Introduction to Probability Theory

Hao WU

Contents

References

[Chu01] Kai Lai Chung. *A Course in Probability Theory.* Academic Press, Inc., San Diego, CA, third edition, 2001.

[Dur10] Rick Durrett. *Probability: Theory and Examples.* Cambridge University Press, Cambridge, fourth edition, 2010.

[LPW08] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, Providence, Rhode Island, 2008.

1 Probability Space and Random Variable

1.1 Probability space

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ contains three elements:

- The space Ω: this is a non-empty set. It can be viewed as the set of all possible outcomes.
- The σ-field $\mathcal{F}$: this can be viewed as a collection of all the events.
- The probability measure $\mathbb{P}$: this is a function from $\mathcal{F}$ to $[0,1]$. It gives a probability to each event.

Definition 1.1.1. *Suppose $\mathcal{F}$ is a non-empty collection of subsets of Ω.*

- *It is a field, if it is closed under complementation and closed under union:*

$$A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}.$$

$$A_1, A_2 \in \mathcal{F} \Longrightarrow A_1 \cup A_2 \in \mathcal{F}.$$

- *It is a monotone class if*

$$A_j \in \mathcal{F}, A_j \subset A_{j+1}, 1 \leqslant j < \infty \Longrightarrow \cup_j A_j \in \mathcal{F}.$$

$$A_j \in \mathcal{F}, A_j \supset A_{j+1}, 1 \leqslant j < \infty \Longrightarrow \cap_j A_j \in \mathcal{F}.$$

- *It is a σ-field if it is closed under complementation and closed under countable union:*

$$A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}.$$

$$A_j \in \mathcal{F}, 1 \leqslant j < \infty \Longrightarrow \cup_j A_j \in \mathcal{F}.$$

Lemma 1.1.2. *A field is a σ-field if and only if it is a monotone class.*

Note that the collection $\{\varnothing, \Omega\}$ is a σ-field, and we call it the trivial σ-field; the collection of all subsets of Ω is a σ-field, and we call it the total σ-field. Suppose J is an index-set (not necessarily countable), and $\mathcal{F}_j$ is a σ-field for each $j \in J$, then $\cap_{j \in J} \mathcal{F}_j$ is also a σ-field.

Definition 1.1.3. *Given any collection $\mathcal{C}$ of sets, the σ-field (resp. monotone class) generated by $\mathcal{C}$ is the intersection of all σ-fields (resp. monotone classes) containing $\mathcal{C}$.*

Lemma 1.1.4. *Suppose $\mathcal{A}$ is a field. Denote by $\mathcal{F}$ the σ-field generated by $\mathcal{A}$ and by $\mathcal{G}$ the monotone class generated by $\mathcal{A}$. Then $\mathcal{F} = \mathcal{G}$.*

Proof. Since $\mathcal{F}$ is a monotone class, we have $\mathcal{G} \subset \mathcal{F}$. To prove $\mathcal{F} \subset \mathcal{G}$, it is sufficient to show that $\mathcal{G}$ is a σ-field. By Lemma 1.1.2, it is sufficient to show $\mathcal{G}$ is a field, i.e. to show $\mathcal{G}$ is closed under intersection and is closed under complementation.

We first show that $\mathcal{G}$ is closed under intersection. Define

$$\mathcal{G}_1 = \{E \in \mathcal{G} : E \cap F \in \mathcal{G}, \forall F \in \mathcal{A}\}, \quad \mathcal{G}_2 = \{E \in \mathcal{G} : E \cap F \in \mathcal{G}, \forall F \in \mathcal{G}\}.$$

For $\mathcal{G}_1$, we see $\mathcal{A} \subset \mathcal{G}_1$ since $\mathcal{A}$ is a field, and $\mathcal{G}_1$ is a monotone class. By the minimality of $\mathcal{G}$, we find $\mathcal{G} \subset \mathcal{G}_1$ and hence $\mathcal{G} = \mathcal{G}_1$. For $\mathcal{G}_2$, we see $\mathcal{A} \subset \mathcal{G}_2$ since $\mathcal{G} = \mathcal{G}_1$, and $\mathcal{G}_2$ is a monotone class. Again we find $\mathcal{G} = \mathcal{G}_2$. Hence $\mathcal{G}$ is closed under intersection.

Next, we show that $\mathcal{G}$ is closed under complementation. Define

$$\mathcal{G}_3 = \{E \in \mathcal{G} : E^c \in \mathcal{G}\}.$$

Since $\mathcal{A} \subset \mathcal{G}_3$ and $\mathcal{G}_3$ is a monotone class, we find $\mathcal{G} = \mathcal{G}_3$ as desired. $\square$

Example 1.1.5. *The union of a countable collection of σ-fields $\{\mathcal{F}_j\}$ such that $\mathcal{F}_j \subset \mathcal{F}_{j+1}$ need not be a σ-fields. For example, $\Omega = \mathbb{Z}_{>0}$ and $\mathcal{F}_j$ is the σ-field generated by $\{\{1\}, \{2\}, \dots, \{j\}\}$. For each j, the σ-field $\mathcal{F}_j$ is finite; but the σ-field generated by $\cup_j \mathcal{F}_j$ is nolonger countable.*

In contrast, the intersection of σ-fields is always a σ-field. Suppose J is an index set and $\mathcal{F}_j$ is a σ-field for each $j \in J$. The intersection $\cap_{j\in J}\mathcal{F}_j$ is a σ-field. But the union need not be a σ-field. We denote by $\vee_{j\in J}\mathcal{F}_j$ the σ-field generated by $\cup_{j\in J}\mathcal{F}_j$.

Definition 1.1.6. *Suppose $\mathcal{F}$ is a σ-field on Ω. A probability measure $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0,1]$ satisfying the following axioms:*

- $\mathbb{P}[E] \geqslant 0$ *for all* $E \in \mathcal{F}$;
- $\mathbb{P}[\Omega] = 1$.
- *If* $\{E_j\}_j$ *is a countable collection of pairwise disjoint sets in* $\mathcal{F}$, *then* $\mathbb{P}[\cup_j E_j] = \sum_j \mathbb{P}[E_j]$.

These axioms imply the following consequences:

- $\mathbb{P}[E^c] = 1 - \mathbb{P}[E]$.
- $\mathbb{P}[E \cup F] + \mathbb{P}[E \cap F] = \mathbb{P}[E] + \mathbb{P}[F]$.
- Continuity: if $E_n \uparrow E$ or $E_n \downarrow E$ then $\mathbb{P}[E_n] \to \mathbb{P}[E]$.
- $\mathbb{P}[\cup_j E_j] \leqslant \sum_j \mathbb{P}[E_j]$.

Example 1.1.7. *Suppose Ω is a countable set: $\Omega = \{\omega_j, j \in J\}$ where J is countable. Let $\mathcal{F}$ be the total σ-field. Suppose $\{p_j, j \in J\}$ is a sequence of numbers satisfying*

$$p_j \geqslant 0, \forall j \in J; \quad \sum_j p_j = 1.$$

Define $\mathbb{P} : \mathcal{F} \to [0,1]$ as follows:

$$\mathbb{P}[E] := \sum_{j:\, \omega_j \in E} p_j, \quad \forall E \in \mathcal{F}.$$

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

Example 1.1.8. *Let $\mathcal{U} = (0,1]$ and let $\mathcal{C}$ be the collection of intervals $(a,b]$ where $0 \leqslant a < b \leqslant 1$. Let $\mathcal{B}$ be the σ-field generated by $\mathcal{C}$ and let Leb be the Lebesgue measure on $(0,1]$. Then $(\mathcal{U}, \mathcal{B}, \mathrm{Leb})$ is a probability measure.*

Theorem 1.1.9 (Carathéodory's Extension Theorem)**.** *Suppose $\mathcal{F}_0$ is a field and $\mathcal{F}$ is the σ-field generated by $\mathcal{F}_0$. Suppose μ is a probability measure on $\mathcal{F}_0$. Then there exists a unique probability measure on $\mathcal{F}$ that coincides with μ on $\mathcal{F}_0$.*

Proof. Proof of existence: reading. Proof of uniqueness: Suppose μ and ν are probability measures on $\mathcal{F}$ such that $\mu = \nu$ on $\mathcal{F}_0$. Then $\mu = \nu$ on $\mathcal{F}$.

Define

$$\mathcal{C} = \{E \in \mathcal{F} : \mu[E] = \nu[E]\}.$$

We see $\mathcal{F}_0 \subset \mathcal{C}$ by the hypothesis, and we find $\mathcal{C}$ is a monotone class. Lemma 1.1.4 gives that $\mathcal{F} \subset \mathcal{C}$ which completes the proof. $\square$

Let us discuss the probability measures on $\mathbb{R} = (-\infty, \infty)$. Let $\mathcal{C}$ be the collection of intervals of the form $(a, b]$ with $a < b$. The field $\mathcal{B}_0$ generated by $\mathcal{C}$ consists of finite union of disjoint sets of the form $(a, b], (-\infty, a]$ or (b, ∞). Denote by $\mathcal{B}$ the σ-field generated by $\mathcal{C}$. It coincides with the Borel field on $\mathbb{R}$. However, the Borel-Lebesgue measure is not a probability measure on $\mathbb{R}$.

The question of probability measures on $\mathbb{R}$ is closely related to distribution functions. A function $F : \mathbb{R} \to [0, 1]$ is a distribution function if it is increasing and right-continuous with $F(-\infty) = 0$ and $F(+\infty) = 1$.

Proposition 1.1.10. *Each probability measure μ on $\mathbb{R}$ uniquely determines a distribution function F through:*

$$\mu((-\infty, x]) = F(x), \quad \forall x \in \mathbb{R}. \tag{1.1.1}$$

Conversely, given a distribution function F, there exists a unique probability measure μ on $\mathbb{R}$ satisfying (1.1.1).

It is clear that the function F defined by a probability measure μ via (1.1.1) is a distribution function. Moreover, we have $\mu((a, b]) = F(b) - F(a)$ for $a < b$; and $\mu(\{x\}) = F(x) - F(x-)$ for $x \in \mathbb{R}$. The converse direction is a particular case of the extension theorem.

Corollary 1.1.11. *If two probability measures on $\mathbb{R}$ agree on all intervals of the form $(a, b]$ with $a < b$, then they agree on $\mathcal{B}$.*

Definition 1.1.12. *The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete if any subset of a set in $\mathcal{F}$ with $\mathbb{P}[F] = 0$ also belongs to $\mathcal{F}$.*

We call a set $F \in \mathcal{F}$ a null set if $\mathbb{P}[F] = 0$. We say a property holds almost surely if it holds except on a null set. Any probability space can be completed by the following theorem. What is the advantage of completion? Suppose a property holds almost surely, i.e. it holds outside a certain set N with $\mathbb{P}[N] = 0$. Then the set on which it fails to hold is a subset of N, not necessarily in $\mathcal{F}$. However, we sometimes need the measurability of the exact exceptional set to proceed.

Theorem 1.1.13. *Given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there exists a complete space $(\Omega, \overline{\mathcal{F}}, \overline{\mathbb{P}})$ such that $\mathcal{F} \subset \overline{\mathcal{F}}$ and $\overline{\mathbb{P}} = \mathbb{P}$ on $\mathcal{F}$.*

Proof. Denote by $\mathcal{N}$ the collection of sets that are subsets of null sets, and define

$$\overline{\mathcal{F}} = \{E \subset \Omega : E \Delta F \in \mathcal{N}, \text{ for some } F \in \mathcal{F}\}.$$

We can check that $\mathcal{F} \subset \overline{\mathcal{F}}$ and $\overline{\mathcal{F}}$ is a σ-field. Define $\overline{\mathbb{P}}$ on $\overline{\mathcal{F}}$ as follows:

$$\overline{\mathbb{P}}[E] = \mathbb{P}[F],$$

where $F \in \mathcal{F}$ is any set such that $E \Delta F \in \mathcal{N}$. This is well-defined, because

$$F_1 \Delta F_2 = (E \Delta F_1) \Delta (E \Delta F_2)$$

which implies $F_1 \Delta F_2 \in \mathcal{N}$ and $\mathbb{P}[F_1] = \mathbb{P}[F_2]$. In order to check $(\Omega, \overline{\mathcal{F}}, \overline{\mathbb{P}})$ is a completion of $(\Omega, \mathcal{F}, \mathbb{P})$, it remains to check:

- $\overline{\mathbb{P}} = \mathbb{P}$ on $\mathcal{F}$: this is clear.
- $\overline{\mathbb{P}}$ is a probability measure on $\overline{\mathcal{F}}$: exercise.
- $\overline{\mathbb{P}}$ is complete: in fact, we can show that if $\overline{\mathbb{P}}[E] = 0$ for $E \in \overline{\mathcal{F}}$, then $E \in \mathcal{N}$. Hence any subset of E belongs to $\mathcal{N} \subset \overline{\mathcal{F}}$.

□

Hereafter, we always assume the probability space is complete.

1.2 Random variable

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Denote by $\mathbb{R} = (-\infty, \infty)$ and by $\mathcal{B}$ the Borel-field on $\mathbb{R}$.

Definition 1.2.1. *A real-valued random variable is a function $X : \Omega \to \mathbb{R}$ such that*

$$X^{-1}(B) \in \mathcal{F}, \quad \forall B \in \mathcal{B}.$$

In other words, a random variable is just a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$.

This definition of random variable is the one we mostly use. But for logical reasons, we sometimes need its generalization, see [Chu01, Section 3.1].

Note that, if X is a random variable and f is a Borel measurable function on $(\mathbb{R}, \mathcal{B})$, then $f(X)$ is also a random variable. The indicator function $\mathbb{1}_A : \Omega \to \{0, 1\}$ is a random variable if and only if $A \in \mathcal{F}$.

Lemma 1.2.2. *X is a random variable if and only if*

$$X^{-1}((-\infty, x]) \in \mathcal{F}, \quad \forall x \in \mathbb{R}.$$

Proof. Only need to show the "only if" part. Define

$$\mathcal{C} = \{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{F}\}.$$

We can check $\mathcal{C}$ is a σ-field. By the hypothesis, $\mathcal{C}$ contains $\{(-\infty, x] : x \in \mathbb{R}\}$ which generates $\mathcal{B}$. This implies $\mathcal{C} = \mathcal{B}$ as desired. □

Lemma 1.2.3. *If $\{X_j, j \geqslant 1\}$ is a sequence of random variables, then*

$$\inf_j X_j, \quad \sup_j X_j, \quad \liminf_j X_j, \quad \limsup_j X_j$$

are random variables. Note that they are everywhere well-defined, but they are not necessarily finite-valued with probability one. If they are not finite-valued, then we need to use the generalized definition of random variables.

Proof. Note that

$$\{\inf_j X_j \geqslant x\} = \cap_j \{X_j \geqslant x\}, \quad \{\sup_j X_j \leqslant x\} = \cap_j \{X_j \leqslant x\}, \quad \forall x \in \mathbb{R}.$$

Combining with Lemma 1.2.2, we see that $\inf_j X_j, \sup_j X_j$ are random variables.

Note that

$$\liminf_j X_j = \sup_n (\inf_{j \geqslant n} X_j), \quad \limsup_j X_j = \inf_n (\sup_{j \geqslant n} X_j).$$

These guarantee that they are random variables. □

Definition 1.2.4. *Each random variable X induces a probability measure μ on $(\mathbb{R}, \mathcal{B})$ by the following correspondence:*

$$\mu[B] = \mathbb{P}[X^{-1}(B)] = \mathbb{P}[X \in B], \quad \forall B \in \mathcal{B}.$$

The measure μ is called the law (or the distribution) of X, denoted by $\mathcal{L}(X)$; its associated distribution function is called the distribution function of X, denoted by F_X.

Specifically, the distribution function F of X is given by

$$F(x) = \mathbb{P}[X \leqslant x], \quad \forall x \in \mathbb{R}.$$

The random variable X determines μ and hence F; whereas, its converse is obviously false. A family of random variables having the same distribution is said to be identically distributed.

Example 1.2.5. *Suppose* $(\mathcal{U}, \mathcal{B}, \mathrm{Leb})$ *is the probability measure in Example 1.1.8. The functions* $X(\omega) = \omega$ *and* $Y(\omega) = 1 - \omega$ *are random variables. They are not identical but they are identically distributed; their common law is* Leb.

Next, we discuss the density of distribution function. To this end, we need to discuss absolute continuity.

Definition 1.2.6. *A function* $F : I \subset \mathbb{R} \to \mathbb{R}$ *is absolutely continuous if, for every* $\epsilon > 0$*, there exists* $\delta > 0$ *such that whenever a finite sequence of pairwise disjoint intervals* (x_k, y_k) *of* I *satisfies* $\sum_k (y_k - x_k) < \delta$*, then*

$$\sum_k |F(y_k) - F(x_k)| < \epsilon.$$

Lemma 1.2.7. *The following conditions of* F *on a compact interval* $I = [a, b]$ *are equivalent.*

(1) F *is absolutely continuous.*

(2) F *has derivative* F' *almost everywhere, the derivative is Lebesgue integrable, and*

$$F(x) = F(a) + \int_a^x F'(y)dy, \quad \forall x \in [a, b].$$

(3) There exists a Lebesgue integrable function g *on* $[a, b]$ *such that*

$$F(x) = F(a) + \int_a^x g(y)dy, \quad \forall x \in [a, b].$$

Moreover, we have the following relation: suppose functions are defined on a compact interval, then

$$\begin{aligned}&\{\text{continuously differentiable}\} \subset \{\text{Lipschitz continuous}\} \subset \{\text{absolutely continuous}\}\\ &\subset \{\text{differentiable almost everywhere}\}.\end{aligned}$$

Suppose the distribution function F is absolutely continuous, then there exists an integrable function p such that

$$F(b) = F(a) + \int_a^b p(x)dx.$$

We define the function p as the density function. Note that it is defined up to a zero-measure set. As we usually consider the integral of p, its value on a zero-measure set does not contribute. Thus there is no ambiguity.

Definition (Definition 1.2.6 bis)**.** *A probability measure* μ *on* $(\mathbb{R}, \mathcal{B})$ *is absolutely continuous (with respect to Lebesgue measure) if, for every measurable set* A*,* $\mathrm{Leb}[A] = 0$ *implies* $\mu[A] = 0$*.*

Lemma (Lemma 1.2.7 bis)**.** *The following conditions of a probability measure* μ *are equivalent.*

(1) μ *is absolutely continuous.*

(2) For every $\epsilon > 0$*, there exists* $\delta > 0$ *such that*

$$\mu[A] \leqslant \epsilon, \quad \textit{as long as } \mathrm{Leb}[A] \leqslant \delta.$$

(3) There exists a Lebesgue integrable function p *such that*

$$\mu[A] = \int_A p(x)dx, \quad \textit{for all Borel sets } A.$$

Example 1.2.8 (Uniform distribution).

$$\mathbb{P}[X \leqslant x] = \begin{cases} 0, & x \leqslant 0, \\ x, & 0 \leqslant x \leqslant 1, \\ 1, & x > 1. \end{cases} \qquad p(x) = \begin{cases} 0, & x \leqslant 0, \\ 1, & 0 \leqslant x \leqslant 1, \\ 0, & x \geqslant 1. \end{cases}$$

Example 1.2.9 (Exponential distribution with parameter $\lambda > 0$).

$$\mathbb{P}[X \leqslant x] = \begin{cases} 0, & x \leqslant 0, \\ 1 - e^{-\lambda x}, & x \geqslant 0. \end{cases} \qquad p(x) = \begin{cases} 0, & x \leqslant 0, \\ \lambda e^{-\lambda x}, & x \geqslant 0. \end{cases}$$

We denote this law by $\mathrm{Exp}(\lambda)$.

Example 1.2.10 (Normal distribution).

$$\mathbb{P}[X \leqslant x] = \int_{\infty}^{x} p(y)dy, \qquad \textit{where } p(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

We denote this law by $\mathcal{N}(0,1)$. *Suppose* $X \sim \mathcal{N}(0,1)$, *for* $m \in \mathbb{R}, \sigma > 0$, *we see that* $\sigma X + m$ *has density*

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

We denote this law by $\mathcal{N}(m, \sigma)$.

Example 1.2.11 (Bernoulli distribution with parameter $p \in (0,1)$).

$$\mathbb{P}[X = 1] = p, \quad \mathbb{P}[X = 0] = 1 - p.$$

Example 1.2.12 (Poisson distribution with parameter $\lambda > 0$).

$$\mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \textit{for } k = 0, 1, 2, \ldots.$$

We denote this law by $\mathrm{Poisson}(\lambda)$.

Example 1.2.13 (Geometric distribution with success probability $p \in (0,1)$).

$$\mathbb{P}[N = k] = p(1-p)^{k-1}, \quad \textit{for } k = 1, 2, \ldots.$$

Next, we will discuss random vector. This is just a vector each of whose components is a random variable. It is sufficient to consider the case of two dimensions.

Recall that the Borel field $\mathcal{B}^2$ on $\mathbb{R}^2$ is the σ-field generated by rectangles of the form $(a, b] \times (c, d]$. It is also generated by product sets of the form $B_1 \times B_2$ where $B_1, B_2 \in \mathcal{B}$. Let X, Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The random vector (X, Y) induces a probability measure on $\mathcal{B}^2$:

$$\nu[A] = \mathbb{P}[(X, Y) \in A], \quad \forall A \in \mathcal{B}^2. \tag{1.2.1}$$

If X, Y are random variables, and f is a Borel measurable function on $(\mathbb{R}^2, \mathcal{B}^2)$, then $f(X, Y)$ is also a random variable.

