislav Kargin

Probability Theory

Lecture Notes

Binghamton University
May 2026

Въ нѣ́дра вмета́ется жре́бій,
ѿ Гдⷭ́ же всѣ́ оу҃правле́нїе є҆гѡ̀.

Proverbs 16:33

Contents

Comments are welcome!

Figure 1: Applications of probability theory

Chapter 1

Review of Measure Theory

This chapter reviews the measure-theoretic foundations on which the rest of the course is built. The main goal of measure theory is to develop a general theory of "size" and integration on abstract spaces that is stable under countable operations—limits, sums, products—and then use it to analyze functions via L^p spaces and convergence theorems. Probability theory is, to a large extent, measure theory with total mass 1 and a stochastic interpretation: measure theory provides the language and tools (extension theorems, integration, convergence, products, conditioning) that make infinite probabilistic constructions and limit arguments precise.

We begin with probability spaces, σ-algebras, and measures, then turn to the Carathéodory extension theorem and Dynkin's π-λ machinery, and finish with the Lebesgue–Stieltjes theorem and Kolmogorov's extension theorem.

1.1 Probability Spaces

Probability space, σ-algebra, and probability measure

Definition 1.1. A **measurable space** is a pair $(\Omega, \mathcal{F})$, where Ω is a set and $\mathcal{F}$ is a σ-algebra. The sets in $\mathcal{F}$ are called **measurable**.

To connect it to the further development, $\mathcal{F}$ is a collection of all subsets which we are able to measure with a probability measure. What is a σ-algebra?

Definition 1.2. A **σ-algebra** $\mathcal{F}$ is a class of subsets of a set Ω that contains $\emptyset$ and is closed under taking complements and countable unions. That is,

It is also often called **σ-field**.

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Note that by DeMorgan's law, $\mathcal{F}$ is also closed under countable intersections:

$$A_1, A_2, ... \in \mathcal{F} \Rightarrow \cap_i A_i \in \mathcal{F}.$$

(or **field** in the different terminology)

An **algebra** $\mathcal{A}$ is a collection of subsets which contains $\emptyset$, and which is closed under complements and **finite unions**. (In a σ-algebra $\mathcal{F}$, the finite unions are also possible because $\emptyset \in \mathcal{F}$ and one can take the empty set for all A_k with $k > n$.)

If $\mathcal{A}$ is an algebra, we denote by $\sigma(\mathcal{A})$ the smallest σ-algebra that contains $\mathcal{A}$ and say that $\mathcal{A}$ **generates** $\sigma(\mathcal{A})$.

Another object which is often used is a **ring**. In measure theory, $\mathcal{R}$ is a ring if it is a collection of sets that contains empty set and is closed under finite unions and differences, i.e., if $A, B \in \mathcal{R}$ then $A \setminus B \in \mathcal{R}$. An algebra is always a ring but the converse is not true. The algebra is a ring with Ω. (You can check these statements as an exercise.)

Example 1.3. Let $\Omega = \mathbb{Z} =$ the integers. Let $\mathcal{R}$ be the finite subsets of $\mathbb{Z}$. This is a ring but not an algebra. Let $\mathcal{A}$ be the subsets of $\mathbb{Z}$ which are either finite or cofinite, (i.e. either A or A^c is finite. This is an algebra but not a σ-algebra [why?]. What σ - algebra does this algebra generate?

Example 1.4. The collection of all subsets of $\mathbb{R}$, $2^{\mathbb{R}}$, is a σ - algebra.

Example 1.5. The class of all finite disjoint unions of intervals $(a, b]$ in $\Omega = \mathbb{R}$ is a ring but not an algebra. If we allow $a = -\infty$ and $b = \infty$, it will be an algebra but not a σ - algebra. The σ-algebra that it generates is called the **Borel σ-algebra** $\mathcal{B}(\mathbb{R})$. It is possible to show that it is strictly smaller than the class of all subsets of $\mathbb{R}$, $2^{\mathbb{R}}$.

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. A function $X : \Omega_1 \to \Omega_2$ is called **measurable** if the inverse image of every measurable set is measurable.

$$X^{-1}(A) \equiv \{\omega : X(\omega) \in A\} \in \mathcal{F}_1 \text{ for all } A \in \mathcal{F}_2$$

In the literature, random variables are usually denoted either by uppercase Latin letters, or by lowercase Greek letters. We will use both conventions.

For conciseness we will also write $\{X \in A\}$ or $(X \in A)$ as a shorthand for $\{\omega : X(\omega) \in A\}$.

A **random variable** X is a measurable function $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(R))$. Sometimes, we will use this term to designate measurable functions that take values in $\mathbb{R}^* = [-\infty, +\infty]$ with the σ-algebra generated by the sets $[-\infty, a)$, (a, b) and $(b, \infty]$.

A **random vector** is a measurable function that takes values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, so a random variable is a particular case of a random vector.

Definition 1.6. A **measure** is a function $\mu : \mathcal{F} \to [0, \infty]$ that satisfies the following conditions:

1. $\mu(\emptyset) = 0$.
2. If $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets, then $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$.

A **probability measure** μ is a measure with additional property $\mu(\Omega) = 1$.

A measure μ is **finite** if $\mu(\Omega) < \infty$ and **σ-finite**, if there is a countable sequence $A_1, A_2, \ldots \in \mathcal{F}$, such that $\sigma(A_i) < \infty$ and $\cup_i A_i = \Omega$.

A **null set** is a set $N \in \mathcal{F}$ such that $\mu(N) = 0$.

Example 1.7 (Zero measure). $\mu(A) = 0$ for any $A \in \mathcal{F}$.

Example 1.8 (Counting measure). $\mu(A) = \#A$, where $\#A$ is the number of elements in A. In particular $\mu(A) = \infty$ if A is infinite. Is it σ-finite on $\mathbb{R}$?

Example 1.9 (Dirac (point mass)). For an $x \in \Omega$ and $A \in \mathcal{F}$, $\delta_x A = \mathbf{1}_{x \in A}$.

A set $A \in \mathcal{F}$ is an **atom** of measure μ if A has no measurable subsets of strictly smaller positive measure. A point $x \in \Omega$ is an atom if $\mu(\{x\}) > 0$.

A measure is **purely atomic** (or **discrete**) if the union of its atoms has the full measure:

$$\mu\Big(\bigcup\{A : A \text{ is an atom of } \mu\}\Big) = \mu(\Omega).$$

A measure is called **non-atomic (atomless)** if it has no atoms. Is there an atomless, σ-finite measure on Borel σ-algebra $\mathcal{B}(\mathbb{R})$? The answer will be given by the Carathéodory Theorem.

Definition 1.10 (Probability space)**.** A **measure space** is a triple $(\Omega, \mathcal{F}, \mu)$, where $\mathcal{F}$ is a σ-algebra, and μ is a measure. If μ is a probability measure, then $(\Omega, \mathcal{F}, \mu)$ is called a **probability measure space** or simply **probability space**.

Countable additivity has several useful consequences collected in the next theorem. Indeed, one of the main reasons we insist on countable additivity rather than mere finite additivity is precisely to guarantee these properties. Before stating the theorem, we fix some notation. We write $A_n \uparrow A$ to mean that $A_1 \subset A_2 \subset \ldots$ and $\bigcup_i A_i = A$; correspondingly, $\mu(A_n) \uparrow \mu(A)$ means that $\mu(A_1) \leq \mu(A_2) \leq \ldots$ and $\lim_{i\to\infty} \mu(A_i) = \mu(A)$. The notations $A_n \downarrow A$ and $\mu(A_n) \downarrow \mu(A)$ are defined analogously.

Theorem 1.11. *Let μ be a measure on a σ-algebra $\mathcal{F}$.*

1. *Monotonicity: If $A \subset B$ then $\mu(A) \leq \mu(B)$.*
2. *Countable subadditivity: If $A_1, A_2 \ldots$ and $\bigcup_{k=1}^{\infty} A_k$ lie in $\mathcal{F}$, then*
$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mu(A_k).$$
3. *Continuity from below: If A_n and A lie in $\mathcal{F}$ and $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$.*
4. *Continuity from above: If A_n and A lie in $\mathcal{F}$, $\mu(A_1) < \infty$, and $A_n \downarrow A$, then $\mu(A_n) \downarrow \mu(A)$.*

(Also see Theorem 1.1.1 in Durrett.)

Proof. (1) If $A \subset B$, then $B = A \cup (B \setminus A)$ is a disjoint union, so by countable additivity $\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A)$.

(2) Define $B_1 = A_1$ and $B_k = A_k \setminus \bigcup_{i<k} A_i$ for $k \geq 2$. Then $\bigcup_{k\geq 1} A_k = \bigsqcup_{k\geq 1} B_k$ is a disjoint union and $B_k \subset A_k$, hence

$$\mu\left(\bigcup_{k\geq 1} A_k\right) \;=\; \sum_{k\geq 1} \mu(B_k) \;\leq\; \sum_{k\geq 1} \mu(A_k).$$

(3) If $A_n \uparrow A$, set $C_1 = A_1$ and $C_k = A_k \setminus A_{k-1}$ for $k \geq 2$. Then $A = \bigsqcup_{k\geq 1} C_k$ and $\mu(A_n) = \sum_{k\leq n} \mu(C_k) \uparrow \sum_{k\geq 1} \mu(C_k) = \mu(A)$.

(4) If $A_n \downarrow A$ and $\mu(A_1) < \infty$, put $B_n = A_1 \setminus A_n$. Then $B_n \uparrow A_1 \setminus A$. By (3), $\mu(B_n) \uparrow \mu(A_1 \setminus A) = \mu(A_1) - \mu(A)$, hence $\mu(A_n) = \mu(A_1) - \mu(B_n) \downarrow \mu(A)$. □

Remark 1.12. The finiteness hypothesis in **continuity from above** is necessary: for Lebesgue measure λ, let $A_n = (n, \infty)$. Then $A_n \downarrow \varnothing$ but $\lambda(A_n) = \infty \not\downarrow 0$.

1.2 Construction of measures: Carathéodory Theorem

Carathéodory Theorem

How do we construct a measure? Here is the typical path: Define a pre-measure μ_0 on a semialgebra (e.g., finite disjoint unions of half-open intervals), extend to an outer measure, and then to a measure on the Carathéodory-measurable sets. In practice you often first enlarge a semialgebra to an algebra/ring and give μ_0 there; Carathéodory then produces the target measure (e.g., Lebesgue measure).

The you typically want to prove that this is a unique extension and have a tool so that the properties that can be easily checked for the pre-measure on the semialgebra can be lifted to the properties of the measure on the entire σ-algebra. This is done using Dynkin's $\pi - \lambda$-machinery.

Our first goal is to understand the concepts in this general description.

Definition 1.13. A collection $\mathcal{S}$ of sets $S \in \Omega$ is said to be a **semialgebra** if

1. $\Omega \in \mathcal{S}$,

2. $\mathcal{S}$ is closed under intersection, i.e., $A, B \in \mathcal{S}$ implies $A \cap B \in \mathcal{S}$, and

3. if $A \in \mathcal{S}$ then its complement A^c is a finite disjoint union of sets in $\mathcal{S}$.

(If we remove the requirement (1), and replace the requirement (3) with "for every $A, B \in \mathcal{S}$, $A \setminus B$ is a finite disjoint union of sets in $\mathcal{S}$", then we obtain the definition of **semiring**. Every semialgebra is a semiring, but the converse is not true. Both semialgebras snd semirings can be used as starting collections of sets in the Carathéodory theorem.

Example 1.14. $\mathcal{S}$ is the empty set plus all sets of the form

$$(a_1, b_1] \times \cdots \times (a_d, b_d] \subset \mathbb{R}^d, \text{ where } -\infty < a_i < b_i < \infty.$$

This is a semiring but not a semialgebra. [Why?]

Now define $\mathcal{S} = \{ \prod_{i=1}^d (a_i, b_i] : \ -\infty \le a_i < b_i \le \infty \}$ in $\Omega = \mathbb{R}^d$ (with the convention $(a, \infty] = (a, \infty)$). Then $\Omega \in \mathcal{S}$, intersections stay in $\mathcal{S}$, and complements of rectangles split into finitely many disjoint such rectangles.

We already defined what is an **algebra**. In Def. 1.13 we replace the requirement (3) by "if $A \in \mathcal{S}$ then its complement $A^c \in \mathcal{S}$. In particular this implies that not only all finite intersections but also all finite unions belong to $\mathcal{S}$.

We can always extend a semialgebra to an algebra.

Lemma 1.15. *If $\mathcal{S}$ is a semialgebra, then $\overline{S} :=$ {finite disjoint unions of sets in $\mathcal{S}$} is an algebra, called the* ***algebra generated by*** *$\mathcal{S}$.*

Proof: Exercise.

Example 1.16. Consider semialgebra $\mathcal{S}$ of intervals in $\mathbb{R}$ as defined in Example 1.14. Then $\overline{\mathcal{S}}_1$ is the empty set and all sets of the form:

$$\bigcup_{i=1}^{k} (a_i, b_i], \text{ where } -\infty \le a_i < b_i \le \infty$$

We can also extend an algebra $\mathcal{S} = \mathcal{A}$ to a σ-algebra. We simply define $\sigma(\mathcal{A})$ as an intersection of all σ-algebras that contain $\mathcal{A}$. This intersection is non-empty since it contains the σ-algebra of all subsets of Ω and one can check that this intersection is indeed a σ-algebra. It is clear that it is the smallest σ-algebra containing $\mathcal{A}$.

Hence, after we extended a semialgebra $\mathcal{S}$ to algebra $\overline{\mathcal{S}}$, we can further extend the algebra $\overline{\mathcal{S}}$ to the σ-algebra $\sigma(\overline{\mathcal{S}})$.

Example 1.17. If $\mathcal{S}_d$ is the semialgebra from Example 1.14, then the σ-algebra $\sigma(\overline{\mathcal{S}}_d)$ generated by $\mathcal{S}_d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$ and is called the **Borel σ-algebra** on $\mathbb{R}^d$. Its elements are called the **Borel sets** of $\mathbb{R}^d$.

Fact: There exist sets in $\mathbb{R}^d$, which are not Borel. (We will not construct one here.) One quick reason: $\mathcal{B}(\mathbb{R})$ has cardinality $|\mathbb{R}|$, while it can be proved that the cardinality of the set of all subsets of $\mathbb{R}$, $2^{\mathbb{R}}$, is strictly larger.

[More generally, for a topological space X, the **Borel σ-algebra** $\mathcal{B}(X)$ is the σ-algebra generated by the open sets of X (equivalently, by the closed sets).]

Now, we can address the question when we can define a suitable function on a semialgebra and extend it to a measure on the generated σ-algebra. The answer is given by the Carathéodory Theorem. This is a difficult Theorem and we give it without proof. (See Durrett or other textbooks if you are interested.)

Definition 1.18 (Premeasure)**.** Let $\mathcal{A} \subseteq 2^\Omega$ be an algebra (or a semi-algebra) of sets. A set function $\mu_0 : \mathcal{A} \to [0, \infty]$ is a **premeasure** if

1. $\mu_0(\varnothing) = 0$;
2. for any pairwise disjoint $(A_n)_{n\geq 1} \subset \mathcal{A}$ with $\bigsqcup_{n\geq 1} A_n \in \mathcal{A}$, one has

$$\mu_0\left(\bigsqcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu_0(A_n).$$

[The difference with the definition of the measure, is that σ-additivity is only required for countable disjoint sums which still belong to the algebra (or semi-algebra)].

We say μ_0 is **σ-finite** if $\Omega = \bigcup_{k\geq 1} E_k$ for some $E_k \in \mathcal{A}$ with $\mu_0(E_k) < \infty$.

Theorem 1.19 (Carathéodory, semialgebra version)**.** *Let $\mathcal{S}$ be a semialgebra on Ω and let $\mu : \mathcal{S} \to [0, \infty]$ with $\mu(\varnothing) = 0$ satisfy:*

(C-1) ***Finite additivity on disjoint unions:*** *if $S = \bigsqcup_{i=1}^m S_i$ with $S, S_i \in \mathcal{S}$, then $\mu(S) = \sum_{i=1}^m \mu(S_i)$.*

(C-2) ***σ-subadditivity on countable disjoint partitions:*** *if $S = \bigsqcup_{i=1}^\infty S_i$ with $S, S_i \in \mathcal{S}$, then $\mu(S) \leq \sum_{i=1}^\infty \mu(S_i)$.*

Then:

1. *(**Premeasure**) μ is countably additive on such partitions; hence it is a **premeasure** on $\mathcal{S}$. It admits a unique finitely additive extension $\overline{\mu}$ to $\overline{\mathcal{S}}$ (the algebra generated by $\mathcal{S}$), given by sums over disjoint decompositions.*
2. *(**Measure extension**) There exists a measure ν on $\sigma(\mathcal{S})$ with $\nu\restriction_{\overline{\mathcal{S}}} = \overline{\mu}$. If $\overline{\mu}$ is σ-finite on $\overline{\mathcal{S}}$, this extension is unique.*

Remark 1.20 (On (C-2)). Assuming (C-1), if $S = \bigsqcup_{i\geq 1} S_i$ with $S, S_i \in \mathcal{S}$, then for each n,

$$\sum_{i=1}^n \mu(S_i) = \mu\left(\bigcup_{i=1}^n S_i\right) \leq \mu(S).$$

Letting $n \uparrow \infty$ gives $\sum_{i\geq 1} \mu(S_i) \leq \mu(S)$. Together with (C-2), $\mu(S) \leq \sum_{i\geq 1} \mu(S_i)$, we obtain

$$\mu(S) = \sum_{i\geq 1} \mu(S_i).$$

Hence (C-1)+(C-2) imply that μ is countably additive on disjoint partitions that remain in $\mathcal{S}$; i.e., μ is a premeasure on $\mathcal{S}$.

Remark 1.21. Often the Carathéodory theorem is formulated as the existence and uniqueness of the measure extension from algebra $\overline{\mathcal{S}}$ to σ-algebra $\sigma(\overline{\mathcal{S}})$, i.e., the statement (ii) in our theorem.

This theorem gives a convenient tool for checking that a function on $\mathcal{S}$ can be extended to $\sigma(\mathcal{S})$. The main point is that it is enough to check σ-additivity for decompositions of sets in $\mathcal{S}$ as a disjoint countable union of sets in $\mathcal{S}$, not for significantly more complicated decompositions of sets in $\sigma(\mathcal{S})$ into disjoint unions of sets in $\sigma(\mathcal{S})$. Exercise 1.43 gives an instructive example of what happens when the premeasure condition fails.

Why not all sets are measurable? Invariance vs. additivity

Example 1.22 (No rotation–invariant **countably additive** probability on all subsets of S^1). Assume there is a probability measure μ on $\mathcal{P}(S^1)$ that is invariant under all rotations and countably additive. Identify S^1 with $\mathbb{R}/\mathbb{Z}$. Declare $x \sim y$ iff $x - y \in \mathbb{Q}$. By the Axiom of Choice, pick a set $A \subset S^1$ containing exactly one representative of each equivalence class and, for $q \in \mathbb{Q} \cap [0,1)$, set $A_q := A + q \pmod 1$. Then $\{A_q\}_{q \in \mathbb{Q} \cap [0,1)}$ are pairwise disjoint and $\bigcup_{q \in \mathbb{Q} \cap [0,1)} A_q = S^1$. By rotation invariance, $\mu(A_q)$ is constant in q. If $\mu(A_q) = 0$, then $\mu(S^1) = \sum_q 0 = 0$, contradiction; if $\mu(A_q) = c > 0$, then $\mu(S^1) = \sum_q c = +\infty$, contradiction. Hence such μ cannot exist.

Remark 1.23. The contradiction uses **countable** additivity. In contrast, because S^1 is an amenable group, there do exist **finitely additive** rotation–invariant probability measures on **all** subsets of S^1 (nonconstructive). They cannot be countably additive.

Example 1.24 (Banach–Tarski (statement) and a corollary for S^2). There is a partition of S^2 into finitely many pairwise disjoint pieces which, after applying only rotations, can be reassembled into two copies of S^2 (Banach–Tarski paradox; uses the Axiom of Choice and the non-amenability of SO(3)).

Corollary 1.25 (No rotation–invariant finitely additive extension of area to all subsets of S^2)**.** *There is no finitely additive set function $m : \mathcal{P}(S^2) \to [0, \infty)$ such that (i) $m(S^2) = 1$; (ii) $m(gE) = m(E)$ for all rotations g and all $E \subseteq S^2$; and (iii) $m(\{x\}) = 0$ for every point $x \in S^2$.*

Idea. If such m existed, apply it to a Banach–Tarski decomposition $S^2 = \bigsqcup_{i=1}^{k} E_i$ and the rotated copies that form two disjoint spheres: by invariance and finite additivity, $1 = m(S^2) = \sum_i m(E_i) = \sum_i m(g_i E_i) = m(\text{one copy}) = m(\text{two copies}) = 2$, a contradiction. The hypothesis $m(\{x\}) = 0$ avoids pathologies from removing/counting finitely many exceptional points. □

In short, on S^1, invariance + countable additivity on all subsets is impossible; on S^2, even finite additivity + invariance (with points null) is impossible – hence we restrict domains.

1.3 Dynkin's $\pi - \lambda$ and monotone class theorems

The uniqueness part of the Carathéodory theorem is proved using so-called Dynkin's π- and λ-systems. Since they are useful in other contexts, we give some additional details on them.

Definition 1.26. A system of sets $\mathcal{A}$ is called a **π-system** if for any $A_1, A_2 \in \mathcal{A}$, we have $A_1 \cap A_2 \in \mathcal{A}$.

Definition 1.27. A collection of subsets $\mathcal{D}$ of set Ω is called a **λ-system** (or d-system, or Dynkin's system) if

1. $\Omega \in \mathcal{D}$
2. If $A \in \mathcal{D}$ and $B \in \mathcal{D}$, $A \subset B \Rightarrow B - A \in \mathcal{D}$

3. If $A_n \in \mathcal{D}$ and $A_n \uparrow A \Rightarrow A \in \mathcal{D}$

Here $A_n \uparrow A$ means that $A_1 \subset A_2 \subset \ldots$ and $\bigcup_n A_n = A$.

Note that every λ-system also satisfy this property:
If $A_n \in \mathcal{D}$ and $A_n \downarrow A \Rightarrow A \in \mathcal{D}$. This is a consequence of the de Morgan law.

The third axiom is also often formulated in the following way:

(iii) if $A_1, A_2, \ldots \in \mathcal{D}$ are pairwise disjoint, then $\bigcup_{n\geq 1} A_n \in \mathcal{D}$.

Now trivially, any σ-algebra is a λ-system. The converse is not true (see Exercise 1.49).

Table 1.1: Families of sets: closure properties

Family	$\boldsymbol{\Omega}$ req.	$\cap$	Complement	Finite $\cup$	Countable $\cup$
π-system	no	✓	—	—	—
λ-system (Dynkin)	✓	no	$B \setminus A$ if $A \subseteq B$	—	✓ (disjoint)
Semialgebra	✓	✓	finite disjoint union	often	—
Algebra (field)	✓	✓	✓	✓	—
σ-algebra	✓	✓	✓	✓	✓

Notes. (1) For a λ-system, "Complement'' means: if $A \subseteq B \in \mathcal{D}$ then $B \setminus A \in \mathcal{D}$; arbitrary complements need not lie in $\mathcal{D}$. (2) The "Countable $\cup$'' for a λ-system is only for **pairwise disjoint** sequences. (3) In a semialgebra $\mathcal{S}$, each $\Omega \setminus A$ ($A \in \mathcal{S}$) must split into a **finite disjoint union** of members of $\mathcal{S}$; closure under finite unions is not required and may fail in general (it holds for the standard rectangle examples in $\mathbb{R}^d$). (4) In a σ-algebra, closure under countable unions implies closure under countable intersections.

Table 1.2: Typical uses (by family)

Family	**Typical use**
π-system	Uniqueness via π–λ; extend independence from generators.
λ-system	Partner for π–λ; lift properties from a π-system to $\sigma(\cdot)$.
Semialgebra	Starting domain for premeasures (Carathéodory extension).
Algebra	Convenient premeasure domain before extension.
σ-algebra	Target domain on which the measure lives.

Notes. (1) **Uniqueness from generators:** if two σ-finite measures agree on a π-system $\mathcal{A}$, then they agree on $\sigma(\mathcal{A})$ (by the π–λ theorem). (2) **Independence from generators:** if $\mathcal{A}_1, \mathcal{A}_2$ are π-systems with $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ for all $A_i \in \mathcal{A}_i$, then $\sigma(\mathcal{A}_1)$ and $\sigma(\mathcal{A}_2)$ are independent. (3) **Construction pipeline:** define a premeasure on a semialgebra/algebra and extend by Carathéodory to a measure on the generated σ-algebra.

Tiny examples.

- **π-system:** all half-open intervals $(a, b] \subset \mathbb{R}$ (closed under finite intersections).
- **λ-system:** for a fixed probability $\mathbb{P}$ and fixed C, the family $\mathcal{D}_C := \{A : \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)\}$ is a λ-system.
- **Semialgebra:** finite disjoint unions of axis-aligned half-open rectangles $\prod_{i=1}^{d}(a_i, b_i] \subset \mathbb{R}^d$ (including infinite endpoints).

- **Algebra:** finite unions of such rectangles (hence closed under complements and finite unions).
- **σ-algebra:** the Borel sets $\mathcal{B}(\mathbb{R}^d) = \sigma(\text{open sets})$.

The importance of λ-systems comes from the following observation.

Lemma 1.28. *If an algebra $\mathcal{A}$ is a λ-system, then it is a σ-algebra.*

Proof. Consider $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$. Define $B_n := \bigcup_{i=1}^{n} A_i \in \mathcal{A}$. It is clear that $B_n \subset B_{n+1}$. Consequently, by the property (3) of a λ-system, $B_n \uparrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$. Similarly, one can show that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$. Therefore, $\mathcal{A}$ is a σ-algebra. □

Theorem 1.29 (Dynkin's π-λ Theorem)**.** *Let $\mathcal{A}$ be a π-system and $\mathcal{D}$ be a λ-system. If $\mathcal{A} \subseteq \mathcal{D}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{D}$.*

Proof. The key here is to show that $\mathcal{D}$ is not only a λ-system but also an algebra. Then, the statement of the theorem will hold by Lemma 1.28.

Let $\mathcal{D}$ be the **smallest λ-system** containing $\mathcal{A}$:

$$\mathcal{D} := \bigcap \{\mathcal{E} : \mathcal{E} \text{ is a } \lambda\text{-system and } \mathcal{A} \subseteq \mathcal{E}\}.$$

Every σ-algebra is a λ-system, and by minimality of $\mathcal{D}$, we have that $\mathcal{A} \subseteq \mathcal{D} \subseteq \sigma(\mathcal{A})$. We aim to prove that this minimal $\mathcal{D}$ is a σ-algebra and therefore $\mathcal{D} = \sigma(\mathcal{A})$ by minimality of $\sigma(\mathcal{A})$.

Step 1: Intersections with generators stay in $\mathcal{D}$. Fix $B \in \mathcal{A}$ and define

$$\mathcal{D}_B := \{ E \subseteq \Omega : E \cap B \in \mathcal{D} \}.$$

We check that $\mathcal{D}_B$ is a λ-system:

- $\Omega \in \mathcal{D}_B$ since $\Omega \cap B = B \in \mathcal{A} \subseteq \mathcal{D}$.
- If $E_1 \subseteq E_2$ lie in $\mathcal{D}_B$, then $(E_2 \setminus E_1) \cap B = (E_2 \cap B) \setminus (E_1 \cap B) \in \mathcal{D}$ because $\mathcal{D}$ is a λ-system. Thus $E_2 \setminus E_1 \in \mathcal{D}_B$.
- If (E_n) are pairwise disjoint in $\mathcal{D}_B$, then $\big(\bigcup_n E_n\big) \cap B = \bigsqcup_n (E_n \cap B) \in \mathcal{D}$, so $\bigcup_n E_n \in \mathcal{D}_B$.

Moreover, since $\mathcal{A}$ is a π-system, for each $A \in \mathcal{A}$ we have $A \cap B \in \mathcal{A} \subseteq \mathcal{D}$ (by assumption), hence $\mathcal{A} \subseteq \mathcal{D}_B$. By the minimality of $\mathcal{D}$ (intersection of all λ-systems containing $\mathcal{A}$), we conclude

$$\mathcal{D} \subseteq \mathcal{D}_B \qquad \text{for every } B \in \mathcal{A}.$$

Hence,

$$E \in \mathcal{D} \subset \mathcal{D}_B,\ B \in \mathcal{A} \implies E \cap B \in \mathcal{D}. \tag{1.1}$$

Step 2: $\mathcal{D}$ is closed under finite intersections. Fix $E \in \mathcal{D}$ and define

$$\mathcal{G}_E := \{ B \subseteq \Omega : E \cap B \in \mathcal{D} \}.$$

Similar to above one can check that $\mathcal{G}_E$ is a λ-system. By (1.1), for every $A \in \mathcal{A}$ we have $E \cap A \in \mathcal{D}$, hence $\mathcal{A} \subseteq \mathcal{G}_E$. By minimality of $\mathcal{D}$, $\mathcal{D} \subseteq \mathcal{G}_E$. Thus,

$$E \in \mathcal{D},\ B \in \mathcal{D} \implies E \cap B \in \mathcal{D} \subseteq \mathcal{G}_E,$$

i.e., $\mathcal{D}$ is closed under finite intersections.

Step 3: $\mathcal{D}$ is closed under finite unions. Note that $\Omega \in \mathcal{D}$ by definition of the λ-system. For $A, B \in \mathcal{D}$, we have $A \cup B = (A^c \cap B^c)^c \in \mathcal{D}$ since $A^c, B^c \in \mathcal{D}$ and $\mathcal{D}$ is closed under intersections.

Hence, the minimal λ-system $\mathcal{D}$ is an algebra, and by Lemma 1.28, it is a σ-algebra. Hence $\sigma(\mathcal{A}) = \mathcal{D}$. As we have seen in the beginning of the proof, this implies the statement of the Dynkin theorem.

□

The main idea behind $\pi - \lambda$ trick is that in many situations you want to push a property checked on simple generators to the whole σ-algebra. Define the class of sets where the property holds; show this class is a λ-system. If your generators form a π-system, the $\pi - \lambda$ theorem upgrades "holds on generators" $\to$ "holds on the σ-algebra."

Here is the example of this trick in action.

Theorem 1.30 (Uniqueness from a π-system under σ-finiteness)**.** *Let $\mathcal{A}$ be a π-system on Ω, and let μ, ν be measures on $\sigma(\mathcal{A})$ such that $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$. Assume there exist $E_n \in \mathcal{A}$ with $\Omega = \bigcup_{n \geq 1} E_n$ and $\mu(E_n) = \nu(E_n) < \infty$ for all n. Then $\mu = \nu$ on $\sigma(\mathcal{A})$.*

Proof. Step 1 (local λ-system). Fix n. Define

$$\mathcal{D}_n := \{E \subseteq \Omega :\ \mu(E \cap E_n) = \nu(E \cap E_n)\}.$$

(This is the system of sets on which measures locally agree.) We claim $\mathcal{D}_n$ is a λ-system: (i) $\Omega \in \mathcal{D}_n$ because $\mu(E_n) = \nu(E_n)$; (ii) if $E_1 \subseteq E_2$ lie in $\mathcal{D}_n$, then using finiteness of $\mu(E_2 \cap E_n)$ we have

$$\begin{aligned} \mu((E_2 \setminus E_1) \cap E_n) &= \mu(E_2 \cap E_n) - \mu(E_1 \cap E_n) \\ &= \nu(E_2 \cap E_n) - \nu(E_1 \cap E_n) = \nu((E_2 \setminus E_1) \cap E_n); \end{aligned}$$

(iii) if (E_k) are pairwise disjoint in $\mathcal{D}_n$, then

$$\mu\Big(\bigcup_k E_k \cap E_n\Big) = \sum_k \mu(E_k \cap E_n) = \sum_k \nu(E_k \cap E_n) = \nu\Big(\bigcup_k E_k \cap E_n\Big).$$

Thus $\mathcal{D}_n$ is a λ-system.

Step 2 (verify on generators; apply π–λ). For $A \in \mathcal{A}$, $A \cap E_n \in \mathcal{A}$ and therefore

$$\mu(A \cap E_n) = \nu(A \cap E_n),$$

so $\mathcal{A} \subseteq \mathcal{D}_n$. By the π–λ theorem, $\sigma(\mathcal{A}) \subseteq \mathcal{D}_n$. Hence for every $E \in \sigma(\mathcal{A})$,

$$\mu(E \cap E_n) = \nu(E \cap E_n) \qquad \text{for all } n. \tag{$*$}$$

Step 3 (disjointify the σ-finite cover; sum). Let $F_1 := E_1$ and $F_n := E_n \setminus \bigcup_{k<n} E_k$ for $n \geq 2$. Then (F_n) are pairwise disjoint, each $F_n \subseteq E_n$, and $\Omega = \bigsqcup_{n\geq 1} F_n$. For $E \in \sigma(\mathcal{A})$,

$$\begin{aligned}\mu(E) &= \sum_{n\geq 1} \mu(E \cap F_n) = \sum_{n\geq 1} \mu(E \cap E_n \cap F_n) \overset{(*)}{=} \sum_{n\geq 1} \nu(E \cap E_n \cap F_n) \\ &= \sum_{n\geq 1} \nu(E \cap F_n) = \nu(E),\end{aligned}$$

using countable additivity on disjoint unions. Thus $\mu = \nu$ on $\sigma(\mathcal{A})$. □

Exercise 1.50 shows that the π-system hypothesis cannot be dropped: two measures can agree on a generating collection $\mathcal{C}$ with $\sigma(\mathcal{C}) = \mathcal{F}$ and yet disagree on $\mathcal{F}$.

Another useful tool is the monotone class theorem.

Definition 1.31 (Monotone class)**.** A family $\mathcal{M} \subseteq 2^\Omega$ is a **monotone class** if:

- whenever $A_1 \subseteq A_2 \subseteq \cdots$ with each $A_n \in \mathcal{M}$, then $\bigcup_{n\geq 1} A_n \in \mathcal{M}$;
- whenever $A_1 \supseteq A_2 \supseteq \cdots$ with each $A_n \in \mathcal{M}$, then $\bigcap_{n\geq 1} A_n \in \mathcal{M}$.

Theorem 1.32 (Monotone class (sets))**.** *If $\mathcal{A}$ is an algebra on Ω and $\mathcal{M}$ is a monotone class (i.e. closed under increasing unions and decreasing intersections) with $\mathcal{A} \subseteq \mathcal{M}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{M}$.*

The proof is in Appendix A.2.

The following variant for measurable functions is often more directly useful.

Theorem 1.33 (Monotone class theorem for functions)**.** *Let $\mathcal{H}$ be a vector space of bounded real-valued functions on Ω such that:*

1. *$\mathbf{1}_\Omega \in \mathcal{H}$;*
2. *$\mathbf{1}_E \in \mathcal{H}$ for all E in a π-system $\mathcal{A}$;*
3. *if $0 \leq f_n \uparrow f$ pointwise with f bounded and each $f_n \in \mathcal{H}$, then $f \in \mathcal{H}$.*

Then $\mathcal{H}$ contains every bounded $\sigma(\mathcal{A})$-measurable function.

The proof is in Appendix A.3.

1.4 Lebesgue–Stieltjes + Kolmogorov's extension

The Caratéodory Theorem can be used to define measures on the real line.

Measures on $(\mathbb{R}, \mathcal{B}(R))$ are called **Borel** measures.

Definition 1.34. A **Stieltjes (distribution) function** $F(x)$ is a map $\mathbb{R} \to \mathbb{R}$ which is

(i) nondecreasing,

(ii) right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$.

Theorem 1.35 (Lebesgue–Stieltjes)**.** *To each Stieltjes function F there corresponds a unique Borel measure μ_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that*

$$\mu_F\big((a,b]\big) = F(b) - F(a) \qquad (-\infty \le a < b \le \infty).$$

Conversely, for any Borel measure μ on $\mathbb{R}$, the function $F_\mu(x) := \mu\big((-\infty, x]\big)$ is Stieltjes and $\mu = \mu_{F_\mu}$.

When $F(x) = x$ the measure is called **Lebesgue measure**.
Sketch of the proof

1. Check pre-measure.
2. Apply Caratéodory.
3. For converse, check that $F_\mu(x)$ is non-decreasing and right-continuous and apply uniqueness.

The full proof is in Appendix A.4.

Completion

The Lebesgue measure can be extended to a larger class of sets than Borel sets by adding all sets that are contained in Borel sets of measure zero and assigning measure zero to them. By passing to a minimal containing σ-algebra, one obtains a σ-algebra of **Lebesgue**-measurable sets. This σ-algebra is larger than the Borel σ-algebra, and the Lebesgue measure can be extended to this larger σ-algebra.

Kolmogorov's extension theorem

Definition 1.36 (Cylinder σ-algebra and finite-dimensional laws)**.** Let T be an index set (e.g. $T = \mathbb{N}$) and let $(E, \mathcal{E})$ be a standard Borel space (e.g., $E = \mathbb{R}^d$ with its Borel σ-algebra). For finite $J \subset T$, write $\pi_J : E^T \to E^J$ for the coordinate projection, $\pi_J(\omega) = (\omega_t)_{t \in J}$. The **cylinder σ-algebra** on E^T is $\mathcal{E}^{\otimes T} := \sigma\{\pi_J^{-1}(A) : J \subset T \text{ finite}, A \in \mathcal{E}^{\otimes J}\}$.

A **finite-dimensional distribution** (f.d.d.) on E^T is a probability measure μ_J on $(E^J, \mathcal{E}^{\otimes J})$ for some finite $J \subset T$.

Definition 1.37 (Consistency (projectivity))**.** A family $\{\mu_J : J \subset T, |J| < \infty\}$ of f.d.d.'s is **consistent** if for all finite $K \subset J \subset T$,

$$\mu_K = \mu_J \circ \pi_{J \to K}^{-1}, \qquad \text{where } \pi_{J \to K} : E^J \to E^K \text{ is the coordinate projection.}$$

(Equivalently: marginalizing μ_J down to K gives μ_K.)

Theorem 1.38 (Kolmogorov extension / consistency)**.** *Let $(E, \mathcal{E})$ be a standard Borel space and T an arbitrary index set. If $\{\mu_J\}_{J \subset T,\ |J| < \infty}$ is a consistent family of finite-dimensional distributions on $(E^J, \mathcal{E}^{\otimes J})$, then there exists a **unique** probability measure $\mathbb{P}$ on $(E^T, \mathcal{E}^{\otimes T})$ such that*

$$\mathbb{P} \circ \pi_J^{-1} = \mu_J \quad \textit{for every finite } J \subset T.$$

Remark 1.39. **How you use it.** Define the f.d.d.'s you want (often by independence/product on rectangles), check the consistency condition, and invoke Theorem 1.38 to get a full process law on $(E^T, \mathcal{E}^{\otimes T})$. Uniqueness is with respect to the cylinder σ-algebra $\mathcal{E}^{\otimes T}$.

Example 1.40 (Countable Bernoulli product $\{0,1\}^{\mathbb{N}}$). Take $E = \{0,1\}$ with $\mathcal{E} = 2^E$, fix $p \in [0,1]$, and for each finite $J \subset \mathbb{N}$ define μ_J on E^J by independence with parameter p:

$$\mu_J\big(\{x \in E^J : \ x_t = a_t \ \forall t \in J\}\big) \ = \ \prod_{t \in J} p^{a_t}(1-p)^{1-a_t}, \qquad a_t \in \{0,1\}.$$

These μ_J are consistent (marginalization just drops factors), so by Theorem 1.38 there is a unique $\mathbb{P}$ on $\big(\{0,1\}^{\mathbb{N}},\, 2^{\{0,1\}^{\mathbb{N}}}\big)$ with $\mathbb{P} \circ \pi_J^{-1} = \mu_J$ for all finite J.

1.5 Exercises

Homework 1 — Measure Bootcamp (4 problems)

Exercise 1.41 (1. Uniqueness from half–lines). Let μ, ν be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mu((-\infty, a]) = \nu((-\infty, a])$ for every rational a.

1. Show $\sigma\{(-\infty, a] : a \in \mathbb{Q}\} = \mathcal{B}(\mathbb{R})$.
2. Conclude $\mu = \nu$ on $\mathcal{B}(\mathbb{R})$. (Give a proof for this step.)

Exercise 1.42 (2. Stieltjes measure: locate the atoms (new computation)). Let F be nondecreasing, right–continuous with $F(\pm\infty)$ finite, and let μ_F be its Stieltjes measure.

1. Prove $\mu_F(\{x\}) = F(x) - F(x-)$ where $F(x-) := \lim_{y \uparrow x} F(y)$.
2. Take the concrete

$$F(x) = \begin{cases} 0, & x \le 0, \\ \frac{1}{3}, & 0 < x \le 1, \\ \frac{1}{3} + \frac{x-1}{3}, & 1 < x \le 4, \\ 1, & x > 4, \end{cases}$$

and compute $\mu_F(\{0\}),\ \mu_F((0,1]),\ \mu_F(\{1\}),\ \mu_F((1,4]),\ \mu_F(\mathbb{R})$.

Exercise 1.43 (3. When Carathéodory fails on $\mathbb{Q}$). Let $\Omega = \mathbb{Q} \cap (0,1]$, $\mathcal{S} = \{(a,b]_{\mathbb{Q}}\}$ and $\mu((a,b]_{\mathbb{Q}}) = b - a$. Construct pairwise disjoint $E_n \in \overline{\mathcal{S}}$ with $\bigcup_n E_n = \Omega$ and $\sum_n \mu(E_n) < \varepsilon$ for an arbitrary $\varepsilon \in (0,1)$. Conclude that μ is **not** a premeasure on $\overline{\mathcal{S}}$, so no countably additive extension exists. Which hypothesis of the Carathéodory extension theorem fails here, and why?

Exercise 1.44 (4. Kolmogorov with a twist). For $p \in (0,1)$ and finite $J \subset \mathbb{N}$, define $\mathbb{P}_J$ on $\{0,1\}^J$ by independent Bern(p).

1. Verify consistency under coordinate projections.

2. Define the cylinder semiring $\mathcal{S} = \{\pi_J^{-1}(A)\}$ and the premeasure $\overline{\mu}(\pi_J^{-1}A) := \mathbb{P}_J(A)$. Prove $\overline{\mu}$ is σ–additive on **this** $\mathcal{S}$ by continuity at $\varnothing$ (you may quote the criterion, but carry out the compact/diagonal step for $\{0,1\}^{\mathbb{N}}$ explicitly).

3. State the resulting Kolmogorov extension conclusion precisely.

Additional Exercises

Exercise 1.45 (Continuity and where it can fail). Let μ be a measure on $(\Omega, \mathcal{F})$.

1. Prove continuity from above: if $A_n \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_n) \downarrow \mu(A)$.

2. Give a counterexample showing the hypothesis $\mu(A_1) < \infty$ is necessary.

3. Suppose μ is only finitely additive on an algebra $\mathcal{A}$. Identify precisely where the standard proof of continuity from below $A_n \uparrow A \Rightarrow \mu(A_n) \uparrow \mu(A)$ breaks.

Exercise 1.46 (Counting-measure continuity). On $(\mathbb{N}, 2^{\mathbb{N}})$ with counting measure μ, let $A_n = \{n, n+1, \dots\}$. Verify $A_n \downarrow \varnothing$ and $\mu(A_n) = \infty$ for all n. Which hypothesis of continuity from above fails?

Exercise 1.47 (Equivalent formulation of continuity from below). Let μ be a measure on $(\Omega, \mathcal{F})$. Show that continuity from below is equivalent to: if $A_n \uparrow A$ then $\mu(A \setminus A_n) \downarrow 0$.

Exercise 1.48 (Equality case in subadditivity). Let μ be a measure on $(\Omega, \mathcal{F})$ and let $A_1, A_2, \dots \in \mathcal{F}$. Prove that

$$\mu\left(\bigcup_{k\geq 1} A_k\right) = \sum_{k\geq 1} \mu(A_k)$$

holds if and only if $\mu\big(A_k \cap \bigcup_{i<k} A_i\big) = 0$ for every k (i.e., the sets are disjoint up to null sets).

Exercise 1.49 (λ-system that is not an algebra). Let $\Omega = \{1,2,3,4\}$. Give an example of a λ-system of subsets of Ω which is not an algebra.

Exercise 1.50 (Why the π-system hypothesis matters). Give an example of two probability measures $\mu \neq \nu$ on $\mathcal{F} =$ all subsets of $\{1,2,3,4\}$ that agree on a collection of sets $\mathcal{C}$ with $\sigma(\mathcal{C}) = \mathcal{F}$, i.e., the smallest σ-algebra containing $\mathcal{C}$ is $\mathcal{F}$.

Exercise 1.51 (Disjointification on a semialgebra). Let $\mathcal{S}$ be a semialgebra on Ω and let $\overline{\mathcal{S}}$ be the family of finite unions of sets in $\mathcal{S}$.

1. Show that every $E \in \overline{\mathcal{S}}$ can be written as a **finite disjoint union** of members of $\mathcal{S}$.

2. Conclude that $\overline{\mathcal{S}}$ is an algebra (closed under complement and finite unions).

3. Give an example of a family closed under finite unions but not under relative complements (hence not a ring).

Exercise 1.52 (Kolmogorov for a two-state Markov chain). Let $E = \{0, 1\}$, let π be an initial distribution on E, and let $K = \big(K_{ij}\big)_{i,j\in E}$ be a stochastic matrix. For finite $J = \{1, \dots, n\}$ define $\mathbb{P}_J$ on E^J by

$$\mathbb{P}_J(x_1, \dots, x_n) = \pi(x_1) \prod_{k=1}^{n-1} K_{x_k, x_{k+1}}.$$

1. Show the family $(\mathbb{P}_J)$ is consistent under coordinate projections.

2. Let $\mathcal{S} = \{\pi_J^{-1}(A) : A \subset E^J,\ J \Subset \mathbb{N}\}$ and define the premeasure $\overline{\mu}(\pi_J^{-1}A) := \mathbb{P}_J(A)$. Prove $\overline{\mu}$ is σ-additive on $\mathcal{S}$ via continuity at $\varnothing$.

3. Conclude (Kolmogorov) that there exists a unique $\mathbb{P}$ on $(E^{\mathbb{N}}, 2^{E^{\mathbb{N}}})$ with $\mathbb{P} \circ \pi_J^{-1} = \mathbb{P}_J$ for all finite J.

Exercise 1.53 (Product rectangles premeasure). On $\mathbb{R}^d$, let $\mathcal{S}$ be the semiring of half-open rectangles $R = \prod_{i=1}^d (a_i, b_i]$ with $-\infty < a_i < b_i \leq \infty$ and define $\mu(R) := \prod_{i=1}^d (b_i - a_i)$ (interpreting $b_i - a_i = \infty$ if needed).

1. Show μ is well-defined on $\mathcal{S}$ (independent of representation by finite disjoint unions).

2. Prove finite additivity on disjoint unions in $\mathcal{S}$.

3. Prove σ-additivity on $\mathcal{S}$ via the continuity-at-$\varnothing$ criterion (you may assume inner regularity on $\mathbb{R}^d$).

Exercise 1.54 (Wiener measure (Gaussian consistency)). For $n \geq 1$, let $\mathbb{P}_n$ be the mean-zero Gaussian measure on $\mathbb{R}^n$ with covariance $\Sigma_{ij} = \min\{i, j\}$, $1 \leq i, j \leq n$.

1. Show that $(\mathbb{P}_n)$ is consistent under the projections $(x_1, \dots, x_{n+1}) \mapsto (x_1, \dots, x_n)$.

2. Let $\mathcal{S} = \{\pi_{1:n}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^n),\ n \in \mathbb{N}\}$. Define $\overline{\mu}(\pi_{1:n}^{-1}B) := \mathbb{P}_n(B)$ and prove $\overline{\mu}$ is σ-additive on $\mathcal{S}$ (use continuity at $\varnothing$ and inner regularity of Borel Gaussians on $\mathbb{R}^n$).

3. Conclude (Kolmogorov) that there exists a probability $\mathbb{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F})$ such that $\mathbb{P} \circ \pi_{1:n}^{-1} = \mathbb{P}_n$ for all n; this is **Wiener measure** on the cylinder σ-algebra.

Chapter 2

Convergence Theorems and Inequalities

he central technical challenge of integration theory is understanding when limits and integrals can be interchanged. This chapter presents the three fundamental convergence theorems—the Monotone Convergence Theorem, Fatou's Lemma, and the Dominated Convergence Theorem—together with the key inequalities of Jensen, Hölder, and Minkowski that underpin the L^p space framework. These tools are indispensable throughout probability theory: they justify passing expectations through limits, establish moment bounds, and provide the analytic backbone for the limit theorems in later chapters.

2.1 Expected Value

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 2.1. Let $X : \Omega \to \mathbb{R}$ be a $\mathcal{B}$-measurable random variable. The **expected value** of X is defined by

$$\mathbb{E}(X) := \int_\Omega X \, d\mathbb{P}(\omega) \equiv \int_\Omega X(\omega) \, \mathbb{P}(d\omega) \tag{2.1}$$

The integral is defined as in Lebesgue integration, whenever $\int |X| \, d\mathbb{P} < \infty$.

We also define

$$\|X\|_p := \Big(\mathbb{E}|X|^p\Big)^{1/p}.$$

We defined the expectation of r.v. $X(\omega)$ as the integral with respect to the probability measure $\mathbb{P}$. For a Lebesgue-Stieltjes measure we can write it using the distribution function $F(x) = \mu_X((-\infty, x]) = \mathbb{P}(\{\omega : X(\omega) \le x\})$.

By a variant of the change of variable formula, we can write:

$$\mathbb{E}X = \int_\Omega X(\omega) \, d\mathbb{P}(\omega) = \int_{\mathbb{R}} x \, dF(x),$$

where the last integral is the Stieltjes-Lebesgue integral. If the function $F(x)$ is absolutely continuous and has the derivative $f(x)$, then

$$\mathbb{E}X = \int_{\mathbb{R}} x f(x)\, dx.$$

Another useful formula applies if $X \geq 0$. Then

$$\mathbb{E}X = \int_0^\infty x\, dF(x) = -\int_0^\infty x\, d(1 - F(x)).$$

By integrating the last equation by parts, we get the formula

$$\mathbb{E}X = \int_0^\infty (1 - F(x))\, dx = \int_0^\infty \mathbb{P}(X \geq x)\, dx. \tag{2.2}$$

2.2 Monotone Convergence Theorem (MCT)

Often we approximate a random variable X by a sequence of random variables X_n. For example, we can consider truncations X_n when $X_n = X$ if $X < n$ and $X_n = n$ if $X \geq n$. Clearly $X_n(\omega) \to X(\omega)$ for all ω. Is it true that $\mathbb{E}X_n \to \mathbb{E}X$?

In order to calculate expectations of random variables which are limits of other random variables, there is a set of very useful theorems. In fact, the countable additivity of probability measure is mostly needed to justify these theorems.

We will often say that a property $P(\omega)$ holds **almost surely (a.s.)** or **almost everywhere (a.e.)**, if the measure of the set of ω where this property does not hold is 0.

Notation. We write $f_n \uparrow f$ if f_n increases pointwise to f, i.e.

$$f_n(\omega) \leq f_{n+1}(\omega) \text{ for all } n \quad \text{and} \quad \lim_{n\to\infty} f_n(\omega) = f(\omega)$$

for every $\omega \in \Omega$ (or for μ-a.e. ω, when working up to null sets).

Theorem 2.2 (Monotone Convergence Theorem (MCT))**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f_n : \Omega \to [0, \infty]$ be measurable. Assume $f_n(\omega) \uparrow f(\omega)$ for a.e. ω. Then*

$$\int_\Omega f_n\, d\mu \uparrow \int_\Omega f\, d\mu,$$

i.e. $\lim_{n\to\infty} \int f_n\, d\mu = \int f\, d\mu$ (allowing the value $+\infty$).

Note the positivity assumption on f_n.

Proof. Set $I_n := \int f_n\, d\mu$. Since $0 \leq f_n \leq f_{n+1} \leq f$, by monotonicity of the integral we have $0 \leq I_n \leq I_{n+1} \leq \int f\, d\mu$. Hence $I_n \uparrow L$ for some $L \leq \int f\, d\mu$. It remains to prove $\int f\, d\mu \leq L$. Recall the definition of integral for nonnegative measurable functions:

$$\int f\, d\mu \;=\; \sup\Big\{ \int s\, d\mu :\; s \text{ simple}, 0 \leq s \leq f \Big\}.$$

Goal: Show every simple $s \le f$ satisfies $\int s\,d\mu \le L$. Then the supremum $\le L$.

Fix a simple function $s \le f$ and fix $c \in (0,1)$. Write $s = \sum_{k=1}^{m} a_k \mathbf{1}_{A_k}$ with $a_k \ge 0$ and pairwise disjoint $A_k \in \mathcal{A}$. For each n define

$$A_{k,n} := A_k \cap \{f_n \ge c \cdot a_k\}.$$

(This is the part of A_k where f_n has "caught up" to height $c \cdot a_k$.) Because $f_n \uparrow f$ and $f \ge s = a_k > c \cdot a_k$ on A_k, we have $A_{k,n} \uparrow A_k$ (up to a null set). By continuity from below of measure,

$$\mu(A_{k,n}) \uparrow \mu(A_k).$$

Now define the simple function $s_n := \sum_{k=1}^{m} c \cdot a_k \mathbf{1}_{A_{k,n}}$. Then $0 \le s_n \le f_n$ (since on $A_{k,n}$ we have $f_n \ge c \cdot a_k$ by definition), so

$$\int s_n\,d\mu \;\le\; \int f_n\,d\mu.$$

But also

$$\int s_n\,d\mu \;=\; \sum_{k=1}^{m} c \cdot a_k\,\mu(A_{k,n}) \;\xrightarrow[n\to\infty]{}\; \sum_{k=1}^{m} c \cdot a_k\,\mu(A_k) \;=\; c\int s\,d\mu.$$

Taking limits in $\int s_n \le \int f_n$ yields $c\int s\,d\mu \le L$. Since $c \in (0,1)$ was arbitrary, letting $c \uparrow 1$ gives $\int s\,d\mu \le L$. Since this holds for every simple $s \le f$, taking the supremum gives $\int f\,d\mu \le L$. Therefore $L = \int f\,d\mu$, as claimed. □

In probability, we use this theorem in the following form. Suppose random variables $X_n \uparrow X$ a.e., then $\mathbb{E}X_n \to \mathbb{E}X$.

Here is an example of MCT in action.

Corollary 2.3 (Absolute continuity of the integral (via MCT))**.** *If $f \in L^1(\mu)$, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\mu(E) < \delta \quad\Longrightarrow\quad \int_E |f|\,d\mu < \varepsilon.$$

Proof. Fix $\varepsilon > 0$. Since $f \in L^1$, by MCT applied to $|f|\mathbf{1}_{\{|f|\le M\}} \uparrow |f|$ as $M \to \infty$, we have

$$\int_{\{|f|>M\}} |f|\,d\mu \longrightarrow 0 \qquad (M \to \infty).$$

Choose M so that $\int_{\{|f|>M\}} |f|\,d\mu < \varepsilon/2$, and set $\delta := \varepsilon/(2M)$. Then for any measurable E,

$$\int_E |f|\,d\mu \le \int_{E\cap\{|f|\le M\}} |f|\,d\mu + \int_{E\cap\{|f|>M\}} |f|\,d\mu \le M\,\mu(E) + \int_{\{|f|>M\}} |f|\,d\mu.$$

If $\mu(E) < \delta$, then $M\mu(E) < \varepsilon/2$, hence $\int_E |f|\,d\mu < \varepsilon$. □

2.3 Fatou lemma and DCT

What happens if the convergence is not monotone?

Example 2.4. Define X_n on $[0,1]$ as $X_n = n\mathbf{1}_{(0,1/n)}$. Then $X_n \xrightarrow{a.s.} 0$ but

$$\lim_{n\to\infty} \mathbb{E}(X_n) = 1 > 0. \tag{2.3}$$

Remark: Remember this example to get the right sign in the inequality given by the following result.

Mnemonic: "Fatou looses weight."

Lemma 2.5 (Fatou)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f_n : \Omega \to [0,\infty]$ be measurable. Then*

$$\int_\Omega \liminf_{n\to\infty} f_n \, d\mu \;\le\; \liminf_{n\to\infty} \int_\Omega f_n \, d\mu.$$

In particular, for r.v.s. $X_n \ge 0$, if $X_n \to X$ a.e., then $\mathbb{E}X \le \liminf_{n\to\infty} \mathbb{E}X_n$.

Note that the functions are assumed to be positive.

Proof. For each $k \ge 1$ define

$$g_k(\omega) := \inf_{n\ge k} f_n(\omega).$$

Then g_k is measurable, $0 \le g_k \le g_{k+1}$ (so $g_k \uparrow$), and moreover

$$\lim_{k\to\infty} g_k(\omega) = \sup_{k\ge 1} \inf_{n\ge k} f_n(\omega) = \liminf_{n\to\infty} f_n(\omega).$$

By the Monotone Convergence Theorem,

$$\int_\Omega \liminf_{n\to\infty} f_n \, d\mu = \lim_{k\to\infty} \int_\Omega g_k \, d\mu.$$

Fix k. Since $g_k \le f_n$ for all $n \ge k$, we have

$$\int_\Omega g_k \, d\mu \le \inf_{n\ge k} \int_\Omega f_n \, d\mu.$$

Taking $k \to \infty$ gives

$$\lim_{k\to\infty} \int_\Omega g_k \, d\mu \le \lim_{k\to\infty} \inf_{n\ge k} \int_\Omega f_n \, d\mu = \liminf_{n\to\infty} \int_\Omega f_n \, d\mu,$$

which is exactly Fatou's lemma. □

However, if we have some additional uniform control on the sequence of functions f_n, then we can obtain the convergence of expectations.

Theorem 2.6 (Dominated Convergence Theorem (DCT))**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Suppose $f_n : \Omega \to \mathbb{R}$ are measurable, $f_n \to f$ a.e., and there exists an integrable function $g \in L^1(\mu)$ (that is, such that $\int_\Omega |g| \, d\mu < \infty$) such that*

$$|f_n| \le g \qquad \text{a.e. for all } n.$$

Then $f \in L^1(\mu)$ and

$$\int_\Omega f_n \, d\mu \;\longrightarrow\; \int_\Omega f \, d\mu, \qquad \text{equivalently} \qquad \int_\Omega |f_n - f| \, d\mu \;\longrightarrow\; 0.$$

Proof. First, since $|f_n| \le g$ and $f_n \to f$ a.e., we have $|f| \le g$ a.e., hence $f \in L^1(\mu)$.

Define the nonnegative functions

$$u_n := g + f_n \ge 0, \qquad v_n := g - f_n \ge 0.$$

Then $u_n \to g + f$ and $v_n \to g - f$ a.e. By Fatou's lemma applied to (u_n),

$$\int_\Omega (g+f)\,d\mu \;\le\; \liminf_{n\to\infty} \int_\Omega (g+f_n)\,d\mu = \int_\Omega g\,d\mu + \liminf_{n\to\infty} \int_\Omega f_n\,d\mu.$$

Rearranging,

$$\int_\Omega f\,d\mu \;\le\; \liminf_{n\to\infty} \int_\Omega f_n\,d\mu.$$

Similarly, applying Fatou to (v_n) gives

$$\int_\Omega (g-f)\,d\mu \;\le\; \liminf_{n\to\infty} \int_\Omega (g-f_n)\,d\mu = \int_\Omega g\,d\mu - \limsup_{n\to\infty} \int_\Omega f_n\,d\mu,$$

so

$$\limsup_{n\to\infty} \int_\Omega f_n\,d\mu \;\le\; \int_\Omega f\,d\mu.$$

Combining the two inequalities,

$$\int_\Omega f\,d\mu \;\le\; \liminf_{n\to\infty} \int_\Omega f_n\,d\mu \;\le\; \limsup_{n\to\infty} \int_\Omega f_n\,d\mu \;\le\; \int_\Omega f\,d\mu,$$

hence $\int f_n\,d\mu \to \int f\,d\mu$.

Finally, apply the same argument to the dominated sequence $|f_n - f| \le 2g$ with $|f_n - f| \to 0$ a.e. to obtain $\int |f_n - f|\,d\mu \to 0$. □

Corollary 2.7 (Bounded Convergence)**.** *If the sequence (X_n) of random variables is uniformly bounded and if $X_n \xrightarrow{a.s.} X$ as $n \to \infty$, then*

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

Here is an example of how DCT is typically applied.

Example 2.8 (Truncation by bounded approximants (DCT workhorse))*.* Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X \in L^1(\mathbb{P})$. Define the truncations

$$X_n := X\,\mathbf{1}_{\{|X|\le n\}}, \qquad n \ge 1.$$

Then $X_n \to X$ a.s. (indeed pointwise), and $|X_n| \le |X|$ with $|X| \in L^1$. By the Dominated Convergence Theorem,

$$\mathbb{E}[X_n] \to \mathbb{E}[X] \qquad \text{and} \qquad \mathbb{E}[\,|X_n - X|\,] \to 0.$$

(Equivalently, $X_n \to X$ in L^1.)

Exercise 2.42 shows that the convergence of integrals can hold even without an integrable dominating function, illustrating that domination is sufficient but not necessary. A sufficient and necessary condition is the Uniform Integrability.

Definition 2.9 (Uniform integrability)**.** A family of integrable functions $\mathcal{X} \subset L^1(\mathbb{P})$ is **uniformly integrable (UI)** if

$$\lim_{K\to\infty} \sup_{X\in\mathcal{X}} \mathbb{E}\big[\,|X|\,\mathbf{1}_{\{|X|>K\}}\,\big] = 0.$$

Equivalently, for every $\varepsilon > 0$ there exists K such that $\sup_{X\in\mathcal{X}} \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>K\}}\big] < \varepsilon$.

Theorem 2.10 (Vitali)**.** *On a probability space, if $X_n \to X$ in probability and $\{X_n\}$ is uniformly integrable, then*

$$\mathbb{E}|X_n - X| \to 0.$$

In particular, $\mathbb{E}[X_n] \to \mathbb{E}[X]$ whenever $X \in L^1$. Conversely, $X_n \to X$ in L^1 implies $\{X_n\}$ is uniformly integrable.

2.4 Inequalities

Theorem 2.11 (Jensen's inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and let $X \in L^1$. Assume $\varphi(X) \in L^1$ (equivalently, $\mathbb{E}|\varphi(X)| < \infty$). Then*

$$\varphi(\mathbb{E}X) \le \mathbb{E}[\varphi(X)]. \tag{2.4}$$

Sketch. A standard fact from convex analysis is that a convex function is the supremum of its supporting affine functions (see Appendix B.4). Concretely, there exists a **countable** family of affine functions $L_n(x) = a_n x + b_n$ such that

$$\varphi(x) = \sup_{n\ge 1} L_n(x) \qquad \text{for all } x \in \mathbb{R}.$$

(One may take supporting lines with rational slopes/intercepts, or with tangency points in $\mathbb{Q}$.)

Fix n. By linearity of expectation,

$$L_n(\mathbb{E}X) = a_n\mathbb{E}X + b_n = \mathbb{E}[a_nX + b_n] = \mathbb{E}[L_n(X)].$$

Since $L_n \le \varphi$ pointwise, monotonicity of expectation yields

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

Therefore $L_n(\mathbb{E}X) \le \mathbb{E}[\varphi(X)]$ for all n, and taking the supremum over n gives

$$\varphi(\mathbb{E}X) = \sup_n L_n(\mathbb{E}X) \le \mathbb{E}[\varphi(X)].$$

□

Example 2.12 (Direction check). Take $\varphi(x) = x^2$ (convex). Then Jensen gives

$$(\mathbb{E}X)^2 \le \mathbb{E}[X^2],$$

i.e. $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2 \ge 0$.

Example 2.13. Take $\varphi(x) = |x|$. For $X \in L^1$, Jensen gives

$$|\mathbb{E}X| \le \mathbb{E}|X|.$$

Theorem 2.14 (Hölder's inequality)**.** *Let $p, q \in [1, \infty]$ satisfy $1/p + 1/q = 1$. If $X \in L^p$ and $Y \in L^q$, then $XY \in L^1$ and*

$$\mathbb{E}|XY| \le \|X\|_p \, \|Y\|_q.$$

Here $\|X\|_r = (\mathbb{E}|X|^r)^{1/r}$ for $1 \le r < \infty$, and

$$\|X\|_\infty = \inf\{M \ge 0 : \ \mathbb{P}(|X| > M) = 0\}.$$

Proof. If $p = 1$ and $q = \infty$, then $|XY| \le |X| \|Y\|_\infty$ a.s., hence

$$\mathbb{E}|XY| \le \|Y\|_\infty \, \mathbb{E}|X| = \|X\|_1 \|Y\|_\infty.$$

The case $p = \infty$, $q = 1$ is symmetric. Assume now $1 < p, q < \infty$.

We use Young's inequality: for $a, b \ge 0$,

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}. \tag{2.5}$$

(See Appendix B.5 for a proof.)

If $\|X\|_p = 0$ or $\|Y\|_q = 0$ the claim is trivial. Otherwise set

$$U := \frac{|X|}{\|X\|_p}, \qquad V := \frac{|Y|}{\|Y\|_q},$$

so that $\mathbb{E}[U^p] = \mathbb{E}[V^q] = 1$. Applying (2.5) pointwise to (U, V) and taking expectations,

$$\mathbb{E}[UV] \le \frac{1}{p}\mathbb{E}[U^p] + \frac{1}{q}\mathbb{E}[V^q] = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying by $\|X\|_p \|Y\|_q$ gives

$$\mathbb{E}|XY| = \|X\|_p \|Y\|_q \, \mathbb{E}[UV] \le \|X\|_p \|Y\|_q.$$

Equality case. For $1 < p, q < \infty$, equality holds iff $U^p = V^q$ a.s., i.e. $|X|^p = c\,|Y|^q$ a.s. for some $c > 0$ (and the signs of X, Y are aligned). □

Example 2.15 (Cauchy-Schwarz Inequality). The special case $p = q = 2$ is the Cauchy-Schwarz inequality.

$$\mathbb{E}(|XY|) \le (\mathbb{E}(X^2)\mathbb{E}(Y^2))^{1/2} \tag{2.6}$$

Example 2.16 (Lyapunov's Inequality, or $L^p \subset L^q$ on a probability space if $p > q$). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $1 \le q < p < \infty$. Apply Hölder to $|X|^q \cdot 1$ with exponents $\frac{p}{q}$ and $\frac{p}{p-q}$:

$$\mathbb{E}|X|^q = \mathbb{E}\big(|X|^q \cdot 1\big) \le \big(\mathbb{E}(|X|^{q \cdot p/q})\big)^{q/p} \big(\mathbb{E}(1^{p/(p-q)})\big)^{(p-q)/p} = (\mathbb{E}|X|^p)^{q/p}.$$

Hence

$$\|X\|_q \le \|X\|_p. \tag{2.7}$$

Remark: "$L^p \subset L^q$ for $p > q$" is in general not true on the spaces that have infinite measure.

Theorem 2.17 (Minkowski's Inequality (triangle inequality in L^p))**.** *Let $1 \leq p \leq \infty$. If $X, Y \in L^p$, then $X + Y \in L^p$ and*

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

This is inequality is a consequence of Hölder's inequality. You are asked to prove it as an exercise. (See hints in Appendix.)

Example 2.18 (A simple check of Minkowski with cancellation). Let ε be a Rademacher random variable, i.e. $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$. Take

$$X := 1, \qquad Y := \varepsilon.$$

Then for any $1 \leq p < \infty$,

$$\|X\|_p = (\mathbb{E}|1|^p)^{1/p} = 1, \qquad \|Y\|_p = (\mathbb{E}|\varepsilon|^p)^{1/p} = 1.$$

Moreover,

$$\|X + Y\|_p = \left(\mathbb{E}|1 + \varepsilon|^p\right)^{1/p} = \left(\tfrac{1}{2} \cdot |2|^p + \tfrac{1}{2} \cdot |0|^p\right)^{1/p} = \left(\tfrac{1}{2}\, 2^p\right)^{1/p} = 2^{1-1/p}.$$

Thus Minkowski reads

$$\|X + Y\|_p = 2^{1-1/p} \leq 2 = \|X\|_p + \|Y\|_p,$$

and the inequality is strict for $1 < p < \infty$ because of cancellation when $\varepsilon = -1$.

2.5 Exercises

Homework 2

Problem 2.19 (Differentiation Under the Integral Sign)**.** Let $f : \mathbb{R} \times [a, b] \to \mathbb{R}$ satisfy the following conditions:

(i) For each $t \in [a, b]$, the function $x \mapsto f(x, t)$ is integrable with respect to Lebesgue measure on $\mathbb{R}$.

(ii) For almost every $x \in \mathbb{R}$, the partial derivative $\frac{\partial f}{\partial t}(x, t)$ exists for all $t \in [a, b]$.

(iii) There exists $g \in L^1(\mathbb{R})$ such that $\left|\frac{\partial f}{\partial t}(x, t)\right| \leq g(x)$ for all $t \in [a, b]$ and a.e. $x \in \mathbb{R}$.

Define $F(t) = \int_{\mathbb{R}} f(x, t)\, dx$ for $t \in [a, b]$.

(a) Prove that F is differentiable on $[a, b]$ and

$$F'(t) = \int_{\mathbb{R}} \frac{\partial f}{\partial t}(x, t)\, dx.$$

(b) Apply this result to compute

$$\frac{d}{dt}\int_0^\infty e^{-tx}\sin(x)\,dx \quad \text{for } t>0,$$

and use your answer to evaluate $\int_0^\infty \frac{\sin x}{x}\,dx$.

Problem 2.20 (Expectation via Tail Probabilities)**.** Let $X \geq 0$ be a nonnegative random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

(a) Using MCT (or Fubini-Tonelli), prove that

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\,dt.$$

(b) Let X be a nonnegative **integer-valued** random variable. Prove that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

(c) Suppose $X \geq 0$ and $p > 0$. Prove that

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

(d) Apply part (c) to show: if $\mathbb{E}[X^p] < \infty$ for some $p > 0$, then $\mathbb{P}(X > t) = o(t^{-p})$ as $t \to \infty$.

Problem 2.21 (Reverse Fatou's Lemma)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space.

(a) **(Reverse Fatou's Lemma)** Let $f_n : \Omega \to \mathbb{R}$ be measurable functions. Suppose there exists $g \in L^1(\mu)$ such that $f_n \leq g$ a.e. for all n. Prove that

$$\limsup_{n\to\infty} \int_\Omega f_n\,d\mu \leq \int_\Omega \limsup_{n\to\infty} f_n\,d\mu.$$

(b) Give an example showing that the inequality can be strict.

(c) Suppose $f_n \to f$ a.e. and there exist $g, h \in L^1(\mu)$ such that $g \leq f_n \leq h$ a.e. for all n. Prove that $\int f_n\,d\mu \to \int f\,d\mu$.

Problem 2.22 (Scheffé's Lemma – Convergence of Densities)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f_n, f \in L^1(\mu)$ with $f_n \geq 0$, $f \geq 0$.

(a) Prove that if $f_n \to f$ a.e. and $\int_\Omega f_n \, d\mu \to \int_\Omega f \, d\mu$, then

$$\int_\Omega |f_n - f| \, d\mu \to 0.$$

This is **Scheffé's Lemma**.

Hint: Consider $(f_n - f)^+ = \max(f_n - f, 0)$ and $(f_n - f)^- = \max(f - f_n, 0)$ separately. Use Fatou's lemma.

(b) Give an example showing that the hypothesis $\int f_n \to \int f$ cannot be dropped, even when $f_n \to f$ a.e.

(c) **Application:** Let $X_1, X_2, \ldots$ be random variables with densities f_n with respect to Lebesgue measure, and let X have density f. Suppose $f_n(x) \to f(x)$ for a.e. x. Prove that $X_n \xrightarrow{d} X$ (convergence in distribution).

Note: Scheffé's lemma says something much stronger when $\int f_n = \int f = 1$.

Additional Practice Exercises

Problem 2.23 (Computing Limits via DCT)**.** *These are the types of integrals you should be able to handle after mastering DCT.*
Evaluate the following limits, justifying each step:

(a) $\displaystyle\lim_{n\to\infty} \int_0^1 \frac{nx^{n-1}}{1+x} \, dx$

(b) $\displaystyle\lim_{n\to\infty} \int_0^\infty \frac{n}{1+n^2x^2} \, dx$

(c) $\displaystyle\lim_{n\to\infty} \int_0^n \left(1 - \frac{x}{n}\right)^n e^{x/2} \, dx$

Problem 2.24 (Interchanging sum and integral)**.** *This result is used constantly—it justifies term-by-term integration of series.*
Prove that if $g_m \geq 0$ then

$$\int \sum_{m=0}^{\infty} g_m \, d\mu = \sum_{m=0}^{\infty} \int g_m \, d\mu. \tag{2.8}$$

Problem 2.25 (L^1 Convergence Does Not Imply A.E. Convergence)**.** *This counterexample (the "typewriter sequence") is a classic that every probabilist should know.*
Give an example of a sequence $f_n \in L^1([0,1])$ such that $\|f_n\|_1 \to 0$ but $f_n(x) \not\to 0$ for **any** $x \in [0,1]$.

Problem 2.26 (DCT with convergence in measure)**.** *This extension of DCT is essential for applications in probability theory.*
Let $X_n \to X$ in probability (or in measure) and assume there exists $Y \in L^1$ such that $|X_n| \le Y$ a.s. for all n. Prove that

$$\mathbb{E}[X_n] \to \mathbb{E}[X].$$

Problem 2.27 (Continuity of the integral over sets)**.** *A direct application of MCT that appears frequently in measure theory arguments.*
Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f \ge 0$ be measurable.

1. If $A_n \uparrow A$, prove $\int_{A_n} f \, d\mu \uparrow \int_A f \, d\mu$.
2. If $A_n \downarrow A$ and $\int_{A_1} f \, d\mu < \infty$, prove $\int_{A_n} f \, d\mu \downarrow \int_A f \, d\mu$.

Problem 2.28 (Truncation and L^1-approximation)**.** *The truncation technique is a workhorse for reducing problems about general L^1 functions to bounded ones.*
Let $X \in L^1(\mathbb{P})$ and define $X_n = X\mathbf{1}_{\{|X| \le n\}}$.

1. Show $X_n \to X$ a.s. and $|X_n| \le |X|$.
2. Use DCT to prove $\mathbb{E}|X_n - X| \to 0$.
3. Conclude: for every $\varepsilon > 0$ there exists a bounded random variable Z such that $\mathbb{E}|X - Z| < \varepsilon$.

Problem 2.29 (Uniform Integrability Criterion)**.** *This equivalent characterization of UI is often easier to verify in practice.*
Let (X_n) be a sequence of random variables on a probability space. Prove that the following are equivalent:

(i) $\{X_n\}$ is uniformly integrable.

(ii) $\sup_n \mathbb{E}[|X_n|] < \infty$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\mathbb{P}(A) < \delta \Rightarrow \sup_n \mathbb{E}[|X_n|\mathbf{1}_A] < \varepsilon$.

Problem 2.30 (Moments and Tail Decay)**.** *Understanding the relationship between moments and tail behavior is fundamental in probability.*
Let $X \ge 0$ be a random variable.

(a) Prove that if $\mathbb{P}(X > t) \le Ce^{-\lambda t}$ for all $t \ge 0$ (exponential tail), then $\mathbb{E}[X^p] < \infty$ for all $p > 0$.

(b) Prove that if $\mathbb{P}(X > t) \le Ct^{-\alpha}$ for all $t \ge 1$ (polynomial tail with $\alpha > 0$), then $\mathbb{E}[X^p] < \infty$ for all $p < \alpha$.

(c) Give an example showing that $\mathbb{E}[X^p] < \infty$ does **not** imply $\mathbb{P}(X > t) \leq Ct^{-q}$ for any $q > p$ (i.e., the $o(t^{-p})$ in Problem 2.20(d) cannot be improved to $O(t^{-(p+\varepsilon)})$ in general).

Problem 2.31 (Null sets have zero integral)**.** *A basic fact that underlies the "a.e." flexibility in measure theory.*
Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $A \in \mathcal{F}$ with $\mu(A) = 0$.

(a) Show that for every measurable $f \geq 0$,

$$\int_A f \, d\mu = 0.$$

(b) Deduce that if $f \in L^1(\mu)$, then $\int_A f \, d\mu = 0$.

Problem 2.32 (Jensen's inequality: quick corollaries)**.** *Two immediate consequences that are used constantly.*
Assume Jensen's inequality. Prove:

1. $|\mathbb{E}X| \leq \mathbb{E}|X|$ for $X \in L^1$.
2. If $X \in L^p$ with $p \geq 1$, then $\|X\|_1 \leq \|X\|_p$ on a probability space.

Problem 2.33 (No L^p inclusion on infinite measure spaces)**.** *This shows that the inclusion $L^p \subset L^q$ for $p > q$ relies crucially on finite measure.*
On $(\mathbb{R}, \lambda)$ let $1 \leq q < p < \infty$.

1. Give an example of $f \in L^p(\mathbb{R})$ but $f \notin L^q(\mathbb{R})$.
2. Give an example of $g \in L^q(\mathbb{R})$ but $g \notin L^p(\mathbb{R})$.

Problem 2.34 (A basic DCT success)**.** *A simple example to build intuition before tackling harder ones.*
Let $f_n(x) = x^n$ on $[0, 1]$ with Lebesgue measure.

1. Find $f(x) = \lim_{n\to\infty} f_n(x)$ for all $x \in [0, 1]$.
2. Use DCT to justify $\int_0^1 f_n(x) \, dx \to \int_0^1 f(x) \, dx$.
3. Compute $\int_0^1 x^n \, dx$ explicitly and check your conclusion.

Problem 2.35 (n^α-spike family)**.** *Explores how the parameter α controls whether DCT applies.*
On $([0, 1], \mathcal{B}, \lambda)$ define $f_n(x) = n^\alpha \mathbf{1}_{[0,1/n]}(x)$.

1. Show $f_n(x) \to 0$ for all $x \in (0, 1]$.

2. Compute $\int_0^1 f_n(x)\,dx$ and determine for which α it tends to 0.
3. Compute $g(x) = \sup_n f_n(x)$ and determine for which α the function g is integrable.

Problem 2.36 (Minkowski's inequality from Hölder's)**.** Assume Hölder's inequality is known. Let $p \geq 1$ and $X, Y \in L^p$. Prove Minkowski's inequality:

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

Problem 2.37 ($L^p \subset L^q$ on finite measure spaces (Lyapunov inequality))**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with $\mu(\Omega) < \infty$. Let $1 \leq q < p \leq \infty$ and $f \in L^p(\mu)$. Prove that $f \in L^q(\mu)$ and

$$\|f\|_{L^q(\mu)} \leq \mu(\Omega)^{\frac{1}{q}-\frac{1}{p}} \|f\|_{L^p(\mu)}.$$

Problem 2.38 (Strict Inequality in Fatou)**.** Construct an explicit sequence (f_n) of nonnegative measurable functions on $[0, 1]$ such that:

(i) $f_n(x) \to f(x)$ for all $x \in [0, 1]$,

(ii) $\int_0^1 \liminf_n f_n\,dx < \liminf_n \int_0^1 f_n\,dx$,

(iii) Both sides of the inequality are finite and positive.

Problem 2.39 (A.E. Convergence Does Not Imply L^1 Convergence)**.** Give an example of a sequence $f_n \in L^1([0, 1])$ such that $f_n \to 0$ a.e. but $\|f_n\|_1 = \int |f_n|\,dx \not\to 0$.

Problem 2.40 (Rare-event spike: DCT failure)**.** On $([0, 1], \mathcal{B}, \lambda)$ define $X_n(x) = n\,\mathbf{1}_{(0,1/n)}(x)$.

1. Show that $X_n \to 0$ a.e.
2. Compute $\mathbb{E}[X_n]$.
3. Show directly that domination fails, that is, that there is no $Y \in L^1$ such that $|X_n| \leq Y$ a.e. for all n.
4. (Optional) Show $\{X_n\}$ is not uniformly integrable by proving that for every $K > 0$,
$$\sup_{n\geq 1} \mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] = 1.$$

Problem 2.41 (Uniform integrability of a specific family)**.** Let $t_n = \sum_{k=1}^{n} 4^{-k}$ and $A_n = (t_{n-1}, t_n]$. Define $f_n(x) = \frac{4^n}{n}\mathbf{1}_{A_n}(x)$ on $(0,1)$. Show that $\{f_n : n \geq 1\}$ is uniformly integrable.

Problem 2.42 (Convergence of integrals without an L^1 dominator)**.** Let $(0,1)$ carry Lebesgue measure λ. Define

$$t_0 := 0, \qquad t_n := \sum_{k=1}^{n} 2^{-2k} = \sum_{k=1}^{n} 4^{-k} = \frac{1}{3}\bigl(1 - 4^{-n}\bigr),$$

and set the disjoint intervals $A_n := (t_{n-1}, t_n] \subset (0,1)$, so that $\lambda(A_n) = 4^{-n}$. Define $f_n(x) := \frac{4^n}{n}\,\mathbf{1}_{A_n}(x)$.

1. Show that $f_n(x) \to 0$ for every $x \in (0,1)$.
2. Show that $\int_0^1 f_n(x)\,dx \to 0$.
3. Show that no integrable function g dominates all f_n.

Chapter 3

Conditional Expectation

In elementary probability, the conditional probability of an event A given an event B is defined by $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, but this formula breaks down when $\mathbb{P}(B) = 0$—for instance, when conditioning on $\{X = x\}$ for a continuous random variable X. Kolmogorov's insight was to replace conditioning on events with conditioning on σ-algebras, defining the conditional expectation $\mathbb{E}[X \mid \mathcal{F}]$ as a random variable characterized by a projection-like property. This chapter develops the theory of conditional expectation: its existence via the Radon–Nikodym theorem, its key algebraic properties (linearity, tower law, pull-out, conditional Jensen), and its interpretation as the L^2-projection onto the subspace of $\mathcal{F}$-measurable random variables. We also introduce regular conditional distributions, which recover the intuitive "conditional density" picture within the rigorous framework.

3.1 Definition of conditional expectation

We present the definition of conditional expectation due to Kolmogorov (1933).

Definition 3.1 (Conditional expectation)**.** Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space, let $\mathcal{F} \subset \mathcal{F}_0$ be a sub-σ-field, and let $X \in L^1(\mathcal{F}_0)$. A random variable Z is called a **conditional expectation of X given $\mathcal{F}$**, written $Z = \mathbb{E}[X \mid \mathcal{F}]$, if

(i) Z is $\mathcal{F}$-measurable and $Z \in L^1(\mathcal{F})$;

(ii) for every $A \in \mathcal{F}$,
$$\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X\,\mathbf{1}_A]\,.$$

Any such Z is called a **version** of $\mathbb{E}[X \mid \mathcal{F}]$. Moreover, if Z_1 and Z_2 are two versions, then $Z_1 = Z_2$ $\mathbb{P}$-a.s.

The question immediately arises whether such random variable Z exists. The proof of the existence relies on the Radon-Nikodym theorem.

Definition 3.2 (Absolute continuity)**.** Let μ, ν be measures on the same measurable space $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** with respect to μ, and write $\nu \ll \mu$, if for every $A \in \mathcal{F}$,

$$\mu(A) = 0 \implies \nu(A) = 0.$$

Definition 3.3 (σ-finiteness)**.** A measure μ on $(\Omega, \mathcal{F})$ is **σ-finite** if $\Omega = \bigcup_{n\geq 1} A_n$ with $A_n \in \mathcal{F}$ and $\mu(A_n) < \infty$ for all n.

Note that every probability measure is finite, hence σ-finite.

Theorem 3.4 (Radon–Nikodym)**.** *Let μ and ν be σ-finite measures on $(\Omega, \mathcal{F})$ with $\nu \ll \mu$. Then there exists an $\mathcal{F}$-measurable function $f : \Omega \to [0, \infty]$ such that*

$$\nu(A) = \int_A f \, d\mu \qquad \text{for all } A \in \mathcal{F}.$$

The function f is unique μ-a.e. and is denoted by $f = \frac{d\nu}{d\mu}$.

A sketch of a proof is in Appendix C.1.

Theorem 3.5 (Existence of conditional expectation via Radon–Nikodym)**.** *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space, let $\mathcal{F} \subset \mathcal{F}_0$, and let $X \in L^1(\mathcal{F}_0)$. Then there exists an $\mathcal{F}$-measurable random variable $Z \in L^1(\mathcal{F})$ such that*

$$\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X\,\mathbf{1}_A] \qquad \text{for all } A \in \mathcal{F}.$$

Moreover, Z is unique up to $\mathbb{P}$-a.s. equality. We denote any such Z by $Z = \mathbb{E}[X \mid \mathcal{F}]$.

In other words,

$$\mathbb{E}(X|\mathcal{F}) = \frac{d\nu}{d(\mathbb{P}|_{\mathcal{F}})}, \text{ where } \nu(A) = \int_A X \, d\mathbb{P},$$

for $A \in \mathcal{F}$. So conditional expectation is literally a Radon–Nicodym derivative.

Proof. First assume $X \geq 0$. Define a set function $\nu : \mathcal{F} \to [0, \infty)$ by

$$\nu(A) := \mathbb{E}[X\,\mathbf{1}_A] = \int_A X \, d\mathbb{P}, \qquad A \in \mathcal{F}.$$

Then ν is a finite measure on $(\Omega, \mathcal{F})$: indeed, $\nu(\varnothing) = 0$ and for pairwise disjoint $(A_n) \subset \mathcal{F}$,

$$\begin{aligned}\nu\Big(\bigsqcup_{n\geq 1} A_n\Big) &= \int X\,\mathbf{1}_{\sqcup_n A_n}\, d\mathbb{P} = \int X \sum_{n\geq 1} \mathbf{1}_{A_n}\, d\mathbb{P} = \sum_{n\geq 1} \int X\,\mathbf{1}_{A_n}\, d\mathbb{P} \\ &= \sum_{n\geq 1} \nu(A_n),\end{aligned}$$

where we used monotone convergence (since $\sum_{n\geq 1} \mathbf{1}_{A_n}$ increases to $\mathbf{1}_{\cup_n A_n}$ and $X \geq 0$). Also $\nu \ll \mathbb{P}|_{\mathcal{F}}$: if $\mathbb{P}(A) = 0$ then $\nu(A) = \int_A X \, d\mathbb{P} = 0$.

By the Radon–Nikodym theorem, there exists an $\mathcal{F}$-measurable function $Z \geq 0$ such that $\nu(A) = \int_A Z \, d\mathbb{P}$ for all $A \in \mathcal{F}$. In particular,

$$\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[X\,\mathbf{1}_A] \qquad \forall\, A \in \mathcal{F}.$$

Moreover, taking $A = \Omega$ gives $\mathbb{E}[Z] = \nu(\Omega) = \mathbb{E}[X] < \infty$, so $Z \in L^1$.

Now let $X \in L^1$ be arbitrary. Write $X = X^+ - X^-$ with $X^\pm \geq 0$ and $\mathbb{E}[X^+] + \mathbb{E}[X^-] < \infty$. Apply the previous construction to X^+ and X^- to obtain $\mathcal{F}$-measurable $Z^+, Z^- \in L^1$ with $\mathbb{E}[Z^\pm \mathbf{1}_A] = \mathbb{E}[X^\pm \mathbf{1}_A]$ for all $A \in \mathcal{F}$. Set $Z := Z^+ - Z^-$. Then Z is $\mathcal{F}$-measurable, $Z \in L^1$, and for all $A \in \mathcal{F}$,

$$\mathbb{E}[Z\,\mathbf{1}_A] = \mathbb{E}[Z^+\,\mathbf{1}_A] - \mathbb{E}[Z^-\,\mathbf{1}_A] = \mathbb{E}[X^+\,\mathbf{1}_A] - \mathbb{E}[X^-\,\mathbf{1}_A] = \mathbb{E}[X\,\mathbf{1}_A].$$

Uniqueness: if Z_1, Z_2 are $\mathcal{F}$-measurable integrable random variables satisfying the displayed identity, then for $D := Z_1 - Z_2$ we have $\mathbb{E}[D\,\mathbf{1}_A] = 0$ for all $A \in \mathcal{F}$. Taking $A = \{D > 0\} \in \mathcal{F}$ yields $\mathbb{E}[D\,\mathbf{1}_{\{D>0\}}] = 0$, hence $D\,\mathbf{1}_{\{D>0\}} = 0$ a.s., so $\mathbb{P}(D > 0) = 0$. Similarly $\mathbb{P}(D < 0) = 0$. Therefore $D = 0$ a.s., i.e. $Z_1 = Z_2$ a.s. □

3.2 Properties

Lemma 3.6 (Equivalent characterization)**.** *Let $X \in L^1(\mathcal{F}_0)$ and let $\mathcal{F} \subset \mathcal{F}_0$. A random variable Z is a version of $\mathbb{E}[X \mid \mathcal{F}]$ iff Z is $\mathcal{F}$-measurable, $Z \in L^1$, and*

$$\mathbb{E}[Y\,Z] = \mathbb{E}[Y\,X] \quad \textit{for every bounded } \boldsymbol{\mathcal{F}}\textbf{-measurable } \textit{random variable } Y.$$

Equivalently, it suffices to require the identity for $Y = \mathbf{1}_A$ with $A \in \mathcal{F}$.

(See a proof in Appendix C.2.)

Theorem 3.7 (Tower property)**.** *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space, and let $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}_0$ be sub-σ-fields. If $X \in L^1(\mathcal{F}_0)$, then*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}\big] = \mathbb{E}[X \mid \mathcal{H}] \qquad \textit{a.s.}$$

(See Exercise 3.31.)

Corollary 3.8 (Taking expectations)**.** *For $X \in L^1$, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.*

Proof. Apply Theorem 3.7 with $\mathcal{H} = \{\varnothing, \Omega\}$. □

Corollary 3.9 (Conditioning on a larger σ-field does nothing)**.** *If X is $\mathcal{H}$-measurable and $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{H}] = X$ and $\mathbb{E}[X \mid \mathcal{G}] = X$ a.s.*

Lemma 3.10 (Pull-out (taking out what is known))**.** *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}_0$. If $X \in L^1(\mathcal{F}_0)$ and Y is bounded and $\mathcal{G}$-measurable, then*

$$\mathbb{E}[Y\,X \mid \mathcal{G}] = Y\,\mathbb{E}[X \mid \mathcal{G}] \qquad \textit{a.s.}$$

More generally, if Y is $\mathcal{G}$-measurable and $Y\,X \in L^1$, the same identity holds.

(See Exercise 3.36.)

Theorem 3.11 (Conditional Jensen inequality)**.** *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and $\mathcal{G} \subset \mathcal{F}_0$. Let X be an integrable real-valued random variable and let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be convex and Borel measurable. Assume either $\varphi(X) \geq 0$ a.s. or $\varphi(X) \in L^1$. Then*

$$\varphi\big(\mathbb{E}[X \mid \mathcal{G}]\big) \;\leq\; \mathbb{E}\big[\varphi(X) \mid \mathcal{G}\big] \qquad \text{a.s.}$$

(See a proof in Appendix C.2.)

Corollary 3.12 (L^p-contraction)**.** *Let $1 \leq p < \infty$ and $X \in L^p(\mathcal{F}_0)$. Then $\mathbb{E}[X \mid \mathcal{G}] \in L^p(\mathcal{G})$ and*

$$\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p.$$

Moreover, the operator $T : X \mapsto \mathbb{E}[X \mid \mathcal{G}]$ is 1-Lipschitz on L^p:

$$\|\mathbb{E}[X \mid \mathcal{G}] - \mathbb{E}[Y \mid \mathcal{G}]\|_p \leq \|X - Y\|_p.$$

For $p = \infty$, $\|\mathbb{E}[X \mid \mathcal{G}]\|_\infty \leq \|X\|_\infty$.

(See Exercise 3.37.)

3.3 L^2-projection characterization of conditional expectation

Theorem 3.13. *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}_0$ be a sub-σ-field. If $X \in L^2(\mathcal{F}_0)$ and $Z := \mathbb{E}[X \mid \mathcal{G}]$, then $Z \in L^2(\mathcal{G})$ and*

$$\langle X - Z,\, Y\rangle_{L^2} \equiv \mathbb{E}[(X - Z)Y] = 0 \qquad \text{for all } Y \in L^2(\mathcal{G}).$$

Consequently, Z is the orthogonal projection of X onto the closed subspace $L^2(\mathcal{G}) \subset L^2(\mathcal{F}_0)$. Equivalently,

$$\mathbb{E}\big[(X - Z)^2\big] \;=\; \min_{Y \in L^2(\mathcal{G})} \mathbb{E}\big[(X - Y)^2\big],$$

and Z is the unique minimizer (up to $\mathbb{P}$-a.s. equality).

So, $Z = \mathbb{E}[X \mid \mathcal{G}]$ is the unique $\mathcal{G}$-measurable r.v. that minimizes mean square error.

Proof. **Step 1: $Z \in L^2(\mathcal{G})$.** Since $Z = \mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, it remains to show $Z \in L^2$. By Corollary 3.12 $\mathbb{E}[Z^2] \leq \mathbb{E}[X^2] < \infty$, hence $Z \in L^2(\mathcal{G})$.

Step 2: Orthogonality against bounded test functions. Let Y be bounded and $\mathcal{G}$-measurable. By Lemma 3.6,

$$\mathbb{E}[YZ] = \mathbb{E}[YX].$$

Therefore $\mathbb{E}[(X - Z)Y] = 0$ for all bounded $\mathcal{G}$-measurable Y.

Step 3: Extend to all $Y \in L^2(\mathcal{G})$ by truncation. Fix $Y \in L^2(\mathcal{G})$. Define the truncations

$$Y_n := (-n) \vee (Y \wedge n), \qquad n \geq 1.$$

Then each Y_n is bounded and $\mathcal{G}$-measurable, so Step 2 gives $\mathbb{E}[(X-Z)Y_n] = 0$ for all n. Moreover, $Y_n \to Y$ in L^2 (since $|Y_n - Y| \leq |Y|$ and $|Y_n - Y| \to 0$ pointwise, one checks $\mathbb{E}[(Y_n - Y)^2] \to 0$). Using Cauchy–Schwarz and $X - Z \in L^2$,

$$\big|\mathbb{E}[(X-Z)(Y_n - Y)]\big| \leq \|X - Z\|_2 \, \|Y_n - Y\|_2 \xrightarrow[n\to\infty]{} 0.$$

Hence

$$\mathbb{E}[(X-Z)Y] = \lim_{n\to\infty} \mathbb{E}[(X-Z)Y_n] = 0.$$

Thus $X - Z \perp L^2(\mathcal{G})$.

Step 4: Best approximation / projection. For any $Y \in L^2(\mathcal{G})$ we have the orthogonal decomposition

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

with $(X-Z) \perp (Z-Y)$ since $Z - Y \in L^2(\mathcal{G})$. Therefore,

$$\|X - Y\|_2^2 = \|X - Z\|_2^2 + \|Z - Y\|_2^2 \geq \|X - Z\|_2^2,$$

so Z minimizes $\mathbb{E}[(X-Y)^2]$ over $Y \in L^2(\mathcal{G})$. If Y is another minimizer then $\|Z - Y\|_2^2 = 0$, hence $Y = Z$ a.s. □

Remark 3.14. The set $L^2(\mathcal{G})$ is indeed a closed subspace of $L^2(\mathcal{F}_0)$ (limits of $\mathcal{G}$-measurable functions are $\mathcal{G}$-measurable), so abstract Hilbert space theory guarantees the existence of orthogonal projections. Our proof instead verified the projection property directly from the Radon–Nikodym construction.

Gaussian Hilbert space.

Centered Gaussian random variables form a Hilbert space with inner product $\langle X, Y \rangle = \mathrm{Cov}(X, Y)$. For jointly Gaussian variables, the conditional expectation $\mathbb{E}[X \mid Y]$ equals the orthogonal projection of X onto the one-dimensional subspace spanned by Y in this Hilbert space.

This is stronger than what Theorem 3.13 guarantees: a priori, $\mathbb{E}[X \mid \sigma(Y)]$ is the projection onto all of $L^2(\sigma(Y))$, which includes nonlinear functions of Y. The simplification occurs because for jointly Gaussian variables, uncorrelatedness implies independence: the residual $X - \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}\mathrm{ar}(Y)} Y$ is uncorrelated with Y, hence independent of Y, hence orthogonal to **every** function of Y in L^2.

Example 3.15 (Jointly Gaussian: conditional expectation is affine). Let (X, Y) be jointly Gaussian with $\mathbb{V}\mathrm{ar}(Y) > 0$ and $X \in L^2$. Then

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}\mathrm{ar}(Y)} \big(Y - \mathbb{E}[Y]\big) \qquad \text{a.s.}$$

In particular, if $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, then $\mathbb{E}[X \mid Y] = \frac{\mathrm{Cov}(X,Y)}{\mathbb{V}\mathrm{ar}(Y)} Y$.

(See Exercise 3.29.)

Summary of key properties

- Tower: $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$ for $\mathcal{H} \subset \mathcal{G}$
- Pull-out: $\mathbb{E}[YX \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$ if Y is $\mathcal{G}$-measurable
- Contraction: $\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p$
- Projection: $\mathbb{E}[X \mid \mathcal{G}]$ minimizes $\mathbb{E}[(X-Y)^2]$ over $Y \in L^2(\mathcal{G})$

3.4 Regular Conditional Probabilities

We defined the conditional **expectations**. How do we compute them in practice? We need something like **conditional probability distributions** which we used in earlier courses to calculate the conditional expectations.

The following theorem is rather general but you can assume for simplicity that $S = T = \mathbb{R}$ and $\mathcal{S} = \mathcal{T} = \mathcal{B}$, the Borel σ-algebra. If you want to know what is a standard Borel space, see Definition A.17 in Appendix.

Theorem 3.16 (Regular conditional distribution (existence on standard Borel spaces))**.** *Let $X : \Omega \to S$ and $Y : \Omega \to T$ be random variables, where $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are measurable spaces. Assume that $(S, \mathcal{S})$ is a **standard Borel space** (in particular, any Polish space with its Borel σ-field, e.g. $S = \mathbb{R}^d$ with $\mathcal{S} = \mathcal{B}(\mathbb{R}^d)$). Then there exists a map*

$$K : T \times \mathcal{S} \to [0, 1],$$

*called a **regular conditional distribution** of X given Y, such that:*

*(i) (**Kernel property**) For each $y \in T$, the set function $B \mapsto K(y, B)$ is a probability measure on $(S, \mathcal{S})$.*

*(ii) (**Measurability in** y) For each $B \in \mathcal{S}$, the map $y \mapsto K(y, B)$ is $\mathcal{T}$-measurable.*

*(iii) (**Correct conditioning**) For each $B \in \mathcal{S}$, the random variable $K(Y, B)$ is a version of $\mathbb{P}(X \in B \mid \sigma(Y))$, i.e.*

$$\mathbb{E}\big[\mathbf{1}_{\{X \in B\}} \mathbf{1}_A\big] = \mathbb{E}\big[K(Y, B)\, \mathbf{1}_A\big] \qquad \text{for all } A \in \sigma(Y).$$

Equivalently, for all $A \in \mathcal{T}$ and $B \in \mathcal{S}$,

$$\mathbb{P}(Y \in A,\ X \in B) = \int_A K(y, B)\, \mathbb{P}_Y(dy).$$

We may write informally $K(y, \cdot) = \mathbb{P}(X \in \cdot \mid Y = y)$.

Note that the formula $\mathbb{P}(Y \in A, X \in B) = \int_A K(y, B)\, \mathbb{P}_Y(dy)$ is an instance of Fubini's theorem applied to the product measure. See Appendix C.3 for a reminder.

How to think about this theorem

1. **Why this is not automatic from conditional expectation.** For each fixed $B \in \mathcal{S}$, the random variable $\mathbb{P}(X \in B \mid \sigma(Y))$ is $\sigma(Y)$-measurable, hence equals $g_B(Y)$ a.s. for some measurable g_B. The theorem says that one can choose these g_B **simultaneously for all** B in a way that $B \mapsto K(y, B)$ is a genuine probability measure for (almost) every y and $y \mapsto K(y, B)$ is measurable. This is the nontrivial content.

2. **What "standard Borel" buys you.** On nice state spaces (Borel subsets of Polish spaces), probability measures are determined by countable data and measurable selection works well; this prevents pathologies and guarantees existence. For completely arbitrary measurable spaces, such a measurable kernel need not exist.

3. **Versions / "$Y = y$" is only symbolic.** The kernel $K(y, \cdot)$ is only determined for $\mathbb{P}_Y$-a.e. y. Changing $K(y, \cdot)$ on a set of y with $\mathbb{P}_Y$-measure 0 does not affect property (iii). So the notation $\mathbb{P}(\cdot \mid Y = y)$ should be read as "a chosen measurable version".

4. **How you actually use it.** Once K exists, conditional expectations can be written as integrals against conditional laws: for any bounded measurable $f : S \to \mathbb{R}$,

$$\mathbb{E}[f(X) \mid \sigma(Y)] \;=\; \int_S f(x)\, K(Y, dx) \qquad \text{a.s.}$$

   (And similarly for $f \geq 0$ or $f \in L^1$ with the usual integrability assumptions.) This is the form needed for computations and for constructing $\mathbb{P}(X \in \cdot \mid Y)$ in examples.

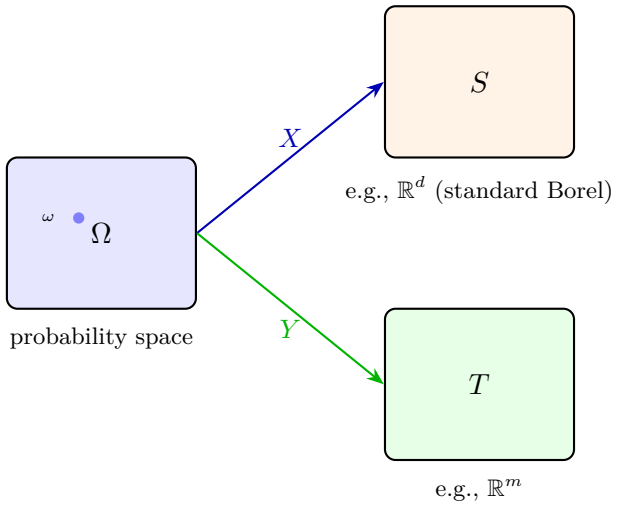
5. **Sanity check: the density case on $\mathbb{R}^2$.** If (X, Y) has a joint density $f_{X,Y}(x, y)$ and $f_Y(y) > 0$, then a version is

$$K(y, B) = \int_B \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx,$$

   which matches the familiar conditional density formula.

Regular Conditional Distribution: A Visual Guide

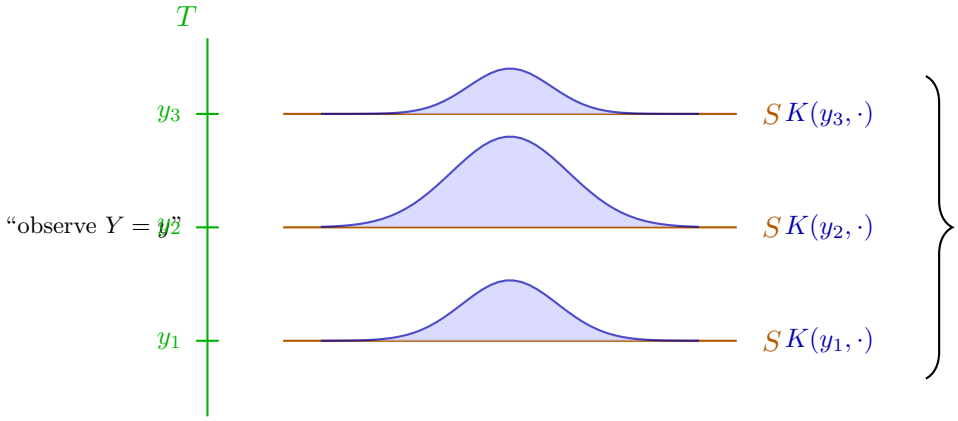
The Setup



Random variables: $X(\omega) \in S$ and $Y(\omega) \in T$ are determined by the outcome $\omega \in \Omega$.

The Kernel $K : T \times \mathcal{S} \to [0, 1]$

The kernel gives a probability measure on S for each $y \in T$



The Three Properties

(i) Kernel property

Fix $y \in T$.

$B \mapsto K(y, B)$ is a probability measure on $(S, \mathcal{S})$.

(ii) Measurability in y

Fix $B \in \mathcal{S}$.

$y \mapsto K(y, B)$ is $\mathcal{T}$-measurable.

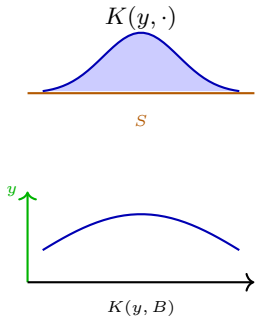
"Probabilities vary measurably with y."

(iii) Correct conditioning

For each $B \in \mathcal{S}$:

$K(Y, B) = \mathbb{P}(X \in B \mid \sigma(Y))$ a.s.

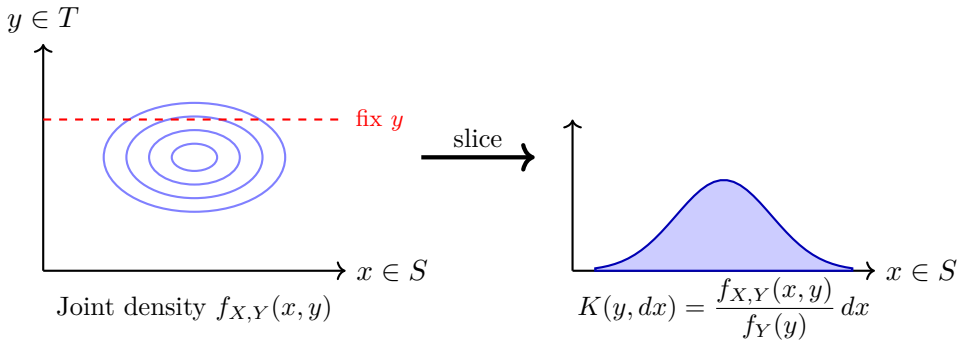
"It's actually the conditional distribution."



$$\mathbb{P}(Y \in A,\, X \in B) = \int_A K(y, B)\, \mathbb{P}_Y(dy)$$

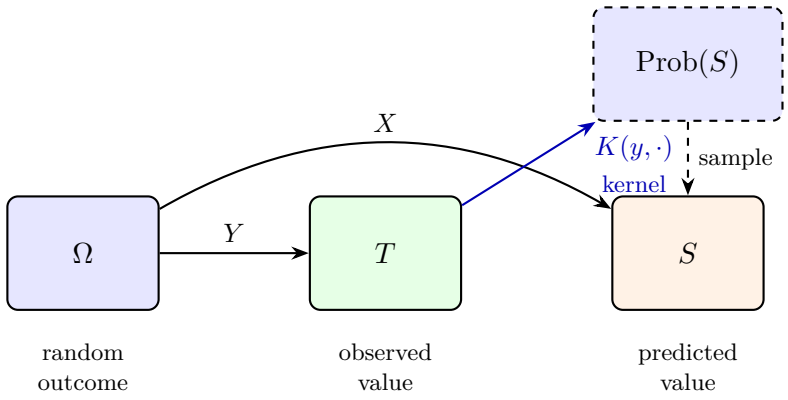
The Density Case (Sanity Check)

If (X, Y) has joint density $f_{X,Y}(x, y)$ with marginal $f_Y(y) > 0$:



Key insight: The kernel $K(y, \cdot)$ "slices" the joint distribution at each value of y and renormalizes to get a probability measure on S.

Summary Diagram



Interpretation: Once we observe $Y = y$, the kernel $K(y, \cdot)$ tells us the distribution of X. We write $K(y, \cdot) = \mathbb{P}(X \in \cdot \mid Y = y)$ informally.

3.5 Examples and warnings

Example 3.17 (Conditioning on a finite partition). Let $\mathcal{G} = \sigma(A_1, \ldots, A_n)$ where $A_1, \ldots, A_n \in \mathcal{F}_0$ form a partition of Ω and $\mathbb{P}(A_i) > 0$ for each i. For $X \in L^1$,

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i=1}^{n} \Big(\frac{\mathbb{E}[X \mathbf{1}_{A_i}]}{\mathbb{P}(A_i)} \Big) \mathbf{1}_{A_i}.$$

(See Exercise 3.32.)

Example 3.18 (Conditioning on a discrete random variable). Let Y take countably many values $(y_k)_{k \geq 1}$ with $\mathbb{P}(Y = y_k) > 0$. Then for $X \in L^1$,

$$\mathbb{E}[X \mid \sigma(Y)] = \sum_{k \geq 1} \mathbb{E}[X \mid Y = y_k] \, \mathbf{1}_{\{Y = y_k\}}, \qquad \mathbb{E}[X \mid Y = y_k] = \frac{\mathbb{E}[X \mathbf{1}_{\{Y = y_k\}}]}{\mathbb{P}(Y = y_k)}.$$

(See Exercise 3.38.)

Example 3.19 (Regular conditional probabilities on $\mathbb{R}$: conditional CDF viewpoint). Let X be $\mathbb{R}$-valued and let $Y : \Omega \to T$ be any random variable. By Theorem 3.16, there exists a kernel $K(y, \cdot)$ on $\mathbb{R}$ such that $K(Y, B) = \mathbb{P}(X \in B \mid \sigma(Y))$ a.s. for all Borel B. Define the **conditional distribution function**

$$F(y, t) := K\big(y, (-\infty, t]\big).$$

Then for each fixed t, the map $y \mapsto F(y, t)$ is measurable, and for $\mathbb{P}_Y$-a.e. y, $t \mapsto F(y, t)$ is a (right-continuous, nondecreasing) distribution function.

(See Exercise 3.39.)

Two warnings

1. **Versions matter.** Conditional expectations and regular conditional laws are only determined up to $\mathbb{P}$-a.s. equality (or $\mathbb{P}_Y$-a.e. equality in the parameter y). Changing a version on a null set produces the same conditional expectation / conditional law.

2. **Why we assumed "standard Borel".** For arbitrary measurable spaces $(S, \mathcal{S})$, a regular conditional distribution $K(y, \cdot)$ need not exist. On standard Borel spaces (in particular $\mathbb{R}^d$ with Borel sets), it does exist.

3.6 Exercises

Homework

Exercise 3.20 (Conditioning on the maximum). Let X, Y be i.i.d. Uniform$[0, 1]$ and let $M = \max(X, Y)$. Compute $\mathbb{E}[X \mid M]$.

Exercise 3.21 (Conditional expectation given a coarsening). Let N be uniform on $\{1, 2, 3, 4, 5, 6\}$ (a fair die) and let $R = N \mod 3$, taking values in $\{0, 1, 2\}$.

(a) Write down the partition of $\Omega = \{1, \dots, 6\}$ induced by $\sigma(R)$.

(b) Compute $\mathbb{E}[N \mid R]$.

(c) Verify the tower property $\mathbb{E}[\mathbb{E}[N \mid R]] = \mathbb{E}[N]$ using your answer.

Exercise 3.22 (Law of total variance). Let $X \in L^2$ and let $\mathcal{G} \subset \mathcal{F}_0$. Prove

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[\mathbb{V}\mathrm{ar}(X \mid \mathcal{G})] + \mathbb{V}\mathrm{ar}(\mathbb{E}[X \mid \mathcal{G}]),$$

where $\mathbb{V}\mathrm{ar}(X \mid \mathcal{G}) := \mathbb{E}[X^2 \mid \mathcal{G}] - (\mathbb{E}[X \mid \mathcal{G}])^2$ is the **conditional variance**.

Exercise 3.23 (Best linear predictor vs. conditional expectation). Let $X, Y \in L^2$ with $\mathbb{V}\mathrm{ar}(Y) > 0$.

(a) Show that the **best linear predictor** of X given Y, defined as the minimizer of

$$\min_{a,b \in \mathbb{R}} \mathbb{E}[(X - a - bY)^2],$$

is given by

$$\hat{X}_{\mathrm{lin}} = \mathbb{E}[X] + \frac{\mathrm{Cov}(X, Y)}{\mathbb{V}\mathrm{ar}(Y)}(Y - \mathbb{E}[Y]).$$

(b) Prove that $\mathbb{E}[(X - \hat{X}_{\mathrm{lin}})^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2]$, with equality if and only if $\mathbb{E}[X \mid Y] = \hat{X}_{\mathrm{lin}}$ a.s.

(c) Give an example where the inequality is strict.

(d) Explain why, for jointly Gaussian (X, Y), we always have equality.

Additional exercises

The following exercises are recommended for deeper understanding. They are ordered roughly by importance for later material in the course.

Exercise 3.24 (Conditional monotone convergence theorem). Let $0 \le X_1 \le X_2 \le \cdots$ be random variables with $X_n \uparrow X$ pointwise. Prove that

$$\mathbb{E}[X_n \mid \mathcal{G}] \uparrow \mathbb{E}[X \mid \mathcal{G}] \quad \text{a.s.}$$

(The limit may be $+\infty$.)

Exercise 3.25 (Conditional Fatou's lemma). Let $(X_n)_{n\ge 1}$ be nonnegative random variables. Prove

$$\mathbb{E}[\liminf_{n\to\infty} X_n \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}] \quad \text{a.s.}$$

Exercise 3.26 (Conditional dominated convergence theorem). Let $X_n \to X$ a.s. and suppose there exists $Y \in L^1$ with $|X_n| \le Y$ for all n. Prove

$$\mathbb{E}[X_n \mid \mathcal{G}] \to \mathbb{E}[X \mid \mathcal{G}] \quad \text{a.s. and in } L^1.$$

Exercise 3.27 (Independence criterion for conditional expectation). Let $X \in L^1$ and let $\mathcal{G} \subset \mathcal{F}_0$ be a sub-σ-field. Prove that X is independent of $\mathcal{G}$ if and only if $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ a.s.

Note: This exercise requires familiarity with independence and the monotone class theorem.

Exercise 3.28 (Conditioning on a sum of exponentials). Let X and Y be independent random variables, each with the standard exponential distribution (density e^{-t} for $t > 0$). Compute the regular conditional distribution $K(s, \cdot) = \mathbb{P}(X \in \cdot \mid X + Y = s)$.

Exercise 3.29 (Gaussian conditioning). Let (X, Y) be jointly Gaussian with $\mathbb{V}\mathrm{ar}(Y) > 0$. Prove that

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] + \frac{\mathrm{Cov}(X, Y)}{\mathbb{V}\mathrm{ar}(Y)}\big(Y - \mathbb{E}[Y]\big).$$

Exercise 3.30 (Conditional expectation and symmetry). Let $X_1, \ldots, X_n$ be i.i.d. random variables in L^1. Show that

$$\mathbb{E}[X_1 \mid X_1 + \cdots + X_n] = \frac{1}{n}(X_1 + \cdots + X_n).$$

Exercise 3.31 (Tower property via defining identity). Let $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}_0$ and $X \in L^1$. Prove directly from the defining property that

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}].$$

Exercise 3.32 (Finite partition formula). Let $\mathcal{G} = \sigma(A_1, \ldots, A_n)$ where (A_i) is a measurable partition of Ω with $\mathbb{P}(A_i) > 0$ for each i. Prove directly from the defining property that for $X \in L^1$,

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i=1}^{n} \frac{\mathbb{E}[X \mathbf{1}_{A_i}]}{\mathbb{P}(A_i)} \mathbf{1}_{A_i}.$$

Exercise 3.33 (Regular conditional distribution in the density case). Let (X, Y) have a joint density $f_{X,Y}$ on $\mathbb{R}^2$ with marginal density $f_Y(y) > 0$ for all y in the support. Define

$$K(y, B) = \int_B \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx, \qquad B \in \mathcal{B}(\mathbb{R}).$$

(a) Show that K is a probability kernel: for each y, $K(y, \cdot)$ is a probability measure, and for each B, $y \mapsto K(y, B)$ is measurable.

(b) Show that $K(Y, B)$ is a version of $\mathbb{P}(X \in B \mid \sigma(Y))$.

Exercise 3.34 (Closedness of $L^2(\mathcal{G})$). Prove that $L^2(\mathcal{G})$ is a closed subspace of $L^2(\mathcal{F}_0)$.

Exercise 3.35 (Pull-out requires measurability). Let $\Omega = \{1, 2, 3, 4\}$ with uniform probability $\mathbb{P}(\{\omega\}) = 1/4$ for each ω. Define $X(\omega) = \omega$ and $\mathcal{G} = \sigma(\{1, 2\}, \{3, 4\})$.

(a) Compute $\mathbb{E}[X \mid \mathcal{G}]$.

(b) Find a random variable Y that is **not** $\mathcal{G}$-measurable such that $\mathbb{E}[YX \mid \mathcal{G}] \neq Y \cdot \mathbb{E}[X \mid \mathcal{G}]$.

(c) Explain why this does not contradict the pull-out lemma.

Exercise 3.36 (Pull-out property). Let $X \in L^1$ and let Y be bounded and $\mathcal{G}$-measurable. Prove $\mathbb{E}[YX \mid \mathcal{G}] = Y\, \mathbb{E}[X \mid \mathcal{G}]$.

Exercise 3.37 (Conditional Jensen $\Rightarrow$ L^p contraction). Assume conditional Jensen for convex φ. Deduce that for $1 \leq p < \infty$ and $X \in L^p$,

$$\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p, \qquad \|\mathbb{E}[X \mid \mathcal{G}] - \mathbb{E}[Y \mid \mathcal{G}]\|_p \leq \|X - Y\|_p.$$

(Optionally: do $p = \infty$ separately.)

Exercise 3.38 (Conditioning on a discrete r.v.). Let $X \in L^1$ and Y take countably many values $(y_k)_{k \geq 1}$ with $\mathbb{P}(Y = y_k) > 0$. Prove Example 3.18, using the formula in Example 3.17 or directly from the definition.

Exercise 3.39 (Regular conditional law on $\mathbb{R}$). Let X be $\mathbb{R}$-valued. Using Theorem 3.16, define $F(y, t) := \mathbb{P}(X \leq t \mid Y = y)$ (as a chosen version). Show that for each fixed t, $y \mapsto F(y, t)$ is measurable, and for $\mathbb{P}_Y$-a.e. y, $t \mapsto F(y, t)$ is a distribution function.

Chapter 4

Independence and Tails

Independence is the structural assumption that drives most of the deep theorems in probability. When random variables are independent, their joint behavior is completely determined by their marginals, and this factorization unlocks powerful tools: extending independence from generators via the π-λ theorem, splitting sums into manageable pieces, and applying multiplicative moment bounds. This chapter introduces independence of σ-algebras and random variables, then develops the theory of tail events—the Kolmogorov and Hewitt–Savage zero-one laws, and the Borel–Cantelli lemmas—which reveal a striking rigidity: for independent sequences, many natural events have probability either zero or one, with no room for intermediate values.

4.1 Independence

Independence via σ-algebras

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 4.1 (Independence of σ-algebras)**.** Two sub-σ-algebras $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ are called **independent** if

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\,\mathbb{P}(H) \qquad \text{for all } G \in \mathcal{G},\ H \in \mathcal{H}.$$

More generally, a family $(\mathcal{G}_i)_{i \in I}$ of sub-σ-algebras is called **independent** if for every finite choice of distinct indices $i_1, \dots, i_k \in I$ and events $G_{i_j} \in \mathcal{G}_{i_j}$,

$$\mathbb{P}\Big(\bigcap_{j=1}^{k} G_{i_j}\Big) = \prod_{j=1}^{k} \mathbb{P}(G_{i_j}).$$

Definition 4.2 (Independence of random variables)**.** Random variables X, Y are **independent** if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent, i.e.

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad \mathbb{P}\big(X \in A,\ Y \in B\big) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

for all $A, B \in \mathcal{B}(\mathbb{R})$. More generally, $(X_i)_{i\in I}$ are independent if $(\sigma(X_i))_{i\in I}$ are independent.

Remark 4.3. The existence of infinite independent sequences with prescribed marginals is guaranteed by Kolmogorov's extension theorem (Theorem 1.38).

π–λ extension lemma for independence

Recall the following definitions.

Definition 4.4 (π-system and λ-system)**.** A class $\mathcal{P} \subset \mathcal{F}$ is a **π-system** if it is closed under finite intersections: $A, B \in \mathcal{P} \Rightarrow A \cap B \in \mathcal{P}$.

A class $\mathcal{L} \subset \mathcal{F}$ is a **λ-system** (Dynkin system) if:

1. $\Omega \in \mathcal{L}$;
2. $A \in \mathcal{L} \Rightarrow A^c \in \mathcal{L}$;
3. if (A_n) are disjoint and $A_n \in \mathcal{L}$, then $\bigsqcup_{n\geq 1} A_n \in \mathcal{L}$.

Remark 4.5. An equivalent system axioms for λ-system is

1. $\Omega \in \mathcal{L}$;
2. if $A, B \in \mathcal{L}$ and $A \subset B$, then $B \setminus A \in \mathcal{L}$;
3. if $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{L}$, then $\bigcup_{n\geq 1} A_n \in \mathcal{L}$.)

Lemma 4.6 (π–λ theorem)**.** *If $\mathcal{P}$ is a π-system, then the smallest λ-system containing $\mathcal{P}$ equals $\sigma(\mathcal{P})$.*

Lemma 4.7 (Independence from generators)**.** *Let $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ be π-systems. Assume that*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \qquad \text{for all } A \in \mathcal{A},\ B \in \mathcal{B}.$$

Then the σ-algebras $\sigma(\mathcal{A})$ and $\sigma(\mathcal{B})$ are independent.

See Exercise 4.28. Use π–λ theorem.

4.2 Tail σ-fields (σ-algebras)

Now that we have the tools to verify independence of σ-algebras, we turn to one of its most striking consequences: events determined by the asymptotic behavior of an independent sequence must have probability 0 or 1.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(X_n)_{n\geq 1}$ be a sequence of random variables.

Definition 4.8 (Tail σ-field)**.** For $m \geq 1$ define the "future" σ-field

$$\mathcal{F}_m^\infty := \sigma(X_m, X_{m+1}, \dots).$$

The **tail σ-field** (or **tail σ-algebra**) of (X_n) is

$$\mathcal{T} := \bigcap_{m=1}^{\infty} \mathcal{F}_m^\infty = \bigcap_{m=1}^{\infty} \sigma(X_m, X_{m+1}, \dots).$$

An event $A \in \mathcal{F}$ is called a **tail event** if $A \in \mathcal{T}$.

Remark 4.9 (Intuition / invariance under finite changes). If $A \in \mathcal{T}$, then membership of ω in A cannot depend on any finite initial segment $(X_1(\omega), \dots, X_{m-1}(\omega))$ for any m; informally, changing finitely many coordinates cannot change whether $\omega \in A$.

Example 4.10 (Canonical tail event: infinitely often). Given events $A_n \in \sigma(X_n)$, define

$$A_n \text{ i.o. } := \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n.$$

Then A_n i.o. $\in \mathcal{T}$ (it depends only on the tail of the sequence).

Example 4.11 (Tail random variable). Let $S_n := \sum_{k=1}^n X_k$ and define

$$L := \limsup_{n \to \infty} \frac{S_n}{n}.$$

Then L is $\mathcal{T}$-measurable: modifying finitely many X_k changes S_n/n by $O(1/n)$, hence does not affect the lim sup.

4.3 Kolmogorov's 0–1 law

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(X_n)_{n \geq 1}$ be a sequence of **independent** random variables. For $m \geq 1$ write

$$\mathcal{F}_1^m := \sigma(X_1, \dots, X_m), \qquad \mathcal{F}_m^\infty := \sigma(X_m, X_{m+1}, \dots), \qquad \mathcal{T} := \bigcap_{m \geq 1} \mathcal{F}_m^\infty.$$

Theorem 4.12 (Kolmogorov's 0–1 law)**.** *If $(X_n)_{n \geq 1}$ are independent, then the tail σ-field $\mathcal{T}$ is **trivial**: for every $A \in \mathcal{T}$,*

$$\mathbb{P}(A) \in \{0, 1\}.$$

Equivalently, every tail event has probability 0 *or* 1.

Proof. Fix $m \geq 1$. Since (X_n) are independent, the σ-algebras $\mathcal{F}_1^m = \sigma(X_1, \dots, X_m)$ and $\mathcal{F}_{m+1}^\infty = \sigma(X_{m+1}, X_{m+2}, \dots)$ are independent. Since $\mathcal{T} \subset \mathcal{F}_{m+1}^\infty$, it follows that $\mathcal{T}$ is independent of $\mathcal{F}_1^m$. Hence for every $A \in \mathcal{T}$ and every $B \in \mathcal{F}_1^m$,

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

Now let

$$\mathcal{F}_{\text{fin}} := \bigvee_{m\geq 1} \mathcal{F}_1^m = \sigma(X_1, X_2, \dots),$$

the σ-algebra generated by all finite initial segments (equivalently, by the entire sequence). By monotone class / π–λ extension (applied in B for fixed A), the identity above extends to all $B \in \mathcal{F}_{\text{fin}}$, so A is independent of $\mathcal{F}_{\text{fin}}$.

In particular, taking $B = A$ (note that $A \in \mathcal{T} \subset \mathcal{F}_{\text{fin}}$), we get

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2,$$

so $\mathbb{P}(A) \in \{0, 1\}$. □

Remark 4.13 (A handy corollary). If Z is $\mathcal{T}$-measurable (a **tail random variable**), then Z is a.s. constant. Indeed, for any $t \in \mathbb{R}$ the event $\{Z \leq t\} \in \mathcal{T}$, hence has probability 0 or 1, which forces the distribution function of Z to be a step function.

Example 4.14. Let $X_1, X_2, \dots$ be independent and $S_n = X_1 + \dots + X_n$. Then $Z_1 = \liminf(S_n/n)$ and $Z_2 = \limsup(S_n/n)$ are almost surely constant (but possibly infinite).

4.4 Borel–Cantelli lemmas

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(A_n)_{n\geq 1}$ be events in $\mathcal{F}$. Recall the event "infinitely often":

$$A_n \text{ i.o. } := \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n,$$

i.e. $\omega \in A_n$ i.o. iff $\omega \in A_n$ for infinitely many n.

Lemma 4.15 (Borel–Cantelli I)**.** *If*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(A_n \text{ i.o.}) = 0.$$

Proof. For each $N \geq 1$,

$$\mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) \leq \sum_{n\geq N} \mathbb{P}(A_n) \xrightarrow[N\to\infty]{} 0$$

by the union bound and the Cauchy criterion for series. Since A_n i.o. $= \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n$, by continuity from above,

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{N\to\infty} \mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) = 0.$$

□

Lemma 4.16 (Borel–Cantelli II (under independence))**.** *Assume that the events $(A_n)_{n\geq 1}$ are independent and*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty.$$

Then

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

Proof. Fix $N \geq 1$. Since the events are independent,

$$\mathbb{P}\Big(\bigcap_{n\geq N} A_n^c\Big) = \lim_{M\to\infty} \mathbb{P}\Big(\bigcap_{n=N}^{M} A_n^c\Big) = \lim_{M\to\infty} \prod_{n=N}^{M} \big(1-\mathbb{P}(A_n)\big),$$

where we used continuity from above and finite independence.

Using $1 - x \leq e^{-x}$ for $x \in [0,1]$, we get

$$\prod_{n=N}^{M} \big(1-\mathbb{P}(A_n)\big) \leq \exp\Big(-\sum_{n=N}^{M} \mathbb{P}(A_n)\Big).$$

Since $\sum_{n\geq 1} \mathbb{P}(A_n) = \infty$, we have $\sum_{n=N}^{M} \mathbb{P}(A_n) \to \infty$ as $M \to \infty$, so the right-hand side tends to 0. Hence

$$\mathbb{P}\Big(\bigcap_{n\geq N} A_n^c\Big) = 0.$$

Taking complements,

$$\mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) = 1 \qquad \text{for every } N \geq 1.$$

Finally,

$$A_n \text{ i.o.} = \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n,$$

and by continuity from above again,

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{N\to\infty} \mathbb{P}\Big(\bigcup_{n\geq N} A_n\Big) = \lim_{N\to\infty} 1 = 1.$$

□

Remark 4.17 (How the 0–1 law interacts). If (A_n) are independent and $A_n \in \sigma(X_n)$ for independent (X_n), then A_n i.o. is a tail event, so Kolmogorov's 0–1 law guarantees $\mathbb{P}(A_n \text{ i.o.}) \in \{0,1\}$. Borel–Cantelli II identifies which of the two occurs when $\sum_n \mathbb{P}(A_n) = \infty$.

Example 4.18 (The assumption of independence in BCL(2) is essential). Consider the space $\Omega = (0,1]$ with the σ-algebra $\mathcal{B}$ of Borel subsets, and the Lebesgue measure as $\mathbb{P}$. The events $A_n = (0, 1/n] \in \mathcal{B}$. Then, $\mathbb{P}(A_n) = 1/n$, $\sum \mathbb{P}(A_n) = \infty$, but $\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(\emptyset) = 0$.

4.5 Hewitt–Savage 0–1 law

Let $(X_n)_{n\ge 1}$ be i.i.d. random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathfrak{S}_{\text{fin}}$ be the group of permutations $\pi : \mathbb{N} \to \mathbb{N}$ that move only finitely many indices (i.e. $\pi(n) = n$ for all but finitely many n). For $\pi \in \mathfrak{S}_{\text{fin}}$ define the permuted sequence

$$(X_n^\pi)_{n\ge 1} := (X_{\pi(n)})_{n\ge 1}.$$

Define the **exchangeable (symmetric) σ-field**

$$\mathcal{I} := \big\{ A \in \sigma(X_1, X_2, \dots) : \ \mathbf{1}_A(X_1, X_2, \dots) = \mathbf{1}_A(X_1^\pi, X_2^\pi, \dots) \ \big\}.$$

a.s. for all $\pi \in \mathfrak{S}_{\text{fin}}$. Equivalently, $A \in \mathcal{I}$ iff A is invariant (a.s.) under every finite permutation of coordinates.

Remark 4.19 (Notation). When we write $\mathbf{1}_A(X_1, X_2, \dots)$, we identify $A \in \sigma(X_1, X_2, \dots) \in \mathcal{F}$ with a Borel set $B \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$ via $A = \{\omega : (X_1(\omega), X_2(\omega), \dots) \in B\}$. The expression $\mathbf{1}_A(X_1, X_2, \dots)$ then means $\mathbf{1}_B(X_1, X_2, \dots)$, and permuting the arguments makes sense:

$$\mathbf{1}_A(X_{\pi(1)}, X_{\pi(2)}, \dots) := \mathbf{1}_B(X_{\pi(1)}, X_{\pi(2)}, \dots).$$

The invariance condition $\mathbf{1}_A(X_1, X_2, \dots) = \mathbf{1}_A(X_{\pi(1)}, X_{\pi(2)}, \dots)$ a.s. thus says that the set $B \subset \mathbb{R}^{\mathbb{N}}$ is invariant under finite permutations of coordinates.

Theorem 4.20 (Hewitt–Savage 0–1 law)**.** *If (X_n) are i.i.d., then $\mathcal{I}$ is trivial: for every $A \in \mathcal{I}$,*

$$\mathbb{P}(A) \in \{0, 1\}.$$

Proof sketch. The idea is to show that every symmetric event A is independent of itself, hence $\mathbb{P}(A) = \mathbb{P}(A)^2$.

Step 1 (Approximation). Since $A \in \sigma(X_1, X_2, \dots)$, for every $\varepsilon > 0$ there exists n and an event $B \in \sigma(X_1, \dots, X_n)$ such that $\mathbb{P}(A \triangle B) < \varepsilon$.

Step 2 (Shift). Define B' by replacing each X_i with X_{n+i} in the definition of B, so $B' \in \sigma(X_{n+1}, \dots, X_{2n})$. Because the (X_n) are i.i.d., $\mathbb{P}(B') = \mathbb{P}(B)$. Moreover, the finite permutation $(1, \dots, 2n) \mapsto (n{+}1, \dots, 2n, 1, \dots, n)$ sends B to B' while preserving A (by symmetry), so $\mathbb{P}(A \triangle B') < \varepsilon$ as well.

Step 3 (Independence). Since $B \in \sigma(X_1, \dots, X_n)$ and $B' \in \sigma(X_{n+1}, \dots, X_{2n})$ are independent,

$$\mathbb{P}(A)^2 \approx \mathbb{P}(B)\,\mathbb{P}(B') = \mathbb{P}(B \cap B') \approx \mathbb{P}(A \cap A) = \mathbb{P}(A),$$

where each "$\approx$" incurs an error at most ε. As $\varepsilon \to 0$, we obtain $\mathbb{P}(A) = \mathbb{P}(A)^2$, so $\mathbb{P}(A) \in \{0, 1\}$. □

Corollary 4.21 (Permutation-invariant random variables are a.s. constant)**.** *If $T = T(X_1, X_2, \dots)$ is measurable and satisfies*

$$T(X_1, X_2, \dots) = T(X_1^\pi, X_2^\pi, \dots) \quad \text{a.s. for all } \pi \in \mathfrak{S}_{\text{fin}},$$

then T is a.s. constant.

Remark 4.22 (Hewitt–Savage as infinite-dimensional concentration).

The Hewitt–Savage 0–1 law can be understood through the lens of concentration of measure in high dimensions.

Finite dimensions: the CLT window.

Let $X_1, X_2, \dots$ be i.i.d.Bernoulli (1/2) and $S_n = \sum_{k=1}^n X_k$. In $\{0,1\}^n$, a permutation-invariant event depends only on S_n. Consider, for example,

$$A_n = \big\{S_n \in [n/2 - \sqrt{n},\ n/2 + \sqrt{n}] \cap \mathbb{Z}\big\}.$$

By the central limit theorem, $\mathbb{P}(A_n) \to \Phi(2) - \Phi(-2) \approx 0.954$, a fixed value outside $\{0,1\}$. More generally, thresholds at scale $\sqrt{n}$ around $n/2$ produce any limiting probability in $(0,1)$.

Infinite dimensions: no CLT window. In contrast, a genuine event of the infinite sequence that is invariant under all finite permutations of $\mathbb{N}$ — such as $\{\limsup_n S_n/n \in [a,b]\}$ — must have probability in $\{0,1\}$. The CLT window that produces intermediate probabilities for finite-dimensional symmetric events simply has no infinite-dimensional counterpart.

Connection to concentration of measure. This can be read as an **exact** version of the concentration of measure philosophy. On finite product spaces $(\Omega^n, \mu^{\otimes n})$, observables with small sensitivity to any single coordinate are **approximately** constant:

$$\mathbb{P}\big(|f - \mathbb{E}f| > \varepsilon\big) \le 2e^{-cn\varepsilon^2} \qquad \text{(McDiarmid/Azuma-type bounds).}$$

Thus "macroscopic" observables in high dimensions are nearly deterministic. Hewitt–Savage is the infinite-dimensional, exact-symmetry analogue:

concentration: many observables are **almost** constant

$$\longleftrightarrow$$

0–1 laws: **invariant** observables are **exactly** constant.

One should not overstate the analogy: for fixed n, even a fully permutation-invariant event can have probability 1/2, as the example above shows. What changes in the infinite product is that invariance under all finite permutations forces measurability with respect to $\mathcal{I}$, which is trivial under i.i.d. assumptions, yielding the strict dichotomy.

A symmetric but not tail event: existence of a global maximum

Let $(X_n)_{n\ge1}$ be i.i.d. bounded real-valued random variables . Write

$$M := \sup_{k\ge1} X_k \in (-\infty, \infty], \qquad E := \{\exists n \ge 1 :\ X_n = M\}.$$

So E is the event that the supremum of the infinite sample is **attained** by one of the coordinates (i.e. the sample has a global maximum).

Proposition 4.23. *The event E is invariant under every finite permutation of coordinates (hence $E \in \mathcal{I}$), but E is not a tail event (in general $E \notin \mathcal{T}$).*

Moreover, assume that the law of X_1 has no atom at its (finite or infinite) right endpoint in the sense that

$$\mathbb{P}(X_1 = x) = 0 \quad \textit{for every } x \textit{ with } \mathbb{P}(X_1 \le x) = 1.$$

Then

$$\mathbb{P}(E) = 0.$$

(If instead $\mathbb{P}(X_1 = b) > 0$ for a finite right endpoint b, then $\mathbb{P}(E) = 1$.)

Proof. **(Symmetry).** A finite permutation π does not change the multiset of values $\{X_n\}$, hence does not change $\sup_{k\ge1} X_k$ nor whether it is attained. Thus E is invariant under finite permutations.

(Not tail). Membership in E can be changed by modifying finitely many coordinates while keeping the entire tail fixed: e.g. fix a tail $(x_m, x_{m+1}, \dots)$ whose supremum is not attained by that tail (such tails exist as sequences of reals), and choose two different initial blocks $(x_1, \dots, x_{m-1})$ so that in one case the overall supremum is attained (put one of the early x_i equal to the overall supremum), and in the other case it is not. Hence E cannot be determined from $\sigma(X_m, X_{m+1}, \dots)$ for any m, so $E \notin \mathcal{T}$.

(Probability). Since the (X_n) are i.i.d., $\mathbb{P}(X_n = M)$ is the same for every n, so it suffices to show $\mathbb{P}(X_1 = M) = 0$.

Note that $\{X_1 = M\} = \{X_1 = x,\ X_m \le x\ \forall\, m \ge 2\}$ on $\{X_1 = x\}$. Because X_1 is independent of $(X_2, X_3, \dots)$,

$$\begin{aligned}\mathbb{P}\big(X_1 = M \,\big|\, X_1 = x\big) &= \mathbb{P}\big(\forall\, m \ge 2\colon X_m \le x\big) \\ &= \lim_{N\to\infty} \prod_{m=2}^{N} \mathbb{P}(X_m \le x) = \lim_{N\to\infty} F(x)^{\,N-1},\end{aligned}$$

where $F(x) := \mathbb{P}(X_1 \le x)$ and we used independence. If $F(x) < 1$, then $F(x)^{N-1} \to 0$; if $F(x) = 1$, then x is (at least) a right endpoint of the distribution, and by the assumption $\mathbb{P}(X_1 = x) = 0$. Therefore

$$\mathbb{P}(X_1 = M) = \int \mathbb{P}(X_1 = M \mid X_1 = x)\, d\mathbb{P}_{X_1}(x) = 0.$$

Finally,

$$\mathbb{P}(E) = \mathbb{P}\Big(\bigcup_{n\ge1} \{X_n = M\}\Big) \le \sum_{n\ge1} \mathbb{P}(X_n = M) = 0.$$

□

Remark 4.24. This is a good illustration of why Hewitt–Savage is different from Kolmogorov's tail 0–1 law: E is symmetric but not tail. Hewitt–Savage implies $\mathbb{P}(E) \in \{0, 1\}$, and the argument above identifies the correct value.

4.6 Tail triviality, mixing and ergodicity

For a stationary sequence (X_n) (i.e. a sequence with shift invariant finite-dimensional distributions), let T denote the left shift and define the shift-invariant σ-field $\mathcal{I}_T := \{A \in \mathcal{F} : T^{-1}A = A\}$. The sequence is called **ergodic** if $\mathcal{I}_T$ is trivial. It is called **strongly mixing** (α-mixing) if

$$\alpha(n) := \sup_{\substack{A\in\mathcal{F}_1^k,\, B\in\mathcal{F}_{k+n}^\infty \\ k\geq 1}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\,\mathbb{P}(B)| \;\longrightarrow\; 0 \quad \text{as } n\to\infty.$$

For stationary sequences,

$$\text{i.i.d.} \implies \text{strongly mixing} \implies \mathcal{T} \text{ trivial} \implies \text{ergodic},$$

and no implication reverses in general.

Strong mixing $\Rightarrow$ $\mathcal{T}$ trivial: Let $A \in \mathcal{T}$. Then $A \in \mathcal{F}_{k+n}^\infty$ for every n, so for any $B \in \mathcal{F}_1^k$,

$$|\mathbb{P}(A\cap B) - \mathbb{P}(A)\,\mathbb{P}(B)| \leq \alpha(n) \to 0.$$

Hence $\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$ for all $B \in \mathcal{F}_1^k$ and all k. By π–λ, A is independent of $\sigma(X_1, X_2, \ldots)$. Since $A \in \mathcal{T} \subset \sigma(X_1, X_2, \ldots)$, the event A is independent of itself, so $\mathbb{P}(A) \in \{0, 1\}$.

$\mathcal{T}$ trivial $\Rightarrow$ ergodic: If $A \in \mathcal{I}_T$, then $T^{-1}A = A$, so $A \in \sigma(X_2, X_3, \ldots) = \mathcal{F}_2^\infty$. Iterating, $A \in \mathcal{F}_m^\infty$ for every m (up to null sets), hence $A \in \mathcal{T}$.

Neither implication reverses: The 1-dependent sequence $X_n = Y_nY_{n+1}$ with (Y_n) i.i.d. has $\alpha(n) = 0$ for $n \geq 2$ (strongly mixing) but is not i.i.d. The deterministic Markov chain on $\{0, 1\}$ with $P(0, 1) = P(1, 0) = 1$, started from $\pi = (1/2, 1/2)$, is ergodic but has $\mathcal{T} = \sigma(X_1)$ non-trivial, since any single X_m determines the entire sequence.

4.7 Exercises

Homework

Problem 4.25 (Warm-up: identifying tail events)**.** Let $(X_n)_{n\geq 1}$ be **independent** random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $S_n := \sum_{k=1}^n X_k$. For each event below, determine (with a brief justification) whether it belongs to the tail σ-field $\mathcal{T} := \bigcap_{m\geq 1} \sigma(X_m, X_{m+1}, \ldots)$.

(a) $\Big\{ \sum_{n=1}^{\infty} X_n \text{ converges} \Big\}$.

(b) $\Big\{ \sup_{n\geq 1} X_n > 7 \Big\}$.

(c) $\{X_n \geq 0 \text{ i.o.}\}$.

(d) $\{X_1 + X_2 > 3\}$.

(e) $\Big\{ \limsup_{n\to\infty} \dfrac{S_n}{n} \le 1 \Big\}$.

Problem 4.26 (Convergence of a random series is a tail event)**.** Let $(X_n)_{n\ge1}$ be independent random variables. Using Kolmogorov's 0–1 law, show that the event

$$A := \Big\{ \sum_{n=1}^{\infty} X_n \text{ converges (to a finite limit)} \Big\}$$

satisfies $\mathbb{P}(A) \in \{0, 1\}$.

Problem 4.27 (Records in an i.i.d. sample)**.** Let $(X_n)_{n\ge1}$ be i.i.d. with a continuous distribution (i.e. the cdf F is continuous, so there are no ties a.s.). Say that X_n is a **record** if

$$X_n > \max(X_1, \dots, X_{n-1}) \qquad (X_1 \text{ is a record by convention}).$$

Define the record events $R_n := \{X_n \text{ is a record}\}$.

(a) Show that $\mathbb{P}(R_n) = \dfrac{1}{n}$.

(b) Show that for the event $E := \{R_n \text{ i.o.}\}$ (i.e. records occur infinitely often), $\mathbb{P}(E) \in \{0, 1\}$, and then show that $\mathbb{P}(E) = 1$.

Problem 4.28 (π–λ independence extension)**.** Let $\mathcal{A}, \mathcal{B} \subset \mathcal{F}$ be π-systems. Assume that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \qquad \forall\, A \in \mathcal{A},\ \forall\, B \in \mathcal{B}.$$

Prove that $\sigma(\mathcal{A})$ and $\sigma(\mathcal{B})$ are independent.

Problem 4.29 (Supremum of i.i.d. exponentials)**.** Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{P}(X_1 > x) = e^{-x}$ for $x \ge 0$ (i.e. $X_1 \sim \text{Exp}(1)$). Show that

$$\limsup_{n\to\infty} \frac{X_n}{\log n} = 1 \qquad \text{a.s.}$$

Additional Exercises

Chapter 5

Convergence Zoo and Uniform Integrability

Probability theory employs several distinct notions of convergence, each suited to different purposes: almost sure convergence captures pathwise behavior in the Strong Law of Large Numbers, convergence in probability captures approximation in a stochastic sense, L^p convergence captures moment behavior, and convergence in distribution underlies the Central Limit Theorem and asymptotic statistics. Knowing when one mode implies another—and when it does not—is essential for applications.

We begin by cataloging the four main modes of convergence and the implications among them, then develop a set of tools for working with weak convergence: the Continuous Mapping Theorem, Slutsky's theorem, and Skorokhod's representation theorem. A central theme is the role of uniform integrability as the bridge between convergence in probability and convergence in L^1.

5.1 Types of Convergence of Random Variables

Definition 5.1 (Almost sure convergence)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X_n, X be real-valued random variables. We say that X_n converges to X **almost surely** (a.s.) if

$$\mathbb{P}\big(\{\omega \in \Omega : \ X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}\big) = 1.$$

We write $X_n \to X$ a.s. or $X_n \xrightarrow{a.s.} X$.

Definition 5.2 (Convergence in probability)**.** We say that X_n converges to X **in probability** if for every $\varepsilon > 0$,

$$\mathbb{P}\big(|X_n - X| > \varepsilon\big) \longrightarrow 0 \qquad (n \to \infty).$$

We write $X_n \to X$ in probability, or $X_n \xrightarrow{\mathbb{P}} X$.

Definition 5.3 (L^p convergence)**.** Fix $p \in [1, \infty)$. We say that X_n converges to X **in** L^p if

$$\mathbb{E}\big[|X_n - X|^p\big] \longrightarrow 0 \qquad (n \to \infty).$$

We write $X_n \to X$ in L^p, or $X_n \xrightarrow{L^p} X$.

Definition 5.4 (Convergence in distribution / weak convergence)**.** We say that X_n converges to X **in distribution** (or **weakly**) if

$$\mathbb{E}[f(X_n)] \longrightarrow \mathbb{E}[f(X)] \qquad (n \to \infty)$$

for every **bounded continuous** function $f : \mathbb{R} \to \mathbb{R}$. We write $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$.

Remark 5.5 (Equivalent 1D CDF formulation). If $F_n(x) = \mathbb{P}(X_n \le x)$ and $F(x) = \mathbb{P}(X \le x)$, then $X_n \Rightarrow X$ is equivalent to

$$F_n(x) \longrightarrow F(x) \qquad \text{for all } x \in \mathbb{R} \text{ at which } F \text{ is continuous.}$$

Definition 5.6 (Weak convergence in $\mathbb{R}^d$)**.** Let X_n, X be $\mathbb{R}^d$-valued random vectors. We say that X_n converges to X **in distribution** if

$$\mathbb{E}[f(X_n)] \longrightarrow \mathbb{E}[f(X)]$$

for every bounded continuous function $f : \mathbb{R}^d \to \mathbb{R}$. We again write $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$.

Remark 5.7 (Properties in common for $\xrightarrow{\mathbb{P}}$, $\xrightarrow{a.s.}$, $\xrightarrow{L^p}$). The following hold for each of these three modes of convergence:

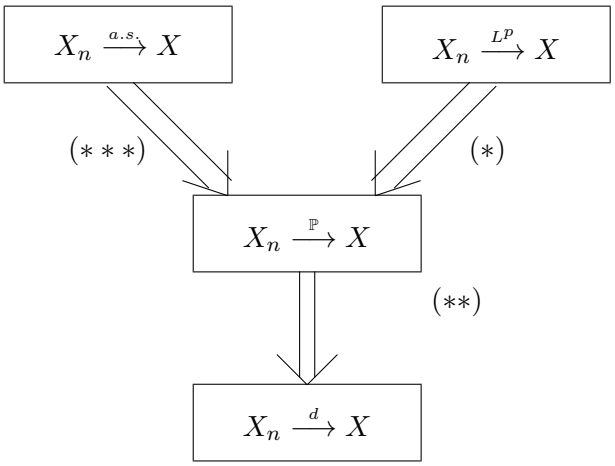
a) The limit is unique up to a.s. equivalence.

b) $X_n \to X \iff (X_n - X) \to 0$ (useful and common reduction).

c) $X_n \to X, Y_n \to Y \implies X_n + Y_n \to X + Y$. For $\xrightarrow{a.s.}$ and $\xrightarrow{\mathbb{P}}$, we also have $X_n Y_n \to XY$; for $\xrightarrow{L^p}$, the product rule requires additional integrability conditions (e.g. Hölder).

d) **Completeness.** The space L^p is complete (Riesz–Fischer theorem), and convergence in probability can be metrized by a complete metric (see text). For a.s. convergence, the situation is different: if $(X_n(\omega))$ is a Cauchy sequence in $\mathbb{R}$ for a.e. ω, then a pointwise limit exists a.e. by completeness of $\mathbb{R}$, but a.s. convergence itself is not metrizable by a single metric on the space of random variables, so "completeness" has a different character here.

Theorem 5.8. *The following property holds among the types of convergence.*



In addition, $X_n \xrightarrow{L^p} X$ *implies that* $X_n \xrightarrow{L^q} X$ *if* $p > q \geq 1$.

Proof. $(***)$ Assume $X_n \to X$ a.s. Fix $\varepsilon > 0$ and set

$$A := \{\omega \in \Omega : \ X_n(\omega) \to X(\omega)\}.$$

Then $\mathbb{P}(A) = 1$. For any $\omega \in A$, by the definition of pointwise convergence there exists $N = N(\omega, \varepsilon)$ such that for all $n \geq N$,

$$|X_n(\omega) - X(\omega)| \leq \varepsilon,$$

hence

$$\mathbf{1}_{\{|X_n(\omega)-X(\omega)|>\varepsilon\}} = 0 \qquad \text{for all } n \geq N.$$

Therefore $\mathbf{1}_{\{|X_n-X|>\varepsilon\}} \to 0$ almost surely.

Since $0 \leq \mathbf{1}_{\{|X_n-X|>\varepsilon\}} \leq 1$, the dominated convergence theorem yields

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}\big[\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] \longrightarrow 0.$$

This is exactly $X_n \to X$ in probability.

$(*)$ can be proved by Markov's inequality:

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\epsilon^p}$$

If the right-hand side converges to zero, then the left-hand side also converges to zero. For $p = \infty$ we need a separate argument; see Exercise 5.35.

$(**)$ Assume $X_n \xrightarrow{\mathbb{P}} X$.

We use the characterization of weak convergence by bounded continuous test functions: $X_n \Rightarrow X$ iff $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all $f \in C_b(\mathbb{R})$.

Fix $f \in C_b(\mathbb{R})$ and set $M := \|f\|_\infty < \infty$.

Step 1: $f(X_n) \xrightarrow{\mathbb{P}} f(X)$. Fix $\varepsilon > 0$. Choose $R > 0$ such that

$$\mathbb{P}(|X| > R) < \varepsilon.$$

Since f is continuous, it is uniformly continuous on the compact interval $[-(R+1),\, R+1]$, hence there exists $\delta \in (0,1)$ such that

$$|x-y|<\delta \ \text{ and } |x|\le R \implies |f(x)-f(y)|<\varepsilon.$$

Then

$$\begin{aligned}\mathbb{P}(|f(X_n)-f(X)|>\varepsilon) &\le \mathbb{P}(|X|>R)+\mathbb{P}\big(|X|\le R,\ |X_n-X|\ge\delta\big)\\ &\le \varepsilon+\mathbb{P}(|X_n-X|\ge\delta)\xrightarrow[n\to\infty]{}\varepsilon,\end{aligned}$$

because $X_n \xrightarrow{\mathbb{P}} X$. Since $\varepsilon>0$ is arbitrary, $\mathbb{P}(|f(X_n)-f(X)|>\varepsilon)\to 0$, i.e. $f(X_n)\xrightarrow{\mathbb{P}} f(X)$.

Step 2: $\mathbb{E}[f(X_n)]\to\mathbb{E}[f(X)]$. Let $Y_n := f(X_n)$ and $Y := f(X)$. Then $Y_n \xrightarrow{\mathbb{P}} Y$ and $|Y_n|\le M$, $|Y|\le M$. For any $\eta>0$,

$$\begin{aligned}\mathbb{E}|Y_n-Y| &= \mathbb{E}\big(|Y_n-Y|\mathbf{1}_{\{|Y_n-Y|>\eta\}}\big)+\mathbb{E}\big(|Y_n-Y|\mathbf{1}_{\{|Y_n-Y|\le\eta\}}\big)\\ &\le 2M\,\mathbb{P}(|Y_n-Y|>\eta)+\eta.\end{aligned}$$

Taking $n\to\infty$ gives $\limsup_{n\to\infty}\mathbb{E}|Y_n-Y|\le\eta$. Since $\eta>0$ is arbitrary, $\mathbb{E}|Y_n-Y|\to 0$, hence

$$|\mathbb{E}[Y_n]-\mathbb{E}[Y]|\le\mathbb{E}|Y_n-Y|\to 0,$$

i.e. $\mathbb{E}[f(X_n)]\to\mathbb{E}[f(X)]$.

Therefore $X_n \Rightarrow X$.

Finally, the claim that $X_n \xrightarrow{L^p} X$ implies that $X_n \xrightarrow{L^q} X$ if $p>q\ge 1$ follows from Lyapunov's inequality 2.7.

□

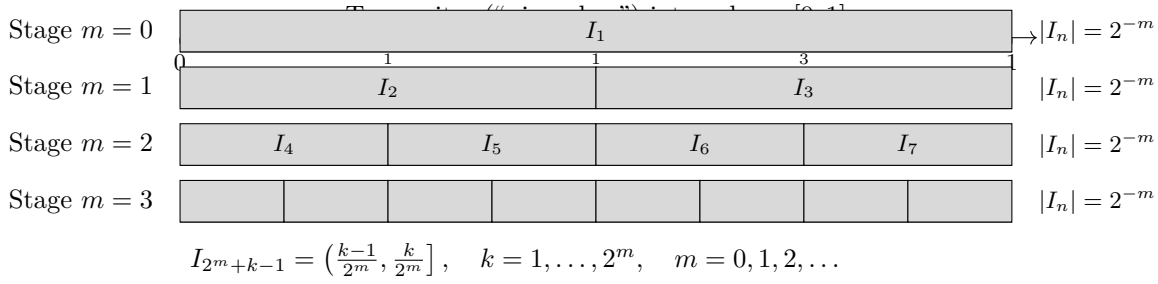


Figure 5.1: The typewriter (piano-key) enumeration of dyadic intervals.

Example 5.9 (Convergence in probability does not imply a.s. convergence (typewriter)). Let $\Omega=(0,1]$ with Lebesgue measure $\mathbb{P}$, and define

$$X_n(\omega) := \mathbf{1}_{I_n}(\omega),$$

where $(I_n)_{n\ge1}$ is the following enumeration of dyadic intervals: for each $m\ge 0$ and $k=1,\dots,2^m$, set

$$I_{2^m+k-1} := \Big(\frac{k-1}{2^m},\frac{k}{2^m}\Big].$$

1. Show that $X_n \to 0$ in probability.
2. Show that X_n does **not** converge to 0 almost surely (in fact, $\mathbb{P}(\limsup_{n\to\infty}\{X_n = 1\}) = 1$).

Remark 5.10 (L^p convergence does not imply a.s. convergence either). The typewriter sequence also converges in L^p for every $p \in [1,\infty)$: $\mathbb{E}|X_n|^p = \mathbb{P}(I_n) = 2^{-m} \to 0$. So this single example simultaneously shows that L^p convergence (and convergence in probability) need not imply almost sure convergence.

The typewriter example shows that convergence in probability does **not** guarantee almost sure convergence. But what partial converse can be salvaged? The answer—and one of the most useful technical tools in probability—is the **subsequence principle**: convergence in probability always yields almost sure convergence along a subsequence.

Proposition 5.11 (Subsequence characterization of convergence in probability)**.** *Let X_n, X be real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

(a) ***Subsequence extraction.*** *If $X_n \xrightarrow{\mathbb{P}} X$, then there exists a subsequence (n_k) such that $X_{n_k} \xrightarrow{a.s.} X$.*

(b) ***Full characterization.*** *$X_n \xrightarrow{\mathbb{P}} X$ if and only if every subsequence of (X_n) has a further subsequence converging to X almost surely.*

(c) ***Monotone upgrade.*** *If $X_n \xrightarrow{\mathbb{P}} X$ and the sequence (X_n) is a.s. monotone nondecreasing (i.e. $X_n \le X_{n+1}$ a.s. for all n), then $X_n \xrightarrow{a.s.} X$.*

Proof of (a). We construct (n_k) by a diagonal argument using Borel–Cantelli. For each $m \ge 1$, since $\mathbb{P}(|X_n - X| > 2^{-m}) \to 0$, we can choose n_m (strictly increasing in m) such that

$$\mathbb{P}\big(|X_{n_m} - X| > 2^{-m}\big) < 2^{-m}.$$

Set $A_m := \{|X_{n_m} - X| > 2^{-m}\}$. Then $\sum_{m=1}^{\infty} \mathbb{P}(A_m) < \infty$, so by the first Borel–Cantelli lemma, $\mathbb{P}(\limsup_m A_m) = 0$. Hence for a.e. ω, for all sufficiently large m we have $|X_{n_m}(\omega) - X(\omega)| \le 2^{-m}$, which gives $X_{n_m}(\omega) \to X(\omega)$. □

Proof of (b). ($\Rightarrow$) If $X_n \xrightarrow{\mathbb{P}} X$, then every subsequence (X_{n_j}) also converges to X in probability. Part (a) applied to this subsequence gives the desired further subsequence.

($\Leftarrow$) By contrapositive: suppose $X_n \not\xrightarrow{\mathbb{P}} X$. Then there exist $\varepsilon > 0$ and a subsequence (n_j) with $\mathbb{P}(|X_{n_j} - X| > \varepsilon) \ge \varepsilon$ for all j. No further subsequence of (X_{n_j}) can converge to X almost surely, since a.s. convergence would imply convergence in probability (hence $\mathbb{P}(|X_{n_{j_k}} - X| > \varepsilon) \to 0$), contradicting $\mathbb{P}(|X_{n_{j_k}} - X| > \varepsilon) \ge \varepsilon$. □

Proof of (c). By (a), extract a subsequence $X_{n_k} \xrightarrow{a.s.} X$. For any index n, choose k such that $n_k \le n \le n_{k+1}$. Since (X_n) is a.s. nondecreasing,

$$X_{n_k} \le X_n \le X_{n_{k+1}} \qquad \text{a.s.}$$

Letting $k \to \infty$, both endpoints converge to X a.s., so $X_n \to X$ a.s. by the squeeze theorem. □

Remark 5.12 (Why this matters). Part (a) is a workhorse: it lets us apply tools that require a.s. convergence (Fatou's lemma, dominated convergence) within proofs that only assume convergence in probability. We use exactly this strategy in the proof of Vitali's theorem below. Part (b) is especially useful for **proving** convergence in probability: instead of a direct ε–δ argument, one can show that every subsequence has a further subsequence converging a.s. to the same limit. Part (c) appears in the proof of the Kolmogorov–Etemadi SLLN (Chapter 6).

5.2 Continuous Mapping Theorem (CMT)

Theorem 5.13 (Continuous mapping: convergence in probability (continuous case))**.** *Let S, T be metric spaces, let X_n, X be S-valued random variables, and let $f : S \to T$ be continuous. If $X_n \xrightarrow{\mathbb{P}} X$, then*

$$f(X_n) \xrightarrow{\mathbb{P}} f(X).$$

Proof. Fix $\varepsilon > 0$ and $\eta > 0$. Since f is continuous, for every $R > 0$ it is uniformly continuous on the compact set $\overline{B}_S(x_0, R)$; in particular, we may argue as follows.

Choose $R > 0$ such that $\mathbb{P}\big(d_S(X, x_0) > R\big) < \eta$ for some fixed $x_0 \in S$. By uniform continuity of f on $\overline{B}_S(x_0, R+1)$, there exists $\delta \in (0, 1)$ such that

$$d_S(x, y) < \delta \text{ and } d_S(x, x_0) \le R \quad \Longrightarrow \quad d_T(f(x), f(y)) < \varepsilon.$$

Hence

$$\begin{aligned}\mathbb{P}\big(d_T(f(X_n), f(X)) > \varepsilon\big) &\le \mathbb{P}\big(d_S(X, x_0) > R\big) + \mathbb{P}\big(d_S(X, x_0) \le R,\ d_S(X_n, X) \ge \delta\big) \\ &\le \eta + \mathbb{P}\big(d_S(X_n, X) \ge \delta\big).\end{aligned}$$

Since $X_n \xrightarrow{\mathbb{P}} X$, the last probability tends to 0. Taking $\limsup_{n\to\infty}$ gives $\limsup_n \mathbb{P}(d_T(f(X_n), f(X)) > \varepsilon) \le \eta$. As $\eta > 0$ is arbitrary, $\mathbb{P}(d_T(f(X_n), f(X)) > \varepsilon) \to 0$. □

Theorem 5.14 (Continuous Mapping Theorem: weak convergence (continuous case))**.** *Let X_n, X be $\mathbb{R}^d$-valued random vectors and let $f : \mathbb{R}^d \to \mathbb{R}^k$ be continuous. If $X_n \Rightarrow X$, then*

$$f(X_n) \Rightarrow f(X).$$

Proof. Fix any $g \in C_b(\mathbb{R}^k)$. Since f is continuous, the composition $g \circ f$ belongs to $C_b(\mathbb{R}^d)$. Hence, by the definition of $X_n \Rightarrow X$,

$$\mathbb{E}[g(f(X_n))] = \mathbb{E}[(g \circ f)(X_n)] \longrightarrow \mathbb{E}[(g \circ f)(X)] = \mathbb{E}[g(f(X))].$$

Since this holds for all $g \in C_b(\mathbb{R}^k)$, we conclude that $f(X_n) \Rightarrow f(X)$. □

Remark 5.15 (Discontinuities). Both versions of the continuous mapping principle admit extensions to certain discontinuous maps: if f is Borel measurable and $\mathbb{P}(X \in \mathrm{Disc}(f)) = 0$, then $X_n \xrightarrow{\mathbb{P}} X$ implies $f(X_n) \xrightarrow{\mathbb{P}} f(X)$, and similarly $X_n \Rightarrow X$ implies $f(X_n) \Rightarrow f(X)$. We will return to these extensions later (after developing the weak convergence toolbox).

Example 5.16 (Using CMT). If $X_n \Rightarrow X$ in $\mathbb{R}^d$, then for any fixed $a \in \mathbb{R}^d$,

$$a^\top X_n \Rightarrow a^\top X,$$

and also

$$\|X_n\| \Rightarrow \|X\|.$$

Indeed, both maps $x \mapsto a^\top x$ and $x \mapsto \|x\|$ are continuous, so the claim follows from Theorem 5.14.

Remark 5.17 (Why discontinuities matter). The conclusion can fail for discontinuous maps. For instance, for $f(x) = \mathbf{1}_{\{x \le 0\}}$, the point 0 is a discontinuity, and one needs an extra condition such as $\mathbb{P}(X = 0) = 0$ to obtain $f(X_n) \Rightarrow f(X)$.

5.3 Slutsky's theorem

Theorem 5.18 (Slutsky). *Let X_n be $\mathbb{R}^d$-valued and Y_n be $\mathbb{R}^m$-valued random vectors. Assume that*

$$X_n \Rightarrow X \qquad \text{and} \qquad Y_n \xrightarrow{\mathbb{P}} c \in \mathbb{R}^m.$$

Then

$$(X_n, Y_n) \Rightarrow (X, c).$$

Remark 5.19. We will use Slutsky as a tool. A proof (using tightness of (X_n) and uniform continuity on compact sets) is given in Appendix E.3.

Corollary 5.20 (Common special cases). *Assume $X_n \Rightarrow X$ and $Y_n \xrightarrow{\mathbb{P}} c$ (real-valued, for simplicity). Then*

$$X_n + Y_n \Rightarrow X + c, \qquad X_n Y_n \Rightarrow cX,$$

and if $c \neq 0$, also

$$\frac{X_n}{Y_n} \Rightarrow \frac{X}{c}.$$

Proof. Apply Theorem 5.18 and the Continuous Mapping Theorem with $h(x, y) = x + y$, $h(x, y) = xy$, and (if $c \neq 0$) $h(x, y) = x/y$. □

Example 5.21 (Preview: studentization via Slutsky). Suppose we are in a situation where an estimator admits an asymptotic normal approximation

$$\sqrt{n}\,(\bar{X}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2),$$

and suppose $S_n \xrightarrow{\mathbb{P}} \sigma$ for some $\sigma > 0$ (e.g. S_n is a consistent estimator of the standard deviation). Then Slutsky implies

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \Rightarrow \mathcal{N}(0,1).$$

We will prove a CLT later; this example shows why Slutsky is indispensable in asymptotic statistics.

5.4 UI and convergence in L^1

Definition 5.22 (Uniform integrability)**.** A family of integrable random variables $\mathcal{F} \subset L^1$ is called **uniformly integrable** (UI) if

$$\lim_{K\to\infty} \sup_{X\in\mathcal{F}} \mathbb{E}\big[\,|X|\,\mathbf{1}_{\{|X|>K\}}\,\big] = 0.$$

A sequence $(X_n)_{n\geq 1}$ is UI if the set $\{X_n : n \geq 1\}$ is UI.

Theorem 5.23 (Vitali convergence theorem (UI + probability $\Rightarrow L^1$))**.** *Let $(X_n)_{n\geq 1}$ be integrable random variables and let X be a random variable. Assume that $X_n \xrightarrow{\mathbb{P}} X$ and that $\{X_n\}$ is uniformly integrable. Then $X \in L^1$ and*

$$\mathbb{E}|X_n - X| \longrightarrow 0,$$

i.e. $X_n \to X$ in L^1.

Corollary 5.24 (Vitali: expectations converge)**.** *If $X_n \xrightarrow{\mathbb{P}} X$ and $\{X_n\}$ is UI, then $X \in L^1$ and*

$$\mathbb{E}[X_n] \to \mathbb{E}[X],$$

provided the X_n are integrable (e.g. $X_n \in L^1$ for all n).

Proof of Thm. 5.23. Fix $\varepsilon > 0$. By uniform integrability, there exists $K < \infty$ such that

$$\sup_{n\geq 1} \mathbb{E}\big[|X_n|\,\mathbf{1}_{\{|X_n|>K\}}\big] < \varepsilon. \tag{5.1}$$

Define truncations

$$X_n^{(K)} := X_n\,\mathbf{1}_{\{|X_n|\leq K\}}, \qquad X^{(K)} := X\,\mathbf{1}_{\{|X|\leq K\}}.$$

Step 1: the truncated part converges in L^1. We claim that $X_n^{(K)} \to X^{(K)}$ in L^1. Fix $\eta > 0$. Since $|X_n^{(K)} - X^{(K)}| \leq 2K$, we have

$$\begin{aligned}
\mathbb{E}|X_n^{(K)} - X^{(K)}| &= \mathbb{E}\Big(|X_n^{(K)} - X^{(K)}|\mathbf{1}_{\{|X_n^{(K)}-X^{(K)}|\leq\eta\}}\Big) \\
&\quad + \mathbb{E}\Big(|X_n^{(K)} - X^{(K)}|\mathbf{1}_{\{|X_n^{(K)}-X^{(K)}|>\eta\}}\Big) \\
&\leq \eta + 2K\,\mathbb{P}\big(|X_n^{(K)} - X^{(K)}| > \eta\big).
\end{aligned} \tag{5.2}$$

Moreover, if $|X_n| \leq K$ and $|X| \leq K$, then $|X_n^{(K)} - X^{(K)}| = |X_n - X|$. Hence

$$\{|X_n^{(K)} - X^{(K)}| > \eta\} \subset \{|X_n - X| > \eta\} \;\cup\; \{|X| > K\} \;\cup\; \{|X_n| > K\}.$$

Therefore

$$\mathbb{P}\big(|X_n^{(K)} - X^{(K)}| > \eta\big) \le \mathbb{P}(|X_n - X| > \eta) + \mathbb{P}(|X| > K) + \mathbb{P}(|X_n| > K). \quad (5.3)$$

Since $X_n \xrightarrow{\mathbb{P}} X$, we have $\mathbb{P}(|X_n - X| > \eta) \to 0$.

It remains to control $\mathbb{P}(|X_n| > K)$ and $\mathbb{P}(|X| > K)$. By Markov,

$$\mathbb{P}(|X_n| > K) \le \frac{1}{K}\mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] \le \frac{\varepsilon}{K} \qquad \text{for all } n,$$

using (5.1). Also, since $X_n \xrightarrow{\mathbb{P}} X$ we can extract a subsequence $X_{n_j} \to X$ a.s. (Proposition 5.11(a)); applying Fatou's lemma along this subsequence gives

$$\mathbb{E}\big[|X|\,\mathbf{1}_{\{|X|>K\}}\big] \le \liminf_{j\to\infty} \mathbb{E}\big[|X_{n_j}|\,\mathbf{1}_{\{|X_{n_j}|>K\}}\big] \le \varepsilon,$$

so in particular $\mathbb{P}(|X| > K) \le \varepsilon/K$ by Markov.

Plugging these bounds into (5.3) and then into (5.2) gives

$$\limsup_{n\to\infty} \mathbb{E}|X_n^{(K)} - X^{(K)}| \le \eta + 2K\Big(0 + \frac{\varepsilon}{K} + \frac{\varepsilon}{K}\Big) = \eta + 4\varepsilon.$$

Since $\eta > 0$ is arbitrary, we obtain

$$\limsup_{n\to\infty} \mathbb{E}|X_n^{(K)} - X^{(K)}| \le 4\varepsilon. \quad (5.4)$$

Step 2: tails are small in L^1. We have

$$\mathbb{E}|X_n - X| \le \mathbb{E}|X_n^{(K)} - X^{(K)}| + \mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] + \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>K\}}\big].$$

Taking $\limsup_{n\to\infty}$, using (5.4), (5.1), and $\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] \le \varepsilon$, we get

$$\limsup_{n\to\infty} \mathbb{E}|X_n - X| \le 4\varepsilon + \varepsilon + \varepsilon = 6\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that $\mathbb{E}|X_n - X| \to 0$.

Finally, $X \in L^1$ follows from $X_n \in L^1$ and $\mathbb{E}|X_n - X| \to 0$ (triangle inequality). □

Lemma 5.25 (L^1 convergence implies uniform integrability)**.** *If $X_n \to X$ in L^1, then the family $\{X_n : n \ge 1\}$ is uniformly integrable.*

Proof. Fix $\varepsilon > 0$. Since $X \in L^1$, choose K_0 such that

$$\mathbb{E}\big[|X|\mathbf{1}_{\{|X|>K_0\}}\big] < \varepsilon.$$

Next choose N such that for all $n \ge N$,

$$\mathbb{E}|X_n - X| < \varepsilon.$$

Let $K := 2K_0$. For $n \ge N$ we have

$$|X_n|\mathbf{1}_{\{|X_n|>2K_0\}} \le |X|\mathbf{1}_{\{|X|>K_0\}} + |X_n - X|.$$

Indeed, on $\{|X_n| > 2K_0\}$ either $|X| > K_0$ or else $|X| \le K_0$, in which case $|X_n - X| \ge |X_n| - |X| > K_0$ and hence $|X_n| \le |X| + |X_n - X| \le K_0 + |X_n - X| \le 2|X_n - X|$ so the displayed bound holds (and in any case the inequality above is valid pointwise). Taking expectations yields, for $n \ge N$,

$$\mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] \le \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>K_0\}}\big] + \mathbb{E}|X_n - X| < 2\varepsilon.$$

For the finitely many indices $n = 1, \dots, N-1$, choose $K \ge 2K_0$ large enough so that

$$\max_{1\le n\le N-1} \mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] < 2\varepsilon,$$

which is possible since each $X_n \in L^1$. Then $\sup_n \mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>K\}}] < 2\varepsilon$, proving uniform integrability. □

Proposition 5.26 (Easy direction in Vitali's Theorem)**.** *If $X_n \to X$ in L^1, then $X_n \xrightarrow{\mathbb{P}} X$ and $\{X_n\}$ is uniformly integrable.*

Proof. Markov gives $X_n \xrightarrow{\mathbb{P}} X$:

$$\mathbb{P}(|X_n - X| > \varepsilon) \le \frac{\mathbb{E}|X_n - X|}{\varepsilon} \to 0.$$

Uniform integrability follows from Lemma 5.25. □

5.5 Skorokhod's Theorem

Theorem 5.27 (Skorokhod representation theorem)**.** *Let (S, d) be a Polish space (complete separable metric space) with its Borel σ-field. Let X_n, X be S-valued random elements such that*

$$X_n \Rightarrow X.$$

Then there exist a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and S-valued random elements $\widetilde{X}_n, \widetilde{X}$ on it such that

$$\widetilde{X}_n \overset{d}{=} X_n, \qquad \widetilde{X} \overset{d}{=} X, \qquad \text{and} \qquad \widetilde{X}_n \to \widetilde{X} \text{ a.s.}$$

Remark 5.28 (How to read/use it). We state this theorem without proof (apply-only). Skorokhod does **not** upgrade $X_n \Rightarrow X$ to a.s. convergence on the **original** probability space. It produces a **new coupling** $(\widetilde{X}_n, \widetilde{X})$ with the same marginal laws and with almost sure convergence. To pass from a.s. convergence to convergence of expectations, one still needs an integrability condition (e.g. uniform integrability).