1.3 Expectation

The concept of "expectation" is the same as integration in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which we briefly review.

A random variable X is discrete if it takes values in a countable set, i.e. there exists a countable set $B \subset \mathbb{R}$ such that $\mathbb{P}[X \in B] = 1$. In this case, we write $B = \{b_j\}_j$ and denote by $\Lambda_j = \{\omega \in \Omega : X(\omega) = b_j\}$ for each j. Suppose X is positive discrete, and define its expectation to be

$$\mathbb{E}[X] = \sum_j b_j \mathbb{P}[\Lambda_j].$$

One can check that this is well-defined.

Suppose X is a positive random variable. For $m, n \in \mathbb{Z}_{>0}$, define

$$\Lambda_m^n = \left\{ \omega : \frac{n}{2^m} \leqslant X < \frac{n+1}{2^m} \right\}.$$

For $m \in \mathbb{Z}_{>0}$, define

$$X_m(\omega) = \sum_n \frac{n}{2^m} \mathbb{1}_{\Lambda_m^n}(\omega).$$

Then X_m is positive discrete and

$$X_m(\omega) \leqslant X_{m+1}(\omega), \quad 0 \leqslant X(\omega) - X_m(\omega) \leqslant \frac{1}{2^m}.$$

The expectation $\mathbb{E}[X_m]$ is defined as above and this is a sequence increasing in m. Define

$$\mathbb{E}[X] = \lim_m \mathbb{E}[X_m].$$

This is well-defined and its definition agrees with the previous definition if X is discrete.

For an arbitrary random variable X, put

$$X = X^+ - X^-, \quad \text{where } X^+ = X \vee 0, \quad X^- = (-X) \vee 0.$$

Both X^+ and X^- are positive random variables, their expectations are defined. If both $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are infinite, we say that the expectation of X does not exist; otherwise, we define

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

In the case of $(\mathcal{U}, \mathcal{B}, \text{Leb})$, the expectation reduces to the ordinary Lebesgue integral.

Now, we collect some basic properties of the expectation. Suppose X, Y, X_n are random variables and a, b are constants.

- $\mathbb{E}[X]$ is finite if and only if $\mathbb{E}[|X|]$ is finite.
- Linearity: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, provided that the right hand side is meaningful.
- Positivity: If $X \geqslant 0$ a.s., then $\mathbb{E}[X] \geqslant 0$.
- Monotonicity: If $X \leqslant Y$ a.s., then $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$.
- Dominated convergence theorem: If $\lim_n X_n = X$ a.s. and $|X_n| \leqslant Y$ a.s. where $\mathbb{E}[Y] < \infty$, then X is integrable and
$$\lim_n \mathbb{E}[|X_n - X|] = 0, \quad \lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

- Monotone convergence theorem: if $X_n \geqslant 0$ and $X_n \uparrow X$ a.s., then

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

- Fatou's lemma: If $X_n \geqslant 0$ a.s., then

$$\mathbb{E}[\liminf_n X_n] \leqslant \liminf_n \mathbb{E}[X_n].$$

- If Λ_n's are disjoint and $\cup_n \Lambda_n = \Omega$, then $\mathbb{E}[X] = \sum_n \mathbb{E}[X \mathbb{1}_{\Lambda_n}]$.

The following inequality is quite helpful in future.

Lemma 1.3.1. *We have*

$$\sum_{n=1}^{\infty} \mathbb{P}[|X| \geqslant n] \leqslant \mathbb{E}[|X|] \leqslant 1 + \sum_{n=1}^{\infty} \mathbb{P}[|X| \geqslant n].$$

In particular, $\mathbb{E}[|X|] < \infty$ *if and only if the above series converges.*

Proof. For $n \geqslant 0$, define $\Lambda_n = \{n \leqslant |X| < n+1\}$. We find (check)

$$\mathbb{E}[|X|] = \sum_{n=0}^{\infty} \mathbb{E}[|X| \mathbb{1}_{\Lambda_n}],$$

where

$$n\mathbb{P}[\Lambda_n] \leqslant \mathbb{E}[|X| \mathbb{1}_{\Lambda_n}] \leqslant (n+1)\mathbb{P}[\Lambda_n].$$

Thus

$$\sum_{n=1}^{\infty} n\mathbb{P}[\Lambda_n] \leqslant \mathbb{E}[|X|] \leqslant 1 + \sum_{n=1}^{\infty} n\mathbb{P}[\Lambda_n].$$

It remains to show

$$\sum_{n=1}^{\infty} n\mathbb{P}[\Lambda_n] = \sum_{n=1}^{\infty} \mathbb{P}[|X| \geqslant n]. \tag{1.3.1}$$

Note that, either side can be infinite.

For large N, we write

$$\begin{aligned}
\sum_{n=1}^{N} n\mathbb{P}[\Lambda_n] &= \sum_{n=1}^{N} n(\mathbb{P}[|X| \geqslant n] - \mathbb{P}[|X| \geqslant n+1]) \\
&= \sum_{n=1}^{N} n\mathbb{P}[|X| \geqslant n] - \sum_{n=2}^{N+1} (n-1)\mathbb{P}[|X| \geqslant n]) \\
&= \sum_{n=1}^{N} \mathbb{P}[|X| \geqslant n] - N\mathbb{P}[|X| \geqslant N+1].
\end{aligned}$$

Thus

$$\sum_{n=1}^{N} n\mathbb{P}[\Lambda_n] \leqslant \sum_{n=1}^{N} \mathbb{P}[|X| \geqslant n] \leqslant \sum_{n=1}^{N} n\mathbb{P}[\Lambda_n] + N\mathbb{P}[|X| \geqslant N+1]. \tag{1.3.2}$$

Note that

$$N\mathbb{P}[|X| \geqslant N+1] \leqslant \mathbb{E}[|X| \mathbb{1}\{|X| \geqslant N+1\}].$$

If $\mathbb{E}[|X|] < \infty$, we have $\mathbb{E}[|X| \mathbb{1}\{|X| \geqslant N+1\}] \to 0$ (check) and hence $N\mathbb{P}[|X| \geqslant N+1] \to 0$. Taking $N \to \infty$ in (1.3.2), we obtain (1.3.1).

If $\mathbb{E}[|X|] = \infty$, then $\sum_1^N n\mathbb{P}[\Lambda_n] \to \infty$ and hence $\sum_1^N \mathbb{P}[|X| \geqslant n] \to \infty$. In this case, we also obtain (1.3.1). □

Corollary 1.3.2. *If X takes only positive integer values, we have*

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}[X \geqslant n].$$

There is a basic relation between the abstract integral with respect to $\mathbb{P}$ on $\mathcal{F}$ on the one hand, and the Lebesgue integral with respect to Leb on $\mathcal{B}$, induced by each random variable. We first give the conclusion in one-dimension, and it is easy to generalize to high dimension.

Theorem 1.3.3. *Suppose X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. It induces the probability space $(\mathbb{R}, \mathcal{B}, \mu)$ as in Definition 1.2.4. For any Borel measurable function f, we have*

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

provided that either side exists.

Proof. If $f = \mathbb{1}_B$ for some $B \in \mathcal{B}$, then we have

$$\mathbb{E}[f(X)] = \mathbb{P}[X \in B], \quad \int_{\mathbb{R}} f(x)\mu[dx] = \mu[B].$$

They are equal by definition. Then the proof can be generalized to simple f, i.e. linear combination of indicator functions. For any arbitrary positive Borel function f, we can define a sequence of simple functions $\{f_m\}_m$ such that $f_m \uparrow f$ everywhere. Since f_m is simple, we have

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

Let $m \to \infty$, by monotone convergence theorem, we obtain the conclusion for f positive. The general case follows as usual. $\square$

Theorem (Theorem 1.3.3 bis)**.** *Suppose (X, Y) is a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$. It induces the probability space $(\mathbb{R}^2, \mathcal{B}^2, \nu)$ as in (1.2.1). For any Borel measurable function f, we have*

$$\mathbb{E}[f(X, Y)] = \int_{\mathbb{R}^2} f(x, y)\nu[dx, dy],$$

provided that either side exists.

Definition 1.3.4. *For any $p \in (0, \infty)$, define*

$$L^p(\Omega, \mathcal{F}, \mathbb{P}) = \{X \text{ random variable on } (\Omega, \mathcal{F}, \mathbb{P}) : \mathbb{E}[|X|^p] < \infty\}.$$

For $X \in L^p$, we call $\mathbb{E}[|X|^p]$ the p-th moment of X.

We are usually interested in $p \geqslant 1$, as L^p is a Banach space when $p \geqslant 1$: we define the norm on L^p by

$$\|X\|_p := \mathbb{E}[|X|^p]^{1/p}.$$

Then one can check that L^p is complete under the norm and hence it is a Banach space.

Note that, for $1 \leqslant p < q$, we have $L^q \subset L^p$ because

$$\mathbb{E}[|X|^p \mathbb{1}\{|X| \geqslant 1\}] \leqslant \mathbb{E}[|X|^q \mathbb{1}\{|X| \geqslant 1\}] \leqslant \mathbb{E}[|X|^q].$$

We collect several well-known inequalities. We suggest the readers to remember the names as well as the inequalities.

- **Chebyshev inequality.** Suppose X is a random variable and φ is a strictly positive and increasing function on $[0, \infty)$. Then for each $x > 0$, we have

$$\mathbb{P}[|X| \geqslant x] \leqslant \frac{\mathbb{E}[\varphi(|X|)]}{\varphi(x)}.$$

- **Hölder inequality.** Suppose X and Y are random variables. For $1 < p < \infty$ and $1/p + 1/q = 1$, we have

$$\mathbb{E}[|XY|] \leqslant \mathbb{E}[|X|^p]^{1/p} \mathbb{E}[|Y|^q]^{1/q}.$$

- **Minkowski inequality.** Suppose X and Y are random variables. For $1 \leqslant p < \infty$, we have

$$\mathbb{E}[|X + Y|^p]^{1/p} \leqslant \mathbb{E}[|X|^p]^{1/p} + \mathbb{E}[|Y|^p]^{1/p},$$

or equivalently

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

- **Jensen's inequality.** If φ is a convex function on $\mathbb{R}$, and X and $\varphi(X)$ are integrable random variables, then

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

Example. *Let us calculate moments of random variables in Examples 1.2.8 to 1.2.13. If the distribution function is absolutely continuous, we denote by p its density, then we have*

$$\mathbb{E}[f(X)] = \int f(x)\mu[dx] = \int f(x)p(x)dx.$$

- *Uniform distribution:*

$$\mathbb{E}[X^n] = \int_0^1 x^n dx = \frac{1}{n+1}.$$

- *Exponential distribution:*

$$\mathbb{E}[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x} dx = n! \lambda^{-n}.$$

- *Normal distribution:*

$$\mathbb{E}[X^{2n-1}] = 0, \quad \mathbb{E}[X^{2n}] = \int x^{2n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = (2n-1)!!.$$

  In particular $\mathrm{var}(X) = 1$.

- *Poisson distribution:*

$$\mathbb{E}[X(X-1)\cdots(X-n+1)] = \lambda^n.$$

  In particular $\mathbb{E}[X] = \lambda$ *and* $\mathrm{var}(X) = \lambda$.

- *Geometric distribution:*

$$\mathbb{E}[N] = \frac{1}{p}, \quad \mathrm{var}(N) = \frac{1-p}{p^2}.$$

The case of $p = 2$ is of particular interest, as it is a Hilbert space: we define the inner product on L^2 by

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

Suppose $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, we define its variance and its deviation by

$$\operatorname{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2], \quad \sigma(X) = \sqrt{\operatorname{var}(X)}.$$

Suppose $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, we define their covariance by

$$\operatorname{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

We say that X and Y are uncorrelated if $\operatorname{cov}(X, Y) = 0$.

Exercise 1.3.5. *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ have finite second moments and they are uncorrelated, then*

$$\operatorname{var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \operatorname{var}(X_j).$$

1.4 Independence

Definition 1.4.1. *The random variables $\{X_j, 1 \leqslant j \leqslant n\}$ are independent if, for any Borel sets $\{B_j, 1 \leqslant j \leqslant n\}$, we have*

$$\mathbb{P}\left[\cap_{j=1}^{n} \{X_j \in B_j\}\right] = \prod_{j=1}^{n} \mathbb{P}[X_j \in B_j].$$

The random variables $\{X_j, j \geqslant 1\}$ are independent if $\{X_j, 1 \leqslant j \leqslant n\}$ are independent for all n.

In terms of the law μ induced by the random vector $(X_1, \ldots, X_n)$ on $(\mathbb{R}^n, \mathcal{B}^n)$, and the laws $\{\mu_j, 1 \leqslant j \leqslant n\}$ induced by each X_j on $(\mathbb{R}, \mathcal{B})$, the independence may be written as

$$\mu[B_1 \times \cdots \times B_n] = \prod_{j=1}^{n} \mu_j[B_j].$$

In other words, if $\{X_j, 1 \leqslant j \leqslant n\}$ are independent, then the induced measure μ is the same as the product measure on the product space.

In terms of the distribution function $F(x_1, \ldots, x_n)$ induced by the random vector $(X_1, \ldots, X_n)$ on $\mathbb{R}^n$, and the distribution functions $\{F_j, 1 \leqslant j \leqslant n\}$ induced by each X_j on $\mathbb{R}$, the independence may be written as

$$F(x_1, \ldots, x_n) = \prod_{j=1}^{n} F_j(x_j).$$

Exercise 1.4.2. *If $\{X_j, 1 \leqslant j \leqslant n\}$ are independent random variables, and $\{f_j, 1 \leqslant j \leqslant n\}$ are Borel measurable functions, then $\{f_j(X_j), 1 \leqslant j \leqslant n\}$ are also independent.*

Generally, let $1 \leqslant n_1 < n_2 < \cdots < n_k = n$, and f_1 be a Borel measurable function of n_1 variables, f_2 be a Borel measurable function of $n_2 - n_1$ variables, ..., f_k be a Borel measurable function of $n_k - n_{k-1}$ variables. Then

$$f_1(X_1, \ldots, X_{n_1}), \quad f_2(X_{n_1+1}, \ldots, X_{n_2}), \quad \ldots, \quad f_k(X_{n_{k-1}+1}, \ldots, X_{n_k})$$

are also independent.

Proposition 1.4.3. *Suppose X and Y are independent random variables and both of them have finite expectations. Then we have*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

We will give two proofs of this theorem. Both of them are instructive.

First proof. First, assume both X and Y are positive discrete:

$$X = \sum_j a_j \mathbb{1}_{A_j}, \quad Y = \sum_j b_j \mathbb{1}_{B_j},$$

where $\{a_j\}_j$ are distinct and $A_j = \{\omega : X(\omega) = a_j\}$, and $\{b_j\}_j$ are distinct and $B_j = \{\omega : Y(\omega) = b_j\}$. Then we have

$$XY = \sum_{j,k} a_j b_k \mathbb{1}_{A_j \cap B_k}.$$

Thus

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{j,k} a_j b_k \mathbb{P}[A_j \cap B_k] && \text{(by def. of expectation)} \\
&= \sum_{j,k} a_j b_k \mathbb{P}[A_j]\mathbb{P}[B_k] && \text{(by the independence)} \\
&= \left(\sum_j a_j \mathbb{P}[A_j]\right)\left(\sum_k b_k \mathbb{P}[B_k]\right) = \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

Next, assume both X and Y are positive. We define the approximations X_m and Y_m the same as the beginning of Section 1.3. Then we have the followings.

- We have $\mathbb{E}[X_m] \uparrow \mathbb{E}[X]$ and $\mathbb{E}[Y_m] \uparrow \mathbb{E}[Y]$.
- Since X and Y are independent, we know that $X_m = 2^{-m}\lfloor 2^m X \rfloor$ and $Y_m = 2^{-m}\lfloor 2^m Y \rfloor$ are independent. Both X_m and Y_m are positive discrete. Thus $\mathbb{E}[X_m Y_m] = \mathbb{E}[X_m]\mathbb{E}[Y_m]$.
- The product $X_m Y_m$ is increasing in m, and $0 \leqslant XY - X_m Y_m \to 0$ as $m \to 0$. By monotone convergence theorem, we have $\mathbb{E}[X_m Y_m] \to \mathbb{E}[XY]$.

Combining these three observations, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ as desired.

Finally, consider random variables X and Y with finite expectation. Write $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ and the rest is as usual. □

Second proof. Let μ_X be the law of X, μ_Y be the law of Y, and μ be the law of (X, Y). Then we have

$$\begin{aligned}
\mathbb{E}[XY] &= \iint_{\mathbb{R}^2} xy\mu[dx, dy] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xy\mu_X[dx]\mu_Y[dy] \\
&= \int_{\mathbb{R}} x\mu_X[dx] \int_{\mathbb{R}} y\mu_Y[dy] = \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

This proof is pretty short! This is because we use Fubini's theorem in the second equal sign. □

Corollary 1.4.4. *If $\{X_j, 1 \leqslant j \leqslant n\}$ are independent random variables with finite expectations, then*

$$\mathbb{E}\left[\prod_{j=1}^n X_j\right] = \prod_{j=1}^n \mathbb{E}[X_j].$$

Proof. Induction, using Exercise 1.4.2. □

Next, we turn to the most exciting question of the section: do independent random variables exist?

Example 1.4.5. *Denote by $\mathcal{U}^n$ the n-dimensional cube:*

$$\mathcal{U}^n = \{(x_1, \ldots, x_n) : 0 \leqslant x_j \leqslant 1, 1 \leqslant j \leqslant n\}.$$

Denote by $\mathcal{B}^n$ the Borel field and Leb^n *the Lebesgue measure. Then $(\mathcal{U}^n, \mathcal{B}^n, \mathrm{Leb}^n)$ is a probability space. Let $\{f_j, 1 \leqslant j \leqslant n\}$ be n Borel measurable functions on $\mathcal{U}$, and set*

$$X_j((x_1, \ldots, x_n)) = f_j(x_j).$$

Then $\{X_j, 1 \leqslant j \leqslant n\}$ are independent random variables. In particular, if $f_j \equiv id$ for all j, then $\{X_j, 1 \leqslant j \leqslant n\}$ are independent identically distributed (i.i.d.) random variables, and their common law is the uniform distribution on $[0, 1]$.

Example 1.4.6. *In the previous example, we construct independent random variables through product space. We can also construct independent random variables on the space $(\mathcal{U}, \mathcal{B}, \mathrm{Leb})$ itself.*

For each real number $x \in (0, 1]$, consider its binary expansion:

$$x = \sum_{n=1}^{\infty} \frac{\epsilon_n}{2^n}, \quad \text{each } \epsilon_n \in \{0, 1\}.$$

Such an expansion is unique except when x is of the form $m/2^n$; the set of such x is countable and hence of zero measure, and we can ignore them. Each ϵ_j is a function of x. Define

$$X_j(x) = \epsilon_j(x).$$

Then $\{X_j, j \geqslant 1\}$ are independent. For any n, and any sequence $(c_1, \ldots, c_n) \in \{0, 1\}^n$, we have

$$\mathbb{P}\left[\cap_{j=1}^n \{X_j = c_j\}\right] = \mathrm{Leb}(x : \epsilon_1(x) = c_1, \ldots, \epsilon_n(x) = c_n) = \frac{1}{2^n};$$

$$\mathbb{P}[X_j = c_j] = \mathrm{Leb}(x : \epsilon_j(x) = c_j) = \frac{1}{2}, \quad \prod_{j=1}^n \mathbb{P}[X_j = c_j] = \frac{1}{2^n}.$$

Theorem 1.4.7. *Suppose $\{\mu_j, j \geqslant 1\}$ is a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$. There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of independent random variables $\{X_j, j \geqslant 1\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ such that the law of X_j is μ_j for each j.*

Proof. For each n, let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be a probability space in which there exists a random variable with law μ_n. Indeed, we can take $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n) = (\mathbb{R}, \mathcal{B}, \mu_n)$ and take the random variable *id*. Define the infinite product space

$$\Omega = \bigtimes_{n=1}^{\infty} \Omega_n.$$

A point in Ω will be written as $\omega = (\omega_1, \omega_2, \ldots)$ where $\omega_n \in \Omega_n$. A set E in Ω is a finite-product set if it is of the form

$$E = \bigtimes_{n=1}^{\infty} F_n,$$

where $F_n \in \mathcal{F}_n$ and all but finitely many F_n's are equal to the corresponding Ω_n's. Let $\mathcal{F}_0$ be the collection of the subsets which are the union of a finite number of disjoint finite-product sets, then $\mathcal{F}_0$ is a field. Let $\mathcal{F}$ be the σ-field generated by $\mathcal{F}_0$. This is called the product σ-field and is denoted by

$$\mathcal{F} = \bigtimes_{n=1}^{\infty} \mathcal{F}_n.$$

Define $\mathbb{P}$ on $\mathcal{F}_0$ as follows. First, for each finite-product set $E = \times_n F_n$, define

$$\mathbb{P}[E] = \prod_{n=1}^{\infty} \mathbb{P}_n[F_n].$$

This is well-defined as all but finitely $\mathbb{P}_n[F_n]$'s equal one. Next, for $E \in \mathcal{F}_0$ and $E = \cup_{k=1}^n E_k$ where E_k's are disjoint finite-product sets, set

$$\mathbb{P}[E] = \sum_{k=1}^{n} \mathbb{P}[E_k].$$

This is well-defined: one can check that if E has two representations of the form above, then the two definitions of $\mathbb{P}[E]$ agree.

For $\mathbb{P}$ defined on $\mathcal{F}_0$ as above, it is positive, $\mathbb{P}[\Omega] = 1$, and it has finite additivity. We will use the extension theorem to extend $\mathbb{P}$ to a probability measure on $\mathcal{F}$. To this end, we only need to check the countable additivity: reading. Now, we have defined the probability measure $\mathbb{P}$ on $\mathcal{F}$. This is called the product measure, and it is denoted by

$$\mathbb{P} = \bigtimes_{n=1}^{\infty} \mathbb{P}_n.$$

So far, we have constructed the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define, for each n,

$$X_n(\omega) = \omega_n.$$

Then X_n is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and its law is given by μ_n: for any $B \in \mathcal{B}$, we need to calculate $\mathbb{P}[X_n \in B]$. Note that

$$\{X_n \in B\} = \bigtimes_{j=1}^{\infty} F_j,$$

where $F_n = B$ and $F_j = \Omega_j$ for $j \neq n$. This is a finite-product set, thus

$$\mathbb{P}[X_n \in B] = \prod_{j=1}^{n} \mathbb{P}_j[F_j] = \mathbb{P}_n[B] = \mu_n[B].$$

It remains to show that $\{X_j, j \geqslant 1\}$ are independent: for all n, $\{X_j, 1 \leqslant j \leqslant n\}$ are independent, because, for any Borel sets $\{B_j, 1 \leqslant j \leqslant n\}$, the set

$$\cap_{j=1}^n \{X_j \in B_j\}$$

is a finite-product set and, by definition,

$$\mathbb{P}\left[\cap_{j=1}^n \{X_j \in B_j\}\right] = \prod_{j=1}^{n} \mathbb{P}_j[X_j \in B_j].$$

□

1.5 Independence continued

Collections of sets $\{\mathcal{A}_j, 1 \leqslant j \leqslant n\}$, where $\mathcal{A}_j \subset \mathcal{F}$, are independent if, for any subset $I \subset \{1, \ldots, n\}$ and for any $A_i \in \mathcal{A}_i$, we have

$$\mathbb{P}\left[\cap_{i \in I} A_i\right] = \prod_{i \in I} \mathbb{P}[A_i].$$

We may assume $\Omega \in \mathcal{A}_j$ for each j, then the above definition is equivalent to the following: $\{\mathcal{A}_j, 1 \leqslant j \leqslant n\}$ are independent if

$$\mathbb{P}\left[\cap_{j=1}^n A_j\right] = \prod_{j=1}^n \mathbb{P}[A_j], \quad \text{for any} A_j \in \mathcal{A}_j, 1 \leqslant j \leqslant n.$$

Let us explain how the above definition relates to independent random variables. Suppose X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, denote by

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}.$$

One can check that this is a σ-field, and X is measurable with respect to $\sigma(X)$. Suppose X and Y are two random variables, then X, Y are independent if and only if $\sigma(X), \sigma(Y)$ are independent. In this sense, the above definition is a generalization of Definition 1.4.1.

Definition 1.5.1. *A collection of sets $\mathcal{A}$ is a π-system if it is closed under intersection.*

Note that a field is a π-system, but a π-system may not be a field.

Theorem 1.5.2. *Suppose $\{\mathcal{A}_j, 1 \leqslant j \leqslant n\}$ are independent and each $\mathcal{A}_j$ is a π-system. Denote by $\sigma(\mathcal{A}_j)$ the σ-field generated by $\mathcal{A}_j$ for each j. Then $\{\sigma(\mathcal{A}_j), 1 \leqslant j \leqslant n\}$ are independent.*

Proof. We may assume $\Omega \in \mathcal{A}_j$ for each j, because $\mathcal{A}_j \cup \{\Omega\}$ is still a π-system and $\{\mathcal{A}_j \cup \{\Omega\}, 1 \leqslant j \leqslant n\}$ are still independent. We will prove that $\{\sigma(\mathcal{A}_1), \mathcal{A}_2, \ldots, \mathcal{A}_n\}$ are independent. If this holds, we can repeat the same argument to show $\{\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), \mathcal{A}_3 \ldots, \mathcal{A}_n\}$ are independent. After n iterations, we obtain the conclusion.

Let $A_j \in \mathcal{A}_j$ for $j = 2, \ldots, n$ and set $F = A_2 \cap \cdots \cap A_n$. Define

$$\mathcal{C} = \{A \in \mathcal{F} : \mathbb{P}[A \cap F] = \mathbb{P}[A]\mathbb{P}[F]\}.$$

From the hypothesis, we know $\mathcal{A}_1 \subset \mathcal{C}$. It suffices to show $\sigma(\mathcal{A}_1) \subset \mathcal{C}$.

- $A \in \mathcal{C} \Longrightarrow A^c \in \mathcal{C}$: since $A \in \mathcal{C}$, we have $\mathbb{P}[A \cap F] = \mathbb{P}[A]\mathbb{P}[F]$, hence

$$\mathbb{P}[A^c \cap F] = \mathbb{P}[F] - \mathbb{P}[A \cap F] = \mathbb{P}[A^c]\mathbb{P}[F].$$

- $B_1, \ldots, B_k \in \mathcal{A}_1 \Longrightarrow \cap_j \hat{B}_j \in \mathcal{C}$ for all $\hat{B}_j \in \{B_j, B_j^c\}$: we will show the case of $k = 2$, and the general case can be proved similarly. Since $\mathcal{A}_1$ is a π-system, we know $B_1 \cap B_2 \in \mathcal{A}_1$, thus

$$\begin{aligned}
\mathbb{P}[B_1 \cap B_2^c \cap F] &= \mathbb{P}[B_1 \cap F] - \mathbb{P}[B_1 \cap B_2 \cap F] \\
&= \mathbb{P}[B_1]\mathbb{P}[F] - \mathbb{P}[B_1 \cap B_2]\mathbb{P}[F] \\
&= \mathbb{P}[B_1 \cap B_2^c]\mathbb{P}[F]. \\
\mathbb{P}[B_1^c \cap B_2^c \cap F] &= \mathbb{P}[F] - \mathbb{P}[B_1 \cap F] - \mathbb{P}[B_2 \cap F] + \mathbb{P}[B_1 \cap B_2 \cap F] \\
&= \mathbb{P}[F] - \mathbb{P}[B_1]\mathbb{P}[F] - \mathbb{P}[B_2]\mathbb{P}[F] + \mathbb{P}[B_1 \cap B_2]\mathbb{P}[F] \\
&= \mathbb{P}[B_1^c \cap B_2^c]\mathbb{P}[F].
\end{aligned}$$

  Note that we are NOT claiming that $\mathcal{C}$ is a field.

- $\mathcal{C}$ is a monotone class: if $B_j \in \mathcal{C}$ and $B_j \subset B_{j+1}$, set $B = \cup_j B_j$. By the continuity of the probability, we have

$$\mathbb{P}[B \cap F] = \lim_j \mathbb{P}[B_j \cap F] = \lim_j \mathbb{P}[B_j]\mathbb{P}[F] = \mathbb{P}[B]\mathbb{P}[F].$$

Denote by $\sigma_0(\mathcal{A}_1)$ the field generated by $\mathcal{A}_1$. The first two items guarantee that $\sigma_0(\mathcal{A}_1) \subset \mathcal{C}$. Combining with Lemma 1.1.4, we find $\sigma(\mathcal{A}_1) \subset \mathcal{C}$. $\square$

Example 1.5.3 (Wald's equation)**.** *Let $\{X_n\}$ be i.i.d. with finite mean and $S_n = \sum_{i=1}^n X_i$. For $k \geqslant 1$, let*

$$\mathcal{F}_k = \sigma(X_j, 1 \leqslant j \leqslant k).$$

Suppose N is a random variable taking positive integer values such that

$$\{N \leqslant k\} \in \mathcal{F}_k, \quad \forall k,$$

and $\mathbb{E}[N] < \infty$. Then we have

$$\mathbb{E}[S_N] = \mathbb{E}[X_1]\mathbb{E}[N].$$

In Example 1.5.3, suppose $B \subset \mathbb{R}$ is a measurable set, and define $N = \min\{n : X_n \in B\}$, then we have

$$\{N \leqslant k\} = \cup_{j=1}^k \{X_j \in B\} \in \mathcal{F}_k, \quad \forall k.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[S_N] &= \sum_{k=1}^{\infty} \mathbb{E}[S_k \mathbb{1}_{\{N=k\}}] = \sum_{k=1}^{\infty} \sum_{j=1}^{k} \mathbb{E}[X_j \mathbb{1}_{\{N=k\}}] \\
&= \sum_{j=1}^{\infty} \sum_{k \geqslant j} \mathbb{E}[X_j \mathbb{1}_{\{N=k\}}] = \sum_{j=1}^{\infty} \mathbb{E}[X_j \mathbb{1}_{\{N \geqslant j\}}].
\end{aligned}$$

(Attention: check that the first equal sign holds.) Note that $\{N \geqslant j\} = \{N \leqslant j-1\}^c \in \mathcal{F}_{j-1}$, thus $\{N \geqslant j\}$ is independent of X_j. Therefore,

$$\begin{aligned}
\mathbb{E}[S_N] &= \sum_{j=1}^{\infty} \mathbb{E}[X_j \mathbb{1}_{\{N \geqslant j\}}] = \sum_{j=1}^{\infty} \mathbb{E}[X_j]\mathbb{P}[N \geqslant j] \\
&= \sum_{j=1}^{\infty} \mathbb{E}[X_1]\mathbb{P}[N \geqslant j] = \mathbb{E}[X_1]\mathbb{E}[N].
\end{aligned}$$

□

Example 1.5.4 (Kolmogorov's 0-1 Law)**.** *Let $\{X_n\}$ be a sequence of independent random variables. Let $\mathcal{G}_n = \sigma(X_k, k \geqslant n)$ and $\mathcal{G}_\infty = \cap_{n \geqslant 1} \mathcal{G}_n$. Then $\mathcal{G}_\infty$ is trivial, i.e. for any $A \in \mathcal{G}_\infty$, we have*

$$\mathbb{P}[A] = 0 \text{ or } 1.$$

In Example 1.5.4, define $S_n = \sum_{j=1}^n X_j$. It is clear that the following events are in $\mathcal{G}_\infty$:

$$\lim_n S_n \text{ exists}, \quad \lim_n \frac{S_n}{n} \text{ exists}, \quad \limsup_n \frac{S_n}{n} > 0.$$

Whereas, the following event is not in $\mathcal{G}_\infty$:

$$\limsup_n S_n > 0.$$

Proof. On the one hand, we have $A \in \mathcal{G}_{n+1}$ for any n. Thus A is independent of $\sigma(X_1, \ldots, X_n)$ for any n. Therefore A is independent of $\sigma(X_n, n \geqslant 1)$ (Exercise: why?). On the other hand, A is measurable with respect to $\sigma(X_n, n \geqslant 1)$. Therefore A is independent of itself. This implies $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$, thus $\mathbb{P}[A] \in \{0, 1\}$. □

If you have learned "probability theory" before, you may have encountered the following definition of "independence": Sets $\{A_j, 1 \leqslant j \leqslant n\} \subset \mathcal{F}$ are independent if, for any subset $I \subset \{1, \ldots, n\}$, we have

$$\mathbb{P}[\cap_{i \in I} A_i] = \prod_{i \in I} \mathbb{P}[A_i].$$

This definition is a particular case in the above definition when $\mathcal{A}_j = \{A_j\}$. We end this section with an instructive example.

Example 1.5.5. *Suppose X_1, X_2, X_3 are i.i.d. Bernoulli random variable with parameter $1/2$. Define*

$$A_1 = \{X_2 = X_3\}, \quad A_2 = \{X_1 = X_3\}, \quad A_3 = \{X_1 = X_2\}.$$

Then the events A_1, A_2, A_3 are pairwise independent, but they are not independent: for any $i \neq j$

$$\mathbb{P}[A_i \cap A_j] = \mathbb{P}[X_1 = X_2 = X_3] = 1/4 = \mathbb{P}[A_i]\mathbb{P}[A_j];$$

however,

$$\mathbb{P}[A_1 \cap A_2 \cap A_3] = \mathbb{P}[X_1 = X_2 = X_3] = 1/4, \quad \mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_3] = 1/8.$$

1.6 Sums of independent random variables

For a random variable X, we denote its law by μ_X and its distribution function by F_X. In this section, we study the sum of two independent random variables. To this end, we need the notation of convolution.

Definition 1.6.1. *The convolution of two distribution functions F_1 and F_2 is defined to be*

$$F(x) = \int F_1(x - y) F_2[dy], \quad \forall x.$$

This is still a distribution function, and we denote it by $F = F_1 * F_2$. The corresponding measure is denoted by $\mu = \mu_1 * \mu_2$.

Lemma 1.6.2. *If X and Y are independent, then we have*

$$\mathbb{P}[X + Y \leqslant z] = \iint \mathbb{1}_{\{x+y \leqslant z\}} \, \mu_X[dx] \mu_Y[dy] = F_X * F_Y(z).$$

In particular, if X has density function, then $X + Y$ has density function which is given by

$$p_{X+Y}(z) = \int p_X(z - y) \mu_Y[dy].$$

In this lemma, if both X and Y have density functions, we $X + Y$ has density function

$$p_{X+Y}(z) = p_X * p_Y(z) := \int p_X(z - y) p_Y(y) dy.$$

Example (Example 1.2.9 continued)**.** *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ are i.i.d. whose common law is exponential with parameter $\lambda > 0$, then $S_n := \sum_{j=1}^n X_j$ has the law of gamma distribution $\Gamma(n, \lambda)$ whose density is given by*

$$p(x) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x} \, \mathbb{1}_{\{x \geqslant 0\}} \, .$$

Proof. For $y \geqslant 0$, let us calculate

$$\begin{aligned}
\mathbb{P}[S_n \leqslant y] &= \int_{x_i \geqslant 0, \sum_1^n x_i \leqslant y} \cdots \int \prod_{i=1}^n \lambda e^{-\lambda x_i} dx_i \\
&= \int_{x_i \geqslant 0, \sum_1^n x_i \leqslant y} \cdots \int \lambda^n e^{-\lambda \sum_1^n x_i} dx_1 \cdots dx_n \\
&= \int_{x_i \geqslant 0, \sum_1^{n-1} x_i \leqslant x \leqslant y} \cdots \int \lambda^n e^{-\lambda x} dx_1 \cdots dx_{n-1} dx \qquad (\text{set } x_i = x_i \text{ for } 1 \leqslant i \leqslant n-1 \text{ and } x = \textstyle\sum_1^n x_i) \\
&= \int_0^y \lambda^n e^{-\lambda x} \frac{x^{n-1}}{(n-1)!} dx.
\end{aligned}$$

This gives the desired density function of S_n. ☐

Exercise (Example 1.2.10 continued)**.** *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ are independent and $X_j \sim \mathcal{N}(m_j, \sigma_j^2)$. Then*

$$\sum_{j=1}^n X_j \sim \mathcal{N}\left(\sum_{j=1}^n m_j, \sum_{j=1}^n \sigma_j^2\right).$$

Theorem (Cramér's theorem)**.** *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ are independent real-valued random variables such that $\sum_{j=1}^n X_j$ has a normal distribution, then all of $\{X_j, 1 \leqslant j \leqslant n\}$ must have normal distributions as well.*

Proof. Bonus. ☐

Example (Example 1.2.12 continued)**.** *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ are independent and $X_j \sim \mathrm{Poisson}(\lambda_j)$.*

$$\sum_{j=1}^n X_j \sim \mathrm{Poisson}\left(\sum_{j=1}^n \lambda_j\right).$$

Proof. It suffices to show the conclusion for $n = 2$. Suppose $X \sim \mathrm{Poisson}(\lambda), Y \sim \mathrm{Poisson}(\mu)$ and X, Y are independent. Let us calculate, for $n \geqslant 0$,

$$\begin{aligned}
\mathbb{P}[X + Y = n] &= \sum_{k=0}^n \mathbb{P}[X = k, Y = n-k] \\
&= \sum_{k=0}^n \mathbb{P}[X = k]\mathbb{P}[Y = n-k] \\
&= \sum_{k=0}^n \frac{\lambda^k}{k!} e^{-\lambda} \frac{\mu^{n-k}}{(n-k)!} e^{-\mu} \\
&= \frac{(\lambda+\mu)^n}{n!} e^{-\lambda-\mu}.
\end{aligned}$$

Thus $X + Y \sim \mathrm{Poisson}(\lambda + \mu)$. ☐

Theorem (Raikov's theorem)**.** *Suppose $\{X_j, 1 \leqslant j \leqslant n\}$ are independent non-negative random variables such that $\sum_{j=1}^n X_j$ has a Poisson distribution, then all of $\{X_j, 1 \leqslant j \leqslant n\}$ must have Poisson distribution as well.*

Proof. Bonus. ☐

1.7 Exercises

Exercise 1.7.1. *Suppose X is a random variable. The median of X is any real number m that satisfies*

$$\mathbb{P}[X \leqslant m] \geqslant 1/2, \quad \textit{and} \quad \mathbb{P}[X \geqslant m] \geqslant 1/2.$$

(1) Show that X has at least one median.

(2) Suppose X has finite mean. Show that m is a median of X if and only if m minimizes

$$\{\mathbb{E}[|X-c|] : c \in \mathbb{R}\}.$$

(3) Suppose X is square integrable. Show that $\mathbb{E}[X]$ minimizes

$$\{\mathbb{E}[(X-c)^2] : c \in \mathbb{R}\}.$$

In particular, $\mathbb{E}[(X-c)^2] \geqslant \mathrm{var}(X)$ for all $c \in \mathbb{R}$.

(4) Suppose X is square integrable, and m is a median. Show that

$$(m - \mathbb{E}[X])^2 \leqslant \mathrm{var}(X).$$

Exercise 1.7.2. *Suppose $\{X_n\}$ are i.i.d. with finite second moment. Define $S_n = \sum_{j=1}^n X_j$. Suppose τ is a positive integer-valued random variable that is independent of $\{X_n\}$, and suppose it has finite second moment.*

(1) Show that

$$\mathbb{E}[S_\tau] = \mathbb{E}[X_1]\mathbb{E}[\tau].$$

(2) Show that

$$\mathrm{var}(S_\tau) = \mathbb{E}[\tau]\,\mathrm{var}(X_1) + \mathrm{var}(\tau)\mathbb{E}[X_1]^2.$$

Exercise 1.7.3. *Suppose $\mathbb{E}[|X|] = a > 0$ and $\mathbb{E}[X^2] = 1$. Show that, for any $\lambda \in (0,1)$,*

$$\mathbb{P}[|X| \geqslant \lambda a] \geqslant (1-\lambda)^2 a^2.$$

Exercise 1.7.4. *Suppose that X, Y, Z are independent and that Z has the same law as $X+Y$. Can we say that $Z - X$ has the same law as Y?*

Exercise 1.7.5. *Suppose that Z is a random variable such that Z is independent of itself. Show that Z is almost surely a constant.*

Exercise 1.7.6 (YCMC2012)**.** *Take two points ξ and η independently with respect to the uniform distribution from the unit interval $[0,1]$. Then in general these two points divide the interval $[0,1]$ into three sub-intervals with lengths X, Y and Z.*

(1) What is the probability that X, Y, Z constitute the lengths of three sides of a triangle in the plane?

(2) What are the distributions of X, Y and Z?

Exercise 1.7.7 (YCMC2012)**.** *Suppose that $\{\xi_k : k = 1, 2, \ldots, n\}$ are i.i.d. with uniform distribution on the interval $[0,1]$. Let $Y = \max\{\xi_k : 1 \leqslant k \leqslant n\}$.*

(1) What is the joint distribution of (ξ_1, Y)

(2) Evaluate the probability $\mathbb{P}[\xi_1 = Y]$.

Exercise 1.7.8 (YCMC2012). *Suppose $S = X_1 + X_2 + \cdots + X_n$ is a sum of independent random variables with X_i distributed as Bernoulli with parameter p_i: $\mathbb{P}[X_i = 1] = p_i$ and $\mathbb{P}[X_i = 0] = 1 - p_i$. Show that $\mathbb{P}[S \text{ is even}] = \frac{1}{2}$ if and only if at least one p_i equals $\frac{1}{2}$.*

Exercise 1.7.9 (YCMC2012). *Let $\{X_i\}$ be i.i.d exponential random variables with rate one. Let N be a geometric random variable with success probability $p \in (0,1)$, i.e. $\mathbb{P}[N = k] = (1-p)^{k-1}p, k = 1, 2, \ldots$, and it is independent of $\{X_i\}$. Find the distribution of $\sum_{i=1}^{N} X_i$.*

Exercise 1.7.10 (YCMC2012). *Let X be a random variable with $\mathbb{E}\left[X^2\right] < \infty$, and $Y = |X|$. Assume that X has a Lebesgue density that is symmetric about 0. Show that random variables X and Y are uncorrelated, but they are not independent.*

Exercise 1.7.11 (YCMC2012). *Let X and Y be two random variables with $|Y| > 0$ a.s.. Let $Z = X/Y$.*

(1) Assume the distribution of (X, Y) has a density $p(x, y)$. What is the density function of Z?