Corollary 5.29 (Weak convergence + UI implies convergence of expectations)**.** *Let $(X_n)_{n\ge1}$ be real-valued random variables. Assume that*

$$X_n \Rightarrow X \qquad \text{and} \qquad \{X_n\}_{n\ge1} \text{ is uniformly integrable.}$$

Then $X \in L^1$ and

$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

5.6 Weak Convergence/Convergence Toolbox

5.6.1 General CMT

Theorem 5.30 (Continuous Mapping Theorem (general form))**.** *Let S, T be metric spaces. Let X_n, X be S-valued random elements such that*

$$X_n \Rightarrow X.$$

Let $f : S \to T$ be Borel measurable and assume

$$\mathbb{P}\big(X \in \mathrm{Disc}(f)\big) = 0.$$

Then

$$f(X_n) \Rightarrow f(X).$$

For proof see Appendix E.2.

5.6.2 Criteria for UI

Definition 5.31 (Equivalent ε–δ form for UI)**.** A family $\mathcal{F} \subset L^1$ is UI iff for every $\varepsilon > 0$ there exists $K < \infty$ such that for all $X \in \mathcal{F}$,

$$\mathbb{E}\big[\, |X|\, \mathbf{1}_{\{|X|>K\}} \,\big] < \varepsilon.$$

Equivalently, for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all measurable A with $\mathbb{P}(A) < \delta$ and all $X \in \mathcal{F}$,

$$\mathbb{E}\big[\, |X|\, \mathbf{1}_{A} \,\big] < \varepsilon.$$

Proposition 5.32 (A useful UI criterion: $L^{1+\delta}$-boundedness)**.** *Let $(X_n)_{n\geq 1}$ be random variables. If there exists $\delta > 0$ such that*

$$\sup_{n\geq 1} \mathbb{E}|X_n|^{1+\delta} < \infty,$$

then the family $\{X_n : n \geq 1\}$ is uniformly integrable.

Proof. Let $M := \sup_n \mathbb{E}|X_n|^{1+\delta} < \infty$. For any $K > 0$ and any n,

$$|X_n|\, \mathbf{1}_{\{|X_n|>K\}} \leq \frac{|X_n|^{1+\delta}}{K^{\delta}} \quad \text{(since on } \{|X_n| > K\} \text{ we have } |X_n| \leq |X_n|^{1+\delta}/K^{\delta}).$$

Taking expectations yields

$$\mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] \leq \frac{\mathbb{E}|X_n|^{1+\delta}}{K^{\delta}} \leq \frac{M}{K^{\delta}}.$$

Taking $\sup_n$ and letting $K \to \infty$ gives

$$\lim_{K\to\infty} \sup_n \mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>K\}}\big] = 0,$$

which is exactly uniform integrability. □

Remark 5.33 (de la Vallée–Poussin criterion). A more general sufficient condition is: if there exists a convex increasing function $\Phi : [0,\infty) \to [0,\infty)$ with $\Phi(x)/x \to \infty$ as $x \to \infty$ such that $\sup_n \mathbb{E}[\Phi(|X_n|)] < \infty$, then $\{X_n\}$ is UI. Proposition 5.32 is the special case $\Phi(t) = t^{1+\delta}$.

5.7 Exercises

Homework

Exercise 5.34 (in probability but not in L^1). Give a counterexample where $X_n \xrightarrow{\mathbb{P}} 0$ but $X_n \not\to 0$ in L^1. (Compute $\mathbb{E}|X_n|$ explicitly.)

Exercise 5.35. Prove that $X_n \to X$ in L^∞ implies $X_n \xrightarrow{\mathbb{P}} X$.

Exercise 5.36 (Convergence in distribution does not imply convergence in probability). Let X be a $\mathcal{N}(0,1)$ random variable. Let $(\varepsilon_n)_{n\geq 1}$ be i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\varepsilon_n = 1) = \mathbb{P}(\varepsilon_n = -1) = 1/2$), independent of X. Define

$$X_n := \varepsilon_n X.$$

1. Show that $X_n \Rightarrow X$.
2. Show that X_n does **not** converge to X in probability.

Exercise 5.37 (CMT can fail at discontinuities). Let $f(x) = \mathbf{1}_{\{x>0\}}$. Give an example with $X_n \to 0$ (even in probability) such that $f(X_n)$ does **not** converge to $f(0)$.

Additional Exercises

Exercise 5.38 (L^p but not L^q for $q > p$). Work on $([0,1], \mathcal{B}, \lambda)$. Fix $1 \leq p < q < \infty$ and define

$$X_n(x) := n^{1/q}\, \mathbf{1}_{(0,1/n)}(x).$$

Show that $X_n \to 0$ in L^p but $X_n \not\to 0$ in L^q.

Exercise 5.39 (Weak $\Rightarrow$ probability when the limit is constant). Assume $X_n \Rightarrow c$ where $c \in \mathbb{R}$ is constant. Prove that $X_n \xrightarrow{\mathbb{P}} c$.

Other modes of convergence for measures

The notions introduced above concern convergence of **random variables**. One can also ask how sequences of **measures** (or laws) converge. Weak convergence of measures is defined by testing against bounded continuous functions; here we briefly introduce two relatives: vague convergence (testing against compactly supported functions) and setwise convergence (testing against all measurable sets). The exercises below explore when these notions agree and when they differ.

Definition 5.40 (Vague convergence)**.** Let E be a locally compact Hausdorff space (e.g. $\mathbb{R}^d$) and let $\mathcal{M}(E)$ denote the set of (Radon) measures on E, i.e. finite on compact sets and inner regular (e.g. Borel probability measure). A sequence $(\mu_n)_{n\geq 1} \subset \mathcal{M}(E)$ is said to converge **vaguely** to $\mu \in \mathcal{M}(E)$ if

$$\int_E f\, d\mu_n \longrightarrow \int_E f\, d\mu \qquad \text{for every } f \in C_c(E),$$

where $C_c(E)$ is the space of **continuous functions on E with compact support**. We write $\mu_n \xrightarrow{v} \mu$.

Exercise 5.41 (Vague but not weak: mass escapes to infinity). Let $\mu_n = \delta_n$ be the Dirac probability measure at n on $\mathbb{R}$.

1. Show that for every $f \in C_c(\mathbb{R})$ one has
$$\int_{\mathbb{R}} f \, d\mu_n \;=\; f(n) \longrightarrow 0.$$
Conclude that $\mu_n \to 0$ **vaguely** (as Radon measures).

2. Show that (μ_n) does **not** converge weakly to any probability measure on $\mathbb{R}$.

3. Show that (μ_n) is not tight, i.e. there is no compact $K \subset \mathbb{R}$ with $\sup_n \mu_n(K^c)$ small.

Exercise 5.42 (When does vague convergence upgrade to weak convergence?). Let (μ_n) be probability measures on $\mathbb{R}^d$ and let μ be a Radon measure on $\mathbb{R}^d$. Assume that
$$\int f \, d\mu_n \to \int f \, d\mu \qquad \forall f \in C_c(\mathbb{R}^d)$$
(i.e. $\mu_n \to \mu$ vaguely) and that $\mu(\mathbb{R}^d) = 1$. Show that μ is a probability measure and that in fact $\mu_n \Rightarrow \mu$ (weak convergence).

Definition 5.43 (Setwise convergence)**.** Let $(E, \mathcal{E})$ be a measurable space and let μ_n, μ be (finite) measures on $(E, \mathcal{E})$. We say that μ_n converges **setwise** to μ if for every measurable set $A \in \mathcal{E}$,
$$\mu_n(A) \longrightarrow \mu(A) \qquad (n \to \infty).$$
Equivalently,
$$\int_E f \, d\mu_n \longrightarrow \int_E f \, d\mu \qquad \text{for every } \textbf{bounded measurable } f : E \to \mathbb{R}.$$

Exercise 5.44 (In probability does **not** imply setwise convergence of laws). Let X be Bernoulli with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$. Let $(\varepsilon_n)_{n \geq 1}$ be i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\varepsilon_n = 1) = \mathbb{P}(\varepsilon_n = -1) = \frac{1}{2}$), independent of X. Define
$$X_n := X + \frac{1}{n} \varepsilon_n.$$
Let $\mu_n = \mathcal{L}(X_n)$ and $\mu = \mathcal{L}(X)$.

1. Show that $X_n \to X$ almost surely (hence $X_n \to X$ in probability).

2. Consider the fixed Borel set $A := (-\infty, 0]$. Compute $\mu(A) = \mathbb{P}(X \in A)$ and $\mu_n(A) = \mathbb{P}(X_n \in A)$.

3. Conclude that μ_n does **not** converge to μ **setwise**, i.e. it is not true that $\mu_n(B) \to \mu(B)$ for every Borel set B.

Does the sequence (X_n) converge weakly?

Chapter 6

SLLN

The **Strong Law of Large Numbers** (SLLN) asserts that sample averages converge **almost surely** to the population mean—not merely in probability, as the Weak Law guarantees. This distinction is fundamental: almost sure convergence means that a single, sufficiently long realization of the experiment will reveal the true mean, which is the theoretical backbone of Monte Carlo simulation, empirical distribution estimation, and statistical inference.

This chapter develops the SLLN in stages of increasing generality. We begin with the Kolmogorov–Etemadi theorem (Section 6.1), which is remarkable for requiring only **pairwise** independence and a finite first moment. We then present Kolmogorov's variance summability SLLN (Section 6.2), which handles non-identically distributed variables under full independence, along with the closely related Three-Series Theorem. Table 6.1 below summarizes the landscape.

Version	**Indep.**	**Moment**	**i.d.?**
4th moment (App.)	mutual	$\mathbb{E}\lvert X_1\rvert^4 < \infty$	yes
Finite variance (App.)	mutual	$\mathbb{E}X_1^2 < \infty$	yes
Var. summability (§6.2)	mutual	$\sum \mathbb{V}\mathrm{ar}(X_n)/n^2 < \infty$	no
Kolmogorov–Etemadi (§6.1)	pairwise	$\mathbb{E}\lvert X_1\rvert < \infty$	yes

Table 6.1: Comparison of SLLN versions. "Indep." = independence assumption; "i.d." = identically distributed. Weaker assumptions appear lower.

The appendix contains proofs of the supporting lemmas (Cesàro, Kronecker), the proof of Kolmogorov's maximal inequality (stated in Section 6.2), the proof of the Kolmogorov–Khinchin convergence theorem (also stated in Section 6.2), the simpler 4th-moment and 2nd-moment proofs of SLLN, the complete proof of the Three-Series Theorem, and a connection between the SLLN and Birkhoff's ergodic theorem.

6.1 Kolmogorov–Etemadi Theorem

The following theorem in the case of i.i.d. random variables was proved by Kolmogorov. Etemadi [Ete81] found an easier proof which proves the theorem under a weaker assumption of pairwise independence.

Statement

Theorem 6.1 (Kolmogorov–Etemadi SLLN)**.** *Let $(X_n)_{n\geq 1}$ be **pairwise independent, identically distributed** random variables with $\mathbb{E}|X_1| < \infty$. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X_1],$$

where $S_n = X_1 + \cdots + X_n$.

Remark 6.1. The remarkable feature of this theorem is that it requires only **pairwise independence** rather than full (mutual) independence. This makes it applicable in settings where full independence fails but pairwise independence holds.

Proof

Step 0: Reduction to the centered case

By considering $X_n - \mathbb{E}[X_1]$ in place of X_n, we may assume without loss of generality that $\mathbb{E}[X_1] = 0$. Our goal is then to show that $S_n/n \to 0$ a.s.

Step 1: Truncation

Define the truncated random variables

$$Y_k = X_k \mathbf{1}_{\{|X_k| \leq k\}}, \qquad T_n = \sum_{k=1}^{n} Y_k.$$

Claim 6.2. *$X_k \neq Y_k$ only finitely often, almost surely.*

Proof of Claim. Since X_1 is integrable, we have

$$\sum_{k=1}^{\infty} \mathsf{P}(X_k \neq Y_k) = \sum_{k=1}^{\infty} \mathsf{P}(|X_k| > k) = \sum_{k=1}^{\infty} \mathsf{P}(|X_1| > k) \leq \mathbb{E}|X_1| < \infty.$$

The final inequality follows from the standard bound $\sum_{k=1}^{\infty} \mathsf{P}(|X| > k) \leq \mathbb{E}|X|$ for any nonnegative random variable. By the first Borel–Cantelli lemma,

$$\mathsf{P}(X_k \neq Y_k \text{ i.o.}) = 0.$$

□

Since $X_k = Y_k$ for all sufficiently large k almost surely, we have

$$\frac{S_n - T_n}{n} \to 0 \text{ a.s.}$$

It therefore suffices to show that $T_n/n \to 0$ a.s.

Step 2: Centering the truncated variables

Define the centered truncated variables

$$\overline{Y}_k = Y_k - \mathbb{E}[Y_k], \qquad \overline{T}_n = \sum_{k=1}^{n} \overline{Y}_k.$$

Claim 6.3. $\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Y_k] \to 0$ *as* $n \to \infty$.

Proof of Claim. By the dominated convergence theorem,

$$\mathbb{E}[Y_k] = \mathbb{E}\big[X_1 \mathbf{1}_{\{|X_1| \le k\}}\big] \xrightarrow{k\to\infty} \mathbb{E}[X_1] = 0.$$

By Cesàro's lemma (Lemma F.2), if a sequence $a_k \to L$, then $\frac{1}{n}\sum_{k=1}^{n} a_k \to L$. Hence

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Y_k] \to 0.$$

□

Since $T_n = \overline{T}_n + \sum_{k=1}^{n} \mathbb{E}[Y_k]$, it now suffices to show that $\overline{T}_n/n \to 0$ a.s.

Step 3: Convergence along a subsequence

Fix $\alpha > 1$ and define the subsequence $n_j = \lfloor \alpha^j \rfloor$ for $j \ge 1$. We will show that $\overline{T}_{n_j}/n_j \to 0$ a.s.

By Chebyshev's inequality,

$$\mathsf{P}\big(|\overline{T}_{n_j}| > \varepsilon n_j\big) \le \frac{\mathbb{V}\mathrm{ar}(\overline{T}_{n_j})}{\varepsilon^2 n_j^2}.$$

Since the (X_k) are pairwise independent, so are the $(\overline{Y}_k)$, and hence

$$\mathbb{V}\mathrm{ar}(\overline{T}_{n_j}) = \sum_{k=1}^{n_j} \mathbb{V}\mathrm{ar}(\overline{Y}_k) = \sum_{k=1}^{n_j} \mathbb{V}\mathrm{ar}(Y_k) \le \sum_{k=1}^{n_j} \mathbb{E}[Y_k^2].$$

Therefore,

$$\mathsf{P}\big(|\overline{T}_{n_j}| > \varepsilon n_j\big) \le \frac{1}{\varepsilon^2 n_j^2} \sum_{k=1}^{n_j} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1| \le k\}}\big].$$

We now sum over j and interchange the order of summation:

$$\sum_{j=1}^{\infty} \frac{1}{n_j^2} \sum_{k=1}^{n_j} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1| \le k\}}\big] = \sum_{k=1}^{\infty} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1| \le k\}}\big] \sum_{\substack{j \ge 1 \\ n_j \ge k}} \frac{1}{n_j^2}.$$

Since $n_j = \lfloor \alpha^j \rfloor \sim \alpha^j$, there exists a constant $C = C(\alpha) > 0$ such that

$$\sum_{\substack{j \ge 1 \\ n_j \ge k}} \frac{1}{n_j^2} \le \frac{C}{k^2}.$$

Thus,

$$\sum_{j=1}^{\infty} \frac{1}{n_j^2} \sum_{k=1}^{n_j} \mathbb{E}[Y_k^2] \le C \sum_{k=1}^{\infty} \frac{\mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1| \le k\}}\big]}{k^2}. \tag{6.1}$$

Lemma 6.4 (Variance Summability Lemma)**.** *For any random variable Z with $\mathbb{E}|Z| < \infty$,*

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}\big[Z^2 \mathbf{1}_{\{|Z| \le k\}}\big]}{k^2} \le 2\mathbb{E}|Z|.$$

Proof. Using Fubini's theorem to interchange the sum and expectation,

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\mathbb{E}\big[Z^2 \mathbf{1}_{\{|Z| \le k\}}\big]}{k^2} &= \mathbb{E}\left[Z^2 \sum_{k \ge |Z|} \frac{1}{k^2} \right] \\
&\le \mathbb{E}\left[Z^2 \cdot \frac{2}{|Z|} \right] \quad (\text{since } \sum_{k \ge m} k^{-2} \le 2/m \text{ for } m \ge 1) \\
&= 2\mathbb{E}|Z|.
\end{aligned}$$

□

Applying Lemma 6.4 to (6.1), we obtain

$$\sum_{j=1}^{\infty} \mathsf{P}\big(|\overline{T}_{n_j}| > \varepsilon n_j\big) \le \frac{2C}{\varepsilon^2} \mathbb{E}|X_1| < \infty.$$

By the first Borel–Cantelli lemma,

$$\mathsf{P}\big(|\overline{T}_{n_j}| > \varepsilon n_j \text{ i.o.}\big) = 0.$$

Since this holds for all $\varepsilon > 0$, we conclude that

$$\frac{\overline{T}_{n_j}}{n_j} \to 0 \text{ a.s.}$$

Step 4: Interpolation between subsequence terms

It remains to extend the convergence from the subsequence (n_j) to all of $\mathbb{N}$. For $n_j \le n < n_{j+1}$, we write

$$\frac{\overline{T}_n}{n} = \frac{\overline{T}_{n_j}}{n} + \frac{\overline{T}_n - \overline{T}_{n_j}}{n}.$$

For the first term, since $n_j/n > n_j/n_{j+1} \to 1/\alpha$, we have

$$\left| \frac{\overline{T}_{n_j}}{n} \right| = \left| \frac{\overline{T}_{n_j}}{n_j} \right| \cdot \frac{n_j}{n} \le \left| \frac{\overline{T}_{n_j}}{n_j} \right| \to 0 \text{ a.s.}$$

For the second term, we bound

$$\left| \frac{\overline{T}_n - \overline{T}_{n_j}}{n} \right| \le \frac{1}{n_j} \sum_{k=n_j+1}^{n_{j+1}} |\overline{Y}_k|.$$

Key observation: The random variables $|Y_k| = |X_k|\mathbf{1}_{\{|X_k|\le k\}}$ are pairwise independent and nonnegative. Although they are **not** identically distributed (the distribution of $|Y_k|$ depends on k), the same subsequence argument from Step 3 applies: what is needed is only the variance summability bound and the Borel–Cantelli lemma, neither of which requires identical distributions. Applying this argument to the nonnegative pairwise independent sequence $(|Y_k|)$ shows that

$$\frac{1}{n_j}\sum_{k=1}^{n_j}|Y_k| \to \mathbb{E}|X_1| \text{ a.s.}$$

Since $|\overline{Y}_k| = |Y_k - \mathbb{E}[Y_k]| \le |Y_k| + |\mathbb{E}[Y_k]| \le |Y_k| + \mathbb{E}|X_1|$ (using $|\mathbb{E}[Y_k]| \le \mathbb{E}|X_1|$), we have

$$\frac{1}{n_j}\sum_{k=n_j+1}^{n_{j+1}}|\overline{Y}_k| \le \frac{1}{n_j}\sum_{k=n_j+1}^{n_{j+1}}|Y_k| + \frac{n_{j+1}-n_j}{n_j}\mathbb{E}|X_1|.$$

For the first part, using the subsequence convergence for $|Y_k|$:

$$\begin{aligned}\frac{1}{n_j}\sum_{k=n_j+1}^{n_{j+1}}|Y_k| &= \frac{n_{j+1}}{n_j}\cdot\frac{1}{n_{j+1}}\sum_{k=1}^{n_{j+1}}|Y_k| - \frac{1}{n_j}\sum_{k=1}^{n_j}|Y_k| \\ &\to \alpha\cdot\mathbb{E}|X_1| - \mathbb{E}|X_1| = (\alpha-1)\mathbb{E}|X_1| \text{ a.s.}\end{aligned}$$

Since $\frac{n_{j+1}-n_j}{n_j} \to \alpha - 1$, combining these bounds gives:

$$\limsup_{n\to\infty}\left|\frac{\overline{T}_n}{n}\right| \le 0 + 2(\alpha-1)\mathbb{E}|X_1| \text{ a.s.}$$

Since $\alpha > 1$ was arbitrary and can be taken arbitrarily close to 1, we conclude that

$$\limsup_{n\to\infty}\left|\frac{\overline{T}_n}{n}\right| = 0,$$

and hence $\overline{T}_n/n \to 0$ a.s.

Conclusion

Combining Steps 1–4, we have shown that

$$\frac{S_n}{n} = \frac{S_n - T_n}{n} + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Y_k] + \frac{\overline{T}_n}{n} \to 0 + 0 + 0 = 0 \text{ a.s.}$$

This completes the proof. □

Summary of Key Techniques

(i) **Truncation and Borel–Cantelli:** Reduces the problem from unbounded to bounded random variables.

(ii) **Centering and Cesàro:** Handles the bias introduced by truncation.

(iii) **Subsequence argument:** Proves convergence along an exponentially growing subsequence $n_j = \lfloor \alpha^j \rfloor$ using Chebyshev's inequality and the variance summability lemma. Crucially, this only requires pairwise independence (for the variance identity).

(iv) **Interpolation:** Extends convergence from the subsequence to all integers by bounding the maximum by a sum of absolute values, then applying the subsequence convergence to $|Y_k|$. The error $(\alpha - 1)\mathbb{E}|X_1|$ can be made arbitrarily small by choosing α close to 1.

(v) **Pairwise independence suffices:** The proof uses only Chebyshev's inequality (which requires pairwise uncorrelatedness for the variance identity) and the Borel–Cantelli lemma. No maximal inequality (like Kolmogorov's inequality in Theorem 6.2) is needed, which is why pairwise independence suffices.

Remark 6.5 (Connection to ergodic theory). There is an elegant alternative perspective on the SLLN: one can derive it as a consequence of Birkhoff's pointwise ergodic theorem by embedding the i.i.d. sequence into the shift dynamical system on the product space $\prod_{i=1}^{\infty} \mathbb{R}$. This viewpoint illuminates why the SLLN is fundamentally a statement about ergodicity of the shift. See the appendix (Section F.6) for details.

6.2 Kolmogorov's SLLN (variance summability version)

The Kolmogorov–Etemadi theorem achieves the SLLN under the minimal moment assumption $\mathbb{E}|X_1| < \infty$, but its proof is long and uses only pairwise independence and Chebyshev's inequality. When **mutual** independence is available, a more powerful tool—Kolmogorov's maximal inequality—leads to a shorter proof and a more general result that applies to non-identically distributed variables. We develop this approach in three stages: the maximal inequality, a convergence theorem for series, and the SLLN.

Kolmogorov's Maximal Inequality

The following inequality sharpens Chebyshev's bound by controlling the **maximum** of partial sums at no extra cost.

Theorem 6.2 (Kolmogorov's Maximal Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{V}\mathrm{ar}(X_i) = \sigma_i^2 < \infty$. Let $S_k = \sum_{j=1}^{k} X_j$ and $s_n^2 = \sum_{j=1}^{n} \sigma_j^2$. Then for any $\ell > 0$,*

$$\mathbb{P}\left(\max_{1 \le k \le n} |S_k| \ge \ell \right) \le \frac{s_n^2}{\ell^2}.$$

Compare this with Chebyshev's inequality, which gives $\mathbb{P}(|S_n| \geq \ell) \leq s_n^2/\ell^2$. The remarkable feature is that replacing $|S_n|$ by $\max_{k\leq n}|S_k|$ does not worsen the bound. This "max costs nothing" property is what enables the upgrade from L^2 convergence to almost sure convergence.

The proof requires mutual independence (not just pairwise) and is given in the appendix (Theorem F.4).

Kolmogorov–Khinchin Convergence Theorem

The maximal inequality has an important consequence for the convergence of series of independent random variables.

Theorem 6.3 (Kolmogorov–Khinchin Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be independent random variables with $\mathbb{E}[X_n] = 0$ for all n. If*

$$\sum_{n=1}^{\infty} \mathbb{V}\mathrm{ar}(X_n) < \infty,$$

then $\sum_{n=1}^{\infty} X_n$ converges almost surely.

Note that the hypothesis $\sum \mathbb{V}\mathrm{ar}(X_n) < \infty$ immediately gives L^2 convergence of the partial sums (since $\mathbb{E}[(S_n - S_m)^2] = \sum_{k=m+1}^{n} \mathbb{V}\mathrm{ar}(X_k) \to 0$). However, L^2 convergence does not imply almost sure convergence in general—recall the typewriter sequence from Chapter 5. The content of the theorem is the upgrade from L^2 to a.s. convergence.

The proof applies Kolmogorov's maximal inequality to the tail $\sup_{k>m}|S_k - S_m|$ and uses a Cauchy criterion argument. For the complete proof, see the appendix (Theorem F.6).

Variance Summability SLLN

The theorem below is the main result of this section. It applies when variables are mutually independent but not necessarily identically distributed.

Theorem 6.4 (Kolmogorov's Variance Summability SLLN)**.** *Let $(X_n)_{n\geq 1}$ be independent random variables with $\mathbb{E}[X_n] = \mu_n$ and finite variances. If*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty,$$

then

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \to 0 \quad \text{a.s.}$$

In particular, if $X_1, X_2, \ldots$ are i.i.d. with mean μ and finite variance σ^2, then $\bar{X}_n \to \mu$ a.s.

The pairwise independence is not sufficient for this theorem (without imposing additional conditions) although the counterexamples are very subtle. See Csörgő, S., Tandori, K., Totik, V. (1983). "On the Strong Law of Large Numbers for Pairwise Independent Random Variables," Acta Math. Hungar. 42, 319–330.

Proof. **Step 1: Reduction to centered variables.**

Let $Y_n = X_n - \mu_n$. Then (Y_n) are independent with $\mathbb{E}[Y_n] = 0$ and $\mathrm{Var}(Y_n) = \mathrm{Var}(X_n)$. We need to show $\frac{1}{n}S_n \to 0$ a.s., where $S_n = \sum_{k=1}^{n} Y_k$.

Step 2: Convergence of the weighted series.

Consider the series $\sum_{n=1}^{\infty} \frac{Y_n}{n}$. We verify the hypotheses of Theorem 6.3:

- *Independence:* The Y_n/n are independent since the $Y_n = X_n - \mu_n$ are independent.
- *Centered:* $\mathbb{E}[Y_n/n] = \mathbb{E}[Y_n]/n = 0$.
- *Variance summability:* $\sum_{n=1}^{\infty} \mathbb{V}\mathrm{ar}(Y_n/n) = \sum_{n=1}^{\infty} \frac{\mathbb{V}\mathrm{ar}(Y_n)}{n^2} = \sum_{n=1}^{\infty} \frac{\mathbb{V}\mathrm{ar}(X_n)}{n^2} < \infty$ by hypothesis.

By the Kolmogorov–Khinchin theorem, $\sum_{n=1}^{\infty} \frac{Y_n}{n}$ converges almost surely.

Step 3: Apply Kronecker's lemma.

By Kronecker's lemma (Lemma F.3 in the appendix; the general form with $b_n = n$ increasing to ∞), since $\sum_{n=1}^{\infty} \frac{Y_n}{n}$ converges a.s. (from Step 2), we obtain

$$\frac{1}{n}\sum_{k=1}^{n} Y_k = \frac{S_n}{n} \to 0 \quad \text{a.s.}$$

This completes the proof. □

Summary of Key Techniques

(i) **Variance summability:** The hypothesis $\sum \mathbb{V}\mathrm{ar}(X_n)/n^2 < \infty$ ensures that $\sum Y_n/n$ satisfies the conditions of Kolmogorov–Khinchin.

(ii) **Kolmogorov–Khinchin (Theorem 6.3):** Upgrades L^2 convergence to almost sure convergence of the series $\sum Y_n/n$, using Kolmogorov's maximal inequality internally.

(iii) **Kronecker's lemma:** Transfers convergence of $\sum Y_n/n$ to convergence of the Cesàro means $S_n/n \to 0$.

Generalization: Kolmogorov's Three-Series Theorem

The variance summability SLLN concerns averages S_n/n, which it reduces to the convergence of the series $\sum Y_n/n$ via Kronecker's lemma. The Kolmogorov–Khinchin theorem, however, applies to **any** series of independent centered variables with summable variances. Combined with truncation and centering techniques, this leads to a complete characterization of when a series of independent random variables converges almost surely.

Theorem 6.5 (Kolmogorov's Three-Series Theorem)**.** *Let $(X_n)_{n\geq 1}$ be independent random variables. Fix any constant $c > 0$ and define the truncated variables $Y_n = X_n \mathbf{1}_{\{|X_n| \leq c\}}$. Then $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following three series all converge:*

(i) $\sum_{n=1}^{\infty} P(|X_n| > c)$

(ii) $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$

(iii) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n)$

Moreover, if these conditions hold for one value of $c > 0$, they hold for all $c > 0$.

The three conditions control the three potential obstacles to convergence: (i) large jumps (truncation error is finite by BC I), (ii) drift (the truncated means don't accumulate), and (iii) fluctuation (the variances are summable). The full proof is given in Section F.5 of the appendix. The sufficiency direction is a clean application of Kolmogorov–Khinchin:

Proof sketch (sufficiency). Assume the three conditions hold. By (i) and the first Borel–Cantelli lemma, $X_n \neq Y_n$ for only finitely many n, so $\sum X_n$ converges a.s. if and only if $\sum Y_n$ does. By (ii), $\sum \mathbb{E}[Y_n]$ converges, so $\sum Y_n$ converges a.s. if and only if $\sum (Y_n - \mathbb{E}[Y_n])$ does. The centered variables $Z_n = Y_n - \mathbb{E}[Y_n]$ are independent with $\mathbb{E}[Z_n] = 0$ and $\sum \mathbb{V}\mathrm{ar}(Z_n) = \sum \mathbb{V}\mathrm{ar}(Y_n) < \infty$ by (iii). By the Kolmogorov–Khinchin theorem (Theorem 6.3), $\sum Z_n$ converges a.s. □

Remark

The theorem is sharp: all three conditions are genuinely needed. For example, $\sum \frac{1}{n}$ diverges even though conditions (i) and (iii) would be trivially satisfied for deterministic $X_n = 1/n$ truncated at any $c > 1$—condition (ii) fails.

Note that just orthogonality (i.e. the assumption of uncorrelatedness) or even pairwise independence rather than full independence of the X_i's is not enough to get an a.s. limit. Counterexamples are hard. On the other hand, if we impose stricter condition on the summability of variances, then SLLN can be recovered for pairwise independent and even uncorrelated random variables. According to classical results of Rademacher-Menchoff, for orthogonal X_i (i.e. mean zero and uncorrelated) the condition

$$\sum_i (\log^2 i)\sigma_i^2 < \infty$$

is sufficient for a.s. convergence of S_n to 0. In the opposite direction, if $b_i \uparrow$ with $b_i = o(\log^2 i)$ there exist orthogonal X_i such that $\sum_i b_i \sigma_i^2 < \infty$ and S_n diverges almost surely.

When we have some additional information on the moments of the random variables X_n, the strong law can be improved by using Kolmogorov's method.

Lemma 6.6 (Generalized Kronecker's Lemma)**.** *If $a_n \uparrow \infty$ and $\sum_{n=1}^{\infty} x_n/a_n$ converges, then*

$$\frac{1}{a_n} \sum_{j=1}^{n} x_j \to 0 \text{ as } n \to \infty.$$

For example, if $a_n = n$, then we recover the usual Kronecker's Lemma. For a proof, see Theorem 2.5.9 in Durrett.

Theorem 6.7. *Let X_1, X_2, ... be i.i.d. with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2 < \infty$, and let $S_n = X_1 + ... + X_n$. If $\epsilon > 0$, then*

$$\frac{S_n}{n^{1/2}(\log n)^{1/2+\epsilon}} \to 0$$

with probability 1 as $n \to \infty$.

Proof. Let $a_n = n^{1/2}(\log n)^{1/2+\epsilon}$ for $n \geq 2$ and $a_1 > 0$. Then,

$$\sum_n \mathbb{V}\mathrm{ar}(X_n/a_n) = \sigma^2 \left(\frac{1}{a_1^2} + \sum_{n\geq 2} \frac{1}{n(\log n)^{1+2\epsilon}} \right) < \infty.$$

So by Theorem 6.5, $\sum_n X_n/a_n$ converges with probability 1, and an application of the Kronecker Lemma delivers the result. □

6.3 Applications of the Strong Law

As our first application, we establish the Glivenko–Cantelli Theorem, which is fundamental in statistics. The question is how we can estimate the cdf of a random variable X if we only have a sample of independent realizations of X, an i.i.d. sample $X_1, \ldots, X_n$.

One possible estimator is the **empirical distribution function**, defined as

$$F_n(\omega, t) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j(\omega) \leq x} = \frac{|\{j : X_j(\omega) \leq x\}|}{n}.$$

In other words, $F_n(t)$ is the fraction of values in the sample $(X_1, \ldots, X_n)$ which are $\leq t$. We wrote $F_n(\omega, t)$ in the definition to emphasize that this is a random function. For conciseness, ω is typically suppressed.

As a function of t (with a fixed ω), $F_n(t)$ is a step function with jumps at X_i and the size of every jump equal to $1/n$.

Theorem 6.8 (Glivenko–Cantelli)**.** *As $n \to \infty$, $F_n(t)$ converges to $F(t)$ uniformly on $\mathbb{R}$ with probability* 1*. That is,*

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - F(t)\right| \xrightarrow{a.s.} 0.$$

Proof. First, let us prove that the convergence holds for every t. Let $Y_n = 1_{X_n \leq t}$. Then Y_n are i.i.d. and

$$\mathbb{E}Y_n = \mathbb{P}(X_n \leq t) = F(t),$$

so by SLLN, $F_n(t) := \frac{1}{n}\sum_{j=1}^{n} Y_j \to F(t)$ for every $t \in \mathbb{R}$.

Some effort is needed to prove that the convergence is uniform. We want to prove that there is a set $\tilde{\Omega} \subset \Omega$, such that $\mathbb{P}(\tilde{\Omega}) = 1$, and for every $\omega \in \tilde{\Omega}$ and $\varepsilon > 0$ one can choose $N(\omega, \varepsilon)$ such that for all $n \geq N$, $\sup_{t\in\mathbb{R}} \left|F_n(t) - F(t)\right| \leq \varepsilon$.

Let Y_n be as above and let $Z_n = 1_{X_n<t}$. Then

$$\mathbb{E}Z_n = \mathbb{P}(X < t) = F(t-) := \lim_{x\uparrow t} F(x).$$

and $F_n(t-) = \frac{1}{n}\sum_{j=1}^n Z_j \xrightarrow{a.s.} F(t-)$.

Fix $k \geq 2$ and introduce $k-1$ points $t_{j,k}$, $1 \leq j \leq k-1$, defined as $t_{j,k} = \inf\{y : F(y) \geq j/k\}$. While it can happen that $F(t_{j,k}) - F(t_{j-1,k}) > 1/k$, we always have

$$F(t_{j,k}-) - F(t_{j-1,k}) \leq 1/k. \qquad (A)$$

Since for a given k, there is only a finite number of points $t_{j,k}$, we can extend the almost sure convergence to all of them. In more detail, we can take the intersection of the sets of ω for which the convergence holds for a fixed point $t_{j,k}$. Let the intersection be denoted Ω_k. Then $\mathbb{P}(\Omega_k) = 1$ and for each $\omega \in \Omega_k$, we can find $N(\omega, \frac{1}{k})$ such that if $n \geq N$, then

$$\big|F_n(t_{j,k}) - F(t_{j,k})\big| \leq \frac{1}{k} \text{ and } \big|F_n(t_{j,k}-) - F(t_{j,k}-)\big| \leq \frac{1}{k}. \qquad (B)$$

If we let $x_{0,k} = -\infty$ and $x_{k,k} = \infty$,then the last two inequalities hold for $j = 0$ or k.

Then for the points between $t_{j,k}$ we can do the following estimate. Let $t_{j-1,k} \leq x \leq t_{j,k}$. Then we use monotonicity and properties (A) and (B) to get:

$$F_n(x) \overset{M}{\leq} F_n(x_{j,k}-) \overset{B}{\leq} F(x_{j,k}-) + \frac{1}{k} \overset{A}{\leq} F(x_{j-1,k}) + \frac{2}{k} \overset{M}{\leq} F(x) + \frac{2}{k}.$$

Similarly,

$$F_n(x) \overset{M}{\geq} F_n(x_{j-1,k}) \overset{B}{\geq} F(x_{j-1,k}) - \frac{1}{k} \overset{A}{\geq} F(x_{j,k}-) - \frac{2}{k} \overset{M}{\geq} F(x) - \frac{2}{k}.$$

and therefore $\sup_x |F_n(x) - F(x)| \leq \frac{2}{k}$. Then we can take $\tilde{\Omega} = \cap_{k=1}^\infty \Omega_k$ and the desired result is proved. □

The second application is **Shannon's theorem**.

Let $X_1, X_2, \ldots$ be i.i.d. with $X_i \in \{1, \ldots, r\}$ and $\mathbb{P}(X_i = k) = p(k) > 0$ for $1 \leq k \leq r$. Here we are thinking of $1, \ldots, r$ as letters of an alphabet, and $X_1, X_2, \ldots$ are the successive letters produced by an information source. The information source is very simple in this example since it produces independent letters. Let $\pi_n(\omega) = p(X_1(\omega)) \ldots p(X_n(\omega))$ be the probability of the realization we observed in the first n trials.

Note that $\pi_n(\omega)$ is random: unless all $p(k)$ are the same, the probability of the first n letters depends on how many letters of each type are in this sequence.

Since $\log_2 \pi_n(\omega)$ is a sum of independent random variables, it follows from the strong law of large numbers that

$$-\frac{1}{n}\log \pi_n(\omega) \xrightarrow{a.s.} H := -\sum_{k=1}^r p(k)\log_2 p(k)$$

The constant H is called the **entropy** of the source and it is a measure of how random it is. A consequence of the almost sure convergence above is the following property: If $\varepsilon > 0$, then as $n \to \infty$

$$\mathbb{P}\Big\{2^{-n(H+\varepsilon)} \leq \pi_n(\omega) \leq 2^{-n(H-\varepsilon)}\Big\} \to 1.$$

So with very large probability a sequence of n initial letters will have probability $\pi_n(\omega) \geq 2^{-n(H+\varepsilon)}$. We can conclude that there are no more than $2^{n(H+\varepsilon)}$ of such sequences, which is typically much smaller than the number of all sequences r^n.

This can be interpreted as saying that if we have all sequences of symbols 0 and 1 that consist of $(H+\varepsilon)n$ letters, we can reliably code the messages of n letters from the alphabet. Here reliably means that the probability that we cannot code a message produced by the informational channel becomes arbitrarily small as n grows.

6.4 Exercises

Homework

Exercise 6.9. Let $(X_n)_{n\geq 1}$ be independent random variables with $P(X_n = n) = P(X_n = -n) = \frac{1}{2n^2}$ and $P(X_n = 0) = 1 - \frac{1}{n^2}$. Show that $\bar{X}_n \to 0$ almost surely.

Exercise 6.10. (SLLN for Random Signs) Let $(a_n)_{n\geq 1}$ be a sequence of real numbers and let $(\varepsilon_n)_{n\geq 1}$ be i.i.d. Rademacher random variables, i.e., $P(\varepsilon_n = 1) = P(\varepsilon_n = -1) = \frac{1}{2}$. Show that $\sum_{n=1}^{\infty} a_n\varepsilon_n$ converges almost surely if and only if $\sum_{n=1}^{\infty} a_n^2 < \infty$.

Exercise 6.11. (Borel's Normal Number Theorem) Let $x \in [0,1]$ be chosen uniformly at random, and let $d_n(x) \in \{0,1,\ldots,9\}$ be the n-th digit in the decimal expansion of x. Prove that for almost every x,

$$\frac{1}{n}\#\{k \leq n : d_k(x) = j\} \to \frac{1}{10} \quad \text{for each } j \in \{0,1,\ldots,9\}.$$

Exercise 6.12. (Converse to SLLN) Let $(X_n)_{n\geq 1}$ be i.i.d. random variables. Suppose $\bar{X}_n \to c$ almost surely for some constant $c \in \mathbb{R}$. Prove that $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = c$.

Additional Exercises

Foundational Exercises

Exercise 6.13. Prove the general form of Kronecker's lemma: if $0 < b_n \uparrow \infty$ and $\sum_{n=1}^{\infty} \frac{a_n}{b_n}$ converges, then $\frac{1}{b_n}\sum_{k=1}^{n} a_k \to 0$.

Exercise 6.14. Let $(X_n)_{n\geq 1}$ be independent random variables with $P(X_n = 2^n) = P(X_n = -2^n) = \frac{1}{2^{2n}}$ and $P(X_n = 0) = 1 - \frac{1}{2^{2n-1}}$. Determine whether $\sum_{n=1}^{\infty} X_n$ converges almost surely.

Exercise 6.15 (Tail-sum lemma). Assume $\mathbb{E}|X| < \infty$. Prove that

$$\sum_{n=1}^{\infty} \mathbb{P}(|X| > n) < \infty.$$

Exercise 6.16 (If averages converge, then increments are small). Let $S_n = \sum_{k=1}^{n} X_k$. Show that if $S_n/n \to L \in \mathbb{R}$, then

$$\frac{X_n}{n} \to 0.$$

Exercise 6.17. Why does the proof of Kolmogorov–Etemadi theorem (Theorem 6.1) fails for pairwise uncorrelated identically distributed random variables $X_1, X_2, \ldots$?

Exercise 6.18. Recall the following notation: for a sequence of events $\{A_n\}$, we define a new event $\{A_n \text{ i.o.}\}$,

$$\{A_n \text{ i.o.}\} = \limsup A_n = \lim_{m\to\infty} \bigcup_{n\geq m} A_n$$
$$= \bigcap_m \bigcup_{n\geq m} A_n.$$

In words, $\{A_n \text{ i.o.}\}$ consists of outcomes ω that repeat infinitely often, that is, they are in infinitely many A_n.

Let $X_1, X_2, \ldots,$ are i.i.d. with $\mathbb{E}|X_i| = \infty$ and $S_n = X_1 + \ldots + X_n$, then

1. $\mathbb{P}\big(|X_n| \geq n \text{ i.o.}\big) = 1, \textit{and}$
2. $\mathbb{P}\Big(\lim \frac{S_n}{n} \text{ exists and } \in (-\infty, \infty)\Big) = 0.$

This exercise shows that the assumption $\mathbb{E}|X| < \infty$ is necessary for the validity of the strong law of large numbers.

Counterexamples (Sharpness of Conditions)

Exercise 6.19. (a) Let $(X_n)_{n\geq 1}$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Show that $\bar{X}_n \to 0$ a.s. and verify that the variance summability condition $\sum \frac{\mathbb{V}\text{ar}(X_n)}{n^2} < \infty$ is satisfied.

(b) Construct i.i.d. random variables $(Y_n)_{n\geq 1}$ with $\mathbb{E}[Y_1] = 0$, $\mathbb{V}\text{ar}(Y_1) = \infty$, but $\mathbb{E}[|Y_1|] < \infty$, so that the SLLN still holds.

Exercise 6.20 (HW: Sharpness of integrability (a concrete counterexample)). Construct i.i.d. (X_n) such that $\mathbb{E}|X_1| = \infty$ and $\frac{S_n}{n}$ does not converge to a finite limit a.s.

One option: take

$$\mathbb{P}(X_1 = k) = \mathbb{P}(X_1 = -k) = \frac{c}{2k^2}, \qquad k \geq 1,$$

with c chosen so total probability is 1.

Exercise 6.21. Let $(X_n)_{n\geq 1}$ be i.i.d. with the symmetric Cauchy distribution, i.e., with density $f(x) = \frac{1}{\pi(1+x^2)}$. Show that $\bar{X}_n$ does not converge almost surely. In fact, prove that $\bar{X}_n \overset{d}{=} X_1$ for all n.

Extensions and Variations

Exercise 6.22. (Marcinkiewicz–Zygmund SLLN) Let $(X_n)_{n\geq 1}$ be i.i.d. with $\mathbb{E}[|X_1|^p] < \infty$ for some $1 \leq p < 2$, and $\mathbb{E}[X_1] = 0$ when $p > 1$. Prove that

$$\frac{S_n}{n^{1/p}} \to 0 \quad \text{a.s.}$$

where $S_n = \sum_{k=1}^n X_k$.

Exercise 6.23. (Weighted Strong Law) Let $(X_n)_{n\geq 1}$ be independent random variables with $\mathbb{E}[X_n] = 0$ and $\mathbb{V}\text{ar}(X_n) \leq C$ for all n. Let $(w_n)_{n\geq 1}$ be positive weights with $W_n = \sum_{k=1}^n w_k \to \infty$. Show that if $\sum_{n=1}^\infty \frac{w_n^2}{W_n^2} < \infty$, then

$$\frac{1}{W_n}\sum_{k=1}^n w_k X_k \to 0 \quad \text{a.s.}$$

Exercise 6.24 (Etemadi maximal inequality). Let $S_k := \sum_{i=1}^k X_i$ for independent random variables (X_i). Show that for every $t > 0$,

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k| \geq 3t\right) \leq 3 \max_{1\leq k\leq n} \mathbb{P}(|S_k| \geq t).$$

Applications

Exercise 6.25. (Monte Carlo Integration) Let $f : [0,1] \to \mathbb{R}$ be measurable with $\int_0^1 f(x)^2\,dx < \infty$. Let $(U_n)_{n\geq 1}$ be i.i.d. Uniform$[0,1]$ random variables. Show that

$$\frac{1}{n}\sum_{k=1}^n f(U_k) \to \int_0^1 f(x)\,dx \quad \text{a.s.}$$

Exercise 6.26. (Convergence of Random Series) Let $(X_n)_{n\geq 1}$ be independent with $X_n \sim \text{Uniform}[-a_n, a_n]$ for a sequence $(a_n)_{n\geq 1}$ of positive reals. Show that $\sum_{n=1}^\infty X_n$ converges almost surely if and only if $\sum_{n=1}^\infty a_n^2 < \infty$.

Exercise 6.27 (6.13). Prove the general form of Kronecker's lemma: if $0 < b_n \uparrow \infty$ and $\sum_{n=1}^\infty \frac{a_n}{b_n}$ converges, then $\frac{1}{b_n}\sum_{k=1}^n a_k \to 0$.

Exercise 6.28 (6.14). Let $(X_n)_{n\geq 1}$ be independent random variables with $P(X_n = 2^n) = P(X_n = -2^n) = \frac{1}{2^{2n}}$ and $P(X_n = 0) = 1 - \frac{1}{2^{2n-1}}$. Determine whether $\sum_{n=1}^\infty X_n$ converges almost surely.

Exercise 6.29 (6.19). (a) Let $(X_n)_{n\geq 1}$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Show that $\bar{X}_n \to 0$ a.s. and verify that the variance summability condition $\sum \frac{\mathbb{V}\text{ar}(X_n)}{n^2} < \infty$ is satisfied.

(b) Construct i.i.d. random variables $(Y_n)_{n\geq 1}$ with $\mathbb{E}[Y_1] = 0$, $\mathbb{V}\text{ar}(Y_1) = \infty$, but $\mathbb{E}[|Y_1|] < \infty$, so that the SLLN still holds.

Exercise 6.30 (6.21). Let $(X_n)_{n\geq 1}$ be i.i.d. with the symmetric Cauchy distribution. Show that $\bar{X}_n$ does not converge almost surely. In fact, prove that $\bar{X}_n \overset{d}{=} X_1$ for all n.

Exercise 6.31 (6.22). (Marcinkiewicz–Zygmund SLLN) Let $(X_n)_{n\geq 1}$ be i.i.d. with $\mathbb{E}[|X_1|^p] < \infty$ for some $1 \leq p < 2$, and $\mathbb{E}[X_1] = 0$ when $p > 1$. Prove that $\frac{S_n}{n^{1/p}} \to 0$ a.s.

Exercise 6.32 (6.23). (Weighted Strong Law) Let $(X_n)_{n\geq 1}$ be independent random variables with $\mathbb{E}[X_n] = 0$ and $\mathbb{V}\mathrm{ar}(X_n) \leq C$ for all n. Let $(w_n)_{n\geq 1}$ be positive weights with $W_n = \sum_{k=1}^n w_k \to \infty$. Show that if $\sum_{n=1}^\infty \frac{w_n^2}{W_n^2} < \infty$, then $\frac{1}{W_n}\sum_{k=1}^n w_k X_k \to 0$ a.s.

Exercise 6.33 (6.10). (SLLN for Random Signs) Let $(a_n)_{n\geq 1}$ be a sequence of real numbers and let $(\varepsilon_n)_{n\geq 1}$ be i.i.d. Rademacher random variables. Show that $\sum_{n=1}^\infty a_n\varepsilon_n$ converges almost surely if and only if $\sum_{n=1}^\infty a_n^2 < \infty$.

Exercise 6.34 (6.11). (Borel's Normal Number Theorem) Let $x \in [0,1]$ be chosen uniformly at random, and let $d_n(x) \in \{0,1,\ldots,9\}$ be the n-th digit in the decimal expansion of x. Prove that for almost every x,

$$\frac{1}{n}\#\{k \leq n : d_k(x) = j\} \to \frac{1}{10} \quad \text{for each } j \in \{0,1,\ldots,9\}.$$

Exercise 6.35 (6.25). (Monte Carlo Integration) Let $f : [0,1] \to \mathbb{R}$ be measurable with $\int_0^1 f(x)^2\,dx < \infty$. Let $(U_n)_{n\geq 1}$ be i.i.d. Uniform$[0,1]$. Show that $\frac{1}{n}\sum_{k=1}^n f(U_k) \to \int_0^1 f(x)\,dx$ a.s.

Exercise 6.36 (6.26). (Convergence of Random Series) Let $(X_n)_{n\geq 1}$ be independent with $X_n \sim$ Uniform$[-a_n, a_n]$. Show that $\sum_{n=1}^\infty X_n$ converges almost surely if and only if $\sum_{n=1}^\infty a_n^2 < \infty$.

Exercise 6.37 (6.18). Recall the following notation: for a sequence of events $\{A_n\}$, we define a new event $\{A_n \text{ i.o.}\}$,

$$\begin{aligned}\{A_n \text{ i.o.}\} = \limsup A_n &= \lim_{m\to\infty} \bigcup_{n\geq m} A_n \\ &= \bigcap_m \bigcup_{n\geq m} A_n.\end{aligned}$$

In words, $\{A_n \text{ i.o.}\}$ consists of outcomes ω that repeat infinitely often, that is, they are in infinitely many A_n.

Let $X_1, X_2, \ldots,$ are i.i.d. with $\mathbb{E}|X_i| = \infty$ and $S_n = X_1 + \ldots + X_n$, then

1. $\mathbb{P}\big(|X_n| \geq n \text{ i.o.}\big) = 1, \textit{and}$
2. $\mathbb{P}\Big(\lim \frac{S_n}{n} \text{ exists and } \in (-\infty,\infty)\Big) = 0.$

Exercise 6.38 (6.12). (Converse to SLLN) Let $(X_n)_{n\geq 1}$ be i.i.d. random variables. Suppose $\bar{X}_n \to c$ a.s. for some constant $c \in \mathbb{R}$. Prove that $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = c$.

Exercise 6.39 (Pairwise independence does not suffice for SLLN). Construct pairwise independent (but not mutually independent) random variables $(X_n)_{n\geq 1}$ with $\mathbb{E}[X_n] = 0$ and $\mathbb{V}\mathrm{ar}(X_n) = 1$ such that $\bar{X}_n \not\to 0$ almost surely.

Chapter 7

Characteristic Functions

Characteristic functions are one of the most powerful tools in probability theory. They provide a complete description of the distribution of a random variable and, crucially, transform convolutions into products—making them indispensable for studying sums of independent random variables.

7.1 Definition and Basic Properties

Definition 7.1 (Characteristic function)**.** For a random variable X, the **characteristic function** (ch.f.) is defined by

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\,\mathbb{E}[\sin(tX)], \quad t \in \mathbb{R}.$$

More generally, for a probability measure μ on $\mathbb{R}$, we define

$$\varphi_\mu(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x).$$

Note that characteristic functions always exist: since $|e^{itx}| = 1$ for all $t, x \in \mathbb{R}$, the integral is always finite.

Proposition 7.2 (Basic properties)**.** *Let φ be the characteristic function of a random variable X. Then:*

1. $\varphi(0) = 1$.
2. $|\varphi(t)| \le 1$ *for all* $t \in \mathbb{R}$.
3. $\varphi(-t) = \overline{\varphi(t)}$ *for all* $t \in \mathbb{R}$ *(Hermitian symmetry).*

These properties follow immediately from the definition: (1) $\varphi(0) = \mathbb{E}[e^0] = 1$; (2) $|\varphi(t)| = |\mathbb{E}[e^{itX}]| \le \mathbb{E}[|e^{itX}|] = 1$; (3) $\varphi(-t) = \mathbb{E}[e^{-itX}] = \mathbb{E}[\overline{e^{itX}}] = \overline{\mathbb{E}[e^{itX}]} = \overline{\varphi(t)}$.

Theorem 7.1 (Uniform continuity)**.** *Every characteristic function is uniformly continuous on* $\mathbb{R}$.

Proof. For any $s, t \in \mathbb{R}$, we have

$$\begin{aligned} |\varphi(t) - \varphi(s)| &= |\mathbb{E}[e^{itX} - e^{isX}]| \\ &\leq \mathbb{E}[|e^{itX} - e^{isX}|] \\ &= \mathbb{E}[|e^{isX}| \cdot |e^{i(t-s)X} - 1|] \\ &= \mathbb{E}[|e^{i(t-s)X} - 1|]. \end{aligned}$$

Now, $|e^{i\theta} - 1| = 2|\sin(\theta/2)| \leq \min(2, |\theta|)$ for all $\theta \in \mathbb{R}$. Thus $|e^{i(t-s)X} - 1| \leq 2$ and $|e^{i(t-s)X} - 1| \to 0$ as $t - s \to 0$ for each fixed X. By the Dominated Convergence Theorem,

$$\mathbb{E}[|e^{i(t-s)X} - 1|] \to 0 \quad \text{as } t - s \to 0.$$

Since the bound depends only on $t - s$ (not on s itself), the convergence is uniform. □

Proposition 7.3 (Real-valued characteristic functions)**.** *A characteristic function φ is real-valued if and only if X has a symmetric distribution, i.e., $X \overset{d}{=} -X$.*

Proof. If $X \overset{d}{=} -X$, then

$$\varphi(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[e^{it(-X)}] = \mathbb{E}[e^{-itX}] = \overline{\varphi(t)},$$

so $\varphi(t)$ equals its own conjugate and is therefore real.

Conversely, if $\varphi(t)$ is real for all t, then $\varphi(t) = \overline{\varphi(t)} = \varphi(-t)$. By the uniqueness theorem for characteristic functions (Theorem 7.4), this implies $X \overset{d}{=} -X$. □

Proposition 7.4 (Affine transformation)**.** *If $Y = aX + b$ for constants $a, b \in \mathbb{R}$, then*

$$\varphi_Y(t) = e^{itb}\varphi_X(at).$$

Proof. Direct computation:

$$\varphi_Y(t) = \mathbb{E}[e^{itY}] = \mathbb{E}[e^{it(aX+b)}] = e^{itb}\mathbb{E}[e^{i(at)X}] = e^{itb}\varphi_X(at).$$

□

Theorem 7.2 (Product formula for independent random variables)**.** *If X and Y are independent random variables, then*

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t).$$

More generally, if $X_1, \ldots, X_n$ are independent, then

$$\varphi_{X_1+\cdots+X_n}(t) = \prod_{k=1}^{n} \varphi_{X_k}(t).$$

Proof. Since X and Y are independent, so are e^{itX} and e^{itY} (as measurable functions of independent random variables). Therefore,

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}] \cdot \mathbb{E}[e^{itY}] = \varphi_X(t) \cdot \varphi_Y(t).$$

The general case follows by induction. □

Remark 7.5. The product formula converts the convolution of distributions into ordinary multiplication of characteristic functions. This is the key computational advantage of working with characteristic functions.

Examples

Example 7.6 (Rademacher/Symmetric Bernoulli/ Coin flips). For the Rademacher random variable with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$, we have

$$\varphi(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t.$$

Note that φ is real-valued, consistent with the symmetric distribution.

Example 7.7 (Poisson). For the Poisson distribution with mean $\lambda > 0$, we have

$$\varphi(t) = \sum_{k=0}^{\infty} e^{itk} \cdot \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.$$

Example 7.8 (Gaussian). The normal distribution $N(\mu, \sigma^2)$ has characteristic function

$$\varphi(t) = e^{i\mu t - \sigma^2 t^2/2}.$$

Proof. By the affine transformation property, it suffices to prove the result for $N(0, 1)$. We compute:

$$\begin{aligned}\varphi(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2 + itx}\, dx.\end{aligned}$$

Completing the square in the exponent: $-x^2/2 + itx = -\frac{1}{2}(x - it)^2 - t^2/2$. Thus

$$\varphi(t) = e^{-t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-it)^2/2}\, dx.$$

The integral equals $\sqrt{2\pi}$ by a standard contour integration argument (shifting the contour of integration in the complex plane), giving $\varphi(t) = e^{-t^2/2}$. (For details, see Section G.1.)

For $N(\mu, \sigma^2)$, if $Z \sim N(0, 1)$ then $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$, so

$$\varphi_X(t) = e^{i\mu t}\varphi_Z(\sigma t) = e^{i\mu t} e^{-\sigma^2 t^2/2} = e^{i\mu t - \sigma^2 t^2/2}.$$

□

Example 7.9 (Uniform). For the uniform distribution on $[a, b]$,

$$\varphi(t) = \frac{1}{b-a}\int_a^b e^{itx}\,dx = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

In particular, for the uniform distribution on $[-1, 1]$:

$$\varphi(t) = \frac{e^{it} - e^{-it}}{2it} = \frac{\sin t}{t}.$$

Example 7.10 (Exponential). For the exponential distribution with rate $\lambda = 1$ (density e^{-x} on $[0, \infty)$),

$$\varphi(t) = \int_0^\infty e^{itx}e^{-x}\,dx = \int_0^\infty e^{-(1-it)x}\,dx = \frac{1}{1-it}.$$

Note that $|\varphi(t)| = 1/\sqrt{1+t^2} < 1$ for $t \neq 0$, and φ is not real-valued, reflecting the asymmetric distribution.

Lemma 7.11 (Characteristic function of a mixture)**.** *If random variables with distribution functions $F_1, \ldots, F_n$ have characteristic functions $\varphi_1, \ldots, \varphi_n$, and $\lambda_1, \ldots, \lambda_n \geq 0$ satisfy $\lambda_1 + \cdots + \lambda_n = 1$, then the distribution with cdf $\sum_{i=1}^n \lambda_i F_i$ has characteristic function $\sum_{i=1}^n \lambda_i \varphi_i$.*

Proof. This follows from linearity of integration:

$$\int e^{itx}\,d\Big(\sum_{i=1}^n \lambda_i F_i\Big)(x) = \sum_{i=1}^n \lambda_i \int e^{itx}\,dF_i(x) = \sum_{i=1}^n \lambda_i \varphi_i(t).$$

□

Example 7.12 (Bilateral exponential (Laplace distribution)). Consider a random variable with density $\frac{1}{2}e^{-|x|}$ on $\mathbb{R}$. This can be viewed as a 50-50 mixture of an exponential on $[0, \infty)$ and its reflection on $(-\infty, 0]$. Using Lemma 7.11 and the affine transformation property:

$$\varphi(t) = \frac{1}{2}\cdot\frac{1}{1-it} + \frac{1}{2}\cdot\frac{1}{1+it} = \frac{1}{2}\cdot\frac{(1+it)+(1-it)}{(1-it)(1+it)} = \frac{1}{1+t^2}.$$

Note that φ is real-valued, consistent with the symmetric density.

7.2 Moments and Derivatives

The derivatives of the characteristic function at the origin encode the moments of the distribution.

Theorem 7.3 (Moments and derivatives)**.** *Suppose $\mathbb{E}[|X|^n] < \infty$ for some positive integer n. Then φ is n times differentiable, and*

$$\varphi^{(k)}(t) = \mathbb{E}[(iX)^k e^{itX}] \quad \text{for } k = 1, \ldots, n.$$

In particular, $\varphi^{(k)}(0) = i^k \mathbb{E}[X^k]$.

Proof. We prove differentiability and the formula by induction. For the base case, consider

$$\frac{\varphi(t+h)-\varphi(t)}{h} = E\left[\frac{e^{i(t+h)X}-e^{itX}}{h}\right] = E\left[e^{itX}\cdot\frac{e^{ihX}-1}{h}\right].$$

As $h \to 0$, the integrand converges pointwise to $e^{itX}\cdot iX$. To apply Dominated Convergence, we need a bound. Using the inequality $|e^{i\theta}-1| \le |\theta|$:

$$\left|e^{itX}\cdot\frac{e^{ihX}-1}{h}\right| \le \frac{|hX|}{|h|} = |X|,$$

which is integrable by assumption. Thus $\varphi'(t) = \mathbb{E}[iXe^{itX}]$.

The higher derivatives follow by repeating the argument, using $\mathbb{E}[|X|^k] < \infty$ for $k \le n$. □

Corollary 7.13 (Taylor expansion)**.** *If $\mathbb{E}[X^2] < \infty$, then as $t \to 0$:*

$$\varphi(t) = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + o(t^2).$$

If $\mathbb{E}[X] = 0$ (centered random variable), this simplifies to

$$\varphi(t) = 1 - \frac{t^2}{2}\mathbb{E}[X^2] + o(t^2) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2),$$

where $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$.

Proof using Theorem 7.3. By Theorem 7.3, the assumption $\mathbb{E}[X^2] < \infty$ implies that φ is twice differentiable with $\varphi(0) = 1$, $\varphi'(0) = i\mathbb{E}[X]$, and $\varphi''(0) = -\mathbb{E}[X^2]$. The result then follows from Taylor's theorem with Peano remainder:

Theorem 1 (Taylor's theorem with Peano remainder)**.** *Let $f : \mathbb{R} \to \mathbb{C}$ be n times differentiable at a point $a \in \mathbb{R}$. Then*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + o((x-a)^n) \quad \text{as } x \to a.$$

Note that this only requires f to be n times differentiable at the single point a—we do not need $f^{(n)}$ to exist in a neighborhood of a or to be continuous.

Applying this with $n = 2$, $a = 0$, and $f = \varphi$:

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \frac{\varphi''(0)}{2}t^2 + o(t^2) = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + o(t^2).$$

□

Remark 7.14. We can also prove this Corollary directly using the Taylor expansion of $e^{i\theta}$ and Dominated Convergence. The proof is given in Appendix G.5. This approach provides explicit error bounds that are useful for proving the Central Limit Theorem.

Remark 7.15. The Taylor expansion in Corollary 7.13 is the key tool for proving the Central Limit Theorem via characteristic functions. The expansion shows that for small t, the characteristic function of a centered random variable with variance σ^2 is approximately $1 - \sigma^2 t^2/2$, which is close to $e^{-\sigma^2 t^2/2}$—the characteristic function of $N(0, \sigma^2)$.

7.3 Inversion Formula and Uniqueness

For absolutely continuous measures, we can write

$$\varphi(t) = \int e^{itx} f(x)\, dx,$$

where $f(x)$ is the density of measure μ. This is the Fourier transform of the density $f(x)$ and, when φ is integrable, we can apply the inverse Fourier transform to obtain

$$f(x) = \frac{1}{2\pi} \int e^{-itx} \varphi(t)\, dt.$$

Then, one can write

$$\begin{aligned} F(b) - F(a) &= \frac{1}{2\pi} \int_a^b dx \int e^{-itx} \varphi(t)\, dt = \frac{1}{2\pi} \int \varphi(t)\, dt \int_a^b e^{-itx}\, dx \\ &= \frac{1}{2\pi} \int \varphi(t) \frac{e^{-itb} - e^{-ita}}{-it}\, dt \\ &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \varphi(t) \frac{e^{-itb} - e^{-ita}}{-it}\, dt \end{aligned}$$

It turns out that in its last form this **inversion formula** holds also for general probability measures μ (not necessarily absolutely continuous), provided that a and b are continuity points of the distribution function $F(x)$.

More generally, we have the following theorem:

Theorem 7.16 (Inversion Formula)**.** *For all $a < b$,*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \varphi(t) \frac{e^{-itb} - e^{-ita}}{-it}\, dt = \mu(a,b) + \frac{1}{2}\Big(\mu(\{a\}) + \mu(\{b\})\Big).$$

The proof is in Appendix G.2.

Theorem 7.4 (Uniqueness)**.** *If two probability measures μ and ν have the same characteristic function, then $\mu = \nu$.*

Proof. The inversion formula determines $F(b) - F(a)$ for all continuity points $a < b$. Since the set of discontinuities of any distribution function is at most countable, and the continuity points are dense, the distribution function F is uniquely determined. Hence $\mu = \nu$. □

Corollary 7.17 (Density recovery)**.** *If the characteristic function φ is integrable, i.e., $\int_{\mathbb{R}} |\varphi(t)|\, dt < \infty$, then the distribution has a bounded continuous density given by*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, dt.$$

Proof. Define $f(x)$ by the formula above. Since $|\varphi(t)|$ is integrable, f is well-defined and continuous (by Dominated Convergence). To verify this is indeed the density, one checks that $\int_a^b f(x)\,dx = F(b) - F(a)$ using Fubini's theorem and the inversion formula. The details are left as an exercise. □

Example 7.18. For the bilateral exponential (Laplace) distribution with $\varphi(t) = 1/(1+t^2)$, we have $\int_{\mathbb{R}} |\varphi(t)|\,dt = \pi < \infty$. The density recovery formula gives

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx}}{1+t^2}\,dt = \frac{1}{2}e^{-|x|},$$

which can be verified by contour integration.

7.4 Lévy's Continuity Theorem

The following theorem is the key tool connecting characteristic functions to convergence in distribution. It will be essential for proving the Central Limit Theorem.

Theorem 7.5 (Lévy's continuity theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables with characteristic functions $(\varphi_n)_{n\geq 1}$.*

1. *If $X_n \xrightarrow{d} X$ for some random variable X with characteristic function φ, then $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.*

2. *Conversely, if $\varphi_n(t) \to \psi(t)$ for all $t \in \mathbb{R}$, where ψ is continuous at $t = 0$, then there exists a random variable X with characteristic function ψ, and $X_n \xrightarrow{d} X$.*

Remark 7.19. The continuity assumption at $t = 0$ in part (2) is essential. Without it, the pointwise limit of characteristic functions need not be a characteristic function. See the exercises for a counterexample.

Proof. **Part (1): Convergence in distribution implies pointwise convergence of CFs.**

For each fixed t, the functions $x \mapsto \cos(tx)$ and $x \mapsto \sin(tx)$ are bounded and continuous. By the definition of convergence in distribution,

$$\varphi_n(t) = \mathbb{E}[\cos(tX_n)] + i\,\mathbb{E}[\sin(tX_n)] \to \mathbb{E}[\cos(tX)] + i\,\mathbb{E}[\sin(tX)] = \varphi(t).$$

Part (2): Pointwise convergence to a function continuous at 0 implies convergence in distribution.

This direction is more involved and proceeds in three steps. Let F_n denote the distribution function of X_n.

Step 1: The sequence (X_n) is tight.

Recall that (X_n) is **tight** if for every $\varepsilon > 0$, there exists $M > 0$ such that $\mathbb{P}(|X_n| \leq M) \geq 1 - \varepsilon$ for all n.

We use the following lemma:

Lemma 7.20. *For any random variable X with characteristic function φ,*

$$\mathbb{P}(|X| > M) \le \frac{M}{2} \int_{-2/M}^{2/M} (1 - \operatorname{Re} \varphi(t))\, dt.$$

Proof of Lemma. We compute:

$$\begin{aligned}
\int_{-2/M}^{2/M} (1 - \operatorname{Re} \varphi(t))\, dt &= \int_{-2/M}^{2/M} \mathbb{E}[1 - \cos(tX)]\, dt \\
&= E\left[\int_{-2/M}^{2/M} (1 - \cos(tX))\, dt\right] \\
&= E\left[\frac{4}{M} - \frac{2\sin(2X/M)}{X}\right] \\
&= \frac{4}{M} E\left[1 - \frac{\sin(2X/M)}{2X/M}\right],
\end{aligned}$$

where the interchange is justified by Fubini (the integrand is bounded).

Now we use the key observation: for all $u \in \mathbb{R}$,

$$1 - \frac{\sin u}{u} \ge 0,$$

and for $|u| \ge 2$,

$$1 - \frac{\sin u}{u} \ge 1 - \frac{|\sin u|}{|u|} \ge 1 - \frac{1}{2} = \frac{1}{2}.$$

Setting $u = 2X/M$, we see that $|X| > M$ implies $|u| > 2$, and hence $1 - \frac{\sin u}{u} \ge \frac{1}{2}$. Therefore,

$$\frac{4}{M} E\left[1 - \frac{\sin(2X/M)}{2X/M}\right] \ge \frac{4}{M} E\left[\frac{1}{2} \cdot \mathbf{1}_{|X|>M}\right] = \frac{2}{M}\mathbb{P}(|X| > M),$$

which rearranges to the claimed inequality. □

Now we prove tightness. We have $\psi(0) = \lim_n \varphi_n(0) = 1$. Since ψ is continuous at 0, for any $\varepsilon > 0$ there exists $\delta > 0$ such that $|t| \le \delta$ implies $|\psi(t) - 1| < \varepsilon$.

We claim this gives $\operatorname{Re} \psi(t) > 1 - \varepsilon$ for $|t| \le \delta$. Indeed, since ψ is a pointwise limit of characteristic functions and $|\varphi_n(t)| \le 1$ for all n, we have $|\psi(t)| \le 1$, hence $\operatorname{Re} \psi(t) \le 1$. Therefore $1 - \operatorname{Re} \psi(t) \ge 0$. Now,

$$1 - \operatorname{Re} \psi(t) = |1 - \operatorname{Re} \psi(t)| \le |1 - \psi(t)| < \varepsilon,$$

which gives $\operatorname{Re} \psi(t) > 1 - \varepsilon$.

Set $M = 2/\delta$, so that $2/M = \delta$. In order to apply Lemma 7.20 we need to bound the integral

$$\int_{-\delta}^{\delta} (1 - \operatorname{Re} \varphi_n(t))\, dt.$$

For each fixed t, we have $\varphi_n(t) \to \psi(t)$, hence $1 - \operatorname{Re}\varphi_n(t) \to 1 - \operatorname{Re}\psi(t)$. Since $|1 - \operatorname{Re}\varphi_n(t)| \leq 2$ for all n and t, the Dominated Convergence Theorem gives

$$\int_{-\delta}^{\delta} (1 - \operatorname{Re}\varphi_n(t))\, dt \to \int_{-\delta}^{\delta} (1 - \operatorname{Re}\psi(t))\, dt < 2\delta \cdot \varepsilon.$$

Thus for all n sufficiently large,

$$\int_{-\delta}^{\delta} (1 - \operatorname{Re}\varphi_n(t))\, dt < 4\delta\varepsilon.$$

By Lemma 7.20 with $M = 2/\delta$:

$$\mathbb{P}(|X_n| > M) \leq \frac{M}{2} \int_{-2/M}^{2/M} (1 - \operatorname{Re}\varphi_n(t))\, dt = \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \operatorname{Re}\varphi_n(t))\, dt < \frac{4\delta\varepsilon}{\delta} = 4\varepsilon.$$

Since ε is arbitrary, (X_n) is tight.

Step 2: Every subsequential limit has characteristic function ψ.

We use the following classical result:

Theorem 7.6 (Helly's selection theorem)**.** *Let (F_n) be a sequence of distribution functions. Then there exists a subsequence (F_{n_k}) and a right-continuous nondecreasing function F with $0 \leq F(x) \leq 1$ for all x, such that $F_{n_k}(x) \to F(x)$ at every continuity point of F.*

The proof uses a diagonalization argument: first extract a subsequence converging at all rationals, then extend to all continuity points. The limiting function F is nondecreasing and right-continuous, but may not be a proper distribution function—we might have $\lim_{x\to-\infty} F(x) > 0$ or $\lim_{x\to+\infty} F(x) < 1$ (mass escaping to $\pm\infty$).

However, tightness rules this out: if (X_n) is tight, then for any $\varepsilon > 0$ there exists M such that $F_n(M) - F_n(-M) \geq 1 - \varepsilon$ for all n. Passing to the limit along the subsequence (at continuity points), $F(M) - F(-M) \geq 1 - \varepsilon$. Since ε is arbitrary, F is a proper distribution function.

Now, suppose $F_{n_k}(x) \to F(x)$ at continuity points, where F is the distribution function of some random variable Y. This means $X_{n_k} \xrightarrow{d} Y$. By Part (1), the characteristic function of Y is

$$\varphi_Y(t) = \lim_{k\to\infty} \varphi_{n_k}(t) = \psi(t).$$

Step 3: The limit is unique.

By the uniqueness theorem (Theorem 7.4), any two random variables with characteristic function ψ have the same distribution. Thus all subsequential limits agree in distribution, and the full sequence converges: $X_n \xrightarrow{d} X$ where X is a random variable with characteristic function ψ. □

Remark 7.21. The proof reveals why continuity at 0 is crucial: it provides the tightness estimate via Lemma 7.20. Without tightness, mass could escape to infinity, and the limiting object might not correspond to a proper random variable.

Corollary 7.22. *Let (X_n) be a sequence of random variables with characteristic functions (φ_n). If $\varphi_n(t) \to \varphi(t)$ for all t, where φ is the characteristic function of a random variable X, then $X_n \xrightarrow{d} X$.*

Proof. Since φ is a characteristic function, it is continuous (in particular, continuous at 0). The result follows from Theorem 7.5. □

7.5 Exercises

Homework

Exercise 7.23 (A limit theorem via characteristic functions). Let $X_n \sim \text{Binomial}(n, \lambda/n)$ for a fixed $\lambda > 0$ and $n \geq \lambda$.

(a) Calculate the characteristic function of X_n.

(b) Find $\lim_{n\to\infty} \varphi_{X_n}(t)$ for all $t \in \mathbb{R}$.

(c) Identify the limit as the characteristic function of a random variable from a well-known family of distributions, and use Lévy's continuity theorem to conclude that $X_n \xrightarrow{d} X$, where X is …

Exercise 7.24 (Which are characteristic functions?). For each of the following functions $\varphi : \mathbb{R} \to \mathbb{R}$, determine whether φ is the characteristic function of some random variable. Justify your answer.

(a) $\varphi(t) = e^{-t^4}$.

(b) $\varphi(t) = e^{-|t|}$.

(c) $\varphi(t) = \cos^2(t)$.

Exercise 7.25 (Inversion formula for integer-valued random variables). Let X be an integer-valued random variable with characteristic function φ.

(a) Show that for every $k \in \mathbb{Z}$,

$$\mathbb{P}(X = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt}\, \varphi(t)\, dt.$$

(b) Let $X \sim \text{Poisson}(\lambda)$. Using $\varphi(t) = e^{\lambda(e^{it}-1)}$ and the formula from (a), verify that $\mathbb{P}(X = 0) = e^{-\lambda}$.

(c) Let X_1, X_2 be independent with $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$. Using characteristic functions and the uniqueness theorem, find the distribution of $X_1 + X_2$.

Exercise 7.26 (Identifying a distribution via characteristic functions). Let X and Y be independent random variables such that $X + Y \sim N(\mu, \sigma^2)$ and $X \sim N(\mu_1, \sigma_1^2)$.

(a) Using the product formula for characteristic functions, find $\varphi_Y(t)$ and identify the distribution of Y.

(b) Show that the condition $\sigma^2 \geq \sigma_1^2$ is automatic: it is impossible to have $\sigma_1^2 > \sigma^2$ under the given hypotheses.

(c) Let $Z_1, \ldots, Z_n$ be i.i.d. $N(\mu, \sigma^2)$. Using characteristic functions, find the distribution of $\bar{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$.

Exercise 7.27 (Convergence to a constant via characteristic functions). Let $(X_n)_{n\geq 1}$ be a sequence of random variables with characteristic functions $(\varphi_n)_{n\geq 1}$.

(a) Show that $X_n \xrightarrow{\mathbb{P}} 0$ if and only if $\varphi_n(t) \to 1$ for all $t \in \mathbb{R}$.

(b) Suppose instead that $\varphi_n(t) \to 1$ only for all t in some interval $(-\delta, \delta)$ with $\delta > 0$. Must $X_n \xrightarrow{\mathbb{P}} 0$? Prove or give a counterexample.

Additional Exercises

Exercise 7.28 (Triangular distribution via convolution). Let U_1, U_2 be independent random variables, each uniformly distributed on $[-1/2, 1/2]$.

(a) Compute the characteristic function of U_1.

(b) Using the product formula, find the characteristic function of $X = U_1 + U_2$.

(c) Show that X has the triangular density $f(x) = 1 - |x|$ on $(-1, 1)$ and zero otherwise.

Exercise 7.29 (Cauchy distribution). The Cauchy distribution has density $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

(a) Show that $\mathbb{E}[|X|] = \infty$ for a Cauchy random variable X.

(b) Using contour integration, show that the characteristic function is $\varphi(t) = e^{-|t|}$.

(c) Using the product formula, show that if X_1, X_2 are independent Cauchy, then $\frac{X_1+X_2}{2}$ is also Cauchy.

Exercise 7.30 (Necessity of continuity at zero in Lévy's theorem). Let X_n be uniformly distributed on $[-n, n]$.

(a) Show that $\varphi_n(t) = \frac{\sin(nt)}{nt}$.

(b) Show that $\varphi_n(t) \to \psi(t)$ for all $t \in \mathbb{R}$, where

$$\psi(t) = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{if } t \neq 0. \end{cases}$$

(c) Verify that ψ is not continuous at $t = 0$.

(d) Explain why X_n does not converge in distribution to any random variable.

Exercise 7.31 (Characteristic function determines the distribution).
(Challenging) Let X and Y be random variables with $\varphi_X(t) = \varphi_Y(t)$ for all t in some interval $(-\delta, \delta)$ with $\delta > 0$. Must X and Y have the same distribution? Prove or give a counterexample.

Exercise 7.32 (Symmetric random variables). Let X be a random variable with characteristic function φ.

(a) Define $X' = -X$. Express $\varphi_{X'}(t)$ in terms of $\varphi_X(t)$.

(b) Suppose X' is an independent copy of X (i.e., X' is independent of X and has the same distribution). Show that $Y = X - X'$ has a real-valued characteristic function.

(c) Give an example where X does not have a symmetric distribution, but $Y = X - X'$ does.

Exercise 7.33 (Gamma distribution). The Gamma distribution $\Gamma(\alpha, \lambda)$ with shape $\alpha > 0$ and rate $\lambda > 0$ has density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

(a) Show that the characteristic function is $\varphi(t) = \left(\frac{\lambda}{\lambda - it}\right)^\alpha$.

(b) Using characteristic functions, show that if $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$ are independent, then $X + Y \sim \Gamma(\alpha + \beta, \lambda)$.

(c) The exponential distribution with rate λ is $\Gamma(1, \lambda)$. Use (b) to show that the sum of n independent $\mathrm{Exp}(\lambda)$ random variables has distribution $\Gamma(n, \lambda)$.

Exercise 7.34 (Moments from characteristic functions). Let X be a random variable with $\mathbb{E}[X^4] < \infty$ and characteristic function φ.

(a) Express $\mathbb{E}[X]$, $\mathbb{E}[X^2]$, $\mathbb{E}[X^3]$, and $\mathbb{E}[X^4]$ in terms of derivatives of φ at $t = 0$.

(b) Verify your formulas for $X \sim N(0, 1)$, using $\varphi(t) = e^{-t^2/2}$.

Exercise 7.35 (Non-negative characteristic functions).
(Challenging) Suppose $\varphi(t) \geq 0$ for all $t \in \mathbb{R}$. Show that $\varphi(t) = |\psi(t)|^2$ for some characteristic function ψ.

Exercise 7.36 (Pólya's criterion).
(Challenging) A function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to satisfy **Pólya's criterion** if:

- φ is continuous,
- $\varphi(0) = 1$,
- $\varphi(t) = \varphi(-t)$ (symmetry),
- φ is convex on $[0, \infty)$,

- $\lim_{t\to\infty} \varphi(t) = 0$.

(a) Verify that $\varphi(t) = \max(1 - |t|, 0)$ satisfies Pólya's criterion.

(b) Show that this φ is the characteristic function of the distribution with density $f(x) = \frac{1-\cos x}{\pi x^2}$.

Remark: Pólya's theorem states that any function satisfying Pólya's criterion is a characteristic function. The proof is beyond our scope, but this provides a useful sufficient condition.

Exercise 7.37 (Tightness and characteristic functions).
(Challenging) Let (X_n) be a sequence of random variables with characteristic functions (φ_n). Show that if (X_n) is tight and $\varphi_n(t) \to \varphi(t)$ for all t in a dense subset of $\mathbb{R}$, then $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.

Exercise 7.38 (Characteristic functions and absolute continuity).
(Challenging)

(a) Show that if $\varphi \in L^1(\mathbb{R})$ (i.e., $\int_{-\infty}^{\infty} |\varphi(t)|\, dt < \infty$), then the corresponding distribution has a bounded continuous density.

(b) Give an example of a distribution with a continuous density whose characteristic function is not in $L^1(\mathbb{R})$.

(c) Give an example of a characteristic function $\varphi \in L^1(\mathbb{R})$ such that the corresponding density is not differentiable everywhere.

Exercise 7.39 (Convergence to a point mass). Let (X_n) be a sequence of random variables with characteristic functions (φ_n).

(a) Show that $X_n \xrightarrow{d} c$ (a constant) if and only if $\varphi_n(t) \to e^{itc}$ for all $t \in \mathbb{R}$.

(b) Show that $X_n \xrightarrow{\mathbb{P}} c$ if and only if $\varphi_n(t) \to e^{itc}$ for all $t \in \mathbb{R}$.

Chapter 8

Central Limit Theorems

he Central Limit Theorem is one of the most celebrated results in probability theory. It explains why the normal distribution appears so frequently in nature: whenever a random quantity arises as the sum of many small, independent contributions, its distribution is approximately Gaussian, regardless of the distributions of the individual terms.

8.1 The Classical CLT

We begin with the i.i.d. case, proved using the characteristic function machinery developed in the previous chapter.

Theorem 8.1 (Classical CLT)**.** *Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Proof. Without loss of generality, we may assume $\mu = 0$ (replace X_i by $X_i - \mu$). Let $\varphi(t) = \mathbb{E}e^{itX_1}$ be the characteristic function of X_1. By independence, the characteristic function of $S_n/(\sigma\sqrt{n})$ is

$$\varphi_n(t) := \mathbb{E}\exp\left(it\frac{S_n}{\sigma\sqrt{n}}\right) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

By the Taylor expansion of characteristic functions (Corollary 7.13),

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2) \quad \text{as } t \to 0.$$

Hence, for fixed t as $n \to \infty$,

$$\varphi_n(t) = \left[1 - \frac{\sigma^2 t^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \to e^{-t^2/2}.$$

The convergence in the last step uses the fact that if z_n are complex numbers with $z_n \to z$, then $(1 + z_n/n)^n \to e^z$ (see Lemma H.1 in the Appendix, or Theorem 3.4.2 in Durrett).

Since $e^{-t^2/2}$ is the characteristic function of $N(0,1)$, Lévy's continuity theorem (Theorem 7.5) implies $S_n/(\sigma\sqrt{n}) \xrightarrow{d} N(0,1)$. □

Remark 8.2. Theorem 2.4.1 in Durrett shows that pairwise independence suffices for the Strong Law of Large Numbers. It is noteworthy that pairwise independence is **not** sufficient for the CLT; see Example 3.4.9 in Durrett for a counterexample.

Remark 8.3 (Two approaches to CLT). There are two classical approaches to proving Central Limit Theorems:

1. The **characteristic function method**: show that the characteristic functions of the normalized sums converge to $e^{-t^2/2}$, then apply Lévy's continuity theorem. This is the approach used above.

2. **Lindeberg's swapping method**: use smooth test functions $f \in \mathbf{C}_b^3(\mathbb{R})$ and a telescoping argument that replaces summands one by one with matched Gaussian random variables. This powerful method does not require characteristic functions and yields the more general Lindeberg CLT (Theorem 8.6).

We will develop the second approach in the sections that follow.

8.2 Triangular Arrays

If we study the sums of random variables which are independent but not necessarily identically distributed, then the language of triangular arrays is useful. Throughout this section we shall study the sequence of sums

$$S_i = \sum_{j=1}^{n_i} X_{ij}$$

obtained by summing the rows of a **triangular array** of random variables

$$\begin{array}{l} X_{11}, X_{12}, \ldots, X_{1n_1} \\ X_{21}, X_{22}, \ldots\ldots, X_{2n_2} \\ X_{31}, X_{32}, \ldots\ldots\ldots, X_{3n_3} \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \end{array}$$

It will be assumed throughout that the triangular arrays we consider satisfy **three standard conditions**:

1. For each i, the n_i random variables $X_{i1}, X_{i2}, \ldots, X_{in_i}$ in the ith row are mutually independent.

2. $\mathbb{E}(X_{ij}) = 0$ for all i, j.

3. $\sum_{j=1}^{n_i} \mathbb{E}X_{ij}^2 = 1$ for all i.

Some remarks about these conditions:

- It is **not** assumed that random variables within each row are identically distributed.
- It is **not** assumed that different rows are independent. In fact, they often are not: a common application arises from the study of partial sums $S_n = X_1 + X_2 + \ldots + X_n$ of a fixed sequence, where $X_{ij} := X_j/s_n$ and $s_n^2 = \sum_{j=1}^{n} \mathbb{V}\mathrm{ar}(X_j)$.
- It will usually be the case that $n_i \to \infty$ as $i \to \infty$.

8.3 The Lindeberg Condition

Definition 8.4 (Lindeberg condition)**.** A triangular array satisfying the three standard conditions is said to satisfy the **Lindeberg condition** if

$$\forall \epsilon > 0, \quad \lim_{i\to\infty} \sum_{j=1}^{n_i} \mathbb{E}\big[X_{ij}^2 \mathbf{1}(|X_{ij}| > \epsilon)\big] = 0. \tag{8.1}$$

The Lindeberg condition makes precise in what sense the individual summands must be negligible relative to the sum for the CLT to hold. It says that for arbitrarily small fixed $\epsilon > 0$, the contributions to the total variance coming from values of $|X_{ij}|$ exceeding ϵ vanish as $i \to \infty$.

An important consequence is that the individual summand variances must themselves become uniformly negligible:

Lemma 8.5. *If a triangular array satisfies the three standard conditions and the Lindeberg condition* (8.1)*, then*

$$\lim_{i\to\infty} \max_{1\le j\le n_i} \mathbb{E}X_{ij}^2 = 0. \tag{8.2}$$

In particular, $n_i \to \infty$ as $i \to \infty$.

Proof. Fix $\epsilon > 0$. Since $X_{ij}^2 \le \epsilon^2 + X_{ij}^2 \mathbf{1}(|X_{ij}| > \epsilon)$, taking expectations and maximizing over j gives

$$\max_{1\le j\le n_i} \mathbb{E}X_{ij}^2 \le \epsilon^2 + \sum_{j=1}^{n_i} \mathbb{E}\big[X_{ij}^2 \mathbf{1}(|X_{ij}| > \epsilon)\big]. \tag{8.3}$$

Under the Lindeberg condition, the sum on the right vanishes as $i \to \infty$, so $\limsup_{i\to\infty} \max_j \mathbb{E}X_{ij}^2 \le \epsilon^2$. Since $\epsilon > 0$ was arbitrary, (8.2) follows. Together with $\sum_{j=1}^{n_i} \mathbb{E}X_{ij}^2 = 1$, this forces $n_i \to \infty$. □

Condition (8.2) is closely related to the **Feller condition** (uniform asymptotic negligibility), which we discuss after the statement of Lindeberg's theorem.

Theorem 8.6 (Lindeberg's CLT)**.** *Suppose that a triangular array satisfies the three standard conditions and the Lindeberg condition* (8.1)*. Then* $S_i \xrightarrow{d} N(0,1)$.

The proof, which refines the argument of Section 8.5, is given in Appendix H.2.

Definition 8.7 (Uniformly asymptotically negligible (UAN))**.** A triangular array is said to be **uniformly asymptotically negligible** if

$$\forall \epsilon > 0, \quad \lim_{i \to \infty} \max_{1 \le j \le n_i} \mathbb{P}(|X_{ij}| > \epsilon) = 0. \tag{8.4}$$

The Lindeberg condition implies UAN (by Chebyshev's inequality applied to each summand), but not conversely. UAN captures the idea that no single summand dominates the sum.

A beautiful converse to Lindeberg's theorem is due to Feller:

Theorem 8.8 (Feller's converse)**.** *If a triangular array satisfies the three standard conditions and is UAN, then* $S_i \xrightarrow{d} N(0,1)$ *if and only if the Lindeberg condition* (8.1) *holds.*

Proof. See Billingsley, Theorem 27.4, or Kallenberg, 5.12. □

Thus, under the mild UAN assumption, the Lindeberg condition is both necessary and sufficient for the CLT. This makes the Lindeberg condition the essentially sharp condition for Gaussian limits in the triangular array setting.

8.4 The Lyapunov Condition

A condition stronger—but often easier to verify—than Lindeberg's is the **Lyapunov condition**:

$$\exists\, \delta > 0 \text{ such that } \lim_{i \to \infty} \sum_{j=1}^{n_i} \mathbb{E}|X_{ij}|^{2+\delta} = 0. \tag{8.5}$$

Lemma 8.9. *Lyapunov's condition implies Lindeberg's condition.*

Proof. Fix $\epsilon > 0$. For any random variable X with $|X| > \epsilon$, we have $X^2 = |X|^{2+\delta}/|X|^{\delta} \le |X|^{2+\delta}/\epsilon^{\delta}$. Therefore,

$$\mathbb{E}\big[X^2 \mathbf{1}(|X| > \epsilon)\big] \le \frac{\mathbb{E}|X|^{2+\delta}}{\epsilon^{\delta}}.$$

Summing over j and taking $i \to \infty$ gives

$$\sum_{j=1}^{n_i} \mathbb{E}\big[X_{ij}^2 \mathbf{1}(|X_{ij}| > \epsilon)\big] \le \frac{1}{\epsilon^{\delta}} \sum_{j=1}^{n_i} \mathbb{E}|X_{ij}|^{2+\delta} \to 0. \qquad \square$$

Theorem 8.10 (Lyapunov's CLT)**.** *If a triangular array satisfies the three standard conditions and the Lyapunov condition* (8.5)*, then* $S_i \xrightarrow{d} N(0,1)$.

This is an immediate corollary of Lindeberg's CLT (Theorem 8.6) and Lemma 8.9. However, we give a direct proof below for the case $\delta = 1$, as it illustrates the key ideas of the Lindeberg swapping method in a simpler setting.

8.5 Proof of Lyapunov's CLT via the Swapping Method

8.5.1 Preliminary facts

We prove Lyapunov's CLT using the Lindeberg swapping method and need two preliminary facts.

Lemma 8.11. *If $X \sim N(0,\sigma^2)$ and $Y \sim N(0,\tau^2)$ are independent, then $X+Y \sim N(0,\sigma^2+\tau^2)$.*

Proof. Either use the convolution formula for densities, or use characteristic functions: $\varphi_{X+Y}(t) = e^{-\sigma^2 t^2/2} \cdot e^{-\tau^2 t^2/2} = e^{-(\sigma^2+\tau^2)t^2/2}$. □

Lemma 8.12. *Let $Y_1, Y_2, \ldots$ and Y be real-valued random variables. Then $Y_i \xrightarrow{d} Y$ if and only if $\lim_{i\to\infty} \mathbb{E}f(Y_i) = \mathbb{E}f(Y)$ for all $f \in \mathbf{C}_b^3(\mathbb{R})$, the set of functions from $\mathbb{R}$ to $\mathbb{R}$ with three bounded continuous derivatives.*

Proof. ($\Rightarrow$) By definition, $Y_i \xrightarrow{d} Y$ means $\mathbb{E}f(Y_i) \to \mathbb{E}f(Y)$ for every $f \in C_b(\mathbb{R})$. Since $\mathbf{C}_b^3(\mathbb{R}) \subset C_b(\mathbb{R})$, the conclusion follows.

($\Leftarrow$) For each fixed $t \in \mathbb{R}$, the functions $x \mapsto \cos(tx)$ and $x \mapsto \sin(tx)$ belong to $\mathbf{C}_b^3(\mathbb{R})$ (indeed, they are bounded with bounded derivatives of all orders). By hypothesis,

$$\mathbb{E}[\cos(tY_i)] \to \mathbb{E}[\cos(tY)] \quad \text{and} \quad \mathbb{E}[\sin(tY_i)] \to \mathbb{E}[\sin(tY)].$$

Therefore $\varphi_{Y_i}(t) = \mathbb{E}[e^{itY_i}] \to \mathbb{E}[e^{itY}] = \varphi_Y(t)$ for every $t \in \mathbb{R}$. Since φ_Y is the characteristic function of Y, it is continuous (in particular at 0). By Lévy's continuity theorem (Theorem 7.5), $Y_i \xrightarrow{d} Y$. □

8.5.2 The swapping argument

For convenience, we restate the result we will prove. Recall that a triangular array satisfies the three standard conditions if the entries in each row are independent, centered, and have variances summing to 1, so that the row sum $S_i = \sum_{j=1}^{n_i} X_{ij}$ has mean 0 and variance 1.

Theorem 2 (Lyapunov's CLT for $\delta = 1$)**.** *Suppose a triangular array satisfies the three standard conditions and*

$$\lim_{i\to\infty} \sum_{j=1}^{n_i} \mathbb{E}|X_{ij}|^3 = 0. \tag{8.6}$$

Then $S_i \xrightarrow{d} N(0,1)$.

The idea is to compare the row sum S_i with a Gaussian random variable by replacing summands one at a time, using smooth test functions to measure the difference.

Proof. Fix a row of the triangular array. Write $X_1, \ldots, X_n$ for the entries (suppressing the row index), with $\mathbb{E}X_j = 0$, $\mathbb{E}X_j^2 = \sigma_j^2$, and $\sum_{j=1}^n \sigma_j^2 = 1$. The row sum is $S = X_1 + \cdots + X_n$.

Construct independent Gaussian variables $Z_j \sim N(0, \sigma_j^2)$, jointly independent of all X_j's. By Lemma 8.11, the Gaussian sum $T := Z_1 + \cdots + Z_n \sim N(0, 1)$.

Hybrid sums. Define $H_0, H_1, \ldots, H_n$ by replacing X's with Z's one at a time:

$$\begin{aligned} S = H_0 &:= X_1 + X_2 + X_3 + \ldots + X_n, \\ H_1 &:= Z_1 + X_2 + X_3 + \ldots + X_n, \\ H_2 &:= Z_1 + Z_2 + X_3 + \ldots + X_n, \\ &\vdots \\ T = H_n &:= Z_1 + Z_2 + Z_3 + \ldots + Z_n. \end{aligned}$$

In general, $H_j = Z_1 + \cdots + Z_j + X_{j+1} + \cdots + X_n$, so that H_{j-1} and H_j differ only in the jth slot:

$$H_{j-1} = R_j + X_j, \qquad H_j = R_j + Z_j,$$

where $R_j := Z_1 + \cdots + Z_{j-1} + X_{j+1} + \cdots + X_n$ collects the common terms. By construction, R_j is independent of both X_j and Z_j.

Telescoping. We want to show that $\mathbb{E}f(S)$ is close to $\mathbb{E}f(T)$ for all $f \in \mathbf{C}_b^3(\mathbb{R})$ with uniform bound K on f and its first three derivatives. By the triangle inequality,

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \le \sum_{j=1}^{n} |\mathbb{E}f(H_{j-1}) - \mathbb{E}f(H_j)|. \tag{8.7}$$

Bounding each swap. Taylor-expanding $f(R_j + X_j)$ and $f(R_j + Z_j)$ to third order:

$$\begin{aligned} f(R_j + X_j) &= f(R_j) + X_j f'(R_j) + \frac{X_j^2}{2} f''(R_j) + \frac{X_j^3}{6} f'''(\alpha_j), \\ f(R_j + Z_j) &= f(R_j) + Z_j f'(R_j) + \frac{Z_j^2}{2} f''(R_j) + \frac{Z_j^3}{6} f'''(\beta_j), \end{aligned}$$

where α_j lies between R_j and $R_j + X_j$ and β_j between R_j and $R_j + Z_j$.

Taking expectations and subtracting, the zeroth-order terms cancel trivially, and the first- and second-order terms cancel by matching moments: R_j is independent of X_j and Z_j, while $\mathbb{E}X_j = \mathbb{E}Z_j = 0$ and $\mathbb{E}X_j^2 = \mathbb{E}Z_j^2 = \sigma_j^2$. Only the third-order remainders survive:

$$|\mathbb{E}f(H_{j-1}) - \mathbb{E}f(H_j)| = \left| \mathbb{E}\frac{X_j^3}{6} f'''(\alpha_j) - \mathbb{E}\frac{Z_j^3}{6} f'''(\beta_j) \right| \le \frac{K}{6}\big(\mathbb{E}|X_j|^3 + \mathbb{E}|Z_j|^3\big). \tag{8.8}$$

(The expectation of the X_j^3 term exists by Lyapunov's condition (8.6), and the Z_j^3 term has finite moments of all orders.)

Controlling the Gaussian moments. Let $c := \mathbb{E}|Z|^3$ where $Z \sim N(0,1)$. Then $\mathbb{E}|Z_j|^3 = c\,\sigma_j^3$. Jensen's inequality gives $\sigma_j^3 = (\mathbb{E}X_j^2)^{3/2} \le \mathbb{E}|X_j|^3$, so $\mathbb{E}|Z_j|^3 \le c\,\mathbb{E}|X_j|^3$.

Summing (8.8) over j via (8.7):

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \le \frac{K(1+c)}{6} \sum_{j=1}^{n} \mathbb{E}|X_j|^3. \tag{8.9}$$

Conclusion. Restoring the row index i, the bound (8.9) reads

$$|\mathbb{E}f(S_i) - \mathbb{E}f(T)| \le \frac{K(1+c)}{6} \sum_{j=1}^{n_i} \mathbb{E}|X_{ij}|^3, \tag{8.10}$$

where $T \sim N(0,1)$ for every row (since $\sum_j \sigma_{ij}^2 = 1$). The right-hand side tends to zero by (8.6).

By Lemma 8.12, $S_i \xrightarrow{d} N(0,1)$ as $i \to \infty$. □

8.6 The Cramér–Wold Device

The results above concern convergence in distribution on $\mathbb{R}$. In many applications—multivariate statistics, random vectors, stochastic processes—one needs convergence in $\mathbb{R}^d$. The Cramér–Wold device reduces the d-dimensional problem to one-dimensional projections.

8.6.1 The Cramér–Wold theorem

Recall that $X_n \xrightarrow{d} X$ in $\mathbb{R}^d$ means $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for every bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$.

Theorem 8.13 (Cramér–Wold device)**.** *Let X_n and X be random vectors in $\mathbb{R}^d$. Then*

$$X_n \xrightarrow{d} X \quad \text{in } \mathbb{R}^d$$

if and only if

$$\langle t, X_n\rangle \xrightarrow{d} \langle t, X\rangle \quad \text{in } \mathbb{R}$$

for every $t \in \mathbb{R}^d$, where $\langle t, x\rangle = \sum_{k=1}^d t_k x_k$ denotes the standard inner product.

Proof. ($\Rightarrow$) If $X_n \xrightarrow{d} X$ in $\mathbb{R}^d$, then for any fixed $t \in \mathbb{R}^d$ the map $x \mapsto \langle t, x\rangle$ is continuous, so $\langle t, X_n\rangle \xrightarrow{d} \langle t, X\rangle$ by the continuous mapping theorem.

($\Leftarrow$) Assume $\langle t, X_n\rangle \xrightarrow{d} \langle t, X\rangle$ for every $t \in \mathbb{R}^d$. We show that the characteristic functions of X_n converge to that of X.

The characteristic function of X_n in $\mathbb{R}^d$ is

$$\varphi_{X_n}(t) = \mathbb{E}[e^{i\langle t, X_n\rangle}], \quad t \in \mathbb{R}^d.$$

This is simply the characteristic function of the real-valued random variable $\langle t, X_n\rangle$ evaluated at 1, i.e., $\varphi_{X_n}(t) = \varphi_{\langle t, X_n\rangle}(1)$.

By assumption, $\langle t, X_n\rangle \xrightarrow{d} \langle t, X\rangle$ for each fixed t. Lévy's continuity theorem (Theorem 7.5, part (1)) implies

$$\varphi_{\langle t,X_n\rangle}(s) \to \varphi_{\langle t,X\rangle}(s) \quad \text{for all } s \in \mathbb{R}.$$

Taking $s = 1$:

$$\varphi_{X_n}(t) \to \varphi_X(t) \quad \text{for all } t \in \mathbb{R}^d.$$

Since φ_X is the characteristic function of the random vector X, it is continuous (in particular at $t = 0$). By the multivariate version of Lévy's continuity theorem, $X_n \xrightarrow{d} X$ in $\mathbb{R}^d$. □

Remark 8.14. The power of Cramér–Wold is that it reduces a d-dimensional convergence question to infinitely many one-dimensional questions—which are often tractable via the one-dimensional CLT and characteristic function tools.

8.6.2 The Multivariate CLT

Theorem 8.15 (Multivariate CLT)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_1] = \mu \in \mathbb{R}^d$ and covariance matrix $\Sigma = \mathrm{Cov}(X_1)$ (assumed finite). Then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu) \xrightarrow{d} N(0, \Sigma),$$

where $N(0, \Sigma)$ denotes the d-dimensional Gaussian with mean 0 and covariance Σ.

Proof. Write $Y_k = X_k - \mu$, so $\mathbb{E}[Y_k] = 0$ and $\mathrm{Cov}(Y_k) = \Sigma$. Let $W_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k$.

By Cramér–Wold (Theorem 8.13), it suffices to show that for every $t \in \mathbb{R}^d$,

$$\langle t, W_n\rangle \xrightarrow{d} \langle t, Z\rangle, \quad \text{where } Z \sim N(0, \Sigma).$$

Now, $\langle t, W_n\rangle = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \langle t, Y_k\rangle$. The random variables $\langle t, Y_k\rangle$ are i.i.d. with

$$\mathbb{E}[\langle t, Y_k\rangle] = 0, \qquad \mathbb{V}\mathrm{ar}(\langle t, Y_k\rangle) = t^T \Sigma\, t.$$

If $t^T \Sigma\, t = 0$, then $\langle t, Y_k\rangle = 0$ a.s. and both sides are the point mass at 0.

If $t^T \Sigma\, t > 0$, the classical CLT (Theorem 8.1) gives

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \langle t, Y_k\rangle \xrightarrow{d} N(0, t^T \Sigma\, t).$$

Since $\langle t, Z\rangle \sim N(0, t^T \Sigma\, t)$ when $Z \sim N(0, \Sigma)$, the one-dimensional limits match for every t. □

8.6.3 Examples

Example 8.16 (Joint convergence of sample mean and variance). Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}[X_1] = \mu$, $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2$, and $\mathbb{E}[X_1^4] < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ and $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{X}_n)^2$.

Consider the random vectors $Y_k = \big(X_k - \mu, (X_k-\mu)^2 - \sigma^2\big)^T \in \mathbb{R}^2$. These are i.i.d. with $\mathbb{E}[Y_k] = 0$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix},$$

where $\mu_3 = \mathbb{E}[(X_1-\mu)^3]$ and $\mu_4 = \mathbb{E}[(X_1-\mu)^4]$.

By the multivariate CLT:

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ \widehat{\sigma}_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} N(0, \Sigma).$$

In particular, $\sqrt{n}(\bar{X}_n - \mu)$ and $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2)$ converge jointly (not just marginally) to a bivariate Gaussian.

Example 8.17 (Multinomial proportions). Suppose items are classified into d categories with probabilities $p_1, \ldots, p_d$ (where $\sum_k p_k = 1$). Let N_k be the number of items in category k out of n total. Define $\hat{p}_k = N_k/n$. Then

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma_{jk} = p_j(\delta_{jk} - p_k)$ and $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_d)^T$. This follows from applying the multivariate CLT to the indicator vectors $Y_i = (e_{c_i} - p)$ where c_i is the category of item i and e_k is the kth standard basis vector.

8.7 Berry–Esseen Theorem

The CLT tells us that $F_n(x) \to \Phi(x)$, but says nothing about the rate. The Berry–Esseen theorem provides a quantitative bound.

Theorem 8.18 (Berry–Esseen). *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_1^2] = \sigma^2 > 0$, and $\mathbb{E}[|X_1|^3] = \rho < \infty$. If F_n denotes the distribution function of $S_n/(\sigma\sqrt{n})$ and Φ denotes the standard normal distribution function, then*

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le \frac{C\rho}{\sigma^3\sqrt{n}},$$

where C is an absolute constant ($C \le 0.4748$ is the best known bound, due to Shevtsova, 2011).

The proof uses Fourier-analytic smoothing techniques and is omitted. See Durrett, Theorem 3.4.10, or Feller, Volume II, Chapter XVI.

Remark 8.19. The Berry–Esseen bound has the following consequences:

1. The rate $O(1/\sqrt{n})$ is optimal: it cannot be improved in general.
2. The bound is most useful when ρ/σ^3 is moderate. For symmetric distributions, $\rho/\sigma^3 \ge 1$ by Jensen's inequality.
3. For the binomial approximation (X_i Bernoulli), the bound gives $O(1/\sqrt{n})$, which matches the known asymptotics.

8.8 The Delta Method and Anscombe's Theorem

8.8.1 The Delta Method

The CLT tells us the asymptotic distribution of $\sqrt{n}(\overline{X}_n - \mu)$. In practice one often needs the asymptotic distribution of $\sqrt{n}\big(g(\overline{X}_n) - g(\mu)\big)$ for a smooth function g. The **delta method** answers this question via a first-order Taylor expansion.

Theorem 8.20 (Delta method — univariate)**.** *Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$. If $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ with $g'(\theta) \neq 0$, then*

$$\sqrt{n}\big(g(T_n) - g(\theta)\big) \xrightarrow{d} N\big(0,\ \sigma^2[g'(\theta)]^2\big).$$

Proof. By Taylor's theorem,

$$g(T_n) - g(\theta) = g'(\theta)(T_n - \theta) + r_n(T_n - \theta),$$

where $r_n(h)/h \to 0$ as $h \to 0$. Since $T_n - \theta \xrightarrow{\mathbb{P}} 0$ (convergence in distribution to a constant implies convergence in probability), we have $r_n(T_n - \theta)/(T_n - \theta) \xrightarrow{\mathbb{P}} 0$.

Write

$$\sqrt{n}\big(g(T_n) - g(\theta)\big) = g'(\theta) \cdot \sqrt{n}(T_n - \theta) + \sqrt{n}\, r_n(T_n - \theta).$$

The remainder satisfies

$$\sqrt{n}\, r_n(T_n - \theta) = \frac{r_n(T_n - \theta)}{T_n - \theta} \cdot \sqrt{n}(T_n - \theta)$$

(defined to be 0 when $T_n = \theta$). The first factor converges to 0 in probability, and the second converges in distribution to $N(0, \sigma^2)$. By Slutsky's theorem (Theorem 5.18), the product converges to 0 in probability, hence in distribution. Another application of Slutsky gives the result. □

Remark 8.21 (What if $g'(\theta) = 0$?). When $g'(\theta) = 0$ and $g''(\theta) \neq 0$, the delta method gives $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} 0$, which is uninformative. In this case the correct scaling is n rather than $\sqrt{n}$:

$$n\big(g(T_n) - g(\theta)\big) \xrightarrow{d} \frac{g''(\theta)}{2}\, \sigma^2 \chi^2_1.$$

This is the **second-order delta method**; see Exercise 8.37.

Example 8.22 (Variance-stabilizing transform for proportions). Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) with $0 < p < 1$. The CLT gives $\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} N(0, p(1-p))$.

The asymptotic variance $p(1-p)$ depends on the unknown p, which complicates inference. The **arcsin transform** $g(x) = \arcsin(\sqrt{x})$ has $g'(x) = 1/(2\sqrt{x(1-x)})$, so the delta method gives

$$\sqrt{n}\big(\arcsin\sqrt{\hat{p}_n} - \arcsin\sqrt{p}\big) \xrightarrow{d} N\left(0, \frac{1}{4}\right).$$

The asymptotic variance $1/4$ is **independent of** p: this is a variance-stabilizing transformation.

Theorem 8.23 (Delta method — multivariate)**.** *Suppose* $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \Sigma)$ *in* $\mathbb{R}^d$*. If* $g : \mathbb{R}^d \to \mathbb{R}^k$ *is differentiable at* θ *with Jacobian matrix* $J = Dg(\theta) \in \mathbb{R}^{k \times d}$*, then*

$$\sqrt{n}\big(g(T_n) - g(\theta)\big) \xrightarrow{d} N(0,\, J\Sigma J^T) \quad \text{in } \mathbb{R}^k.$$

Proof. Apply the one-dimensional argument to each component, or use $\langle t, g(T_n) - g(\theta)\rangle \approx \langle J^T t,\, T_n - \theta\rangle$ together with Cramér–Wold (Theorem 8.13). □

Example 8.24 (Asymptotic distribution of the sample correlation)*.* Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. bivariate with finite fourth moments. Define

$$T_n = \Big(\frac{1}{n}\sum X_i,\ \frac{1}{n}\sum Y_i,\ \frac{1}{n}\sum X_i^2,\ \frac{1}{n}\sum Y_i^2,\ \frac{1}{n}\sum X_iY_i\Big)^T \in \mathbb{R}^5.$$

The multivariate CLT gives $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \Sigma_5)$ for the appropriate 5×5 covariance matrix Σ_5.

The sample correlation is $r_n = g(T_n)$ where g expresses the Pearson correlation in terms of these five averages. Applying the multivariate delta method and simplifying (a computation left to Exercise 8.31) gives the classical result: if $\rho = \operatorname{Corr}(X, Y)$, then

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0,\, (1 - \rho^2)^2)$$

whenever (X, Y) is bivariate normal. Fisher's z-transform $g(r) = \tanh^{-1}(r)$ is a variance-stabilizing transformation analogous to the arcsin in Example 8.22:

$$\sqrt{n}\big(\tanh^{-1}(r_n) - \tanh^{-1}(\rho)\big) \xrightarrow{d} N(0, 1).$$

8.8.2 CLT for Randomly Stopped Sums: Anscombe's Theorem

In many applications the number of summands is itself random: in sequential analysis, one collects data until a stopping rule fires; in renewal theory, the number of arrivals by time t is a random variable. The following theorem, due to Anscombe (1952), extends the CLT to this setting.

Theorem 8.25 (Anscombe's theorem)**.** *Let* $X_1, X_2, \ldots$ *be i.i.d. with* $\mathbb{E}[X_1] = \mu$ *and* $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 \in (0, \infty)$*. Let* $(N_n)_{n \ge 1}$ *be positive integer-valued random variables satisfying* $N_n / n \xrightarrow{\mathbb{P}} \lambda$ *for some* $\lambda > 0$*. Then*

$$\frac{S_{N_n} - N_n \mu}{\sigma \sqrt{N_n}} \xrightarrow{d} N(0, 1),$$

where $S_k = X_1 + \cdots + X_k$*.*

Remark 8.26. In the general statement, N_n need not be independent of the X_i's. However, the full proof requires uniform control of the partial sum process (specifically, that fluctuations of S_m over intervals $m \in [\lambda n - \delta n,\, \lambda n + \delta n]$ are negligible), which is most naturally obtained from Doob's maximal inequality for martingales—a tool we will develop in Chapter 9. We prove below the important special case where N_n is independent of $(X_i)_{i \ge 1}$; the proof uses only the tools available to us now.

Proof of Theorem 8.25 when $N_n \perp\!\!\!\perp (X_i)_{i\geq 1}$. Write $Z_k := (S_k - k\mu)/(\sigma\sqrt{k})$, so the classical CLT gives $\mathbb{E}[f(Z_k)] \to \mathbb{E}[f(Z)]$ as $k \to \infty$, for every bounded continuous $f : \mathbb{R} \to \mathbb{R}$, where $Z \sim N(0,1)$.

Since N_n is independent of (X_i), conditioning on N_n gives

$$\mathbb{E}\big[f(Z_{N_n})\big] = \sum_{k=1}^{\infty} \mathbb{E}\big[f(Z_k)\big]\, \mathbb{P}(N_n = k).$$

We want to show this converges to $\mathbb{E}[f(Z)]$. Fix $\varepsilon > 0$. Since $\mathbb{E}[f(Z_k)] \to \mathbb{E}[f(Z)]$ as $k \to \infty$, we can choose K large enough that $|\mathbb{E}[f(Z_k)] - \mathbb{E}[f(Z)]| < \varepsilon$ for all $k \geq K$. Splitting the sum at K:

$$\begin{aligned}\big|\mathbb{E}[f(Z_{N_n})] - \mathbb{E}[f(Z)]\big| &= \Big|\sum_{k=1}^{\infty} \big(\mathbb{E}[f(Z_k)] - \mathbb{E}[f(Z)]\big)\mathbb{P}(N_n = k)\Big| \\ &\leq \underbrace{2\|f\|_\infty\, \mathbb{P}(N_n < K)}_{\text{small tail: few terms, but CLT may not yet be accurate}} \\ &\quad + \underbrace{\varepsilon \cdot \mathbb{P}(N_n \geq K)}_{\text{bulk: many terms, each contributing} \leq \varepsilon}.\end{aligned}$$

Since $N_n/n \xrightarrow{\mathbb{P}} \lambda > 0$, we have $N_n \xrightarrow{\mathbb{P}} \infty$, so $\mathbb{P}(N_n < K) \to 0$ as $n \to \infty$ for each fixed K. Therefore $\limsup_n |\mathbb{E}[f(Z_{N_n})] - \mathbb{E}[f(Z)]| \leq \varepsilon$. Since $\varepsilon > 0$ was arbitrary, $Z_{N_n} \xrightarrow{d} N(0,1)$. □

An immediate and useful corollary restates the conclusion in terms of the sample mean:

Corollary 8.27. *Under the hypotheses of Theorem 8.25,*

$$\sqrt{n}\big(\overline{X}_{N_n} - \mu\big) \xrightarrow{d} N(0,\, \sigma^2/\lambda),$$

where $\overline{X}_k := S_k/k$.

Proof. Write $\sqrt{n}(\overline{X}_{N_n} - \mu) = Z_{N_n} \cdot \sigma\sqrt{N_n}/\sqrt{n} = Z_{N_n} \cdot \sigma\sqrt{N_n/n}$. By Theorem 8.25, $Z_{N_n} \xrightarrow{d} N(0,1)$. Since $N_n/n \xrightarrow{\mathbb{P}} \lambda$, the continuous mapping theorem gives $\sqrt{N_n/n} \xrightarrow{\mathbb{P}} \sqrt{\lambda}$. By Slutsky's theorem (Theorem 5.18),

$$\sqrt{n}(\overline{X}_{N_n} - \mu) \xrightarrow{d} \sigma\sqrt{\lambda}\, N(0,1) = N(0, \sigma^2\lambda).$$ □

Remark 8.28. The asymptotic variance $\sigma^2\lambda$ in Corollary 8.27 may seem counterintuitive: if $\lambda > 1$, we average **more** than n observations, yet the variance is **larger** than σ^2. The resolution is that $\sqrt{n}$ is not the natural scaling for $N_n \approx \lambda n$ observations. Rescaling by $\sqrt{N_n}$ instead gives the cleaner statement $\sqrt{N_n}(\overline{X}_{N_n} - \mu) \xrightarrow{d} N(0, \sigma^2)$, which is just Anscombe's theorem rewritten. The $\sqrt{n}$-scaling is useful when n represents a budget or time horizon and N_n is the random number of observations collected within that budget.

Exercises

Homework

Exercise 8.29 (Delta method for the log-mean). Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}[X_1] = \mu > 0$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$.

(a) Use the delta method to find the asymptotic distribution of $\sqrt{n}(\log \overline{X}_n - \log \mu)$.

(b) Suppose $X_i \sim \mathrm{Exp}(\lambda)$ (so $\mu = 1/\lambda$, $\sigma^2 = 1/\lambda^2$). Simplify your answer.

(c) Use the result of (b) to construct an approximate 95% confidence interval for $\log(1/\lambda)$ based on $\overline{X}_n$.

Exercise 8.30 (Lindeberg verification: sums with growing variances). Let $X_1, X_2, \dots$ be independent with $\mathbb{E}[X_j] = 0$ and $\mathbb{E}[X_j^2] = \sigma_j^2$, where $\sigma_j^2 = j^\alpha$ for some $\alpha > 0$. Assume that $|X_j| \le C\,\sigma_j$ for a constant C independent of j (i.e., the variables are uniformly bounded relative to their standard deviations). Let $s_n^2 = \sum_{j=1}^n \sigma_j^2$.

(a) Show that $s_n^2 \sim n^{\alpha+1}/(\alpha+1)$ as $n \to \infty$.

(b) Set up the triangular array $X_{n,j} = X_j / s_n$ and verify the three standard conditions.

(c) Verify the Lindeberg condition directly, and conclude that $S_n/s_n \xrightarrow{d} N(0,1)$.

Exercise 8.31 (Asymptotic distribution of the sample correlation). Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. bivariate normal with $\mathbb{E}X = \mathbb{E}Y = 0$, $\mathbb{V}\mathrm{ar}(X) = \mathbb{V}\mathrm{ar}(Y) = 1$, and $\mathrm{Corr}(X,Y) = \rho$ with $|\rho| < 1$.

(a) Define $A_n = \frac{1}{n}\sum X_i^2$, $B_n = \frac{1}{n}\sum Y_i^2$, $C_n = \frac{1}{n}\sum X_i Y_i$. Express the sample correlation r_n in terms of A_n, B_n, C_n.

(b) Apply the multivariate CLT to $(X_i^2, Y_i^2, X_iY_i)^T$ to find $\sqrt{n}\big((A_n, B_n, C_n)^T - (1, 1, \rho)^T\big) \xrightarrow{d} N(0, \Sigma)$. You will need the moments $\mathbb{E}[X^4]$, $\mathbb{E}[X^2Y^2]$, $\mathbb{E}[X^3Y]$, $\mathbb{E}[X^2(XY)]$ etc. for the bivariate normal.

Useful fact: for (X,Y) standard bivariate normal with correlation ρ: $\mathbb{E}[X^4] = 3$, $\mathbb{E}[X^2Y^2] = 1 + 2\rho^2$, $\mathbb{E}[X^3Y] = 3\rho$, $\mathbb{E}[XY^3] = 3\rho$.

(c) Apply the multivariate delta method to $g(a,b,c) = c/\sqrt{ab}$ to show

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N\big(0,\ (1-\rho^2)^2\big).$$

Exercise 8.32 (Berry–Esseen for the binomial). Let $X_1, X_2, \dots$ be i.i.d. Bernoulli with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$, where $0 < p < 1$.

(a) Compute $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$ and $\rho = \mathbb{E}|X_1 - p|^3$.

(b) Apply the Berry–Esseen theorem to find an explicit bound on $\sup_x |F_n(x) - \Phi(x)|$ in terms of n and p.

(c) For $p = 1/2$ and $n = 100$, compute the numerical value of the Berry–Esseen bound. Compare with the exact value obtained by computation.

Exercise 8.33 (CLT for fixed points of a random permutation). Let π be a uniformly random permutation of $\{1, 2, \dots, n\}$, and let $W_n = \sum_{i=1}^{n} \mathbf{1}(\pi(i) = i)$ be the number of fixed points.

(a) Define $X_i = \mathbf{1}(\pi(i) = i)$. Compute $\mathbb{E}[X_i]$, $\mathbb{V}\mathrm{ar}(X_i)$, $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$, and $\mathbb{V}\mathrm{ar}(W_n)$.

(b) The X_i are **not** independent. Nevertheless, show that $W_n \xrightarrow{d} \mathrm{Poisson}(1)$ by computing $\mathbb{E}[e^{tW_n}]$ using inclusion-exclusion.

(c) Since $\mathbb{V}\mathrm{ar}(W_n) = 1$ for all n, the CLT scaling $\sqrt{n}$ is inappropriate. Instead, show directly from (b) that the Poisson(1) limit emerges. Compare: does the normal distribution or the Poisson distribution give a better approximation for moderate n?

This example illustrates that dependent indicators with pairwise covariances $O(1/n^2)$ can converge to a Poisson limit rather than a Gaussian one when the variance stays bounded.

Additional Exercises

Exercise 8.34 (Verifying the Lindeberg condition for i.i.d. sums). Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. Define the triangular array $X_{ij} = X_j/(\sigma\sqrt{i})$ for $j = 1, \dots, i$.

(a) Verify that the three standard conditions hold.

(b) Show that the Lindeberg condition is satisfied.

(c) Conclude that Theorem 8.1 follows from Theorem 8.6.

Exercise 8.35 (CLT for non-identically distributed summands). Let $X_1, X_2, \dots$ be independent (but not identically distributed) with $\mathbb{E}[X_j] = 0$, $\mathbb{E}[X_j^2] = \sigma_j^2$, and $\mathbb{E}[|X_j|^3] < \infty$ for each j. Let $s_n^2 = \sum_{j=1}^{n} \sigma_j^2 \to \infty$.

Show that if $\sum_{j=1}^{n} \mathbb{E}|X_j|^3 / s_n^3 \to 0$, then $S_n/s_n \xrightarrow{d} N(0, 1)$.

Exercise 8.36 (Cramér–Wold does not extend to infinite dimensions). Show that the Cramér–Wold theorem can fail in infinite-dimensional spaces. Specifically, find a sequence of probability measures (μ_n) on ℓ^2 such that for every continuous linear functional f the pushforward $f_*\mu_n$ converges weakly on $\mathbb{R}$, but (μ_n) does not converge weakly on ℓ^2.

Exercise 8.37 (Second-order delta method). Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ and g is twice differentiable at θ with $g'(\theta) = 0$ and $g''(\theta) \neq 0$.

(a) Show that $n\big(g(T_n) - g(\theta)\big) \xrightarrow{d} \frac{g''(\theta)}{2}\, \sigma^2\, \chi_1^2$.

(b) Apply this to $g(x) = (x - \mu)^2$ with $T_n = \overline{X}_n$ (where X_i are i.i.d. with mean μ and variance σ^2) to find the asymptotic distribution of $n(\overline{X}_n - \mu)^2$. Compare with what you get from squaring the CLT directly.

Exercise 8.38 (When the CLT fails: Cauchy sums). Let $X_1, X_2, \dots$ be i.i.d. standard Cauchy (density $f(x) = 1/(\pi(1+x^2))$).

(a) Show that $\mathbb{E}[|X_1|] = \infty$.

(b) Show that the characteristic function of X_1 is $\varphi(t) = e^{-|t|}$.

(c) Prove that $\overline{X}_n = (X_1 + \cdots + X_n)/n$ is itself standard Cauchy for every n. In particular, no normalization of the partial sums converges to a Gaussian.

This is the simplest example of a **stable distribution**: one that is preserved (up to location and scale) under summation.

Exercise 8.39 (Variance-stabilizing transformations). Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, v(\theta))$ where $v(\theta) > 0$ is a known function of θ.

(a) Show that $g(\theta) := \int_0^\theta \frac{du}{\sqrt{v(u)}}$ is a variance-stabilizing transformation, meaning that $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0,1)$ regardless of θ.

(b) Apply this to the Poisson case: $X_1, \dots, X_n$ i.i.d. Poisson(λ), $T_n = \overline{X}_n$, $v(\lambda) = \lambda$. Show that the VST is $g(\lambda) = 2\sqrt{\lambda}$, and write the resulting asymptotic statement.

(c) Apply this to the binomial case: $T_n = \hat{p}_n$, $v(p) = p(1-p)$. Verify that $g(p) = \arcsin(\sqrt{p})$ (up to a constant factor), recovering Example 8.22.

Exercise 8.40 (Applying Anscombe: renewal CLT). Let $X_1, X_2, \dots$ be i.i.d. positive random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$, and let $S_k = X_1 + \cdots + X_k$. Define the **renewal counting process**

$$N(t) := \max\{k \geq 0 : S_k \leq t\}, \qquad t > 0.$$

(a) Show that $N(t)/t \to 1/\mu$ almost surely as $t \to \infty$.

(b) Use Anscombe's theorem (Theorem 8.25) to show that

$$\frac{N(t) - t/\mu}{\sigma\, t^{1/2}/\mu^{3/2}} \xrightarrow{d} N(0,1) \quad \text{as } t \to \infty.$$

(c) Interpret the result: if X_i represents the time between successive bus arrivals, what does part (b) say about the number of buses arriving in a long time interval $[0, t]$?

Exercise 8.41 (How many voters to poll?). A pollster wants to estimate the fraction p of voters who support a candidate. She samples n voters independently (with replacement from the population).

(a) Using the CLT and the Berry–Esseen bound, determine the smallest n such that

$$\mathbb{P}\big(|\hat{p}_n - p| \leq 0.03\big) \geq 0.95$$

uniformly over all $p \in (0,1)$. This is the origin of the "± 3 percentage points" margin of error in political polling.

(b) The pollster observes $\hat{p}_n = 0.52$ with $n = 1000$. Use the delta method (via the log-odds transform $g(p) = \log \frac{p}{1-p}$) to construct a 95% confidence interval for the **odds ratio** $p/(1-p)$.

Exercise 8.42 (Failure of CLT when Lindeberg fails). Construct a triangular array satisfying the three standard conditions and the UAN condition, for which the Lindeberg condition **fails** and the row sums S_i do **not** converge in distribution to $N(0,1)$.

1. For row i, let $n_i = i$ and let $X_{i1}, \dots, X_{ii}$ be i.i.d. with

$$\mathbb{P}(X_{ij} = 1) = \mathbb{P}(X_{ij} = -1) = \frac{1}{2i}, \qquad \mathbb{P}(X_{ij} = 0) = 1 - \frac{1}{i}.$$

   Verify the three standard conditions.

2. Show that the UAN condition holds but the Lindeberg condition fails for every $\varepsilon < 1$.

3. Compute the characteristic function $\varphi_{S_i}(t)$ and show that

$$\varphi_{S_i}(t) \to e^{\cos t - 1} \quad \text{as } i \to \infty.$$

   Identify this as the characteristic function of a compound Poisson random variable: $W = Y_1 + \cdots + Y_N$, where $N \sim \text{Poisson}(1)$ and the Y_j are i.i.d. taking values ± 1 with equal probability, independent of N.

4. Verify that $\mathbb{V}\text{ar}(W) = 1$, consistent with the three standard conditions, but that W is integer-valued and hence not Gaussian.

Chapter 9

Martingales: Foundations

Martingales are one of the most powerful tools in modern probability theory. Named after a class of betting strategies, they formalize the idea of a "fair game"—a stochastic process whose future expected value, given all currently available information, equals its present value. This simple definition leads to deep consequences: convergence theorems, maximal inequalities, and a second proof of the Strong Law of Large Numbers that illuminates the result from an entirely different angle.

The theory builds directly on the conditional expectation machinery developed in Chapter 3. There, we learned to compute $\mathbb{E}(X \mid \mathcal{G})$ for a single σ-algebra $\mathcal{G}$; here, we study what happens when the conditioning σ-algebra **grows with time**, modelling the accumulation of information.

This chapter and the chapters that follows are organized as follows. Section 9.1 introduces filtrations and stopping times—the language of "information available at time n" and "decisions made without looking into the future." Section 9.2 defines martingales and develops basic properties, including examples, martingale transforms, and the Doob decomposition. Section 10.1 studies the interplay between martingales and stopping times through the optional stopping theorem and its applications (gambler's ruin, Wald's identities). Section 10.2 presents Doob's maximal inequality and the Azuma–Hoeffding concentration inequality. Chapter 11 develops the convergence theory: first the L^2-bounded and L^1-bounded convergence theorems (via the upcrossing inequality), then the role of uniform integrability in upgrading a.s. convergence to L^1 convergence and in characterizing martingales as conditional expectations. Finally, Chapter 12 introduces backward martingales and uses them to give a second, elegant proof of the SLLN.

9.1 Filtrations and Stopping Times

Filtrations, stopping times, and martingales are all related to the idea of "the information available at the present time." This is represented by an increasing family of σ-algebras indexed by time and random variables measurable with

As a reminder, σ-algebras and σ-fields mean the same thing.

respect to these σ-algebras.

Filtrations

Definition 9.1. A **filtration** $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ is an increasing sequence of σ-algebras. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a filtration $\{\mathcal{F}_n\}_{n\geq 0}$ is called a **filtered probability space**.

We use $\mathcal{F}_\infty$ to denote $\sigma(\bigcup_{n=0}^{\infty} \mathcal{F}_n)$.

The σ-algebra $\mathcal{F}_n$ is often interpreted as the collection of events that are "determined by time n." For example, if $X_1, X_2, \dots$ is a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then this sequence defines a **natural filtration**

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \quad \mathcal{F}_n = \sigma(X_1, \dots, X_n), \quad n \geq 1.$$

Unless stated otherwise, when we work with a sequence of random variables, we use the natural filtration.

Definition 9.2. A sequence of random variables $\{X_n\}$ is **adapted** to a filtration $\{\mathcal{F}_n\}$ if X_n is $\mathcal{F}_n$-measurable for all $n \geq 0$. (We sometimes abuse notation and write $X_n \in \mathcal{F}_n$.)

A sequence $\{X_n\}$ is **predictable** if X_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$. Predictability is a stronger requirement: X_n is known one step before time n.

Stopping times

Definition 9.3. A **stopping time** T is a random variable $T : \Omega \to \mathbb{Z}^+ \cup \{\infty\}$ such that for every $n < \infty$, $\{T = n\} \in \mathcal{F}_n$.

Intuitively, the decision to stop at time n must be based only on the information available at time n—you cannot look into the future. If X_i represents your win or loss at time i in a casino, then your decision to stop after round n should not depend on X_i for $i > n$.

Remark 9.4. An equivalent definition requires $\{T \leq n\} \in \mathcal{F}_n$ for all n. Since $\{T \leq n\} = \bigcup_{k=0}^{n} \{T = k\}$, this is clearly equivalent.

Here are examples of valid stopping times:

- **Constant stopping time:** $T(\omega) = k$ for all ω.
- **First entrance time:** $T = \inf\{n \geq 0 : X_n \in A\}$ for a Borel set A.
- **Composition with non-decreasing functions:** If T is a stopping time and $f\colon \mathbb{Z}^+ \to \mathbb{Z}^+$ is non-decreasing with $f(t) \geq t$ for all t, then $T' = f(T)$ is again a stopping time.
- **Min and max:** If T_1 and T_2 are stopping times, so are $T_1 \wedge T_2$ and $T_1 \vee T_2$.

The last property implies that any stopping time T can be written as an increasing limit of bounded stopping times: $T_n = T \wedge n \uparrow T$.

Example 9.5 (Non-examples of stopping times). Consider the natural filtration generated by random variables $X_1, \ldots, X_N$.

(a) The random variable $T =$ first $i \leq N$ such that $X_i = \max_{1 \leq j \leq N} X_j$ is **not** a stopping time: the event

$$\{T = n\} = \{X_1 < X_n, \ldots, X_{n-1} < X_n\} \cap \{X_{n+1} \leq X_n, \ldots, X_N \leq X_n\}$$

requires knowledge of $X_{n+1}, \ldots, X_N$ and hence $\{T = n\} \notin \mathcal{F}_n$ in general.

(b) The **last** visit to a set A, i.e., $T = \sup\{n : X_n \in A\}$, is not a stopping time.

(c) A non-mathematical example: "cook toast until 10 seconds before it starts to smoke" is not a stopping rule, because it requires knowledge of a future event.

The σ-algebra at a stopping time

Just as we have σ-algebras $\mathcal{F}_n$ associated with deterministic times, we have a σ-algebra $\mathcal{F}_T$ associated with any stopping time T, representing the information available at time T.

Definition 9.6. For a stopping time T, we define

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty : A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for each } n \geq 0\}.$$

One can verify that $\mathcal{F}_T$ is indeed a σ-algebra and that T is $\mathcal{F}_T$-measurable. Moreover, if $\{X_n\}$ is adapted, then X_T is $\mathcal{F}_T$-measurable on the event $\{T < \infty\}$. Exercise 9.20 asks for further properties of $\mathcal{F}_T$.

9.2 Martingales – Basic Results

An $\{\mathcal{F}_n\}$-adapted sequence of random variables $\{X_n\}$ is a **martingale** with respect to $\{\mathcal{F}_n\}$ if

1. $\mathbb{E}|X_n| < \infty$ for every $n \geq 0$, and
2. $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for every $n \geq 0$.

It is useful to think about a martingale X_n as total winnings in a fair game: given the history up to time n, the expected value of the next observation equals the current value.

From the definition and the tower property of conditional expectations, it is immediate that $\mathbb{E}(X_p|\mathcal{F}_n) = X_n$ for all $p > n$. In particular, $\mathbb{E}(X_n) = \mathbb{E}(X_0)$ for every n.

Definition 9.7. An adapted sequence $\{X_n\}$ with $\mathbb{E}|X_n| < \infty$ for all n is called a **submartingale** if $X_n \leq \mathbb{E}(X_{n+1}|\mathcal{F}_n)$, and a **supermartingale** if $X_n \geq \mathbb{E}(X_{n+1}|\mathcal{F}_n)$.

A supermartingale is not an especially "super" thing: it represents winnings in a losing game where your expected wealth in the next period is less

than it is now. Note that if $\{X_n\}$ is a submartingale then $\{-X_n\}$ is a supermartingale and vice versa. For that reason we will usually formulate results either only for submartingales or only for supermartingales; the other case follows by obvious transformation. A martingale is both a submartingale and a supermartingale.

For a submartingale, $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$, so the means form a non-decreasing sequence. A submartingale can be seen as a stochastic analogue of a non-decreasing sequence. We will see later that if a submartingale is bounded from above, then it converges almost surely.

Martingale differences

We can decompose any martingale into its increments. Define $Y_{i+1} = X_{i+1} - X_i$. Then

$$\mathbb{E}(Y_{i+1}|\mathcal{F}_i) = 0 \quad \text{for every } i \geq 0.$$

Such a sequence $\{Y_i\}$ is called a **martingale difference sequence**. Conversely, if $\{Y_i\}$ is a martingale difference sequence, then

$$X_n = X_0 + Y_1 + Y_2 + \cdots + Y_n$$

is a martingale. This representation is the analogue of writing a function as its initial value plus a sum of increments.

Examples of martingales

Example 9.8 (Sums of independent random variables). Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\mathbb{E}\xi_i = 0$ for all $i \geq 1$. Define $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$. Let S_0 be a constant and $S_n = S_0 + \sum_{i=1}^n \xi_i$. Then $\{S_n\}$ is a martingale:

$$\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}\xi_{n+1} = S_n.$$

Example 9.9 (Quadratic martingale). In the setting of the previous example, assume additionally that $\mathbb{V}\mathrm{ar}(\xi_i) = \sigma^2 < \infty$ for all i. Then $M_n = S_n^2 - n\sigma^2$ is a martingale. Indeed, $S_{n+1}^2 = S_n^2 + 2S_n\xi_{n+1} + \xi_{n+1}^2$, and since ξ_{n+1} is independent of $\mathcal{F}_n$:

$$\mathbb{E}(S_{n+1}^2 - (n+1)\sigma^2|\mathcal{F}_n) = S_n^2 + 2S_n \cdot 0 + \sigma^2 - (n+1)\sigma^2 = S_n^2 - n\sigma^2.$$

Example 9.10 (Exponential martingale). Let $Y_1, Y_2, \ldots$ be non-negative i.i.d. random variables with $\mathbb{E}Y_i = 1$, and let $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then $M_n = \prod_{i=1}^n Y_i$ is a martingale:

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = M_n\mathbb{E}(Y_{n+1}) = M_n.$$

An important special case arises when we have i.i.d. random variables ξ_i with moment generating function $\varphi(\theta) = \mathbb{E}(e^{\theta\xi_i}) < \infty$. Setting $Y_i = e^{\theta\xi_i}/\varphi(\theta)$, we obtain the **exponential martingale**

$$M_n = \frac{e^{\theta S_n}}{[\varphi(\theta)]^n},$$

which is in fact a one-parameter family of martingales indexed by θ. This martingale plays a central role in large deviations theory and will reappear in the gambler's ruin problem (Section 10.1).

Example 9.11 (Doob's martingale / Learning martingale). If X is an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{\mathcal{F}_n\}$ is a filtration, then

$$M_n = \mathbb{E}(X|\mathcal{F}_n)$$

is a martingale with respect to $\{\mathcal{F}_n\}$. This follows directly from the tower property:

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \mathbb{E}\big(\mathbb{E}(X|\mathcal{F}_{n+1}) \,\big|\, \mathcal{F}_n\big) = \mathbb{E}(X|\mathcal{F}_n) = M_n.$$

This is called **Doob's martingale** (or the **learning martingale**): as the filtration grows, we learn more and more about X. We will see in Section 11.4 that every uniformly integrable martingale is of this form.

Example 9.12 (Pólya urn). An urn initially contains one red ball and one blue ball. At each step, a ball is drawn uniformly at random, then returned to the urn together with one new ball of the same color. Let R_n be the number of red balls after n draws (so $R_0 = 1$ and the total number of balls after n draws is $n+2$). Define

$$M_n = \frac{R_n}{n+2}.$$

Then M_n is a martingale with respect to $\mathcal{F}_n = \sigma(R_1, \dots, R_n)$. Indeed, given $\mathcal{F}_n$, the $(n+1)$st draw adds a red ball with probability $R_n/(n+2)$, so

$$\begin{aligned}\mathbb{E}(M_{n+1}|\mathcal{F}_n) &= \frac{1}{n+3}\Big[(R_n+1)\cdot\frac{R_n}{n+2} + R_n\cdot\frac{n+2-R_n}{n+2}\Big] \\ &= \frac{R_n}{n+3}\Big[\frac{R_n+1}{n+2} + \frac{n+2-R_n}{n+2}\Big] = \frac{R_n(n+3)}{(n+3)(n+2)} = \frac{R_n}{n+2} = M_n.\end{aligned}$$

Since $M_n \in [0,1]$, we will show in Section 11 that M_n converges a.s. to a limit M_∞. In fact, $M_\infty \sim \mathrm{Uniform}(0,1)$ see Appendix L for a proof (and a generalization to arbitrary initial compositions).

Example 9.13 (Branching process). In a Galton–Watson branching process, we start with $Z_0 = 1$ individual. Each individual in generation n independently produces a random number of offspring with common distribution $\{p_k\}_{k\ge 0}$, and Z_n denotes the total number of individuals in generation n. Let $\mu = \sum k p_k$ be the mean offspring number, and suppose $0 < \mu < \infty$. Define $\mathcal{F}_n = \sigma(Z_0, Z_1, \dots, Z_n)$. Then

$$M_n = \frac{Z_n}{\mu^n}$$

is a non-negative martingale. To verify this, write $Z_{n+1} = \sum_{i=1}^{Z_n} \xi_i^{(n+1)}$ where $\xi_i^{(n+1)}$ are i.i.d. with mean μ, independent of $\mathcal{F}_n$. Then

$$\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = Z_n \cdot \mu,$$

and so $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = Z_n\mu/\mu^{n+1} = Z_n/\mu^n = M_n$.

We will show in Section 11 that every non-negative martingale converges a.s. to a limit. In particular, $M_n \xrightarrow{a.s.} W$ for some $W \geq 0$. However, the limit can be trivial: if $\mu \leq 1$ (and the offspring distribution is not degenerate at 1), then $Z_n = 0$ eventually and so $W = 0$ a.s. Even when $\mu > 1$, one may have $W = 0$ a.s. unless sufficient moments exist. The Kesten–Stigum theorem asserts that $\mathbb{P}(W > 0) > 0$ (conditional on non-extinction) if and only if $\sum p_k k \log k < \infty$.

Martingale transforms

A martingale transform is the discrete stochastic analogue of a stochastic integral. It formalizes the idea of a gambling strategy: a_n is the bet size chosen before round n, and $Y_n = X_n - X_{n-1}$ is the outcome.

Definition 9.14. Suppose $\{X_n\}$ is a martingale with respect to $\{\mathcal{F}_n\}$, with differences $Y_n = X_n - X_{n-1}$, and $\{a_n\}$ is a predictable sequence (i.e., a_n is $\mathcal{F}_{n-1}$-measurable). The **martingale transform** $(a \cdot X)_n$ is defined by

$$(a \cdot X)_n = \sum_{k=1}^{n} a_k Y_k.$$

Lemma 9.15. *If $\{X_n\}$ is an $\mathcal{F}_n$-martingale and $\{a_n\}$ is a predictable sequence such that $a_n Y_n$ is integrable for each n, then $(a \cdot X)$ is an $\mathcal{F}_n$-martingale.*

Proof. Let $X'_n = (a \cdot X)_n$. Since $a_n \in \mathcal{F}_{n-1}$ and X_n is a martingale:

$$\begin{aligned}\mathbb{E}(X'_n - X'_{n-1}|\mathcal{F}_{n-1}) &= \mathbb{E}(a_n Y_n|\mathcal{F}_{n-1}) = a_n \mathbb{E}(Y_n|\mathcal{F}_{n-1}) \\ &= a_n(\mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}) = 0.\end{aligned}$$

□

Remark 9.16. If $\{X_n\}$ is a submartingale and $\{a_n\}$ is predictable with $a_n \geq 0$, then $(a \cdot X)$ is also a submartingale. (The non-negativity of a_n preserves the direction of the inequality.) This is intuitively clear: a non-negative betting strategy applied to a favorable game remains favorable.

Generating submartingales from martingales

Lemma 9.17. *Suppose $\{(X_n, \mathcal{F}_n)\}$ is a martingale and φ is a convex function such that $\mathbb{E}|\varphi(X_n)| < \infty$ for every n. Then $\{(\varphi(X_n), \mathcal{F}_n)\}$ is a submartingale.*

Proof. By Jensen's inequality for conditional expectations:

$$\varphi(X_n) = \varphi\big(\mathbb{E}(X_{n+1}|\mathcal{F}_n)\big) \leq \mathbb{E}\big(\varphi(X_{n+1})|\mathcal{F}_n\big).$$

□

In particular, for any $p \geq 1$, if $\{X_n\}$ is a martingale with $\mathbb{E}|X_n|^p < \infty$ for all n, then $\{|X_n|^p\}$ is a submartingale. Taking $p = 2$, we see that $\mathbb{E}(X_n^2)$ is non-decreasing in n, i.e., the variance of a martingale can only grow with time.

Doob decomposition

Every submartingale can be uniquely decomposed into a martingale plus a predictable non-decreasing process. This is the discrete-time analogue of the Doob–Meyer decomposition in continuous time.

Theorem 9.18 (Doob Decomposition)**.** *Let $\{X_n\}$ be a submartingale with respect to $\{\mathcal{F}_n\}$. Then there exist a martingale $\{M_n\}$ and a predictable non-decreasing sequence $\{A_n\}$ with $A_0 = 0$ such that*

$$X_n = M_n + A_n \quad \text{for all } n \geq 0.$$

Moreover, this decomposition is unique (up to the choice of $M_0 = X_0$).

Proof. Existence. Define $A_0 = 0$ and for $n \geq 1$,

$$A_n = \sum_{k=1}^{n} \big(\mathbb{E}(X_k|\mathcal{F}_{k-1}) - X_{k-1}\big).$$

Since $\{X_n\}$ is a submartingale, each summand $\mathbb{E}(X_k|\mathcal{F}_{k-1}) - X_{k-1} \geq 0$, so $\{A_n\}$ is non-decreasing. Also, A_n is $\mathcal{F}_{n-1}$-measurable (in fact each increment is $\mathcal{F}_{k-1}$-measurable), so $\{A_n\}$ is predictable.

Set $M_n = X_n - A_n$. Then $M_0 = X_0$ and

$$\begin{aligned}\mathbb{E}(M_n - M_{n-1}|\mathcal{F}_{n-1}) &= \mathbb{E}(X_n - X_{n-1} - (A_n - A_{n-1})|\mathcal{F}_{n-1}) \\ &= \mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1} - \big(\mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}\big) = 0.\end{aligned}$$

So $\{M_n\}$ is a martingale.

Uniqueness. If $X_n = M_n + A_n = M'_n + A'_n$ are two such decompositions, then $M_n - M'_n = A'_n - A_n$. The left side is a martingale; the right side is predictable (both A'_n and A_n are predictable). A predictable martingale $\{D_n\}$ satisfies $D_n \in \mathcal{F}_{n-1}$ and $\mathbb{E}(D_n|\mathcal{F}_{n-1}) = D_{n-1}$, hence $D_n = D_{n-1}$ a.s. for all n. Thus $D_n = D_0$ a.s. for all n, which gives $A_n = A'_n$ and $M_n = M'_n$ a.s. □

Remark 9.19. For a martingale, $A_n = 0$ for all n, so the Doob decomposition is trivial. The decomposition becomes interesting for submartingales: the predictable part A_n captures the systematic "drift" while M_n captures the "noise." This is analogous to writing a function as a monotone function plus an oscillatory part. Exercise 9.26 provides practice with computing the decomposition explicitly.

9.3 Exercises

Homework

Exercise 9.20 (Properties of $\mathcal{F}_T$). Show the following properties of $\mathcal{F}_T$:

(a) If $T = k$ is a constant stopping time, then $\mathcal{F}_T = \mathcal{F}_k$.

(b) If S and T are stopping times with $S \leq T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

(c) If S and T are stopping times, then $\mathcal{F}_{S\wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$.

Exercise 9.21 (Likelihood ratio martingale). Let $X_1, X_2, \ldots$ be i.i.d. with density f with respect to Lebesgue measure, and let g be another density with $f(x) > 0$ whenever $g(x) > 0$. Define

$$L_n = \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)}, \qquad L_0 = 1.$$

(a) Show that $\{L_n\}$ is a non-negative martingale with respect to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ (under the probability measure induced by f).

(b) What is $\mathbb{E}(L_n)$?

(c) The ratio $g(x)/f(x)$ is called the **likelihood ratio**. Give an interpretation of L_n in the context of hypothesis testing: an observer sees data $X_1, \ldots, X_n$ and wants to decide whether the true density is f (null hypothesis) or g (alternative).

Exercise 9.22 (Running maximum is a submartingale). Let $\{X_n\}$ be a submartingale with respect to $\{\mathcal{F}_n\}$. Define $\bar{X}_n = \max_{0\le k\le n} X_k$. Show that $\{\bar{X}_n\}$ is also a submartingale.

Exercise 9.23 (Predictable quadratic variation). Let $\{M_n\}$ be a martingale with $\mathbb{E}(M_n^2) < \infty$ for all n. By Lemma 9.17, $\{M_n^2\}$ is a submartingale.

(a) Write down the Doob decomposition $M_n^2 = N_n + \langle M\rangle_n$. Show that

$$\langle M\rangle_n = \sum_{k=1}^{n} \mathbb{E}\big((M_k - M_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big).$$

The process $\langle M\rangle_n$ is called the **predictable quadratic variation** (or **angle bracket process**) of M.

(b) Show that $\mathbb{E}(\langle M\rangle_n) = \mathbb{E}(M_n^2) - \mathbb{E}(M_0^2)$. Conclude that if M_0 is a constant, then $\mathbb{E}(\langle M\rangle_n) = \mathbb{V}\mathrm{ar}(M_n)$.

(c) Compute $\langle M\rangle_n$ when $M_n = S_n = \sum_{i=1}^{n} \xi_i$ with ξ_i i.i.d., $\mathbb{E}\xi_i = 0$, $\mathbb{V}\mathrm{ar}(\xi_i) = \sigma^2$. Relate your answer to Example 9.9.

Additional Exercises

Exercise 9.24 (Computing $\mathcal{F}_T$ explicitly). Consider two coin flips with $\Omega = \{HH, HT, TH, TT\}$ and the natural filtration

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \quad \mathcal{F}_1 = \sigma(\text{first flip}), \quad \mathcal{F}_2 = 2^{\Omega}.$$

Define $T(\omega) = 1$ if the first flip is H, and $T(\omega) = 2$ if the first flip is T.

(a) Verify that T is a stopping time.

(b) List all elements of $\mathcal{F}_T$. How many are there? Compare with $|\mathcal{F}_1|$ and $|\mathcal{F}_2|$.

(c) Explain intuitively why $\mathcal{F}_T$ contains more information than $\mathcal{F}_1$ but less than $\mathcal{F}_2$.

Exercise 9.25 (Rounding a stopping time: ceiling vs. floor). Let T be a stopping time with respect to a filtration $\{\mathcal{F}_n\}$.

(a) Show that $T' = 2\lceil T/2 \rceil$ is a stopping time. (Here $\lceil \cdot \rceil$ is the ceiling function, so T' rounds T up to the nearest even number.)

(b) Show that $T'' = 2\lfloor T/2 \rfloor$ is **not** a stopping time in general, by giving a concrete counterexample.

(c) Explain how both parts are consistent with the general principle stated in Section 9.1: if $f\colon \mathbb{Z}^+ \to \mathbb{Z}^+$ is non-decreasing with $f(t) \geq t$ for all t, then $f(T)$ is a stopping time.

Exercise 9.26 (Doob decomposition examples). Find the Doob decomposition for the following submartingales:

(a) $X_n = S_n^2$, where S_n is a random walk with i.i.d. increments ξ_i satisfying $\mathbb{E}\xi_i = 0$ and $\mathbb{V}\mathrm{ar}(\xi_i) = \sigma^2$.

(b) $X_n = (S_n^+)^2$, where S_n is a symmetric random walk and $x^+ = \max(x, 0)$.

Exercise 9.27 (Pólya urn: a second martingale). Consider the Pólya urn from Example 9.12. Let R_n denote the number of red balls and $B_n = n + 2 - R_n$ the number of blue balls after n draws.

(a) Show that
$$W_n = \frac{R_n \cdot B_n}{(n+2)(n+3)}$$
is a martingale with respect to $\mathcal{F}_n = \sigma(R_1, \ldots, R_n)$.

(b) Express W_n in terms of $M_n = R_n/(n+2)$ and simplify. Compute W_0 and use the martingale property to find $\mathbb{E}(M_\infty(1 - M_\infty))$, where $M_\infty = \lim_{n\to\infty} M_n$. (The limit exists a.s. by Section 11.)

Exercise 9.28 (The cubic martingale and polynomial martingales). Let $\xi_1, \xi_2, \ldots$ be i.i.d. with $\mathbb{E}(\xi_i) = 0$, $\mathbb{E}(\xi_i^2) = \sigma^2$, $\mathbb{E}(\xi_i^3) = \gamma$, and $\mathbb{E}(\xi_i^4) < \infty$. Let $S_n = \sum_{i=1}^n \xi_i$ with $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$.

(a) Show that
$$M_n = S_n^3 - 3n\sigma^2 S_n - n\gamma$$
is a martingale. (This is the "cubic analogue" of Example 9.9.)

(b) Can you construct a degree-4 polynomial martingale in S_n? What moment conditions are needed?

Remark 9.29 (Polynomial martingales of arbitrary degree). The construction in Exercise 9.28 is part of a general pattern. Let $\psi(\theta) = \log \mathbb{E}(e^{\theta\xi_1})$ be the cumulant generating function, with cumulants $\kappa_1 = 0$, $\kappa_2 = \sigma^2$, $\kappa_3 = \gamma$, etc. Since the exponential martingale $M_n(\theta) = e^{\theta S_n - n\psi(\theta)}$ (Example 9.10) is

a martingale for each θ, formally differentiating k times with respect to θ at $\theta = 0$ produces a polynomial martingale of degree k in S_n. The first few are:

$$\begin{aligned} k=1: &\quad S_n, \\ k=2: &\quad S_n^2 - n\kappa_2, \\ k=3: &\quad S_n^3 - 3n\kappa_2 S_n - n\kappa_3. \end{aligned}$$

The degree-k martingale requires $\mathbb{E}|\xi_1|^k < \infty$. In practice, one does not need the general formula via cumulants: it is simpler to expand $(S_n+\xi_{n+1})^k$ directly and match terms, as in the exercise.

Exercise 9.30 (Doob's martingale and convexity). Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\{\mathcal{F}_n\}$ a filtration. Define Doob's martingale $M_n = \mathbb{E}(X \mid \mathcal{F}_n)$ (Example 9.11). Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be convex with $\mathbb{E}|\varphi(X)| < \infty$.

(a) Show that $\mathbb{E}(\varphi(M_n))$ is non-decreasing in n and satisfies $\mathbb{E}(\varphi(M_n)) \le \mathbb{E}(\varphi(X))$ for all n. Thus:

$$\varphi(\mathbb{E}X) \le \mathbb{E}(\varphi(M_1)) \le \mathbb{E}(\varphi(M_2)) \le \cdots \le \mathbb{E}(\varphi(X)).$$

(b) Taking $\varphi(x) = |x|^p$ for $p \ge 1$, conclude that $\sup_n \mathbb{E}|M_n|^p \le \mathbb{E}|X|^p$.

(c) Interpret: as the filtration grows, M_n becomes a progressively better approximation to X. Why does $\mathbb{E}(\varphi(M_n))$ increase?

Chapter 10

Martingales: Optional Stopping and Inequalities

This chapter develops two fundamental themes in martingale theory. The first is the **optional stopping theorem**, which asks: if a martingale is a "fair game" at every deterministic time, does fairness persist when we stop at a random time chosen based on the observed history? The answer is yes—under appropriate integrability conditions—and the result has striking applications, most notably to gambler's ruin and Wald's identities. The second theme is **martingale inequalities**: Doob's maximal inequality controls the probability that a submartingale ever exceeds a threshold, and the Azuma–Hoeffding inequality provides exponential concentration bounds for martingales with bounded increments. We conclude with an application of Doob's inequality to complete the proof of Anscombe's theorem from Chapter 8.

10.1 Stopped Martingales and Optional Stopping

We know that if $\{X_n\}$ is a martingale, then $\mathbb{E}(X_n) = \mathbb{E}(X_0)$ for every fixed time n. A fair game cannot produce any expected gain at a deterministic time. What happens if we use a stopping rule to decide when to quit?

The stopped process is a martingale

Theorem 10.1 (Stopped martingale theorem)**.** *If $\{X_n\}$ is a martingale and T is a stopping time, then the stopped process $\{X_{n\wedge T}\}$ is also a martingale.*

Proof. Define $a_n = \mathbf{1}_{\{T\geq n\}} = \mathbf{1}_{\{T>n-1\}}$. Since $\{T > n-1\} = \{T \leq n-1\}^c \in \mathcal{F}_{n-1}$, the sequence $\{a_n\}$ is predictable. Moreover, a_n is bounded. We claim that

$$X_{n\wedge T} - X_0 = \sum_{k=1}^{n} a_k(X_k - X_{k-1}) = (a \cdot X)_n.$$

To see this, consider two cases. If $T \geq n$, then $a_k = 1$ for all $k \leq n$, so the sum equals $\sum_{k=1}^{n}(X_k - X_{k-1}) = X_n - X_0 = X_{n\wedge T} - X_0$. If $T = m < n$, then $a_k = 1$ for $k \leq m$ and $a_k = 0$ for $k > m$, so the sum equals $\sum_{k=1}^{m}(X_k - X_{k-1}) = X_m - X_0 = X_T - X_0 = X_{n\wedge T} - X_0$. By Lemma 9.15, $(a \cdot X)$ is a martingale, so $\{X_{n\wedge T}\}$ is a martingale. □

In particular, $\mathbb{E}(X_{n\wedge T}) = \mathbb{E}(X_0)$ for every n.

Optional stopping: when does $\mathbb{E}(X_T) = \mathbb{E}(X_0)$?

Suppose $\mathbb{P}(T < \infty) = 1$. Then X_T is well-defined and $X_{T\wedge n} \to X_T$ a.s. as $n \to \infty$. If we could exchange limit and expectation, we would conclude

$$\mathbb{E}(X_T) = \mathbb{E}\Big(\lim_{n\to\infty} X_{T\wedge n}\Big) = \lim_{n\to\infty} \mathbb{E}(X_{T\wedge n}) = \mathbb{E}(X_0).$$

However, this exchange is not always valid.

Exercise 10.20 asks you to verify that this stopping time is finite a.s.

Example 10.2. In the setting of the previous exercise, $S_T = 1$ with probability 1, so $\mathbb{E}(S_T) = 1 \neq 0 = \mathbb{E}(S_0)$. Despite the fact that we play a fair game, this stopping rule produces a sure gain of one dollar! The resolution is that the exchange of limit and expectation is not justified.

Exercise 10.21 asks you to show that $\mathbb{E}(T) = \infty$.

The following result gives sufficient conditions for the optional stopping identity to hold.

Theorem 10.3 (Optional Stopping Theorem)**.** *Let $\{X_n\}$ be a martingale and T a stopping time with $\mathbb{P}(T < \infty) = 1$. Then $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ provided any one of the following conditions holds:*

(i) Bounded stopping time: $T \leq N$ a.s. for some constant N.

(ii) Bounded stopped process: $|X_{T\wedge n}| \leq C$ a.s. for all n and some constant C.

(iii) Dominated stopped process: $|X_{T\wedge n}| \leq Y$ a.s. for all n and some integrable random variable Y.

(iv) Uniform integrability: the family $\{X_{T\wedge n}\}_{n\geq 0}$ is uniformly integrable.

Proof. In each case, $X_{T\wedge n} \to X_T$ a.s. and we justify the exchange of limit and expectation.

(i) If $T \leq N$, then $X_T = X_{T\wedge N}$ and $\mathbb{E}(X_T) = \mathbb{E}(X_{T\wedge N}) = \mathbb{E}(X_0)$ since $\{X_{n\wedge T}\}$ is a martingale by Theorem 10.1.

(ii) and (iii) follow from the bounded convergence theorem and the dominated convergence theorem, respectively.

(iv) Uniform integrability plus a.s. convergence implies L^1 convergence by the Vitali convergence theorem (Chapter 5), so $\mathbb{E}(X_{T\wedge n}) \to \mathbb{E}(X_T)$. □

Remark 10.4. Condition (iv) is the most general: it is also **necessary** for the conclusion $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ to hold for all martingales stopped at T. This will be discussed further in Section 11.4.

Gambler's ruin

Suppose that $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with

$$\mathbb{P}(X_i = 1) = p, \qquad \mathbb{P}(X_i = -1) = q = 1 - p.$$

Let $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$. Define the stopping time

$$\tau = \inf\{n \geq 0 : S_n = A \text{ or } S_n = -B\},$$

where $A, B > 0$ are integers. Here S_n represents the wealth of a gambler who wins or loses one dollar each round, and τ is the first time the gambler's wealth reaches A (target) or $-B$ (ruin).

We want to find $\mathbb{P}(S_\tau = A)$—the probability that the gambler wins A dollars before losing B dollars—and $\mathbb{E}(\tau)$.

Step 1: τ is finite a.s.

This argument works for any $p > 0$. The probability that the gambler wins $A+B$ times in a row is at least $p^{A+B} > 0$. Consider the events E_k, $k = 0, 1, \ldots$, that the gambler wins $A+B$ consecutive games on rounds $[k(A+B)+1, \ldots, (k+1)(A+B)]$. These events are independent and $\sum_k \mathbb{P}(E_k) = \infty$, so by the second Borel–Cantelli lemma, $\mathbb{P}(E_k \text{ i.o.}) = 1$. If any E_k occurs, then the walk must have exited $[-B, A]$ by time $(k+1)(A+B)$. Hence $\mathbb{P}(\tau < \infty) = 1$.

Moreover, τ has exponentially decaying tails:

$$\mathbb{P}(\tau > n(A+B)) \leq (1 - p^{A+B})^n,$$

which implies that all moments of τ are finite. In particular, $\mathbb{E}(\tau) < \infty$.

(See Exercise 10.22.)

Step 2: Fair game ($p = 1/2$)

When $p = 1/2$, the random walk S_n is a martingale. Since $|S_{n\wedge\tau}| \leq \max(A, B)$, the stopped process is bounded and Theorem 10.3(ii) gives

$$0 = \mathbb{E}(S_0) = \mathbb{E}(S_\tau) = A\,\mathbb{P}(S_\tau = A) - B\,\mathbb{P}(S_\tau = -B).$$

Since $\mathbb{P}(S_\tau = -B) = 1 - \mathbb{P}(S_\tau = A)$, we solve to find

$$\boxed{\mathbb{P}(S_\tau = A) = \frac{B}{A+B}.}$$

Note that $\mathbb{P}(S_\tau = A) \to 1$ as $B \to \infty$: with unlimited capital, you will eventually win A dollars in a fair game.

To find $\mathbb{E}(\tau)$, we use the quadratic martingale $M_n = S_n^2 - n$ from Example 9.9 (with $\sigma^2 = 1$). The stopped process $M_{n\wedge\tau}$ satisfies

$$|M_{n\wedge\tau}| = |S_{n\wedge\tau}^2 - (n \wedge \tau)| \leq \max(A, B)^2 + \tau.$$

Since $\mathbb{E}(\tau) < \infty$, this is an integrable dominating function, and Theorem 10.3(iii) gives $\mathbb{E}(M_\tau) = 0$, i.e.,

$$\mathbb{E}(S_\tau^2) = \mathbb{E}(\tau).$$

Computing $\mathbb{E}(S_\tau^2)$ directly:

$$\mathbb{E}(S_\tau^2) = A^2 \cdot \frac{B}{A+B} + B^2 \cdot \frac{A}{A+B} = AB.$$

Hence

$$\boxed{\mathbb{E}(\tau) = AB.}$$

Note that $\mathbb{E}(\tau) \to \infty$ as $B \to \infty$.

Remark 10.5 (Distribution of τ via the exponential martingale). The linear and quadratic martingales gave us $\mathcal{P}(S_\tau = A)$ and $\mathbb{E}(\tau)$. Can we extract more—say, the full distribution of τ? Yes: the **exponential martingale** $M_n^{(\theta)} = e^{\theta S_n}/(\cosh\theta)^n$ (a special case of Example 9.10) is a martingale for every $\theta \in \mathbb{R}$, and applying optional stopping to it produces the Laplace transform of τ. See Appendix I for the derivation and consequences.

Step 3: Unfair game ($p \neq 1/2$)

When $p \neq 1/2$, the random walk S_n is no longer a martingale. We need a different martingale. This can be obtained as a special case of the exponential martingale (Example 9.10), or by the following direct argument.

We seek a function h such that $h(S_n)$ is a martingale. This requires $h(x) = p\,h(x+1) + q\,h(x-1)$. Trying $h(x) = z^x$ gives the equation $z = pz^2 + q$, whose roots are $z = 1$ and $z = q/p$. Since $p \neq 1/2$, these are distinct, and we conclude that $M_n = (q/p)^{S_n}$ is a martingale with $M_0 = 1$.

The stopped process $M_{n\wedge\tau}$ is bounded (since $-B \leq S_{n\wedge\tau} \leq A$), so by Theorem 10.3(ii):

$$1 = \mathbb{E}(M_0) = \mathbb{E}(M_\tau) = \mathbb{P}(S_\tau = A)\,(q/p)^A + \mathbb{P}(S_\tau = -B)\,(q/p)^{-B}.$$

Writing $r = q/p$ and using $\mathbb{P}(S_\tau = -B) = 1 - \mathbb{P}(S_\tau = A)$:

$$1 = \mathbb{P}(S_\tau = A) \cdot r^A + (1 - \mathbb{P}(S_\tau = A)) \cdot r^{-B},$$

and solving:

$$\boxed{\mathbb{P}(S_\tau = A) = \frac{1 - r^{-B}}{r^A - r^{-B}} = \frac{(q/p)^B - 1}{(q/p)^{A+B} - 1}.}$$

The second form is obtained by multiplying numerator and denominator by r^B.

Exercise 10.23 asks you to analyze the limiting behavior as $B \to \infty$.

Wald's first identity

Theorem 10.6 (Wald's First Identity)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}|X_1| < \infty$, $S_n = X_1 + \cdots + X_n$, and $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. If T is a stopping time for $\{\mathcal{F}_n\}$ with $\mathbb{E}(T) < \infty$, then*

$$\mathbb{E}(S_T) = \mathbb{E}(X_1) \cdot \mathbb{E}(T).$$

Proof. We write

$$\mathbb{E}(S_T) = \mathbb{E}\Big(\sum_{n=1}^{T} X_n\Big) = \mathbb{E}\Big(\sum_{n=1}^{\infty} X_n \mathbf{1}_{\{T \geq n\}}\Big) = \sum_{n=1}^{\infty} \mathbb{E}(X_n \mathbf{1}_{\{T \geq n\}}).$$

The exchange of sum and expectation is justified by Fubini's theorem (or Tonelli's theorem applied to $|X_n|\mathbf{1}_{\{T \geq n\}}$, whose sum has expectation $\mathbb{E}|X_1| \cdot \mathbb{E}(T) < \infty$).

For each n, we condition on $\mathcal{F}_{n-1}$:

$$\mathbb{E}(X_n \mathbf{1}_{\{T \geq n\}}) = \mathbb{E}\big(\mathbb{E}(X_n \mathbf{1}_{\{T \geq n\}} | \mathcal{F}_{n-1})\big) = \mathbb{E}\big(\mathbf{1}_{\{T \geq n\}} \mathbb{E}(X_n | \mathcal{F}_{n-1})\big),$$

where the last equality holds because $\{T \geq n\} = \{T \leq n-1\}^c \in \mathcal{F}_{n-1}$. (At the end of period $n-1$ we know whether we have stopped, but we do not yet know the outcome of round n.) Since X_n is independent of $\mathcal{F}_{n-1}$:

$$\mathbb{E}(S_T) = \sum_{n=1}^{\infty} \mathbb{E}\big(\mathbf{1}_{\{T \geq n\}} \mathbb{E}(X_1)\big) = \mathbb{E}(X_1) \sum_{n=1}^{\infty} \mathbb{P}(T \geq n) = \mathbb{E}(X_1) \cdot \mathbb{E}(T).$$

□

Remark 10.7. Wald's identity does **not** require $\mathbb{E}(X_1) = 0$. When $\mathbb{E}(X_1) = 0$, it reduces to $\mathbb{E}(S_T) = 0$, which is the optional stopping result for the martingale S_n. Thus Wald's identity can be viewed as an extension of optional stopping to the non-centered case.

The following generalization is left as Exercise 10.24. Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(X_1) = 0$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$. If T is a stopping time with $\mathbb{E}(T) < \infty$, show that $\mathbb{E}(S_T^2) = \sigma^2 \mathbb{E}(T)$.

10.2 Martingale Inequalities

Doob's maximal inequality

The following result is the martingale generalization of Kolmogorov's maximal inequality. It controls the probability that a submartingale ever exceeds a given threshold.

Theorem 10.8 (Doob's Maximal Inequality)**.** *Suppose $\{X_i\}_{i=0}^n$ is a non-negative submartingale. Let $E = \big\{\max_{0 \leq i \leq n} X_i \geq \lambda\big\}$. Then*

$$\mathbb{P}(E) \leq \frac{1}{\lambda} \mathbb{E}(\mathbf{1}_E X_n) \leq \frac{1}{\lambda} \mathbb{E} X_n. \tag{10.1}$$

Usually one only needs the weaker inequality $\mathbb{P}(E) \le \frac{1}{\lambda}\mathbb{E}X_n$, but the sharper middle inequality with $\mathbb{E}(\mathbf{1}_E X_n)$ is useful in the proof of Doob's L^p inequality.

Proof. We decompose E into disjoint events E_j, where E_j is the event that the process first reaches λ at time j:

$$E_j = \{X_0 < \lambda, \dots, X_{j-1} < \lambda,\, X_j \ge \lambda\}.$$

Note that $E_j \in \mathcal{F}_j$.

For each j, we have the chain of inequalities

$$\mathbb{P}(E_j) \le \frac{1}{\lambda}\int_{E_j} X_j \, d\mathbb{P} \le \frac{1}{\lambda}\int_{E_j} X_n \, d\mathbb{P}.$$

The first inequality is Markov's inequality restricted to E_j. The second uses the submartingale property: since $\mathbb{E}(X_n \mid \mathcal{F}_j) \ge X_j$ and $E_j \in \mathcal{F}_j$,

$$\mathbb{E}(\mathbf{1}_{E_j} X_n) = \mathbb{E}\big(\mathbf{1}_{E_j}\mathbb{E}(X_n \mid \mathcal{F}_j)\big) \ge \mathbb{E}(\mathbf{1}_{E_j} X_j).$$

Summing over $j = 0, 1, \dots, n$ and using $E = \bigsqcup_j E_j$:

$$\mathbb{P}(E) \le \frac{1}{\lambda}\int_E X_n \, d\mathbb{P} \le \frac{1}{\lambda}\int X_n \, d\mathbb{P},$$

where the last inequality uses $X_n \ge 0$. □

Corollary 10.9 (L^p version of Doob's inequality)**.** *If $\{X_n\}$ is a martingale with $\mathbb{E}|X_n|^p < \infty$ for all n and some $p \ge 1$, then*

$$\mathbb{P}\Big(\max_{0\le i\le n} |X_i| \ge \lambda\Big) \le \frac{1}{\lambda^p}\mathbb{E}|X_n|^p. \tag{10.2}$$

Proof. By Lemma 9.17, $\{|X_i|^p\}$ is a non-negative submartingale. Apply Theorem 10.8 to this submartingale with threshold λ^p. □

For $p = 2$, this gives a generalization of Kolmogorov's maximal inequality (Chapter 6) from sums of independent random variables to arbitrary martingales.

Remark 10.10. Doob's maximal inequality can also be stated in L^p norm form: if $\{X_n\}$ is a martingale with $\sup_n \mathbb{E}|X_n|^p < \infty$ for some $p > 1$, then

$$\mathbb{E}\Big(\max_{0\le i\le n} |X_i|^p\Big) \le \Big(\frac{p}{p-1}\Big)^p \mathbb{E}|X_n|^p.$$

This is Doob's L^p maximal inequality; The proof is given in Appendix J. The constant $\big(\frac{p}{p-1}\big)^p$ diverges as $p \to 1$, reflecting the fact that the L^1 case requires additional assumptions (uniform integrability) for good behavior.

Azuma–Hoeffding inequality

The Azuma–Hoeffding inequality is a concentration inequality for martingales with bounded differences. It generalizes the classical Hoeffding inequality for sums of bounded independent random variables.

Theorem 10.11 (Azuma–Hoeffding Inequality)**.** *Let $\{Y_n\}$ be a martingale with respect to $\{\mathcal{F}_n\}$, and suppose there exist constants $K_1, K_2, \dots$ such that $|Y_n - Y_{n-1}| \le K_n$ a.s. for all $n \ge 1$. Then for every $x > 0$,*

$$\mathbb{P}(|Y_n - Y_0| \ge x) \le 2 \exp\Big(- \frac{x^2}{2 \sum_{i=1}^{n} K_i^2} \Big). \tag{10.3}$$

The proof relies on the following lemma.