(2) Assume X and Y are independent, and $X \sim \mathcal{N}(0, 1)$ and Y is uniform on $(0, 1)$. Give the density function of Z.

Exercise 1.7.12 (YCMC2013). *Suppose that $0 \leqslant X \leqslant 1$ is a random variable. For what distributions of X does $\operatorname{var}(X)$ have the largest value?*

Exercise 1.7.13 (YCMC2013). *Suppose that X and Y are i.i.d with normal distribution $\mathcal{N}(0, 1)$. Give the distribution of*

$$\left(\frac{X}{\sqrt{X^2+Y^2}}, \frac{Y}{\sqrt{X^2+Y^2}}\right).$$

Exercise 1.7.14 (YCMC2013). *Suppose that X, Y, Z are i.i.d with uniform distribution on $[0, 1]$. Show that $(XY)^Z$ also has the uniform distribution on $[0, 1]$.*

Exercise 1.7.15 (YCMC2014). *Given two independent random variables X and Y such that Y has the uniform law on $[0, 1]$ and $\mathbb{P}[X = 0] = \mathbb{P}[X = 1/2] = 1/2$. Show that $W := X + 1/2Y$ has the uniform law on $[0, 1]$.*

Exercise 1.7.16 (YCMC2014). *Suppose Z has the exponential law with parameter one. Let $[Z]$ and $\{Z\}$ be the integral and fractional parts of Z, i.e., $Z = [Z] + \{Z\}$ with $[Z] \in \mathbb{Z}$ and $\{Z\} \in [0, 1)$. Show that $[Z]$ and $\{Z\}$ are independent and determine their laws.*

Exercise 1.7.17 (YCMC2014). *Let X be a real-valued random variable such that for all smooth functions $f : \mathbb{R} \to \mathbb{R}$ with compact support we have $\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)]$. Show that X has the standard normal distribution.*

Exercise 1.7.18 (YCMC2015). *Suppose X and Y are independent integrable random variables and $\mathbb{E}[X] = 0$. Show that $\mathbb{E}[|X + Y|] \geqslant \mathbb{E}[|Y|]$.*

Exercise 1.7.19 (YCMC2015). *Suppose $\{X_i : i \in \mathbb{N}\}$ are i.i.d. with exponential law of parameter one. For $x > 0$, define $N(x) = \inf\{n : \sum_{i=1}^{n} X_i > x\}$. Calculate the mean of $N(x)$.*

Exercise 1.7.20 (YCMC2016). *Choose 2016 points on the circle $x^2 + y^2 = 1$ at random. Interpret them as cuts that divide the circle into 2016 arcs. Compute the expected length of the arc that contains the point $(1, 0)$. How about the variance ?*

Exercise 1.7.21 (YCMC2016). *Let $N \geqslant 2$ be an integer, and let X be a random variable taking values in $\{0, 1, 2, ...\}$ such that $\mathbb{P}[X \equiv k (\mathrm{mod} N)] = \frac{1}{N}$ for all $k \in \{0, 1, ..., N-1\}$. Compute $\mathbb{E}\left[e^{i(2\pi m)X/N}\right]$ for all integers $m \geqslant 1$.*

Exercise 1.7.22 (YCMC2016). *Let $b > a > 0$ be real numbers. Let X be a random variable taking values in $[a, b]$, and let $Y = \frac{1}{X}$. Determine the set of all possible values of $\mathbb{E}[X] \times \mathbb{E}[Y]$.*

2 Convergence Concepts

We will study the convergence of a sequence of random variables $\{X_n, n \geqslant 1\}$ (or $\{X_n\}$ for short). When we say "convergence", we mean "convergence to a finite limit". Recall that we assumed the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to be complete. In this section, we will introduce the following different concepts of convergence:

- almost sure convergence,
- convergence in probability,
- convergence in L^p with $p \in [1, \infty)$,
- convergence in distribution (also called convergence in law, weak convergence);

and we will study the relation between them.

2.1 Convergence: almost sure, in probability, and in L^p

Definition 2.1.1 (Almost sure convergence (a.s.))**.** *The sequence of random variables $\{X_n\}$ converges a.s. to the random variable X if there exists a null set $\mathcal{N}$ such that*

$$\lim_n X_n(\omega) = X(\omega), \quad \forall \omega \in \Omega \backslash \mathcal{N}. \tag{2.1.1}$$

Lemma 2.1.2. *The sequence $\{X_n\}$ converges a.s. to X if and only if, for any $\epsilon > 0$, we have*

$$\lim_{m \to \infty} \mathbb{P}[|X_n - X| \leqslant \epsilon, \ \forall n \geqslant m] = 1. \tag{2.1.2}$$

Proof. From (2.1.1) to (2.1.2). Suppose $\{X_n\}$ converges to X on Ω_0 with $\mathbb{P}[\Omega_0] = 1$. For $\epsilon > 0$, define

$$A_m(\epsilon) = \cap_{n=m}^{\infty} \{|X_n - X| \leqslant \epsilon\}.$$

The sequence of events $\{A_m(\epsilon)\}$ is increasing in m, and we find

$$\Omega_0 \subset \cup_m A_m(\epsilon).$$

By monotone convergence theorem, we have $\lim_m \mathbb{P}[A_m(\epsilon)] = 1$ as desired.

From (2.1.2) to (2.1.1). Define $A_m(\epsilon)$ in the same way as above. Set

$$A = \cap_{k \geqslant 1} \cup_{m \geqslant 1} A_m(2^{-k}).$$

By the hypothesis, we have $\lim_m \mathbb{P}[A_m(2^{-k})] = 1$. Since the events $\{A_m(2^{-k})\}$ is increasing in m, we have $\mathbb{P}[\cup_m A_m(2^{-k})] = 1$. This gives that $\mathbb{P}[A] = 1$. One can check that $\{X_n\}$ converges to X on A. This completes the proof. □

Definition 2.1.3 (Convergence in probability)**.** *The sequence $\{X_n\}$ converges in probability to the random variable X if, for every $\epsilon > 0$, we have*

$$\lim_n \mathbb{P}[|X_n - X| > \epsilon] = 0.$$

As a consequence of Lemma 2.1.2, we see that almost sure convergence implies convergence in probability. But the converse is false, see Example 2.1.6

Definition 2.1.4 (Convergence in L^p)**.** *Assume $p \geqslant 1$. The sequence $\{X_n\}$ converges in L^p to the random variable X if $X_n \in L^p, X \in L^p$ and*

$$\lim_n \mathbb{E}[|X_n - X|^p] = 0.$$

Lemma 2.1.5. *Assume $p > 0$. If $X_n \to X$ in L^p, then $X_n \to X$ in probability.*

Proof. Assume $X_n \to X$ in L^p. For any $\epsilon > 0$, we have

$$\mathbb{P}[|X_n - X| > \epsilon] \leqslant \epsilon^{-p}\mathbb{E}[|X_n - X|^p] \to 0.$$

□

Example 2.1.6. *Almost sure convergence does not imply convergence in L^p.*

In $(\mathcal{U}, \mathcal{B}, \text{Leb})$, define

$$X_n(\omega) = \begin{cases} 2^n, & \text{if } \omega \in (0, 1/n), \\ 0, & \text{otherwise.} \end{cases}$$

We have $X_n \to 0$ almost surely, but $\mathbb{E}[X_n^p] = 2^{np}/n \to \infty$ as $n \to \infty$ for any $p > 0$.

Example 2.1.7. *Convergence in L^p does not imply almost sure convergence.*

In $(\mathcal{U}, \mathcal{B}, \text{Leb})$, let $\varphi_{k,j}$ be the indicator function of the interval

$$\left(\frac{j-1}{k}, \frac{j}{k}\right), \quad k \geqslant 1,\, 1 \leqslant j \leqslant k.$$

Order these functions first according to k increasing, and then for each k according to j increasing, into one sequence φ_{k_n, j_n}. Set $X_n = \varphi_{k_n, j_n}$. Then we have

$$\mathbb{E}[X_n^p] = \frac{1}{k_n} \to 0.$$

So $X_n \to 0$ in L^p. But $\{X_n\}$ does not converge. For each ω and every k, there exists j such that $\varphi_{kj}(\omega) = 1$. Thus there exist infinitely many n such that $X_n(\omega) = 1$. Similarly, there exist infinitely many n such that $X_n(\omega) = 0$. Thus $\{X_n(\omega)\}$ does not converge. In other words, the set on which $\{X_n\}$ converges is empty.

Example 2.1.8 (L^2 weak law)**.** *Let $\{X_n\}$ be independent random variables with $\mathbb{E}[X_i] = m$ and $\text{var}(X_i) \leqslant C < \infty$. Set $S_n = \sum_{j=1}^n X_j$. Then*

$$\frac{S_n}{n} \to m, \quad \text{in } L^2.$$

Proof. We observe that

$$\mathbb{E}[(S_n/n - m)^2] = \text{var}(S_n/n) = \frac{1}{n^2}\,\text{var}(S_n) = \frac{1}{n^2}\sum_{j=1}^n \text{var}(X_j) \leqslant \frac{C}{n} \to 0.$$

□

Example 2.1.9 (Polynomial approximation)**.** *Let f be a continuous function on $[0,1]$. Define the polynomial:*

$$f_n(x) = \sum_{j=0}^n \binom{n}{j} x^j (1-x)^{n-j} f(j/n).$$

This is called Bernstein polynomial of degree n associated to f. Then we have

$$\sup_{x \in [0,1]} |f_n(x) - f(x)| \to 0, \quad n \to \infty.$$

Proof. Suppose $\{X_j, j \geqslant 1\}$ are i.i.d. Bernoulli variable with parameter $p \in [0,1]$: $\mathbb{P}[X_j = 1] = p, \mathbb{P}[X_j = 0] = 1-p$, and set $S_n = \sum_{j=1}^n X_j$. Then we find

$$\mathbb{E}[X_j] = p, \quad \mathrm{var}(X_j) = p(1-p), \quad \mathbb{E}[S_n] = np, \quad \mathrm{var}(S_n) = np(1-p).$$

Moreover,

$$\mathbb{P}[S_n = j] = \binom{n}{j} p^j (1-p)^{n-j}, \quad \text{thus } f_n(p) = \mathbb{E}[f(S_n/n)].$$

Note that f is bounded: $|f| \leqslant M$, and it is uniformly continuous: for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| \leqslant \epsilon$ as long as $|x-y| \leqslant \delta$. Thus

$$|f_n(p) - f(p)| \leqslant \mathbb{E}[|f(S_n/n) - f(p)|] \leqslant \epsilon + 2M\mathbb{P}[|S_n/n - p| > \delta].$$

Note that

$$\mathbb{P}[|S_n/n - p| > \delta] \leqslant \frac{\mathrm{var}(S_n)}{n^2\delta^2} = \frac{p(1-p)}{n\delta^2} \leqslant \frac{1}{4n\delta^2}.$$

Thus

$$|f_n(p) - f(p)| \leqslant \epsilon + \frac{M}{2n\delta^2}.$$

This gives the conclusion. □

Example 2.1.10 (Coupon collecting)**.** *Let $\{X_n\}$ be i.i.d. uniform on $\{1, 2, \ldots, N\}$. Let T_N be the first time n that $\#\{X_1, \ldots, X_n\} = N$. Then*

$$\frac{T_N}{N\log N} \to 1, \quad \textit{in } L^2.$$

Proof. For $1 \leqslant k \leqslant N$, define τ_k to be the first time n that $\#\{X_1, \ldots, X_n\} = k$. It is clear that $\tau_1 = 1$, and that $Y_k := \tau_k - \tau_{k-1}$ satisfies geometric distribution:

$$\mathbb{P}[Y_k \geqslant m] = \left(\frac{k-1}{N}\right)^{m-1}.$$

Moreover, $\{Y_k, 1 \leqslant k \leqslant N\}$ are independent and $T_N = \sum_{k=1}^N Y_k$. Let us calculate:

$$\mathbb{E}[Y_k] = \left(1 - \frac{k-1}{N}\right)^{-1}, \quad \mathrm{var}(Y_k) \leqslant \left(1 - \frac{k-1}{N}\right)^{-2}.$$

$$\mathbb{E}[T_N] = N\sum_{k=1}^N \frac{1}{k} \sim N\log N, \quad \mathrm{var}(T_N) \leqslant N^2 \sum_{k=1}^N \frac{1}{k^2} \leqslant CN^2.$$

Denote by

$$err_N := N\sum_{k=1}^N \frac{1}{k} - N\log N = o(N\log N).$$

We have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{T_N}{N\log N} - 1\right)^2\right] &= \frac{\mathbb{E}[(T_N - N\log N)^2]}{(N\log N)^2} \\
&= \frac{\mathrm{var}(T_N) + err_N^2}{(N\log N)^2} \\
&\leqslant \frac{CN^2 + err_N^2}{(N\log N)^2} \to 0, \quad N \to \infty.
\end{aligned}$$

□

2.2 Borel Cantelli lemma

Definition 2.2.1. *Let $\{E_n\}$ be a sequence of subsets in $\mathcal{F}$. Define*

$$\limsup_n E_n = \cap_{m=1}^{\infty} \cup_{n \geqslant m} E_n, \quad \liminf_n E_n = \cup_{m=1}^{\infty} \cap_{n \geqslant m} \mathbb{E}_n.$$

Note that

$$\liminf_n E_n = \left(\limsup_n E_n^c\right)^c,$$

so that, in a sense, one of the two notions suffices. We will focus on $\limsup_n E_n$.

Lemma 2.2.2. *A point belongs to $\limsup_n E_n$ if and only if it belongs to infinitely many terms of the sequence $\{E_n, n \geqslant 1\}$. In more intuitive language: the event $\limsup_n E_n$ occurs if and only if the events E_n occur infinitely many often, and we write*

$$\{\limsup_n E_n\} = \{E_n, i.o.\}.$$

Proof. If ω belongs to infinitely many E_n's, then it belongs to

$$F_m := \cup_{n \geqslant m} E_n, \quad \text{for every } m.$$

Thus it belongs to

$$\cap_{m \geqslant 1} F_m = \limsup_n E_n.$$

Conversely, if $\omega \in \limsup_n E_n$, then it belongs to F_m for every m. If ω only belongs to finite many E_n's, then there exists $N(\omega)$ such that $\omega \notin E_n$ for $n \geqslant N(\omega)$, then $\omega \notin F_m$ for $m \geqslant N(\omega)$, contradiction. □

As an illustration of the convenience of the new notions, we may restate Lemma 2.1.2 as follows.

Lemma (Lemma 2.1.2 bis)**.**

$$X_n \to X \; a.s. \quad \textit{if and only if} \quad \mathbb{P}[|X_n - X| > \epsilon \; i.o.] = 0, \forall \epsilon > 0.$$

Theorem 2.2.3 (Borel Cantelli lemma)**.**

- *For arbitrary sequence $\{E_n\}$, we have*

$$\sum_n \mathbb{P}[E_n] < \infty \Longrightarrow \mathbb{P}[E_n \; i.o.] = 0. \tag{2.2.1}$$

- *If the events $\{E_n\}$ are independent, we have*

$$\sum_n \mathbb{P}[E_n] = \infty \Longrightarrow \mathbb{P}[E_n \; i.o.] = 1. \tag{2.2.2}$$

Proof of Theorem 2.2.3—(2.2.1)*.* Since $\sum_n \mathbb{P}[E_n] < \infty$, we have

$$\mathbb{P}[F_m] \leqslant \sum_{n \geqslant m} \mathbb{P}[E_n] \to 0, \quad \text{as } m \to \infty.$$

By monotone convergence theorem, we have

$$\mathbb{P}[E_n \; i.o.] = \lim_m \mathbb{P}[F_m] = 0.$$

□

Corollary 2.2.4. *Convergence in probability implies almost sure convergence along subsequence.*

Proof. Suppose $X_n \to X$ in probability. Then we have, for any $\epsilon > 0$,

$$\lim_n \mathbb{P}[|X_n - X| > \epsilon] = 0.$$

For each k, there exists n_k such that

$$\mathbb{P}\left[|X_{n_k} - X| > 2^{-k}\right] \leqslant 2^{-k},$$

and that $n_k \uparrow \infty$ as $k \uparrow \infty$. Thus

$$\sum_k \mathbb{P}\left[|X_{n_k} - X| > 2^{-k}\right] < \infty.$$

By Borel Cantelli lemma, we have

$$\mathbb{P}\left[|X_{n_k} - X| > 2^{-k} \ i.o.\right] = 0.$$

In other words, there exits Ω_0 with $\mathbb{P}[\Omega_0] = 1$ such that the following holds. For each $\omega \in \Omega_0$, there exists $K(\omega)$ such that

$$|X_{n_k} - X| \leqslant 2^{-k}, \quad \forall k \geqslant K(\omega).$$

Then it is immediate that $X_{n_k}(\omega) \to X(\omega)$ as $k \to \infty$. Hence $X_{n_k} \to X$ on Ω_0. □

We will give another application of Borel-Cantelli lemma.

Example 2.2.5. *Suppose $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[X_j] = m$ and $\mathbb{E}[X_j^4] < \infty$. Set $S_n = \sum_{j=1}^n X_j$. Then*

$$\frac{S_n}{n} \to m, \quad a.s.$$

Proof. We may assume $m = 0$. Let us calculate $\mathbb{E}[S_n^4]$:

$$\mathbb{E}[S_n^4] = \mathbb{E}\left[\sum_{1 \leqslant i,j,k,l \leqslant n} X_i X_j X_k X_l\right] = n\mathbb{E}[X_1^4] + 3(n^2 - n)\mathbb{E}[X_1^2]^2 \leqslant Cn^2.$$

Chebyshev's inequality gives

$$\mathbb{P}\left[\frac{|S_n|}{n} > \epsilon\right] \leqslant \frac{C}{n^2\epsilon^4}.$$

Summing over n is finite. Thus Borel-Cantelli lemma implies

$$\mathbb{P}\left[\frac{|S_n|}{n} > \epsilon \ i.o.\right] = 0.$$

Thus

$$\mathbb{P}\left[\cup_{k=1}^\infty \left\{\frac{|S_n|}{n} > 2^{-k} \ i.o.\right\}\right] = 0.$$

This gives $S_n/n \to 0$ almost surely. □

Proof of Theorem 2.2.3—(2.2.2). It suffices to show

$$\mathbb{P}[\liminf_n E_n^c] = \lim_m \mathbb{P}[\cap_{n \geqslant m} E_n^c] = 0. \tag{2.2.3}$$

Since $\{E_n^c\}$ are independent, we have, for $m \leqslant m'$,

$$\mathbb{P}[\cap_{n=m}^{m'} E_n^c] = \prod_{n=m}^{m'} \mathbb{P}[E_n^c] = \prod_{n=m}^{m'} (1 - \mathbb{P}[E_n])$$
$$\leqslant \prod_{n=m}^{m'} \exp(-\mathbb{P}[E_n]) = \exp\left(-\sum_{n=m}^{m'} \mathbb{P}[E_n]\right).$$

Let $m' \to \infty$, the right hand side goes to zero, combining with monotone convergence theorem, we have

$$\mathbb{P}[\cap_{n\geqslant m} E_n^c] = 0,$$

which gives (2.2.3) as desired. □

Theorem 2.2.6. *The implication* (2.2.2) *remains true if the events* $\{E_n\}$ *are pairwise independent.*

Proof. Denote by I_n the indicator function of E_n. Set

$$p_n = \mathbb{E}[I_n] = \mathbb{P}[E_n], \quad S_n = \sum_{j=1}^{n} I_j.$$

Then the hypothesis is equivalent to

$$\mathbb{E}[S_n] \to \infty, \quad \text{as } n \to \infty, \tag{2.2.4}$$

and the conclusion is equivalent to

$$\mathbb{P}\left[\lim_n S_n = \infty\right] = 1. \tag{2.2.5}$$

By Chebyshev's inequality, we have, for any $A > 0$,

$$\mathbb{P}[|S_n - \mathbb{E}[S_n]| \leqslant A\sigma(S_n)] \geqslant 1 - \frac{\sigma(S_n)^2}{A^2\sigma(S_n)^2} = 1 - \frac{1}{A^2}.$$

By pairwise independence, we have

$$\sigma(S_n)^2 = \sum_{j=1}^{n} \mathrm{var}(I_j) = \sum_{j=1}^{n} (p_j - p_j^2) \leqslant \mathbb{E}[S_n].$$

Since $\mathbb{E}[S_n] \to \infty$, we find $\sigma(S_n) = o(\mathbb{E}[S_n])$. Thus, there exists $n_0(A)$ such that,for $n \geqslant n_0(A)$, we have

$$\mathbb{P}\left[S_n \geqslant \frac{1}{2}\mathbb{E}[S_n]\right] \geqslant 1 - \frac{1}{A^2}.$$

Since $\{S_n\}$ is an increasing sequence, we have

$$\mathbb{P}\left[\lim_m S_m \geqslant \frac{1}{2}\mathbb{E}[S_n]\right] \geqslant 1 - \frac{1}{A^2}.$$

Let $n \to \infty$, we have

$$\mathbb{P}\left[\lim_m S_m = \infty\right] \geqslant 1 - \frac{1}{A^2}.$$

Let $A \to \infty$, we obtain (2.2.5). □

2.3 Weak convergence

Definition 2.3.1. *A sequence of measures $\{\mu_n\}$ converges weakly to a measure μ if*

$$\mu_n((a,b]) \to \mu((a,b]), \quad \text{for all continuity points } a, b \text{ of } \mu. \tag{2.3.1}$$

We denote by $\mu_n \Longrightarrow \mu$.