Lemma 10.12 (Hoeffding's lemma)**.** *Suppose X is a random variable with $\mathbb{E}(X) = 0$ and $|X| \le c$ a.s. Then for every $t > 0$,*

$$\mathbb{E}(e^{tX}) \le \cosh(tc) \le e^{t^2 c^2 / 2}.$$

Proof. By convexity of e^{tx} and the constraint $x \in [-c, c]$:

$$e^{tx} \le \frac{c + x}{2c}\, e^{tc} + \frac{c - x}{2c}\, e^{-tc}.$$

Taking expectations and using $\mathbb{E}(X) = 0$:

$$\mathbb{E}(e^{tX}) \le \frac{1}{2}(e^{tc} + e^{-tc}) = \cosh(tc).$$

The inequality $\cosh(u) \le e^{u^2/2}$ follows from comparing Taylor series: $\cosh(u) = \sum_{k=0}^{\infty} \frac{u^{2k}}{(2k)!} \le \sum_{k=0}^{\infty} \frac{u^{2k}}{2^k k!} = e^{u^2/2}$, since $(2k)! \ge 2^k k!$. □

Proof of Theorem 10.11. Let $D_k = Y_k - Y_{k-1}$ denote the martingale differences, so $|D_k| \le K_k$ a.s. and $\mathbb{E}(D_k \mid \mathcal{F}_{k-1}) = 0$. By Markov's inequality applied to $e^{t(Y_n - Y_0)}$, for every $t > 0$:

$$\mathbb{P}(Y_n - Y_0 \ge x) \le e^{-tx}\, \mathbb{E}\big[e^{t(Y_n - Y_0)}\big].$$

We bound the expectation by peeling off the last increment:

$$\begin{aligned} \mathbb{E}\big[e^{t(Y_n - Y_0)}\big] &= \mathbb{E}\Big[\mathbb{E}\big(e^{tD_n} \cdot e^{t(Y_{n-1} - Y_0)} \,\big|\, \mathcal{F}_{n-1}\big)\Big] \\ &= \mathbb{E}\Big[e^{t(Y_{n-1} - Y_0)}\, \mathbb{E}\big(e^{tD_n} \,\big|\, \mathcal{F}_{n-1}\big)\Big]. \end{aligned}$$

Since $\mathbb{E}(D_n \mid \mathcal{F}_{n-1}) = 0$ and $|D_n| \le K_n$, Lemma 10.12 applied conditionally gives

$$\mathbb{E}(e^{tD_n} \mid \mathcal{F}_{n-1}) \le e^{t^2 K_n^2 / 2}.$$

Therefore

$$\mathbb{E}\big[e^{t(Y_n - Y_0)}\big] \le e^{t^2 K_n^2/2}\, \mathbb{E}\big[e^{t(Y_{n-1} - Y_0)}\big].$$

Iterating this argument n times:

$$\mathbb{P}(Y_n - Y_0 \geq x) \leq \exp\Big(- tx + \frac{t^2}{2} \sum_{i=1}^{n} K_i^2 \Big).$$

Optimizing over $t > 0$: the minimum occurs at $t^* = x / \sum_{i=1}^n K_i^2$, giving

$$\mathbb{P}(Y_n - Y_0 \geq x) \leq \exp\Big(- \frac{x^2}{2 \sum_{i=1}^n K_i^2} \Big).$$

Applying the same argument to $\{-Y_n\}$ gives the same bound for $\mathbb{P}(Y_0 - Y_n \geq x)$, and the result follows by a union bound. □

Remark 10.13. When $K_i = K$ for all i, the Azuma–Hoeffding inequality becomes

$$\mathbb{P}(|Y_n - Y_0| \geq x) \leq 2e^{-x^2/(2nK^2)},$$

which gives sub-Gaussian tails with variance proxy nK^2.

Application: Concentration for bin packing

The Azuma–Hoeffding inequality is particularly powerful in combination with the **method of bounded differences**, which applies to functions of independent random variables that are not too sensitive to any single coordinate. The following example illustrates this technique.

Suppose we have n objects of independent random sizes $X_1, X_2, \ldots, X_n$, each drawn from the same distribution on $[0, 1]$, and an unlimited supply of unit-size boxes. Let B_n be the minimum number of boxes needed to pack all objects. It is known that there exists a constant β (depending on the distribution of X_i) such that $B_n/n \xrightarrow{a.s.} \beta$. We want to show that B_n concentrates around its mean.

Define $Y_i = \mathbb{E}(B_n \mid \mathcal{F}_i)$ where $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$. Then $\{Y_i\}_{i=0}^n$ is a martingale (this is Doob's martingale from Example 9.11), with $Y_0 = \mathbb{E}(B_n)$ and $Y_n = B_n$.

We claim that $|Y_i - Y_{i-1}| \leq 1$ for each i. To see this, let $B_n(i)$ denote the number of boxes needed to pack all objects **except** the ith. Clearly $B_n(i) \leq B_n \leq B_n(i)+1$ (the ith object can always be placed in a separate box). Taking conditional expectations:

$$\mathbb{E}(B_n(i) \mid \mathcal{F}_{i-1}) \leq Y_{i-1} \leq \mathbb{E}(B_n(i) \mid \mathcal{F}_{i-1}) + 1,$$
$$\mathbb{E}(B_n(i) \mid \mathcal{F}_i) \leq Y_i \leq \mathbb{E}(B_n(i) \mid \mathcal{F}_i) + 1.$$

The key observation is that $B_n(i)$ does not depend on X_i, so $\mathbb{E}(B_n(i) \mid \mathcal{F}_{i-1}) = \mathbb{E}(B_n(i) \mid \mathcal{F}_i)$. Combining the two inequalities gives $|Y_i - Y_{i-1}| \leq 1$.

By the Azuma–Hoeffding inequality with $K_i = 1$:

$$\mathbb{P}\big(|B_n - \mathbb{E}B_n| \geq x\big) \leq 2 \exp\Big(- \frac{x^2}{2n} \Big).$$

Setting $x = \varepsilon n$ and using $\mathbb{E}B_n/n \to \beta$:

$$\mathbb{P}(|B_n - \beta n| \geq \varepsilon n) \leq 2 \exp\big(- \tfrac{1}{2}\varepsilon^2 n(1 + o(1)) \big).$$

Thus the number of boxes concentrates around βn with deviations that are exponentially unlikely.

Exercise 10.19 applies this technique to the chromatic number of a random graph.

Application: Proof of Anscombe's theorem

In Chapter 8 we stated Anscombe's theorem (CLT for randomly stopped sums) in full generality, but only proved it when N_n is independent of the summands. We promised that the general case would follow from Doob's maximal inequality. We now deliver on that promise.

Theorem 3 (Anscombe's theorem, general case)**.** *Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 \in (0, \infty)$. Let $(N_n)_{n \geq 1}$ be positive integer-valued random variables (not necessarily independent of the X_i's) satisfying $N_n/n \xrightarrow{\mathbb{P}} \lambda$ for some $\lambda > 0$. Then*

$$\frac{S_{N_n} - N_n \mu}{\sigma \sqrt{N_n}} \xrightarrow{d} N(0,1).$$

Proof. Let $M_k = S_k - k\mu = \sum_{i=1}^{k} (X_i - \mu)$. This is a martingale with $\mathbb{E}(M_k^2) = k\sigma^2$. Write $m_n = \lfloor \lambda n \rfloor$. We decompose:

$$\frac{M_{N_n}}{\sigma\sqrt{N_n}} = \underbrace{\frac{M_{m_n}}{\sigma\sqrt{m_n}}}_{\text{(I)}} \cdot \underbrace{\frac{\sqrt{m_n}}{\sqrt{N_n}}}_{\text{(II)}} + \underbrace{\frac{M_{N_n} - M_{m_n}}{\sigma\sqrt{N_n}}}_{\text{(III)}}.$$

Term (I) converges to $N(0,1)$ in distribution by the classical CLT, since $m_n \to \infty$.

Term (II) satisfies $m_n/N_n = (\lambda n + O(1))/N_n \xrightarrow{\mathbb{P}} 1$ (since $N_n/n \xrightarrow{\mathbb{P}} \lambda$), so $\sqrt{m_n/N_n} \xrightarrow{\mathbb{P}} 1$.

By Slutsky's theorem, the product (I) $\times$ (II) $\xrightarrow{d} N(0,1)$. It remains to show:

Term (III) converges to 0 in probability. Fix $\varepsilon > 0$ and $\delta > 0$. Then

$$\mathbb{P}\Big(\frac{|M_{N_n} - M_{m_n}|}{\sigma\sqrt{N_n}} > \varepsilon\Big) \leq \mathbb{P}(|N_n - m_n| > \delta n) + \mathbb{P}\Big(\max_{|k - m_n| \leq \delta n} |M_k - M_{m_n}| > \varepsilon\sigma\sqrt{(\lambda - \delta)n}\Big),$$

where on the second event we used $N_n \geq (\lambda - \delta)n$ (which holds on the complement of the first event, for n large enough).

The first probability vanishes as $n \to \infty$ since $N_n/n \xrightarrow{\mathbb{P}} \lambda$.

For the second probability, we apply Corollary 10.9 (with $p = 2$) to the martingale $M'_j = M_{m_n + j} - M_{m_n}$, $j = 0, 1, \dots, \lfloor \delta n \rfloor$:

$$\mathbb{P}\Big(\max_{0 \leq j \leq \lfloor \delta n \rfloor} |M'_j| \geq \varepsilon\sigma\sqrt{(\lambda - \delta)n}\Big) \leq \frac{\mathbb{E}|M'_{\lfloor \delta n \rfloor}|^2}{\varepsilon^2 \sigma^2 (\lambda - \delta) n} = \frac{\lfloor \delta n \rfloor \sigma^2}{\varepsilon^2 \sigma^2 (\lambda - \delta) n} \leq \frac{\delta}{\varepsilon^2 (\lambda - \delta)}.$$

A similar bound handles $\max_{0\le j\le \lfloor \delta n\rfloor} |M_{m_n-j} - M_{m_n}|$. Therefore

$$\limsup_{n\to\infty} \mathbb{P}\Big(\frac{|M_{N_n} - M_{m_n}|}{\sigma\sqrt{N_n}} > \varepsilon \Big) \le \frac{2\delta}{\varepsilon^2(\lambda - \delta)}.$$

Since $\delta > 0$ was arbitrary, letting $\delta \to 0$ gives the desired conclusion. □

Remark 10.14. The heart of the argument is the maximal inequality estimate on Term (III): Doob's inequality gives us uniform control of the martingale M_k over a window of width δn around m_n, with a bound proportional to δ. This is precisely the "uniform control of the partial sum process" mentioned in Chapter 8.

10.3 Exercises

Homework

Exercise 10.15 (The stopped walk is not uniformly integrable). Let $(S_n)_{n\ge 0}$ be the simple symmetric random walk starting at $S_0 = 0$, and let $T = \inf\{n : S_n = 1\}$. It is known that $\mathbb{P}(T < \infty) = 1$ (see Exercise 10.20). Show directly that the family $\{S_{T\wedge n}\}_{n\ge 0}$ is **not** uniformly integrable.

Exercise 10.16 (Expected duration of gambler's ruin: unfair game). In the gambler's ruin setting with $p \neq 1/2$, use Wald's first identity to show that

$$\mathbb{E}(\tau) = \frac{A\,\mathbb{P}(S_\tau = A) - B\,\mathbb{P}(S_\tau = -B)}{p - q},$$

where $\mathbb{P}(S_\tau = A)$ is the ruin probability computed in the text.

(a) When $p > 1/2$, show that $\mathbb{E}(\tau) \to A/(p-q)$ as $B \to \infty$. Interpret this: why should a gambler with an edge and unlimited capital expect to win A dollars in finite expected time?

(b) When $p < 1/2$, show that $\mathbb{E}(\tau) \to \infty$ as $B \to \infty$.

(c) Verify that in the limit $p \to 1/2$, the formula recovers $\mathbb{E}(\tau) = AB$.

Exercise 10.17 (Pattern matching via martingales). Let $\xi_1, \xi_2, \dots$ be i.i.d. uniform on an alphabet of size s. Fix a pattern $p = p_1 p_2 \cdots p_\ell$ and let

$$T = \inf\{n \ge \ell : (\xi_{n-\ell+1}, \dots, \xi_n) = (p_1, \dots, p_\ell)\}$$

be the first time the pattern appears.

(a) Show that $\mathbb{E}(T) < \infty$.

(b) Consider the following **gambling team**: at each time $k \ge 1$, a new gambler G_k enters the casino with \$1. At time k, he bets his entire wealth that $\xi_k = p_1$; if he wins, his wealth is multiplied by s, otherwise he is out. If he survived, at time $k+1$ he bets his entire \$$s$ that $\xi_{k+1} = p_2$;

and so on. Let W_n denote the total wealth of all gamblers (including those who went home with \$0) at time n. Show that

$$M_n = W_n - n$$

is a martingale with $M_0 = 0$.

(c) Verify that the optional stopping theorem applies to M and T, and conclude that

$$\mathbb{E}(T) = \sum_{j=1}^{\ell} s^j \cdot \mathbf{1}\big\{(p_1, \dots, p_j) = (p_{\ell-j+1}, \dots, p_\ell)\big\}.$$

(d) Apply the formula to compute $\mathbb{E}(T)$ for

(i) the pattern HH (two heads in a row) with a fair coin;

(ii) the pattern HT with a fair coin.

Compare the two answers and explain heuristically why one is larger than the other.

Exercise 10.18 (Method of bounded differences). Let $X_1, \dots, X_n$ be independent random variables with X_i taking values in a measurable space $\mathcal{X}_i$, and let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be a bounded measurable function satisfying the **bounded-differences condition**: there exist constants $c_1, \dots, c_n > 0$ such that for every i and every $x_1, \dots, x_n, x_i'$,

$$\big|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)\big| \le c_i.$$

Prove that for every $t > 0$,

$$\mathbb{P}\big(|f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n)| \ge t\big) \le 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\Big).$$

Exercise 10.19 (Concentration of the chromatic number). Let G be a random graph on n vertices where each edge is included independently with probability p. Let $\chi(G)$ denote the chromatic number. Show that

$$\mathbb{P}(|\chi(G) - \mathbb{E}\chi(G)| \ge t) \le 2e^{-t^2/(2n)}.$$

Additional Exercises

Exercise 10.20 (Hitting time is finite). Consider a simple symmetric random walk S_n starting at $S_0 = 0$ and let $T = \inf\{n : S_n = 1\}$. Show that $\mathbb{P}(T < \infty) = 1$.

Exercise 10.21 (Hitting time has infinite expectation). Show that $\mathbb{E}(T) = \infty$ for the stopping time in Exercise 10.20.

Exercise 10.22 (Gambler's ruin: finite expected duration). In the gambler's ruin setting, verify the tail bound $\mathbb{P}(\tau > n) \le (1 - p^{A+B})^{\lfloor n/(A+B)\rfloor}$ and show that $\mathbb{E}(\tau) \le \sum_{n=0}^{\infty}(n+1)(1-p^{A+B})^n < \infty$.

Exercise 10.23 (Gambler's ruin: limiting probabilities). Analyze the limiting behavior of $\mathbb{P}(S_\tau = A)$ as $B \to \infty$ in both cases $p > 1/2$ and $p < 1/2$. Interpret the results.

Exercise 10.24 (Wald's Second Identity). Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}(X_1) = 0$ and $\mathbb{V}\text{ar}(X_1) = \sigma^2 < \infty$. If T is a stopping time with $\mathbb{E}(T) < \infty$, show that $\mathbb{E}(S_T^2) = \sigma^2 \mathbb{E}(T)$.

Chapter 11

Martingale Convergence Theorems

When does a martingale converge? This chapter answers the question in stages of increasing depth. We begin with L^2-bounded martingales, where orthogonality of increments gives both a.s. and L^2 convergence. We then prove Doob's fundamental result that every L^1-bounded martingale converges almost surely, using the upcrossing inequality—a beautifully combinatorial argument. However, L^1 boundedness alone does not guarantee L^1 convergence; for that, we need uniform integrability. The chapter develops the theory of uniformly integrable martingales, culminating in the closure theorem: a martingale is UI if and only if it has the form $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ for some integrable X_∞. We close with regular stopping times, which extend the optional stopping identity to unbounded stopping times.

11.1 L^2-bounded martingale convergence

Our first convergence theorem is a generalization of Kolmogorov's theorem on the convergence of series of independent random variables. The proof uses orthogonality of martingale differences and Doob's maximal inequality.

Theorem 11.1 (L^2 Martingale Convergence)**.** *Suppose that $\{M_n\}$ is a martingale satisfying $\sup_{n\geq 0} \mathbb{E}(M_n^2) \leq B < \infty$. Then there exists a random variable M_∞ with $\mathbb{E}(M_\infty^2) \leq B$ such that $M_n \xrightarrow{a.s.} M_\infty$ and $M_n \xrightarrow{L^2} M_\infty$.*

Proof. Let $D_k = M_k - M_{k-1}$ denote the martingale differences.

Step 1: Orthogonality. For $i < j$, the differences D_i and D_j are orthogonal in L^2:

$$\mathbb{E}(D_i D_j) = \mathbb{E}\big(D_i\, \mathbb{E}(D_j \mid \mathcal{F}_{j-1})\big) = \mathbb{E}(D_i \cdot 0) = 0.$$

Therefore, assuming $M_0 = 0$ for simplicity (the general case follows by con-

sidering $M_n - M_0$),

$$\mathbb{E}(M_n^2) = \sum_{k=1}^{n} \mathbb{E}(D_k^2) \le B, \tag{11.1}$$

so the series $\sum_{k=1}^{\infty} \mathbb{E}(D_k^2)$ converges.

Step 2: Almost sure convergence. For any pair $a < b$, define

$$A_{ab} = \Big\{\omega : \liminf_{n\to\infty} M_n(\omega) \le a < b \le \limsup_{n\to\infty} M_n(\omega)\Big\}.$$

The event that M_n fails to converge is $A = \bigcup_{a<b,\, a,b\in\mathbb{Q}} A_{ab}$, so it suffices to show $\mathbb{P}(A_{ab}) = 0$ for each rational $a < b$.

Setting $\varepsilon = (b-a)/2$, we have for all $m \ge 0$:

$$A_{ab} \subset \Big\{ \sup_{n\ge m} |M_n - M_m| \ge \varepsilon\Big\}.$$

Applying Corollary 10.9 with $p = 2$ to the martingale $M'_k = M_{m+k} - M_m$:

$$\mathbb{P}(A_{ab}) \le \frac{1}{\varepsilon^2} \sup_{n\ge m} \mathbb{E}(|M_n - M_m|^2) = \frac{1}{\varepsilon^2} \sum_{k=m+1}^{\infty} \mathbb{E}(D_k^2), \tag{11.2}$$

where the equality uses orthogonality. Since the series $\sum \mathbb{E}(D_k^2)$ converges, the tail $\sum_{k=m+1}^{\infty} \mathbb{E}(D_k^2) \to 0$ as $m \to \infty$. Since m is arbitrary, $\mathbb{P}(A_{ab}) = 0$.

Step 3: L^2 convergence. Let M_∞ denote the a.s. limit. By orthogonality,

$$\mathbb{E}(|M_\infty - M_n|^2) = \sum_{k=n+1}^{\infty} \mathbb{E}(D_k^2) \to 0 \quad \text{as } n \to \infty,$$

so $M_n \xrightarrow{L^2} M_\infty$. In particular, $\mathbb{E}(M_\infty^2) = \lim_{n\to\infty} \mathbb{E}(M_n^2) \le B$. □

Remark 11.2 (On the passage to the infinite sum in Step 3). The identity

$$\mathbb{E}\big((M_\infty - M_n)^2\big) = \sum_{k=n+1}^{\infty} \mathbb{E}(D_k^2) \tag{11.3}$$

used in Step 3 deserves a closer look. Orthogonality of the differences gives the **finite** identity

$$\mathbb{E}\big((M_N - M_n)^2\big) = \sum_{k=n+1}^{N} \mathbb{E}(D_k^2), \qquad n < N, \tag{11.4}$$

which is a purely algebraic consequence of $\mathbb{E}(D_i D_j) = 0$ for $i \ne j$. Passing from (11.4) to (11.3) requires letting $N \to \infty$ inside an expectation, and almost sure convergence $M_N \xrightarrow{a.s.} M_\infty$ alone does **not** justify this: one must either bound the pre-limit quantity (Fatou) or exhibit an L^2 limit by another route (completeness).

Route 1: Fatou plus norm continuity. From $M_N \xrightarrow{a.s.} M_\infty$ we get $(M_N - M_n)^2 \to (M_\infty - M_n)^2$ a.s., so Fatou's lemma gives the **inequality**

$$\mathbb{E}\big((M_\infty - M_n)^2\big) \le \liminf_{N\to\infty} \mathbb{E}\big((M_N - M_n)^2\big) = \sum_{k=n+1}^{\infty} \mathbb{E}(D_k^2),$$

with the last step following from (11.4). Since $\sum_k \mathbb{E}(D_k^2) < \infty$, the right-hand side tends to 0 as $n \to \infty$. This already proves $M_n \xrightarrow{L^2} M_\infty$ and, in particular, $M_\infty \in L^2$. Once L^2 convergence is established, continuity of the L^2 norm upgrades the inequality to the equality (11.3):

$$\|M_\infty - M_n\|_2 = \lim_{N\to\infty} \|M_N - M_n\|_2 = \Big(\sum_{k=n+1}^{\infty} \mathbb{E}(D_k^2)\Big)^{1/2}.$$

Route 2: completeness of L^2. A more structural argument bypasses a.s. convergence entirely. From (11.4), for $n < m$,

$$\|M_m - M_n\|_2^2 = \sum_{k=n+1}^{m} \mathbb{E}(D_k^2) \to 0 \quad \text{as } n, m \to \infty,$$

because $\sum_k \mathbb{E}(D_k^2) < \infty$. Hence $\{M_n\}$ is Cauchy in L^2. By completeness of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, there exists $\widetilde{M}_\infty \in L^2$ with $M_n \xrightarrow{L^2} \widetilde{M}_\infty$, and by continuity of $\|\cdot\|_2$,

$$\|\widetilde{M}_\infty - M_n\|_2^2 = \lim_{m\to\infty} \|M_m - M_n\|_2^2 = \sum_{k=n+1}^{\infty} \mathbb{E}(D_k^2).$$

It remains to identify $\widetilde{M}_\infty$ with the a.s. limit M_∞. Both L^2 convergence and a.s. convergence imply convergence in probability, so $M_n \xrightarrow{\mathbb{P}} M_\infty$ and $M_n \xrightarrow{\mathbb{P}} \widetilde{M}_\infty$; by uniqueness of the limit in probability, $\widetilde{M}_\infty = M_\infty$ a.s.

Moral. Finite orthogonality is algebra; the infinite-sum identity (11.3) is a limit theorem. In Route 1 we control the a.s. limit by an L^2 quantity via Fatou; in Route 2 we produce an L^2 limit directly from completeness and only at the end match it with the a.s. limit.

Remark 11.3 (Extension to sub- and supermartingales). Theorem 11.1 extends to L^2-bounded submartingales and supermartingales: if $\{X_n\}$ is a submartingale (or supermartingale) with $\sup_n \mathbb{E}(X_n^2) < \infty$, then X_n converges a.s. and in L^2 to some $X_\infty \in L^2$. The supermartingale case follows from the submartingale case by considering $-X_n$. A clean proof in the non-negative case — which uses Doob's maximal inequality and covers the main applications — is given in Appendix K; the general case uses the Doob decomposition and can be found in Durrett (2019), §4.4.

11.2 Upcrossing inequality and L^1-bounded convergence

The L^2-bounded convergence theorem above requires a second moment condition. The following deeper result, due to Doob, requires only L^1 boundedness.

The proof uses a completely different technique: counting **upcrossings**.

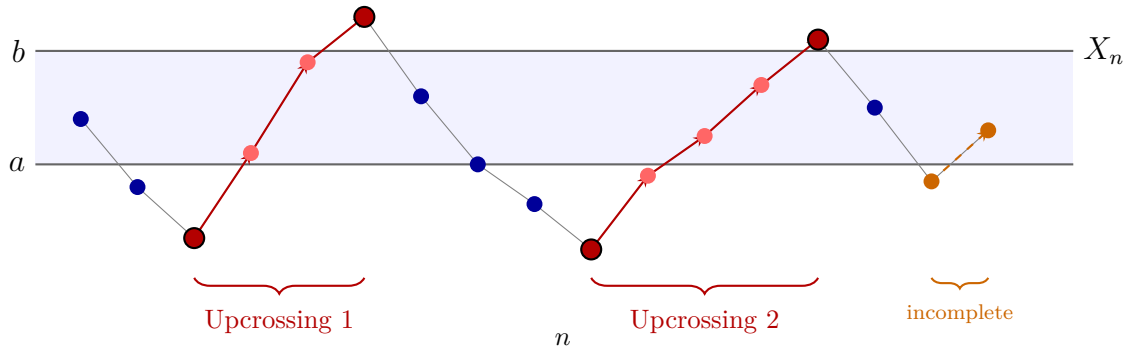


Figure 11.1: Upcrossings of the interval $[a, b]$. The red path segments show two completed upcrossings (from below a to above b). The dashed orange segment shows an incomplete upcrossing at the end.

Suppose $\{X_n\}$ is a sequence of real numbers and $a < b$. An **upcrossing** of the interval $[a, b]$ is a pair of times $s < t$ such that $X_s \le a$ and $X_t \ge b$, with no previous upcrossing completing after time s (see Figure 11.1). More precisely, we define the upcrossing times inductively: let $\sigma_1 = \inf\{n \ge 0 : X_n \le a\}$, $\tau_1 = \inf\{n > \sigma_1 : X_n \ge b\}$, and for $k \ge 2$, $\sigma_k = \inf\{n > \tau_{k-1} : X_n \le a\}$, $\tau_k = \inf\{n > \sigma_k : X_n \ge b\}$. Then $U_{a,b}(n)$ is the number of completed upcrossings by time n, and $U_{a,b} = \lim_{n\to\infty} U_{a,b}(n)$.

The intuition is that of a stock trader who buys when the price drops to a and sells when it rises to b. Each completed upcrossing yields a profit of at least $b - a$.

Theorem 11.4 (Doob's Upcrossing Inequality)**.** *Let $\{X_n\}$ be a supermartingale. Then for every $a < b$ and every n,*

$$\mathbb{E}\, U_{a,b}(n) \le \frac{\mathbb{E}(a - X_n)^+}{b - a} \le \frac{|a| + \mathbb{E}|X_n|}{b - a}. \tag{11.5}$$

This theorem implies that for L^1-bounded supermartingales, the total number of upcrossings is finite a.s. Indeed, $\mathbb{E}\, U_{a,b}(n)$ is non-decreasing in n and bounded by (11.5). By the monotone convergence theorem, $\mathbb{E}\, U_{a,b} = \lim_{n\to\infty} \mathbb{E}\, U_{a,b}(n) < \infty$, which implies $U_{a,b} < \infty$ a.s.

Proof. Define a predictable sequence $\{C_n\}$ that represents the "upcrossing trading strategy": $C_n = 1$ when the trader holds the stock at time n (i.e., the price has dropped below a and has not yet risen above b), and $C_n = 0$ otherwise. Formally, $C_1 = \mathbf{1}_{\{X_0 \le a\}}$ and for $n \ge 2$:

$$C_n = \mathbf{1}_{\{C_{n-1}=1\}} \mathbf{1}_{\{X_{n-1} < b\}} + \mathbf{1}_{\{C_{n-1}=0\}} \mathbf{1}_{\{X_{n-1} \le a\}}.$$

In words: the trader stays in ($C_n = 1$) if he was in and the price hasn't reached b yet, or enters ($C_n = 1$) if he was out and the price drops to a or below. Note that C_n is $\mathcal{F}_{n-1}$-measurable, so $\{C_n\}$ is predictable.

Define $Y_n = (C \cdot X)_n = \sum_{k=1}^{n} C_k (X_k - X_{k-1})$. This is the trader's cumulative profit. Since $\{X_n\}$ is a supermartingale and $0 \le C_n \le 1$ is predictable, $\{Y_n\}$ is also a supermartingale, so $\mathbb{E}(Y_n) \le \mathbb{E}(Y_0) = 0$.

On the other hand, each completed upcrossing contributes at least $b-a$ to the profit, while the only possible loss comes from an incomplete upcrossing at the end (if the trader is holding the stock at time n with the price below b). This loss is at most $(a - X_n)^+$. Therefore:

$$Y_n \geq (b-a)\, U_{a,b}(n) - (a - X_n)^+.$$

Taking expectations and using $\mathbb{E}(Y_n) \leq 0$:

$$(b-a)\, \mathbb{E}\, U_{a,b}(n) \leq \mathbb{E}(a - X_n)^+.$$

The second inequality in (11.5) follows from $(a-x)^+ \leq |a| + |x|$. □

We are now ready for the main convergence theorem.

Definition 11.5. A supermartingale $\{X_n\}$ is **L^1-bounded** if $\sup_n \mathbb{E}|X_n| < \infty$.

Theorem 11.6 (Doob's Martingale Convergence Theorem)**.** *If $\{X_n\}$ is an L^1-bounded supermartingale, then X_n converges almost surely to a limit X_∞ with $\mathbb{E}|X_\infty| < \infty$.*

Proof. By the upcrossing inequality, for every pair of rationals $a < b$, the number of upcrossings $U_{a,b}$ is finite a.s. Define

$$A_{ab} = \Big\{\omega : \liminf_{n\to\infty} X_n(\omega) \leq a < b \leq \limsup_{n\to\infty} X_n(\omega)\Big\}.$$

On the event A_{ab}, the sequence X_n crosses the interval $[a,b]$ infinitely often, so $U_{a,b} = \infty$. Since $U_{a,b} < \infty$ a.s., we have $\mathbb{P}(A_{ab}) = 0$.

The event that X_n fails to converge to a limit in $[-\infty, \infty]$ is

$$A = \bigcup_{a<b,\, a,b\in\mathbb{Q}} A_{ab}.$$

Since this is a countable union of null sets, $\mathbb{P}(A) = 0$, so X_n converges a.s. to a limit $X_\infty \in [-\infty, \infty]$.

It remains to show that X_∞ is integrable. By Fatou's lemma:

$$\mathbb{E}(|X_\infty|) = \mathbb{E}(\liminf_n |X_n|) \leq \liminf_n \mathbb{E}(|X_n|) \leq \sup_n \mathbb{E}(|X_n|) < \infty.$$

In particular, $|X_\infty| < \infty$ a.s. □

Corollary 11.7. *Every non-negative supermartingale converges almost surely to an integrable limit.*

Proof. If $\{X_n\}$ is a non-negative supermartingale, then $\mathbb{E}|X_n| = \mathbb{E}(X_n) \leq \mathbb{E}(X_0) < \infty$, so it is L^1-bounded. □

This corollary justifies the claims made earlier about the Pólya urn (Example 9.12) and branching process (Example 9.13) martingales.

Example 11.8 (L^1-bounded martingale that does not converge in L^1). Let S_n be a symmetric simple random walk starting at $S_0 = 1$, and let

$$N = \inf\{n : S_n = 0\}.$$

Set $X_n = S_{N \wedge n}$. By Theorem 10.1, $\{X_n\}$ is a martingale, and stopping at 0 keeps it non-negative.

The martingale is L^1-bounded. This is the point of the example, so we verify it explicitly. Because $X_n \geq 0$, we have $|X_n| = X_n$, and the martingale property gives

$$\mathbb{E}|X_n| \;=\; \mathbb{E}(X_n) \;=\; \mathbb{E}(X_0) \;=\; 1 \qquad \text{for every } n \geq 0.$$

Hence $\sup_n \mathbb{E}|X_n| = 1 < \infty$. Note how cheaply this came: non-negativity collapses $\mathbb{E}|X_n|$ to $\mathbb{E}(X_n)$, and the martingale property pins the latter to $\mathbb{E}(X_0)$. No estimate, no inequality.

Almost sure convergence. Since $\{X_n\}$ is L^1-bounded, Doob's theorem (Theorem 11.6) gives $X_n \xrightarrow{a.s.} X_\infty$ for some integrable X_∞. By recurrence of the symmetric walk, $\mathbb{P}(N < \infty) = 1$, so $X_\infty = S_N = 0$ a.s.

Failure of L^1 convergence. If $X_n \to X_\infty$ in L^1, then in particular $\mathbb{E}(X_n) \to \mathbb{E}(X_\infty)$. But $\mathbb{E}(X_n) = 1$ for every n while $\mathbb{E}(X_\infty) = 0$, so L^1 convergence fails.

This is the example to keep in mind: L^1-boundedness alone is **not** enough to upgrade a.s. convergence to L^1 convergence, even for a martingale, even one as well-behaved as a non-negative one. The mechanism of failure is that the family $\{X_n\}$ is not uniformly integrable: a unit of expected mass is supported on the shrinking event $\{N > n\}$ at growing values of S_n, and escapes to infinity rather than landing on the limit. Uniform integrability is exactly the condition that closes this gap, and it is the subject of Section 11.4.

11.3 L^p convergence for $p > 1$

The situation is much better for $p > 1$ than for $p = 1$: L^p boundedness implies both a.s. convergence and L^p convergence. The key reason is that L^p boundedness for $p > 1$ implies uniform integrability.

Theorem 11.9 (L^p Martingale Convergence, $p > 1$)**.** *Let $\{X_n\}$ be a martingale with $\sup_n \mathbb{E}|X_n|^p < \infty$ for some $p > 1$. Then there exists $X_\infty \in L^p$ such that $X_n \xrightarrow{a.s.} X_\infty$ and $X_n \xrightarrow{L^p} X_\infty$.*

Proof. Since $\mathbb{E}|X_n| \leq (\mathbb{E}|X_n|^p)^{1/p} \leq C$ (by Jensen's inequality), the martingale is L^1-bounded. By Doob's convergence theorem (Theorem 11.6), $X_n \xrightarrow{a.s.} X_\infty$ for some integrable X_∞.

It remains to show L^p convergence. The key tool is Doob's L^p maximal inequality (stated in the remark after Corollary 10.9), which gives

$$\mathbb{E}\Big(\sup_{k \geq 0} |X_k|^p\Big) \leq \Big(\frac{p}{p-1}\Big)^p \sup_n \mathbb{E}|X_n|^p < \infty.$$

In particular, $X^* = \sup_{k \geq 0} |X_k|$ is in L^p, and since $|X_\infty| \leq X^*$ a.s., we have $|X_n - X_\infty|^p \leq (2X^*)^p$ for all n, with $(2X^*)^p$ integrable. Since $X_n \to X_\infty$ a.s., we have $|X_n - X_\infty|^p \to 0$ a.s., and the dominated convergence theorem gives $\mathbb{E}|X_n - X_\infty|^p \to 0$. □

Remark 11.10. For $p = 2$, this result is consistent with Theorem 11.1, but the proof there was more elementary (using orthogonality directly). The present proof works for all $p > 1$ but relies on Doob's L^p maximal inequality, which we stated without proof in Section 10.2.

The case $p = 1$ is genuinely different: L^1 boundedness does **not** imply uniform integrability, and does not guarantee L^1 convergence, as Example 11.8 shows. The additional condition needed for L^1 convergence is the subject of the next section.

11.4 Uniformly Integrable Martingales

We have seen that L^1-bounded martingales converge almost surely, but may fail to converge in L^1 (Example 11.8). The missing ingredient is uniform integrability. Recall from Chapter 5 that a family $\mathcal{C}$ of random variables is **uniformly integrable** (UI) if

$$\sup_{X \in \mathcal{C}} \mathbb{E}(|X| \mathbf{1}_{\{|X| \geq K\}}) \to 0 \quad \text{as } K \to \infty,$$

and that the Vitali convergence theorem states: $X_n \to X$ in L^1 if and only if $X_n \xrightarrow{\mathbb{P}} X$ and $\{X_n\}$ is UI. In this section, we apply these ideas to martingales.

When is a martingale uniformly integrable?

There are two important sufficient conditions for a martingale to be UI.

Condition 1: L^p boundedness for $p > 1$. If $\sup_n \mathbb{E}|X_n|^p < \infty$ for some $p > 1$, then $\{X_n\}$ is UI. This was proved in Chapter 5 (the de la Vallée-Poussin criterion). In this case, Theorem 11.9 already gives L^p (and hence L^1) convergence.

Condition 2: Doob's martingale. If $X_n = \mathbb{E}(X \mid \mathcal{F}_n)$ for some integrable X, then $\{X_n\}$ is UI. This is the more interesting case, and we prove it below.

Theorem 11.11. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ range over all sub-σ-algebras of $\mathcal{F}$. Then the family $\{\mathbb{E}(X \mid \mathcal{G}) : \mathcal{G} \subset \mathcal{F}\}$ is uniformly integrable.*

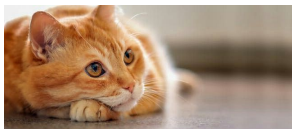
The proof uses the following lemma, which says that integration against an L^1 function is "absolutely continuous" with respect to $\mathbb{P}$.

Lemma 11.12 (Absolute continuity of integration)**.** *If $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\mathbb{P}(A) < \delta \implies \mathbb{E}(|X|; A) < \varepsilon$$

for all $A \in \mathcal{F}$. (Here $\mathbb{E}(Y; A) = \mathbb{E}(Y \mathbf{1}_A)$.)

Proof. Suppose the conclusion fails. Then for some $\varepsilon_0 > 0$, there exist sets $A_n \in \mathcal{F}$ with $\mathbb{P}(A_n) < 2^{-n}$ and $\mathbb{E}(|X|; A_n) \geq \varepsilon_0$.



Let

$$H = \limsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n = \{\omega : \omega \in A_n \text{ for infinitely many } n\}.$$

By the first Borel–Cantelli lemma (since $\sum \mathbb{P}(A_n) < \infty$), $\mathbb{P}(H) = 0$.

However, $|X|\mathbf{1}_{A_n} \leq |X|$ which is integrable, so by the reverse Fatou lemma:

$$\mathbb{E}(|X|; H) \geq \limsup_{n\to\infty} \mathbb{E}(|X|; A_n) \geq \varepsilon_0.$$

But $\mathbb{P}(H) = 0$ implies $\mathbb{E}(|X|; H) = 0$, a contradiction. □

Proof of Theorem 11.11. Write $X_{\mathcal{G}} = \mathbb{E}(X \mid \mathcal{G})$. Let $\varepsilon > 0$. Choose $\delta > 0$ as in Lemma 11.12, and choose K so that $K^{-1}\mathbb{E}|X| < \delta$.

By Jensen's inequality for conditional expectations, $|X_{\mathcal{G}}| \leq \mathbb{E}(|X| \mid \mathcal{G})$, so $\mathbb{E}|X_{\mathcal{G}}| \leq \mathbb{E}|X|$. By Markov's inequality:

$$\mathbb{P}(|X_{\mathcal{G}}| > K) \leq \frac{\mathbb{E}|X_{\mathcal{G}}|}{K} \leq \frac{\mathbb{E}|X|}{K} < \delta.$$

Since $\{|X_{\mathcal{G}}| > K\} \in \mathcal{G}$, we can use Jensen's inequality and the definition of conditional expectation:

$$\begin{aligned}\mathbb{E}(|X_{\mathcal{G}}|; \{|X_{\mathcal{G}}| > K\}) &\leq \mathbb{E}\Big[\mathbb{E}(|X| \mid \mathcal{G}); \{|X_{\mathcal{G}}| > K\}\Big] \\ &= \mathbb{E}(|X|; \{|X_{\mathcal{G}}| > K\}) < \varepsilon,\end{aligned}$$

where the last inequality follows from Lemma 11.12 since $\mathbb{P}(|X_{\mathcal{G}}| > K) < \delta$. This holds for every $\mathcal{G}$, so the family is UI. □

Corollary 11.13. *If $X \in L^1$ and $X_n = \mathbb{E}(X \mid \mathcal{F}_n)$, then $\{X_n\}_{n\geq 0}$ is a uniformly integrable martingale.*

The closure theorem

The remarkable fact is that the converse also holds: **every** UI martingale arises as a Doob martingale. This gives a complete characterization.

Theorem 11.14 (Closure Theorem for UI Martingales)**.** *Let $\{X_n\}$ be a uniformly integrable martingale with respect to $\{\mathcal{F}_n\}$. Then:*

(i) $X_\infty = \lim_{n\to\infty} X_n$ exists a.s. and in L^1.

(ii) For every $n \geq 0$, $X_n = \mathbb{E}(X_\infty \mid \mathcal{F}_n)$ a.s.

Proof. (i) Since UI implies L^1-boundedness, Doob's convergence theorem (Theorem 11.6) gives $X_n \xrightarrow{a.s.} X_\infty$ for some integrable X_∞. Since $\{X_n\}$ is UI and $X_n \xrightarrow{a.s.} X_\infty$, the Vitali convergence theorem (Chapter 5) upgrades this to L^1 convergence.

(ii) Fix $n \geq 0$ and let $A \in \mathcal{F}_n$. For any $r \geq n$, the martingale property gives $\mathbb{E}(X_r; A) = \mathbb{E}(X_n; A)$. On the other hand,

$$|\mathbb{E}(X_r; A) - \mathbb{E}(X_\infty; A)| \leq \mathbb{E}|X_r - X_\infty| \to 0$$

as $r \to \infty$, by the L^1 convergence from part (i). Therefore

$$\mathbb{E}(X_\infty; A) = \lim_{r\to\infty} \mathbb{E}(X_r; A) = \mathbb{E}(X_n; A)$$

for every $A \in \mathcal{F}_n$. By the definition of conditional expectation, this means $X_n = \mathbb{E}(X_\infty \mid \mathcal{F}_n)$ a.s. □

Remark 11.15. Combining Corollary 11.13 and Theorem 11.14, we have the following equivalence for a martingale $\{X_n\}$:

$$\{X_n\} \text{ is UI} \iff X_n = \mathbb{E}(X_\infty \mid \mathcal{F}_n) \text{ for some } X_\infty \in L^1.$$

This justifies calling Doob's martingale (Example 9.11) the "canonical" or "most general" form of a UI martingale. The variable X_∞ is called the **closure** of the martingale.

An important consequence is Lévy's upward theorem (Exercise 11.23): if the filtration $\mathcal{F}_n$ increases to $\mathcal{F}_\infty$, then $\mathbb{E}(X \mid \mathcal{F}_n) \to \mathbb{E}(X \mid \mathcal{F}_\infty)$ a.s. and in L^1.

Regular stopping times

In many applications, the martingale $\{X_n\}$ itself is not uniformly integrable, but the **stopped** martingale $\{X_{n\wedge T}\}$ is. This motivates the following definition.

Definition 11.16. A stopping time T is called **regular** for the martingale $\{X_n\}$ if the stopped martingale $\{X_{n\wedge T}\}$ is uniformly integrable.

If T is regular for $\{X_n\}$, then by the closure theorem, $X_{n\wedge T} \to X_T$ in L^1 and $X_{n\wedge T} = \mathbb{E}(X_T \mid \mathcal{F}_n)$. In particular, $\mathbb{E}(X_T) = \mathbb{E}(X_0)$, so the optional stopping identity holds.

Example 11.17. Any bounded stopping time $T \leq N$ is regular for any martingale, since $\{X_{n\wedge T}\}$ is eventually constant and hence trivially UI. More generally, if $|X_{n\wedge T}| \leq C$ a.s. for all n, then T is regular.

The following theorem gives a useful criterion for regularity in the i.i.d. setting.

Theorem 11.18. *Let $S_n = Y_1 + \cdots + Y_n$, where Y_k are i.i.d. with $\mathbb{E}Y_k = 0$ and $\mathbb{E}|Y_k| < \infty$. Let T be a stopping time with $\mathbb{E}(T) < \infty$. Then:*

(i) T is regular for $\{S_n\}$.

(ii) If additionally $\mathbb{E}(Y_1^2) = \sigma^2 < \infty$, then T is regular for $\{S_n^2 - n\sigma^2\}$.

Proof. (i) Since $\mathbb{E}(T) < \infty$, we have $\mathbb{P}(T < \infty) = 1$ and $S_{T\wedge n} \xrightarrow{a.s.} S_T$. We will show that $S_{T\wedge n} \to S_T$ in L^1, which implies that $\{S_{T\wedge n}\}$ is UI (by the Vitali theorem in reverse: L^1 convergence implies UI).

We bound

$$|S_{T\wedge n} - S_T| \le \sum_{j=n+1}^{\infty} |Y_j| \mathbf{1}_{\{j\le T\}} =: \xi_{n+1}.$$

Clearly $\xi_{n+1} \to 0$ a.s. (since the sum is zero once $n \ge T$), and $\xi_{n+1} \le \xi_1$ for all n. It suffices to show $\mathbb{E}(\xi_1) < \infty$, after which $\mathbb{E}(\xi_{n+1}) \to 0$ by dominated convergence. We compute:

$$\mathbb{E}(\xi_1) = \sum_{j=1}^{\infty} \mathbb{E}\big(|Y_j| \mathbf{1}_{\{j\le T\}}\big).$$

Since $\{j \le T\} = \{T \le j-1\}^c \in \mathcal{F}_{j-1}$ and Y_j is independent of $\mathcal{F}_{j-1}$:

$$\mathbb{E}\big(|Y_j| \mathbf{1}_{\{j\le T\}}\big) = \mathbb{E}|Y_j| \cdot \mathbb{P}(j \le T) = \mathbb{E}|Y_1| \cdot \mathbb{P}(T \ge j).$$

Summing: $\mathbb{E}(\xi_1) = \mathbb{E}|Y_1| \cdot \mathbb{E}(T) < \infty$.

(ii) We show that $S_{T\wedge n} \to S_T$ in L^2, which implies L^1 convergence and hence regularity for both $\{S_n\}$ and $\{S_n^2 - n\sigma^2\}$.

By orthogonality (since $Y_j \mathbf{1}_{\{j\le T\}}$ and $Y_i \mathbf{1}_{\{i\le T\}}$ are uncorrelated for $i \ne j$, using the independence of Y_j from $\mathcal{F}_{j-1} \supset \sigma(\{j \le T\})$):

$$\mathbb{E}(S_{T\wedge n} - S_T)^2 = \sum_{j=n+1}^{\infty} \mathbb{E}\big(Y_j^2 \mathbf{1}_{\{j\le T\}}\big) = \mathbb{E}(Y_1^2) \sum_{j=n+1}^{\infty} \mathbb{P}(T \ge j) \to 0$$

as $n \to \infty$, since $\sum_{j=1}^{\infty} \mathbb{P}(T \ge j) = \mathbb{E}(T) < \infty$.

Now, $S_{T\wedge n} \to S_T$ in L^2 implies $S_{T\wedge n}^2 \to S_T^2$ in L^1 (since $\big|\|f_n\|_2^2 - \|f\|_2^2\big| = \big|\langle f_n - f, f_n + f\rangle\big| \le \|f_n - f\|_2 \|f_n + f\|_2$, and both factors are bounded). Also, $T \wedge n \to T$ in L^1 by monotone convergence. Therefore $S_{T\wedge n}^2 - (T \wedge n)\sigma^2 \to S_T^2 - T\sigma^2$ in L^1, so $\{S_{n\wedge T}^2 - (n \wedge T)\sigma^2\}$ is UI. □

Remark 11.19. Theorem 11.18 gives a unified perspective on the optional stopping arguments of Chapter 10.1. In gambler's ruin, the martingales $\{S_{n\wedge\tau}\}$ and $\{M_{n\wedge\tau}\}$ were handled directly via boundedness and domination—conditions (ii) and (iii) of the optional stopping theorem. The present theorem subsumes both: from $\mathbb{E}(\tau) = AB < \infty$, it concludes that τ is regular for $\{S_n\}$ and $\{S_n^2 - n\}$, which is condition (iv)—strictly more general than (ii) and (iii). For applications where the stopped process is **neither** bounded **nor** dominated by an integrable random variable, this UI route via $\mathbb{E}T < \infty$ becomes essential.

Example 11.20 (Wald's second identity). Let $Y_1, Y_2, \ldots$ be i.i.d. with $\mathbb{E}Y_1 = 0$ and $\mathbb{V}\mathrm{ar}(Y_1) = \sigma^2 < \infty$, and set $S_n = Y_1 + \cdots + Y_n$. Let T be a stopping time (with respect to $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$) satisfying $\mathbb{E}T < \infty$. Then

$$\mathbb{E}S_T^2 = \sigma^2 \mathbb{E}T.$$

Consider the quadratic martingale $M_n = S_n^2 - n\sigma^2$. The stopped process $M_{T\wedge n}$ is **not** uniformly bounded: on the event $\{T > n\}$, the term $n\sigma^2$ grows

without limit. Likewise, no obvious integrable random variable dominates $\sup_n |M_{T\wedge n}|$—a direct dominator would require $\mathbb{E}\sup_{k\le T} S_k^2 < \infty$, which by Doob's L^2 maximal inequality (applied to the stopped martingale) is equivalent to $\sigma^2\mathbb{E}T < \infty$, exactly the regularity condition. So neither condition (ii) nor (iii) of the optional stopping theorem applies cleanly.

The clean argument is via regularity. By Theorem 11.18(ii), T is regular for $\{M_n\}$, hence $M_{T\wedge n} \to M_T$ in L^1. Therefore

$$0 = \mathbb{E}M_0 = \lim_{n\to\infty} \mathbb{E}M_{T\wedge n} = \mathbb{E}M_T = \mathbb{E}S_T^2 - \sigma^2\mathbb{E}T,$$

which gives the identity.

11.5 Lévy's 0-1 law

As a further application of UI martingales, we derive Lévy's 0-1 law, which strengthens Kolmogorov's 0-1 law by describing how conditional probabilities of tail events converge.

Theorem 11.21 (Lévy's 0-1 law)**.** *Let $X_1, X_2, \ldots$ be independent random variables, and let A be a tail event, i.e., $A \in \mathcal{T} = \bigcap_n \sigma(X_{n+1}, X_{n+2}, \ldots)$. Then*

$$\mathbb{P}(A \mid X_1, \ldots, X_n) \xrightarrow{a.s.} \mathbf{1}_A \quad \text{as } n \to \infty.$$

Proof. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The process $M_n = \mathbb{P}(A \mid \mathcal{F}_n) = \mathbb{E}(\mathbf{1}_A \mid \mathcal{F}_n)$ is Doob's martingale for $\mathbf{1}_A$. By Theorem 11.11, $\{M_n\}$ is UI, and by the closure theorem (Theorem 11.14), $M_n \to \mathbb{E}(\mathbf{1}_A \mid \mathcal{F}_\infty)$ a.s., where $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$.

Since $A \in \mathcal{T}$ and $X_1, X_2, \ldots$ are independent, A is independent of $\mathcal{F}_n$ for every n. By a monotone class argument (or Lévy's upward theorem, Exercise 11.23), $\mathbf{1}_A$ is $\mathcal{F}_\infty$-measurable (since $A \in \sigma(X_1, X_2, \ldots) = \mathcal{F}_\infty$). Therefore $\mathbb{E}(\mathbf{1}_A \mid \mathcal{F}_\infty) = \mathbf{1}_A$, and

$$\mathbb{P}(A \mid X_1, \ldots, X_n) \to \mathbf{1}_A \quad \text{a.s.}$$

□

Remark 11.22. Lévy's 0-1 law says more than Kolmogorov's: not only is $\mathbb{P}(A) \in \{0, 1\}$, but the conditional probability $\mathbb{P}(A \mid X_1, \ldots, X_n)$ converges to the indicator $\mathbf{1}_A$ almost surely. In other words, as we observe more and more data, our uncertainty about any tail event vanishes completely.

11.6 Exercises

Homework

Exercise 11.23 (Lévy's upward theorem). Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\{\mathcal{F}_n\}$ be a filtration with $\mathcal{F}_\infty = \sigma\big(\bigcup_n \mathcal{F}_n\big)$. Show that

$$\mathbb{E}(X \mid \mathcal{F}_n) \to \mathbb{E}(X \mid \mathcal{F}_\infty) \qquad \text{a.s. and in } L^1.$$

Exercise 11.24 (Convergence of conditional probabilities, and Kolmogorov's 0–1 law).

(a) Let $A \in \mathcal{F}_\infty$ (where $\mathcal{F}_\infty = \sigma\big(\bigcup_n \mathcal{F}_n\big)$). Show that

$$\mathbb{P}(A \mid \mathcal{F}_n) \to \mathbf{1}_A \qquad \text{a.s.}$$

(b) As an application, deduce **Kolmogorov's 0–1 law**: if $X_1, X_2, \ldots$ are independent random variables, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, and A is a tail event (i.e., $A \in \bigcap_n \sigma(X_{n+1}, X_{n+2}, \ldots)$), then $\mathbb{P}(A) \in \{0, 1\}$.

Exercise 11.25 (Bayesian consistency for Bernoulli observations). Let θ be a random variable taking values in $(0, 1)$ with some prior distribution π, and conditional on θ, let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(θ) random variables. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Show that the posterior mean is consistent:

$$\mathbb{E}(\theta \mid \mathcal{F}_n) \to \theta \qquad \text{a.s.}$$

Exercise 11.26 (Lévy's extension of Borel–Cantelli). Let $\{A_n\}$ be a sequence of events with $A_n \in \mathcal{F}_n$ for each $n \geq 1$. Let $p_n = \mathbb{P}(A_n \mid \mathcal{F}_{n-1})$ (a random variable measurable with respect to $\mathcal{F}_{n-1}$).

(a) Show that $M_n = \sum_{k=1}^{n} (\mathbf{1}_{A_k} - p_k)$ is a martingale with respect to $\{\mathcal{F}_n\}$, with $\mathbb{E}(M_n^2) = \sum_{k=1}^{n} \mathbb{E}\big(p_k(1-p_k)\big) \leq \sum_{k=1}^{n} \mathbb{E}(p_k)$.

(b) Prove that on the event $\big\{\sum_{n=1}^{\infty} p_n < \infty\big\}$, only finitely many A_n occur, i.e., $\sum_n \mathbf{1}_{A_n} < \infty$ a.s.

(c)* Prove that on the event $\big\{\sum_{n=1}^{\infty} p_n = \infty\big\}$, infinitely many A_n occur, i.e., $\sum_n \mathbf{1}_{A_n} = \infty$ a.s.

(c) Conclude that almost surely, the events $\big\{\sum_n p_n = \infty\big\}$ and $\big\{\sum_n \mathbf{1}_{A_n} = \infty\big\}$ coincide. (This generalizes the two Borel–Cantelli lemmas: when the A_n are independent, $p_n = \mathbb{P}(A_n)$ is deterministic, recovering the classical statements.)

Exercise 11.27 (Galton–Watson: non-extinction and the limit W_∞). Let $\{Z_n\}_{n \geq 0}$ be a Galton–Watson branching process with $Z_0 = 1$ and offspring distribution ξ satisfying $\mathbb{E}\xi = \mu$, $\mathbb{V}\mathrm{ar}(\xi) = \sigma^2 < \infty$. Set $W_n = Z_n / \mu^n$.

(a) Verify that $\{W_n\}$ is a martingale with respect to $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$, and that $\mathbb{E}(W_n) = 1$ for all n.

(b) Compute $\mathbb{V}\mathrm{ar}(W_n)$. Show, when $\mu > 1$, that

$$\mathbb{V}\mathrm{ar}(W_n) = \frac{\sigma^2}{\mu(\mu - 1)} \big(1 - \mu^{-n}\big),$$

and conclude $\sup_n \mathbb{E}(W_n^2) < \infty$.

(c) Conclude (assuming $\mu > 1$) that $W_n \to W_\infty$ a.s. and in L^2, with $\mathbb{E}(W_\infty) = 1$.

(d) Show that on the event $\{\text{extinction}\} = \{Z_n = 0 \text{ eventually}\}$, $W_\infty = 0$. Conclude that $\mathbb{P}(\text{non-extinction}) > 0$ when $\mu > 1$.

Additional Exercises

Exercise 11.28 (The example revisited: $\{S_{N\wedge n}\}$ is not UI). Let S_n be the symmetric random walk with $S_0 = 1$, and let $N = \inf\{n : S_n = 0\}$. Set $X_n = S_{N\wedge n}$, the martingale of Example 11.8. Show directly (without invoking Vitali's theorem) that $\{X_n\}$ is not uniformly integrable.

Exercise 11.29 (Strongest optional stopping for UI martingales). Let $\{X_n\}$ be a uniformly integrable martingale with closure $X_\infty = \lim_n X_n$, and let $S \le T$ be stopping times (possibly taking the value ∞). Show that

$$\mathbb{E}(X_T \mid \mathcal{F}_S) = X_S \qquad \text{a.s.}$$

In particular, $\mathbb{E}(X_T) = \mathbb{E}(X_S) = \mathbb{E}(X_0)$.

Exercise 11.30 (L^2-bounded characterization). Let $\{M_n\}$ be a martingale with $M_0 \in L^2$, and let $D_k = M_k - M_{k-1}$ denote the martingale differences. Show that

$$\sup_{n\ge 0} \mathbb{E}(M_n^2) < \infty \iff \sum_{k=1}^{\infty} \mathbb{E}(D_k^2) < \infty.$$

In this case, identify $\sup_n \mathbb{E}(M_n^2)$ in terms of $\mathbb{E}(M_0^2)$ and $\sum_k \mathbb{E}(D_k^2)$.

Exercise 11.31 (Likelihood ratio martingale). Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $(\Omega, \mathcal{F})$, and let $\{\mathcal{F}_n\}$ be a filtration. Let $\mathbb{P}_n, \mathbb{Q}_n$ denote the restrictions of $\mathbb{P}, \mathbb{Q}$ to $\mathcal{F}_n$, and assume $\mathbb{Q}_n \ll \mathbb{P}_n$ for every n, with Radon–Nikodym derivative $L_n = d\mathbb{Q}_n/d\mathbb{P}_n$.

(a) Show that $\{L_n\}$ is a non-negative martingale under $\mathbb{P}$, with $\mathbb{E}_{\mathbb{P}}(L_n) = 1$.

(b) Conclude that $L_n \to L_\infty$ $\mathbb{P}$-a.s. for some $L_\infty \ge 0$ with $\mathbb{E}_{\mathbb{P}}(L_\infty) \le 1$.

(c) (Loss of mass.) Let $X_1, X_2, \dots$ be i.i.d. Bernoulli, and let $\mathbb{P}$ correspond to parameter $1/2$ while $\mathbb{Q}$ corresponds to parameter $2/3$. Compute L_n in terms of $S_n = X_1 + \cdots + X_n$, and show that $L_n \to 0$ $\mathbb{P}$-a.s. (so the inequality in (b) is strict).

Exercise 11.32 (Galton–Watson: extinction probability). Continuing the setting of Exercise 11.27, let $\varphi(s) = \mathbb{E}(s^\xi)$ for $s \in [0,1]$ be the offspring generating function, and let $q^* \in [0,1]$ be a fixed point of φ (i.e., $\varphi(q^*) = q^*$).

(a) Show that $M_n = (q^*)^{Z_n}$ is a non-negative bounded martingale, hence converges a.s. and in L^p for every $p \ge 1$.

(b) Compute the a.s. limit M_∞: show that on $\{\text{extinction}\}$, $M_\infty = 1$, and on $\{\text{non-extinction}\}$ (assuming $q^* < 1$), $M_\infty = 0$.

(c) Apply the martingale property to conclude that the extinction probability $q = \mathbb{P}(\text{extinction})$ is a fixed point of φ. (One can show further — not part of this exercise — that q equals the smallest fixed point in $[0,1]$.)

Exercise 11.33 (Convergence of random series). Let $X_1, X_2, \dots$ be independent random variables with $\mathbb{E}(X_k) = 0$ and $\sum_{k=1}^{\infty} \mathbb{V}\mathrm{ar}(X_k) < \infty$. Show that $\sum_{k=1}^{\infty} X_k$ converges almost surely.

Chapter 12

Backward Martingales

A **backward martingale** (or **reversed martingale**) is a martingale indexed by the negative integers. This simple change of direction leads to remarkably strong convergence results: every backward martingale is automatically uniformly integrable, and hence converges both a.s. and in L^1. As an application, we obtain a second proof of the Strong Law of Large Numbers—one that is arguably the most elegant proof available.

12.1 Definition and basic properties

Definition 12.1. A **backward filtration** is a decreasing sequence of σ-algebras $\mathcal{G}_0 \supset \mathcal{G}_1 \supset \mathcal{G}_2 \supset \cdots$. An integrable sequence $\{M_n\}_{n\geq 0}$ is a **backward martingale** with respect to $\{\mathcal{G}_n\}$ if M_n is $\mathcal{G}_n$-measurable for each n and

$$\mathbb{E}(M_n \mid \mathcal{G}_{n+1}) = M_{n+1} \quad \text{for all } n \geq 0.$$

Note the direction: $\mathcal{G}_{n+1} \subset \mathcal{G}_n$, so we are conditioning on **less** information as the index increases—each step, we forget a little more. I prefer to call this an "Alzheimer martingale," but the mathematical community insists on **backward**.

Equivalently, re-indexing by $-n$, the process M_{-n} is an ordinary martingale with respect to the filtration $\mathcal{G}_{-n}$.

Theorem 12.2 (Backward martingale convergence)**.** *Let $\{M_n\}$ be a backward martingale with respect to a backward filtration $\{\mathcal{G}_n\}$. Let $\mathcal{G}_\infty = \bigcap_{n=0}^{\infty} \mathcal{G}_n$. Then:*

(i) $\{M_n\}$ is uniformly integrable.

(ii) $M_n \to M_\infty = \mathbb{E}(M_0 \mid \mathcal{G}_\infty)$ a.s. and in L^1.

Proof. (i) By the backward martingale property, $\mathbb{E}(M_0 \mid \mathcal{G}_n) = M_n$ for every $n \geq 0$. (This follows from iterating the defining relation: $\mathbb{E}(M_0 \mid \mathcal{G}_1) = M_1$, $\mathbb{E}(M_1 \mid \mathcal{G}_2) = M_2$, and by the tower property $\mathbb{E}(M_0 \mid \mathcal{G}_2) = M_2$, etc.)

Therefore the family $\{M_n\}_{n\geq 0}$ is contained in $\{\mathbb{E}(M_0 \mid \mathcal{G}) : \mathcal{G} \subset \mathcal{F}\}$, which is uniformly integrable by Theorem 11.11.

(ii) From part (i) we have the explicit representation $M_n = \mathbb{E}(M_0 \mid \mathcal{G}_n)$ for every $n \geq 0$. We use this to obtain both convergence and identification of the limit.

A.s. convergence. Re-indexing $\widetilde{M}_k := M_{-k}$ for $k \in \{\ldots, -2, -1, 0\}$, we obtain a martingale on the index set $\{\ldots, -2, -1, 0\}$ with respect to the increasing filtration $\widetilde{\mathcal{F}}_k := \mathcal{G}_{-k}$. Since $\{M_n\}$ is UI, it is L^1-bounded, so the upcrossing inequality (which applies on any countable totally ordered index set) gives a.s. convergence of $\widetilde{M}_k$ as $k \to -\infty$, i.e., $M_n \to M_\infty$ a.s. for some integrable M_∞.

L^1 convergence. UI together with a.s. convergence gives L^1 convergence by Vitali's theorem.

Identification of the limit. For any $A \in \mathcal{G}_\infty = \bigcap_m \mathcal{G}_m$, we have $A \in \mathcal{G}_n$ for every n, so by part (i):

$$\mathbb{E}(M_n; A) = \mathbb{E}\big(\mathbb{E}(M_0 \mid \mathcal{G}_n); A\big) = \mathbb{E}(M_0; A).$$

Passing $n \to \infty$ using L^1 convergence: $\mathbb{E}(M_\infty; A) = \mathbb{E}(M_0; A)$.

It remains to verify that M_∞ is $\mathcal{G}_\infty$-measurable. For each fixed k, the tail $\{M_n\}_{n\geq k}$ consists of $\mathcal{G}_k$-measurable random variables (since $\mathcal{G}_n \subset \mathcal{G}_k$ for $n \geq k$), so the a.s. limit M_∞ is $\mathcal{G}_k$-measurable. This holds for every k, hence M_∞ is $\mathcal{G}_\infty$-measurable. Combined with the integration condition above, $M_\infty = \mathbb{E}(M_0 \mid \mathcal{G}_\infty)$ a.s. □

Remark 12.3. The key difference from forward martingales is that uniform integrability comes for free: backward martingales are **always** UI, with no additional moment assumptions. This is because every backward martingale is automatically a Doob martingale (of the form $\mathbb{E}(M_0 \mid \mathcal{G}_n)$).

12.2 Application: Second proof of the SLLN

We now give a proof of the Strong Law of Large Numbers that is completely different from the truncation-based proof in Chapter 6. The argument is short and illuminating: it reveals the SLLN as a consequence of the backward martingale convergence theorem.

Theorem 12.4 (SLLN via backward martingales)**.** *Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}(X_1) \quad \text{as } n \to \infty,$$

where $S_n = X_1 + \cdots + X_n$.

Proof. Without loss of generality, assume $\mathbb{E}(X_1) = 0$ (otherwise replace X_i by $X_i - \mathbb{E}(X_1)$).

Step 1: Construct a backward martingale. Define the backward filtration

$$\mathcal{G}_n = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots), \quad n \geq 1.$$

Note that $\mathcal{G}_1 \supset \mathcal{G}_2 \supset \cdots$ (knowing $S_n, S_{n+1}, \ldots$ gives more information than knowing $S_{n+1}, S_{n+2}, \ldots$).

We claim that $M_n = S_n/n$ is a backward martingale with respect to $\{\mathcal{G}_n\}$.

First, $M_n = S_n/n$ is $\mathcal{G}_n$-measurable since S_n is the first generator of $\mathcal{G}_n$.

Second, we must show $\mathbb{E}(M_n \mid \mathcal{G}_{n+1}) = M_{n+1}$, i.e., $\mathbb{E}(S_n/n \mid \mathcal{G}_{n+1}) = S_{n+1}/(n+1)$.

To compute $\mathbb{E}(S_n \mid \mathcal{G}_{n+1})$, we write $S_n = S_{n+1} - X_{n+1}$. Since S_{n+1} is $\mathcal{G}_{n+1}$-measurable:

$$\mathbb{E}(S_n \mid \mathcal{G}_{n+1}) = S_{n+1} - \mathbb{E}(X_{n+1} \mid \mathcal{G}_{n+1}).$$

Now, by the symmetry of the i.i.d. assumption, each X_i plays the same role in S_{n+1}. More precisely, for any $1 \le i \le n+1$, $S_{n+1} = S_{n+1}$ regardless of which summand we label X_i, so by the exchangeability of $(X_1, \ldots, X_{n+1})$:

$$\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \mathbb{E}(X_j \mid \mathcal{G}_{n+1}) \quad \text{for all } 1 \le i, j \le n+1.$$

Since $\sum_{i=1}^{n+1} \mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \mathbb{E}(S_{n+1} \mid \mathcal{G}_{n+1}) = S_{n+1}$, we conclude

$$\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \frac{S_{n+1}}{n+1} \quad \text{for each } i \le n+1.$$

Therefore $\mathbb{E}(S_n \mid \mathcal{G}_{n+1}) = S_{n+1} - S_{n+1}/(n+1) = nS_{n+1}/(n+1)$, and

$$\mathbb{E}(M_n \mid \mathcal{G}_{n+1}) = \frac{1}{n} \cdot \frac{nS_{n+1}}{n+1} = \frac{S_{n+1}}{n+1} = M_{n+1}.$$

Step 2: Identify the limit. By Theorem 12.2, $M_n \to M_\infty = \mathbb{E}(M_1 \mid \mathcal{G}_\infty)$ a.s. and in L^1, where

$$\mathcal{G}_\infty = \bigcap_{n=1}^{\infty} \mathcal{G}_n.$$

We claim that $\mathcal{G}_\infty$ is trivial. First note that $\mathcal{G}_n = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$, since $S_{n+k} = S_n + X_{n+1} + \cdots + X_{n+k}$ and conversely $X_{n+k} = S_{n+k} - S_{n+k-1}$. We show that $\mathcal{G}_\infty$ is contained in the exchangeable σ-algebra $\mathcal{E}$ from Chapter 4. Recall that $\mathcal{E}$ consists of all events invariant under finite permutations of $(X_1, X_2, \ldots)$.

Let π be any finite permutation, say one that moves only $X_1, \ldots, X_N$. For $n \ge N$, the partial sum $S_n = X_1 + \cdots + X_n$ is unchanged by π (permuting summands does not change the sum), and $X_{n+1}, X_{n+2}, \ldots$ are untouched. Therefore $\mathcal{G}_n = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$ is invariant under π for all $n \ge N$, and hence $\mathcal{G}_\infty = \bigcap_n \mathcal{G}_n$ is invariant under π. Since π was arbitrary, $\mathcal{G}_\infty \subset \mathcal{E}$.

By the Hewitt–Savage 0-1 law (Chapter 4), $\mathcal{E}$ is trivial under the i.i.d. assumption, so $\mathcal{G}_\infty$ is also trivial. Therefore $M_\infty = \mathbb{E}(M_1 \mid \mathcal{G}_\infty) = \mathbb{E}(M_1) = \mathbb{E}(X_1) = 0$ a.s.

We conclude that $S_n/n \to 0$ a.s. $\square$

Remark 12.5. Compared to the Kolmogorov–Etemadi proof in Chapter 6, this proof has several notable features:

- It requires **full** independence and identical distribution (not just pairwise independence), since Step 1 uses exchangeability of $(X_1, \dots, X_{n+1})$ and Step 2 uses the Hewitt–Savage 0-1 law.
- It uses no truncation and no variance bounds—only the machinery of backward martingales and the 0-1 law.
- The proof naturally reveals **why** the limit is $\mathbb{E}(X_1)$: it is because S_n/n is a conditional expectation of X_1, and the conditioning σ-algebra shrinks to something trivial.

Exercise 12.14 explores the L^2 convergence rate of the backward martingale S_n/n.

12.3 De Finetti's theorem

De Finetti's theorem is one of the most striking applications of backward martingales. It says that an infinite exchangeable sequence is, conditional on its tail behavior, an i.i.d. sequence. The result has deep philosophical significance in Bayesian statistics: it justifies the representation of subjective beliefs about exchangeable observables as mixtures of i.i.d. models.

Recall from Chapter 4 that a sequence $(X_n)_{n\geq 1}$ is **exchangeable** if

$$(X_1, \dots, X_k) \stackrel{d}{=} (X_{\pi(1)}, \dots, X_{\pi(k)})$$

for every k and every permutation π of $\{1, \dots, k\}$, and that $\mathcal{I}$ denotes the exchangeable σ-field of events invariant under all finite permutations.

Theorem 12.6 (Strong law for exchangeable sequences)**.** *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence with $\mathbb{E}|X_1| < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}(X_1 \mid \mathcal{I}) \quad \text{as } n \to \infty,$$

with convergence also in L^1.

Proof. The argument is identical to the i.i.d. case (Theorem 12.4) up to the identification of the limit.

Define $\mathcal{G}_n = \sigma(S_n, S_{n+1}, S_{n+2}, \dots)$, where $S_n = X_1 + \cdots + X_n$. The same exchangeability argument as in the i.i.d. proof shows that $M_n = S_n/n$ is a backward martingale with respect to $\{\mathcal{G}_n\}$: for any $1 \leq i, j \leq n+1$, exchangeability gives

$$\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \mathbb{E}(X_j \mid \mathcal{G}_{n+1}),$$

and summing yields $\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = S_{n+1}/(n+1)$.

By Theorem 12.2, $M_n \to Z$ a.s. and in L^1 for some integrable Z, and as in the i.i.d. proof, $\mathcal{G}_\infty \subset \mathcal{I}$.

Identification of the limit. We claim that $Z = \mathbb{E}(X_1 \mid \mathcal{I})$ a.s.

(a) Z is $\mathcal{I}$-measurable. As an a.s. limit of $\mathcal{G}_n$-measurable random variables (with $\mathcal{G}_n \supset \mathcal{G}_\infty$ for every n), the variable Z is $\mathcal{G}_\infty$-measurable. Since $\mathcal{G}_\infty \subset \mathcal{I}$, it is also $\mathcal{I}$-measurable.

(b) Integration condition. For any $A \in \mathcal{I}$ and any $n \geq 1$:

$$\mathbb{E}(M_n; A) = \frac{1}{n}\mathbb{E}(S_n; A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i; A).$$

Since A is invariant under finite permutations and $(X_1, \ldots, X_n)$ is exchangeable, $\mathbb{E}(X_i; A) = \mathbb{E}(X_1; A)$ for every i.[1] Therefore

$$\mathbb{E}(M_n; A) = \mathbb{E}(X_1; A).$$

Passing $n \to \infty$ using L^1 convergence of M_n:

$$\mathbb{E}(Z; A) = \mathbb{E}(X_1; A) \quad \text{for every } A \in \mathcal{I}.$$

Combining (a) and (b), $Z = \mathbb{E}(X_1 \mid \mathcal{I})$ a.s. □ □

Remark 12.7. When (X_n) is i.i.d., the Hewitt–Savage 0-1 law makes $\mathcal{I}$ trivial, and the limit collapses to the constant $\mathbb{E}(X_1)$. Theorem 12.4 is therefore the special case of Theorem 12.6 obtained by reducing $\mathbb{E}(X_1 \mid \mathcal{I})$ to $\mathbb{E}(X_1)$.

Theorem 12.8 (De Finetti, $\{0,1\}$-valued case)**.** *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence with $X_n \in \{0,1\}$. Then there exists a $[0,1]$-valued random variable Θ, measurable with respect to $\mathcal{I}$, such that*

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k \mid \mathcal{I}) = \Theta^s(1-\Theta)^{k-s} \quad \text{a.s.}$$

for every k and every $(x_1, \ldots, x_k) \in \{0,1\}^k$ with $s = x_1 + \cdots + x_k$. Equivalently, conditional on $\mathcal{I}$, the variables (X_n) are i.i.d. Bernoulli(Θ).

Proof. Define

$$\Theta := \mathbb{E}(X_1 \mid \mathcal{I}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{a.s.}$$

by Theorem 12.6. Since each $X_i \in \{0,1\}$, the limit Θ takes values in $[0,1]$ and is $\mathcal{I}$-measurable.

Fix $(x_1, \ldots, x_k) \in \{0,1\}^k$ with $s = \sum x_i$. Define the U-statistic

$$V_N := \frac{1}{(N)_k} \sum_{(i_1,\ldots,i_k)}^{*} \mathbf{1}_{\{X_{i_1}=x_1,\ldots,X_{i_k}=x_k\}}, \qquad (N)_k := N(N-1)\cdots(N-k+1),$$

where the starred sum runs over ordered k-tuples of distinct indices in $\{1, \ldots, N\}$.

Step 1 (combinatorial identity). By exchangeability, the indicator $\mathbf{1}_{\{X_{i_1}=x_1,\ldots,X_{i_k}=x_k\}}$ depends only on the values $X_{i_1}, \ldots, X_{i_k}$ and which of them equal 1. Counting the number of ordered k-tuples $(i_1, \ldots, i_k)$ for which exactly the s positions corresponding to the 1's in the pattern are filled by indices with $X = 1$ (and the remaining $k - s$ positions by indices with $X = 0$):

$$\sum_{(i_1,\ldots,i_k)}^{*} \mathbf{1}_{\{X_{i_1}=x_1,\ldots,X_{i_k}=x_k\}} = (T_N)_s \cdot (N - T_N)_{k-s},$$

[1] To verify: let π be the transposition of indices 1 and i. Then $(X_1, X_2, \ldots) \overset{d}{=} (X_1^\pi, X_2^\pi, \ldots)$ by exchangeability, and $\mathbf{1}_A$ is invariant under π, so $\mathbb{E}(X_1\mathbf{1}_A) = \mathbb{E}(X_1^\pi \mathbf{1}_A) = \mathbb{E}(X_i \mathbf{1}_A)$.

where $T_N = \sum_{i=1}^N X_i$ counts the 1's in the first N samples. Therefore

$$V_N = \frac{(T_N)_s \, (N - T_N)_{k-s}}{(N)_k}.$$

Step 2 (asymptotic limit). By Theorem 12.6, $T_N/N \to \Theta$ a.s. For falling factorials, $(M)_j = M^j(1 + O(1/M))$ as $M \to \infty$, so

$$V_N \to \Theta^s(1-\Theta)^{k-s} \quad \text{a.s.}$$

Step 3 (SLLN for U-statistics). Independently, V_N is the empirical average of the symmetric kernel $f(y_1, \ldots, y_k) = \mathbf{1}_{\{y_1 = x_1, \ldots, y_k = x_k\}}$ evaluated on all distinct k-tuples. The strong law of large numbers for U-statistics on exchangeable sequences (see, e.g., Aldous [Ald85], Section 3, or Kallenberg [Kal05], Theorem 1.18) gives

$$V_N \to \mathbb{E}\big(f(X_1, \ldots, X_k) \mid \mathcal{I}\big) = \mathbb{P}(X_1 = x_1, \ldots, X_k = x_k \mid \mathcal{I}) \quad \text{a.s.}$$

Conclusion. The two limits in Steps 2 and 3 must agree, giving the claimed identity. □ □

Remark 12.9. The general (non-Bernoulli) form of de Finetti's theorem states that any exchangeable sequence is, conditional on $\mathcal{I}$, an i.i.d. sequence with some random distribution ν. The proof uses Theorem 12.6 together with a regular conditional probability argument; see Durrett, Section 5.6.

Remark 12.10 (Bayesian interpretation). De Finetti's theorem provides the foundational justification for Bayesian inference about exchangeable observations. If a statistician treats observations (X_n) as exchangeable (a weaker and more defensible assumption than i.i.d.), the theorem shows that this is mathematically equivalent to specifying a prior distribution over an unknown "parameter" Θ together with a model in which the X_n are conditionally i.i.d. given Θ. The parameter is not assumed; it emerges from the exchangeability structure.

References for this section.

- D. Aldous, **Exchangeability and Related Topics**, in: École d'Été de Probabilités de Saint-Flour XIII—1983, Lecture Notes in Math. **1117**, Springer, 1985, pp. 1–198. (The standard reference; Section 3 covers de Finetti and U-statistic SLLNs.)

- O. Kallenberg, **Probabilistic Symmetries and Invariance Principles**, Springer, 2005. (Chapter 1 contains a self-contained treatment of exchangeable sequences and de Finetti's theorem.)

- R. Durrett, **Probability: Theory and Examples**, 5th ed., Cambridge University Press, 2019, §5.6. (Concise treatment within a standard graduate textbook.)

12.4 Exercises

Homework

Exercise 12.11 (Lévy's downward theorem). Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}_0 \supset \mathcal{G}_1 \supset \mathcal{G}_2 \supset \cdots$ be a decreasing sequence of σ-algebras with $\mathcal{G}_\infty = \bigcap_{n=0}^{\infty} \mathcal{G}_n$. Show that

$$\mathbb{E}(X \mid \mathcal{G}_n) \to \mathbb{E}(X \mid \mathcal{G}_\infty) \qquad \text{a.s. and in } L^1.$$

Exercise 12.12 (SLLN strengthened to L^1 convergence). Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}|X_1| < \infty$. Show that

$$\frac{S_n}{n} \to \mathbb{E}(X_1) \qquad \text{in } L^1$$

(in addition to the a.s. convergence of Theorem 12.4).

Exercise 12.13 (U-statistics). Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in some measurable space $(\mathcal{X}, \mathcal{B})$, and let $h : \mathcal{X}^k \to \mathbb{R}$ be a symmetric measurable function (i.e., h is invariant under permutation of its arguments) with $\mathbb{E}|h(X_1, \ldots, X_k)| < \infty$. For $n \geq k$, define the **U-statistic**

$$U_n = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} h(X_{i_1}, \ldots, X_{i_k}).$$

Define the **empirical measure** $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, the random probability measure on $\mathcal{X}$ that places mass $1/n$ at each observed value. Define the backward filtration

$$\mathcal{T}_n = \sigma\big(\hat{\mu}_n, X_{n+1}, X_{n+2}, \ldots\big), \qquad n \geq k,$$

the smallest σ-algebra making the random measure $\hat{\mu}_n$ and each of the variables $X_{n+1}, X_{n+2}, \ldots$ measurable.

Remark on the structure of $\mathcal{T}_n$. The σ-algebra $\sigma(\hat{\mu}_n)$ generated by the random measure consists of events of the form $\{\hat{\mu}_n \in \Gamma\}$ for Γ a measurable set in the space of probability measures on $\mathcal{X}$; concretely, it is generated by the events $\{\hat{\mu}_n(B) \leq c\}$ as B ranges over Borel subsets of $\mathcal{X}$ and c over $[0, 1]$. Equivalently, $\sigma(\hat{\mu}_n)$ is the σ-algebra of **permutation-symmetric** events on $(X_1, \ldots, X_n)$ — those invariant under permutation of the first n indices, since $\hat{\mu}_n$ depends on $(X_1, \ldots, X_n)$ only through the multiset $\{X_1, \ldots, X_n\}$. Equivalently again, $\sigma(\hat{\mu}_n)$ is generated by all symmetric measurable functions of $X_1, \ldots, X_n$ (e.g., S_n, $\sum_i X_i^2$, or the order statistics in the case $\mathcal{X} = \mathbb{R}$). Combining with the tail, $\mathcal{T}_n$ contains the symmetric information about $X_1, \ldots, X_n$ together with the full information about $X_{n+1}, X_{n+2}, \ldots$.

(a) Show that $\mathcal{T}_n \supset \mathcal{T}_{n+1}$, and that U_n is $\mathcal{T}_n$-measurable.

(b) Show that $\mathbb{E}(U_n \mid \mathcal{T}_{n+1}) = U_{n+1}$, so $\{U_n\}_{n \geq k}$ is a backward martingale with respect to $\{\mathcal{T}_n\}_{n \geq k}$.

(c) Conclude that

$$U_n \to \mathbb{E}\big(h(X_1, \ldots, X_k)\big) \qquad \text{a.s. and in } L^1.$$

Additional Exercises

Exercise 12.14 (SLLN in L^p). Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(X_1) = \mu$ and $\mathbb{E}|X_1|^p < \infty$ for some $p \geq 1$.

(a) For $p = 2$, compute $\mathbb{E}\big((S_n/n - \mu)^2\big)$ explicitly and conclude $S_n/n \to \mu$ in L^2 at rate $1/\sqrt{n}$.

(b) For general $p \geq 1$, show that $S_n/n \to \mu$ in L^p.

Exercise 12.15 ($\mathcal{G}_\infty$ and the exchangeable σ-algebra). In the proof of Theorem 12.4, the chapter argues informally that $\mathcal{G}_\infty \subset \mathcal{E}$, where $\mathcal{E}$ is the exchangeable σ-algebra of $(X_1, X_2, \dots)$.

Make this argument rigorous. Specifically: let π be a finite permutation of the positive integers, and let $T_\pi : \Omega \to \Omega$ be the corresponding measurable map (identifying Ω with $\mathbb{R}^{\mathbb{N}}$ via $X_1, X_2, \dots$). Show that for every $A \in \mathcal{G}_\infty$, $T_\pi^{-1}(A) = A$ (up to a null set).

Exercise 12.16 (Sample variance via U-statistics). Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(X_1^2) < \infty$, and let $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$. Define the (unbiased) sample variance

$$V_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \qquad \bar{X}_n = \frac{S_n}{n}.$$

(a) Show that V_n can be written as a U-statistic of degree 2 with kernel $h(x, y) = (x - y)^2/2$:

$$V_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \frac{(X_i - X_j)^2}{2}.$$

(b) Conclude from Exercise 12.13 that $V_n \to \sigma^2$ a.s. and in L^1.

Exercise 12.17 (The exchangeability lemma in the SLLN proof). In the proof of Theorem 12.4, the chapter invokes "by exchangeability" to deduce that

$$\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \frac{S_{n+1}}{n+1} \qquad \text{a.s., for each } 1 \leq i \leq n+1,$$

where $\mathcal{G}_{n+1} = \sigma(S_{n+1}, X_{n+2}, X_{n+3}, \dots)$. Fill in the rigorous argument:

(a) Show that for any $i, j \in \{1, 2, \dots, n+1\}$, $\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = \mathbb{E}(X_j \mid \mathcal{G}_{n+1})$ a.s.

(b) Conclude $\mathbb{E}(X_i \mid \mathcal{G}_{n+1}) = S_{n+1}/(n+1)$ a.s.

Chapter 13

Where to Go from Here

This book covers what is sometimes called the "classical core" of measure-theoretic probability: the construction of measures, integration and convergence theorems, conditional expectation, independence, the laws of large numbers, the central limit theorem, and the basic theory of discrete-time martingales. These topics form the language in which most of modern probability is spoken. But they are very much a beginning, not an end. The purpose of this short chapter is to sketch a few of the directions in which the subject continues, with pointers to standard references. The selection is personal and far from exhaustive; it reflects the directions a student leaving this course is most likely to encounter, whether in research or in applications.

The directions below are roughly grouped by how they relate to the material of this course: some sharpen results we have already proved, some relax our assumptions, some lift the theory into continuous time or into high dimensions, and some build bridges to neighbouring fields.

13.1 Sharper than the LLN: concentration and large deviations

Chapter 6 tells us that $S_n/n \to \mathbb{E}(X_1)$ almost surely. This is a qualitative statement: it says nothing about **how close** S_n/n is to its mean for a given n, nor how unlikely large deviations are. Two enormous bodies of theory address these questions.

Concentration inequalities provide non-asymptotic bounds of the form

$$\mathbb{P}\big(|f(X_1,\dots,X_n) - \mathbb{E}f| \geq t\big) \leq 2e^{-ct^2/v},$$

where f is some function of independent (or weakly dependent) random variables and v is a variance proxy. The simplest example is Hoeffding's inequality for sums of bounded independent variables; from there the theory branches into Bernstein-type inequalities (which exploit variance), McDiarmid's bounded-differences inequality (for general functions of independent inputs), the Efron–Stein inequality, Azuma's inequality for martingales (which we may meet

briefly in this course), Talagrand's convex distance inequality, and the entropy method based on logarithmic Sobolev inequalities. The unifying message is striking: in many natural problems, "most" of the randomness cancels out, and the typical fluctuation of a function of n independent inputs is controlled by the maximal effect of any single input.

This circle of ideas is one of the most actively used parts of probability today, with central roles in statistics, machine learning, information theory, theoretical computer science, and random matrix theory. The standard references are Boucheron, Lugosi, and Massart, **Concentration Inequalities** (Oxford, 2013), and Vershynin, **High-Dimensional Probability** (Cambridge, 2018).

Large deviation theory addresses the same question on a logarithmic scale. Cramér's theorem states that for i.i.d. variables with finite moment generating function,

$$\frac{1}{n}\log \mathbb{P}\big(S_n/n \geq a\big) \to -I(a),$$

where the rate function I is the Legendre transform of the cumulant generating function. Sanov's theorem extends this to empirical measures, and Varadhan's lemma turns the rate function into a tool for evaluating exponential integrals. The general framework—large deviation principles—applies to a wide class of stochastic systems, from random walks and queues to interacting particle systems and stochastic PDEs.

Standard references: Dembo and Zeitouni, **Large Deviations Techniques and Applications**, 2nd ed. (Springer, 1998); den Hollander, **Large Deviations** (AMS, 2000).

13.2 Sharper than the CLT: rates of convergence

Chapter 8 proves that $(S_n - n\mu)/(\sigma\sqrt{n}) \xrightarrow{d} \mathcal{N}(0,1)$ but says nothing about **how fast** the convergence is. The Berry–Esseen theorem gives the first quantitative answer: under a finite third moment, the Kolmogorov distance between the distribution of the standardised sum and the normal distribution is at most $C\mathbb{E}|X_1|^3/(\sigma^3\sqrt{n})$. Edgeworth expansions go further, providing polynomial corrections in powers of $1/\sqrt{n}$.

A more recent and remarkably flexible approach is **Stein's method**, introduced in 1972. The starting point is the observation that a random variable W is standard normal if and only if $\mathbb{E}\big[f'(W) - Wf(W)\big] = 0$ for all sufficiently smooth f. To bound the distance between the distribution of some W and the normal distribution, one solves the differential equation $f'(w) - wf(w) = h(w) - \mathbb{E}h(Z)$ and bounds the left-hand side. The method requires no characteristic functions and is particularly powerful when the summands are dependent: it has produced sharp central limit theorems for U-statistics, graph colourings, the number of fixed points of a random permutation, and many other settings where Fourier methods are awkward or unavailable. Analogous machinery has been developed for the Poisson approximation (the Chen–Stein method) and for many other target distributions.

Standard references: Chen, Goldstein, and Shao, **Normal Approximation by Stein's Method** (Springer, 2011); Ross, **Fundamentals of Stein's Method** (**Probab. Surv.**, 2011).

13.3 Beyond independence: ergodic theory and stationary processes

The strong law of large numbers is fundamentally a statement about i.i.d. sequences. Many natural sequences, however, are not independent: outputs of a Markov chain, observations of a dynamical system, time series in economics or climate science. The right generalization replaces independence by **stationarity**.

Birkhoff's ergodic theorem states that for a measure-preserving transformation T on a probability space and an integrable function f,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k\omega) \to \mathbb{E}(f \mid \mathcal{I}) \quad \text{a.s.},$$

where $\mathcal{I}$ is the σ-algebra of T-invariant events. When T is **ergodic**, the limit collapses to $\mathbb{E}f$, and we recover an SLLN-style statement for stationary sequences. The Kingman subadditive ergodic theorem extends this to subadditive functionals and is the foundation for many results in random walks, percolation, and products of random matrices (the Furstenberg–Kesten theorem on Lyapunov exponents).

This material is the gateway to ergodic theory proper, which fuses probability with dynamical systems and has deep connections with number theory and combinatorics (e.g. Furstenberg's proof of Szemerédi's theorem). At the more applied end, it underlies the theory of stationary time series and the consistency of many statistical estimators in time-series settings.

Standard references: Walters, **An Introduction to Ergodic Theory** (Springer, 1982); Einsiedler and Ward, **Ergodic Theory with a View towards Number Theory** (Springer, 2011); Krengel, **Ergodic Theorems** (de Gruyter, 1985).

A closely related circle of ideas concerns **exchangeable** sequences—sequences whose joint distribution is invariant under permutations. De Finetti's theorem (Chapter 12) shows that an infinite exchangeable sequence is a mixture of i.i.d. sequences; this is the mathematical foundation of Bayesian statistics. The far-reaching extension of these ideas to arrays and graphs (the Aldous–Hoover theorem, graphon limits) connects to modern combinatorics; see Aldous, **Exchangeability and Related Topics** (Saint-Flour XIII, Springer, 1985), and Lovász, **Large Networks and Graph Limits** (AMS, 2012).

13.4 Continuous time: Brownian motion and stochastic calculus

Almost everything in this book takes place in discrete time. The most important next step for many students is the construction and analysis of **stochastic**

processes in continuous time.

The central object is **Brownian motion**: a process $(B_t)_{t\geq 0}$ with continuous paths, independent Gaussian increments, and $B_0 = 0$. It arises as a scaling limit of random walks (Donsker's invariance principle), as the unique continuous Gaussian process with covariance $\mathbb{E}(B_s B_t) = s \wedge t$, and as the canonical model of diffusion. Its sample paths are nowhere differentiable yet have well-defined quadratic variation, which forces a new differential calculus. **Itô's calculus** provides the framework: the Itô integral $\int_0^t H_s\, dB_s$ is built using L^2 approximation against the quadratic variation $\langle B\rangle_t = t$, and Itô's formula plays the role of the chain rule. From here one constructs solutions to **stochastic differential equations** of the form $dX_t = b(X_t)\, dt + \sigma(X_t)\, dB_t$, which model everything from particle diffusion to interest rates.

The theory rests heavily on continuous-time martingales, whose machinery (the Doob–Meyer decomposition, optional stopping, L^p inequalities) parallels the discrete theory of Chapters 9–12. Connections to PDE are deep: solutions of parabolic and elliptic equations admit probabilistic representations via Brownian motion (the Feynman–Kac formula), and conversely the laws of diffusions are characterized by PDE.

Standard references: Karatzas and Shreve, **Brownian Motion and Stochastic Calculus**, 2nd ed. (Springer, 1991); Le Gall, **Brownian Motion, Martingales, and Stochastic Calculus** (Springer, 2016); Revuz and Yor, **Continuous Martingales and Brownian Motion**, 3rd ed. (Springer, 1999).

13.5 High dimensions and random matrices

Many problems in modern statistics, signal processing, and machine learning involve random vectors or matrices whose dimension is itself large—comparable to, or larger than, the sample size. Classical fixed-dimension intuitions can fail dramatically in this regime, and a separate theory has emerged.

The theory of **high-dimensional probability** studies concentration of norms of random vectors, suprema of Gaussian and sub-Gaussian processes, covering numbers, and uniform laws of large numbers. Tools include Gaussian concentration (Borell's inequality), generic chaining, and the matrix Bernstein inequalities. Many results stand as natural generalizations of one-dimensional concentration with explicit dimension dependence.

The theory of **random matrices** studies the spectra of $n \times n$ matrices with random entries as $n \to \infty$. The Wigner semicircle law and the Marchenko–Pastur law play roles analogous to the central limit theorem: they describe the limiting spectral distribution of broad classes of random matrices. Beyond the bulk of the spectrum, the largest eigenvalue obeys the Tracy–Widom distribution, with universal behaviour across many models. Random matrix theory has remarkable connections with integrable systems, the Riemann zeta function, free probability, and combinatorics, and is now a routine tool in statistics (PCA, high-dimensional inference) and physics.

Standard references: Vershynin, **High-Dimensional Probability** (Cambridge, 2018); Wainwright, **High-Dimensional Statistics** (Cambridge, 2019); Anderson, Guionnet, and Zeitouni, **An Introduction to Random Matri-**

ces (Cambridge, 2010); Tao, **Topics in Random Matrix Theory** (AMS, 2012).

13.6 Probability in service of statistics and computation

Probability is, of course, the foundation of statistics. Two areas in particular are worth singling out.

Empirical processes and statistical learning theory ask uniform questions: how well does the empirical measure $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ approximate the true distribution, **uniformly** over a class of test functions or events? The basic Glivenko–Cantelli theorem (uniform convergence over half-lines) gives a foretaste; the general theory, built around Vapnik–Chervonenkis dimension, Rademacher complexities, and metric entropy, is the probabilistic backbone of statistical learning theory and provides risk bounds for classification, regression, and density estimation. Standard references: van der Vaart and Wellner, **Weak Convergence and Empirical Processes** (Springer, 1996); Giné and Nickl, **Mathematical Foundations of Infinite-Dimensional Statistical Models** (Cambridge, 2016).

Markov chain Monte Carlo provides the computational counterpart. To sample from a complicated probability distribution π on a large state space, one constructs a Markov chain whose stationary distribution is π and runs it for a sufficient number of steps. The theoretical questions are: does the chain converge, and how fast? **Mixing time** theory answers the second question with tools ranging from coupling and conductance to spectral gaps and log-Sobolev inequalities. The standard reference is Levin and Peres, **Markov Chains and Mixing Times**, 2nd ed. (AMS, 2017).

13.7 A few further directions worth a look

The following are not, strictly speaking, "foundational" continuations of this course, but each is a thriving and beautiful area in its own right, accessible to a student with the background developed here.

Optimal transport and Wasserstein distances have become a standard language for comparing probability measures. They give a geometry on the space of measures, with deep connections to PDE, geometry, and statistics; see Villani, **Topics in Optimal Transportation** (AMS, 2003), and Santambrogio, **Optimal Transport for Applied Mathematicians** (Birkhäuser, 2015).

Free probability, developed by Voiculescu, provides a non-commutative analogue of classical probability in which independence is replaced by "freeness". It is the right framework for understanding limits of large random matrices, and has unexpected applications in operator algebras and combinatorics; see Mingo and Speicher, **Free Probability and Random Matrices** (Springer, 2017).

Probabilistic combinatorics applies probabilistic reasoning to deterministic combinatorial questions: random graphs (Erdős–Rényi, configuration models, preferential attachment), the probabilistic method, threshold phenomena, percolation. See Alon and Spencer, **The Probabilistic Method**, 4th ed. (Wiley, 2016), and van der Hofstad, **Random Graphs and Complex Networks**, vols. 1–2 (Cambridge, 2017–2024).

Interacting particle systems and stochastic spatial models include percolation, the contact process, the voter model, exclusion processes, and many others; see Liggett, **Interacting Particle Systems** (Springer, 1985), and Grimmett, **Probability on Graphs**, 2nd ed. (Cambridge, 2018).

Stochastic analysis on manifolds, rough paths, and regularity structures extend stochastic calculus to settings where classical Itô theory breaks down (very irregular driving signals, stochastic PDEs); see Friz and Hairer, **A Course on Rough Paths**, 2nd ed. (Springer, 2020).

Closing remark

A graduate course in probability is necessarily selective: the field is too large for any one book or one semester to cover. What I hope this course has given you is fluency in the basic vocabulary—measures, integrals, conditional expectations, characteristic functions, martingales—and a sense of how these objects fit together. With that vocabulary in hand, the directions sketched above are open to you. Pick one that matches your taste or the questions you want to answer in your own research, and dive in.

The list above is in no way complete. Probability is a living subject; its boundaries with analysis, combinatorics, geometry, statistics, and computer science are constantly shifting, and new connections are made every year. The best advice is the same as for any area of mathematics: read papers, talk to people, and work through examples until the abstractions become tangible.

Appendix A

Measure Theory

A.1 Carathéodory theorem

Since this theorem belongs mainly to the measure theory we will not give the full proof. For the full proof see Durrett's or Shiryaev's books. However we indicate some ideas.

The uniqueness of the measure extension was proved in Thm. 1.30. We will give only a sketch of the existence proof.

The existence part of the Carathéodory theorem: Sketch of the proof.

Step 1. Define a set function on **all** subsets of Ω, which is called the **outer measure**:

$$\mu^*(A) = \inf_{A \subset \bigcup A_j} \sum_j \mu(A_j), \tag{A.1}$$

where the infinum is taken over all countable collections A_j of sets from $\mathcal{A}$ that cover A. Without loss of generality we can assume that A_j are disjoint. (Replace A_j by $A_j \setminus \bigcup_{i=1}^{i=j-1} A_i$)).

Step 2. Show that μ^* has the following properties:

1. The set function μ^* is countably subadditive, that is,
$$\mu^*\left(\bigcup_j A_j\right) \leq \sum_j \mu^*(A_j).$$
2. For $A \in \mathcal{A}$, $\mu^*(A) \leq \mu(A)$. (Trivial)
3. For $A \in \mathcal{A}$, $\mu^*(A) \geq \mu(A)$. (Here we need to use the countable additivity of μ on (A).)

Step 3. Define a set E to be measurable if

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

holds for all sets A, and establish the following properties for the class $\mathcal{M}$ of measurable sets: The class of measurable sets $\mathcal{M}$ is a σ - algebra and μ^* is countably additive measure on it.

Step 4. Finally, show that $\mathcal{A} \subset \mathcal{M}$. This implies that $\sigma(\mathcal{A}) \subset \mathcal{M}$ and μ^* is an extension of μ from $\mathcal{A}$ to $\sigma(\mathcal{A})$.

□

A.2 Monotone Class Theorem (sets)

Theorem A.1 (Monotone Class Theorem (sets))**.** *Let $\mathcal{A}$ be an algebra on Ω, and let $\mathcal{M}$ be a monotone class with $\mathcal{A} \subseteq \mathcal{M}$. Then*

$$\sigma(\mathcal{A}) \subseteq \mathcal{M}.$$

Equivalently, the monotone class generated by an algebra $\mathcal{A}$ coincides with $\sigma(\mathcal{A})$.

Proof. Let

$$\mathcal{M}^* := \bigcap \{\mathcal{N} : \ \mathcal{N} \text{ is a monotone class and } \mathcal{A} \subseteq \mathcal{N}\}$$

be the **minimal monotone class** generated by $\mathcal{A}$. By definition $\mathcal{M}^*$ is a monotone class with $\mathcal{A} \subseteq \mathcal{M}^*$, and for any monotone class $\mathcal{M}$ with $\mathcal{A} \subseteq \mathcal{M}$ we have $\mathcal{M}^* \subseteq \mathcal{M}$. We will prove that $\mathcal{M}^* = \sigma(\mathcal{A})$; then the desired inclusion $\sigma(\mathcal{A}) \subseteq \mathcal{M}$ follows immediately from $\mathcal{M}^* \subseteq \mathcal{M}$.

Step 1 (intersections with generators). Fix $B \in \mathcal{A}$ and set

$$\mathcal{M}_B^* := \{E \subseteq \Omega : \ E \cap B \in \mathcal{M}^*\}.$$

Then $\mathcal{M}_B^*$ is a monotone class: if $E_n \uparrow E$ (or $E_n \downarrow E$), then $E_n \cap B \uparrow E \cap B$ (resp. $\downarrow$), hence $E \in \mathcal{M}_B^*$ since $\mathcal{M}^*$ is monotone. Because $\mathcal{A}$ is a π-system, $A \cap B \in \mathcal{A} \subseteq \mathcal{M}^*$ for every $A \in \mathcal{A}$, so $\mathcal{A} \subseteq \mathcal{M}_B^*$. By the **minimality of** $\mathcal{M}^*$, we conclude

$$\mathcal{M}^* \subseteq \mathcal{M}_B^* \qquad \text{for each } B \in \mathcal{A}.$$

Equivalently,

$$E \in \mathcal{M}^*, \ B \in \mathcal{A} \implies E \cap B \in \mathcal{M}^*. \tag{A.2}$$

Step 2 (closure under finite intersections). Fix $E \in \mathcal{M}^*$ and define

$$\mathcal{H}_E := \{B \subseteq \Omega : \ E \cap B \in \mathcal{M}^*\}.$$

Exactly as above, $\mathcal{H}_E$ is a monotone class. Moreover, (A.2) gives $A \in \mathcal{A} \Rightarrow E \cap A \in \mathcal{M}^*$, hence $\mathcal{A} \subseteq \mathcal{H}_E$. By minimality of $\mathcal{M}^*$ we obtain $\mathcal{M}^* \subseteq \mathcal{H}_E$. Therefore for all $E, B \in \mathcal{M}^*$ we have $E \cap B \in \mathcal{M}^*$, i.e. $\mathcal{M}^*$ is closed under finite intersections.

Step 3 (complements) Let $\mathcal{K} := \{E \in \mathcal{M}^* : E^c \in \mathcal{M}^*\}$. Class $\mathcal{K}$ is a monotone class: if $E_n \uparrow E$ with each $\mathbb{E}_n^c \in \mathcal{M}^*$, then $E^c = \cup_n E_n^c \in \mathcal{M}^*$ by decreasing-closure. Similarly for $E_n \downarrow E$.

Since $\mathcal{A}$ is an algebra, $A^c \in \mathcal{A} \in \mathcal{M}^*$ for all $A \in \mathcal{A}$; hence $\mathcal{A} \subseteq \mathcal{K}$. By minimality $\mathcal{M}^* \in \mathcal{K}$, so $\mathcal{M}^*$ is closed under complements.

Step 4 (σ-algebra) With complements and finite intersections, $\mathcal{M}^*$ is an algebra. It is also closed under increasing unions since it is a monotone class.

Now consider an arbitrary sequence of $E_n \in \mathcal{M}^*$. Set $U_n = \cup_{k=1}^n E_k$. Then $U_n \in \mathcal{M}^*$ by algebra property, and then $\cup_n E_n = \lim_n U_n \in \mathcal{M}^*$ by monotone closure property of monotone class. Thus $\mathcal{M}^*$ is a σ-algebra containing algebra $\mathcal{A}$.

Consequently $\sigma(\mathcal{A}) \subseteq \mathcal{M}^* \subseteq \mathcal{M}$.

Step 5 (Equality) Every σ-algebra is a monotone class and $\sigma(\mathcal{A})$ contains $\mathcal{A}$, so by minimality $\mathcal{M}^* \subseteq \sigma(\mathcal{A})$. Therefore $\mathcal{M}^* = \sigma(\mathcal{A})$. □

A.3 Monotone Class Theorem (functions)

Theorem A.2 (Monotone Class Theorem (functions))**.** *Let $\mathcal{H}$ be a vector space of bounded real-valued functions on Ω such that:*

1. $\mathbf{1}_\Omega \in \mathcal{H}$*;*
2. $\mathbf{1}_E \in \mathcal{H}$ *for all E in a π-system $\mathcal{A}$;*
3. *if $0 \le f_n \uparrow f$ pointwise with f bounded and each $f_n \in \mathcal{H}$, then $f \in \mathcal{H}$.*

Then $\mathcal{H}$ contains every bounded $\sigma(\mathcal{A})$-measurable function.

Proof. **Step 1 (Reduce to indicators).** It suffices to show $\mathbf{1}_E \in \mathcal{H}$ for all $E \in \sigma(\mathcal{A})$. Indeed, once we have all indicators, $\mathcal{H}$ contains all bounded simple $\sigma(\mathcal{A})$-measurable functions (by the vector space property), and then every bounded measurable $f \ge 0$ is the pointwise increasing limit of bounded simple functions, so $f \in \mathcal{H}$ by (3). For general bounded measurable f, write $f = f^+ - f^-$.

Step 2 (Build a monotone class of sets). Define

$$\mathcal{M} := \{E \subseteq \Omega : \ \mathbf{1}_E \in \mathcal{H}\}.$$

We verify that $\mathcal{M}$ is a monotone class containing $\mathcal{A}$:

- $\mathcal{A} \subseteq \mathcal{M}$ by assumption (2), and $\Omega \in \mathcal{M}$ by (1).
- **Complements:** if $E \in \mathcal{M}$, then $\mathbf{1}_{E^c} = \mathbf{1}_\Omega - \mathbf{1}_E \in \mathcal{H}$ (since $\mathcal{H}$ is a vector space containing $\mathbf{1}_\Omega$), so $E^c \in \mathcal{M}$.
- **Increasing unions:** if $E_n \uparrow E$ with each $E_n \in \mathcal{M}$, then $0 \le \mathbf{1}_{E_n} \uparrow \mathbf{1}_E$ pointwise and $\mathbf{1}_E$ is bounded, so $\mathbf{1}_E \in \mathcal{H}$ by (3), hence $E \in \mathcal{M}$.
- **Decreasing intersections:** if $E_n \downarrow E$, then $E_n^c \uparrow E^c$; by the two properties above, $E^c \in \mathcal{M}$, hence $E \in \mathcal{M}$.

Step 3 (Finite intersections). If $E, F \in \mathcal{M}$, then $\mathbf{1}_{E\cap F} = \mathbf{1}_E \cdot \mathbf{1}_F$. We need this product to lie in $\mathcal{H}$. Note that $\mathbf{1}_E \cdot \mathbf{1}_F = \mathbf{1}_E + \mathbf{1}_F - \mathbf{1}_{E\cup F}$. Since $E \cup F = (E^c \cap F^c)^c$, it suffices to show $\mathcal{M}$ is closed under finite intersections.

For $B \in \mathcal{A}$ fixed, define $\mathcal{M}_B := \{E \in \mathcal{M} : \ E \cap B \in \mathcal{M}\}$. Then $\mathcal{M}_B$ is a monotone class (if $E_n \uparrow E$ with $E_n \cap B \in \mathcal{M}$, then $E_n \cap B \uparrow E \cap B$, so $E \cap B \in \mathcal{M}$; similarly for $\downarrow$). Since $\mathcal{A}$ is a π-system, $\mathcal{A} \subseteq \mathcal{M}_B$. By the monotone class theorem for sets, $\sigma(\mathcal{A}) \subseteq \mathcal{M}_B$.

Since this holds for every $B \in \mathcal{A}$, a second application (fixing $E \in \sigma(\mathcal{A}) \cap \mathcal{M}$ and defining $\mathcal{H}_E := \{B \in \mathcal{M} : \ E \cap B \in \mathcal{M}\}$) gives $\sigma(\mathcal{A}) \subseteq \mathcal{H}_E$. Hence $\mathcal{M} \cap \sigma(\mathcal{A})$ is closed under finite intersections.

Step 4 (Conclude). From Steps 2 and 3, $\mathcal{M}$ is a monotone class containing the algebra generated by $\mathcal{A} \cup \{\Omega\}$ (complements + finite intersections + Ω). By the monotone class theorem for sets, $\sigma(\mathcal{A}) \subseteq \mathcal{M}$. By Step 1, $\mathcal{H}$ contains all bounded $\sigma(\mathcal{A})$-measurable functions. □

A.4 Proof of the Lebesgue–Stieltjes Theorem

Lemma A.3. *Let $\mathcal{S}$ be a semialgebra and let μ satisfy* (C-1) *(finite additivity on disjoint unions) on $\mathcal{S}$. Let $\overline{\mathcal{S}}$ be the algebra generated by $\mathcal{S}$ (finite disjoint unions of members of $\mathcal{S}$), and define $\overline{\mu} : \overline{\mathcal{S}} \to [0, \infty]$ by*

$$\text{if } A = \bigsqcup_{j=1}^{m} S_j \text{ with } S_j \in \mathcal{S}, \quad \overline{\mu}(A) := \sum_{j=1}^{m} \mu(S_j).$$

Then:

(a) If $A, B_i \in \overline{\mathcal{S}}$ and $A = \bigsqcup_{i=1}^{n} B_i$, then $\overline{\mu}(A) = \sum_{i=1}^{n} \overline{\mu}(B_i)$.

(b) If $A, B_i \in \overline{\mathcal{S}}$ and $A \subset \bigcup_{i=1}^{n} B_i$, then $\overline{\mu}(A) \leq \sum_{i=1}^{n} \overline{\mu}(B_i)$.

Proof. (a) For each i, pick a disjoint decomposition $B_i = \bigsqcup_{j=1}^{m_i} S_{ij}$ with $S_{ij} \in \mathcal{S}$. Since the B_i are pairwise disjoint and each $S_{ij} \subset B_i$, the family $\{S_{ij}\}_{i,j}$ is pairwise disjoint and

$$A = \bigsqcup_{i=1}^{n} B_i = \bigsqcup_{i=1}^{n} \bigsqcup_{j=1}^{m_i} S_{ij}.$$

By the definition of $\overline{\mu}$ and (C-1),

$$\overline{\mu}(A) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mu(S_{ij}) = \sum_{i=1}^{n} \overline{\mu}(B_i),$$

as claimed.

Before (b) note a basic consequence of (a): **monotonicity on $\overline{\mathcal{S}}$.** If $D, E \in \overline{\mathcal{S}}$ with $D \subset E$, then $E = D \sqcup (E \setminus D)$ with $E \setminus D \in \overline{\mathcal{S}}$, so $\overline{\mu}(E) = \overline{\mu}(D) + \overline{\mu}(E \setminus D) \geq \overline{\mu}(D)$.

(b) From $A \subset \bigcup_{i=1}^{n} B_i$ build a disjoint refinement of the cover by setting

$$C_1 := B_1, \qquad C_k := B_k \setminus \bigcup_{i<k} B_i \quad (k = 2, \ldots, n).$$

Then $C_k \in \overline{\mathcal{S}}$, the C_k are pairwise disjoint, and $\bigcup_{k=1}^{n} C_k = \bigcup_{k=1}^{n} B_k$. Hence

$$A = \bigsqcup_{k=1}^{n} \big(A \cap C_k\big),$$

with each $A \cap C_k \in \overline{\mathcal{S}}$. By part (a),

$$\overline{\mu}(A) = \sum_{k=1}^{n} \overline{\mu}(A \cap C_k) \leq \sum_{k=1}^{n} \overline{\mu}(C_k) \leq \sum_{k=1}^{n} \overline{\mu}(B_k),$$

where the first inequality uses monotonicity and the second uses $C_k \subset B_k$. $\square$

Proof of Thm. 1.35. We define μ on the semialgebra $\mathcal{S}$ from example 1.14 with $d = 1$ by the formula $\mu\big((a, b]\big) = F(b) - F(a)$. For $a = -\infty$ or $b = \infty$, we extend $F(x)$ by defining:

$$F(-\infty) = \lim_{x \downarrow -\infty} F(x) \text{ and } F(\infty) = \lim_{x \uparrow \infty} F(x).$$

We are going to apply Theorem 1.19. Condition C-1 (finite additivity on disjoint unions) is easy to check by induction on the number of elements in the cover and we concentrate on condition C-2 (σ-subadditivity on countable disjoint partitions).

Suppose first that $-\infty < a < b < \infty$, and $(a, b] \subset \cup_{i=1}^{\infty}(a_i, b_i]$ where (without loss of generality) $-\infty < a_i < b_i < \infty$. (We use a slightly more lax condition $(a, b] \subset \cup_{i=1}^{\infty}(a_i, b_i]$ instead of $(a, b] \subset \sqcup_{i=1}^{\infty}(a_i, b_i]$.)

We aim to use a compactness argument to convert a countable cover of $(a, b]$ to a finite cover of $(a, b]$. This cannot be done directly since $(a, b]$ is not closed and $(a_i, b_i]$ are not open. To circumvent this, we want to shorten the open interval $(a, b]$ to a closed interval $[a + \delta, b]$ with an arbitrary small loss in measure; so for a given $\varepsilon > 0$, choose $\delta > 0$ so that

$$F(a + \delta) < F(a) + \varepsilon,$$

We also want to extend the intervals of the cover to open intervals, so choose $\eta_i > 0$ so that

$$F(b_i + \eta_i) < F(b_i) + \frac{\varepsilon}{2^i}.$$

We can find necessary δ and η_i because $F(x)$ is right-continuous.

Then, by compactness, we can find a finite subcover of $[a + \delta, b]$. Let us denote it (α_j, β_j), $1 \leq j \leq J$. Then, the finite union of $(\alpha_j, \beta_j]$ covers $(a + \delta, b]$ and by Lemma A.3,

$$F(b) - F(a + \delta) \leq \sum_{j=1}^{J} \big(F(\beta_j) - F(\alpha_j)\big) \leq \sum_{i=1}^{\infty} \big(F(b_i + \eta_i) - F(a_i)\big).$$

This implies that

$$F(b) - F(a) \leq 2\varepsilon + \sum_{i=1}^{\infty} \big(F(b_i) - F(a_i)\big).$$

Since $\varepsilon > 0$ is arbitrary this proves that condition C-2 holds if $-\infty < a < b < \infty$. To cover the case when either $a = -\infty$ or $b = \infty$, we can take a finite interval $(A, B] \subset (a, b]$ and note that for this interval

$$F(B) - F(A) \leq \sum_{i=1}^{\infty} \big(F(b_i) - F(a_i)\big).$$

by what we already proved. Then we pass to an appropriate limit in this inequality. For example, if $b = \infty$ and $a < \infty$, then we take $A = a$, and write:

$$F(\infty) - F(a) = \lim_{B \uparrow b} F(B) - F(a) \leq \sum_{i=1}^{\infty} \big(F(b_i) - F(a_i)\big).$$

Other cases are covered similarly. By Caratéodory Theorem, this shows the existence of the extending measure. The uniqueness follows because this measure is σ-finite. For example we can take a countable cover of $\mathbb{R}$ by intervals $(n, n+1]$ where $n \in \mathbb{Z}$. The measure of each of these intervals is clearly finite. □

A.5 Measures on $\mathbb{R}^d$

Can one generalize Theorem 1.35 to construct measures on $\mathbb{R}^d$? In fact, we only aim to define probability measures, that is, the measures, for which $\mu(\Omega) = 1$.

A suitable semialgebra was constructed in Example 1.14 and we need a suitable generalization of the Stieltjes measure function. Then, we take a function $F : \mathbb{R}^d \to \mathbb{R}$ and we require that

$$\mu\big((-\infty, b_1] \times \ldots \times (-\infty, b_d]\big) = F(b_1, \ldots, b_d).$$

How does this formula extends to finite rectangles $(a_1, b_1] \times \ldots \times (a_d, b_d]$?

For this extension we introduce a new notation. Let

$$A = (a_1, b_1] \times \ldots \times (a_d, b_d],$$
$$V(A) = \{a_1, b_1\} \times \ldots \times \{a_d, b_d\},$$

where $-\infty < a_i < b_i < \infty$. For $v \in V(A)$, let

$$\operatorname{sgn}(v) = (-1)^{\#\text{ of } a\text{'s in } v},$$

and let

$$\Delta_A F = \sum_{v \in V(A)} \operatorname{sgn}(v) F(v).$$

Then, it turns out that the additivity properties of measure μ imply that

$$\mu(A) = \Delta_A F.$$

For example, for $d = 2$,

$$\mu\big((a_1, b_1] \times (a_2, b_2]\big) = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2).$$

Definition A.4. A function $F : \mathbb{R}^d \to [0, 1]$ is called a **cumulative distribution function** if it has the following properties:

(i) F is nondecreasing, i.e, if $x \leq y$ then $F(x) \leq F(y)$. (Here we use the vector notation, $x = (x_1, \ldots, x_d)$, and $x \leq y$ means $x_i \leq y_i$ for all i.)

(ii) F is right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$.

(iii) If $x_n \downarrow -\infty$, i.e. each coordinate does, then $F(x_n) \downarrow 0$. If $x_n \uparrow +\infty$, then $F(x_n) \uparrow 1$.

(iv) $\Delta_A F \geq 0$ for all rectangles A.

Theorem A.5. *Suppose $F : \mathbb{R}^d \to [0,1]$ is a valid cumulative distribution function. Then there is a unique probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ so that $\mu(A) = \Delta_A F$ for all finite rectangles A.*

For the proof of this theorem see Theorem 1.1.11 in Durrett.

Example A.6. If we take function

$$F(x_1, \ldots, x_d) = \begin{cases} 1, & \text{if } 0 \leq x_i \leq 1 \text{ for all } i = 1, \ldots, d, \\ 0, & \text{otherwise,} \end{cases}$$

then all conditions of Theorem A.5 are satisfied and the resulting measure is called the Lebesgue measure on the cube $[0,1]^d$.

0	2/3	1
0	0	2/3
0	0	0

Figure A.1

Example A.7. This example illustrate that condition (iv) is not redundant.

Consider function

$$F(x_1, x_2) = \begin{cases} 1, & \text{if } x_1, x_2 \geq 1 \\ \frac{2}{3}, & \text{if } x_1 \geq 1 \text{ and } 0 \leq x_2 < 1 \\ \frac{2}{3}, & \text{if } 0 \leq x_1 < 1 \text{ and } x_2 \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(See illustration in Fig. A.1.) Then conditions (i) – (iii) in the definition of the cdf (cumulative distribution function) are satisfied but we have

$$\mu\big((1/2, 1] \times (1/2, 1]\big) = 1 - \frac{2}{3} - \frac{2}{3} + 0 = -\frac{1}{3}.$$

which contradicts the positivity of measure μ.

A.6 Regularity of Borel measures on $\mathbb{R}^n$

Definition A.8 (Tightness, inner/outer regularity)**.** Let μ be a finite Borel measure on $\mathbb{R}^n$.

- **Tight** means: for every $\varepsilon > 0$ there exists a compact K with $\mu(K) > 1 - \varepsilon$.
- **Inner regular on a class $\mathcal{C}$** means:

$$\mu(E) = \sup\{\mu(K) : K \subset E,\ K \text{ compact}\}$$

  for all $E \in \mathcal{C}$.
- **Outer regular** means: $\mu(E) = \inf\{\mu(U) : E \subset U,\ U \text{ open}\}$ for all Borel E.

Lemma A.9 (Tightness on $\mathbb{R}^n$)**.** *Every Borel probability measure μ on $\mathbb{R}^n$ is tight.*

Proof. Let $B_R = \{x : \|x\| \le R\}$. Then $B_R \uparrow \mathbb{R}^n$, hence by continuity from below, $\mu(B_R) \uparrow \mu(\mathbb{R}^n) = 1$. For given $\varepsilon > 0$ choose R with $\mu(B_R) > 1 - \varepsilon$. Since B_R is compact, we are done. □

Lemma A.10 (Inner regularity on open sets)**.** *If $U \subset \mathbb{R}^n$ is open and μ is a finite Borel measure, then*

$$\mu(U) = \sup_{m \in \mathbb{N}} \mu(K_m), \qquad K_m := \big\{x \in U : \operatorname{dist}(x, U^{\mathrm{c}}) \ge 1/m\big\} \cap \overline{B_m},$$

where $\overline{B_m} = \{x : \|x\| \le m\}$. Each K_m is compact and $K_m \uparrow U$.

Proof. K_m is closed and bounded, hence compact; clearly $K_m \subset U$. Given $x \in U$, $\operatorname{dist}(x, U^{\mathrm{c}}) > 0$ and $\|x\| < \infty$, so for large m we have $x \in K_m$. Thus $K_m \uparrow U$ and by continuity from below, $\mu(K_m) \uparrow \mu(U)$. □

Lemma A.11 (Outer regular on closed sets)**.** *Let μ be a finite Borel measure on $\mathbb{R}^n$. If $F \subset \mathbb{R}^n$ is closed and $F^\varepsilon := \{x : \operatorname{dist}(x, F) < \varepsilon\}$, then $F^\varepsilon \downarrow F$ and*

$$\mu(F) = \inf_{\varepsilon > 0} \mu(F^\varepsilon).$$

Proof. F^ε are open and decrease to F. Since $\mu(F^1) < \infty$, continuity from above gives $\mu(F) = \lim_{\varepsilon \downarrow 0} \mu(F^\varepsilon) = \inf_{\varepsilon > 0} \mu(F^\varepsilon)$. □

Theorem A.12 (Inner regular on all Borel sets)**.** *Let μ be a Borel probability on $\mathbb{R}^n$. For every Borel $E \subset \mathbb{R}^n$ and every $\varepsilon > 0$ there exists a compact $K \subset E$ with $\mu(E \setminus K) < \varepsilon$.*

Proof. Fix $\varepsilon > 0$. By tightness, choose R so that $\mu(\overline{B_R}^{\,c}) < \varepsilon/2$. Cover the compact set $\overline{B_R}$ by finitely many open balls $U_1, \dots, U_m$. By Lemma A.10, for each i pick a compact $K_i \subset U_i$ with

$$\mu\big((E \cap U_i) \setminus K_i\big) < \frac{\varepsilon}{2m}.$$

Set $K := \bigcup_{i=1}^{m} \big(K_i \cap E\big)$, which is compact and contained in E. Then

$$\begin{aligned} \mu(E \setminus K) \;&\le\; \mu(E \cap \overline{B_R}^{\,c}) + \mu\Big(\big(E \cap \overline{B_R}\big) \setminus \bigcup_{i=1}^{m} (K_i \cap E)\Big) \\ &\le\; \frac{\varepsilon}{2} + \sum_{i=1}^{m} \mu\big((E \cap U_i) \setminus K_i\big) \;<\; \varepsilon. \end{aligned}$$

□

Corollary A.13 (Working form used in Kolmogorov)**.** *For any Borel set $B \subset \mathbb{R}^n$ and $\delta > 0$ there exists a compact $A \subset B$ with $\mu(B \setminus A) \le \delta$. Equivalently: every Borel probability on $\mathbb{R}^n$ is inner regular.*

A.7 Kolmogorov's extension theorem with proof

Lemma A.14 (Continuity at $\emptyset \iff \sigma$-additivity on an algebra)**.** *Let $\mathcal{R}$ be an algebra of subsets of Ω, and let $\mu : \mathcal{R} \to [0, \infty)$ be finite (i.e. $\mu(E) < \infty$ for all $E \in \mathcal{R}$) and finitely additive. Then the following are equivalent:*

(i) *μ is σ-additive on $\mathcal{R}$.*

(ii) *(**continuity at $\emptyset$**) If $E_n \downarrow \emptyset$ with $E_n \in \mathcal{R}$, then $\mu(E_n) \downarrow 0$.*

Proof. **(i)⇒(ii).** If $E_n \downarrow \emptyset$, set $F_1 := E_1 \setminus E_2$ and $F_n := E_n \setminus E_{n+1}$ for $n \ge 2$. Then $E_1 = \bigsqcup_{n \ge 1} F_n$ and (countable) additivity gives $\mu(E_1) = \sum_{n \ge 1} \mu(F_n)$; moreover $\mu(E_m) = \sum_{n \ge m} \mu(F_n)$, hence $\mu(E_m) \downarrow 0$.

(ii)⇒(i). First note that (ii) implies **continuity from below**: if $E_n \uparrow E$ with $E, E_n \in \mathcal{R}$, then $E \setminus E_n \downarrow \emptyset$, so $\mu(E_n) = \mu(E) - \mu(E \setminus E_n) \uparrow \mu(E)$.

Now let $(A_k)_{k \ge 1} \subset \mathcal{R}$ be pairwise disjoint with $A := \bigcup_{k \ge 1} A_k \in \mathcal{R}$. Set $U_N := \bigcup_{k=1}^{N} A_k \in \mathcal{R}$. By finite additivity, $\mu(U_N) = \sum_{k=1}^{N} \mu(A_k)$, and by continuity from below, $\mu(U_N) \uparrow \mu(A)$. Hence $\mu(A) = \sum_{k \ge 1} \mu(A_k)$. □

Corollary A.15 (Semiring version)**.** *Let $\mathcal{S}$ be a semiring on Ω, and let $\mu : \mathcal{S} \to [0, \infty)$ be finite and finitely additive. Then the following are equivalent:*

(i) *μ is σ-additive on $\mathcal{S}$ (**i.e.**, for pairwise disjoint $(A_k) \subset \mathcal{S}$ with $A := \bigsqcup_{k \ge 1} A_k \in \mathcal{S}$, one has $\mu(A) = \sum_{k \ge 1} \mu(A_k)$).*

(ii) *(**continuity at $\emptyset$ on $\mathcal{S}$**) If $E_n \downarrow \emptyset$ with $E_n \in \mathcal{S}$, then $\mu(E_n) \downarrow 0$.*

Proof. **(i)⇒(ii)** is as in Lemma A.14, using that each difference $E_n \setminus E_{n+1}$ is a **finite disjoint union** of members of $\mathcal{S}$ (semiring property), so disjointification still lies within the domain for σ-additivity.

For **(ii)⇒(i)**, pass to the algebra $\overline{\mathcal{S}}$ of all finite disjoint unions of sets in $\mathcal{S}$ and extend μ uniquely to a finitely additive $\overline{\mu} : \overline{\mathcal{S}} \to [0, \infty)$ by $\overline{\mu}(\bigsqcup_{j=1}^{m} S_j) := \sum_{j=1}^{m} \mu(S_j)$. If $C_n \downarrow \emptyset$ with $C_n \in \overline{\mathcal{S}}$, write a fixed finite disjoint decomposition

$C_1 = \bigsqcup_{j=1}^{m} S_j$ with $S_j \in \mathcal{S}$. Then $C_n \cap S_j \downarrow \emptyset$ in $\mathcal{S}$ for each j, so $\mu(C_n \cap S_j) \downarrow 0$ by (ii). Consequently,

$$\overline{\mu}(C_n) = \sum_{j=1}^{m} \mu(C_n \cap S_j) \ \downarrow \ 0,$$

i.e. $\overline{\mu}$ is continuous at $\emptyset$ on $\overline{\mathcal{S}}$. By Lemma A.14 (applied to the algebra $\overline{\mathcal{S}}$), $\overline{\mu}$ is σ-additive on $\overline{\mathcal{S}}$. In particular, for disjoint $(A_k) \subset \mathcal{S}$ with $A = \bigsqcup_{k\geq 1} A_k \in \mathcal{S} \subset \overline{\mathcal{S}}$,

$$\mu(A) = \overline{\mu}(A) = \sum_{k\geq 1} \overline{\mu}(A_k) = \sum_{k\geq 1} \mu(A_k).$$

□

Theorem A.16 (Kolmogorov extension on $\mathbb{R}^{\mathbb{N}}$)**.** *Let $(\mathbb{P}_n)_{n\geq 1}$ be probability measures on $\mathcal{B}(\mathbb{R}^n)$ such that*

$$\mathbb{P}_{n+1}(B \times \mathbb{R}) = \mathbb{P}_n(B) \qquad \forall\, B \in \mathcal{B}(\mathbb{R}^n),\ \forall n \geq 1.$$

Then there exists a unique probability measure $\mathbb{P}$ on

$$\big(\mathbb{R}^{\mathbb{N}},\ \mathcal{F} := \sigma\{\pi_{1:n}^{-1}(B) :\ B \in \mathcal{B}(\mathbb{R}^n),\ n \in \mathbb{N}\}\big)$$

with $\mathbb{P} \circ \pi_{1:n}^{-1} = \mathbb{P}_n$ for every $n \geq 1$.

Proof. Step 1 (premeasure on cylinders). Let $\mathcal{S} := \{\pi_{1:n}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^n),\ n \in \mathbb{N}\}$. Define $\overline{\mu} : \mathcal{S} \to [0,1]$ by

$$\overline{\mu}\big(\pi_{1:n}^{-1}(B)\big) := \mathbb{P}_n(B).$$

This is well-defined: if $\pi_{1:n}^{-1}(B) = \pi_{1:m}^{-1}(C)$ with $m \geq n$, then $C = B \times \mathbb{R}^{m-n}$ and by consistency $\mathbb{P}_m(C) = \mathbb{P}_n(B)$. Finite additivity on $\mathcal{S}$ follows since $\mathcal{S}$ is a semiring and each $\mathbb{P}_n$ is a measure.

Step 2 (σ-additivity via continuity at $\varnothing$). By Corollary A.15, a finite premeasure on a semiring is σ-additive iff it is continuous at $\varnothing$ for decreasing sequences from the domain. Let $(\widetilde{B}_k) \subset \mathcal{S}$ decrease with $\bigcap_k \widetilde{B}_k = \varnothing$; suppose, to the contrary, that $\lim_k \overline{\mu}(\widetilde{B}_k) = \delta > 0$. After discarding finitely many initial terms, assume $\overline{\mu}(\widetilde{B}_k) \geq \delta$ for all k.

Write each cylinder as $\widetilde{B}_k = \pi_{1:n_k}^{-1}(B_k)$ with $B_k \in \mathcal{B}(\mathbb{R}^{n_k})$, and (reindexing if needed) assume $n_k \uparrow$.

By inner regularity of Borel probabilities on $\mathbb{R}^n$, choose compact $A_k \subset B_k$ with $\mathbb{P}_{n_k}(B_k \setminus A_k) \leq \delta/2^{k+1}$. Set $\widetilde{A}_k := \pi_{1:n_k}^{-1}(A_k) \in \mathcal{S}$ and define the decreasing cylinders

$$\widetilde{C}_k := \bigcap_{j=1}^{k} \widetilde{A}_j \in \mathcal{S}.$$

Then $\widetilde{C}_k \subseteq \widetilde{B}_k$ and

$$\overline{\mu}(\widetilde{B}_k \setminus \widetilde{C}_k) \leq \sum_{j=1}^{k} \overline{\mu}(\widetilde{B}_j \setminus \widetilde{A}_j) = \sum_{j=1}^{k} \mathbb{P}_{n_j}(B_j \setminus A_j) \leq \delta/2,$$

so $\overline{\mu}(\widetilde{C}_k) \geq \delta/2$ for all k; in particular, each $\widetilde{C}_k \neq \emptyset$.

Each $\widetilde{C}_k$ depends only on the first n_k coordinates and has compact, nonempty base

$$C_k := \{x_{1:n_k} \in \mathbb{R}^{n_k} : (x_{1:j}) \in A_j \text{ for all } 1 \leq j \leq k\} = \Big(\bigcap_{j=1}^{k-1} (A_j \times \mathbb{R}^{n_k - n_j})\Big) \cap A_k,$$

hence C_k is (closed subset of) the compact set A_k, thus compact.

Pick $x^{(k)} \in C_k$ for each k. By compactness of C_1, the sequence of first coordinates has a convergent subsequence; pass to that subsequence and use compactness of C_2 to extract a further subsequence with the first two coordinates convergent; continue diagonally. This yields $x^* = (x_1^*, x_2^*, \ldots) \in \mathbb{R}^{\mathbb{N}}$ with $(x_1^*, \ldots, x_j^*) \in A_j$ for all j, i.e. $x^* \in \bigcap_k \widetilde{C}_k$.

But $\widetilde{C}_k \subseteq \widetilde{B}_k$ and $\bigcap_k \widetilde{B}_k = \varnothing$, a contradiction. Hence $\overline{\mu}(\widetilde{B}_k) \downarrow 0$, proving σ-additivity on $\mathcal{S}$.

Step 3 (uniqueness). The class $\mathcal{S}$ is a π-system generating $\mathcal{F}$. If $\mathbb{P}'$ is another probability on $(\mathbb{R}^{\mathbb{N}}, \mathcal{F})$ with the same finite-dimensional marginals, then $\mathbb{P}$ and $\mathbb{P}'$ agree on $\mathcal{S}$, hence on $\mathcal{F}$ by the π–λ theorem. □

Similar to the previous theorem, the assumption of countable additivity in the Caratheodory theorem can be checked for infinite products of probability measures on finite sets, and on $[0, 1]$. However, it should be noted that it is not automatically satisfied for arbitrary $(\Omega_i, \mathcal{F}_i)$. Usually one requires additionally that $(\Omega_i, \mathcal{F}_i)$ is a metric space with Borel-sigma algebra $\mathcal{F}$ and that every $\mathbb{P}_i$ satisfy an appropriate condition.

Definition A.17 (Standard Borel space)**.** A measurable space $(\Omega, \mathcal{F})$ is **standard Borel** if there exists a Polish topology τ on Ω (complete separable metric topology) such that $\mathcal{F} = \mathcal{B}(\Omega, \tau)$, the Borel σ-algebra of (Ω, τ). Equivalently, $(\Omega, \mathcal{F})$ is measurably isomorphic to $(B, \mathcal{B}(B))$ for some Borel subset B of a Polish space (e.g. $B \subset \mathbb{R}$ with the subspace Borel σ-algebra).

Why we care (probability "state spaces"). If each one-step state space $(\Omega_t, \mathcal{F}_t)$ is standard Borel, then: (i) countable products $\prod_t (\Omega_t, \mathcal{F}_t)$ are standard Borel; (ii) Kolmogorov extension holds cleanly (existence/uniqueness on the product σ-algebra); (iii) regular conditional probabilities and disintegrations exist; (iv) powerful measurable selection theorems apply.

Canonical examples.

$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, any countable set with all subsets, $\{0, 1\}^{\mathbb{N}}$, $\mathbb{N}^{\mathbb{N}}$, the Cantor set C with $\mathcal{B}(C)$,

any Borel subset $B \subset \mathbb{R}^m$ with the induced $\mathcal{B}(B)$, and any countable product / countable disjoint union of the above.

Non-examples.

- $(\mathbb{R}, \mathcal{L})$ where $\mathcal{L}$ is the **Lebesgue** σ-algebra (completed Borel): not standard Borel (no Polish topology has $\mathcal{L}$ as its Borel sets).
- $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ (all subsets), or an uncountable set with the countable–cocountable σ-algebra: not standard Borel.

Basic structure facts (useful to remember). - Any uncountable standard Borel spaces are Borel-isomorphic to each other (in particular, to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$). Thus, up to measurable isomorphism, standard Borel spaces come in three "sizes": finite, countably infinite, or continuum. - Standard Borel $\Rightarrow$ Radon: every Borel probability on a standard Borel space is tight and inner regular w.r.t. some compatible Polish topology on the same underlying set. - Closure: Borel subspaces, Borel images under Borel isomorphisms, countable products, and countable disjoint unions of standard Borel spaces are standard Borel.

Link to Kolmogorov extension. Working with standard Borel $(\Omega_t, \mathcal{F}_t)$ ensures that each finite product $(\prod_{t\in J} \Omega_t, \bigotimes_{t\in J} \mathcal{F}_t)$ is again standard Borel; hence its Borel probabilities are inner regular (compact approximation exists), which is exactly what we used to turn the cylinder premeasure into a measure (continuity at $\varnothing$) and to get uniqueness via the π–λ theorem.

A.8 Random Variables

A.8.1 Random variables as measurable functions

An example of a random variable (measurable function) is the **indicator function** of an event $A \in \mathcal{F}_1$ defined as

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

Here $\Omega_2 = \{0, 1\}$ and $\mathcal{F}_2 = \{\emptyset, \{0\}, \{1\}, \Omega_2\}$.

Conversely, the indicator function of a set $A \subset \Omega$ which is not measurable, is not a random variable.

A map from a topological space to another topological space is called **Borel measurable** (or $\mathcal{B}$-measurable) if it is measurable with respect to the Borel σ-algebras on these spaces.

A map from $\mathbb{R}^d$ to a topological space S is called **Lebesgue measurable** if the pre-images of Borel sets are Lebesgue measurable.

The point of the definition of the Borel and Lebesgue measurable functions is to ensure that the events $\{X \in A\}$ have a well-defined probability for sufficiently nice sets of elements in Ω_2, that is, for all Borel sets A.

One needs some tools to check that a function is measurable. First, if $\mathcal{F}_2$ is generated by a semi-algebra $\mathcal{S}$ then it is enough to check that inverse images of the sets in $\mathcal{S}$.

More generally, let $\mathcal{S}$ denote a collection of subsets of Ω (not necessarily a semialgebra) and $\sigma(\mathcal{S})$ denote the minimal σ-algebra that contains all sets in $\mathcal{S}$.

Lemma A.18. *Let $X : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$. If $X^{-1}(A) \in \mathcal{F}_1$ for all $A \in \mathcal{S}$ and $\mathcal{F}_2 = \sigma(\mathcal{S})$, then X is measurable.*

Proof. Observe that the operation of taking inverse image, $A \to X^{-1}(A)$, preserves unions, intersections and complements:

$$X^{-1}\big(\cup_i A_i\big) = \cup_i X^{-1}(A_i), \tag{A.3}$$

$$X^{-1}(A^c) = \Big(X^{-1}(A)\Big)^c. \tag{A.4}$$

Let $\mathcal{T}$ be a collection of sets A in $\mathcal{F}_2$, for which $X^{-1}(A) \in \mathcal{F}_1$. By assumption, $\mathcal{S} \subset \mathcal{T}$. In addition, $\mathcal{T}$ must be a σ-algebra. Indeed, if not, then we could find some sets in $\mathcal{T}$ that would violate the axioms of σ-algebra, and by applying properties (A.3) and (A.4) we would find sets in $\mathcal{F}_1$ which would also violate these axioms. However, this is impossible because $\mathcal{F}_1$ is a σ-algebra by assumption.

Since $\mathcal{F}_2 = \sigma(\mathcal{S})$ is the minimal σ-algebra that contains $\mathcal{S}$, we conclude that $\mathcal{T} = \mathcal{F}_2$, and therefore X is measurable. □

By using this lemma we observe that **the continuous maps are $\mathcal{B}$-measurable.** Indeed, the Borel σ-algebra is generated by open sets and the pre-image of an open set by continuous map is open.

However the class of $\mathcal{B}$-measurable functions is significantly larger, since the pre-images of the open sets are not required to be open, as in the case of continuous functions, but only required to be Borel sets. For example, one can check that monotone right-continuous functions are measurable.

To extend our collection of measurable functions further one needs an additional tool.

The next lemma is obvious from definitions.

Lemma A.19. *If maps $X_1 \colon (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ and $X_2 \colon (\Omega_2, \mathcal{F}_2) \to (\Omega_3, \mathcal{F}_3)$ are measurable, then their composition $X_2 \circ X_1 \colon (\Omega_1, \mathcal{F}_1) \to (\Omega_3, \mathcal{F}_3)$ is also measurable.*

In particular, the composition of Borel measurable functions is Borel measurable.

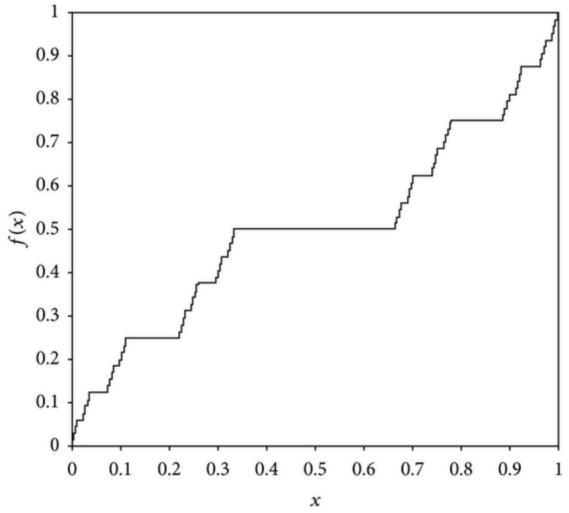
Example A.20. By using Lemmas A.18 and A.19, it is easy to see that if X is a random variable, then X^n, $\log(X)$ and many other functions of X are measurable.

Example A.21 (The composition of Lebesgue measurable functions is not necessarily Lebesgue measurable.).

This a somewhat involved example that shows that the Borel measurability is sometimes preferable over the Lebesgue measurability. As many counterexamples it involves the Cantor set.

Recall that the **Cantor set** $C \subset [0, 1]$ is defined by removing $(1/3, 2/3)$ from $[0, 1]$ and then iteratively removing the middle third of each interval that remains.

Exercise A.22. The Cantor set C is closed. The Lebesgue measure of C is 0.



Define the function F by setting $F(x) = 0$ for $x \leq 0$, $F(x) = 1$ for $x \geq 1$, $F(x) = 1/2$ for $x \in [1/3, 2/3]$, then $F(x) = 1/4$ for $x \in [1/9, 2/9]$, $F(x) = 3/4$ for $x \in [7/9, 8/9]$, and so on.

It can be checked that $F(x)$ is a non-decreasing continuous function, which is called the **Cantor-Lebesgue function** or the **Cantor staircase**.

Let $f\colon [0,1] \to [0,1]$ be the Cantor-Lebesgue function restricted to the interval $[0,1]$. This is a monotonic and continuous function, and the image $f(C)$ of the Cantor set C is all of $[0,1]$. Define $g(x) = x + f(x)$. Then $g\colon [0,1] \to [0,2]$ is a strictly monotonic and continuous map, so its inverse $h = g^{-1}$ is continuous, too.

Observe that $g(C)$ has measure one in $[0,2]$: this is because f is constant on every interval in the complement of C, so g maps such an interval to an interval of the same length. It follows that there is a non-Lebesgue measurable subset A of $g(C)$. (This is by Vitali's theorem: a subset of $\mathbb{R}$ is a Lebesgue null set if and only if all its subsets are Lebesgue measurable. So if all subsets of $g(C)$ are Lebesgue measurable, then $g(C)$ has measure null, contradiction.)

Put $B = g^{-1}(A) \subset C$. Then B is a Lebesgue measurable set as a subset of the Lebesgue null set C, so the characteristic function 1_B of B is Lebesgue measurable.

The function $k = 1_B \circ h$ is the composition of the Lebesgue measurable function 1_B and the continuous function h, but k is not Lebesgue measurable, since $k^{-1}(1) = (1_B \circ h)^{-1}(1) = h^{-1}(B) = g(B) = A$.

Now we prove a result that gives a criterion for measurability of functions of several variables.

The product of measurable spaces $(S_1, \mathcal{S}_1)$, ..., $(S_n, \mathcal{S}_n)$ is the set $S_1 \times \ldots \times S_n$ with the σ-algebra which is generated by products $A_1 \times \ldots \times A_n$, where $A_1 \in \mathcal{S}_1, \ldots, A_n \in \mathcal{S}_n$.

Theorem A.23. *Suppose $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$ and $(S_i, \mathcal{S}_i)$, $i = 1, \ldots, n$, are measurable spaces, that $X_i\colon \Omega \to S_i$, are measurable, and f is a measurable map from $(S_1, \mathcal{S}_1) \times \ldots \times (S_n, \mathcal{S}_n)$ to $(\Omega_2, \mathcal{F}_2)$. Then $f(X_1, X_2, \ldots, X_n)$ is a measurable map from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$.*

Proof. In view of Lemma A.19, it suffices to prove that the map $X\colon \Omega \to S_1 \times \ldots \times S_n$, defined as $\omega \to \big(X_1(\omega), \ldots, X_n(\omega)\big)$, is measurable. We use Lemma A.18 and observe that for any $A_1 \in \mathcal{S}_1$, ..., $A_n \in \mathcal{S}_2$,

$$X^{-1}(A_1 \times \ldots \times A_n) \equiv \{X \in A_1 \times \ldots \times A_n\} = \bigcap_{i=1}^{n} \{X_i \in A_i\} \in \mathcal{F}_1,$$

and therefore X is measurable. □

Here is an application of this theorem

Theorem A.24. *If $X_1, \ldots, X_n$ are random vectors, then $X_1 + \ldots + X_n$ is a random vector.*

Proof. In view of Theorem A.23, it is enough to check that $f(x_1, \ldots, x_n) = x_1 + \ldots + x_n$ is a measurable map. This holds because this function is continuous. □

Here is another useful result. Roughly speaking it says that limits of measurable functions are measurable.

Theorem A.25. *If X_1, X_2, ..., are random variables, then so are*

$$\inf_n X_n, \quad \sup_n X_n \quad \limsup_n X_n \quad \liminf_n X_n.$$

Proof. Let us check the statements for inifmums. We apply Lemma A.18. Observe that $\inf X_n < a$ if and only if there is at least one $X_n < a$. That is,

$$\{\inf_n X_n < a\} = \bigcup_n \{X_n < a\} \in \mathcal{F},$$

which shows measurability of $\inf X_n$.

Then, by definition,

$$\liminf_{n\to\infty} X_n = \sup_n \Big(\inf_{m \geq n} X_m \Big).$$

If $Y_n = \inf_{m\geq n} X_m$, then Y_n is a random variable by what we just proved about infimums. Then $\sup_n Y_n$ is also a random variable. □

One imprecision in the above theorem is that even if X_n random variables, that is, they take values in $\mathbb{R} = (-\infty, \infty)$, the infimums and supremums can values $-\infty$ and $+\infty$. For this reason it is useful to allow random variables to take values in $\mathbb{R}^* = [+\infty, -\infty]$.

A.8.2 Almost sure convergence

From Theorem A.25, we see that the set

$$\Omega_o = \{\omega : \lim_{n\to\infty} X_n \text{ exists. } \} = \{\omega : \limsup_{n\to\infty} X_n - \liminf_{n\to\infty} X_n = 0\}$$

is a measurable set. If $\mathbb{P}(\Omega_o) = 1$, then we say that X_n converges **almost surely** or **a.s.** for short. In order to define a limit on the whole set, we define the limit random variable as

$$X_\infty = \limsup_{n\to\infty} X_n.$$

We write $X_n \xrightarrow{a.s.} X$ if X differs from X_∞ on a set of measure 0.

A.8.3 σ-algebras generated by functions

Let $(X_i, i \in I)$ be a family of functions that take a set Ω to measurable spaces $(S_i, \mathcal{S}_i)$, $i \in I$. Here, $I \neq \varphi$ is an arbitrary index set (i.e., possibly uncountable).

For every X_i we can consider $X_i^{-1}(\mathcal{S}_i)$, the collection of inverse images for all sets in $\mathcal{S}_i$. One can check that these collections are σ-algebras.

Then the smallest σ-algebra generated by $X_i^{-1}(\mathcal{S}_i)$ is called the **σ-algebra generated by** $(X_i, i \in I)$ and denoted by $\sigma(X_i, i \in I)$. We can also define it as the smallest σ-algebra on Ω with respect to which each X_i is measurable.

This construction is usually used when we have X_i that arrive one after another and want to know how the corresponding σ-algebras increase:

$$\sigma(X_1) \subset \sigma(X_1, X_2) \subset \sigma(X_1, X_2, X_3) \subset \dots$$

This is called the **filtration** of σ - algebras.

A.8.4 Distributions

If X is a real r.v. defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then X induces a probability measure on $(\mathbb{R}, \mathcal{B})$ called its **distribution measure**. By definition, $\mu(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$, as a function of Borel sets B of $\mathbb{R}$. To show that μ is a probability measure one needs to check countable-additivity. However, it is simply inherited from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Namely for disjoint B_i's,

$$\begin{aligned}\mu(\sqcup_i B_i) &= \mathbb{P}[X^{-1}(\sqcup_i B_i)] \\ &= \mathbb{P}(\sqcup_i X^{-1}(B_i)) = \sum_i \mathbb{P}(X^{-1}(B_i)) = \sum_i \mu(B_i)\end{aligned}$$

The distribution measure of a r.v. X can described by its **cumulative distribution function** (cdf), $F(x) = \mathbb{P}\big(\{\omega : X(\omega) \leq x\}\big)$.

Theorem A.26. *A cdf F of every probability measure on $\mathbb{R}$ has the following properties:*

1. *F is a non-decreasing function of x.*
2. *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$*
3. *F is right continuous, i.e., $\lim_{y\downarrow x} F(y) = F(x)$*

Proof. Let us prove that F is right continuous. Observe that if $y \downarrow x$ then $\{X \leq y\} \downarrow \{X \leq x\}$. Then, for measure $\mathbb{P}$ we can use the continuity from above property from Theorem 1.11 to conclude that

$$\mathbb{P}(\{X \leq y\}) \downarrow \mathbb{P}(\{X \leq y\}),$$

which is exactly what we wanted to prove.

Refer to Theorem 1.2.1 in Durrett for the complete proof of this theorem. □

Theorem A.27. *If F satisfies the properties of Theorem A.26, then it is the distribution function of a random variable that takes values in $\mathbb{R}$.*

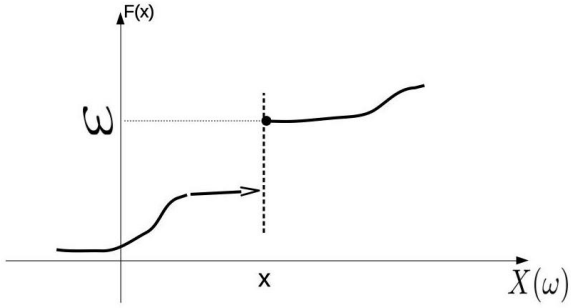


Figure A.2: Construction of a r.v. with a given CDF

Proof. (We are following the proof of Theorem 1.2.2 in Durrett.) The proof is by construction. Let $F : \mathbb{R} \to [0,1]$ have properties $1, 2, 3$ in Theorem A.26. We will construct a random variable defined on $(\Omega = (0,1], \mathcal{B}((0,1]), \mathbb{P})$, where $\mathbb{P}$ denotes the Lebesgue measure, and show that it has the distribution function F.

The idea of the following proof is that for every $\omega \in (0,1]$, we trying to define $X(\omega)$ as the inverse of the function $F(x)$. This is not always possible so we do it in a sophisticated way.

We define

$$X(\omega) = \sup\{x : F(x) < \omega\}.$$

One needs to check the measurability of this function, that is, the fact that the inverse image $X^{-1}(A)$ of every Borel set A is Borel. By Lemma A.18, it is enough to check that $\{\omega : X(\omega) < x\}$ is Borel for every $x \in \mathbb{R}$. Note that $X(\omega)$ is non-decreasing. Therefore the set $\{\omega : X(\omega) < x\}$ is either $(0, \alpha)$ or $(0, \alpha]$ for some $\alpha \in (0,1]$ and these are Borel sets.

Then, note that if we manage to show that the following sets are equal,

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

then the definitions of the distribution function and the Lebesgue measur imply that

$$F_X(x) := \mathbb{P}(\omega : X(\omega) \leq x) = \mathbb{P}(\omega : \omega \leq F(x)) = F(x).$$

To check the set equality above, observe that $\omega \leq F(x)$ means tautologically that x is outside of the set $\{y : F(y) < \omega\}$ which means that $x \geq \alpha = X(\omega)$. Therefore,

$$\{\omega : \omega \leq F(x)\} \subset \{\omega : X(\omega) \leq x\}.$$

On the other hand, if $\omega > F(x)$, then since F is right continuous, there is an $\epsilon > 0$ so that $\omega > F(x+\epsilon)$ which means that $x+\varepsilon$ is in the set $\{y : F(y) < \omega\}$.

It follows that x is strictly less than the supremum of this set, $X(\omega)$, that is $X(\omega) \geq x + \epsilon > x$. This completes the proof of the set equality stated above and therefore completes the proof of the theorem. □

When the cdf (cumulative distribution measure) of a random variable X has the form

$$F(x) = \int_{-\infty}^{x} f(t)\, dt,$$

then we say that X has **density function** f, or that the distribution measure of X is **absolutely continuous** with respect to the Lebesgue measure on $\mathbb{R}$.

A probability measure $\mathbb{P}$ (or its associated distribution function) is said to be **discrete** if there is a countable set S with $\mathbb{P}(S^c) = 0$.

For example, $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x \geq 0$ is the cdf of a discrete probability measure, which is called **point mass at 0** or an **atom** at 0.

Example A.28 (Singular measures).

Recall the Cantor staircase function $F(x)$ from Example A.21. Since this function is continuous and non-decreasing, it is clear that it is a valid distribution function.

From the definition, we see that $dF/dx = 0$ for every x in the complement of the Cantor set C. As the Lebesgue measure of C is zero, we see that the derivative of F is zero except on a set of zero Lebesgue measure. In particular, for this distribution function, there is no function f for which $F(x) = \int_{-\infty}^{x} f(t)\, dt$.

At the same time, since $F(x)$ is continuous, the probability of each point equals zero, so this measure is not **discrete**.

Such distribution functions are called **Lebesgue singular distribution functions** and the corresponding measures are called **singular measures**.

Even discrete distribution functions can be quite complex.

Example A.29 (Distribution function with a dense subset of discontinuities).

Let $q_1, q_2, \ldots$ be an enumeration of the rational numbers and set

$$F(x) = \sum_{i=1}^{\infty} 2^{-i} \mathbf{1}_{[q_i, \infty)}(x).$$

Clearly, such F is non-decreasing, with limits 0 and 1 as $x \to -\infty$ and $x \to \infty$, respectively. It is not hard to check that F is also right continuous, hence a distribution function, whereas by construction F is discontinuous at each rational number.

Appendix B

Convergence Theorems and Inequalities

B.1 Lebesgue Integral

We recall here how the Lebesgue integral is defined for a general σ-finite measure μ on Ω. This is done in steps by considering a larger and larger class of functions.

1. Simple functions
2. Bounded functions
3. Non-negative functions
4. General measurable functions

At every step one have to check that the following properties hold:

(i) If $\varphi \geq 0$ almost everywhere ("a.e." for short), that is, the inequality can be violated only on a set of measure 0, then $\int \varphi \, d\nu \geq 0$.

(ii) For any $a \in \mathbb{R}$, $\int a\varphi \, d\mu = a \int \varphi \, d\mu$.

(iii) $\int (\varphi + \psi) \, d\mu = \int \varphi \, d\mu + \int \psi \, d\mu$.

Step 1: Simple random variables. A function φ is called simple if it can be written as $\varphi = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ for some disjoint Borel sets A_i with $\mu(A_i) < \infty$. Then the integral is defined as

$$\int \varphi \, d\mu = \sum_{i=1}^{n} c_i \mathbb{P}(A_i) \tag{B.1}$$

works. One can verify that the definition does not depend on the choice of the A_i in the decomposition and the properties (i) - (iii) hold.

Step 2. Compactly supported bounded random variables.

Here we assume that f is a bounded (measurable) function. By compactly supported we mean here that f vanishes outside of a set E with $\mu(E) < \infty$. Then one defines

$$\int f\,d\mu = \sup_{\varphi \le f} \int \varphi\,d\nu = \inf_{\psi \ge f} \int \psi\,d\nu, \tag{B.2}$$

where φ and ψ are simple functions. Here one needs to check that the supremum and infimums in the definition are equal to each other. In the proof of this fact it is essential that $f(x)$ is bounded. This is the crucial part of the construction so we give some details. As a consequence of properties (i) - (iii), for all $\varphi \le f \le \psi$, we have

$$\int \varphi\,d\mu \le \int \psi\,d\mu,$$

and, therefore,

$$\sup_{\varphi \le f} \int \varphi\,d\mu \le \inf_{\psi \ge f} \int \psi\,d\mu.$$

To prove the opposite inequality, suppose $|f| \le M$ and let

$$E_k = \left\{ x \in E : \frac{(k-1)M}{n} < f(x) \le \frac{kM}{n} \right\} \text{ for } -n \le k \le n$$
$$\varphi_n(x) = \sum_{k=-n}^{n} \frac{(k-1)M}{n} 1_{E_k} \text{ and } \psi_n(x) = \sum_{k=-n}^{n} \frac{kM}{n} 1_{E_k}.$$

Note that $\varphi_n(x)$ and $\psi_n(x)$ are simple functions (in particular, they are measurable), and that $\psi_n - \varphi_n = (M/n)1_E$, so

$$\int \psi_n - \varphi_n\,d\mu = \frac{M}{n}\mu(E).$$

Therefore,

$$\begin{aligned} \sup_{\varphi \le f} \int \varphi\,d\mu \ge \int \varphi_n\,d\mu &= \int \psi_n\,d\mu - \frac{M}{n}\mu(E) \\ &\ge \inf_{\psi \ge f} \int \psi\,d\mu - \frac{M}{n}\mu(E). \end{aligned}$$

Taking the limit $n \to \infty$ shows that

$$\sup_{\varphi \le f} \int \varphi\,d\mu \ge \inf_{\psi \ge f} \int \psi\,d\mu,$$

and completes the proof.

One also need to prove that (i) - (iii) hold. See Durrett for the complete proof.

Step 3: Nonnegative random variables.

In this case the definition is as follows.

$$\int f\,d\mu = \sup_{0\le h\le f} \int h\,d\mu,$$

where h are bounded and compactly supported. Again, one needs to check that properties (i) – (iii) hold.

Step 4: functions f, for which $\int |f|, d\mu < \infty$. Such functions are called **integrable**. Define

$$f^+(x) = f(x) \vee 0 \text{ and } f^-(x) = (-f(x)) \vee 0,$$

where $a \vee b := max(a, b)$. Then the integral of f is defined by

$$\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu.$$

The right hand side is well defined because $f^+, f^- \le |f|$ and therefore the integrals of these functions are finite. Again, one can check that properties (i) – (iii) hold.

Note that it is also possible to extend the definition in the last step to the case when only one of f^+ and f^- has a finite integral. In this case, f is called quasi-integrable and $\int f\,d\mu$ is either $+\infty$ or $-\infty$.

In the particular case when μ is the Lebesgue measure the integral is often denoted $\int f(x)\,dx$, and in the case when μ is the Lebesgue-Stieltjes measure on $\mathbb{R}$ with the Stieltjes function $G(x)$ the integral is often denoted $\int f(x)\,dG(x)$.

If μ is a counting measure on a countable set Ω, then one can write $\sum_{i\in\Omega} f(i)$ for $\int f\,d\mu$.

Another piece of notation is that for a set $E \subset \Omega$, one defines

$$\int_E f\,d\mu := \int f \cdot 1_E\,d\mu,$$

where 1_E is the indicator function of the set E.

B.2 DCT with convergence in measure

First, we introduce an additional concept for convergence of functions, the convergence in measure.

Definition B.1. A sequence of measurable functions f_n is said to converge to a measurable function f **in measure**, if

$$\lim_{n\to\infty} \mu\big(\{\omega : |f_n(\omega) - f(\omega)| \ge \varepsilon\}\big) = 0,$$

for every $\varepsilon > 0$.

When we talk about probability measures, this mode of convergence is called the convergence of random variables **in probability**. We will denote this convergence as $f_n \xrightarrow{\mu} f$ or $f_n \xrightarrow{\mathbb{P}} f$.

The convergence in measure is weaker than the almost sure convergence.

Example B.2. Consider the interval $\Omega = [0,1]$, divide it successively into $2, 3, 4, \ldots$ parts, and enumerate the intervals in succession. That is, $I_1 = [0, \frac{1}{2}]$, $I_2 = [\frac{1}{2}, 1]$, $I_3 = [0, \frac{1}{3}]$, $I_4 = [\frac{1}{3}, \frac{2}{3}]$, $I_5 = [\frac{2}{3}, 1]$, and so on. If $f_n(x) = 1_{I_n}(x)$, then it easy to check that f_n tends to 0 in measure but not almost everywhere. Indeed, $\mu(|f_n| > \epsilon) \to 0$ for any fixed $\varepsilon > 0$ and therefore $f_n \xrightarrow{\mu} 0$, but for any ω and arbitrary N we can find $n > N$ so that $f_n(\omega) = 1$. Hence for any $\omega \in \Omega$, the sequence $f_n(\omega)$ does not converge to 0 (or any other value). In particular, f_n does not converge to 0 almost surely.

This raises the question if the Dominated Convergence Theorem holds if functions only converge in measure, not almost surely.

Lemma B.3 (Subsequence principle for convergence in measure)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with $\mu(\Omega) < \infty$. If $f_n \to f$ in measure, then there exists a subsequence (f_{n_k}) such that $f_{n_k} \to f$ a.e.*

Proof. For $m \geq 1$ choose $n(m)$ such that

$$\mu\big(\{|f_{n(m)} - f| > 2^{-m}\}\big) < 2^{-m}.$$

Let $A_m := \{|f_{n(m)} - f| > 2^{-m}\}$. Then $\sum_m \mu(A_m) < \infty$, so by Borel–Cantelli, $\mu(\limsup_m A_m) = 0$. Hence for a.e. ω, for all large m we have $|f_{n(m)}(\omega) - f(\omega)| \leq 2^{-m}$, which implies $f_{n(m)}(\omega) \to f(\omega)$. □

Theorem B.4 (Dominated convergence with convergence in measure)**.** *Let $(\Omega, \mathcal{A}, \mu)$ satisfy $\mu(\Omega) < \infty$. Assume $f_n \to f$ in measure and $|f_n| \leq g$ a.e. for all n, where $g \in L^1(\mu)$. Then $f \in L^1(\mu)$ and*

$$\int |f_n - f|\, d\mu \to 0, \qquad \textit{in particular} \qquad \int f_n\, d\mu \to \int f\, d\mu.$$

Proof. From $f_n \to f$ in measure, extract a subsequence $f_{n_k} \to f$ a.e. (previous lemma). Also $|f| \leq g$ a.e. (since a further a.e.convergent subsequence forces this), hence $f \in L^1$.

Apply the (a.e.) Dominated Convergence Theorem to the subsequence:

$$\int |f_{n_k} - f|\, d\mu \to 0.$$

If the full sequence did not satisfy $\int |f_n - f|\, d\mu \to 0$, then there would exist $\varepsilon > 0$ and a subsequence (f_{m_j}) with $\int |f_{m_j} - f|\, d\mu \geq \varepsilon$ for all j. But (f_{m_j}) still converges to f in measure, so it has a further subsequence converging a.e., and by the (a.e.) DCT its integrals $\int |f_{m_{j_\ell}} - f|\, d\mu$ must tend to 0, a contradiction. Therefore $\int |f_n - f|\, d\mu \to 0$. □

B.3 Function spaces

Probability theory is closely related to functional analysis. In particular, we can often use the fact that random variables with a given number of moments can be thought of as belonging to some functional spaces. The most important for us will be Banach and Hilbert spaces.

Definition B.5. (i) Let X be a normed linear space with norm $||\cdot||_X$. If X is complete with respect to the induced metric $d(x,y) := ||x-y||_X$, it is called a **Banach space**.

(ii) If in addition norm $||\cdot||_X$ arises from an inner product $(\cdot,\cdot)_X$, then X is called a **Hilbert space**.

An example of Banach spaces are $\mathcal{L}^p(\Omega,\mu)$ spaces. Let $(\Omega,\mathcal{F},\mu)$ be a measure space.

Then $\mathcal{L}^p(\Omega,\mu)$ is a space of measurable functions f, which have a finite p-moment, $\int |f|^p < \infty$, factored by the following equivalence relation $f \sim g$ in $^p \iff f = g$ μ-everywhere.

Let $||X||_p := (\int ||X||^p)^{1/p}$ be the ^{p} **norm** of X. Define convergence in L^p as follows:

$$X_n \overset{\mathcal{L}^p}{\to} X \iff ||X_n - X||_p \to 0 \tag{B.3}$$

It can be shown that $\mathcal{L}^p$ is complete for $p \geq 1$, i.e. if

$$\lim_{n,m\to\infty} ||X_n - X_m||_p = 0 \Rightarrow \exists \text{ a r.v. } X \text{ s.t. } X_n \overset{\mathcal{L}^p}{\to} X. \tag{B.4}$$

Therefore $\mathcal{L}^p$ spaces are Banach spaces for $p \geq 1$. (For $p < 1$, $||\cdot||_p$ is not a norm.)

For $p = 2$, the space $\mathcal{L}^2$ is a Hilbert space with the inner product $(f,g) = \int fg \, d\mu$.

For $p = 1$, the space $\mathcal{L}^1$ is the space of all integrable functions.

For $p = \infty$, $\mathcal{L}^\infty$ is the space of essentially bounded functions with the norm,

$$||X||_\infty = \inf\{M : \mathbb{P}(|X| > M) = 0\}.$$

In general, the $\mathcal{L}^p$ spaces with larger p are more restrictive and are easier to handle.

B.4 Supporting lines and a countable representation of convex functions

Definition B.6 (Subgradient and subdifferential)**.** Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and let $x_0 \in \mathbb{R}$. A number $m \in \mathbb{R}$ is called a **subgradient** of φ at x_0 if

$$\varphi(x) \;\geq\; \varphi(x_0) + m(x - x_0) \qquad \text{for all } x \in \mathbb{R}. \tag{B.5}$$

The set of all subgradients at x_0 is the **subdifferential** and is denoted

$$\partial\varphi(x_0) := \{m \in \mathbb{R} : \; m \text{ satisfies (B.5)}\}.$$

Any affine function $L(x) = \varphi(x_0) + m(x - x_0)$ with $m \in \partial\varphi(x_0)$ is called a **supporting line** (or **supporting affine minorant**) of φ at x_0.

Lemma B.7 (Existence of subgradients on $\mathbb{R}$)**.** *If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then for every $x_0 \in \mathbb{R}$ the subdifferential $\partial\varphi(x_0)$ is a nonempty interval. More precisely, the one-sided derivatives*

$$\varphi'_-(x_0) := \lim_{h \downarrow 0} \frac{\varphi(x_0) - \varphi(x_0 - h)}{h}, \qquad \varphi'_+(x_0) := \lim_{h \downarrow 0} \frac{\varphi(x_0 + h) - \varphi(x_0)}{h}$$

exist in $[-\infty, +\infty]$ and satisfy $\varphi'_-(x_0) \le \varphi'_+(x_0)$, and

$$\partial\varphi(x_0) = [\varphi'_-(x_0), \varphi'_+(x_0)] \cap \mathbb{R}.$$

Proof sketch. For convex φ, the secant slopes are monotone: for $a < x < b$,

$$\frac{\varphi(x) - \varphi(a)}{x - a} \;\le\; \frac{\varphi(b) - \varphi(x)}{b - x}.$$

This monotonicity implies the existence of the one-sided derivative limits and the inequality $\varphi'_-(x_0) \le \varphi'_+(x_0)$.

Fix any $m \in [\varphi'_-(x_0), \varphi'_+(x_0)]$. For $x > x_0$, the monotonicity of secant slopes gives

$$\frac{\varphi(x) - \varphi(x_0)}{x - x_0} \;\ge\; \varphi'_+(x_0) \;\ge\; m,$$

so $\varphi(x) \ge \varphi(x_0) + m(x - x_0)$. Similarly, for $x < x_0$ one gets

$$\frac{\varphi(x) - \varphi(x_0)}{x - x_0} \;\le\; \varphi'_-(x_0) \;\le\; m,$$

which again rearranges to $\varphi(x) \ge \varphi(x_0) + m(x - x_0)$. Thus $m \in \partial\varphi(x_0)$ and $\partial\varphi(x_0) \neq \varnothing$. □

Theorem B.8 (Countable supporting-line representation)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex. For each rational $q \in \mathbb{Q}$ choose one $m(q) \in \partial\varphi(q)$ and define*

$$L_q(x) := \varphi(q) + m(q)(x - q).$$

Then $L_q(x) \le \varphi(x)$ for all x and

$$\varphi(x) = \sup_{q \in \mathbb{Q}} L_q(x) \qquad \text{for all } x \in \mathbb{R}.$$

Proof. The inequality $L_q \le \varphi$ holds by the definition of $m(q) \in \partial\varphi(q)$.

Fix $x \in \mathbb{R}$. Let $q_n \in \mathbb{Q}$ be any sequence with $q_n \to x$. Since convex functions on $\mathbb{R}$ are continuous, $\varphi(q_n) \to \varphi(x)$. Also, subgradients are locally bounded: on any compact interval I, all $m \in \partial\varphi(t)$ with $t \in I$ lie in a bounded interval (this follows from the monotonicity of secant slopes). Hence for n large, $m(q_n)$ is bounded and therefore

$$L_{q_n}(x) = \varphi(q_n) + m(q_n)(x - q_n) \longrightarrow \varphi(x) \quad \text{as } n \to \infty.$$

Thus $\sup_{q \in \mathbb{Q}} L_q(x) \ge \limsup_{n \to \infty} L_{q_n}(x) = \varphi(x)$. Together with $\sup_{q \in \mathbb{Q}} L_q(x) \le \varphi(x)$, we conclude equality. □

B.5 Details for Hölder's and Minkowski's Inequalities

In the proof of H'older inequality we used the following result.

Lemma B.9 (Young's inequality)**.** *Let $1 < p, q < \infty$ with $1/p + 1/q = 1$. Then for all $a, b \geq 0$,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Proof. Fix $a \geq 0$ and consider $F(b) = \frac{a^p}{p} + \frac{b^q}{q} - ab$ for $b \geq 0$. Then $F'(b) = b^{q-1} - a$ and $F''(b) = (q-1)b^{q-2} \geq 0$, so F is convex and minimized when $b^{q-1} = a$, i.e. $b = a^{1/(q-1)} = a^{p-1}$. At this point,

$$F(b_{\min}) = \frac{a^p}{p} + \frac{(a^{p-1})^q}{q} - a \cdot a^{p-1} = \frac{a^p}{p} + \frac{a^p}{q} - a^p = 0.$$

Hence $F(b) \geq 0$ for all $b \geq 0$, which is the claim. □

Proof of Minkowski's inequality

Proof. The case $p = \infty$ is immediate:

$$|X + Y| \leq |X| + |Y| \leq \|X\|_\infty + \|Y\|_\infty \quad \text{a.s.}$$

hence $\|X + Y\|_\infty \leq \|X\|_\infty + \|Y\|_\infty$.

Assume now $1 \leq p < \infty$. The case $p = 1$ is just $\mathbb{E}|X + Y| \leq \mathbb{E}|X| + \mathbb{E}|Y|$. So assume $p > 1$ and let q be the conjugate exponent, $1/p + 1/q = 1$.

Write

$$|X + Y|^p = |X + Y| \cdot |X + Y|^{p-1}.$$

Using $|X + Y| \leq |X| + |Y|$ and linearity,

$$\mathbb{E}|X + Y|^p \leq \mathbb{E}\big(|X|\,|X + Y|^{p-1}\big) + \mathbb{E}\big(|Y|\,|X + Y|^{p-1}\big).$$

Apply Hölder to each term with exponents (p, q):

$$\mathbb{E}\big(|X|\,|X + Y|^{p-1}\big) \leq \|X\|_p \,\big\||X + Y|^{p-1}\big\|_q,$$

and similarly with Y. Since $(p-1)q = p$, we have

$$\big\||X + Y|^{p-1}\big\|_q = \Big(\mathbb{E}\big(|X + Y|^{(p-1)q}\big)\Big)^{1/q} = \Big(\mathbb{E}|X + Y|^p\Big)^{1/q} = \|X + Y\|_p^{p/q}.$$

Therefore,

$$\mathbb{E}|X + Y|^p \leq (\|X\|_p + \|Y\|_p)\,\|X + Y\|_p^{p/q}.$$

If $\|X + Y\|_p = 0$ we are done. Otherwise divide by $\|X + Y\|_p^{p/q}$ to get

$$\|X + Y\|_p^{p - p/q} \leq \|X\|_p + \|Y\|_p.$$

But $p - p/q = p(1 - 1/q) = p(1/p) = 1$, hence $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$. □

B.6 Markov-type inequalities

Theorem B.10 (Markov's Inequality)**.** *If $X \geq 0,\ a > 0$, then*

$$\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a \tag{B.6}$$

Proof. Integrate $\mathbf{1}_{X \geq a} \leq X/a$. Then, the stated result follows by the monotonicity and linearity properties of the integral. □

This result can be generalized.

Theorem B.11 (Generalized Markov's Inequality)**.** *Suppose $\varphi : \mathbb{R} \to \mathbb{R}$ is non-negative. Let $A \subset \mathbb{R}$ and let $l_A = \inf\{\varphi(y) : y \in A\}$, that is, l_A is the lower bound on values of φ on A. Then,*

$$\mathbb{P}(X \in A) \leq \mathbb{E}\varphi(X)/l_A.$$

Proof. The definition of l_A and the fact that $\varphi \geq 0$ imply that

$$l_A 1_{(X \in A)} \leq \varphi(X) 1_{(X \in A)} \leq \varphi(X).$$

So taking expected values gives the required result. □

If we apply this to $\varphi(x) = x^2$ and $A = (-a, a)^c$, the we will get the usual Chebyshev inequality (without centering):

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}X^2}{a^2}.$$

We also have the following corollary of Theorem B.11.

Corollary B.12. *Let $\psi : \mathbb{R} \to \mathbb{R}^+$ be non-decreasing. Then*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\psi(X)}{\psi(a)} \tag{B.7}$$

Example B.13. If we take $\psi(x) = e^{tx}$, we find that for every $t \geq 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}e^{tX}}{e^{ta}} = e^{-ta} m_X(t), \tag{B.8}$$

where $m_X(t)$ is the moment generating function of X.

Here is an application.

Theorem B.14. *Suppose X_i are independent random variables with*

$$\mathbb{P}(X_i = 1) = P(X_i = -1) = \frac{1}{2},$$

and set $S_n = X_1 + \ldots X_n$. Then, for each $a > 0$,

$$\mathbb{P}(S_n \geq a) \leq e^{-\frac{a^2}{2n}}.$$

Proof. By (B.8),

$$\begin{aligned}\mathbb{P}(S_n \geq a) &\leq e^{-ta} m_{S_n}(t) = e^{-ta}(m_{X_1}(t))^n = e^{-ta}\Big(\frac{e^t + e^{-t}}{2}\Big)^n \\ &\leq e^{-ta+t^2 n/2}.\end{aligned}$$

The last inequality follows from the inequality $(e^t + e^{-t})/2 \leq e^{t^2/2}$ which can be obtained by expanding the exponentials in series and comparing the coefficients. Then, optimizing the inequality over t, we take $t = a/n$ and find that

$$\mathbb{P}(S_n \geq a) \leq e^{-\frac{a^2}{2n}},$$

as claimed. □

For example, if we use symmetry and take $a = tn$, then we find that $\mathbb{P}(|S_n| > tn) < 2e^{-\frac{t^2 n}{2}}$, which is much better than we could get from Chebyshev's inequality.