We have several remarks about this definition.

- The requirement that the convergence only holds for continuity points of μ is essential. For instance, suppose μ_n is uniform on $(0, 1/n)$ with total mass one, the sequence converges weakly to δ_0. However, we do not have the convergence at the discontinuous point $x = 0$.

- Denote by F_n the distribution function of μ_n and by F the distribution function of μ. Then (2.3.1) is equivalent to

  $$F_n(x) \to F(x), \quad \text{for all continuity point } x \text{ of } F.$$

  Denote by $\mathcal{C}_F$ the set of continuity points of F. Note that $\mathbb{R}\backslash\mathcal{C}_F$ is countable: Since F is increasing and right-continuous, if F is discontinuous at two distinct points x and y, the two open intervals $(F(x-), F(x))$ and $(F(y-), F(y))$ are disjoint. This implies that the set of discontinuous points are countable.

- In the definition, we do not require $\{\mu_n\}$ or μ to be probability measures. Usually, we are interested in the case when $\{\mu_n\}$ are probability measures. However, even if $\{\mu_n\}$ are probability measures, the limit may nolonger be a probability measure. This is why we give the definition in a general form.

- The set of probability measures is "compact", i.e. any sequence of probability measures has a weakly convergent subsequence, but the limiting measure may nolonger be "probability measure". To give the precise statement, we need the following definition: a measure μ on $(\mathbb{R}, \mathcal{B})$ is a subprobability measure if $\mu(\mathbb{R}) \leqslant 1$.

Proposition 2.3.2 (Helly's extraction principle)**.** *Given any sequence of subprobability measures, there is a subsequence that converges weakly to a subprobability measure.*

Proof. It is more convenient to work with the distribution functions: for $n \geqslant 1$, define

$$F_n(x) = \mu_n(-\infty, x], \quad \forall x \in \mathbb{R}.$$

Since μ_n is a subprobability measure, the function F_n is increasing and right-continuous with $F_n(-\infty) = 0$ and $F_n(+\infty) = \mu_n(\mathbb{R}) \leqslant 1$.

Denote by $\mathbb{Q}$ the set of rational numbers and enumerate it as $\{r_k, k \geqslant 1\}$. Consider the sequence $\{F_n(r_1), n \geqslant 1\}$. It is bounded, and hence there exists convergent subsequence, denoted by $F_{1n}(r_1) \to G(r_1)$. Consider the sequence $\{F_{1n}(r_2), n \geqslant 1\}$. It is bounded, and contains a convergent subsequence, denoted by $F_{2n}(r_2) \to G(r_2)$. Note that $\{F_{2n}\}$ is a subsequence of $\{F_{1n}\}$, thus $F_{2n}(r_1) \to G(r_1)$. Continue in this way, we obtain

$$\begin{array}{llllll}
F_{11}, & F_{12}, & \cdots, & F_{1n}, & \cdots, & \text{converging at } r_1; \\
F_{21}, & F_{22}, & \cdots, & F_{2n}, & \cdots, & \text{converging at } r_1, r_2; \\
\cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots; \\
F_{n1}, & F_{n2}, & \cdots, & F_{nn}, & \cdots, & \text{converging at } r_1, r_2, \ldots, r_n; \\
\cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots.
\end{array}$$

Choose the diagonal sequence $\{F_{nn}, n \geqslant 1\}$. We assert that it converges along all r_j's. Define

$$G(r_j) = \lim_n F_{nn}(r_j), \quad \forall j \geqslant 1.$$

It is clear that the function G is defined on $\mathbb{Q}$ and it is increasing on $\mathbb{Q}$. Set

$$F(x) = \inf\{G(r) : x < r \in \mathbb{Q}\}, \quad \forall x \in \mathbb{R}.$$

The function F has the following properties.

- First, F is increasing because G is increasing on $\mathbb{Q}$.
- Second, F is right-continuous. To see this, one needs to check that, for each $x \in \mathbb{R}$, for any $\epsilon > 0$, there exists $\delta > 0$ such that $F(y) \leqslant F(x) + \epsilon$ as long as $y \leqslant x + \delta$. This is true because, by the definition, one can find $q \in \mathbb{Q}$ such that $x < q$ and $F(x) \leqslant G(q) \leqslant F(x) + \epsilon$. Then for any $y \leqslant q$, we have $F(y) \leqslant G(q) \leqslant F(x) + \epsilon$.
- Finally, we will show that
$$\lim_n F_{nn}(x) = F(x), \quad \forall x \in \mathcal{C}_F. \tag{2.3.2}$$
For any $p < p' < x < q' < q$ with $p, p', q, q' \in \mathbb{Q}$, we have
$$\begin{aligned} F(p) \leqslant G(p') = \lim_n F_{nn}(p') &\leqslant \liminf_n F_{nn}(x) \\ &\leqslant \limsup_n F_{nn}(x) \leqslant \lim_n F_{nn}(q') = G(q') \leqslant F(q). \end{aligned}$$
Thus, for any $p < x < q$ with $p, q \in \mathbb{Q}$, we have
$$F(p) \leqslant \liminf_n F_{nn}(x) \leqslant \limsup_n F_{nn}(x) \leqslant F(q).$$
Let $p \uparrow x$ and $q \downarrow x$ with $p, q \in \mathbb{Q}$, since x is a continuity point of F, we obtain (2.3.2).

Now, let μ be the unique measure on $\mathbb{R}$ such that $\mu(a, b] = F(b) - F(a)$ for F's continuity points a and b. By (2.3.2), we see that μ_{nn} converges weakly to μ as desired. □

Proposition 2.3.3. *Suppose $\{\mu_n\}$ is a sequence of subprobability measures. If every weakly convergent subsequence converges to the same limit μ, then $\mu_n \Longrightarrow \mu$.*

Proof. We prove by contradiction. If $\{\mu_n\}$ do not converge to μ, there exists continuity points a, b of μ such that $\mu_n(a, b] \nrightarrow \mu(a, b]$. Consider the sequence $\{\mu_n(a, b], n \geqslant 1\}$, it is bounded, and hence contains a convergent subsequence, denoted by $\mu_{n_k}(a, b] \to L \neq \mu(a, b]$. Consider the sequence $\{\mu_{n_k}\}$, it is a sequence of subprobability measures, by Proposition 2.3.2, there exists a convergent subsequence, denoted by $\mu_{n_{k_j}}$. By the hypothesis, we have $\mu_{n_{k_j}} \Longrightarrow \mu$. In particular, $\mu_{n_{k_j}}(a, b] \to \mu(a, b]$. This is a contradiction. □

In the above two propositions, we work with subprobability measures, and the reason is that, the subsequential limit of sequence of probability measures may nolonger be a probability measure. If we require the subsequential limit to be a probability measure, we need to impose the tightness on the sequence of probability measures.

Definition 2.3.4. *A family of probability measures $\{\mu_\alpha, \alpha \in \mathcal{A}\}$ is tight if, for any $\epsilon > 0$, there exists a finite interval I,*

$$\inf_{\alpha \in \mathcal{A}} \mu_\alpha(I) \geqslant 1 - \epsilon.$$

Theorem 2.3.5. *Let $\{\mu_\alpha, \alpha \in \mathcal{A}\}$ be a family of probability measures. In order that any sequence contains a subsequence which converges weakly to a probability measure, it is necessary and sufficient that the family is tight.*

This statement can also be phrased as follows: a family of probability measures is relatively compact if and only if it is tight.

Proof. Suppose the family is tight. Proposition 2.3.2 asserts that any sequence $\{\mu_n\}$ contains a convergent subsequence $\mu_{n_k} \Longrightarrow \mu$. It remains to show that $\mu(\mathbb{R}) = 1$. For any $\epsilon > 0$, since the family is tight, there is a finite interval I such that $\mu_{n_k}(I) \geqslant 1 - \epsilon$. We can find two continuity points a, b of μ such that $I \subset (a, b)$. Then we have

$$\mu(a, b] = \lim_k \mu_{n_k}(a, b] \geqslant \lim_k \mu_{n_k}(I) \geqslant 1 - \epsilon.$$

Thus $\mu(\mathbb{R}) \geqslant 1 - \epsilon$. Let $\epsilon \to 0$, we have $\mu(\mathbb{R}) = 1$.

Conversely, we prove by contradiction. If the family is not tight, then there exists $\epsilon_0 > 0$ such that for each interval $I_n = (-n, n)$, there exists μ_n in the family such that

$$\mu_n(I_n) \leqslant 1 - \epsilon_0, \quad \forall n.$$

Proposition 2.3.2 asserts that $\{\mu_n\}$ contains a convergent subsequence $\mu_{n_k} \Longrightarrow \mu$. On the one hand, μ is a probability measure by the hypothesis. Thus there exist continuity points a, b of μ such that $\mu(a, b] \geqslant 1 - \epsilon_0/2$. Thus

$$\lim_k \mu_{n_k}(a, b] = \mu(a, b] \geqslant 1 - \epsilon_0/2.$$

On the other hand, since $n_k \to \infty$, we have $I_{n_k} \to \mathbb{R}$, thus $(a, b] \subset I_{n_k}$ for k large enough. Therefore,

$$\lim_k \mu_{n_k}(a, b] \leqslant \liminf_k \mu_{n_k}(I_{n_k}) \leqslant 1 - \epsilon_0.$$

Contradiction. $\square$

Next, we will discuss other criterion of the weak convergence. This has to do with classes of continuous functions on $\mathbb{R}$. We first collection some related notations.

$$\begin{aligned}
C_c &= \{\text{continuous functions which vanish outside a compact set}\},\\
C_0 &= \{\text{continuous functions } f \text{ such that } f(x) \to 0 \text{ as } |x| \to \infty\},\\
C_b &= \{\text{bounded continuous functions}\},\\
C &= \{\text{continuous functions}\}.
\end{aligned}$$

We have $C_c \subset C_0 \subset C_b \subset C$. It is well known that C_0 is the closure of C_c with respect to uniform convergence.

Proposition 2.3.6. *Suppose $\{\mu_n\}$ and μ are probability measures. Then $\mu_n \Longrightarrow \mu$ if and only if*

$$\lim_n \int f(x)\mu_n[dx] \to \int f(x)\mu[dx], \quad \forall f \in C_b. \tag{2.3.3}$$

Proof. Suppose (2.3.3) holds. For any continuity points a, b of μ, for any $\epsilon > 0$, let $f_\epsilon = 1$ on (a, b), and $f_\epsilon = 0$ on $(\infty, a - \epsilon) \cup (b + \epsilon, \infty)$, and f_ϵ is linear on $[a - \epsilon, a]$ and $[b, b + \epsilon]$. Then $f_\epsilon \in C_b$, and

$$\limsup_n \mu_n(a, b] \leqslant \lim_n \int f_\epsilon(x)\mu_n[dx] = \int f_\epsilon(x)\mu[dx] \leqslant \mu(a - \epsilon, b + \epsilon).$$

Let $\epsilon \to 0$, since a, b are continuity points of μ, we have

$$\limsup_n \mu_n(a, b] \leqslant \mu(a, b].$$

Similarly, we can show

$$\liminf_n \mu_n(a, b] \geqslant \mu(a, b].$$

These give $\mu_n \Longrightarrow \mu$.

Suppose $\mu_n \implies \mu$. Denote by $\mathcal{C}_\mu$ the set of continuity points of μ. By the definition, we know that (2.3.3) holds for $f = \mathbb{1}_{(a,b]}$ with $a, b \in \mathcal{C}_\mu$. We first show that (2.3.3) holds for $f \in C_c$ by constructing approximations of f. Since $f \in C_c$, we may assume $supp(f) \subset [a, b]$. Note that f is uniformly continuous, thus, for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| \leqslant \epsilon$ as long as $|x - y| \leqslant \delta$. Since $\mathcal{C}_\mu$ is dense, there exists N and $\{a_j, 1 \leqslant j \leqslant N\}$ such that

$$a_j \in \mathcal{C}_\mu, \quad 1 \leqslant j \leqslant N; \quad a_j \leqslant a_{j+1} < a_j + \delta, \quad 1 \leqslant j \leqslant N-1; \quad a_1 < a, b < a_N.$$

Define f_ϵ as follows:

$$f_\epsilon = \sum_{j=1}^{N-1} f(a_j)\, \mathbb{1}_{(a_j, a_{j+1}]}.$$

Since f is uniformly continuous, we have $\sup_{x\in\mathbb{R}} |f(x) - f_\epsilon(x)| \leqslant \epsilon$. Thus

$$\begin{aligned}\left|\int f d\mu_n - \int f d\mu\right| &\leqslant \int |f - f_\epsilon| d\mu_n + \left|\int f_\epsilon d\mu_n - \int f_\epsilon d\mu\right| + \int |f - f_\epsilon| d\mu \\ &\leqslant 2\epsilon + \left|\int f_\epsilon d\mu_n - \int f_\epsilon d\mu\right|,\end{aligned}$$

where

$$\int f_\epsilon d\mu_n - \int f_\epsilon d\mu = \sum_{j=1}^{N-1} f(a_j)\left(\mu_n(a_j, a_{j+1}] - \mu(a_j, a_{j+1}]\right) \to 0$$

because $a_j \in \mathcal{C}_\mu$. This implies

$$\limsup_n \left|\int f d\mu_n - \int f d\mu\right| \leqslant 2\epsilon.$$

Let $\epsilon \to 0$, we obtain (2.3.3) for $f \in C_c$.

Generally, consider $f \in C_b$. Suppose $\sup_x |f(x)| \leqslant M$. For any $\epsilon > 0$, there exist $a, b \in \mathcal{C}_\mu$ such that $\mu((a, b]^c) \leqslant \epsilon$ and $\mu_n((a, b]^c) \leqslant 2\epsilon$ for n large enough. Define $f_\epsilon = f$ on (a, b), $f_\epsilon = 0$ on $(-\infty, a - \epsilon) \cup (b + \epsilon, \infty)$, and f_ϵ linear on $(a - \epsilon, a)$ and $(b, b + \epsilon)$. Then $f_\epsilon \in C_c$, $\sup_x |f(x) - f_\epsilon(x)| \leqslant 2M$. We have

$$\begin{aligned}\left|\int f d\mu_n - \int f d\mu\right| &\leqslant \int |f - f_\epsilon| d\mu_n + \left|\int f_\epsilon d\mu_n - \int f_\epsilon d\mu\right| + \int |f - f_\epsilon| d\mu \\ &\leqslant 4M\epsilon + \left|\int f_\epsilon d\mu_n - \int f_\epsilon d\mu\right| + 2M\epsilon,\end{aligned}$$

where

$$\left|\int f_\epsilon d\mu_n - \int f_\epsilon d\mu\right| \to 0,$$

because $f_\epsilon \in C_c$. This gives

$$\limsup_n \left|\int f d\mu_n - \int f d\mu\right| \leqslant 6M\epsilon.$$

Let $\epsilon \to 0$, we obtain (2.3.3) for $f \in C_b$. □

Remark 2.3.7. *Assume the same assumption as in Proposition 2.3.6. From the above proof, we see that $\mu_n \implies \mu$ if and only if* (2.3.3) *holds for all $f \in C_c$.*

A function f on $\mathbb{R}$ is lower semicontinuous if

$$f(x) \leqslant \liminf_{y\to x, y\neq x} f(y), \quad \forall x.$$

A function f is upper semicontinuous if $-f$ is lower semicontinuous. There are several equivalent definitions of lower/upper semicontinuous, but the following characterization is the most useful: f is bounded and lower semicontinuous if and only if there exists a sequence $f_n \in C_b$ which increases to f everywhere.

Corollary 2.3.8. *Suppose $\{\mu_n\}$ and μ are probability measures. Then the following statements are equivalent.*

$$\mu_n \Longrightarrow \mu. \tag{2.3.4}$$

$$\lim_n \int f d\mu_n = \int f d\mu, \quad \forall f \in C_b. \tag{2.3.5}$$

$$\liminf_n \int f d\mu_n \geqslant \int f d\mu, \quad \forall \textit{bounded lower semicontinuous } f. \tag{2.3.6}$$

$$\limsup_n \int f d\mu_n \leqslant \int f d\mu, \quad \forall \textit{bounded upper semicontinuous } f. \tag{2.3.7}$$

Proof. We already have (2.3.4) and (2.3.5) are equivalent. It is clear that (2.3.6) and (2.3.7) are equivalent. It remains to show (2.3.5) and (2.3.6) are equivalent.

From (2.3.5) to (2.3.6). Since f is lower semicontinuous, there exists a sequence $f_k \in C_b$ such that $f_k \uparrow f$. We have, for each k,

$$\liminf_n \int f d\mu_n \geqslant \liminf_n \int f_k d\mu_n = \int f_k d\mu.$$

By monotone convergence theorem, we have

$$\liminf_n \int f d\mu_n \geqslant \lim_k \int f_k d\mu = \int f d\mu.$$

From (2.3.6) to (2.3.5). Since (2.3.6) holds, we have (2.3.7) holds. For any $f \in C_b$, it is both lower semicontinuous and upper semicontinuous, thus (2.3.6) and (2.3.7) imply (2.3.5). □

Corollary 2.3.9. *Suppose $\{\mu_n\}$ and μ are probability measures. Then the following statements are equivalent.*

$$\mu_n \Longrightarrow \mu. \tag{2.3.8}$$

$$\liminf_n \mu_n(O) \geqslant \mu(O), \quad \forall \textit{open set } O. \tag{2.3.9}$$

$$\limsup_n \mu_n(K) \leqslant \mu(K), \quad \forall \textit{closed set } K. \tag{2.3.10}$$

Proof. From (2.3.8) to (2.3.9): The indicator function on open set $\mathbb{1}_O$ is bounded lower semicontinuous. Since (2.3.8) implies (2.3.6), we see (2.3.9) holds.

It is clear that (2.3.9) and (2.3.10) are equivalent. It remains to show (2.3.9) implies (2.3.8): for any μ's continuity points a and b, we have

$$\liminf_n \mu_n(a,b] \geqslant \liminf_n \mu_n(a,b) \geqslant \mu(a,b), \quad \text{(by (2.3.9))}$$

$$\limsup_n \mu_n(a,b] \leqslant \limsup_n \mu_n[a,b] \leqslant \mu[a,b]. \quad \text{(by (2.3.10))}$$

Since a, b are μ's continuity points, we have $\mu(a,b) = \mu[a,b]$. Thus $\lim_n \mu_n(a,b] = \mu(a,b]$ as desired. □

2.4 Convergence in distribution

Definition 2.4.1. *A sequence of random variables $\{X_n\}$ converges in distribution to a random variable X if $\mathcal{L}(X_n) \Longrightarrow \mathcal{L}(X)$. We denote by $X_n \overset{d}{\Longrightarrow} X$, or $X_n \to X$ in distribution.*

We first discuss the relation between convergence in distribution and convergence in probability.

Lemma 2.4.2. *Convergence in probability implies convergence in distribution.*

Proof. Suppose $X_n \to X$ in probability. It suffices to show that $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for any $f \in C_c$.

Since $f \in C_c$, we know that it is bounded. Suppose $|f| \leqslant M$. Furthermore, it is uniformly continuous, i.e. for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| \leqslant \epsilon$ as long as $|x - y| \leqslant \delta$. Thus

$$\mathbb{E}[|f(X_n) - f(X)|] \leqslant \epsilon \mathbb{P}[|X_n - X| \leqslant \delta] + 2M\mathbb{P}[|X_n - X| > \delta].$$

Let $n \to \infty$, we have

$$\limsup_n \mathbb{E}[|f(X_n) - f(X)|] \leqslant \epsilon.$$

Let $\epsilon \to 0$, we have $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as desired. □

Lemma 2.4.3. *Suppose X_n converges to a constant c in distribution. Then $X_n \to c$ in probability.*

Proof. The limiting distribution μ is dirac, and its continuity points is $\mathbb{R}\backslash\{c\}$. In particular, $c - \epsilon$ and $c + \epsilon$ are continuity points. Thus

$$\mathbb{P}[|X_n - c| > \epsilon] = \mathbb{P}[X_n \in (-\infty, c - \epsilon)] + \mathbb{P}[X_n \in (c + \epsilon, \infty)] \to 0, \quad n \to \infty.$$

This gives the convergence in probability. □

Convergence of random variables in distribution is merely a convenience of speech, it does not have the usual properties associated with convergence. Suppose $X_n \stackrel{d}{\Longrightarrow} X$, the random variables $\{X_n\}$ and X may be not in the same probability space! Suppose $X_n \stackrel{d}{\Longrightarrow} X$ and $Y_n \stackrel{d}{\Longrightarrow} Y$ and suppose they are in the same probability space, it does not follow by any means that $X_n + Y_n$ will converge in distribution to $X + Y$. Nevertheless, the following simple situation still holds.

Lemma 2.4.4. *Suppose $X_n \to X$ in distribution, and $Y_n \to 0$ in distribution, then*

(1) $X_n + Y_n \to X$ in distribution.