Marzienkewicz inequalities

Section 3.8 in Allan Gut's textbook

Appendix C

Conditional Expectations

C.1 Proof Sketch for Radon–Nikodym Theorem

Proof sketch for Radon–Nikodym theorem (finite-measure case). Assume $\mu(\Omega) < \infty$ and $\nu(\Omega) < \infty$.

1. For each rational $q \geq 0$, consider the signed measure $\nu - q\mu$ and take a Hahn decomposition $\Omega = P_q \sqcup N_q$ such that $(\nu - q\mu)(A) \geq 0$ for all $A \subset P_q$ and $(\nu - q\mu)(A) \leq 0$ for all $A \subset N_q$.

2. One checks the monotonicity $q_1 < q_2 \Rightarrow P_{q_2} \subset P_{q_1}$. Define

$$f(\omega) := \sup\{q \in \mathbb{Q}_{\geq 0} : \ \omega \in P_q\} \in [0, \infty].$$

3. Using the defining property of the Hahn decomposition and approximation by rationals, show that for all $A \in \mathcal{F}$,

$$\int_A f \, d\mu \leq \nu(A) \quad \text{and} \quad \nu(A) \leq \int_A f \, d\mu,$$

   hence equality.

4. Uniqueness follows: if $\int_A f \, d\mu = \int_A g \, d\mu$ for all A, then $f = g$ μ-a.e.

The σ-finite general case reduces to the finite case by restricting to sets of finite measure. □

C.2 Conditional Expectation: Proofs of Additional Results

Lemma 1 (Equivalent characterization)**.** *Let $X \in L^1(\mathcal{F}_0)$ and let $\mathcal{F} \subset \mathcal{F}_0$. A random variable Z is a version of $\mathbb{E}[X \mid \mathcal{F}]$ iff Z is $\mathcal{F}$-measurable, $Z \in L^1$, and*

$$\mathbb{E}[YZ] = \mathbb{E}[YX] \quad \textit{for every bounded } \mathcal{F}\textbf{-measurable } \textit{random variable } Y.$$

Equivalently, it suffices to require the identity for $Y = \mathbf{1}_A$ with $A \in \mathcal{F}$.

Proof. Assume first that $Z = \mathbb{E}[X \mid \mathcal{F}]$ in the sense of Definition 3.1. Let Y be bounded and $\mathcal{F}$-measurable. We first treat $Y \geq 0$ simple: $Y = \sum_{k=1}^{m} a_k \mathbf{1}_{A_k}$ with $a_k \geq 0$ and $A_k \in \mathcal{F}$. Then

$$\mathbb{E}[YZ] = \sum_{k=1}^{m} a_k \, \mathbb{E}[Z\mathbf{1}_{A_k}] = \sum_{k=1}^{m} a_k \, \mathbb{E}[X\mathbf{1}_{A_k}] = \mathbb{E}[YX].$$

For general bounded $Y \geq 0$, choose simple $Y_n \uparrow Y$ pointwise. Since Y is bounded, $0 \leq Y_n \leq \|Y\|_\infty$, and $Z, X \in L^1$, dominated convergence gives $\mathbb{E}[Y_n Z] \to \mathbb{E}[YZ]$ and $\mathbb{E}[Y_n X] \to \mathbb{E}[YX]$, hence $\mathbb{E}[YZ] = \mathbb{E}[YX]$. For an arbitrary bounded (signed) Y, write $Y = Y^+ - Y^-$ and apply the above to $Y^\pm$.

Conversely, suppose Z is $\mathcal{F}$-measurable, $Z \in L^1$, and $\mathbb{E}[YZ] = \mathbb{E}[YX]$ for all bounded $\mathcal{F}$-measurable Y. Taking $Y = \mathbf{1}_A$ for $A \in \mathcal{F}$ yields

$$\mathbb{E}[Z\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] \qquad \forall\, A \in \mathcal{F},$$

which is exactly the defining property of $\mathbb{E}[X \mid \mathcal{F}]$.

Finally, the last sentence follows since the class of bounded $\mathcal{F}$-measurable Y contains all indicators $\mathbf{1}_A$, and conversely the indicator condition implies the bounded-test-function condition by the approximation argument above. □

Theorem 4 (Conditional Jensen inequality)**.** *Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and $\mathcal{G} \subset \mathcal{F}_0$. Let X be an integrable real-valued random variable and let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be convex and Borel measurable. Assume either $\varphi(X) \geq 0$ a.s. or $\varphi(X) \in L^1$. Then*

$$\varphi\big(\mathbb{E}[X \mid \mathcal{G}]\big) \;\leq\; \mathbb{E}\big[\varphi(X) \mid \mathcal{G}\big] \qquad \textit{a.s.}$$

Proof. A standard fact from convex analysis is that a convex Borel function φ is the pointwise supremum of a family of affine functions dominated by φ: there exists a (possibly countable) collection $\{(a_i, b_i)\}_{i \in I} \subset \mathbb{R}^2$ such that

$$\varphi(t) = \sup_{i \in I} (a_i t + b_i), \qquad t \in \mathbb{R},$$

and $a_i t + b_i \leq \varphi(t)$ for all t and all i. Fix i. From $a_i X + b_i \leq \varphi(X)$ and monotonicity of conditional expectation,

$$a_i \, \mathbb{E}[X \mid \mathcal{G}] + b_i = \mathbb{E}[a_i X + b_i \mid \mathcal{G}] \leq \mathbb{E}[\varphi(X) \mid \mathcal{G}] \qquad \text{a.s.}$$

Taking the supremum over i and using the representation of φ gives

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) = \sup_i (a_i \, \mathbb{E}[X \mid \mathcal{G}] + b_i) \leq \mathbb{E}[\varphi(X) \mid \mathcal{G}] \qquad \text{a.s.}$$

□

C.3 Fubini-Tonelli Theorem

Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be σ-finite measure spaces. Set $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu_1 \otimes \mu_2)$.

Theorem C.1 (Fubini - Tonelli)**.** *Let $f : X \times Y \to [-\infty, \infty]$ be $\mathcal{A} \otimes \mathcal{B}$-measurable.*

(Tonelli) *If $f \geq 0$, then the functions $x \mapsto \int_Y f(x,y)\,\mu_2(dy)$ and $y \mapsto \int_X f(x,y)\,\mu_1(dx)$ are measurable (possibly $+\infty$) and*

$$\begin{aligned}\int_X \left(\int_Y f(x,y)\,\mu_2(dy) \right) \mu_1(dx) &= \int_{X\times Y} f\, d(\mu_1 \otimes \mu_2) \\ &= \int_Y \left(\int_X f(x,y)\,\mu_1(dx) \right) \mu_2(dy).\end{aligned}$$

(Fubini) *If $\int_{X\times Y} |f|\, d(\mu_1 \otimes \mu_2) < \infty$, then for μ_1-a.e. x and μ_2-a.e. y the inner integrals are finite, and the same equalities hold with all integrals finite.*

Appendix D

Independence and Tails

D.1 Hewitt–Savage 0–1 law

Exchangeability and the symmetric σ-field

Let $(X_n)_{n\geq 1}$ be S-valued random variables, where S is a Polish space (e.g. $S = \mathbb{R}$).

Definition D.1 (Exchangeable sequence)**.** The sequence $(X_n)_{n\geq 1}$ is **exchangeable** if for every $k \geq 1$ and every permutation π of $\{1, \dots, k\}$,

$$(X_1, \dots, X_k) \overset{d}{=} (X_{\pi(1)}, \dots, X_{\pi(k)}).$$

Definition D.2 (Symmetric (exchangeable) σ-field)**.** Let $\mathfrak{S}_{\text{fin}}$ be the group of permutations of $\mathbb{N}$ that move only finitely many indices. Define

$$\mathcal{I} := \big\{A \in \sigma(X_1, X_2, \dots) : \ \mathbf{1}_A(X_1, X_2, \dots) = \mathbf{1}_A(X_{\pi(1)}, X_{\pi(2)}, \dots) \text{ a.s.}\big\}.$$

for all $\pi \in \mathfrak{S}_{\text{fin}}$. Events in $\mathcal{I}$ are called **symmetric** (or **exchangeable**).

Direct proof of the Hewitt–Savage 0–1 law

We give a self-contained proof that does not rely on de Finetti's theorem.

Proof of the Hewitt–Savage 0–1 law (direct). Let $A \in \mathcal{I}$. We will show $\mathbb{P}(A) = \mathbb{P}(A)^2$.

Step 1 (Approximation by finitely many coordinates). Since $A \in \sigma(X_1, X_2, \dots) = \bigvee_{n\geq 1} \sigma(X_1, \dots, X_n)$, for every $\varepsilon > 0$ there exists $n \geq 1$ and an event $B \in \sigma(X_1, \dots, X_n)$ with

$$\mathbb{P}(A \triangle B) < \varepsilon.$$

(This follows from the fact that $\bigcup_{n\geq 1} \sigma(X_1, \dots, X_n)$ is an algebra that generates $\sigma(X_1, X_2, \dots)$; any event in the generated σ-algebra can be approximated in probability by events in the algebra.)

Step 2 (Shifting indices). Write $B = \{(X_1, \dots, X_n) \in C\}$ for some Borel set $C \subset \mathbb{R}^n$, and define

$$B' := \{(X_{n+1}, \dots, X_{2n}) \in C\} \in \sigma(X_{n+1}, \dots, X_{2n}).$$

Since the X_i are i.i.d., $\mathbb{P}(B') = \mathbb{P}(B)$.

Now consider the permutation π that swaps blocks:

$$\pi(i) := \begin{cases} i+n & \text{if } 1 \le i \le n, \\ i-n & \text{if } n+1 \le i \le 2n, \\ i & \text{if } i > 2n. \end{cases}$$

This is a finite permutation, and π maps B to B' (it replaces $(X_1, \dots, X_n)$ with $(X_{n+1}, \dots, X_{2n})$ in the defining condition). Since (X_i) are i.i.d. (hence exchangeable) and $A \in \mathcal{I}$, the event A is a.s. invariant under π. Moreover, because π merely permutes the i.i.d. sequence, the joint law is preserved:

$$(X_{\pi(1)}, X_{\pi(2)}, \dots) \overset{d}{=} (X_1, X_2, \dots).$$

Therefore

$$\mathbb{P}(A \triangle B') = \mathbb{P}\big(\pi^{-1}(A) \triangle \pi^{-1}(B')\big) = \mathbb{P}(A \triangle B) < \varepsilon,$$

where the first equality uses the distributional invariance and $\pi^{-1}(A) = A$ a.s., and $\pi^{-1}(B') = B$.

Step 3 (Independence and conclusion). Since $B \in \sigma(X_1, \dots, X_n)$ and $B' \in \sigma(X_{n+1}, \dots, X_{2n})$, independence gives $\mathbb{P}(B \cap B') = \mathbb{P}(B)\,\mathbb{P}(B')$. We estimate:

$$\begin{aligned} \big|\mathbb{P}(A) - \mathbb{P}(B)\big| &\le \mathbb{P}(A \triangle B) < \varepsilon, \\ \big|\mathbb{P}(A) - \mathbb{P}(B')\big| &\le \mathbb{P}(A \triangle B') < \varepsilon, \\ \big|\mathbb{P}(A) - \mathbb{P}(B \cap B')\big| &\le \mathbb{P}(A \triangle (B \cap B')) \le \mathbb{P}(A \triangle B) + \mathbb{P}(A \triangle B') < 2\varepsilon, \end{aligned}$$

where the last line uses $A \triangle (B \cap B') \subset (A \triangle B) \cup (A \triangle B')$ (since $\omega \in A$ iff $\omega \in B \cap B'$ whenever ω belongs to both B and B' and to A; more precisely, $A = (A \cap B \cap B') \cup (A \setminus (B \cap B'))$ and the symmetric difference is controlled termwise).

Combining:

$$\big|\mathbb{P}(A) - \mathbb{P}(A)^2\big| \le \big|\mathbb{P}(A) - \mathbb{P}(B \cap B')\big| + \big|\mathbb{P}(B)\,\mathbb{P}(B') - \mathbb{P}(A)^2\big| < 2\varepsilon + 2\varepsilon = 4\varepsilon.$$

(For the second term: $|\mathbb{P}(B)\,\mathbb{P}(B') - \mathbb{P}(A)^2| \le |\mathbb{P}(B) - \mathbb{P}(A)|\,\mathbb{P}(B') + \mathbb{P}(A)\,|\mathbb{P}(B') - \mathbb{P}(A)| < \varepsilon + \varepsilon = 2\varepsilon$.)

Since $\varepsilon > 0$ was arbitrary, $\mathbb{P}(A) = \mathbb{P}(A)^2$, so $\mathbb{P}(A) \in \{0, 1\}$. □

Proof of Hewitt-Savage via de Finetti (representation + identification of $\mathcal{I}$)

Theorem D.3 (de Finetti, standard form)**.** *If $(X_n)_{n \ge 1}$ is exchangeable, then there exists a **random** probability measure M on S such that, conditional on M, the variables (X_n) are i.i.d. with common law M. Equivalently, for every $k \ge 1$ and Borel sets $B_1, \dots, B_k \subset S$,*

$$\mathbb{P}(X_1 \in B_1, \dots, X_k \in B_k) = \mathbb{E}\Big[\prod_{j=1}^{k} M(B_j)\Big].$$

Remark D.4 (A concrete way to define M). One can (and typically does) choose M to be $\mathcal{I}$-measurable by setting, for Borel $B \subset S$,

$$M(B) \;:=\; \mathbb{P}(X_1 \in B \mid \mathcal{I}) \quad \text{(a version of conditional probability)}.$$

Then M is a random probability measure and the conditional i.i.d. statement holds: given $\mathcal{I}$ (hence given M), the coordinates are i.i.d. with law M.

Remark D.5 (Key structural fact). A standard strengthening of de Finetti says that, up to $\mathbb{P}$-null sets,

$$\mathcal{I} \;=\; \sigma(M) \qquad \text{(completion understood)}.$$

Informally: **the only source of randomness left after quotienting by finite-permutation symmetry is the directing measure M.** See, e.g., Durrett or Kallenberg for a full proof.

Hewitt–Savage for i.i.d. as a corollary

Theorem 5 (Hewitt–Savage 0–1 law)**.** *If $(X_n)_{n\geq 1}$ are i.i.d., then $\mathcal{I}$ is trivial: for every $A \in \mathcal{I}$,*

$$\mathbb{P}(A) \in \{0, 1\}.$$

Proof using de Finetti + LLN. Let μ be the (deterministic) law of X_1. Since i.i.d. implies exchangeable, we may apply de Finetti and obtain a directing random measure M.

Step 1: $M = \mu$ a.s. Fix a bounded continuous function $f : S \to \mathbb{R}$. Let

$$A_n(f) := \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

By the (ordinary) strong law of large numbers for i.i.d. variables with law μ,

$$A_n(f) \;\xrightarrow[n\to\infty]{}\; \int f \, d\mu \qquad \text{a.s.}$$

On the other hand, conditional on M the variables are i.i.d. with law M, hence by the conditional strong law,

$$A_n(f) \;\xrightarrow[n\to\infty]{}\; \int f \, dM \qquad \text{a.s.}$$

Therefore $\int f \, dM = \int f \, d\mu$ a.s. for every bounded continuous f. Choosing a **countable** convergence-determining class of such f (e.g. bounded Lipschitz functions with rational parameters), we conclude that $M = \mu$ a.s.

Step 2: conclude triviality of $\mathcal{I}$. By the structural fact $\mathcal{I} = \sigma(M)$ (up to null sets) and Step 1, we have

$$\mathcal{I} = \sigma(\mu),$$

which is the trivial σ-field (since μ is deterministic). Hence every $A \in \mathcal{I}$ has probability 0 or 1. □

Remark D.6. This proof makes clear why exchangeability alone does **not** imply a 0–1 law: for a general exchangeable sequence the directing measure M is random, and $\mathcal{I} = \sigma(M)$ is typically nontrivial.

Appendix E

Convergence Zoo and Uniform Integrability

E.1 Portmanteau inequalities and continuity sets

Throughout this appendix, (S, d) is a metric space with its Borel σ-field. We use the definition

$$\mu_n \Rightarrow \mu \quad \Longleftrightarrow \quad \int f\, d\mu_n \to \int f\, d\mu \ \text{ for every } f \in C_b(S).$$

Theorem E.1 (Portmanteau (inequalities))**.** *Assume $\mu_n \Rightarrow \mu$.*

1. *For every closed set $F \subset S$,*

$$\limsup_{n\to\infty} \mu_n(F) \le \mu(F).$$

2. *For every open set $G \subset S$,*

$$\liminf_{n\to\infty} \mu_n(G) \ge \mu(G).$$

Proof. For a set $A \subset S$, write $d(x, A) := \inf_{y\in A} d(x, y)$.

(1) Closed sets. Fix a closed $F \subset S$. For $m \ge 1$ define

$$f_m(x) := \big(1 - m\, d(x, F)\big)_+ \;=\; \max\{0,\, 1 - m\, d(x, F)\}.$$

Then $f_m \in C_b(S)$, $0 \le f_m \le 1$, and $f_m \downarrow \mathbf{1}_F$ pointwise:

- if $x \in F$, then $d(x, F) = 0$ so $f_m(x) = 1$ for all m;
- if $x \notin F$, then $d(x, F) > 0$ so for $m > 1/d(x, F)$ we have $f_m(x) = 0$.

Also $\mathbf{1}_F \le f_m$, hence $\mu_n(F) \le \int f_m\, d\mu_n$. Therefore,

$$\limsup_{n\to\infty} \mu_n(F) \le \limsup_{n\to\infty} \int f_m\, d\mu_n = \int f_m\, d\mu,$$

using $\mu_n \Rightarrow \mu$ and $f_m \in C_b(S)$. Finally, by monotone convergence (since $f_m \downarrow \mathbf{1}_F$ and $0 \le f_m \le 1$),

$$\int f_m \, d\mu \downarrow \int \mathbf{1}_F \, d\mu = \mu(F).$$

Letting $m \to \infty$ yields $\limsup_n \mu_n(F) \le \mu(F)$.

(2) Open sets. Fix an open $G \subset S$ and set $F := G^c$, which is closed. For $m \ge 1$ define

$$g_m(x) := \min\{1,\, m\, d(x, F)\}.$$

Then $g_m \in C_b(S)$, $0 \le g_m \le 1$, and $g_m \uparrow \mathbf{1}_G$ pointwise:

- if $x \in G$, then $d(x, F) > 0$, so for m large enough $m\, d(x, F) \ge 1$ and $g_m(x) = 1$;
- if $x \in F$, then $d(x, F) = 0$ and $g_m(x) = 0$ for all m.

Also $g_m \le \mathbf{1}_G$, hence $\int g_m \, d\mu_n \le \mu_n(G)$. Therefore,

$$\liminf_{n\to\infty} \mu_n(G) \ge \liminf_{n\to\infty} \int g_m \, d\mu_n = \int g_m \, d\mu.$$

Letting $m \to \infty$ and using monotone convergence gives

$$\int g_m \, d\mu \uparrow \mu(G),$$

hence $\liminf_n \mu_n(G) \ge \mu(G)$. □

Corollary E.2 (Continuity sets)**.** *If $\mu_n \Rightarrow \mu$ and A is Borel with $\mu(\partial A) = 0$, then*

$$\mu_n(A) \to \mu(A).$$

Equivalently, if $X_n \Rightarrow X$ and $\mathbb{P}(X \in \partial A) = 0$, then $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$.

Proof. Apply Theorem E.1 to $F = \overline{A}$ and $G = A^\circ$:

$$\mu(A^\circ) \le \liminf_n \mu_n(A) \le \limsup_n \mu_n(A) \le \mu(\overline{A}).$$

But $\overline{A} \setminus A^\circ \subset \partial A$, so $\mu(\overline{A}) = \mu(A^\circ)$ when $\mu(\partial A) = 0$. Hence $\mu_n(A) \to \mu(A)$. □

Lemma E.3 (Reverse Portmanteau: inequalities ⇒ weak convergence)**.** *Let (S, d) be a metric space and let ν_n, ν be probability measures on S. Assume that for every closed set $F \subset S$,*

$$\limsup_{n\to\infty} \nu_n(F) \le \nu(F). \tag{$*$}$$

Then $\nu_n \Rightarrow \nu$, i.e.

$$\int f \, d\nu_n \to \int f \, d\nu \qquad \text{for all } f \in C_b(S).$$

Proof. From $(*)$ we also get the open-set inequality: for open G let $F = G^c$ (closed) and use $\liminf(1 - a_n) = 1 - \limsup a_n$:

$$\liminf_{n\to\infty} \nu_n(G) = \liminf_{n\to\infty} \big(1 - \nu_n(F)\big) = 1 - \limsup_{n\to\infty} \nu_n(F) \geq 1 - \nu(F) = \nu(G). \tag{$**$}$$

Step 1: continuity sets. For any Borel A,

$$\nu(A^\circ) \leq \liminf_n \nu_n(A) \leq \limsup_n \nu_n(A) \leq \nu(\overline{A})$$

(using $(**)$ for A° and $(*)$ for $\overline{A}$). Hence if $\nu(\partial A) = 0$ then $\nu_n(A) \to \nu(A)$.

Step 2: convergence of integrals for nonnegative bounded continuous f. Let $f \in C_b(S)$ with $0 \leq f \leq M$. For $t \in \mathbb{R}$ set $A_t := \{x : f(x) > t\}$, which is open. Moreover, $\partial A_t \subset \{x : f(x) = t\}$. The set of t with $\nu(f = t) > 0$ is at most countable (the events $\{f = t\}$ are disjoint), so for all but countably many t we have $\nu(\partial A_t) = 0$, hence from Step 1:

$$\nu_n(A_t) \to \nu(A_t) \qquad \text{for a.e. } t \in [0, M].$$

Using the layer-cake formula (valid for $0 \leq f \leq M$),

$$\int f \, d\nu_n = \int_0^M \nu_n(f > t)\, dt, \qquad \int f \, d\nu = \int_0^M \nu(f > t)\, dt,$$

and dominated convergence (since $0 \leq \nu_n(f > t) \leq 1$), we get $\int f \, d\nu_n \to \int f \, d\nu$.

Step 3: general $f \in C_b(S)$. Write $f = f^+ - f^-$ and apply Step 2 to $f^\pm$ (both bounded, continuous, nonnegative). □

Supporting lemmas

Lemma E.4 (Only countably many level sets can have positive mass)**.** *Let $(S, \mathcal{B}, \nu)$ be a probability space and let $f : S \to \mathbb{R}$ be measurable. Then the set*

$$T := \{t \in \mathbb{R} : \ \nu(f = t) > 0\}$$

is at most countable.

Proof. For $t \in \mathbb{R}$ set $E_t := \{x \in S : \ f(x) = t\}$, so that the sets $(E_t)_{t\in\mathbb{R}}$ are pairwise disjoint. Let $a_t := \nu(E_t)$.

For each $k \in \mathbb{N}$ define

$$I_k := \{t \in \mathbb{R} : \ a_t \geq 2^{-k}\}.$$

We claim I_k is finite. Indeed, if $t_1, \ldots, t_N \in I_k$ are distinct, then disjointness gives

$$1 \ \geq \ \nu\Big(\bigcup_{j=1}^N E_{t_j}\Big) \ = \ \sum_{j=1}^N \nu(E_{t_j}) \ \geq \ N \cdot 2^{-k},$$

hence $N \le 2^k$. So $|I_k| \le 2^k$.

Finally, if $a_t > 0$, choose k such that $2^{-k} \le a_t$; then $t \in I_k$. Thus

$$T \subset \bigcup_{k=1}^{\infty} I_k,$$

a countable union of finite sets, hence T is countable. □

Lemma E.5 (Boundary of a strict superlevel set)**.** *Let (S, d) be a metric space and let $f : S \to \mathbb{R}$ be continuous. For any $t \in \mathbb{R}$, set*

$$A_t := \{x \in S : \ f(x) > t\}.$$

Then

$$\partial A_t \subset \{x \in S : \ f(x) = t\}.$$

Proof. Fix t and let $x \in \partial A_t$. Then there exist sequences $x_n \in A_t$ and $y_n \notin A_t$ such that $x_n \to x$ and $y_n \to x$. Hence

$$f(x_n) > t, \qquad f(y_n) \le t.$$

By continuity of f at x,

$$f(x) = \lim_{n\to\infty} f(x_n) \ge t, \qquad f(x) = \lim_{n\to\infty} f(y_n) \le t,$$

so $f(x) = t$. □

Corollary E.6 (For a.e. level, $\nu(\partial A_t) = 0$)**.** *Let (S, d) be a metric space, let ν be a Borel probability measure on S, and let $f \in C_b(S)$ with $0 \le f \le M$. Then for all $t \in [0, M]$ except at most countably many,*

$$\nu(\partial A_t) = 0, \qquad A_t = \{f > t\}.$$

Proof. By Lemma E.5, $\partial A_t \subset \{f = t\}$, hence

$$\nu(\partial A_t) \le \nu(f = t).$$

By Lemma E.4, the set of t with $\nu(f = t) > 0$ is countable, so for all other t we have $\nu(\partial A_t) = 0$. □

E.2 Proofs of the general CMT

Theorem E.7 (Continuous Mapping Theorem (general form))**.** *Let S, T be metric spaces, and let X_n, X be S-valued random elements such that $X_n \Rightarrow X$. Let $f : S \to T$ be Borel measurable and assume*

$$\mathbb{P}\big(X \in \mathrm{Disc}(f)\big) = 0,$$

where $\mathrm{Disc}(f)$ *is the set of discontinuity points of f. Then*

$$f(X_n) \Rightarrow f(X).$$

Proof. Let $\mu_n = \mathcal{L}(X_n)$ and $\mu = \mathcal{L}(X)$. Fix a closed set $C \subset T$ and set $B := f^{-1}(C) \subset S$. Then

$$\mathbb{P}\big(f(X_n) \in C\big) = \mu_n(B).$$

By Portmanteau (closed-set inequality),

$$\limsup_{n\to\infty} \mu_n(B) \le \mu(\overline{B}).$$

We claim that $\overline{B} \subset B \cup \mathrm{Disc}(f)$. Indeed, if $x \in \overline{B}$ and $x \notin \mathrm{Disc}(f)$, pick $x_k \in B$ with $x_k \to x$; continuity of f at x gives $f(x_k) \to f(x)$, and since $f(x_k) \in C$ and C is closed, we get $f(x) \in C$, i.e. $x \in B$.

Therefore,

$$\mu(\overline{B}) \le \mu(B) + \mu(\mathrm{Disc}(f)) = \mathbb{P}\big(f(X) \in C\big) + \mathbb{P}\big(X \in \mathrm{Disc}(f)\big) = \mathbb{P}\big(f(X) \in C\big).$$

Combining,

$$\limsup_{n\to\infty} \mathbb{P}\big(f(X_n) \in C\big) \le \mathbb{P}\big(f(X) \in C\big) \qquad \text{for every closed } C \subset T.$$

By reverse Portmanteau, this implies $f(X_n) \Rightarrow f(X)$. □

Theorem E.8 (Continuous Mapping Theorem for convergence in probability (general form))**.** *Let S, T be metric spaces, and let X_n, X be S-valued random elements such that $X_n \xrightarrow{\mathbb{P}} X$. Let $f : S \to T$ be Borel measurable and assume*

$$\mathbb{P}\big(X \in \mathrm{Disc}(f)\big) = 0.$$

Then

$$f(X_n) \xrightarrow{\mathbb{P}} f(X).$$

Proof. Fix $\varepsilon > 0$ and $\eta > 0$. Fix a point $x_0 \in S$.

Step 1: localize. Choose $R > 0$ such that

$$\mathbb{P}\big(d_S(X, x_0) > R\big) < \eta.$$

Step 2: discard the bad set. For each $m \ge 1$, define the set of points in $\overline{B}_S(x_0, R)$ where f oscillates by at least ε at scale $1/m$:

$$A_m := \Big\{ x \in \overline{B}_S(x_0, R) : \sup_{\substack{y \in S: \\ d_S(x,y) < 1/m}} d_T(f(x), f(y)) \ge \varepsilon \Big\}.$$

If $x \in \overline{B}_S(x_0, R)$ and $x \notin \mathrm{Disc}(f)$, then f is continuous at x, so $x \notin A_m$ for all sufficiently large m. Therefore $A_m \setminus \mathrm{Disc}(f) \downarrow \emptyset$ as $m \to \infty$, and since $\mathbb{P}(X \in \mathrm{Disc}(f)) = 0$,

$$\begin{aligned} \mathbb{P}(X \in A_m) &= \mathbb{P}(X \in A_m \setminus \mathrm{Disc}(f)) + \mathbb{P}(X \in A_m \cap \mathrm{Disc}(f)) \\ &\le \mathbb{P}(X \in A_m \setminus \mathrm{Disc}(f)) + 0 \downarrow 0. \end{aligned}$$

Choose m large enough so that $\mathbb{P}(X \in A_m) < \eta$, and set $\delta := 1/m$. By definition of A_m, for every $x \in \overline{B}_S(x_0, R)\setminus A_m$ and every $y \in S$ with $d_S(x,y) < \delta$:

$$d_T(f(x), f(y)) < \varepsilon. \tag{E.1}$$

Step 3: conclude. On the event $\{d_S(X, x_0) \le R\}\cap\{X \notin A_m\}\cap\{d_S(X_n, X) < \delta\}$, (E.1) applies with $x = X(\omega)$ and $y = X_n(\omega)$, giving $d_T(f(X_n), f(X)) < \varepsilon$. Therefore

$$\begin{aligned}\mathbb{P}\big(d_T(f(X_n), f(X)) > \varepsilon\big) &\le \mathbb{P}\big(d_S(X, x_0) > R\big) + \mathbb{P}(X \in A_m) + \mathbb{P}\big(d_S(X_n, X) \ge \delta\big)\\ &< \eta + \eta + \mathbb{P}\big(d_S(X_n, X) \ge \delta\big).\end{aligned}$$

Since $X_n \xrightarrow{\mathbb{P}} X$, we have $\mathbb{P}(d_S(X_n, X) \ge \delta) \to 0$. Taking $\limsup_{n\to\infty}$:

$$\limsup_{n\to\infty} \mathbb{P}\big(d_T(f(X_n), f(X)) > \varepsilon\big) \le 2\eta.$$

Since $\eta > 0$ is arbitrary, $\mathbb{P}(d_T(f(X_n), f(X)) > \varepsilon) \to 0$. □

Remark E.9. Compare with the proof for the continuous case (Theorem 5.13 in the main text): there, continuity of f on the **entire** compact set $\overline{B}_S(x_0, R+1)$ gives a single δ by uniform continuity, free of charge. Here, f may be discontinuous on this set, so we must first discard the "bad" set A_m where the continuity modulus is too coarse. The key point is that $\mathbb{P}(X \in A_m) \to 0$ because $\mathbb{P}(X \in \mathrm{Disc}(f)) = 0$.

E.3 Proof of Slutsky's theorem

Theorem E.10 (Slutsky)**.** *Let X_n be $\mathbb{R}^d$-valued and Y_n be $\mathbb{R}^m$-valued random vectors. Assume that*

$$X_n \Rightarrow X \qquad \text{and} \qquad Y_n \xrightarrow{\mathbb{P}} c \in \mathbb{R}^m.$$

Then

$$(X_n, Y_n) \Rightarrow (X, c).$$

Consequently, for every continuous map $h : \mathbb{R}^{d+m} \to \mathbb{R}^k$,

$$h(X_n, Y_n) \Rightarrow h(X, c).$$

Proof. Fix $\Phi \in C_b(\mathbb{R}^{d+m})$ and write $M := \|\Phi\|_\infty$. We will show

$$\mathbb{E}[\Phi(X_n, Y_n)] \to \mathbb{E}[\Phi(X, c)],$$

which implies $(X_n, Y_n) \Rightarrow (X, c)$ by the definition of weak convergence.

Step 1: a tightness estimate for (X_n). Choose $R > 0$ such that $\mathbb{P}(\|X\| > R) < \varepsilon$, where $\|\cdot\|$ is the Euclidean norm. Let $\psi_R : \mathbb{R}^d \to [0,1]$ be a continuous function such that

$$\psi_R(x) = 0 \text{ if } \|x\| \le R, \qquad \psi_R(x) = 1 \text{ if } \|x\| \ge R+1.$$

Then $\mathbf{1}_{\{\|x\|\geq R+1\}} \leq \psi_R(x) \leq \mathbf{1}_{\{\|x\|>R\}}$, hence

$$\mathbb{P}(\|X_n\| \geq R+1) \leq \mathbb{E}[\psi_R(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[\psi_R(X)] \leq \mathbb{P}(\|X\| > R) < \varepsilon.$$

In particular, for n large enough,

$$\mathbb{P}(\|X_n\| \geq R+1) < 2\varepsilon. \tag{E.2}$$

Step 2: uniform continuity on a compact set. Consider the compact set

$$K := \overline{B}_{R+1}(0) \times \overline{B}_1(c) \ \subset \ \mathbb{R}^{d+m}.$$

Since Φ is continuous, it is uniformly continuous on K. Hence there exists $\delta \in (0,1)$ such that

$$\|x\| \leq R+1, \ \|y-c\| \leq \delta \quad \Longrightarrow \quad |\Phi(x,y) - \Phi(x,c)| < \varepsilon. \tag{E.3}$$

Step 3: compare $\Phi(X_n, Y_n)$ to $\Phi(X_n, c)$. Write

$$\mathbb{E}[\Phi(X_n,Y_n)] - \mathbb{E}[\Phi(X,c)] = \Big(\mathbb{E}[\Phi(X_n,Y_n)] - \mathbb{E}[\Phi(X_n,c)]\Big) + \Big(\mathbb{E}[\Phi(X_n,c)] - \mathbb{E}[\Phi(X,c)]\Big).$$

For the second term, the map $x \mapsto \Phi(x,c)$ is bounded and continuous on $\mathbb{R}^d$, so $X_n \Rightarrow X$ implies

$$\mathbb{E}[\Phi(X_n,c)] \to \mathbb{E}[\Phi(X,c)]. \tag{E.4}$$

For the first term, split the expectation according to the events $\{\|X_n\| \leq R+1\}$ and $\{\|Y_n - c\| \leq \delta\}$:

$$\begin{aligned}\big|\mathbb{E}[\Phi(X_n,Y_n)] - \mathbb{E}[\Phi(X_n,c)]\big| &\leq \mathbb{E}\big[|\Phi(X_n,Y_n) - \Phi(X_n,c)|\, \mathbf{1}_{\{\|X_n\|\leq R+1, \ \|Y_n-c\|\leq\delta\}}\big] \\ &\quad + 2M\,\mathbb{P}(\|X_n\| > R+1) + 2M\,\mathbb{P}(\|Y_n - c\| > \delta).\end{aligned}$$

By (E.3), the first expectation is at most ε. Using (E.2) and $Y_n \xrightarrow{\mathbb{P}} c$ (so $\mathbb{P}(\|Y_n - c\| > \delta) \to 0$), we get for n large enough

$$\big|\mathbb{E}[\Phi(X_n,Y_n)] - \mathbb{E}[\Phi(X_n,c)]\big| \leq \varepsilon + 4M\varepsilon + o(1). \tag{E.5}$$

Combining (E.4) and (E.5), we obtain

$$\limsup_{n\to\infty} \big|\mathbb{E}[\Phi(X_n,Y_n)] - \mathbb{E}[\Phi(X,c)]\big| \leq (1+4M)\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $\mathbb{E}[\Phi(X_n,Y_n)] \to \mathbb{E}[\Phi(X,c)]$ for all $\Phi \in C_b(\mathbb{R}^{d+m})$, hence $(X_n,Y_n) \Rightarrow (X,c)$.

Finally, if $h : \mathbb{R}^{d+m} \to \mathbb{R}^k$ is continuous, then for any $g \in C_b(\mathbb{R}^k)$, $g \circ h \in C_b(\mathbb{R}^{d+m})$, so

$$\mathbb{E}[g(h(X_n,Y_n))] \to \mathbb{E}[g(h(X,c))],$$

i.e. $h(X_n,Y_n) \Rightarrow h(X,c)$. □

Appendix F

SLLN

F.1 SLLN with finite fourth moment

It is not difficult to prove SLLN for i.i.d random variables if the existence of the finite fourth moment is assumed.

Theorem F.1. *If X_1, ..., X_n, ..., is a sequence of independent identically distributed random variables with $\mathbb{E}|X_i|^4 = C < \infty$, then*

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = \mathbb{E}(X_1)$$

with probability 1.

Proof. We can assume without loss of generality that $\mathbb{E}[X_i] = 0$. Otherwise, just take $Y_i = X_i - \mathbb{E}[X_i]$.

A simple calculation shows

$$\mathbb{E}[(S_n)^4] = n\mathbb{E}[(X_1)^4] + 3n(n-1)\mathbb{E}[(X_1)^2]^2 \leq nC + 3n^2\sigma^4,$$

and by applying a Chebychev type inequality using fourth moments,

$$\mathbb{P}[\frac{|S_n|}{n} \geq \delta] = \mathbb{P}[|S_n|^4 \geq (n\delta)^4] \leq \frac{nC + 3n^2\sigma^4}{n^4\delta^4}.$$

Hence,

$$\sum_{n=1}^{\infty} \mathbb{P}\left[\frac{|S_n|}{n} \geq \delta\right] < \infty,$$

and we can now apply the Borel-Cantelli Lemma (BC I) to conclude that these events will happen only finitely many times. □

F.2 Cesàro's and Kronecker's lemmas

Lemma F.2 (Cesàro's Lemma)**.** *Let $(a_n)_{n\geq 1}$ be a sequence of real numbers. If $a_n \to L$ as $n \to \infty$, then*

$$\frac{1}{n}\sum_{k=1}^{n} a_k \to L.$$

Proof. Let $\varepsilon > 0$. Since $a_n \to L$, there exists $N \in \mathbb{N}$ such that $|a_k - L| < \varepsilon$ for all $k > N$. For $n > N$, we split the sum:

$$\left|\frac{1}{n}\sum_{k=1}^{n} a_k - L\right| = \left|\frac{1}{n}\sum_{k=1}^{n}(a_k - L)\right| \leq \frac{1}{n}\sum_{k=1}^{N}|a_k - L| + \frac{1}{n}\sum_{k=N+1}^{n}|a_k - L|.$$

Let $C_N = \sum_{k=1}^{N}|a_k - L|$, a fixed finite constant. The first term satisfies $C_N/n \to 0$ as $n \to \infty$. For the second term, since $|a_k - L| < \varepsilon$ for all $k > N$:

$$\frac{1}{n}\sum_{k=N+1}^{n}|a_k - L| < \frac{(n-N)\varepsilon}{n} < \varepsilon.$$

Choosing n large enough that $C_N/n < \varepsilon$, we obtain

$$\left|\frac{1}{n}\sum_{k=1}^{n} a_k - L\right| < 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the result follows. □

Lemma F.3 (Kronecker)**.** *If $\sum_{j=1}^{\infty}\frac{y_j}{j}$ converges, then*

$$\frac{1}{n}\sum_{j=1}^{n} y_j \to 0.$$

Proof. Let $S_n = \sum_{j=1}^{n}\frac{y_j}{j}$ for $n \geq 1$ and $S_0 = 0$. Summation by parts gives

$$\sum_{j=1}^{n} y_j = \sum_{j=1}^{n} j(S_j - S_{j-1}) = nS_n - \sum_{j=1}^{n} S_{j-1}.$$

Hence,

$$\frac{1}{n}\sum_{j=1}^{n} y_j = S_n - \frac{1}{n}\sum_{j=1}^{n} S_{j-1}.$$

By assumption, $S_n \to x$ for some $x \in \mathbb{R}$. By Cesàro's lemma (Lemma F.2), $\frac{1}{n}\sum_{j=1}^{n} S_{j-1} \to x$. Therefore,

$$\frac{1}{n}\sum_{j=1}^{n} y_j \to x - x = 0.$$ □

F.3 Kolmogorov's Maximal Inequality

Theorem F.4 (Kolmogorov's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{V}\mathrm{ar}(X_i) = \sigma_i^2 < \infty$. Let $S_k = \sum_{j=1}^{k} X_j$, $s_n^2 = \sum_{j=1}^{n}\sigma_j^2$, and*

$$T_n(\omega) = \max_{1\leq k\leq n}|S_k(\omega)|.$$

Then for any $\ell > 0$,

$$\mathbb{P}(T_n \geq \ell) \leq \frac{s_n^2}{\ell^2}.$$

The important point here is that the estimate depends only on s_n^2 and not on the number of summands. The Chebyshev bound on S_n is

$$\mathbb{P}(|S_n| \geq \ell) \leq \frac{s_n^2}{\ell^2},$$

and therefore taking the maximum over k costs nothing.

Proof. Define the disjoint events

$$E_k = \{|S_1| < \ell, \ldots, |S_{k-1}| < \ell, |S_k| \geq \ell\}, \quad k = 1, \ldots, n.$$

Then $\{T_n \geq \ell\} = \bigcup_{k=1}^n E_k$ is a disjoint union. On E_k we have $|S_k| \geq \ell$, so $S_k^2 \geq \ell^2$, and thus

$$\mathbb{P}(E_k) \leq \frac{1}{\ell^2} \int_{E_k} S_k^2 \, d\mathbb{P}.$$

Key observation: We claim that

$$\int_{E_k} S_k^2 \, d\mathbb{P} \leq \int_{E_k} S_n^2 \, d\mathbb{P}.$$

To see this, write $S_n = S_k + (S_n - S_k)$ and expand:

$$S_n^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2.$$

Since $(S_n - S_k)^2 \geq 0$, it suffices to show that $\int_{E_k} S_k(S_n - S_k) \, d\mathbb{P} = 0$.

Claim: $\displaystyle\int_{E_k} S_k(S_n - S_k) \, d\mathbb{P} = 0.$

Proof of Claim: The event E_k depends only on $(X_1, \ldots, X_k)$, as does S_k. Therefore $S_k \mathbf{1}_{E_k}$ is measurable with respect to $\sigma(X_1, \ldots, X_k)$. On the other hand, $S_n - S_k = X_{k+1} + \cdots + X_n$ is measurable with respect to $\sigma(X_{k+1}, \ldots, X_n)$.

By independence of $(X_1, \ldots, X_k)$ and $(X_{k+1}, \ldots, X_n)$, the random variables $S_k \mathbf{1}_{E_k}$ and $S_n - S_k$ are independent. Therefore:

$$\int_{E_k} S_k(S_n - S_k) \, d\mathbb{P} = \mathbb{E}[S_k \mathbf{1}_{E_k}] \cdot \mathbb{E}[S_n - S_k] = \mathbb{E}[S_k \mathbf{1}_{E_k}] \cdot 0 = 0,$$

where we used $\mathbb{E}[S_n - S_k] = \sum_{j=k+1}^n \mathbb{E}[X_j] = 0$. ◇

Thus $\int_{E_k} S_k^2 \, d\mathbb{P} \leq \int_{E_k} S_n^2 \, d\mathbb{P}$, and summing over k:

$$\mathbb{P}(T_n \geq \ell) = \sum_{k=1}^n \mathbb{P}(E_k) \leq \frac{1}{\ell^2} \sum_{k=1}^n \int_{E_k} S_n^2 \, d\mathbb{P} = \frac{1}{\ell^2} \int_{T_n \geq \ell} S_n^2 \, d\mathbb{P} \leq \frac{\mathbb{E}[S_n^2]}{\ell^2} = \frac{s_n^2}{\ell^2}.$$

□

Remark F.5 (On the independence assumption). The proof requires **mutual independence**, not merely pairwise independence. The key step is showing that $S_k \mathbf{1}_{E_k}$ and $S_n - S_k$ are independent, which allows us to factor the expectation $\mathbb{E}[S_k \mathbf{1}_{E_k}(S_n - S_k)]$. This factorization relies on the σ-fields $\sigma(X_1, \ldots, X_k)$ and $\sigma(X_{k+1}, \ldots, X_n)$ being independent, which follows from mutual independence but not from pairwise independence alone.

To see why pairwise independence is insufficient, consider the following example. Let X_1, X_2 be independent Rademacher random variables (taking values ± 1 with probability $1/2$ each), and let $X_3 = X_1 X_2$. Then any two of X_1, X_2, X_3 are independent, but X_3 is not independent of $\sigma(X_1, X_2)$. In general, pairwise independence does not imply that functions of disjoint groups of variables are independent.

F.4 Kolmogorov–Khinchin Convergence Theorem

The following theorem is the key tool that converts variance summability into almost sure convergence of a series. It powers both the variance summability SLLN (Section 6.2) and the sufficiency direction of the Three-Series Theorem (Section F.5).

Theorem F.6 (Kolmogorov–Khinchin Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be independent random variables with $\mathbb{E}[X_n] = 0$ for all n. If*

$$\sum_{n=1}^{\infty} \mathbb{V}\mathrm{ar}(X_n) < \infty,$$

then $\sum_{n=1}^{\infty} X_n$ converges almost surely.

Remark F.7. The partial sums $S_n = \sum_{k=1}^{n} X_k$ automatically converge in L^2, since

$$\mathbb{E}[(S_n - S_m)^2] = \sum_{k=m+1}^{n} \mathbb{V}\mathrm{ar}(X_k) \to 0 \quad \text{as } m, n \to \infty.$$

However, L^2 convergence does not imply almost sure convergence in general (recall the typewriter sequence from Chapter 5). The content of the theorem is the upgrade from L^2 to a.s., which is achieved by Kolmogorov's maximal inequality.

Proof. Let $S_n = \sum_{k=1}^{n} X_k$. Since $\mathbb{R}$ is complete, S_n converges if and only if it is Cauchy. We use Kolmogorov's maximal inequality (Theorem F.4) to show that the partial sums are almost surely Cauchy.

Step 1: Bounding the tail oscillation.

For $m < n$, the random variables $X_{m+1}, \ldots, X_n$ are independent with mean zero. By Kolmogorov's maximal inequality applied to this block,

$$\mathbb{P}\left(\max_{m<k\leq n} |S_k - S_m| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \sum_{k=m+1}^{n} \mathbb{V}\mathrm{ar}(X_k).$$

Step 2: Taking $n \to \infty$.

The events $\{\max_{m<k\leq n} |S_k - S_m| > \varepsilon\}$ are increasing in n, so by continuity of probability,

$$\mathbb{P}\left(\sup_{k>m} |S_k - S_m| > \varepsilon \right) = \lim_{n\to\infty} \mathbb{P}\left(\max_{m<k\leq n} |S_k - S_m| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \sum_{k>m} \mathbb{V}\mathrm{ar}(X_k).$$

Since $\sum_{k=1}^{\infty} \mathbb{V}\mathrm{ar}(X_k) < \infty$, the tail $\sum_{k>m} \mathbb{V}\mathrm{ar}(X_k) \to 0$ as $m \to \infty$.

Step 3: Cauchy criterion.

For each $\varepsilon > 0$, define the events

$$A_m^{\varepsilon} = \left\{ \sup_{k>m} |S_k - S_m| > \varepsilon \right\}.$$

These events are decreasing in m (as m increases, the supremum is taken over a smaller set). By Step 2, $\mathbb{P}(A_m^{\varepsilon}) \to 0$, so by continuity from above,

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} A_m^{\varepsilon} \right) = \lim_{m\to\infty} \mathbb{P}(A_m^{\varepsilon}) = 0.$$

Step 4: Conclusion.

The event that S_n does not converge (equivalently, is not Cauchy) is contained in:

$$\{S_n \text{ is not Cauchy}\} \subseteq \bigcup_{j=1}^{\infty} \bigcap_{m=1}^{\infty} A_m^{1/j}.$$

By Step 3 and a countable union bound,

$$\mathbb{P}(S_n \text{ does not converge}) \leq \sum_{j=1}^{\infty} \mathbb{P}\left(\bigcap_{m=1}^{\infty} A_m^{1/j} \right) = 0.$$

Therefore $\sum_{n=1}^{\infty} X_n$ converges almost surely. □

Remark F.8. The proof reveals why Kolmogorov's maximal inequality is essential: it controls the **supremum** $\sup_{k>m} |S_k - S_m|$, not just $|S_n - S_m|$ for a single n. This "max costs nothing" feature is what enables the Cauchy criterion argument. Chebyshev's inequality alone gives only $\mathbb{P}(|S_n - S_m| > \varepsilon) \leq \varepsilon^{-2} \sum_{k=m+1}^{n} \mathbb{V}\mathrm{ar}(X_k)$, which controls each fixed n but does not directly yield almost sure convergence.

F.5 Proof of Kolmogorov's three series theorem

Theorem 6 (Kolmogorov's Three-Series Theorem (restated))**.** *Let $(X_n)_{n\geq 1}$ be independent random variables. Fix any constant $c > 0$ and define the truncated variables $Y_n = X_n \mathbf{1}_{\{|X_n| \leq c\}}$. Then $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following three series all converge:*

(i) $\sum_{n=1}^{\infty} P(|X_n| > c)$

(ii) $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$

(iii) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n)$

Moreover, if these conditions hold for one value of $c > 0$, they hold for all $c > 0$.

Proof of Sufficiency

Assume the three series converge.

Step 1: Truncation has finite effect.

By condition (i) and the first Borel–Cantelli lemma, $P(|X_n| > c \text{ i.o.}) = 0$. Hence $X_n = Y_n$ for all sufficiently large n, almost surely. It therefore suffices to show $\sum Y_n$ converges a.s.

Step 2: Center the truncated variables.

Let $Z_n = Y_n - \mathbb{E}[Y_n]$. Then (Z_n) are independent with $\mathbb{E}[Z_n] = 0$ and

$$\operatorname{Var}(Z_n) = \operatorname{Var}(Y_n).$$

By condition (iii), $\sum_{n=1}^{\infty} \operatorname{Var}(Z_n) < \infty$.

Step 3: Apply Kolmogorov–Khinchin.

The (Z_n) are independent with $\mathbb{E}[Z_n] = 0$ and $\sum_{n=1}^{\infty} \mathbb{V}\mathrm{ar}(Z_n) = \sum_{n=1}^{\infty} \mathbb{V}\mathrm{ar}(Y_n) < \infty$ by condition (iii). By the Kolmogorov–Khinchin convergence theorem (Theorem F.6), $\sum_{n=1}^{\infty} Z_n$ converges almost surely.

Step 4: Conclusion.

We have $\sum Z_n$ converges a.s. By condition (ii), $\sum \mathbb{E}[Y_n]$ converges. Therefore,

$$\sum_{n=1}^{\infty} Y_n = \sum_{n=1}^{\infty} Z_n + \sum_{n=1}^{\infty} \mathbb{E}[Y_n]$$

converges a.s. Since $X_n = Y_n$ eventually, $\sum X_n$ converges a.s. □

Proof of Necessity

Assume $\sum_{n=1}^{\infty} X_n$ converges almost surely.

Step 1: Condition (i) holds.

Since $\sum X_n$ converges a.s., we have $X_n \to 0$ a.s. Thus $P(|X_n| > c \text{ i.o.}) = 0$. The events $\{|X_n| > c\}$ are independent, so by the second Borel–Cantelli lemma (contrapositive), $\sum P(|X_n| > c) < \infty$.

Step 2: The truncated series converges a.s.

By Step 1 and Borel–Cantelli I, $X_n = Y_n$ eventually a.s. Since $\sum X_n$ converges, $\sum Y_n$ converges a.s.

Step 3: Condition (iii) holds via symmetrization.

Let (X_n') be an independent copy of (X_n), and set $Y_n' = X_n' \mathbf{1}_{\{|X_n'| \le c\}}$. The symmetrized variables $Y_n - Y_n'$ are independent and symmetric.

Since $\sum Y_n$ and $\sum Y_n'$ both converge a.s., so does $\sum (Y_n - Y_n')$.

For independent symmetric bounded random variables, a.s. convergence of $\sum (Y_n - Y_n')$ implies

$$\sum_{n=1}^{\infty} \mathbb{E}[(Y_n - Y_n')^2] < \infty.$$

(See lemma below.)

Since $\mathbb{E}[(Y_n - Y_n')^2] = 2 \operatorname{Var}(Y_n)$, we obtain $\sum \operatorname{Var}(Y_n) < \infty$.

Step 4: Condition (ii) holds.

Let $Z_n = Y_n - \mathbb{E}[Y_n]$. By condition (iii) and the sufficiency direction applied to (Z_n), the series $\sum Z_n$ converges a.s. Since $\sum Y_n$ converges a.s. (Step 2), we have

$$\sum_{n=1}^{\infty} \mathbb{E}[Y_n] = \sum_{n=1}^{\infty} Y_n - \sum_{n=1}^{\infty} Z_n$$

converges. □

Lemma F.9. *Let $(Z_n)_{n\geq 1}$ be independent symmetric random variables with $|Z_n| \leq M$ for all n. If $\sum_{n=1}^{\infty} Z_n$ converges almost surely, then $\sum_{n=1}^{\infty} \mathbb{E}[Z_n^2] < \infty$.*

Proof. Suppose for contradiction that $\sum_{n=1}^{\infty} \mathbb{E}[Z_n^2] = \infty$. Let $S_n = \sum_{k=1}^{n} Z_k$ and $V_n = \sum_{k=1}^{n} \mathbb{E}[Z_k^2]$. Since the Z_k are symmetric (hence mean zero) and independent,

$$\operatorname{Var}(S_n) = \sum_{k=1}^{n} \mathbb{E}[Z_k^2] = V_n \to \infty.$$

We verify the Lindeberg condition: for any $\varepsilon > 0$,

$$\frac{1}{V_n} \sum_{k=1}^{n} \mathbb{E}\left[Z_k^2 \mathbf{1}_{\{|Z_k| > \varepsilon \sqrt{V_n}\}} \right] \to 0.$$

Since $|Z_k| \leq M$ and $V_n \to \infty$, for all sufficiently large n we have $M < \varepsilon \sqrt{V_n}$, so the indicators vanish and the sum is zero.

By the Lindeberg–Feller central limit theorem,

$$\frac{S_n}{\sqrt{V_n}} \xrightarrow{d} N(0,1).$$

In particular, $P(|S_n| > \sqrt{V_n}) \to P(|Z| > 1) > 0$ for $Z \sim N(0,1)$. Since $V_n \to \infty$, this implies $|S_n| \to \infty$ in probability, contradicting the assumption that S_n converges almost surely. □

Remark F.10 (Application to the Three-Series Theorem). In the necessity proof, the symmetrized variables $Z_n = Y_n - Y_n'$ are independent, symmetric, and bounded by $2c$. Since $\sum Z_n$ converges a.s., Lemma F.9 gives

$$\sum_{n=1}^{\infty} \mathbb{E}[(Y_n - Y_n')^2] < \infty.$$

Since Y_n and Y_n' are independent with the same distribution,

$$\mathbb{E}[(Y_n - Y_n')^2] = \mathbb{E}[Y_n^2] - 2\mathbb{E}[Y_n]\mathbb{E}[Y_n'] + \mathbb{E}[(Y_n')^2] = 2\operatorname{Var}(Y_n),$$

yielding $\sum \operatorname{Var}(Y_n) < \infty$.

Summary of Key Techniques

(i) **Borel–Cantelli lemmas:** Control the truncation error and establish condition (i).

(ii) **Kolmogorov's maximal inequality:** Converts variance summability into a.s. convergence for centered variables.

(iii) **Symmetrization:** Reduces the necessity proof for condition (iii) to symmetric random variables, where variance control is more direct.

F.6 Connection of LLN and Ergodic Theorem

We can think about the SLLN as a consequence of an ergodic theorem.

Recall that if $(M, \mathcal{B}, \mu)$ is a probability space, then $T : M \to M$ is a **measure-preserving** transformation of M, if it is measurable and $\mu(T^{(-1)}A) = \mu(A)$ for any $A \in \mathcal{B}$. Transformation T is called **ergodic** if every invariant subset has measure 0 or 1. The Birkhoff ergodic theorem says that for any measurable function f and ergodic T, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int f(x)\, \mu(dx)$$

for μ-almost all x. This has a striking resemblance to Kolmogorov's SLLN. In fact, it turns out that one can derive the SLLN as a consequence of the ergodic theorem by defining a suitable space M.

If we have a sequence of i.i.d. real-valued random variables X_i defined on $(\Omega, \mathcal{F}, \mathbb{P})$, then can define a product space $M = \prod_{i=1}^{\infty} \mathbb{R}$ with the product σ-algebra $\widehat{\mathcal{B}}$, and the measure μ, which is a product of the distribution measures μ_i of the random variables X_i. We define the functions $Y_i : M \to \mathbb{R}$ that map $(\widehat{\omega}_1, \widehat{\omega}_2, \ldots)$ to $\widehat{\omega}_i$ and observe that X_i and Y_i have the same distribution:

$$\mu(Y_1 \in A_1, \ldots, Y_n \in A_n) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n).$$

We also have a shift operator T on M that maps $(\omega_1, \omega_2, \omega_3, \ldots)$ to $(\omega_2, \omega_3, \omega_4, \ldots)$. It is clear that this operator is measure-preserving in the sense that $\mu(T^{-1}(A)) = \mu(A)$, for any measurable set $A \in \widehat{\mathcal{F}}$.

Lemma F.11. *Transformation T is ergodic.*

This requires a proof and the ideas of the proof are similar to the ideas behind the proof of the Kolmogorov's zero-one theorem. What is used is that μ is a product measure generated by cylinder sets. See a book on ergodic theory for a proof.

In addition, $Y_i(\widehat{\omega}) = \widehat{\omega}_i = Y_1(T^{i-1}\widehat{\omega})$.

So by Birkhoff's ergodic theorem, it follows that

$$\frac{1}{n} \sum_{i=1}^{n} Y_i(\widehat{\omega}) = \frac{1}{n} \sum_{k=0}^{n-1} Y_1(T^k \widehat{\omega}) \to \widehat{\mathbb{E}} Y_1 = \mathbb{E} X_1,$$

μ almost surely. It remains to get back from measure μ and functions Y_i to random variables X_i and measure $\mathbb{P}$.

Let $C \subset M$ be the set of sequences $(x_i) \in M$ such that $\frac{1}{n}\sum_{i=1}^{n} x_i$ converges to $\mathbb{E}(X_1)$. Then the assertion $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to $\mathbb{E}(X_1)$ almost surely means that $\mathbb{P}(\omega, (X_k(\omega)) \in C) = 1$. This is equivalent to the statement that $\mu(\widehat{\omega}, (Y_k(\widehat{\omega})) \in C) = 1$ because X_k and Y_k have the same distribution measure μ on M. And as we have just seen, $\mu(\widehat{\omega}, (Y_k(\widehat{\omega})) \in C) = 1$ because of Birkhoff's ergodic theorem.

F.7 SLLN with finite second moment

It is an insight due to Kolmogorov, that the method of subsequences is still useful for the proof of the almost sure convergence. The idea is to choose a convergent subsequence and to prove that fluctuations of partial sums S_n between the elements of this subsequence converge to zero almost surely.

Theorem F.12. *If $X, X_1, X_2, ...$ are i.i.d. random variables with $E(X^2) = \sigma^2 < \infty$, and $S_n := X_1 + X_2 + ... + X_n$, then*

$$\frac{S_n}{n} \xrightarrow{a.s.} E(X). \tag{F.1}$$

Proof. Without loss of generality we can assume that $E(X) = 0$. Then, as we have seen, $S_{k^2}/k^2 \xrightarrow{a.s.} 0$. Indeed,

$$\mathbb{P}\left(\left|\frac{S_{k^2}}{k^2}\right| > \epsilon\right) < \frac{\sigma^2}{k^2\epsilon^2},$$

and the convergence holds by the Borel-Cantelli lemma, as in our previous theorem.

Now, let us define

$$M_k = \max_{k^2 \le n < (k+1)^2} \left|\frac{S_n}{n} - \frac{S_{k^2}}{k^2}\right|.$$

Since for $k^2 \le n < (k+1)^2$,

$$\left|\frac{S_n}{n}\right| \le \left|\frac{S_{k^2}}{k^2}\right| + \left|\frac{S_n}{n} - \frac{S_{k^2}}{k^2}\right| \le \left|\frac{S_{k^2}}{k^2}\right| + |M_k|,$$

and we know that $\frac{S_{k^2}}{k^2} \xrightarrow{a.s.} 0$, therefore it is enough to prove that $M_k \xrightarrow{a.s.} 0$. (Since if $X_n \xrightarrow{a.s.} 0$ and $Y_n \xrightarrow{a.s.}$, then $X_n + Y_n \xrightarrow{a.s.} 0$.)

For convenience we define

$$D_k := \max_{k^2 \le n < (k+1)^2} |S_n - S_{k^2}|.$$

Then, we have

$$\begin{aligned} M_k &= \max_{k^2 \le n < (k+1)^2} \left|\frac{S_n - S_{k^2}}{n} + \frac{S_{k^2}}{n} - \frac{S_{k^2}}{k^2}\right| \\ &\le \left|\frac{D_k}{k^2}\right| + 2\left|\frac{S_{k^2}}{k^2}\right|. \end{aligned}$$

It follows that it is enough to prove that $D_k/k^2 \xrightarrow{a.s.} 0$.

We have

$$\begin{aligned} D_k^2 &= \max_{1\leq m\leq 2n} (S_{k^2+m} - S_{k^2})^2 \\ &\leq \sum_{m=1}^{2n} (S_{k^2+m} - S_{k^2})^2. \end{aligned}$$

Taking expectations on both sides, we get that

$$\begin{aligned} E(D_k^2) &\leq \sum_{m=1}^{2k} m\sigma^2 = k(2k+1)\sigma^2 \\ &\leq 4k^2\sigma^2, \end{aligned}$$

Hence we get that

$$\begin{aligned} \mathbb{P}\left(\left|\frac{D_k}{k^2}\right| > \epsilon\right) &\leq \frac{E\left(\left(\frac{D_k}{k^2}\right)^2\right)}{\epsilon^2} \\ &\leq \frac{4\sigma^2}{k^2\epsilon^2}. \end{aligned}$$

Hence,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{D_k}{k^2}\right| > \epsilon\right) < \infty$$

By applying the Borel-Cantelli lemma (BC I), we get that $D_k/k^2 \xrightarrow{a.s.} 0$, which completes the proof. □

Appendix G

Characteristic functions

G.1 Contour integration argument needed for CF of Gaussian r.v.

We want to show that

$$I(t) := \int_{-\infty}^{\infty} e^{-(x-it)^2/2}\,dx = \sqrt{2\pi}.$$

Consider the function $f(z) = e^{-z^2/2}$, which is entire (analytic on all of $\mathbb{C}$). For fixed $t \in \mathbb{R}$ and $R > 0$, consider the rectangular contour Γ_R with vertices at $-R$, R, $R + it$, and $-R + it$, traversed counterclockwise.

Since f is entire, Cauchy's theorem gives $\oint_{\Gamma_R} f(z)\,dz = 0$. We decompose this into four integrals:

1. **Bottom edge** (from $-R$ to R along the real axis):

$$I_1(R) = \int_{-R}^{R} e^{-x^2/2}\,dx \to \sqrt{2\pi} \quad \text{as } R \to \infty.$$

2. **Top edge** (from $R + it$ to $-R + it$):

$$I_3(R) = \int_{R}^{-R} e^{-(x+it)^2/2}\,dx = -\int_{-R}^{R} e^{-(x+it)^2/2}\,dx \to -I(t) \quad \text{as } R \to \infty.$$

3. **Right edge** (from R to $R + it$): Parametrize by $z = R + iy$ for $y \in [0, t]$ (assuming $t > 0$; the case $t < 0$ is similar):

$$I_2(R) = \int_{0}^{t} e^{-(R+iy)^2/2} \cdot i\,dy.$$

We have $-(R+iy)^2/2 = -(R^2 - y^2 + 2iRy)/2 = -(R^2 - y^2)/2 - iRy$, so

$$|e^{-(R+iy)^2/2}| = e^{-(R^2-y^2)/2} \le e^{-(R^2-t^2)/2}.$$

Thus $|I_2(R)| \le |t| \cdot e^{-(R^2-t^2)/2} \to 0$ as $R \to \infty$.

4. **Left edge** (from $-R+it$ to $-R$): By a similar estimate, $|I_4(R)| \to 0$ as $R \to \infty$.

Taking $R \to \infty$ in the equation $I_1(R) + I_2(R) + I_3(R) + I_4(R) = 0$, we obtain

$$\sqrt{2\pi} + 0 - I(t) + 0 = 0,$$

hence $I(t) = \sqrt{2\pi}$. This completes the proof that $\varphi(t) = e^{-t^2/2}$ for $N(0,1)$.

For $N(\mu, \sigma^2)$, if $Z \sim N(0,1)$ then $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$, so

$$\varphi_X(t) = e^{i\mu t}\varphi_Z(\sigma t) = e^{i\mu t}e^{-\sigma^2 t^2/2} = e^{i\mu t - \sigma^2 t^2/2}.$$

G.2 Proof of inversion formula

The fundamental fact about characteristic functions is that they uniquely determine the distribution. This is established via the **inversion formula**, which allows us to recover the distribution from its characteristic function.

Theorem G.1 (Inversion formula)**.** *Let φ be the characteristic function of a probability measure μ with distribution function F. If $a < b$ are continuity points of F, then*

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt.$$

Proof. We compute the integral on the right-hand side by interchanging the order of integration. Using Fubini's theorem (justified below):

$$\begin{aligned}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt &= \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \left(\int_{\mathbb{R}} e^{itx}\, d\mu(x) \right) dt \\ &= \int_{\mathbb{R}} \left(\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt \right) d\mu(x).\end{aligned}$$

The inner integral can be evaluated using the formula

$$\int_{-T}^{T} \frac{e^{ity}}{it}\, dt = 2\int_{0}^{T} \frac{\sin(ty)}{t}\, dt,$$

which follows from the fact that $\cos(ty)/t$ is odd. Thus the inner integral equals

$$2\int_0^T \frac{\sin(t(x-a))}{t}\, dt - 2\int_0^T \frac{\sin(t(x-b))}{t}\, dt.$$

Now we use the classical Dirichlet integral:

$$\int_0^\infty \frac{\sin u}{u}\, du = \frac{\pi}{2}.$$

By substitution $u = ty$, we have for $y \neq 0$:

$$\int_0^T \frac{\sin(ty)}{t}\,dt = \operatorname{sgn}(y)\int_0^{T|y|} \frac{\sin u}{u}\,du \to \operatorname{sgn}(y)\cdot\frac{\pi}{2} \quad \text{as } T\to\infty.$$

Define

$$g_T(x) := \frac{1}{2\pi}\left(2\int_0^T \frac{\sin(t(x-a))}{t}\,dt - 2\int_0^T \frac{\sin(t(x-b))}{t}\,dt\right).$$

As $T \to \infty$:

- If $x < a$: both terms contribute $-\pi/2$, so $g_T(x) \to 0$.
- If $a < x < b$: first term contributes $+\pi/2$, second contributes $-\pi/2$, so $g_T(x) \to 1$.
- If $x > b$: both terms contribute $+\pi/2$, so $g_T(x) \to 0$.
- If $x = a$ or $x = b$: $g_T(x) \to 1/2$.

Thus $g_T(x) \to \mathbf{1}_{(a,b)}(x) + \frac{1}{2}\mathbf{1}_{\{a,b\}}(x)$ pointwise. Since $|g_T(x)| \le C$ uniformly (the Dirichlet integral is bounded), the Dominated Convergence Theorem gives

$$\lim_{T\to\infty}\int_{\mathbb{R}} g_T(x)\,d\mu(x) = \int_{\mathbb{R}} \mathbf{1}_{(a,b)}(x)\,d\mu(x) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}).$$

When a and b are continuity points of F, we have $\mu(\{a\}) = \mu(\{b\}) = 0$, so the limit equals $\mu((a,b)) = F(b) - F(a)$.

Justification of Fubini. The integrand satisfies

$$\left|\frac{e^{-ita} - e^{-itb}}{it}e^{itx}\right| \le \frac{|e^{-ita} - e^{-itb}|}{|t|} \le \frac{2}{|t|} \wedge (b-a),$$

where the second bound uses $|e^{i\theta_1} - e^{i\theta_2}| \le |\theta_1 - \theta_2|$. This is integrable over $[-T, T] \times \mathbb{R}$ with respect to $dt \times d\mu(x)$, justifying the interchange. □

G.3 Proof of Helly's Selection Theorem

Theorem G.2 (Helly's selection theorem)**.** *Let $(F_n)_{n\ge1}$ be a sequence of distribution functions. Then there exists a subsequence $(F_{n_k})_{k\ge1}$ and a right-continuous nondecreasing function $F : \mathbb{R} \to [0,1]$ such that $F_{n_k}(x) \to F(x)$ at every continuity point of F.*

Remark G.1. The limiting function F satisfies $0 \le F(x) \le 1$ and is nondecreasing and right-continuous, but it need not be a proper distribution function. Specifically, we may have

$$F(-\infty) := \lim_{x\to-\infty} F(x) > 0 \quad \text{or} \quad F(+\infty) := \lim_{x\to+\infty} F(x) < 1.$$

This corresponds to mass "escaping to infinity." If the sequence (F_n) is tight, meaning that for every $\varepsilon > 0$ there exists $M > 0$ such that $F_n(M) - F_n(-M) \ge 1 - \varepsilon$ for all n, then $F(-\infty) = 0$ and $F(+\infty) = 1$, so F is a proper distribution function.

Proof. The proof proceeds in three steps.

Step 1: Convergence along rationals via diagonalization.

Let $\{r_1, r_2, r_3, \ldots\}$ be an enumeration of the rationals $\mathbb{Q}$. We construct a subsequence along which $F_n(r)$ converges for every $r \in \mathbb{Q}$.

Since $0 \le F_n(r_1) \le 1$ for all n, the sequence $(F_n(r_1))_{n\ge1}$ is bounded. By the Bolzano–Weierstrass theorem, there exists a subsequence $(n_k^{(1)})_{k\ge1}$ such that $\lim_{k\to\infty} F_{n_k^{(1)}}(r_1)$ exists.

Now consider the sequence $(F_{n_k^{(1)}}(r_2))_{k\ge1}$. This is again bounded in $[0,1]$, so there exists a further subsequence $(n_k^{(2)})_{k\ge1}$ of $(n_k^{(1)})_{k\ge1}$ such that $\lim_{k\to\infty} F_{n_k^{(2)}}(r_2)$ exists. Since $(n_k^{(2)})$ is a subsequence of $(n_k^{(1)})$, the limit $\lim_{k\to\infty} F_{n_k^{(2)}}(r_1)$ also exists (and equals the previously obtained limit).

Continuing inductively, for each $j \ge 1$ we obtain a subsequence $(n_k^{(j)})_{k\ge1}$ of $(n_k^{(j-1)})_{k\ge1}$ such that $\lim_{k\to\infty} F_{n_k^{(j)}}(r_i)$ exists for all $i \le j$.

Define the diagonal subsequence $m_k := n_k^{(k)}$. For any fixed j, the sequence $(m_k)_{k\ge j}$ is a subsequence of $(n_k^{(j)})_{k\ge1}$, so $\lim_{k\to\infty} F_{m_k}(r_j)$ exists. Since j was arbitrary, the limit

$$G(r) := \lim_{k\to\infty} F_{m_k}(r)$$

exists for every rational $r \in \mathbb{Q}$. The function $G : \mathbb{Q} \to [0,1]$ is nondecreasing: if $r < s$ are rationals, then $F_{m_k}(r) \le F_{m_k}(s)$ for all k, so $G(r) \le G(s)$.

Step 2: Extension to all of $\mathbb{R}$.

We extend G to a function F on all of $\mathbb{R}$ by defining

$$F(x) := \inf\{G(r) : r \in \mathbb{Q},\, r > x\}.$$

We verify that F has the required properties:

1. F **takes values in** $[0,1]$: Since $0 \le G(r) \le 1$ for all $r \in \mathbb{Q}$, we have $0 \le F(x) \le 1$.

2. F **is nondecreasing**: If $x < y$, then $\{r \in \mathbb{Q} : r > y\} \subset \{r \in \mathbb{Q} : r > x\}$, so $F(x) \le F(y)$.

3. F **is right-continuous**: Fix $x \in \mathbb{R}$ and $\varepsilon > 0$. By definition of infimum, there exists $r \in \mathbb{Q}$ with $r > x$ and $G(r) < F(x) + \varepsilon$. For any $y \in (x, r)$, we have

   $$F(x) \le F(y) \le G(r) < F(x) + \varepsilon,$$

   where the middle inequality holds because $r > y$. Thus $|F(y) - F(x)| < \varepsilon$ for all $y \in (x, r)$, proving right-continuity.

Step 3: Convergence at continuity points of F.

Let x be a continuity point of F. We show that $F_{m_k}(x) \to F(x)$.

Upper bound. For any rational $s > x$, we have $F_{m_k}(x) \le F_{m_k}(s)$ since F_{m_k} is nondecreasing. Taking $k \to \infty$:

$$\limsup_{k\to\infty} F_{m_k}(x) \le G(s).$$

Since this holds for all rational $s > x$:

$$\limsup_{k\to\infty} F_{m_k}(x) \le \inf\{G(s) : s \in \mathbb{Q},\, s > x\} = F(x).$$

Lower bound. For any rational $r < x$, we have $F_{m_k}(x) \ge F_{m_k}(r)$. Taking $k \to \infty$:

$$\liminf_{k\to\infty} F_{m_k}(x) \ge G(r).$$

Since this holds for all rational $r < x$:

$$\liminf_{k\to\infty} F_{m_k}(x) \ge \sup\{G(r) : r \in \mathbb{Q},\, r < x\} =: F(x-),$$

where $F(x-)$ denotes the left limit of F at x.

Conclusion. Since x is a continuity point of F, we have $F(x-) = F(x)$. Combining the bounds:

$$F(x) = F(x-) \le \liminf_{k\to\infty} F_{m_k}(x) \le \limsup_{k\to\infty} F_{m_k}(x) \le F(x).$$

Therefore $\lim_{k\to\infty} F_{m_k}(x) = F(x)$. □

Remark G.2. The set of discontinuity points of any nondecreasing function is at most countable (each discontinuity corresponds to a jump, and there can be at most countably many disjoint intervals of positive length in $[0, 1]$). Thus $F_{m_k}(x) \to F(x)$ for all but countably many x.

G.4 Sufficient condition for the existence of density

Theorem G.3. *If $\int |\varphi(t)|\, dt < \infty$ then μ has bounded continuous density*

$$f(x) = \frac{1}{2\pi} \int e^{-itx} \varphi(t)\, dt$$

Proof. We note that

$$\Big| \frac{e^{-itb} - e^{-ita}}{-it} \Big| = \Big| \int_a^b e^{-ity}\, dy \Big| \le |b - a|,$$

so the integral in the inversion formula converges absolutely under the assumption that $\int |\varphi(t)|\, dt < \infty$ and

$$\mu(a, b) + \frac{1}{2}\mu(\{a, b\}) \le \frac{b - a}{2\pi} \int |\varphi(t)|\, dt.$$

This implies that μ has no point masses. Then we have

$$\begin{aligned}\mu(x, x+h) &= \frac{1}{2\pi}\int \Big(\int_x^{x+h} e^{-ity} dy\Big)\varphi(t)\, dt \\ &= \int_x x+h\Big(\frac{1}{2\pi}\int e^{-ity}\varphi(t)\, dt\Big) dy\end{aligned}$$

by Fubini's theorem, so by definition μ has density function

$$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)\, dt.$$

The dominated convergence theorem implies f is continuous and the proof is complete. □

Theorem G.3 shows that there is some duality between density functions and characteristic functions with the property $\int |\varphi(t)|\, dt < \infty$. Applying this duality to our earlier examples, we can get new examples.

Example G.4 (The Cauchy distribution). The random variable with density $\frac{1}{\pi}(1+x^2)^{-}1$ has ch.f. $\varphi(t) = e^{-|t|}$.

Example G.5 (Polya's distirbution). This is an example where the characteristic function is compactly supported. The density is

$$f(x) = \frac{1-\cos x}{\pi x^2},$$

and the characteristic function is $\varphi(t) = (1-|t|)^+$.

G.5 Taylor Expansion for Characteristic Functions

We give a direct proof of the Taylor expansion for characteristic functions that does not rely on differentiability of φ. The key is to expand $e^{i\theta}$ and control the remainder.

Lemma G.6 (Taylor expansion of $e^{i\theta}$)**.** *For all $\theta \in \mathbb{R}$:*

1. $e^{i\theta} = 1 + i\theta + r_1(\theta)$, *where* $|r_1(\theta)| \le \frac{\theta^2}{2}$ *and* $|r_1(\theta)| \le 2|\theta|$.
2. $e^{i\theta} = 1 + i\theta - \dfrac{\theta^2}{2} + r_2(\theta)$, *where* $|r_2(\theta)| \le \frac{|\theta|^3}{6}$ *and* $|r_2(\theta)| \le \theta^2$.

Proof. **Part (1).** Define $r_1(\theta) = e^{i\theta} - 1 - i\theta$. For the first bound, write

$$e^{i\theta} = \sum_{k=0}^{\infty} \frac{(i\theta)^k}{k!} = 1 + i\theta + \sum_{k=2}^{\infty} \frac{(i\theta)^k}{k!},$$

so

$$|r_1(\theta)| = \left|\sum_{k=2}^{\infty} \frac{(i\theta)^k}{k!}\right| \le \sum_{k=2}^{\infty} \frac{|\theta|^k}{k!} \le \frac{\theta^2}{2}\sum_{j=0}^{\infty} \frac{|\theta|^j}{(j+2)!/(2!)} \le \frac{\theta^2}{2} e^{|\theta|}.$$

For $|\theta| \le 1$, this gives $|r_1(\theta)| \le e\theta^2/2 \le 2\theta^2$. For $|\theta| > 1$, use the crude bound $|r_1(\theta)| \le |e^{i\theta}| + |1| + |\theta| \le 2 + |\theta| \le 2|\theta| + |\theta| \le 3|\theta|$.

A cleaner bound: $|r_1(\theta)| = |e^{i\theta} - 1 - i\theta|$. Using the integral form of the remainder:

$$e^{i\theta} - 1 - i\theta = \int_0^\theta (e^{iu} - 1) \cdot i \, du = \int_0^\theta i \, du \int_0^u ie^{iv} \, dv = -\int_0^\theta \int_0^u e^{iv} \, dv \, du.$$

Thus

$$|r_1(\theta)| \le \int_0^{|\theta|} \int_0^u 1 \, dv \, du = \int_0^{|\theta|} u \, du = \frac{\theta^2}{2}.$$

For the second bound, $|r_1(\theta)| = |e^{i\theta} - 1 - i\theta| \le |e^{i\theta} - 1| + |\theta|$. Using $|e^{i\theta} - 1| = 2|\sin(\theta/2)| \le |\theta|$, we get $|r_1(\theta)| \le 2|\theta|$.

Part (2). Define $r_2(\theta) = e^{i\theta} - 1 - i\theta + \frac{\theta^2}{2}$. From the power series:

$$r_2(\theta) = \sum_{k=3}^{\infty} \frac{(i\theta)^k}{k!}.$$

Thus

$$|r_2(\theta)| \le \sum_{k=3}^{\infty} \frac{|\theta|^k}{k!} \le \frac{|\theta|^3}{6} \sum_{j=0}^{\infty} \frac{|\theta|^j}{(j+3)!/(3!)} \le \frac{|\theta|^3}{6} e^{|\theta|}.$$

For $|\theta| \le 1$, this gives $|r_2(\theta)| \le e|\theta|^3/6 \le |\theta|^3$.

For the second bound, $|r_2(\theta)| = |r_1(\theta) + \theta^2/2| \le |r_1(\theta)| + \theta^2/2 \le \theta^2/2 + \theta^2/2 = \theta^2$. □

Proposition G.7 (Taylor expansion for characteristic functions)**.** *Let X be a random variable.*

1. *If $\mathbb{E}[|X|] < \infty$, then $\varphi(t) = 1 + it\mathbb{E}[X] + o(t)$ as $t \to 0$.*
2. *If $\mathbb{E}[X^2] < \infty$, then $\varphi(t) = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + o(t^2)$ as $t \to 0$.*

Proof. **Part (1).** Using Lemma G.6(1) with $\theta = tX$:

$$\varphi(t) = \mathbb{E}[e^{itX}] = 1 + it\mathbb{E}[X] + \mathbb{E}[r_1(tX)].$$

We need to show $\mathbb{E}[r_1(tX)] = o(t)$, i.e., $\mathbb{E}[r_1(tX)]/t \to 0$ as $t \to 0$.

By the lemma, $|r_1(tX)| \le \min\left(\frac{t^2X^2}{2}, 2|tX|\right)$, so

$$\frac{|r_1(tX)|}{|t|} \le \min\left(\frac{|t|X^2}{2}, 2|X|\right) \le 2|X|.$$

Since $\mathbb{E}[|X|] < \infty$, the right-hand side is integrable. As $t \to 0$,

$$\frac{|r_1(tX)|}{|t|} \le \frac{t^2X^2}{2|t|} = \frac{|t|X^2}{2} \to 0 \quad \text{pointwise.}$$

By the Dominated Convergence Theorem, $\mathbb{E}[|r_1(tX)|]/|t| \to 0$.

Part (2). Using Lemma G.6(2) with $\theta = tX$:

$$\varphi(t) = \mathbb{E}[e^{itX}] = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + \mathbb{E}[r_2(tX)].$$

We need to show $\mathbb{E}[r_2(tX)] = o(t^2)$, i.e., $\mathbb{E}[r_2(tX)]/t^2 \to 0$ as $t \to 0$.

By the lemma, $|r_2(tX)| \le \min\left(\frac{|t|^3|X|^3}{6}, t^2X^2\right)$, so

$$\frac{|r_2(tX)|}{t^2} \le \min\left(\frac{|t||X|^3}{6}, X^2\right) \le X^2.$$

Since $\mathbb{E}[X^2] < \infty$, the right-hand side is integrable. As $t \to 0$,

$$\frac{|r_2(tX)|}{t^2} \le \frac{|t||X|^3}{6} \to 0 \quad \text{pointwise.}$$

By the Dominated Convergence Theorem, $\mathbb{E}[|r_2(tX)|]/t^2 \to 0$. □

Remark G.8. The key advantage of this approach is that we obtain explicit bounds:

- If $\mathbb{E}[X^2] < \infty$, then $|\varphi(t) - 1 - it\mathbb{E}[X]| \le \frac{t^2}{2}\mathbb{E}[X^2]$.
- If $\mathbb{E}[|X|^3] < \infty$, then $\left|\varphi(t) - 1 - it\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2]\right| \le \frac{|t|^3}{6}\mathbb{E}[|X|^3]$.

These bounds are used in proving quantitative versions of the Central Limit Theorem, such as the Berry–Esseen inequality.

Appendix H

Central Limit Theorems

H.1 Complex exponential limit

Lemma H.1. *If (z_n) is a sequence of complex numbers with $z_n \to z$, then*

$$\Big(1 + \frac{z_n}{n}\Big)^n \to e^z.$$

Proof. Write $w_n = z_n/n$. For n large enough, $|w_n| < 1/2$, and we can use the principal branch of the logarithm: $\log(1+w) = w - w^2/2 + O(w^3)$ for $|w| < 1/2$. Then

$$n\log(1+w_n) = n\Big(\frac{z_n}{n} - \frac{z_n^2}{2n^2} + O\Big(\frac{1}{n^3}\Big)\Big) = z_n - \frac{z_n^2}{2n} + O\Big(\frac{1}{n^2}\Big) \to z,$$

since $z_n \to z$. Exponentiating gives the result. □

Remark H.2. A more general version states: if z_n are complex with $nz_n \to z$ and $n|z_n|^2 \to 0$, then $(1+z_n)^n \to e^z$. This version is useful when the $o(1/n)$ remainder in the CF expansion is only known in the weaker sense. See Durrett, Theorem 3.4.2.

H.2 Proof of Lindeberg's CLT

The proof of the full Lindeberg CLT follows the same swapping strategy as the proof of Lyapunov's CLT in Section 8.5, but replaces the crude third-moment bound on the Taylor remainder with a refined estimate that involves only second moments. We restate the theorem for convenience.

Theorem 7 (Lindeberg's CLT)**.** *Suppose that a triangular array satisfies the three standard conditions and the Lindeberg condition: for every $\epsilon > 0$,*

$$\lim_{i\to\infty} \sum_{j=1}^{n_i} \mathbb{E}\big[X_{ij}^2 \mathbf{1}(|X_{ij}| > \epsilon)\big] = 0.$$

Then $S_i \xrightarrow{d} N(0,1)$.

The refined remainder bound

Let $f \in \mathbf{C}_b^3(\mathbb{R})$ with $|f^{(k)}| \le K$ for $k = 0, 1, 2, 3$, and let R, X be real numbers. Define the **Taylor remainder** after cancelling the zeroth-, first-, and second-order terms at R:

$$\Delta(R, X) := f(R + X) - f(R) - Xf'(R) - \frac{X^2}{2} f''(R). \tag{H.1}$$

We have two bounds on $|\Delta|$, coming from Taylor expansions of different order:

Third-order bound. By the third-order Taylor formula with Lagrange remainder,

$$|\Delta(R, X)| = \frac{|X|^3}{6} |f'''(\alpha)| \le \frac{K}{6} |X|^3, \tag{H.2}$$

where α lies between R and $R + X$.

Second-order bound. By the second-order Taylor formula, $f(R+X) = f(R) + Xf'(R) + \frac{X^2}{2} f''(\gamma)$ for some γ between R and $R + X$. Comparing with (H.1):

$$|\Delta(R, X)| = \frac{X^2}{2} |f''(\gamma) - f''(R)| \le KX^2. \tag{H.3}$$

The key idea is to use each bound where it is strongest. Fix $\epsilon > 0$ and split:

$$\begin{aligned} |\Delta(R, X)| &\le |\Delta(R, X)| \mathbf{1}(|X| \le \epsilon) + |\Delta(R, X)| \mathbf{1}(|X| > \epsilon) \\ &\le \frac{K\epsilon}{6} X^2 + KX^2 \mathbf{1}(|X| > \epsilon), \end{aligned} \tag{H.4}$$

using (H.2) with $|X|^3 \le \epsilon |X|^2$ on the first piece and (H.3) on the second. The advantage over the Lyapunov proof: the right-hand side involves only X^2, not $|X|^3$.

The proof

Proof of Theorem 8.6. Fix a row of the triangular array. Write $X_1, \dots, X_n$ for the entries (suppressing the row index), with $\mathbb{E}X_j = 0$, $\mathbb{E}X_j^2 = \sigma_j^2$, and $\sum_{j=1}^n \sigma_j^2 = 1$. The row sum is $S = X_1 + \cdots + X_n$. Construct independent $Z_j \sim N(0, \sigma_j^2)$, jointly independent of all X_j's, and set $T := Z_1 + \cdots + Z_n \sim N(0, 1)$.

As in Section 8.5.2, define hybrid sums $H_j = Z_1 + \cdots + Z_j + X_{j+1} + \cdots + X_n$ for $j = 0, \dots, n$, so that $H_0 = S$ and $H_n = T$. Each consecutive pair differs in the jth slot: $H_{j-1} = R_j + X_j$ and $H_j = R_j + Z_j$, where R_j is independent of both X_j and Z_j.

Fix $f \in \mathbf{C}_b^3(\mathbb{R})$ with $|f^{(k)}| \le K$ for $k = 0, 1, 2, 3$.

Telescoping. By the triangle inequality,

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \le \sum_{j=1}^{n} |\mathbb{E}f(H_{j-1}) - \mathbb{E}f(H_j)|.$$

Bounding each swap. Using the notation (H.1), we write

$$f(R_j + X_j) = f(R_j) + X_j f'(R_j) + \frac{X_j^2}{2} f''(R_j) + \Delta(R_j, X_j),$$

and similarly for $f(R_j+Z_j)$. Taking expectations and subtracting, the zeroth-, first-, and second-order terms cancel by independence and moment matching ($\mathbb{E}X_j = \mathbb{E}Z_j = 0$, $\mathbb{E}X_j^2 = \mathbb{E}Z_j^2 = \sigma_j^2$), leaving

$$|\mathbb{E}f(H_{j-1}) - \mathbb{E}f(H_j)| \le \mathbb{E}|\Delta(R_j, X_j)| + \mathbb{E}|\Delta(R_j, Z_j)|. \tag{H.5}$$

For the X_j term, apply the refined bound (H.4):

$$\mathbb{E}|\Delta(R_j, X_j)| \le \frac{K\epsilon}{6}\,\mathbb{E}X_j^2 + K\,\mathbb{E}\big[X_j^2 \mathbf{1}(|X_j| > \epsilon)\big].$$

For the Z_j term, since Z_j is Gaussian and has moments of all orders, use the third-order bound (H.2):

$$\mathbb{E}|\Delta(R_j, Z_j)| \le \frac{K}{6}\,\mathbb{E}|Z_j|^3 = \frac{Kc}{6}\,\sigma_j^3,$$

where $c := \mathbb{E}|Z|^3$ for $Z \sim N(0,1)$.

Summing over j. Combining and using $\sum_{j=1}^n \mathbb{E}X_j^2 = 1$:

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \le K \sum_{j=1}^n \mathbb{E}\big[X_j^2 \mathbf{1}(|X_j| > \epsilon)\big] + \frac{K\epsilon}{6} + \frac{Kc}{6}\sum_{j=1}^n \sigma_j^3. \tag{H.6}$$

Conclusion. Restoring the row index i, we show that each of the three terms in (H.6) can be made arbitrarily small. Let $\eta > 0$.

First term: For any fixed $\epsilon > 0$, this tends to zero as $i \to \infty$ by the Lindeberg condition (8.1).

Third term: Since $\sigma_{ij}^3 \le (\max_{1\le k\le n_i} \sigma_{ik}) \cdot \sigma_{ij}^2$, we have

$$\sum_{j=1}^{n_i} \sigma_{ij}^3 \le \Big(\max_{1\le k\le n_i} \sigma_{ik}\Big) \underbrace{\sum_{j=1}^{n_i} \sigma_{ij}^2}_{=1} \to 0$$

as $i \to \infty$, by Lemma 8.5.

Second term: This is $\frac{K\epsilon}{6}$, which is at our disposal.

Choose $\epsilon > 0$ so that $\frac{K\epsilon}{6} < \eta/3$, then choose i_0 large enough so that the first and third terms are each less than $\eta/3$ for all $i \ge i_0$. Then $|\mathbb{E}f(S_i) - \mathbb{E}f(T)| < \eta$ for all $i \ge i_0$.

Since $f \in \mathbf{C}_b^3(\mathbb{R})$ was arbitrary, Lemma 8.12 gives $S_i \xrightarrow{d} N(0,1)$. □

Remark H.3. Comparing the Lyapunov and Lindeberg proofs, the only difference is in how the Taylor remainder $\Delta(R_j, X_j)$ is bounded. Lyapunov uses the third-order bound $\frac{K}{6}|X_j|^3$ uniformly, which requires $\mathbb{E}|X_j|^3 < \infty$. Lindeberg uses the third-order bound on $\{|X_j| \le \epsilon\}$ (converting $|X_j|^3$ to ϵX_j^2) and the second-order bound on $\{|X_j| > \epsilon\}$. This splitting is the key insight that replaces the third-moment assumption with the Lindeberg condition.

Appendix I

Distribution of the Exit Time in Gambler's Ruin

In the fair gambler's ruin problem we computed $\mathcal{P}(S_\tau = A) = B/(A+B)$ using the linear martingale S_n and $\mathbb{E}(\tau) = AB$ using the quadratic martingale $S_n^2 - n$. The exponential martingale $M_n^{(\theta)} = e^{\theta S_n}/(\cosh\theta)^n$ encodes even more information: it gives the full Laplace transform of τ, and hence (in principle) the distribution of τ.

Throughout this appendix, S_n denotes the symmetric random walk with $S_0 = 0$ and $\tau = \inf\{n : S_n = A \text{ or } S_n = -B\}$, with $A, B > 0$ integers.

I.1 The Laplace transform of τ

The exponential martingale

For any $\theta \in \mathbb{R}$, the sequence

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

is a martingale. This is a special case of Example 9.10: take the i.i.d. increments $X_i = \pm 1$ with $\varphi(\theta) = \mathbb{E}(e^{\theta X_1}) = \cosh\theta$, and set $M_n = e^{\theta S_n}/(\cosh\theta)^n$. One verifies directly:

$$\mathbb{E}(M_{n+1}^{(\theta)} \mid \mathcal{F}_n) = M_n^{(\theta)} \cdot \mathbb{E}\left(\frac{e^{\theta X_{n+1}}}{\cosh\theta}\right) = M_n^{(\theta)} \cdot \frac{\cosh\theta}{\cosh\theta} = M_n^{(\theta)}.$$

Proposition I.1. *For every $\theta > 0$,*

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\right) = \frac{\cosh\big(\theta(A-B)/2\big)}{\cosh\big(\theta(A+B)/2\big)}. \tag{I.1}$$

Proof. Apply OST to the exponential martingale. Since $S_{n\wedge\tau} \in [-B, A]$, the stopped process satisfies

$$e^{-\theta B} \le e^{\theta S_{n\wedge\tau}} \le e^{\theta A},$$

and $(\cosh\theta)^{n\wedge\tau} \geq 1$ (since $\cosh\theta \geq 1$ and $n \wedge \tau \geq 0$). Therefore $|M_{n\wedge\tau}^{(\theta)}|$ is bounded by $e^{\theta \max(A,B)}$, and Theorem 10.3(ii) applies. On $\{\tau < \infty\}$ (which has probability 1), $M_{n\wedge\tau}^{(\theta)} \to M_\tau^{(\theta)}$ a.s., so

$$1 = \mathbb{E}(M_0^{(\theta)}) = \mathbb{E}(M_\tau^{(\theta)}) = \mathbb{E}\left(\frac{e^{\theta A}}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = A\}}\right) + \mathbb{E}\left(\frac{e^{-\theta B}}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = -B\}}\right). \tag{I.2}$$

Applying the same argument with θ replaced by $-\theta$ and using $\cosh(-\theta) = \cosh\theta$:

$$1 = \mathbb{E}\left(\frac{e^{-\theta A}}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = A\}}\right) + \mathbb{E}\left(\frac{e^{\theta B}}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = -B\}}\right). \tag{I.3}$$

Define

$$\alpha := \mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = A\}}\right), \qquad \beta := \mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\,\mathbf{1}_{\{S_\tau = -B\}}\right).$$

Then (I.2) and (I.3) become the linear system

$$e^{\theta A}\,\alpha + e^{-\theta B}\,\beta = 1,$$
$$e^{-\theta A}\,\alpha + e^{\theta B}\,\beta = 1.$$

Solving by Cramer's rule (determinant $e^{\theta(A+B)} - e^{-\theta(A+B)} = 2\sinh(\theta(A+B))$):

$$\alpha = \frac{e^{\theta B} - e^{-\theta B}}{2\sinh(\theta(A+B))} = \frac{\sinh(\theta B)}{\sinh(\theta(A+B))},$$
$$\beta = \frac{e^{\theta A} - e^{-\theta A}}{2\sinh(\theta(A+B))} = \frac{\sinh(\theta A)}{\sinh(\theta(A+B))}.$$

Adding:

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\right) = \alpha + \beta = \frac{\sinh(\theta A) + \sinh(\theta B)}{\sinh(\theta(A+B))}.$$

Using $\sinh x + \sinh y = 2\sinh\big((x+y)/2\big)\cosh\big((x-y)/2\big)$ in the numerator and $\sinh(2u) = 2\sinh u\cosh u$ in the denominator with $u = \theta(A+B)/2$:

$$\frac{\sinh(\theta A) + \sinh(\theta B)}{\sinh(\theta(A+B))} = \frac{2\sinh\big(\theta(A+B)/2\big)\cosh\big(\theta(A-B)/2\big)}{2\sinh\big(\theta(A+B)/2\big)\cosh\big(\theta(A+B)/2\big)} = \frac{\cosh\big(\theta(A-B)/2\big)}{\cosh\big(\theta(A+B)/2\big)}.$$

□

Setting $s = \log\cosh\theta$ (equivalently $\cosh\theta = e^s$, valid for $s \geq 0$), (I.1) becomes the Laplace transform:

$$\mathbb{E}(e^{-s\tau}) = \frac{\cosh\big(\theta(s)(A-B)/2\big)}{\cosh\big(\theta(s)(A+B)/2\big)}, \qquad \theta(s) = \operatorname{arccosh}(e^s).$$

Recovering known quantities

Example I.2 (The case $\theta \to 0$). As $\theta \to 0$, $(\cosh\theta)^\tau \to 1$, so the left side of (I.1) tends to 1. Expanding both sides to second order in θ should recover $\mathbb{E}(\tau) = AB$. Indeed, $\cosh u = 1 + u^2/2 + O(u^4)$, so

$$\frac{\cosh\big(\theta(A-B)/2\big)}{\cosh\big(\theta(A+B)/2\big)} = \frac{1 + \theta^2(A-B)^2/8 + O(\theta^4)}{1 + \theta^2(A+B)^2/8 + O(\theta^4)} = 1 - \theta^2\,\frac{(A+B)^2 - (A-B)^2}{8} + O(\theta^4).$$

The second-order term is $-\theta^2 \cdot AB/2$ (using $(A+B)^2-(A-B)^2 = 4AB$). On the other hand, expanding $1/(\cosh\theta)^\tau = (1+\theta^2/2+O(\theta^4))^{-\tau} = 1-\tau\theta^2/2+O(\theta^4)$, so

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\right) = 1 - \frac{\theta^2}{2}\mathbb{E}(\tau) + O(\theta^4).$$

Matching the θ^2 coefficients gives $\mathbb{E}(\tau) = AB$, as expected.

Example I.3 (Symmetric case $A = B$). When $A = B$, (I.1) simplifies to

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\right) = \frac{1}{\cosh(\theta A)}.$$

This is a clean closed form. The right side can be expanded as a series using $1/\cosh(x) = 2\sum_{k=0}^{\infty}(-1)^k e^{-(2k+1)x}$, which gives

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^\tau}\right) = 2\sum_{k=0}^{\infty}(-1)^k e^{-(2k+1)\theta A} = 2\sum_{k=0}^{\infty}(-1)^k(\cosh\theta)^{-(2k+1)A/\theta}\ldots$$

(The last step requires more care; inverting the Laplace transform here leads to the spectral decomposition of the discrete Laplacian on $\{-A,\ldots,A\}$ with absorbing boundary conditions.)

I.2 Asymptotic decay of $\mathcal{P}(\tau > n)$

Proposition I.4. *For fixed $A, B \geq 1$,*

$$\mathcal{P}(\tau > n) = C_{A,B}\,\lambda_{A,B}^n\,(1+o(1)) \quad \text{as } n\to\infty,$$

where $\lambda_{A,B} = \cos\big(\pi/(A+B)\big)$ is the largest eigenvalue (in absolute value) of the transition matrix of the random walk on $\{-B+1,\ldots,A-1\}$ with absorbing boundary conditions, and $C_{A,B} > 0$ is an explicit constant.

Sketch: The tail $\mathcal{P}(\tau > n) = \mathcal{P}(S_k \in (-B, A) \text{ for all } k \leq n)$ can be computed via the (n)th power of the sub-stochastic transition matrix P of the walk restricted to the interior. The eigenvalues of P are $\cos(k\pi/(A+B))$ for $k = 1,\ldots,A+B-1$, and the dominant one is $\cos(\pi/(A+B))$. Writing the initial distribution in the eigenbasis gives the stated asymptotic.

Alternatively, this follows from inverting the Laplace transform (I.1): the right side is meromorphic in s with poles at the zeros of $\cosh\big(\theta(s)(A+B)/2\big)$, and the pole closest to $s = 0$ determines the exponential decay rate. The calculation produces the same eigenvalue $\cos\big(\pi/(A+B)\big)$.

First passage time: the case $B = \infty$

When $B = \infty$ the lower barrier disappears and τ becomes the first passage time $T = \inf\{n : S_n = A\}$. This is the limiting case of the gambler with unlimited capital.

Proposition I.5. *For every $\theta > 0$,*

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^T}\right) = e^{-A\theta}. \tag{I.4}$$

Proof. As $B \to \infty$, $\tau_B \uparrow T$ monotonically, so $(\cosh\theta)^{-\tau_B} \downarrow (\cosh\theta)^{-T}$ (since $\cosh\theta > 1$ for $\theta \neq 0$). By the monotone convergence theorem,

$$\mathbb{E}\left(\frac{1}{(\cosh\theta)^T}\right) = \lim_{B\to\infty} \frac{\cosh\big(\theta(A-B)/2\big)}{\cosh\big(\theta(A+B)/2\big)}.$$

For large B, $\cosh(u) \sim \frac{1}{2}e^{|u|}$, so

$$\frac{\cosh\big(\theta(B-A)/2\big)}{\cosh\big(\theta(A+B)/2\big)} \sim \frac{e^{\theta(B-A)/2}}{e^{\theta(A+B)/2}} = e^{-\theta A}. \qquad \square$$

Remark I.6 (Consistency with $\mathbb{E}(T) = \infty$). The function $\theta \mapsto e^{-A|\theta|}$ has a cusp at $\theta = 0$ and is not twice differentiable there. On the other hand, expanding $(\cosh\theta)^{-T} \approx 1 - T\theta^2/2 + \cdots$ gives $\mathbb{E}((\cosh\theta)^{-T}) \approx 1 - \mathbb{E}(T)\theta^2/2 + \cdots$, which would require a smooth quadratic at the origin if $\mathbb{E}(T) < \infty$. The cusp in $e^{-A|\theta|}$ is the signature of $\mathbb{E}(T) = \infty$.

Remark I.7 (Stable laws and tail asymptotics). Setting $s = \log\cosh\theta$ and using $\operatorname{arccosh}(e^s) = \sqrt{2s}(1 + O(s))$ as $s \to 0$, (I.4) becomes

$$\mathbb{E}(e^{-sT}) \approx e^{-A\sqrt{2s}} \quad \text{as } s \to 0^+.$$

This is the Laplace transform of a **stable distribution with index** $1/2$. The classical Tauberian theorem then gives the tail asymptotic

$$\mathcal{P}(T > n) \sim \frac{A}{\sqrt{2\pi}}\, n^{-1/2} \quad \text{as } n \to \infty.$$

The polynomial decay $n^{-1/2}$ is dramatically heavier than the exponential tails $\mathcal{P}(\tau > n) \sim C\,\lambda^n$ of the two-barrier exit time (Proposition I.4). Removing the lower barrier transforms the tail from exponential to polynomial—and this is exactly what makes $\mathbb{E}(T) = \infty$.

Remark I.8 (Connection to Brownian motion). In the diffusive scaling limit ($S_{\lfloor nt\rfloor}/\sqrt{n} \to B_t$, where B_t is standard Brownian motion), T/n^2 converges in distribution to the first passage time $T_a = \inf\{t \geq 0 : B_t = a\}$ with $a = A/\sqrt{n}$. The Laplace transform of T_a is

$$\mathbb{E}(e^{-sT_a}) = e^{-a\sqrt{2s}},$$

which is the **inverse Gaussian** (or Lévy) distribution. Formula (I.4) is its discrete analogue, with $\sqrt{2s}$ replaced by $\operatorname{arccosh}(e^s)$.

Unfair game

When $p \neq 1/2$, the same idea works with the exponential martingale adapted to the unfair walk: $M_n^{(\theta)} = e^{\theta S_n}/\varphi(\theta)^n$, where $\varphi(\theta) = pe^{\theta} + qe^{-\theta}$. Applying OST produces the analogue of (I.1) with $\cosh\theta$ replaced by $\varphi(\theta)$. The result encodes both $\mathcal{P}(S_\tau = A)$ and the Laplace transform of τ simultaneously; the fair case corresponds to $p = q = 1/2$, $\varphi(\theta) = \cosh\theta$.

Exercise I.9. Carry out the analogue of Proposition I.1 for the unfair game with $p \neq 1/2$, obtaining the joint Laplace transform

$$\mathbb{E}\left(\frac{e^{\theta S_\tau}}{\varphi(\theta)^\tau} \right) = 1.$$

Use this together with a second equation (from a different value of θ, or from the $-\theta$ argument if applicable) to express $\mathbb{E}(e^{-s\tau})$ in closed form.

Appendix J

Doob's L^p Maximal Inequality

In Section 10.2 we stated Doob's L^p maximal inequality without proof:

Theorem J.1 (Doob's L^p maximal inequality)**.** *Let $\{X_n\}_{n=0}^N$ be a martingale (or a non-negative submartingale) with $\mathbb{E}|X_n|^p < \infty$ for some $p > 1$. Then*

$$\mathbb{E}\Big(\max_{0\le k\le N} |X_k|^p\Big) \le \Big(\frac{p}{p-1}\Big)^p \mathbb{E}|X_N|^p. \tag{J.1}$$

If $\{X_n\}_{n\ge 0}$ is an L^p-bounded martingale, then letting $N \to \infty$ and using monotone convergence,

$$\mathbb{E}\Big(\sup_{k\ge 0} |X_k|^p\Big) \le \Big(\frac{p}{p-1}\Big)^p \sup_n \mathbb{E}|X_n|^p.$$

We give the proof in this appendix. The argument has two ingredients: the layer-cake (or distribution-function) representation of $\mathbb{E}(Y^p)$ for a non-negative random variable Y, and the strong form of Doob's maximal inequality (the middle inequality in (10.1)), which we proved in Chapter 10.2. The key analytic step is Hölder's inequality at the right moment.

J.1 A lemma on L^p moments: strong maximal inequality

The following lemma isolates the purely analytic part of the argument. It applies to any pair of non-negative random variables satisfying a "strong maximal inequality"; the martingale structure enters only through verifying the hypothesis.

Lemma J.2. *Let $Y, Z \ge 0$ be random variables with $\mathbb{E}(Z^p) < \infty$ for some $p > 1$. Suppose that for every $\lambda > 0$,*

$$\lambda\, \mathbb{P}(Y \ge \lambda) \le \mathbb{E}\big(Z\, \mathbf{1}_{\{Y\ge\lambda\}}\big). \tag{J.2}$$

Then

$$\mathbb{E}(Y^p) \le \Big(\frac{p}{p-1}\Big)^p \mathbb{E}(Z^p). \tag{J.3}$$

Proof. We first give the argument assuming $\mathbb{E}(Y^p) < \infty$; we remove this a posteriori assumption at the end.

Step 1: layer-cake representation. For any non-negative random variable Y and any $p > 0$,

$$\mathbb{E}(Y^p) = \int_0^\infty p\lambda^{p-1}\,\mathbb{P}(Y \geq \lambda)\,d\lambda. \tag{J.4}$$

This is the standard identity, obtained from

$$Y^p = \int_0^Y p\lambda^{p-1}\,d\lambda = \int_0^\infty p\lambda^{p-1}\mathbf{1}_{\{Y\geq\lambda\}}\,d\lambda$$

by taking expectations and applying Fubini's theorem.

Step 2: apply the hypothesis (J.2). Substituting (J.2) into (J.4),

$$\mathbb{E}(Y^p) = \int_0^\infty p\lambda^{p-1}\mathbb{P}(Y \geq \lambda)\,d\lambda \leq \int_0^\infty p\lambda^{p-2}\,\mathbb{E}\big(Z\,\mathbf{1}_{\{Y\geq\lambda\}}\big)\,d\lambda.$$

By Fubini's theorem,

$$\int_0^\infty p\lambda^{p-2}\,\mathbb{E}\big(Z\,\mathbf{1}_{\{Y\geq\lambda\}}\big)\,d\lambda = \mathbb{E}\bigg(Z\int_0^Y p\lambda^{p-2}\,d\lambda\bigg) = \mathbb{E}\Big(Z\cdot\tfrac{p}{p-1}\,Y^{p-1}\Big),$$

where the inner integral $\int_0^Y p\lambda^{p-2}\,d\lambda = \frac{p}{p-1}Y^{p-1}$ uses $p > 1$. Therefore

$$\mathbb{E}(Y^p) \leq \frac{p}{p-1}\,\mathbb{E}\big(Z\,Y^{p-1}\big). \tag{J.5}$$

Step 3: Hölder. Let $q = p/(p-1)$ be the conjugate exponent, so that $q(p-1) = p$. Hölder's inequality applied to (J.5) gives

$$\mathbb{E}(Y^p) \leq \frac{p}{p-1}\,\big(\mathbb{E}(Z^p)\big)^{1/p}\big(\mathbb{E}(Y^{q(p-1)})\big)^{1/q} = \frac{p}{p-1}\,\big(\mathbb{E}(Z^p)\big)^{1/p}\big(\mathbb{E}(Y^p)\big)^{1/q}.$$

Under the assumption $\mathbb{E}(Y^p) < \infty$, we may divide both sides by $(\mathbb{E}(Y^p))^{1/q}$ (if it is zero, (J.3) is trivial). Using $1 - 1/q = 1/p$,

$$\big(\mathbb{E}(Y^p)\big)^{1/p} \leq \frac{p}{p-1}\,\big(\mathbb{E}(Z^p)\big)^{1/p}.$$

Raising both sides to the p-th power gives (J.3).

Step 4: removing the assumption $\mathbb{E}(Y^p) < \infty$. In applications to martingales, Y is a maximum and may a priori fail to be in L^p. We handle this by truncation: replace Y with $Y \wedge M$ for $M > 0$. One checks that the hypothesis (J.2) continues to hold with $Y\wedge M$ in place of Y (since $\{Y\wedge M \geq \lambda\} = \{Y \geq \lambda\}$ for $\lambda \leq M$ and both sides are zero for $\lambda > M$). Since $Y \wedge M \leq M$ is bounded, $\mathbb{E}((Y \wedge M)^p) < \infty$, and the argument above yields

$$\mathbb{E}\big((Y \wedge M)^p\big) \leq \Big(\frac{p}{p-1}\Big)^p\,\mathbb{E}(Z^p).$$

Letting $M \to \infty$ and applying monotone convergence to the left-hand side gives (J.3). □

J.2 Proof of Theorem J.1

Proof of Theorem J.1. Set $Y := \max_{0\le k\le N} |X_k|$ and $Z := |X_N|$. We verify the hypothesis (J.2) of Lemma J.2.

If $\{X_n\}$ is a martingale, then $\{|X_n|\}$ is a non-negative submartingale (Jensen's inequality applied to the convex function $|\cdot|$). Applying the strong form of Doob's maximal inequality — the middle inequality in (10.1) of Theorem 10.8 — to the non-negative submartingale $\{|X_n|\}_{n=0}^N$ with threshold $\lambda > 0$, we obtain

$$\lambda\,\mathbb{P}\Big(\max_{0\le k\le N} |X_k| \ge \lambda\Big) \le \mathbb{E}\Big(|X_N|\,\mathbf{1}_{\{\max_k |X_k|\ge\lambda\}}\Big),$$

which is exactly (J.2) for $Y = \max_{0\le k\le N} |X_k|$ and $Z = |X_N|$.

If $\{X_n\}$ is a non-negative submartingale, the same reasoning applies directly to $\{X_n\}$ itself (without needing to pass through $|\cdot|$), with $Y = \max_{0\le k\le N} X_k$ and $Z = X_N$.

In either case, Lemma J.2 gives

$$\mathbb{E}\Big(\max_{0\le k\le N} |X_k|^p\Big) = \mathbb{E}(Y^p) \le \Big(\frac{p}{p-1}\Big)^p \mathbb{E}(Z^p) = \Big(\frac{p}{p-1}\Big)^p \mathbb{E}|X_N|^p,$$

which is (J.1).

The statement for L^p-bounded martingales $\{X_n\}_{n\ge0}$ follows by letting $N \to \infty$. The sequence $\max_{0\le k\le N} |X_k|^p$ is non-decreasing in N, so monotone convergence gives

$$\mathbb{E}\Big(\sup_{k\ge0} |X_k|^p\Big) = \lim_{N\to\infty} \mathbb{E}\Big(\max_{0\le k\le N} |X_k|^p\Big) \le \Big(\frac{p}{p-1}\Big)^p \sup_n \mathbb{E}|X_n|^p.$$

□

Remark J.3 (On the constant). The constant $p/(p-1)$ in Lemma J.2 is **sharp**: there exist martingales for which equality is approached in (J.1) up to an arbitrarily small factor. Note that $p/(p-1) \to \infty$ as $p \to 1^+$, reflecting the fact that the L^1 analogue of Doob's maximal inequality fails: for $p = 1$, an L^1-bounded martingale need not have an integrable maximal function. The correct replacement at $p = 1$ is the $L \log L$ inequality of Doob, which states that if $\mathbb{E}(|X_N| \log^+ |X_N|) < \infty$, then $\mathbb{E}(\max_{k\le N} |X_k|) \le C(1 + \mathbb{E}(|X_N| \log^+ |X_N|))$ for a universal constant C.

Remark J.4 (What the strong maximal inequality gives us). The middle inequality $\mathbb{P}(E) \le \frac{1}{\lambda}\mathbb{E}(\mathbf{1}_E X_N)$ in Theorem 10.8 — not the weaker $\mathbb{P}(E) \le \mathbb{E}(X_N)/\lambda$ — is essential for this argument. The layer-cake integration produces $\mathbb{E}(Z\,Y^{p-1})$, and it is the random-variable-valued right-hand side $\mathbb{E}(Z\,\mathbf{1}_E)$ that yields this structure after Fubini. The weak inequality $\lambda\mathbb{P}(E) \le \mathbb{E}(Z)$ would only produce $\mathbb{E}(Z) \cdot \int_0^\infty p\lambda^{p-2}\,d\lambda$, a divergent integral.

Appendix K

L^2 Convergence for Submartingales

K.1 L^2 convergence of non-negative submartingales

The L^2 martingale convergence theorem (Theorem 11.1) extends to L^2-bounded submartingales. Because a supermartingale is the negative of a submartingale, the supermartingale version follows automatically. We give here a short proof for the case of **non-negative** submartingales, which covers the most common applications (for instance, when one applies the theorem to $|M_n|^2$ for an L^2-bounded martingale M_n, or to a non-negative supermartingale regarded through its negative).

Theorem K.1 (L^2 convergence for non-negative submartingales)**.** *Let $\{X_n\}$ be a non-negative submartingale with $\sup_n \mathbb{E}(X_n^2) \le B < \infty$. Then there exists a random variable $X_\infty \in L^2$ such that*

$$X_n \xrightarrow{a.s.} X_\infty \qquad \text{and} \qquad X_n \xrightarrow{L^2} X_\infty.$$

Proof. Step 1: almost sure convergence. By the Cauchy–Schwarz inequality,

$$\sup_n \mathbb{E}(X_n) \le \sup_n \big(\mathbb{E}(X_n^2)\big)^{1/2} \le \sqrt{B},$$

so $\{X_n\}$ is L^1-bounded. By Doob's upcrossing convergence theorem (Theorem 11.6) applied to the submartingale $\{X_n\}$, there exists $X_\infty \in L^1$ with $X_n \xrightarrow{a.s.} X_\infty$. Non-negativity of X_n forces $X_\infty \ge 0$ a.s.

Step 2: the maximal function is in L^2. Since $\{X_n\}$ is a non-negative submartingale, so is $\{X_n^2\}$: the function $\varphi(x) = x^2$ is convex and non-decreasing on $[0,\infty)$, so Jensen's inequality (for conditional expectations) gives

$$\mathbb{E}(X_{n+1}^2 \mid \mathcal{F}_n) \ge \varphi\big(\mathbb{E}(X_{n+1} \mid \mathcal{F}_n)\big) \ge \varphi(X_n) = X_n^2.$$

Applying the L^p form of Doob's maximal inequality (the remark after Corollary 10.9) with $p = 2$ — the statement there is for martingales, but the proof uses only that $\{|X_k|^p\}$ is a non-negative submartingale, so it applies verbatim to our non-negative submartingale $\{X_k\}$ — we obtain

$$\mathbb{E}\Big(\max_{0\le k\le n} X_k^2\Big) \le 4\,\mathbb{E}(X_n^2) \le 4B.$$

The sequence $\max_{0\le k\le n} X_k^2$ is non-decreasing in n, so monotone convergence gives

$$\mathbb{E}\Big(\sup_{k\ge 0} X_k^2\Big) \le 4B. \tag{K.1}$$

In particular, the maximal function $X^* := \sup_{k\ge 0} X_k$ belongs to L^2.

Step 3: L^2 convergence. From $0 \le X_n \le X^*$ for all n, passing to the a.s. limit, $0 \le X_\infty \le X^*$ a.s., and therefore

$$|X_n - X_\infty|^2 \le \big(X_n + X_\infty\big)^2 \le (2X^*)^2 = 4(X^*)^2.$$

By (K.1), the dominating random variable $4(X^*)^2$ is integrable. Since $X_n \to X_\infty$ a.s., the dominated convergence theorem yields

$$\mathbb{E}\big(|X_n - X_\infty|^2\big) \longrightarrow 0,$$

i.e. $X_n \xrightarrow{L^2} X_\infty$. In particular, $X_\infty \in L^2$ with $\mathbb{E}(X_\infty^2) \le 4B$. □

Corollary K.2 (L^2 convergence for non-positive supermartingales)**.** *If $\{Y_n\}$ is a non-positive supermartingale with $\sup_n \mathbb{E}(Y_n^2) < \infty$, then Y_n converges a.s. and in L^2 to a limit $Y_\infty \in L^2$.*

Proof. Apply Theorem K.1 to $X_n := -Y_n$, which is a non-negative submartingale with $\mathbb{E}(X_n^2) = \mathbb{E}(Y_n^2)$. □

Remark K.3. It is instructive to observe why the proof of Theorem K.1 does not extend directly to **signed** submartingales. The argument used that $\{X_n\}$ is a non-negative submartingale so that $\{X_n^2\}$ is a submartingale (via Jensen with $\varphi(x) = x^2$ on $[0,\infty)$, which is convex and **non-decreasing**). For a general submartingale, $\varphi(x) = x^2$ is still convex on $\mathbb{R}$ but not monotone, and $\{X_n^2\}$ need not be a submartingale. Consequently, Doob's maximal inequality cannot be applied to $\{X_n^2\}$, and the dominator $X^* = \sup_k X_k$ in the main step above is no longer controlled in L^2 by $\sup_n \mathbb{E}(X_n^2)$.

The correct extension to signed submartingales goes through the Doob decomposition, as we now show.

K.2 The general L^2-bounded submartingale

Theorem K.4 (L^2 convergence, general submartingale)**.** *Let $\{X_n\}$ be a submartingale with $\sup_n \mathbb{E}(X_n^2) \le B < \infty$. Then there exists $X_\infty \in L^2$ with $X_n \xrightarrow{a.s.} X_\infty$ and $X_n \xrightarrow{L^2} X_\infty$.*

The proof uses the Doob decomposition and reduces to a single technical lemma controlling the predictable increasing part in L^2.

The Doob decomposition

By replacing X_n with $X_n - X_0$, we may assume $X_0 = 0$; this costs nothing since X_0 is a fixed random variable in L^2 and the convergence of X_n is equivalent to the convergence of $X_n - X_0$. Write the Doob decomposition as

$$X_n = M_n + A_n,$$

where $M_0 = A_0 = 0$, $\{M_n\}$ is a martingale, and $\{A_n\}$ is predictable non-decreasing with increments

$$\Delta A_k = \mathbb{E}(\Delta X_k \mid \mathcal{F}_{k-1}) \geq 0.$$

Structure of the proof

The crux is the following lemma.

Lemma K.5. $\mathbb{E}(A_\infty^2) \leq 4B$, *and in particular* $A_n \nearrow A_\infty$ *with* $A_\infty \in L^2$.

Taking Lemma K.5 for granted, we complete the proof of Theorem K.4.

Proof of Theorem K.4 given Lemma K.5. Since A_n is non-decreasing, MCT gives $A_n \nearrow A_\infty$ a.s., and Lemma K.5 yields $A_\infty \in L^2$. The dominator A_∞^2 is integrable and $0 \leq A_n \leq A_\infty$, so dominated convergence gives

$$\mathbb{E}\big((A_n - A_\infty)^2\big) \to 0, \qquad \text{i.e. } A_n \xrightarrow{L^2} A_\infty.$$

For the martingale part, $M_n = X_n - A_n$ and hence

$$\|M_n\|_2 \leq \|X_n\|_2 + \|A_n\|_2 \leq \sqrt{B} + \|A_\infty\|_2 \leq \sqrt{B} + 2\sqrt{B} = 3\sqrt{B},$$

uniformly in n. Thus $\{M_n\}$ is L^2-bounded, and Theorem 11.1 gives $M_\infty \in L^2$ with $M_n \xrightarrow{a.s.} M_\infty$ and $M_n \xrightarrow{L^2} M_\infty$.

Setting $X_\infty := M_\infty + A_\infty \in L^2$,

$$X_n = M_n + A_n \xrightarrow{a.s.} M_\infty + A_\infty = X_\infty,$$

and by Minkowski,

$$\|X_n - X_\infty\|_2 \leq \|M_n - M_\infty\|_2 + \|A_n - A_\infty\|_2 \to 0.$$

□

Proof of Lemma K.5

Expanding $X_n^2 = (M_n + A_n)^2$ and taking expectations,

$$\mathbb{E}(X_n^2) = \mathbb{E}(M_n^2) + 2\mathbb{E}(M_n A_n) + \mathbb{E}(A_n^2). \tag{K.2}$$

The cross term $\mathbb{E}(M_n A_n)$ admits the following identity. By summation by parts,

$$M_n A_n = \sum_{k=1}^{n} A_{k-1}\Delta M_k + \sum_{k=1}^{n} M_k \Delta A_k.$$

Since A_{k-1} is $\mathcal{F}_{k-1}$-measurable and $\mathbb{E}(\Delta M_k \mid \mathcal{F}_{k-1}) = 0$, the first sum has expectation 0. For the second, the tower property and predictability of ΔA_k give $\mathbb{E}(M_k \Delta A_k) = \mathbb{E}(M_{k-1} \Delta A_k)$. Hence

$$\mathbb{E}(M_n A_n) = \sum_{k=1}^{n} \mathbb{E}(M_{k-1} \Delta A_k). \tag{K.3}$$

The substantive step is to bound the right-hand side of (K.3). By the Cauchy–Schwarz inequality, applied to the measure $\sum_{k=1}^{n} \Delta A_k \cdot d\mathbb{P}$ on $\{1, \dots, n\} \times \Omega$,

$$\left| \sum_{k} \mathbb{E}(M_{k-1} \Delta A_k) \right| \le \left(\sum_{k} \mathbb{E}(M_{k-1}^2 \, \Delta A_k) \right)^{1/2} \big(\mathbb{E}(A_n)\big)^{1/2}. \tag{K.4}$$

Using Doob's L^2 maximal inequality applied to the martingale M, one shows (the technical step, below) that

$$\sum_{k=1}^{n} \mathbb{E}(M_{k-1}^2 \, \Delta A_k) \le 4 \mathbb{E}(M_n^2) \cdot \mathbb{E}(A_n). \tag{K.5}$$

Combining (K.4) and (K.5),

$$|\mathbb{E}(M_n A_n)| \le 2 \|M_n\|_2 \, \mathbb{E}(A_n).$$

Substituting into (K.2) and using $\mathbb{E}(M_n^2) \ge 0$,

$$\mathbb{E}(A_n^2) \le \mathbb{E}(X_n^2) + 4 \|M_n\|_2 \, \mathbb{E}(A_n) \le B + 4\sqrt{B} \, \mathbb{E}(A_n).$$

Since $\mathbb{E}(A_n) = \mathbb{E}(X_n) - \mathbb{E}(M_n) = \mathbb{E}(X_n) \le \sqrt{B}$, this yields $\mathbb{E}(A_n^2) \le 5B$ uniformly in n. By MCT applied to the non-decreasing sequence A_n^2,

$$\mathbb{E}(A_\infty^2) = \lim_{n} \mathbb{E}(A_n^2) \le 5B.$$

Remark K.6 (On the technical step (K.5)). The inequality (K.5) is the genuine content of the argument; it expresses the fact that the martingale M cannot concentrate excessively on the time points where A grows. A standard proof uses Doob's L^2 maximal inequality in the form $\mathbb{E}(\sup_{k \le n} M_k^2) \le 4\mathbb{E}(M_n^2)$ together with a Fubini-type rearrangement of $\sum_k \mathbb{E}(M_{k-1}^2 \Delta A_k)$. The full details can be found in Durrett (2019), §4.4. The constant 4 in (K.5) is not sharp; what matters for our purpose is only that the right-hand side factors as a product involving $\mathbb{E}(M_n^2)$ and $\mathbb{E}(A_n)$.

Example K.7 (When the non-negative case suffices). Let $\{M_n\}$ be an L^2-bounded martingale. Then $\{|M_n|\}$ is a non-negative submartingale (Jensen applied to $|\cdot|$) with $\sup_n \mathbb{E}(|M_n|^2) = \sup_n \mathbb{E}(M_n^2) < \infty$. Theorem K.1 applied to $|M_n|$ gives $|M_n| \to |M_\infty|$ a.s. and in L^2 — a result already implied by the martingale convergence theorem, but obtained here without ever invoking orthogonality of differences.

Appendix L

The Pólya Urn Martingale and Its Limit

In Example 9.12 we introduced the Pólya urn martingale $M_n = R_n/(n+2)$ and asserted — without proof — that $M_\infty \sim \text{Uniform}(0,1)$. In this appendix we prove this statement, and more generally, we identify the limit distribution for an urn with an arbitrary initial composition.

The argument we give is based on **exchangeability** of the sequence of draws, a concept that we will revisit in Chapter 12 in connection with backward martingales and the Hewitt–Savage 0–1 law. The end of the argument uses the method of moments and a direct computation involving the Beta function; we do not invoke the full de Finetti representation theorem, but the structure of the proof is exactly the de Finetti-style reasoning that this theorem abstracts.

L.1 The general Pólya urn

We consider the urn in slightly greater generality than in Example 9.12. Let $a, b \geq 1$ be integers. The urn starts with a red balls and b blue balls. At each step, a ball is drawn uniformly at random, then returned to the urn together with one new ball of the same color. Let R_n denote the number of red balls after n draws, so that $R_0 = a$, the total number of balls after n draws is $a + b + n$, and the urn in Example 9.12 corresponds to $a = b = 1$.

Set

$$M_n := \frac{R_n}{a+b+n}.$$

The same calculation as in Example 9.12 shows that $\{M_n\}$ is a martingale. Since $M_n \in [0,1]$, it is L^1-bounded (in fact uniformly bounded), so by Doob's convergence theorem there exists $M_\infty \in [0,1]$ with $M_n \xrightarrow{a.s.} M_\infty$; bounded convergence upgrades this to L^1 (and indeed L^p for every $p \geq 1$) convergence.

The goal of this appendix is to prove:

Theorem L.1. *The limit M_∞ is* Beta(a,b)*-distributed, i.e. it has density*

$$f(x) = \frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}, \qquad x \in (0,1),$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function.

Corollary L.2. *In the urn of Example 9.12 (initial composition $a = b = 1$), $M_\infty \sim \mathrm{Uniform}(0,1)$.*

Proof of corollary. $\mathrm{Beta}(1,1)$ has density $x^0(1-x)^0/B(1,1) = 1$ on $(0,1)$. □

L.2 Exchangeability of the sequence of draws

Let $X_i := \mathbf{1}\{\text{the } i\text{-th draw is red}\}$, so that $R_n = a + \sum_{i=1}^n X_i$. The key structural fact about the Pólya urn is that the sequence $(X_1, X_2, \ldots)$ is **exchangeable**.

Definition L.3. A finite sequence of random variables $(Y_1, \ldots, Y_n)$ is **exchangeable** if for every permutation σ of $\{1, \ldots, n\}$,

$$(Y_{\sigma(1)}, \ldots, Y_{\sigma(n)}) \overset{d}{=} (Y_1, \ldots, Y_n).$$

An infinite sequence $(Y_1, Y_2, \ldots)$ is exchangeable if every finite initial segment is exchangeable.

Lemma L.4 (Exchangeability of draws)**.** *For any $n \geq 1$ and any $(x_1, \ldots, x_n) \in \{0,1\}^n$ with $k := \sum_i x_i$,*

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{a(a+1)\cdots(a+k-1)\cdot b(b+1)\cdots(b+n-k-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)}. \tag{L.1}$$

In particular, the probability depends only on k and not on the order of the x_i's. Hence $(X_1, X_2, \ldots)$ is exchangeable.

Proof. We prove (L.1) by unraveling the conditional probabilities. By the chain rule,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i \mid X_1, \ldots, X_{i-1}).$$

Given $X_1, \ldots, X_{i-1}$ with $k_{i-1} := \sum_{j \leq i-1} x_j$ red draws among the first $i-1$, the urn at step i contains $a + k_{i-1}$ red balls out of $a+b+i-1$ total. Therefore

$$\mathbb{P}(X_i = 1 \mid X_1, \ldots, X_{i-1}) = \frac{a + k_{i-1}}{a+b+i-1},$$
$$\mathbb{P}(X_i = 0 \mid X_1, \ldots, X_{i-1}) = \frac{b + (i-1-k_{i-1})}{a+b+i-1}.$$

Multiplying over $i = 1, \ldots, n$: the denominator is $(a+b)(a+b+1)\cdots(a+b+n-1)$, and the numerator is the product of terms of two types. Each step where $x_i = 1$ contributes $a + k_{i-1}$, and these increment by 1 each time $x_j = 1$, giving the product $a(a+1)\cdots(a+k-1)$ after all k red draws. Similarly, the $n-k$ blue draws contribute $b(b+1)\cdots(b+n-k-1)$. This yields (L.1).

The right-hand side of (L.1) depends only on $k = \sum x_i$, so $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is invariant under permutations of the arguments. Hence $(X_1, \ldots, X_n)$ is exchangeable, and since n was arbitrary, the infinite sequence $(X_1, X_2, \ldots)$ is exchangeable. □

Rewriting (L.1) using the identity $a(a+1)\cdots(a+k-1) = \Gamma(a+k)/\Gamma(a)$ and likewise for b:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \frac{\Gamma(a+k)\,\Gamma(b+n-k)\,\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)\,\Gamma(a+b+n)} = \frac{B(a+k, b+n-k)}{B(a,b)}. \tag{L.2}$$

L.3 Identifying the limit distribution

We now use the exchangeability identity (L.2) together with the law of large numbers to determine the distribution of M_∞.

Proof of Theorem L.1. Step 1: the limit is the limiting frequency of red draws. Since $R_n = a + \sum_{i=1}^n X_i$,

$$M_n = \frac{a + \sum_{i=1}^n X_i}{a+b+n} = \frac{a}{a+b+n} + \frac{n}{a+b+n} \cdot \frac{1}{n} \sum_{i=1}^n X_i.$$

The first term tends to 0 and the factor $n/(a+b+n) \to 1$, so

$$M_\infty = \lim_{n\to\infty} \frac{S_n}{n} \qquad \text{a.s.,} \qquad \text{where } S_n := \sum_{i=1}^n X_i.$$

(The limit on the right exists a.s. because the one on the left does; we do not need SLLN here because convergence is given.)

Step 2: a moment formula via the exchangeability identity. Fix integers $k \geq 0$ and $\ell \geq 0$, and consider the specific outcome in which the first k draws are red and the next ℓ are blue:

$$p_{k,\ell} := \mathbb{P}(X_1 = 1, \dots, X_k = 1,\ X_{k+1} = 0, \dots, X_{k+\ell} = 0).$$

By (L.2),

$$p_{k,\ell} = \frac{B(a+k,\ b+\ell)}{B(a,b)} = \frac{1}{B(a,b)} \int_0^1 x^{a+k-1}(1-x)^{b+\ell-1}\,dx. \tag{L.3}$$

Step 3: expressing $p_{k,\ell}$ in terms of M_∞. We claim

$$\mathbb{E}\big(M_\infty^k (1-M_\infty)^\ell\big) = p_{k,\ell} \tag{L.4}$$

for all integers $k, \ell \geq 0$ with $k + \ell \geq 1$. To see this, observe that by exchangeability, for any $n \geq k+\ell$,

$$\mathbb{E}\left(\binom{S_n}{k} \binom{n-S_n}{\ell} \right) = \binom{n}{k+\ell} \binom{k+\ell}{k} \cdot p_{k,\ell},$$

because the left-hand side counts, with weight, the number of ordered pairs (I, J) of disjoint subsets $I, J \subset \{1, \dots, n\}$ with $|I| = k$, $|J| = \ell$, for which $X_i = 1$ on I and $X_j = 0$ on J; by exchangeability each such pair contributes $p_{k,\ell}$, and the number of pairs is $\binom{n}{k}\binom{n-k}{\ell} = \binom{n}{k+\ell}\binom{k+\ell}{k}$.

Now divide both sides by $n^{k+\ell}$. The identity $\binom{n}{r}/n^r \to 1/r!$ as $n \to \infty$ gives

$$\frac{1}{n^{k+\ell}}\binom{n}{k+\ell}\binom{k+\ell}{k} \longrightarrow \frac{1}{k!\,\ell!}.$$

Meanwhile,

$$\frac{1}{n^{k+\ell}}\binom{S_n}{k}\binom{n-S_n}{\ell} = \frac{1}{k!\,\ell!}\left(\frac{S_n}{n}\right)^k\left(\frac{n-S_n}{n}\right)^\ell + O(n^{-1}),$$

where the error is uniform in ω (since $S_n/n \in [0,1]$). Using $S_n/n \to M_\infty$ a.s., the bounded convergence theorem gives

$$\frac{1}{n^{k+\ell}}\,\mathbb{E}\left(\binom{S_n}{k}\binom{n-S_n}{\ell}\right) \longrightarrow \frac{1}{k!\,\ell!}\,\mathbb{E}\big(M_\infty^k(1-M_\infty)^\ell\big).$$

Equating limits yields (L.4).

Step 4: conclusion via moments. Combining (L.4) with (L.3),

$$\begin{aligned}\mathbb{E}\big(M_\infty^k(1-M_\infty)^\ell\big) &= \frac{1}{B(a,b)}\int_0^1 x^{a+k-1}(1-x)^{b+\ell-1}\,dx \\ &= \int_0^1 x^k(1-x)^\ell\cdot\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\,dx,\end{aligned}$$

which is precisely $\mathbb{E}\big(Y^k(1-Y)^\ell\big)$ for $Y \sim \mathrm{Beta}(a,b)$.

Setting $\ell = 0$ we obtain $\mathbb{E}(M_\infty^k) = \mathbb{E}(Y^k)$ for all $k \geq 0$. Since M_∞ and Y are both supported on $[0,1]$, a distribution on a bounded interval is uniquely determined by its moments (by Weierstrass approximation: polynomials are dense in $C([0,1])$), so M_∞ and Y have the same distribution. Hence $M_\infty \sim \mathrm{Beta}(a,b)$. □

Remark L.5 (Connection to de Finetti's theorem). The argument above is a hands-on instance of de Finetti's representation theorem, which states that every infinite exchangeable sequence of $\{0,1\}$-valued random variables is a **mixture** of i.i.d. Bernoulli sequences: there exists a random variable $\Theta \in [0,1]$ such that conditionally on Θ, the X_i's are i.i.d. Bernoulli(Θ). The limiting frequency $S_n/n \to \Theta$ a.s. then gives $\Theta = M_\infty$. In the Pólya urn we have **computed** the mixing distribution explicitly: it is $\mathrm{Beta}(a,b)$.

Remark L.6 (Direct proof for the case $a = b = 1$). For the original urn of Example 9.12 ($a = b = 1$), one can avoid the moment computation entirely by observing that (L.1) specialized to $a = b = 1$ gives

$$\mathbb{P}(S_n = k) = \binom{n}{k}\frac{k!\,(n-k)!}{(n+1)!} = \frac{1}{n+1}, \qquad k = 0,1,\ldots,n.$$

So S_n is uniform on $\{0,1,\ldots,n\}$, and $R_n = 1+S_n$ is uniform on $\{1,2,\ldots,n+1\}$. Hence $M_n = R_n/(n+2)$ is uniformly distributed on $\{1/(n+2), 2/(n+2), \ldots, (n+1)/(n+2)\}$, and as $n \to \infty$ this discrete uniform distribution converges in distribution to $\mathrm{Uniform}(0,1)$. Combined with the almost sure convergence $M_n \to M_\infty$, this identifies $M_\infty \sim \mathrm{Uniform}(0,1)$.

Example L.7 (A forecast interpretation). Theorem L.1 has the following striking interpretation. The fraction M_n of red balls is a martingale — a "fair forecast" of the long-run composition. The theorem says that this forecast converges to an a.s. limit M_∞, and **marginally**, M_∞ has a non-trivial distribution (uniform on $[0,1]$ when $a = b = 1$). So a priori the long-run fraction is maximally uncertain; conditionally on observing the first n draws, our best estimate updates to $R_n/(a+b+n)$. The martingale property says these forecasts are "calibrated" at each step.

Hints to Selected Exercises

Hint for Exercise Exercise 1.41. (Use π–λ on a **countable** generator).

Hint for Exercise Exercise 1.43. If a countably additive extension $\tilde{\mu}$ existed, then for any $q \in \Omega$ the sets $E_n := (q - 1/n, q]_{\mathbb{Q}}$ decrease to $\{q\}$, so $\tilde{\mu}(\{q\}) = \lim_n \tilde{\mu}(E_n) = 0$; hence $\tilde{\mu}(\Omega) = \sum_{q\in\Omega} \tilde{\mu}(\{q\}) = 0$, contradicting $\tilde{\mu}(\Omega) = 1$.

Hint for Exercise 2.26 From convergence in probability extract an a.s. convergent subsequence; apply DCT to the subsequence; then argue that the whole sequence has the same limit.

Hint for Exercise 2.31 For $f \geq 0$ consider $f_n := f \wedge n$ and use MCT on A (or on Ω with $\mathbf{1}_A$).

Hint for Exercise 2.36 For $1 < p < \infty$ let q be conjugate to p and start with $\mathbb{E}|X + Y|^p = \mathbb{E}(|X + Y| \cdot |X + Y|^{p-1})$.

Hint for Exercise 2.37 Apply Hölder to $|f|^q \cdot 1$ with exponents $\frac{p}{q}$ and $\frac{p}{p-q}$.

Hint for Exercise 2.40 For (3): If Y is a dominator, estimate $\int_0^{1/n} Y\, dP$ and find a contradiction with the absolute integrability of the Lebesgue integral.

Hint for Exercise 2.41 Define $N(K) := \min\{n \geq 1 :\ 4^n/n > K\}$, and show that for both $n \geq N(K)$ and $n < N(K)$, $\int |f_n| \mathbf{1}_{\{|f_n|>K\}}\, dx \leq \frac{1}{N(K)}$. Then show that $N(K) \to \infty$ as $K \to \infty$.

Hint for Exercise 3.20 Given $M = m$, decompose according to whether X or Y achieved the maximum. Use symmetry to determine the probability of each case. Note that this problem does **not** fit the standard conditional density framework because the joint distribution of (X, M) has a singular component along the diagonal $\{X = M\}$.

Hint for Exercise 3.21 For (b), use the finite partition formula: on each atom A_i of the partition, $\mathbb{E}[N \mid R]$ equals $\mathbb{E}[N\mathbf{1}_{A_i}]/\mathbb{P}(A_i)$.

Hint for Exercise 3.22 Write $X = (X - \mathbb{E}[X \mid \mathcal{G}]) + \mathbb{E}[X \mid \mathcal{G}]$ and use the L^2-orthogonality from Theorem 3.13. Alternatively, expand $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ and apply the tower property.

Hint for Exercise 3.23 (a) Differentiate with respect to a and b, or use projection onto the two-dimensional subspace $\mathrm{span}\{1, Y\} \subset L^2$.

(b) The conditional expectation minimizes over **all** $\sigma(Y)$-measurable predictors, while the best linear predictor minimizes over a smaller class.

(c) Try Y uniform on $\{-1, 1\}$ and $X = Y^2$.

(d) For jointly Gaussian variables, $L^2(\sigma(Y))$ is spanned by $\{1, Y\}$.

Hint for Exercise 3.24 Set $Z_n := \mathbb{E}[X_n \mid \mathcal{G}]$. Then (Z_n) is increasing and $Z_n \leq \mathbb{E}[X \mid \mathcal{G}]$ by monotonicity. Let $Z := \lim_n Z_n$. To show $Z = \mathbb{E}[X \mid \mathcal{G}]$,

verify the defining property: for $A \in \mathcal{G}$,

$$\mathbb{E}[Z\mathbf{1}_A] = \lim_n \mathbb{E}[Z_n\mathbf{1}_A] = \lim_n \mathbb{E}[X_n\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A],$$

using (unconditional) MCT twice.

Hint for Exercise 3.25 Define $Y_k := \inf_{n\geq k} X_n$. Then $Y_k \uparrow \liminf X_n$ and $Y_k \leq X_n$ for all $n \geq k$. Apply Exercise 3.24 to (Y_k).

Hint for Exercise 3.26 Apply conditional Fatou to $Y + X_n$ and $Y - X_n$.

Hint for Exercise 3.27 For ($\Rightarrow$): if $X \perp\!\!\!\perp \mathcal{G}$, then for any $A \in \mathcal{G}$, the random variables X and $\mathbf{1}_A$ are independent, so $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[X]\mathbb{P}(A)$.

For ($\Leftarrow$): show that $\mathbb{E}[\mathbf{1}_{\{X\in B\}}\mathbf{1}_A] = \mathbb{P}(X \in B)\mathbb{P}(A)$ for all $A \in \mathcal{G}$ and $B \in \mathcal{B}(\mathbb{R})$. Use the test-function characterization and a monotone class argument.

Hint for Exercise 3.28 Let $S = X + Y$. Find the joint density of (X, S) by the change of variables $(x, y) \mapsto (x, x + y)$. The marginal density of S is the convolution of two exponentials. Apply the conditional density formula.

Hint for Exercise 3.29 Let $a := \mathrm{Cov}(X, Y)/\mathbb{V}\mathrm{ar}(Y)$ and set $Z := \mathbb{E}[X] + a\,(Y - \mathbb{E}[Y])$, which is $\sigma(Y)$-measurable. Show that $X - Z$ is uncorrelated with Y. For jointly Gaussian vectors, uncorrelatedness implies independence, so $X - Z \perp\!\!\!\perp Y$. Then for any bounded Borel g,

$$\mathbb{E}[(X - Z)g(Y)] = \mathbb{E}[X - Z]\,\mathbb{E}[g(Y)] = 0,$$

so by the test-function characterization $Z = \mathbb{E}[X \mid Y]$.

Hint for Exercise 3.30 By exchangeability, $\mathbb{E}[X_i \mid X_1 + \cdots + X_n]$ has the same distribution for all i, hence equals the same function of the sum. Now use linearity and the pull-out property.

Hint for Exercise 3.31 Set $U := \mathbb{E}[X \mid \mathcal{G}]$ and $V := \mathbb{E}[U \mid \mathcal{H}]$. For $A \in \mathcal{H}$, we have $A \in \mathcal{G}$, so $\mathbb{E}[V\mathbf{1}_A] = \mathbb{E}[U\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$.

Hint for Exercise 3.32 A $\mathcal{G}$-measurable random variable is constant on each atom A_i, so $\mathbb{E}[X \mid \mathcal{G}] = \sum_i c_i\mathbf{1}_{A_i}$. Plug this into $\mathbb{E}[(\mathbb{E}[X \mid \mathcal{G}])\mathbf{1}_{A_j}] = \mathbb{E}[X\mathbf{1}_{A_j}]$ and solve for c_j.

Hint for Exercise 3.34 If $Y_n \in L^2(\mathcal{G})$ and $Y_n \to Y$ in L^2, extract a subsequence converging a.s. (using convergence in probability and the subsequence principle). The a.s. limit of $\mathcal{G}$-measurable functions is $\mathcal{G}$-measurable.

Hint for Exercise 3.35 For (a), use the finite partition formula. For (b), try $Y = \mathbf{1}_{\{1,3\}}$.

Hint for Exercise 3.36 Use the test-function characterization with $Y\mathbf{1}_A$.

Hint for Exercise 3.37 Apply Theorem 3.11 to the convex function $\varphi(t) = |t|^p$, take the expectation of the result and use the tower property of conditional expectation.

Hint for Exercise 3.38 Apply Example 3.17 to the partition $\{Y = y_k\}$ (first finite truncations, then pass to the limit), or verify the defining property directly on sets $A \in \sigma(Y)$, which are unions of $\{Y = y_k\}$.

Hint for Exercise 4.26 First show that $A \in \mathcal{T}$ by expressing convergence in terms of the Cauchy criterion and verifying that the resulting event is independent of any finite initial segment.

Hint for Exercise 4.27

- By continuity, a.s. all values $X_1, \dots, X_n$ are distinct. Among all $n!$ equally likely orderings of distinct values, how many have X_n as the largest?

- Use the Hewitt–Savage 0–1 law.

- To rule out $\mathbb{P}(E) = 0$, show that for any N, there is positive probability that X_{N+1} exceeds $\max(X_1, \dots, X_N)$.

Hint for Exercise 4.28 Use Dynkin's π–λ theorem twice.

Hint for Exercise 4.29

Fix $\varepsilon \in (0,1)$. Apply Borel–Cantelli to $A_n^+ := \{X_n > (1+\varepsilon)\log n\}$ and $A_n^- := \{X_n > (1-\varepsilon)\log n\}$ separately.

Hint for Exercise 5.35 Recall that $\|Y\|_\infty = \inf\{M : \mathbb{P}(|Y| \le M) = 1\}$ and conclude that $\mathbb{P}(|X_n - X| > \varepsilon) = 0$ for all large n.

Hint for Exercise 5.39 For $\varepsilon > 0$, use the open set $(c-\varepsilon, c+\varepsilon)$ and Portmanteau.

Hint for Exercise 6.9 Compute $\mathbb{E}[X_n]$ and $\mathbb{V}\mathrm{ar}(X_n)$, then verify the variance summability condition $\sum_{n=1}^{\infty} \frac{\mathbb{V}\mathrm{ar}(X_n)}{n^2} < \infty$.

Hint for Exercise 6.10 Apply the Three-Series Theorem. Since the ε_n are bounded, truncation is trivial. Compute the variance of $a_n\varepsilon_n$.

Hint for Exercise 6.11 Define $X_n^{(j)} = \mathbf{1}_{\{d_n = j\}} - \frac{1}{10}$. Show these are independent with mean zero, then apply SLLN.

Hint for Exercise 6.12 For the converse, use the second Borel–Cantelli lemma to show that if $\mathbb{E}[|X_1|] = \infty$, then $|X_n| > n$ infinitely often, which contradicts convergence of $\bar{X}_n$. (See also Exercise 6.18.)

Hint for Exercise 6.13 Follow the proof for the special case $b_n = n$: let $S_n = \sum_{k=1}^{n} \frac{a_k}{b_k}$, write $a_k = b_k(S_k - S_{k-1})$, apply summation by parts, then use a weighted Cesàro argument.

Hint for Exercise 6.14 Use Kolmogorov's Three-Series Theorem with truncation level $c = 1$. Check each of the three conditions separately.

Hint for Exercise 6.15 Use the layer-cake identity $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > t)\,dt$ and compare the integral with a sum over $[n, n+1)$.

Hint for Exercise 6.16 Write $\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n} \cdot \frac{S_{n-1}}{n-1}$.

Hint for Exercise 6.18 Prove that $\mathbb{E}|X_1| \le \sum_{n=0}^{\infty} \mathbb{P}\big(|X_1| > n\big)$ and use Borel–Cantelli II.

Hint for Exercise 6.19 For part (b), consider a distribution with $P(|Y_1| > t) \sim t^{-\alpha}$ for $1 < \alpha < 2$. This gives finite mean but infinite variance.

Hint for Exercise 6.20 (i) Check $\mathbb{E}|X_1| = \infty$. (ii) Prove the lemma: if $S_n/n \to L \in \mathbb{R}$, then $X_n/n \to 0$. (iii) Show $\sum_{n\ge1} \mathbb{P}(|X_1| \ge n) = \infty$, hence $|X_n| \ge n$ i.o. by Borel–Cantelli, so $X_n/n \not\to 0$ a.s.

Hint for Exercise 6.21 Use characteristic functions. The Cauchy distribution has characteristic function $\varphi(t) = e^{-|t|}$. Compute the characteristic function of $\bar{X}_n$.

Hint for Exercise 6.22 Truncate at level $k^{1/p}$: define $Y_k = X_k \mathbf{1}_{\{|X_k| \le k^{1/p}\}}$. Show that the truncation error is negligible using Borel–Cantelli, then apply variance summability to the truncated variables.

Hint for Exercise 6.23 Define $Z_n = w_n X_n$ and apply the variance summability SLLN with the sequence $(b_n) = (W_n)$ in a generalized Kronecker lemma.

Hint for Exercise 6.24 Split the event according to the first time k that $|S_k| \geq 3t$, and compare S_k to the "rest" $S_n - S_k$.

Hint for Exercise 6.25 The random variables $f(U_k)$ are i.i.d. with mean $\int_0^1 f(x)\,dx$ and finite variance (by the square-integrability assumption). Apply the variance summability SLLN.

Hint for Exercise 6.26 For Uniform$[-a, a]$, we have mean zero and variance $a^2/3$. Apply the Three-Series Theorem—the truncation and mean conditions are automatic since the variables are bounded and symmetric.

Hint for Exercise 7.23 Write X_n as a sum of n independent Bernoulli(λ/n) random variables and use the product formula. For (b), recall the standard limit $(1 + z/n)^n \to e^z$.

Hint for Exercise 7.24 Use the necessary conditions for characteristic functions (basic properties, moment–derivative relations) to rule out candidates, and sufficient conditions (product formula, Pólya's criterion) to confirm candidates.

Hint for Exercise 7.25 For (a), write $\varphi(t) = \sum_{j\in\mathbb{Z}} \mathbb{P}(X = j)\, e^{ijt}$ and use the orthogonality relation $\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-k)t}\,dt = \mathbf{1}_{\{j=k\}}$.

Hint for Exercise 7.26 For (a), solve for $\varphi_Y = \varphi_{X+Y}/\varphi_X$. For (b), what constraint does $|\varphi_Y(t)| \leq 1$ impose?

Hint for Exercise 7.27 For (a), recall that convergence to a constant in probability and in distribution are equivalent. For (b), consider what tools from the proof of Lévy's theorem might apply here, and whether it is possible to construct a sequence that escapes to infinity while having CFs converge to 1 on an interval.

Hint for Exercise 7.28 Either use the inversion formula, or compute the density directly via convolution and verify that its CF matches your answer in (b).

Hint for Exercise 7.29 For (b), close the contour in the upper half-plane for $t > 0$ and lower half-plane for $t < 0$. Part (c) shows that the sample mean of Cauchy random variables has the same distribution as a single observation—the Law of Large Numbers fails dramatically.

Hint for Exercise 7.30 For (d), use Part 1 of Lévy's theorem: if $X_n \xrightarrow{d} X$, then $\varphi_n \to \varphi_X$ pointwise, but φ_X must be continuous.

Hint for Exercise 7.31 Consider the analyticity of characteristic functions.

Hint for Exercise 7.32 For (b), use the product formula and part (a). For (c), try $X \sim \text{Exp}(1)$.

Hint for Exercise 7.33 For (a), substitute $u = (\lambda - it)x$ and use $\int_0^\infty u^{\alpha-1}e^{-u}\,du = \Gamma(\alpha)$.

Hint for Exercise 7.34 Use $\varphi^{(k)}(0) = i^k\mathbb{E}[X^k]$.

Hint for Exercise 7.35 If X has characteristic function φ, consider $Y = X_1 - X_2$ where X_1, X_2 are independent copies of some random variable. The full proof requires Bochner's theorem or infinitely divisible distributions, which are beyond this course; focus on understanding the structure and verifying the claim for specific examples.

Hint for Exercise 7.36 For (b), use the inversion formula or recognize this as related to the triangular distribution.
Hint for Exercise 7.37 Use equicontinuity: tightness gives a uniform bound on the modulus of continuity of $\{\varphi_n\}$, then extend convergence from a dense set to all of $\mathbb{R}$.
Hint for Exercise 7.38 For (a), define $f(x) = \frac{1}{2\pi}\int e^{-itx}\varphi(t)\,dt$ and use DCT for continuity. For (b), consider mollifying the uniform distribution. For (c), consider the triangular density.
Hint for Exercise 7.39 Recall that for convergence to a constant, convergence in distribution and convergence in probability are equivalent.
Hint for Exercise 8.29 Apply Theorem 8.20 with $g(x) = \log x$, so $g'(\mu) = 1/\mu$.
Hint for Exercise 8.30 For part (c): the boundedness assumption gives $|X_{n,j}| \le C\sigma_j/s_n$. Show that $\max_j \sigma_j/s_n \to 0$, which implies $\mathbf{1}(|X_{n,j}| > \varepsilon) = 0$ for all j and large n.
Hint for Exercise 8.31 For part (c), compute $\nabla g(1,1,\rho) = (-\rho/2,\, -\rho/2,\, 1)^T$. The key computation is $\nabla g^T \Sigma\, \nabla g$.
Hint for Exercise 8.33 For (a): $\mathbb{E}[X_i] = 1/n$, $\mathbb{E}[X_iX_j] = 1/(n(n-1))$, hence $\mathrm{Cov}(X_i, X_j) = 1/(n^2(n-1))$, and $\mathbb{V}\mathrm{ar}(W_n) = 1$ exactly. For (b): $\mathbb{E}\big[\binom{W_n}{k}\big] = 1/k!$ for each fixed k.
Hint for Exercise 8.34 For (b): show that for each fixed $\varepsilon > 0$, $X_1^2\mathbf{1}(|X_1| > \varepsilon\sigma\sqrt{i}) \to 0$ a.s., then apply DCT.
Hint for Exercise 8.35 Set up the triangular array with $X_{n,j} = X_j/s_n$; verify the Lyapunov condition.
Hint for Exercise 8.36 Use the Riesz representation theorem: every continuous linear functional on ℓ^2 has the form $x \mapsto \langle t, x\rangle$ for some $t \in \ell^2$. Consider $\mu_n = \delta_{e_n}$ where e_n is the nth standard basis vector. Show that $\langle t, e_n\rangle \to 0$ for every $t \in \ell^2$, but (μ_n) is not tight.For tightness, learn more about compact subsets of ℓ^2.
Hint for Exercise 8.37 For part (a): write $g(T_n) - g(\theta) = \frac{1}{2}g''(\theta)(T_n-\theta)^2 + o_P((T_n-\theta)^2)$, then use $\sqrt{n}(T_n - \theta) \xrightarrow{d} Z \sim N(0, \sigma^2)$ and the continuous mapping theorem.
Hint for Exercise 8.38 For (b), compute $\varphi(t) = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)}\,dx$ by contour integration or use the known formula for the Fourier transform of $1/(1+x^2)$. For (c), use independence and the CF of a sum.
Hint for Exercise 8.39 For (a), apply the delta method with $g'(\theta) = 1/\sqrt{v(\theta)}$, so $[g'(\theta)]^2 \cdot v(\theta) = 1$.
Hint for Exercise 8.40 For (a): use $S_{N(t)} \le t < S_{N(t)+1}$, divide by $N(t)$, and apply SLLN. For (b): the event $\{N(t) < k\}$ is the same as $\{S_k > t\}$. Use this duality to convert the CLT for S_k (with random index) into a CLT for $N(t)$.
Hint for Exercise 8.41 For (a): $\mathbb{V}\mathrm{ar}(\hat{p}_n) = p(1-p)/n \le 1/(4n)$, so $\sqrt{n}|\hat{p}_n - p|/\sqrt{p(1-p)} \le \sqrt{n}|\hat{p}_n - p| \cdot 2$. You need $2 \cdot 0.03\sqrt{n} \ge 1.96$, giving $n \ge 1068$.
Hint for Exercise 8.42 For (c): $\varphi_{ij}(t) = 1 - 1/i + \cos(t)/i = 1 + (\cos t - 1)/i$. Use $(1 + z/i)^i \to e^z$. For the compound Poisson identification, condition on N: $\mathbb{E}[e^{itW}] = \mathbb{E}[(\cos t)^N] = e^{\cos t - 1}$.
Hint for Exercise 9.20 For part (b), take $A \in \mathcal{F}_S$ and show that $A \cap \{T \le n\} \in \mathcal{F}_n$ by writing $A \cap \{T \le n\} = A \cap \{S \le n\} \cap \{T \le n\}$.

Hint for Exercise 9.21 For part (a), write $L_{n+1} = L_n \cdot g(X_{n+1})/f(X_{n+1})$. Use independence to compute $\mathbb{E}(g(X_{n+1})/f(X_{n+1}) \mid \mathcal{F}_n) = \mathbb{E}(g(X_{n+1})/f(X_{n+1}))$, and note that $\int (g(x)/f(x))f(x)\,dx = \int g(x)\,dx = 1$.

Hint for Exercise 9.22 Note that $\bar{X}_{n+1} = \max(\bar{X}_n, X_{n+1}) \geq \bar{X}_n$, so the submartingale property $\mathbb{E}(\bar{X}_{n+1} \mid \mathcal{F}_n) \geq \bar{X}_n$ follows immediately. The main work is verifying integrability: $\bar{X}_n^+ \leq \sum_{k=0}^n X_k^+$.

Hint for Exercise 9.23 For (a), start from the Doob decomposition formula and expand $M_k = M_{k-1} + (M_k - M_{k-1})$, square, and take conditional expectations. The cross term vanishes by the martingale property. For (b), take expectations of $M_n^2 = N_n + \langle M \rangle_n$ and use $\mathbb{E}(N_n) = \mathbb{E}(N_0)$.

Hint for Exercise 9.24 $A \in \mathcal{F}_T$ requires $A \cap \{T \leq n\} \in \mathcal{F}_n$ for each n. The binding constraint is $n = 1$: since $\{T \leq 1\} = \{HH, HT\}$, the set $A \cap \{HH, HT\}$ must be in $\mathcal{F}_1$, so it is either $\emptyset$ or $\{HH, HT\}$.

Hint for Exercise 9.25 For (a), $f(t) = 2\lceil t/2 \rceil$ is non-decreasing and satisfies $f(t) \geq t$. For (b), $g(t) = 2\lfloor t/2 \rfloor$ satisfies $g(3) = 2 < 3$, so the sufficient condition fails. Take a simple stopping time with $\mathbb{P}(T = 3) > 0$ and show $\{T'' = 2\} \notin \mathcal{F}_2$.

Hint for Exercise 9.26 For part (a), compare with Example 9.9: $S_n^2 = (S_n^2 - n\sigma^2) + n\sigma^2$. What is the martingale part and what is the predictable part?

Hint for Exercise 9.27 For (a), compute $\mathbb{E}(R_{n+1}B_{n+1} \mid \mathcal{F}_n)$ by conditioning on whether a red or blue ball is drawn: if red is drawn, $R_{n+1}B_{n+1} = (R_n + 1)B_n = R_nB_n + B_n$; similarly for blue.

Hint for Exercise 9.28 For (a), expand $(S_n + \xi_{n+1})^3$ and compute $\mathbb{E}(\cdot \mid \mathcal{F}_n)$ using independence and the moment assumptions. For (b), expand $(S_n + \xi_{n+1})^4$ and determine what must be subtracted from S_n^4 to kill the drift; the fourth cumulant $\kappa_4 = \mathbb{E}(\xi^4) - 3\sigma^4$ will appear.

Hint for Exercise 9.30 Monotonicity follows from Lemma 9.17: $\varphi(M_n)$ is a submartingale. The upper bound uses Jensen for conditional expectations: $\varphi(M_n) = \varphi(\mathbb{E}(X \mid \mathcal{F}_n)) \leq \mathbb{E}(\varphi(X) \mid \mathcal{F}_n)$; now take expectations.

Hint for Exercise 10.15 Decompose $S_{T\wedge n} = S_{T\wedge n}^+ - S_{T\wedge n}^-$. Use $S_{T\wedge n} \leq 1$ to show that $\mathbb{E}(S_{T\wedge n}^+) \to 1$. Use $\mathbb{E}(S_{T\wedge n}) = 0$ to conclude that $\mathbb{E}(S_{T\wedge n}^-) \to 1$. Then show that for any fixed K, the truncated part $\mathbb{E}(S_{T\wedge n}^- \mathbf{1}_{\{S_{T\wedge n}^- \leq K\}}) \to 0$, so the tail $\mathbb{E}(S_{T\wedge n}^- \mathbf{1}_{\{S_{T\wedge n}^- > K\}}) \to 1$.

Hint for Exercise 10.16 For part (c), write $p = (1+\delta)/2$, $q = (1-\delta)/2$, so that $r = q/p = (1-\delta)/(1+\delta)$, and expand r^A and r^{-B} to second order in δ.

Hint for Exercise 10.17 For (a), apply Borel–Cantelli to disjoint blocks of length ℓ. For (c), observe that at any time $n < T$, at most ℓ gamblers have nonzero wealth, each bounded by s^ℓ. For the computation of W_T: gambler G_{T-j+1}, who entered at time $T-j+1$, has placed bets on the letters $p_1, \ldots, p_j$ while the realised letters at times $T-j+1, \ldots, T$ are $p_{\ell-j+1}, \ldots, p_\ell$. He is still in the game iff these two strings agree.

Hint for Exercise 10.18 Apply Azuma–Hoeffding to the Doob martingale $Y_k = \mathbb{E}(f(X_1, \ldots, X_n) \mid \mathcal{F}_k)$ where $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$. To bound $|Y_k - Y_{k-1}|$, use independence to express Y_{k-1} as an expectation over an independent copy X_k' of X_k, and apply the bounded-differences condition pointwise.

Hint for Exercise 10.19 Reveal the edges one vertex at a time: let X_i encode

the edges between vertex $i+1$ and the vertices $\{1,\dots,i\}$. Apply Exercise 10.18 with $c_i = 1$.

Hint for Exercise 10.21 Use the reflection principle or the quadratic martingale $S_n^2 - n$ to relate $\mathbb{E}(T \wedge n)$ to $\mathbb{E}(S_{T\wedge n}^2)$.

Hint for Exercise 10.24 Apply the argument of Wald's first identity to the martingale $M_n = S_n^2 - n\sigma^2$, or use the approach from the gambler's ruin problem.

Hint for Exercise 11.23 The martingale $M_n = \mathbb{E}(X \mid \mathcal{F}_n)$ is UI by Theorem 11.11, so it converges to some M_∞ a.s. and in L^1. To identify M_∞, you need to show $\mathbb{E}(M_\infty; A) = \mathbb{E}(X; A)$ for every $A \in \mathcal{F}_\infty$.

Main subtlety. It is straightforward to verify this identity for $A \in \bigcup_n \mathcal{F}_n$ using the martingale property and L^1 convergence. However, $\bigcup_n \mathcal{F}_n$ is in general only an algebra, **not** a σ-algebra, while $\mathcal{F}_\infty = \sigma\big(\bigcup_n \mathcal{F}_n\big)$ is the generated σ-algebra. The conditional expectation identity must hold for **all** $A \in \mathcal{F}_\infty$, not just on the algebra. Use the π–λ theorem (Dynkin's theorem) to extend the conclusion from $\bigcup_n \mathcal{F}_n$ to $\mathcal{F}_\infty$.

Hint for Exercise 11.24 For (a), apply Exercise 11.23 to $X = \mathbf{1}_A$. For (b), use the fact that a tail event is independent of every $\mathcal{F}_n$, so $\mathbb{P}(A \mid \mathcal{F}_n) = \mathbb{P}(A)$ a.s. Combine with (a).

Hint for Exercise 11.25 By Exercise 11.23 applied to $X = \theta$ (which is integrable since $\theta \in (0,1)$), $\mathbb{E}(\theta \mid \mathcal{F}_n) \to \mathbb{E}(\theta \mid \mathcal{F}_\infty)$ a.s. So it suffices to show $\mathbb{E}(\theta \mid \mathcal{F}_\infty) = \theta$, which reduces to showing θ is $\mathcal{F}_\infty$-measurable. Use the SLLN: conditional on θ, $\bar{X}_n = (X_1 + \cdots + X_n)/n \to \theta$ a.s.

Hint for Exercise 11.26 For (b), introduce the stopping time $T_K = \inf\{n : \sum_{k\le n} p_k > K\}$. The stopped martingale $M_{n\wedge T_K}$ has bounded L^2 norm (since $\sum_{k\le n\wedge T_K} \mathbb{E}(p_k) \le K$), so it converges a.s. On the event $\{\sum_n p_n \le K\}$, $T_K = \infty$, so M_n converges; this implies $\sum \mathbf{1}_{A_k}$ converges. Take $K \to \infty$. For (c), apply (b) to the events A_n^c with conditional probabilities $1 - p_n$, after reformulating; or argue directly: on $\{\sum p_n = \infty, \sum \mathbf{1}_{A_n} < \infty\}$, the martingale M_n would tend to $-\infty$, contradicting a.s. convergence on this event.

Hint for Exercise 11.27 For (a), condition on $\mathcal{F}_n$ and use $Z_{n+1} = \sum_{i=1}^{Z_n} \xi_i^{(n)}$, where the $\xi_i^{(n)}$ are i.i.d. copies of ξ independent of $\mathcal{F}_n$. For (b), use the law of total variance:

$$\mathbb{V}\mathrm{ar}(W_{n+1}) = \mathbb{E}\mathbb{V}\mathrm{ar}(W_{n+1} \mid \mathcal{F}_n) + \mathbb{V}\mathrm{ar}\mathbb{E}(W_{n+1} \mid \mathcal{F}_n) = \mu^{-2(n+1)}\sigma^2\mathbb{E}(Z_n) + \mathbb{V}\mathrm{ar}(W_n).$$

This gives a telescoping sum.

Hint for Exercise 11.28 $X_n \to 0$ a.s. but $\mathbb{E}(X_n) = 1$ for all n. If $\{X_n\}$ were UI, then $X_n \to 0$ would hold in L^1, giving $\mathbb{E}(X_n) \to 0$. Alternatively, show directly that $\sup_n \mathbb{E}(X_n \mathbf{1}_{\{X_n > K\}}) \to 1$ as $n \to \infty$ for every fixed K.

Hint for Exercise 11.29 By the closure theorem (Theorem 11.14), $X_n = \mathbb{E}(X_\infty \mid \mathcal{F}_n)$. First show $X_T = \mathbb{E}(X_\infty \mid \mathcal{F}_T)$: for any $A \in \mathcal{F}_T$, decompose $A = \bigsqcup_{n\in\mathbb{N}\cup\{\infty\}} A \cap \{T = n\}$ and use that $A \cap \{T = n\} \in \mathcal{F}_n$. Then for $A \in \mathcal{F}_S \subset \mathcal{F}_T$, $\mathbb{E}(X_T \mathbf{1}_A) = \mathbb{E}(X_\infty \mathbf{1}_A) = \mathbb{E}(X_S \mathbf{1}_A)$.

Hint for Exercise 11.30 Use orthogonality of martingale differences: $\mathbb{E}(M_n^2) = \mathbb{E}(M_0^2) + \sum_{k=1}^n \mathbb{E}(D_k^2)$.

Hint for Exercise 11.31 For (a), for $A \in \mathcal{F}_n$, $\mathbb{E}_{\mathbb{P}}(L_{n+1}\mathbf{1}_A) = \mathbb{Q}_{n+1}(A) = \mathbb{Q}_n(A) = \mathbb{E}_{\mathbb{P}}(L_n \mathbf{1}_A)$. For (c), $L_n = (4/3)^{S_n}(2/3)^{n-S_n}$. Take logarithm and use

the SLLN under $\mathbb{P}$ to find $\frac{1}{n}\log L_n \to (1/2)\log(8/9) < 0$, so $L_n \to 0$ $\mathbb{P}$-a.s.

Hint for Exercise 11.32 For (a), $\mathbb{E}((q^*)^{Z_{n+1}} \mid \mathcal{F}_n) = \varphi(q^*)^{Z_n} = (q^*)^{Z_n}$ since Z_{n+1} given $\mathcal{F}_n$ is a sum of Z_n i.i.d. copies of ξ. For (b), on extinction $Z_n = 0$ eventually, so $(q^*)^{Z_n} = 1$ eventually; on non-extinction $Z_n \to \infty$. For (c), $\mathbb{E}(M_\infty) = \mathbb{E}(M_0) = q^*$ and $\mathbb{E}(M_\infty) = \mathbb{P}(\text{extinction})$ from (b).

Hint for Exercise 11.33 Apply the L^2-bounded martingale convergence theorem (Theorem 11.1) to the partial sums $S_n = X_1 + \cdots + X_n$ with respect to the natural filtration.

Hint for Exercise 12.11 Verify that $M_n = \mathbb{E}(X \mid \mathcal{G}_n)$ is a backward martingale with respect to $\{\mathcal{G}_n\}$, then apply Theorem 12.2. To identify the limit, use the tower property: since $\mathcal{G}_\infty \subset \mathcal{G}_0$, $\mathbb{E}(M_0 \mid \mathcal{G}_\infty) = \mathbb{E}(\mathbb{E}(X \mid \mathcal{G}_0) \mid \mathcal{G}_\infty) = \mathbb{E}(X \mid \mathcal{G}_\infty)$.

Hint for Exercise 12.12 The same backward martingale machinery used in the proof of Theorem 12.4 gives both a.s. and L^1 convergence directly — backward martingales are uniformly integrable. Extract this conclusion explicitly.

Hint for Exercise 12.13 For (a), $\hat{\mu}_{n+1} = \frac{n}{n+1}\hat{\mu}_n + \frac{1}{n+1}\delta_{X_{n+1}}$. For (b), the key fact is that conditional on $\mathcal{T}_{n+1} = \sigma(\hat{\mu}_{n+1}, X_{n+2}, \ldots)$, the random vector $(X_1, \ldots, X_{n+1})$ is uniformly distributed over all permutations of the multiset specified by $\hat{\mu}_{n+1}$. Hence for any subset $\{i_1 < \cdots < i_k\} \subset \{1, \ldots, n+1\}$, $\mathbb{E}(h(X_{i_1}, \ldots, X_{i_k}) \mid \mathcal{T}_{n+1})$ is the same; averaging over the $\binom{n+1}{k}$ subsets yields U_{n+1}. For (c), Theorem 12.2 gives convergence to $\mathbb{E}(h(X_1, \ldots, X_k) \mid \mathcal{T}_\infty)$. Then $\mathcal{T}_\infty \subset \mathcal{E}$ (the exchangeable σ-algebra) by the same argument as in the proof of Theorem 12.4, and Hewitt–Savage gives the conclusion.

Hint for Exercise 12.14 For (a), $\mathbb{V}\mathrm{ar}(S_n/n) = \mathbb{V}\mathrm{ar}(X_1)/n$ by independence. For (b), apply Exercise 12.12 to the sequence $|X_i|^p$ to get $Y_n := \frac{1}{n}\sum_{i=1}^n |X_i|^p \to \mathbb{E}|X_1|^p$ a.s. and in L^1. Then $\{Y_n\}$ is uniformly integrable. By Jensen, $|S_n/n|^p \le Y_n$, so $\{|S_n/n|^p\}$ is also UI. Combined with the a.s. convergence $S_n/n \to \mu$ from the SLLN, Vitali's theorem upgrades this to L^p convergence (after noting $\{|S_n/n - \mu|^p\}$ is UI by an analogous bound).

Hint for Exercise 12.15 Choose N large enough that π fixes all integers greater than N. For $n \ge N$, the partial sum S_n is invariant under T_π (sum is symmetric), and the sequence $(X_{n+1}, X_{n+2}, \ldots)$ is unchanged by T_π. Hence $\mathcal{G}_n$ is invariant under T_π for every $n \ge N$. Take the intersection over n.

Hint for Exercise 12.16 For (a), expand $\sum_i (X_i - \bar{X}_n)^2$ and $\sum_{i<j}(X_i - X_j)^2$, then use the identity $\sum_{i<j}(X_i - X_j)^2 = n\sum_i X_i^2 - S_n^2$. For (b), $\mathbb{E}(h(X_1, X_2)) = \mathbb{E}((X_1 - X_2)^2)/2 = \mathbb{V}\mathrm{ar}(X_1) = \sigma^2$.

Hint for Exercise 12.17 For (a), it suffices to show $\mathbb{E}(X_i \mathbf{1}_A) = \mathbb{E}(X_j \mathbf{1}_A)$ for every $A \in \mathcal{G}_{n+1}$. Write $\mathbf{1}_A = G(S_{n+1}, X_{n+2}, X_{n+3}, \ldots)$ for some measurable G, and note that S_{n+1} is symmetric in $(X_1, \ldots, X_{n+1})$. Apply the change of variable corresponding to the transposition τ swapping indices i and j, using exchangeability of the i.i.d. sequence under τ. For (b), sum the equality from (a) over $i = 1, \ldots, n+1$ and use $\mathbb{E}(S_{n+1} \mid \mathcal{G}_{n+1}) = S_{n+1}$.

Bibliography

[Ald85] David J. Aldous. "Exchangeability and related topics". In: **École d'Été de Probabilités de Saint-Flour XIII—1983**. Vol. 1117. Lecture Notes in Mathematics. Springer, 1985, pp. 1–198. DOI: 10.1007/BFb0099421.

[Ete81] Nasrollah Etemadi. "An elementary proof of the strong law of large numbers". In: **Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete** 55.1 (1981), pp. 119–122. DOI: 10.1007/BF01013465.

[Kal05] Olav Kallenberg. **Probabilistic Symmetries and Invariance Principles**. Probability and Its Applications. Springer, 2005. DOI: 10.1007/0-387-28861-9.