(2) $X_n Y_n \to 0$ in distribution.

Proof. By Lemma 2.4.3, we know that $Y_n \to 0$ in probability.

We first show (1). Suppose $f \in C_c$. Then f is uniformly continuous and it is bounded: $|f| \leqslant M$. For any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| \leqslant \epsilon$ as long as $|x - y| \leqslant \delta$. Thus

$$\begin{aligned}
&\mathbb{E}[|f(X_n + Y_n) - f(X_n)|] \\
&\leqslant \mathbb{E}[|f(X_n + Y_n) - f(X_n)| \mathbb{1}\{|Y_n| \leqslant \delta\}] + \mathbb{E}[|f(X_n + Y_n) - f(X_n)| \mathbb{1}\{|Y_n| > \delta\}] \\
&\leqslant \epsilon + 2M\mathbb{P}[|Y_n| > \delta].
\end{aligned}$$

Let $n \to \infty$ and then $\epsilon \to 0$, we obtain $\mathbb{E}[|f(X_n + Y_n) - f(X_n)|] \to 0$, and hence $\mathbb{E}[f(X_n + Y_n)] \to \mathbb{E}[f(X)]$ as desired.

Next, we show (2). We choose M large such that $\pm M$ are both continuity points of X. We have

$$\mathbb{P}[|X_n Y_n| > \epsilon] \leqslant \mathbb{P}[|X_n| > M] + \mathbb{P}[|Y_n| > \epsilon/M].$$

Since $X_n \to X$ in distribution, we have $\lim_n \mathbb{P}[|X_n| > M] = \mathbb{P}[|X| > M]$. Since $Y_n \to 0$ in probability, we have $\lim_n \mathbb{P}[|Y_n| > \epsilon/M] = 0$. Thus

$$\limsup_n \mathbb{P}[|X_n Y_n| > \epsilon] \leqslant \mathbb{P}[|X| > M].$$

Let $M \to \infty$ in the way that $\pm M$ are both continuity points of X, we have $\lim_n \mathbb{P}[|X_n Y_n| > \epsilon] = 0$ as desired. □

As a consequence of Lemmas 2.4.3 and 2.4.4, we have the following.

Proposition 2.4.5. *If $X_n \to X$, $\alpha_n \to a$, $\beta_n \to b$ in distribution where a, b are constants, then $\alpha_n X_n + \beta_n \to aX + b$ in distribution.*

Proof. We refer to the two conclusions in Lemma 2.4.4 as (1) and (2) respectively. Combining $X_n \to X$ and $\alpha_n - a \to 0$ in distribution and (2), we have $(\alpha_n - a)X_n \to 0$ in distribution. Combining with $aX_n \to aX$ in distribution and (1), we have $\alpha_n X_n \to aX$ in distribution. Combining with $\beta_n - b \to 0$ in distribution and (1), we have $\alpha_n X_n + \beta_n - b \to aX$ in distribution. This implies $\alpha_n X_n + \beta_n \to aX + b$ in distribution. Exercise: give a direct proof of this proposition. □

Next, let us discuss the relation between convergence in distribution and almost sure convergence. As almost sure convergence implies convergence in probability which implies convergence in distribution, the direction of implication is clear. In a sense, the following theorem gives the reverse direction.

Theorem 2.4.6. *Suppose $X_n \to X$ in distribution. Then there exists a probability space and random variables in the space $\{Y_n\}$ and Y such that*

$$Y_n \to Y \quad a.s., \quad \mathcal{L}(Y_n) = \mathcal{L}(X_n), \quad \mathcal{L}(Y) = \mathcal{L}(X).$$

Proof. Take the probability space $(\mathcal{U}, \mathcal{B}, \mathrm{Leb})$. Denote by F_n the distribution function of X_n. Denote by F the distribution function of X. Define

$$Y_n(x) := \sup\{y : F_n(y) < x\}; \quad Y(x) := \sup\{y : F(y) < x\}.$$

One can check that $\mathcal{L}(Y_n) = \mathcal{L}(X_n)$ and $\mathcal{L}(Y) = \mathcal{L}(X)$. To this end, we only need to show

$$\{x : Y(x) \leqslant z\} = \{x : x \leqslant F(z)\}. \tag{2.4.1}$$

We have the following observations:

- If $x \leqslant F(z)$, then $Y(x) \leqslant z$.
- If $x > F(z)$, since F is right-continuous, there exists $\epsilon > 0$ such that $F(z+\epsilon) < x$, thus $Y(x) \geqslant z+\epsilon$. Therefore, $x > F(z)$ implies $Y(x) > z$.

Combining these two observations, we obtain (2.4.1).

Denote by $\mathcal{C}_F$ the set of continuity points of F. Define $a_x = \sup\{y : F(y) < x\}$ and $b_x = \inf\{y : F(y) > x\}$, and set $\mathcal{U}_0 = \{x : (a_x, b_x) = \varnothing\}$. Since $\{(a_x, b_x) : x \in \mathcal{U} \backslash \mathcal{U}_0\}$ are disjoint open intervals, the set $\mathcal{U} \backslash \mathcal{U}_0$ is at most countable. We will show that $Y_n \to Y$ on $\mathcal{U}_0$.

We first show that $\liminf_n Y_n(x) \geqslant Y(x)$. For $y < Y(x)$ and $y \in \mathcal{C}_F$, we have $F(y) < x$. Since $y \in \mathcal{C}_F$ and $F_n \Longrightarrow F$, we have $F_n(y) < x$ for n large enough, thus $y \leqslant Y_n(x)$ for n large enough. Hence $\liminf_n Y_n(x) \geqslant y$ for any $y < Y(x)$ and $y \in \mathcal{C}_F$. This implies that $\liminf_n Y_n(x) \geqslant Y(x)$.

Next we show that $\limsup_n Y_n(x) \leqslant Y(x)$ for $x \in \mathcal{U}_0$. For $y > Y(x)$ and $y \in \mathcal{C}_F$, we have $F(y) \geqslant x$. Since $x \in \mathcal{U}_0$, we have $F(y) > x$. Since $y \in \mathcal{C}_F$ and $F_n \Longrightarrow F$, we have $F_n(y) > x$ for n large enough, thus $Y_n(x) \leqslant y$ for n large enough. Hence $\limsup_n Y_n(x) \leqslant y$ for any $y > Y(x)$ and $y \in \mathcal{C}_F$. This implies that $\limsup_n Y_n(x) \leqslant Y(x)$. □

2.5 Uniform integrability

In this section, we discuss the relation between convergence in distribution and the convergence in L^1. As convergence in L^1 implies the convergence in probability which implies convergence in distribution, the direction of implication is clear. In reality, we are usually interested in the following question: we have $X_n \to X$ in distribution, and we desire $\mathbb{E}[X_n] \to \mathbb{E}[X]$. This is false in general, but under the condition of "uniform integrability", it is true.

Definition 2.5.1. *A collection $\{X_i, i \in I\}$ of random variables is Uniform Integrable (UI) if*

$$\sup_i \mathbb{E}\left[|X_i| \mathbb{1}_{\{|X_i| \geqslant \alpha]\}}\right] \to 0, \qquad \text{as } \alpha \to \infty.$$

Suppose $X \in L^1$, since

$$\mathbb{E}\left[|X| \mathbb{1}_{\{|X| \geqslant \alpha\}}\right] \to 0, \quad \alpha \to \infty,$$

we know that the family containing only X is UI. From the definition, it is clear that the union of two UI families is still UI. Thus we have the following conclusion.

Lemma 2.5.2. *If the family contains finitely many random variables in L^1, then it is UI.*

Lemma 2.5.3. *(1) A UI family is bounded in L^1.*

(2) If a family of random variables is bounded in L^p for some $p > 1$, then it is UI.

Proof. The first item is clear. We only need to show the second. Suppose $\{X_i, i \in I\}$ is bounded in L^p with $p > 1$, i.e. $\sup_i \mathbb{E}[|X_i|^p] \leqslant C < \infty$. We have

$$\mathbb{E}[|X_i|^p] \geqslant \mathbb{E}\left[|X_i|^p \mathbb{1}_{\{|X_i| > \alpha\}}\right] \geqslant \alpha^{p-1} \mathbb{E}\left[|X_i| \mathbb{1}_{\{|X_i| > \alpha\}}\right].$$

Thus

$$\sup_i \mathbb{E}[|X_i| \mathbb{1}_{\{|X_i| \geqslant \alpha\}}] \leqslant \frac{C}{\alpha^{p-1}} \to 0, \quad \alpha \to \infty.$$

□

Proposition 2.5.4. *Suppose that $X_n, X \in L^1$ and $X_n \to X$ a.s. Then*

$$X_n \to X \text{ in } L^1 \quad \text{if and only if} \quad \{X_n, n \geqslant 1\} \text{ is UI.}$$

Proof of $\Rightarrow$. We will show that, for any $\epsilon > 0$, there exists α such that $\mathbb{E}[|X_n| \mathbb{1}_{\{|X_n| > \alpha\}}] \leqslant 2\epsilon$.

Since $X_n \to X$ in L^1, there exists N such that $\mathbb{E}[|X_n - X|] \leqslant \epsilon$ for all $n \geqslant N$. Moreover, there exists M such that $\mathbb{E}[|X_n|] \leqslant M$. For $\beta > 0$, we have that

$$\mathbb{E}[|X_n| \mathbb{1}_{\{|X_n| > \beta\}}] \leqslant \mathbb{E}[|X_n - X|] + \mathbb{E}[|X| \mathbb{1}_{\{|X_n| > \beta\}}] \leqslant \epsilon + \mathbb{E}[|X| \mathbb{1}_{\{|X_n| > \beta\}}].$$

We have the following claim: Suppose $Z \in L^1$, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\mathbb{E}[|Z| \mathbb{1}_A] \leqslant \epsilon, \quad \text{as long as } \mathbb{P}[A] \leqslant \delta. \tag{2.5.1}$$

Assume (2.5.1) holds, we choose $\beta = M/\delta$, then $\mathbb{P}[|X_n| > \beta] \leqslant \mathbb{E}[|X_n|]/\beta \leqslant \delta$ and therefore,

$$\mathbb{E}[|X| \mathbb{1}_{\{|X_n| > \beta\}}] \leqslant \epsilon.$$

This implies that, for all $n \geqslant N$,

$$\mathbb{E}[|X_n| \mathbb{1}_{\{|X_n| > \beta\}}] \leqslant 2\epsilon.$$

Note that the finite family $\{X_1, ..., X_N\}$ is UI, and we can find β' such that, for all $1 \leqslant n \leqslant N$,

$$\mathbb{E}[|X_n| \mathbb{1}_{\{|X_n| > \beta'\}}] \leqslant \epsilon.$$

Set $\alpha = \beta \vee \beta'$, and it is the desired quantity.

It remains to show (2.5.1). Since $Z \in L^1$, there exists C such that $\mathbb{E}[|Z| \mathbb{1}_{\{|Z| \geqslant C\}}] \leqslant \epsilon/2$. Then we have

$$\mathbb{E}\left[|Z| \mathbb{1}_A\right] \leqslant \mathbb{E}\left[|Z| \mathbb{1}_{\{|Z| \geqslant C\}}\right] + C\mathbb{P}[A] \leqslant \epsilon/2 + C\delta.$$

We can choose $\delta = \epsilon/(2C)$. □

Proof of $\Leftarrow$. It suffices to show that, for any $\epsilon > 0$, there exists N such that $\mathbb{E}[|X_n - X|] \leqslant \epsilon$ for $n \geqslant N$.

Since $\{X_n, n \geqslant 1\}$ is UI and $X \in L^1$, there exists M such that

$$\mathbb{E}\left[|X_n| \mathbb{1}_{\{|X_n|>M\}}\right] \leqslant \epsilon/4, \quad \mathbb{E}\left[|X| \mathbb{1}_{\{|X|>M\}}\right] \leqslant \epsilon/4.$$

Define the cutoff function:

$$\varphi_M(x) = \begin{cases} M, & x \geqslant M; \\ x, & |x| \leqslant M; \\ -M, & x \leqslant -M. \end{cases}$$

Note that $\varphi_M(X_n) \to \varphi_M(X)$ almost surely, hence by dominated convergence theorem, we have $\varphi_M(X_n) \to \varphi_M(X)$ in L^1. Thus, there exists N such that, for all $n \geqslant N$,

$$\mathbb{E}[|\varphi_M(X_n) - \varphi_M(X)|] \leqslant \epsilon/4.$$

Therefore, for all $n \geqslant N$, we have

$$\mathbb{E}[|X_n - X|] \leqslant \mathbb{E}[|\varphi_M(X_n) - \varphi_M(X)|] + \mathbb{E}\left[|X_n| \mathbb{1}_{\{|X_n|>M\}}\right] + \mathbb{E}\left[|X| \mathbb{1}_{\{|X|>M\}}\right] \leqslant \epsilon,$$

as desired. $\square$

From the proof of Proposition 2.5.4, we have several consequences.

Corollary 2.5.5. *(1) Suppose $X_n \to X$ in L^1, then $\{X_n\}$ is UI.*

(2) Suppose $\{X_n, n \geqslant 1\}$ is UI, and that $X_n \to X$ in distribution. Then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

To end this chapter, let us summarize the relation between different notions of convergence.

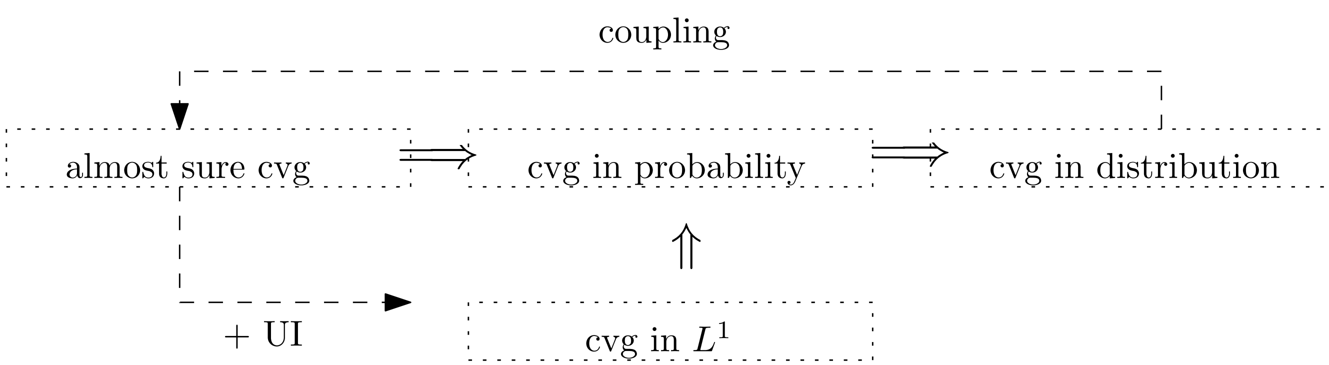


Figure 2.1

2.6 Exercises

Exercise 2.6.1. *Let $\{X_n\}$ be independent. Show that $\sup_n X_n < \infty$ almost surely if and only if*

$$\sum_n \mathbb{P}[X_n > A] < \infty, \quad \textit{for some } A.$$

Exercise 2.6.2. *Suppose $\{A_n\}$ is a sequence of events. Show that*

$$\mathbb{P}[A_n \ i.o.] \geqslant \limsup_n \mathbb{P}[A_n].$$

Exercise 2.6.3. *Let $\{X_n\}$ be i.i.d. Poisson random variables with $\mathbb{E}[X_n] = 1$. Define $S_n = \sum_{j=1}^n X_j$.*

(1) *Show that $S_n/n \to 1$ in L^2.*

(2) *Show that $S_n/n \to 1$ almost surely.*

Exercise 2.6.4. *Suppose f is continuous. If $X_n \to X$ in probability, then $f(X_n) \to f(X)$ in probability.*

Exercise 2.6.5. *Suppose $X_n \downarrow X$ a.s., each X_n is integrable and $\inf_n \mathbb{E}[X_n] > -\infty$, then $X_n \to X$ in L^1.*

Exercise 2.6.6. *Let $\{X_n\}$ be a sequence of independent random variables with mean 0 and variance 1. Show that for any bounded random variable Y we have $\lim_{n\to\infty} \mathbb{E}[X_n Y] = 0$.*

Exercise 2.6.7 (YCMC2012)**.** *Let $\{X_n\}$ be a sequence of random variables satisfying*

$$\lim_{a\to\infty} \sup_{n\geqslant 1} \mathbb{P}[|X_n| > a] = 0.$$

Assume that $Y_n \to 0$ in probability. Show that $X_n Y_n \to 0$ in probability.

Exercise 2.6.8 (YCMC2014)**.** *Let $\{X_n\}$ be a sequence of uncorrelated random variables of mean zero such that*

$$\sum_{n=1}^{\infty} n\mathbb{E}\left[X_n^2\right] < \infty.$$

Show that $S_n = \sum_{i=1}^n X_i$ converges almost surely.

Exercise 2.6.9 (YCMC2015)**.** *Suppose $\{\xi_n\}$ are independent with Bernoulli distribution $\mathbb{P}[\xi_n = 1] = p_n$ and $\mathbb{P}[\xi_n = 0] = 1 - p_n$. Assume $\sum_{n=1}^{\infty} p_n p_{n+1} < +\infty$, show that $\sum_{n=1}^{\infty} \xi_n \xi_{n+1}$ converges almost surely.*

Exercise 2.6.10 (YCMC2015)**.** *Suppose X_n converges in distribution to X. Let $\{N_t, t \geqslant 0\}$ be a set of positive-integer-valued random variables, which is independent of $\{X_n\}$ and converges in probability to ∞ as $t \to \infty$. Show that $\{X_{N_t}\}$ converges in distribution to X as $t \to \infty$.*

Exercise 2.6.11 (YCMC2015)**.** *Let $\{X_n\}$ be independent and $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$.*

(1) *If $\sum X_n^2$ converges in L^1, then $\sum X_n^2$ converges in L^p, for every $p \in [1, \infty)$.*

(2) *Assume that $\mu_n = 0$ for every n. If $\sum \sigma_n^2 = \infty$, then $\mathbb{P}\left[\sum X_n^2 = \infty\right] = 1$.*

Exercise 2.6.12 (YCMC2016)**.** *For each n, let X_n be an exponential random variable with parameter q_n. Suppose that $\{X_n\}$ are independent.*

(1) *What is $\mathbb{E}\left[e^{-X_n}\right]$?*

(2) *Suppose $\sum \frac{1}{q_n} < \infty$, show that $\sum X_n < \infty$ almost surely.*

(3) *Suppose $\sum \frac{1}{q_n} = \infty$, show that $\sum X_n = \infty$ almost surely.*

Exercise 2.6.13 (YCMC2016)**.** *Let $\{X_n\}$ be i.i.d. Prove or disprove: If $\limsup_{n\to\infty} |X_n|/n \leqslant 1$ almost surely, then $\sum_{n=1}^{\infty} \mathbb{P}\left[|X_n| \geqslant n\right] < \infty$.*

Exercise 2.6.14 (YCMC2017)**.** *Let $\{X_n\}$ be positive random variables. Assume that $X_n \to 0$ in probability, and that $\lim_{n\to\infty} \mathbb{E}[X_n] = 2$. Show that $\lim_{n\to\infty} \mathbb{E}[|X_n - 1|]$ exists and compute its value.*

3 Law of Large Numbers

3.1 Weak law of large numbers

The goal of this section is the following "weak law of large numbers".

Theorem 3.1.1. *Let $\{X_n\}$ be i.i.d. with finite mean m. Define $S_n = \sum_{j=1}^n X_j$, then we have*

$$\frac{S_n}{n} \to m, \quad \textit{in probability.}$$

To prove this theorem, we need to truncate X_n's and then show the conclusion for the truncated random variables. For the truncation to work, we need to first understand what are the good truncations.

Definition 3.1.2. *Two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ are equivalent if*

$$\sum_n \mathbb{P}[X_n \neq Y_n] < \infty.$$

Suppose $\{X_n\}$ and $\{Y_n\}$ are equivalent, by Borel-Cantelli lemma, there exists Ω_0 with $\mathbb{P}[\Omega_0] = 1$ such that, for any $\omega \in \Omega_0$, we have $X_n(\omega) = Y_n(\omega)$ for all but finitely many n. Thus it is clear that

- $\sum_n (X_n - Y_n)$ converges almost surely.
- $\frac{1}{n}\sum_{j=1}^n (X_j - Y_j) \to 0$ almost surely.
- $\frac{1}{n}\sum_{j=1}^n X_j \xrightarrow{\text{proba.}} X$ implies $\frac{1}{n}\sum_{j=1}^n Y_j \xrightarrow{\text{proba.}} X$.

Proof of Theorem 3.1.1. Denote by μ the common law of X_n's, and suppose $Z \sim \mu$. Since $Z \in L^1$, we have

$$\sum_n \mathbb{P}[|Z| \geqslant n] < \infty.$$

We introduce random variables Y_n's by truncating X_n's:

$$Y_n = X_n \mathbb{1}_{\{|X_n| \leqslant n\}}.$$

Then

$$\sum_n \mathbb{P}[X_n \neq Y_n] = \sum_n \mathbb{P}[|X_n| > n] = \sum_n \mathbb{P}[|Z| \geqslant n] < \infty.$$

Hence $\{Y_n\}$ and $\{X_n\}$ are equivalent. Define $T_n = \sum_{j=1}^n Y_j$. If we prove

$$\frac{T_n - \mathbb{E}[T_n]}{n} \to 0 \quad \text{in probability,} \tag{3.1.1}$$

then the conclusion follows, because $\mathbb{E}[T_n]/n \to m$ as $n \to \infty$. It remains to show (3.1.1). We see, for any $\epsilon > 0$,

$$\mathbb{P}\left[|T_n - \mathbb{E}[T_n]| \geqslant n\epsilon\right] \leqslant \frac{\operatorname{var}(T_n)}{n^2\epsilon^2}.$$

It suffices to show $\operatorname{var}(T_n) = o(n^2)$. Let us calculate $\operatorname{var}(T_n)$.

$$\operatorname{var}(T_n) = \sum_{j=1}^n \operatorname{var}(Y_j) \leqslant \sum_{j=1}^n \mathbb{E}[Y_j^2] = \sum_{j=1}^n \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z| \leqslant j\}}\right].$$

The most naive estimate is the following:

$$\operatorname{var}(T_n) \leqslant \sum_{j=1}^{n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant j\}}\right] \leqslant \sum_{j=1}^{n} j^2,$$

which is $O(n^3)$. The less naive estimate is the following:

$$\operatorname{var}(T_n) \leqslant \sum_{j=1}^{n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant j\}}\right] \leqslant \sum_{j=1}^{n} j\mathbb{E}\left[|Z| \mathbb{1}_{\{|Z|\leqslant j\}}\right] \leqslant \sum_{j=1}^{n} j\mathbb{E}[|Z|],$$

which is $O(n^2)$. But we desire a control of $o(n^2)$. To improve it, let $\{a_n\}$ be a sequence of integers such that $1 \leqslant a_n \leqslant n$, $a_n \to \infty$, but $a_n = o(n)$. Then we have

$$\begin{aligned}
\operatorname{var}(T_n) &\leqslant \sum_{j=1}^{n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant j\}}\right] = \sum_{j\leqslant a_n} + \sum_{a_n<j\leqslant n} \\
&= \sum_{j\leqslant a_n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant j\}}\right] + \sum_{a_n<j\leqslant n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant a_n\}}\right] + \sum_{a_n<j\leqslant n} \mathbb{E}\left[Z^2 \mathbb{1}_{\{a_n<|Z|\leqslant j\}}\right] \\
&\leqslant \sum_{j\leqslant a_n} a_n\mathbb{E}\left[|Z|\right] + \sum_{a_n<j\leqslant n} a_n\mathbb{E}\left[|Z|\right] + \sum_{a_n<j\leqslant n} n\mathbb{E}\left[|Z| \mathbb{1}_{\{a_n<|Z|\leqslant j\}}\right] \\
&\leqslant O(1)na_n + O(1)n^2\mathbb{E}\left[|Z| \mathbb{1}_{\{|Z|>a_n\}}\right].
\end{aligned}$$

The first term is $na_n = o(n^2)$ because $a_n = o(n)$; the second term is also $o(n^2)$ because $\mathbb{E}\left[|Z| \mathbb{1}_{\{|Z|>a_n\}}\right] \to 0$ since $a_n \to \infty$. Therefore, we have $\operatorname{var}(T_n) = o(n^2)$ as desired. $\square$

The following example explains that the finite expectation in the hypothesis of Theorem 3.1.1 is not a necessary condition for the convergence in probability.

Example 3.1.3. *Let $\{X_n\}$ be i.i.d. with the common law given by*

$$\mathbb{P}[Z = n] = \mathbb{P}[Z = -n] = \frac{c}{n^2 \log n}, \quad n = 2, 3, \ldots,$$

where c is a normalizing constant. Define $S_n = \sum_{j=1}^{n} X_j$. It is clear that $\mathbb{E}[|Z|] = \infty$, but we have

$$\frac{S_n}{n} \to 0, \quad \textit{in probability.}$$

Proof. Define

$$T_n = \sum_{j=1}^{n} X_j \mathbb{1}_{\{|X_j|\leqslant n\}}.$$

(Note that, this is different from the definition in the proof of Theorem 3.1.1, as all X_j in the summation are truncated at the same constant n.) Let us calculate $\mathbb{P}[T_n \neq S_n]$ and $\operatorname{var}(T_n)$.

$$\mathbb{P}[T_n \neq S_n] \leqslant \sum_{j=1}^{n} \mathbb{P}[|X_j| > n] = n\mathbb{P}[|Z| > n] \sim \frac{c}{\log n}.$$

$$\operatorname{var}(T_n) = n\mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant n\}}\right] = n\sum_{j=2}^{n} \frac{c}{\log j} \sim \frac{cn^2}{\log n}.$$

Thus

$$\mathbb{P}\left[\left|\frac{S_n}{n}\right| > \epsilon\right] \leqslant \mathbb{P}\left[\left|\frac{T_n}{n}\right| > \epsilon\right] + \mathbb{P}[T_n \neq S_n] \leqslant \frac{\operatorname{var}(T_n)}{n^2\epsilon^2} + \mathbb{P}[T_n \neq S_n] \to 0.$$

$\square$

3.2 Three series theorem

The goal of this section is Kolmogorov's three series theorem.

Theorem 3.2.1. *Let $\{X_n\}$ be independent random variables and define the truncation for a fixed constant $A > 0$:*

$$Y_n = X_n \mathbb{1}_{\{|X_n| \leqslant A\}}.$$

Then the series $\sum_n X_n$ converges almost surely if and only if the following three series all converge:

$$\sum_n \mathbb{P}[|X_n| > A], \quad \sum_n \mathbb{E}[Y_n], \quad \sum_n \mathrm{var}(Y_n).$$

To show the direction of $\Longleftarrow$, we need the following lemma.

Lemma 3.2.2. *Let $\{X_n\}$ be independent random variables such that $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^2] < \infty$. Define $S_n = \sum_{j=1}^n X_j$. Then we have*

$$\mathbb{P}\left[\max_{1 \leqslant j \leqslant n} |S_j| \geqslant \epsilon\right] \leqslant \frac{\mathbb{E}[S_n^2]}{\epsilon^2}.$$

Proof. Fix $\epsilon > 0$ and define

$$\Lambda = \left\{\max_{1 \leqslant j \leqslant n} |S_j| \geqslant \epsilon\right\}.$$

Define $T = \min\{j : |S_j| \geqslant \epsilon\}$ to be the first time that $|S_j|$ exceeds ϵ, and define $\Lambda_k = \{T = k\}$:

$$\Lambda_k = \left\{\max_{1 \leqslant j \leqslant k-1} |S_j| < \epsilon, |S_k| \geqslant \epsilon\right\}.$$

Note that Λ_k's are disjoint and $\Lambda = \sqcup_{k=1}^n \Lambda_k$. We have

$$\begin{aligned}
\mathbb{E}\left[S_n^2 \mathbb{1}_\Lambda\right] &= \sum_{k=1}^n \mathbb{E}\left[S_n^2 \mathbb{1}_{\Lambda_k}\right] \\
&= \sum_{k=1}^n \mathbb{E}\left[S_k^2 \mathbb{1}_{\Lambda_k} + 2S_k(S_n - S_k) \mathbb{1}_{\Lambda_k} + (S_n - S_k)^2 \mathbb{1}_{\Lambda_k}\right].
\end{aligned}$$

Note that $S_k \mathbb{1}_{\Lambda_k}$ and $S_n - S_k$ are independent, thus

$$\mathbb{E}\left[S_k(S_n - S_k) \mathbb{1}_{\Lambda_k}\right] = \mathbb{E}\left[S_k \mathbb{1}_{\Lambda_k}\right] \mathbb{E}\left[(S_n - S_k)\right] = 0.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[S_n^2 \mathbb{1}_\Lambda] &= \sum_{k=1}^n \mathbb{E}\left[S_k^2 \mathbb{1}_{\Lambda_k} + (S_n - S_k)^2 \mathbb{1}_{\Lambda_k}\right] \\
&\geqslant \sum_{k=1}^n \mathbb{E}\left[S_k^2 \mathbb{1}_{\Lambda_k}\right] \geqslant \sum_{k=1}^n \epsilon^2 \mathbb{P}[\Lambda_k] = \epsilon^2 \mathbb{P}[\Lambda].
\end{aligned}$$

Thus we have $\mathbb{P}[\Lambda] \leqslant \mathbb{E}[S_n^2]/\epsilon^2$, as desired. ☐

Proof of Theorem 3.2.1$\Longleftarrow$. Suppose the three series all converge. Since the first series converges, we have $\sum_n \mathbb{P}[X_n \neq Y_n] < \infty$. Thus $\{X_n\}$ and $\{Y_n\}$ are equivalent, it suffices to show that $\sum_n Y_n$ converges almost surely. Since the second series converges, it suffices to show that $\sum_n (Y_n - \mathbb{E}[Y_n])$ converges almost surely. Let us consider the tail of this series

$$T(n, m) := \sum_{j=n}^m (Y_j - \mathbb{E}[Y_j]).$$

We need to show that, almost surely, the oscillation

$$W_n := \max_{\ell \geqslant k \geqslant n} |T(k,\ell)|$$

is small when n is large.

Fix $\epsilon > 0$, by Lemma 3.2.2, we have

$$\mathbb{P}\left[\max_{n \leqslant j \leqslant m} |T(n,j)| \geqslant \epsilon/2\right] \leqslant 4\epsilon^{-2} \sum_{j=n}^{m} \operatorname{var}(Y_j).$$

Let $m \to \infty$, we have

$$\mathbb{P}\left[\max_{j \geqslant n} |T(n,j)| \geqslant \epsilon/2\right] \leqslant 4\epsilon^{-2} \sum_{j \geqslant n} \operatorname{var}(Y_j).$$

For $\ell \geqslant k \geqslant n$, we have

$$T(k,\ell) = T(n,\ell) - T(n,k).$$

Thus

$$\begin{aligned}
\mathbb{P}[W_n \geqslant \epsilon] &= \mathbb{P}\left[\max_{\ell \geqslant k \geqslant n} |T(k,\ell)| \geqslant \epsilon\right] \\
&\leqslant \mathbb{P}\left[\max_{j \geqslant n} |T(n,j)| \geqslant \epsilon/2\right] \leqslant 4\epsilon^{-2} \sum_{j \geqslant n} \operatorname{var}(Y_j).
\end{aligned}$$

Since the third series converges, we have

$$\lim_n \mathbb{P}[W_n \geqslant \epsilon] = 0.$$

Since the sequence of events $\{W_n \geqslant \epsilon\}$ is decreasing in n, we have

$$\mathbb{P}\left[\lim_n W_n \geqslant \epsilon\right] = 0.$$

Let $\epsilon \to 0$, we have

$$\mathbb{P}\left[\lim_n W_n = 0\right] = 1.$$

This implies the almost sure convergence. □

To show the direction of $\Longrightarrow$, we need the following lemma.

Lemma 3.2.3. *Let $\{X_n\}$ be independent random variables which are bounded: there exists a constant A such that $|X_n| \leqslant A$ almost surely for all n. Define $S_n = \sum_{j=1}^n X_j$. Then we have*

$$\mathbb{P}\left[\max_{1 \leqslant j \leqslant n} |S_j| \leqslant B\right] \leqslant \frac{(2B+A)^2}{\operatorname{var}(S_n)}.$$

Proof. Define $T = \min\{j : |S_j| > B\}$ to be the first time that $|S_j|$ exceeds B. Then we have

$$\{T > k\} = \left\{\max_{1 \leqslant j \leqslant k} |S_j| \leqslant B\right\}, \quad \{T = k\} = \left\{\max_{1 \leqslant j \leqslant k-1} |S_j| \leqslant B, |S_k| > B\right\}.$$

We need to give a upper bound for $\mathbb{P}[T > n] \operatorname{var}(S_n)$. Let us consider the expectation and the variance of S_k on $\{T > k\}$:

$$a_k := \mathbb{E}\left[S_k \mathbb{1}_{\{T>k\}}\right] / \mathbb{P}[T > k], \quad \mathbb{E}\left[(S_k - a_k)^2 \mathbb{1}_{\{T>k\}}\right].$$

It is clear that $|a_k| \leqslant B$. We write

$$
\begin{aligned}
&\mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T>k+1\}}\right] \\
&\quad = \mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T>k\}}\right] - \mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T=k+1\}}\right].
\end{aligned}
$$

For the first term,

$$
\begin{aligned}
&\mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T>k\}}\right] \\
&\quad = \mathbb{E}\left[(S_k-a_k+X_{k+1}-a_{k+1}+a_k)^2\,\mathbb{1}_{\{T>k\}}\right] \\
&\quad = \mathbb{E}\left[(S_k-a_k)^2\,\mathbb{1}_{\{T>k\}}\right] + \mathbb{E}\left[(X_{k+1}-a_{k+1}+a_k)^2\,\mathbb{1}_{\{T>k\}}\right] && \text{(by indep.)} \\
&\quad = \mathbb{E}\left[(S_k-a_k)^2\,\mathbb{1}_{\{T>k\}}\right] + \mathbb{E}\left[(X_{k+1}-a_{k+1}+a_k)^2\right]\mathbb{P}[T>k] && \text{(by indep.)} \\
&\quad \geqslant \mathbb{E}\left[(S_k-a_k)^2\,\mathbb{1}_{\{T>k\}}\right] + \operatorname{var}(X_{k+1})\mathbb{P}[T>k]. && \text{(by Exercise 1.7.1)}
\end{aligned}
$$

For the second term,

$$
\mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T=k+1\}}\right] = \mathbb{E}\left[(S_k+X_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T=k+1\}}\right].
$$

Note that, $|S_k| \leqslant B$ on $\{T=k+1\}$, and $|X_{k+1}| \leqslant A$, and $|a_{k+1}| \leqslant B$. Thus, for the second term,

$$
\mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T=k+1\}}\right] \leqslant (2B+A)^2\mathbb{P}[T=k+1].
$$

Combining the two estimates, we have

$$
\begin{aligned}
&\mathbb{E}\left[(S_{k+1}-a_{k+1})^2\,\mathbb{1}_{\{T>k+1\}}\right] \\
&\quad \geqslant \mathbb{E}\left[(S_k-a_k)^2\,\mathbb{1}_{\{T>k\}}\right] + \operatorname{var}(X_{k+1})\mathbb{P}[T>k] - (2B+A)^2\mathbb{P}[T=k+1].
\end{aligned}
$$

Summing over k, we have

$$
\begin{aligned}
&\mathbb{E}\left[(S_n-a_n)^2\,\mathbb{1}_{\{T>n\}}\right] \\
&\quad \geqslant \mathbb{E}\left[(X_1-a_1)^2\,\mathbb{1}_{\{T>1\}}\right] + \sum_{k=1}^{n-1}\operatorname{var}(X_{k+1})\mathbb{P}[T>k] - (2B+A)^2\mathbb{P}[2\leqslant T\leqslant n] \\
&\quad \geqslant \mathbb{E}\left[(X_1-a_1)^2\,\mathbb{1}_{\{T>1\}}\right] + \left(\operatorname{var}(S_n)-\operatorname{var}(X_1)\right)\mathbb{P}[T>n] - (2B+A)^2\mathbb{P}[2\leqslant T\leqslant n]
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\operatorname{var}(S_n)\mathbb{P}[T>n] \\
&\quad \leqslant \mathbb{E}\left[(S_n-a_n)^2\,\mathbb{1}_{\{T>n\}}\right] + \operatorname{var}(X_1)\mathbb{P}[T>n] + (2B+A)^2\mathbb{P}[2\leqslant T\leqslant n].
\end{aligned}
$$

Note that

$$
\mathbb{E}\left[(S_n-a_n)^2\,\mathbb{1}_{\{T>n\}}\right] = \mathbb{E}[S_n^2\,\mathbb{1}_{\{T>n\}}] - a_n^2\mathbb{P}[T>n] \leqslant B^2\mathbb{P}[T>n].
$$

Thus

$$
\begin{aligned}
\operatorname{var}(S_n)\mathbb{P}[T>n] &\leqslant B^2\mathbb{P}[T>n] + A^2\mathbb{P}[T>n] + (2B+A)^2\mathbb{P}[2\leqslant T\leqslant n] \\
&\leqslant (2B+A)^2.
\end{aligned}
$$

This gives the conclusion. □

Proof of Theorem 3.2.1 $\Longrightarrow$. Suppose $\sum_n X_n$ converges almost surely, then we have

$$
\mathbb{P}[|X_n| > A \ i.o.] = 0.
$$

Then Borel Cantelli lemma guarantees the convergence of the first series. As a consequence, the sequences $\{Y_n\}$ and $\{X_n\}$ are equivalent, hence $\sum_n Y_n$ converges almost surely as well. By Lemma 3.2.3, we have

$$\mathbb{P}\left[\max_{n\leqslant k\leqslant m}|\sum_{j=n}^{k} Y_j|\leqslant 1\right]\leqslant\frac{(A+2)^2}{\sum_{j=n}^{m}\operatorname{var}(Y_j)}.$$

If the third series diverges, then the right hand-side will go to zero as $m\to\infty$. Hence the tail of $\sum_n Y_n$ almost surely would not be bounded by one, so the series could not converge. This confirms the convergence of the third series. By the proof of direction of $\Longleftarrow$, the convergence of the third series implies the convergence of $\sum_n(Y_n-\mathbb{E}[Y_n])$. Combining with the convergence of $\sum_n Y_n$, we have the convergence of the second series. $\square$

Example 3.2.4. *Suppose $\{X_n\}$ are i.i.d. Bernoulli random variables with parameter $1/2$: $\mathbb{P}[X_n=1]=\mathbb{P}[X_n=-1]=1/2$.*

- *Consider the series*
$$\sum_n\frac{X_n}{n}.$$
There is no absolute convergence, but there is almost sure convergence.

- *Consider the series*
$$\sum_n\frac{X_n}{\sqrt{n}}.$$
It diverges almost surely. Note that this is different from $\sum_n(-1)^n/\sqrt{n}$ which does converge.

Proof. By three series theorem, we see that the series $\sum_n X_n/n$ converges almost surely: take $Y_n=X_n/n$ and $A=1$, we have $\mathbb{P}[|X_n/n|>1]=0$, $\mathbb{E}[Y_n]=0$, and $\operatorname{var}(Y_n^2)=1/n^2$.

Next, we consider $\sum_n X_n/\sqrt{n}$: take $Y_n=X_n/\sqrt{n}$ and $A=1$, we have $\mathbb{P}[|X_n/\sqrt{n}|>1]=0$, $\mathbb{E}[Y_n]=0$, and $\operatorname{var}(Y_n)=1/n$. By three series law, we see that

$$\mathbb{P}\left[\sum_n\frac{X_n}{\sqrt{n}}\text{ converges}\right]<1.$$

By Kolmogorov's 0-1 law in Example 1.5.4, we have

$$\mathbb{P}\left[\sum_n\frac{X_n}{\sqrt{n}}\text{ converges}\right]=0.$$

$\square$

3.3 Strong law of large numbers

The goal of this section is the following "strong law of large numbers".

Theorem 3.3.1. *Let $\{X_n\}$ be i.i.d. Define $S_n=\sum_{j=1}^n X_j$. Then we have*

$$\mathbb{E}[|X_1|]<\infty\Longrightarrow\frac{S_n}{n}\to\mathbb{E}[X_1],\ \textit{almost surely;}\tag{3.3.1}$$

$$\mathbb{E}[|X_1|]=\infty\Longrightarrow\limsup_n\frac{|S_n|}{n}=\infty,\ \textit{almost surely.}\tag{3.3.2}$$

Lemma 3.3.2 (Kronecker's lemma). *Let $\{x_n\}$ be a sequence of real numbers, $\{a_n\}$ be a sequence of numbers such that $0 < a_n \uparrow \infty$. Then*

$$\sum_n \frac{x_n}{a_n} \text{ converges } \implies \frac{1}{a_n}\sum_{j=1}^n x_j \to 0.$$

Proof. Define $b_n = \sum_{j=1}^n x_j/a_j$ and set $a_0 = b_0 = 0$. Then we have

$$\frac{1}{a_n}\sum_{j=1}^n x_j = \frac{1}{a_n}\sum_{j=1}^n a_j(b_j - b_{j-1}) = b_n - \frac{1}{a_n}\sum_{j=0}^{n-1}(a_{j+1} - a_j)b_j.$$

Note that $a_{j+1} - a_j \geqslant 0$ and

$$\frac{1}{a_n}\sum_{j=0}^{n-1}(a_{j+1} - a_j) = 1.$$

Combining with the fact that $b_j \to b_\infty$, we have

$$\frac{1}{a_n}\sum_{j=0}^{n-1}(a_{j+1} - a_j)b_j \to b_\infty.$$

Thus

$$\frac{1}{a_n}\sum_{j=1}^n x_j = b_n - \frac{1}{a_n}\sum_{j=0}^{n-1}(a_{j+1} - a_j)b_j \to 0.$$

□

Lemma 3.3.3. *Let $\{X_n\}$ be independent random variables with $\mathbb{E}[X_n] = 0$. Suppose $\{a_n\}$ is a sequence of numbers such that $0 < a_n \uparrow \infty$. Suppose $\varphi : \mathbb{R} \to \mathbb{R}_+$ is positive, even and continuous function on $\mathbb{R}$ such that*

$$\frac{\varphi(x)}{|x|} \uparrow, \quad \frac{\varphi(x)}{x^2} \downarrow, \quad \text{as } |x| \uparrow.$$

Assume φ satisfies the following condition:

$$\sum_n \frac{\mathbb{E}[\varphi(X_n)]}{\varphi(a_n)} < \infty.$$

Then

$$\sum_n \frac{X_n}{a_n} \quad \text{converges almost surely.}$$

Consequently,

$$\frac{1}{a_n}\sum_{j=1}^n X_j \to 0, \quad \text{almost surely.}$$

Note that there is a wide range of choice of φ in Lemma 3.3.3, for instance $\varphi(x) = |x|^p$ with $1 \leqslant p \leqslant 2$.

Proof. Define

$$Y_n = X_n \mathbb{1}_{\{|X_n| \leqslant a_n\}}.$$

We will use three series theorem for $\{X_n/a_n\}$ and $\{Y_n/a_n\}$ with $A = 1$. First of all,

$$\begin{aligned}\sum_n \mathbb{P}\left[\frac{X_n}{a_n} \neq \frac{Y_n}{a_n}\right] &= \sum_n \mathbb{P}[|X_n| > a_n] \\ &= \sum_n \mathbb{P}[\varphi(X_n) > \varphi(a_n)] \leqslant \sum_n \frac{\mathbb{E}[\varphi(X_n)]}{\varphi(a_n)} < \infty.\end{aligned}$$

Second,

$$\begin{aligned}
\sum_n \frac{|\mathbb{E}[Y_n]|}{a_n} &= \sum_n \frac{|\mathbb{E}[X_n \mathbb{1}_{\{|X_n|>a_n\}}]|}{a_n} && \text{(since } \mathbb{E}[X_n]=0\text{)} \\
&\leqslant \sum_n \mathbb{E}\left[\frac{|X_n|}{a_n}\mathbb{1}_{\{|X_n|>a_n\}}\right] \\
&\leqslant \sum_n \mathbb{E}\left[\frac{\varphi(X_n)}{\varphi(a_n)}\right] < \infty. && \text{(since } \varphi(x)/|x| \text{ is increasing)}
\end{aligned}$$

Finally,

$$\begin{aligned}
\sum_n \operatorname{var}(Y_n/a_n) \leqslant \sum_n \mathbb{E}\left[\frac{Y_n^2}{a_n^2}\right] &= \sum_n \mathbb{E}\left[\frac{X_n^2}{a_n^2}\mathbb{1}_{\{|X_n|\leqslant a_n\}}\right] \\
&\leqslant \sum_n \mathbb{E}\left[\frac{\varphi(X_n)}{\varphi(a_n)}\right] < \infty. && \text{(since } \varphi(x)/x^2 \text{ is decreasing)}
\end{aligned}$$

□

Proof of Theorem 3.3.1—(3.3.1). Define

$$Y_n = X_n \mathbb{1}_{\{|X_n|\leqslant n\}}\,.$$

The sequences $\{X_n\}$ and $\{Y_n\}$ are equivalent:

$$\sum_n \mathbb{P}[X_n \neq Y_n] = \sum_n \mathbb{P}[|X_1|>n] < \infty.$$

To apply Lemma 3.3.3 to $\{Y_n - \mathbb{E}[Y_n]\}$ with $a_n = n$ and $\varphi(x) = x^2$, suppose Z has the same law as $\{X_n\}$, we calculate

$$\begin{aligned}
\sum_n \frac{\operatorname{var}(Y_n)}{n^2} \leqslant \sum_n \frac{\mathbb{E}[Y_n^2]}{n^2} &= \sum_n \frac{1}{n^2}\mathbb{E}\left[X^2 \mathbb{1}_{\{|X|\leqslant n\}}\right] \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{E}[X^2 \mathbb{1}_{\{j-1<|X|\leqslant j\}}] \\
&= \sum_{j=1}^{\infty} \mathbb{E}[X^2 \mathbb{1}_{\{j-1<|X|\leqslant j\}}] \sum_{n\geqslant j} \frac{1}{n^2} \\
&= \sum_{j=1}^{\infty} \mathbb{E}[X^2 \mathbb{1}_{\{j-1<|X|\leqslant j\}}] \frac{O(1)}{j} \\
&\leqslant O(1) \sum_{j=1}^{\infty} \mathbb{E}[|X| \mathbb{1}_{\{j-1<|X|\leqslant j\}}] = O(1)\mathbb{E}[|X|] < \infty.
\end{aligned}$$

Applying Lemma 3.3.3 to $\{Y_n - \mathbb{E}[Y_n]\}$, we obtain

$$\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mathbb{E}[Y_j]) \to 0, \quad \text{almost surely.}$$

Therefore

$$\lim_n \frac{S_n}{n} = \lim_n \frac{1}{n}\sum_{j=1}^{n} Y_j = \lim_n \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Y_j] = \mathbb{E}[X].$$

□

Proof of Theorem 3.3.1—(3.3.2). Since $\mathbb{E}[|X_1|] = \infty$, we have $\mathbb{E}[|X_1|/A] = \infty$ for any $A > 0$. Thus

$$\sum_n \mathbb{P}[|X_n| > An] = \infty.$$

By Borel-Cantelli lemma, we have

$$\mathbb{P}[|X_n| > An \ i.o.] = 1.$$

Note that $|S_n - S_{n-1}| = |X_n| > An$ implies $|S_n| > An/2$ or $|S_{n-1}| > An/2$. Thus

$$\mathbb{P}\left[|S_n| \geqslant \frac{An}{2} \ i.o.\right] = 1.$$

This means that, for each A, there exists a null set $\mathcal{N}(A)$ such that

$$\limsup_n \frac{|S_n|}{n} \geqslant \frac{A}{2}, \quad \text{on } \Omega \backslash \mathcal{N}(A).$$

Take $\mathcal{N} = \cup_{m=1}^{\infty} \mathcal{N}(m)$, then it is still a null set and

$$\limsup_n \frac{|S_n|}{n} = \infty, \quad \text{on } \Omega \backslash \mathcal{N}.$$

□

Example (Example 3.1.3 continued)**.** *In this example, we have $S_n/n \to 0$ in probability. However, by strong law of large numbers, we have*

$$\limsup_n \frac{S_n}{n} = +\infty, \quad \liminf \frac{S_n}{n} = -\infty, \quad \textit{almost surely.}$$

Example 3.3.4. *Suppose $\{X_n\}$ are i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathrm{var}(X_1) = 1$. Define $S_n = \sum_{j=1}^n X_j$. For $\epsilon > 0$, we have*

$$\frac{S_n}{n^{1/2}(\log n)^{1/2+\epsilon}} \to 0, \quad \textit{a.s.} \tag{3.3.3}$$

Proof. By strong law of large numbers, we have $S_n/n \to 0$ a.s. Thus (3.3.3) is a better conclusion. For the convergence in (3.3.3), by law of the iterated logarithm in Example 5.5.5, we have

$$\limsup_n \frac{S_n}{\sqrt{2n \log\log n}} = 1, \quad \textit{a.s.}$$

Thus the convergence in (3.3.3) is not far from the best possible.

Next, we show (3.3.3). Set $a_n = n^{1/2}(\log n)^{1/2+\epsilon}$ and $Y_n = \frac{X_n}{a_n} \mathbb{1}_{\{|X_n| \leqslant a_n\}}$. Then we have

$$\sum_n \mathbb{P}[|X_n| > a_n] \leqslant \sum_n \frac{\mathbb{E}[X_n^2]}{a_n^2} = \sum_n \frac{1}{a_n^2} < \infty;$$

$$\sum_n |\mathbb{E}[Y_n]| \leqslant \sum_n \mathbb{E}\left[\frac{|X_n|}{a_n} \mathbb{1}_{\{|X_n| \geqslant a_n\}}\right] \leqslant \sum_n \frac{\mathbb{E}[X_n^2]}{a_n^2} = \sum_n \frac{1}{a_n^2} < \infty;$$

$$\sum_n \mathrm{var}(Y_n) \leqslant \sum_n \frac{\mathbb{E}[X_n^2]}{a_n^2} = \sum_n \frac{1}{a_n^2} < \infty.$$

By three series law, we see $\sum_n X_n/a_n$ converges a.s. Combining with Lemma 3.3.2, we have $S_n/a_n \to 0$ a.s. □

Example 3.3.5. *Suppose $\{X_n\}$ are i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[|X_1|^p] < \infty$ for some $p \in (1,2)$. Define $S_n = \sum_{j=1}^n X_j$. Then*

$$\frac{S_n}{n^{1/p}} \to 0, \quad a.s.$$

Proof. Strong law of large numbers gives $S_n/n \to 0$ a.s. The conclusion here is better. Set $a_n = n^{1/p}$ and $Y_n = X_n \mathbb{1}_{\{|X_n| \leqslant a_n\}}$. Then $\{X_n\}$ and $\{Y_n\}$ are equivalent, because

$$\sum_n \mathbb{P}[|X_n| > a_n] = \sum_n \mathbb{P}[|X_1|^p > n] \leqslant \mathbb{E}[|X_1|^p] < \infty.$$

Define $T_n = \sum_{j=1}^n Y_j$. It suffices to show $T_n/a_n \to 0$ a.s. We will show

$$\sum_n \operatorname{var}(Y_n/a_n) < \infty, \tag{3.3.4}$$

and

$$\frac{\mathbb{E}[T_n]}{a_n} \to 0. \tag{3.3.5}$$

Assuming (3.3.4) holds, by Lemma 3.3.3 with $\varphi(x) = x^2$, we have

$$\frac{T_n - \mathbb{E}[T_n]}{a_n} \to 0, \quad a.s.$$

Combining with (3.3.5), we have $T_n/a_n \to 0$ a.s. Thus it remains to show (3.3.4) and (3.3.5).

Suppose Z has the same law as X_1. We have

$$\begin{aligned}
\sum_{n\geqslant 1} \operatorname{var}(Y_n/a_n) &\leqslant \sum_{n\geqslant 1} \frac{1}{a_n^2}\mathbb{E}[Y_n^2] = \sum_{n\geqslant 1} \frac{1}{a_n^2}\mathbb{E}\left[Z^2 \mathbb{1}_{\{|Z|\leqslant a_n\}}\right] \\
&= \sum_{n\geqslant 1} \frac{1}{a_n^2} \sum_{j=1}^n \mathbb{E}\left[Z^2 \mathbb{1}_{\{a_{j-1}<|Z|\leqslant a_j\}}\right] \\
&= \sum_{j\geqslant 1} \mathbb{E}\left[Z^2 \mathbb{1}_{\{a_{j-1}<|Z|\leqslant a_j\}}\right] \sum_{n\geqslant j} \frac{1}{a_n^2} \\
&= \sum_{j\geqslant 1} \mathbb{E}\left[Z^2 \mathbb{1}_{\{a_{j-1}<|Z|\leqslant a_j\}}\right] O(1) a_j^{p-2} \\
&\leqslant O(1) \sum_{j\geqslant 1} \mathbb{E}\left[|Z|^p \mathbb{1}_{\{a_{j-1}<|Z|\leqslant a_j\}}\right] = O(1)\mathbb{E}[|Z|^p] < \infty.
\end{aligned}$$

This gives (3.3.4).

Finally, we derive (3.3.5): for $1 \leqslant m \leqslant n$,

$$\begin{aligned}
\frac{|\mathbb{E}[T_n]|}{a_n} &\leqslant \frac{1}{a_n} \sum_{j=1}^n \mathbb{E}\left[|Z| \mathbb{1}_{\{|Z|>a_j\}}\right] \\
&\leqslant \frac{1}{a_n} \sum_{j=1}^n a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_j\}}\right] \\
&\leqslant \frac{1}{a_n} \sum_{j=1}^m a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_j\}}\right] + \frac{1}{a_n} \sum_{j=m}^n a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_j\}}\right] \\
&\leqslant \frac{1}{a_n} \sum_{j=1}^m a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_j\}}\right] + \frac{1}{a_n} \sum_{j=m}^n a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_m\}}\right] \\
&\leqslant \frac{1}{a_n} \sum_{j=1}^m a_j^{1-p}\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_j\}}\right] + O(1)\mathbb{E}\left[|Z|^p \mathbb{1}_{\{|Z|>a_m\}}\right].
\end{aligned}$$

We let $n \to \infty$ and then $m \to \infty$, we obtain (3.3.5). $\square$

Example 3.3.6 (Renewal theory)**.** *Suppose $\{X_n\}$ are i.i.d. with $0 < X_1 < \infty$. Let $T_n = \sum_{j=1}^n X_j$. For a concrete situation, consider a janitor who replaces a light bulb the instant it burns out. Suppose a bulb is put in at time 0 and let X_i be the lifetime of the ith light bulb. In this interpretation, T_n is the time that the nth light bulb burns out and $N_t = \sup\{n : T_n \leqslant t\}$ is the number of light bulbs that have burnt out by time t. If $\mathbb{E}[X_1] = \mu \leqslant \infty$, then*

$$\frac{N_t}{t} \to \frac{1}{\mu}, \quad \text{as } t \to \infty, \quad \text{almost surely;} \tag{3.3.6}$$

$$\frac{\mathbb{E}[N_t]}{t} \to \frac{1}{\mu}, \quad \text{as } t \to \infty. \tag{3.3.7}$$

If $\mu = \infty$, then $1/\mu = 0$.

Proof of (3.3.6). By strong law of large number, we have $T_n/n \to \mu$ almost surely. By the definition of N_t, we have $T_{N_t} \leqslant t \leqslant T_{N_t+1}$. Thus

$$\frac{T_{N_t}}{N_t} \leqslant \frac{t}{N_t} \leqslant \frac{T_{N_t+1}}{N_t+1}\frac{N_t+1}{N_t}.$$

Strong law of large number implies that, there exists a null set $\mathcal{N}$ such that

$$\frac{T_n(\omega)}{n} \to \mu \text{ as } n \to \infty, \quad N_t(\omega) \uparrow \infty \text{ as } t \uparrow \infty, \quad \forall \omega \in \Omega \backslash \mathcal{N}.$$

Therefore,

$$\frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \to \mu, \quad \frac{N_t(\omega)+1}{N_t(\omega)} \to 1, \quad \forall \omega \in \Omega \backslash \mathcal{N}.$$

This gives the conclusion. $\square$

Proof of (3.3.7). We have shown the almost sure convergence, in order to have convergence in L^1, we need to check the collection $\{N_t/t : t > 0\}$ is UI. To this end, we will show that it is bounded in L^2.

Since $X_1 > 0$ a.s., there exists $\delta > 0$ such that

$$\mathbb{P}[X_1 \geqslant \delta] = p > 0.$$

Define $X'_n = \delta \mathbb{1}_{\{X_n \geqslant \delta\}}$. The sequence $\{X'_n\}$ is i.i.d. Bernoulli random variables and define T'_n and N'_t for $\{X'_n\}$. Since $X'_n \leqslant X_n$, we have $T'_n \leqslant T_n$ and $N_t \leqslant N'_t$. From Exercise 3.3.7, $\mathbb{E}[(N'_t/t)^2]$ is bounded uniform over t. Thus $\mathbb{E}[(N_t/t)^2] \leqslant \mathbb{E}[(N'_t/t)^2]$ is bounded uniform over t. This completes the proof. $\square$

Exercise 3.3.7. *Suppose $\{X_n\}$ are i.i.d. Bernoulli random variables with parameter $p \in (0,1)$: $\mathbb{P}[X_1 = 1] = p$ and $\mathbb{P}[X_1 = 0] = 1 - p$. Define $S_n = \sum_{j=1}^n X_j$ and $N_t = \sup\{n : S_n \leqslant t\}$. We have $\mathbb{E}[N_t] = O(t)$ and $\mathbb{E}[N_t^2] = O(t^2)$ as $t \to \infty$.*

Proof. For $t > 0$ and $n \geqslant 1$, we have $\{N_t \geqslant n\} = \{S_n \leqslant t\}$. Note that S_n takes values in $\{0, 1, \ldots, n\}$. We have

$$\begin{cases} t \geqslant n, & \mathbb{P}[N_t \geqslant n] = 1, \\ t < n, & \mathbb{P}[N_t \geqslant n] = \mathbb{P}[S_n \leqslant t] = \sum_{j=0}^{[t]} \binom{n}{j} p^j (1-p)^{n-j}. \end{cases}$$

Let us evaluate $\mathbb{E}[N_t]$:

$$\mathbb{E}[N_t] = \sum_n \mathbb{P}[N_t \geqslant n] = [t] + \sum_{n>t}\sum_{j=0}^{[t]} \binom{n}{j} p^j (1-p)^{n-j} = t + O(1).$$

Then we evaluate $\mathbb{E}[N_t^2]$:

$$\begin{aligned}\mathbb{E}[N_t^2] &= \sum_{n\geqslant 1} n^2\mathbb{P}[N_t = n] \\ &= \sum_{n\geqslant 1} n^2\mathbb{P}[N_t \geqslant n-1] - \sum_{n\geqslant 1} n^2\mathbb{P}[N_t \geqslant n] \\ &= 1 + \sum_{n\geqslant 1}(2n+1)\mathbb{P}[N_t \geqslant n] \\ &= O(t^2) + \sum_{n\geqslant t}(2n+1)\sum_{j=0}^{[t]}\binom{n}{j}p^j(1-p)^{n-j} = O(t^2).\end{aligned}$$

□

3.4 Exercises

Exercise 3.4.1. *For arbitrary $\{X_n\}$, if $\sum_n \mathbb{E}[|X_n|] < \infty$, then $\sum_n X_n$ converges almost surely.*

Exercise 3.4.2. *If $\{X_n\}$ is a sequence of independent random variables, then the convergence of the series $\sum_n X_n$ in probability is equivalent to its almost sure convergence.*

Exercise 3.4.3. *Let $\{X_n\}$ be independent Poisson random variables with $\mathbb{E}[X_n] = \lambda_n$. Define $S_n = \sum_{j=1}^n X_j$. If $\sum \lambda_n = \infty$ then $S_n/\mathbb{E}[S_n] \to 1$ almost surely.*

Exercise 3.4.4. *Suppose $\{X_n\}$ satisfies the following assumptions:*

- *There are constants C and μ such that $\operatorname{var}(X_n) \leqslant C$ and $\mathbb{E}[X_n] = \mu$ for all n.*
- *There exists a function $f : \mathbb{N} \to \mathbb{R}_+$ such that $f(n) \to 0$ and $\operatorname{cov}(X_i, X_j) \leqslant f(|i-j|)$ for all i, j.*

Show that

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{in } L^2(\Omega, \mathcal{F}, \mathbb{P}).$$

Exercise 3.4.5. *Suppose $\{X_n\}$ is a sequence of random variables.*

(1) Show that $X_n \to 0$ almost surely if and only if, for any $\epsilon > 0$, $\mathbb{P}[|X_n| > \epsilon \text{ i.o.}] = 0$.

(2) Show that there exists a sequence $a_n \uparrow \infty$ such that X_n/a_n converges to 0 almost surely.

Exercise 3.4.6. *Suppose that $\{X_k : k \in \mathbb{N}\}$ are i.i.d. and set $S_n := \sum_{j=1}^n X_j$. Show that $S_n/n \to 0$ almost surely if and only if the following two conditions are satisfied:*

(1)

$$\frac{S_n}{n} \to 0 \quad \text{in probability,}$$

(2)

$$\frac{S_{2^n}}{2^n} \to 0 \quad \text{a.s.}$$

An alternative set of conditions is (1) as above and

(3) for any $\epsilon > 0$,

$$\sum_n \mathbb{P}\left[|S_{2^{n+1}} - S_{2^n}| > 2^n\epsilon\right] < \infty.$$

Exercise 3.4.7. *If $\{X_n : n \in \mathbb{N}\}$ is a sequence of positive, non-decreasing random variables such that $X_n \to \infty$ almost surely, show that*

$$\limsup_{n\to\infty} \frac{\log X_n}{\log \mathbb{E}[X_n]} \leqslant 1 \quad a.s.$$

Exercise 3.4.8. *Suppose that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$ for pairwise independent events $\{A_n\}$. Let $S_n = \sum_{i=1}^{n} \mathbb{1}_{A_i}$ be the number of events occurring among the first n events.*

(1) Show that $\operatorname{var}(S_n) \leqslant \mathbb{E}[S_n]$ and deduce from it that $S_n/\mathbb{E}[S_n] \to 1$ in probability.

(2) Applying Borel-Cantelli lemma to show that $S_{n_k}/\mathbb{E}[S_{n_k}] \to 1$ almost surely as $k \to \infty$, where $n_k := \min\{n : \mathbb{E}[S_n] \geqslant k^2\}$.

(3) Show that $\mathbb{E}\left[S_{n_{k+1}}\right]/\mathbb{E}\left[S_{n_k}\right] \to 1$ as $k \to \infty$. Deduce that $S_n/\mathbb{E}[S_n] \to 1$ almost surely.

Exercise 3.4.9 (YCMC2013)**.** *Let $\{X_n\}$ be a sequence of random variables.*

(1) Assume that $\sum_{n=0}^{\infty} \mathbb{P}[|X_n| > n] < \infty$. Show that $\limsup_{n\to\infty} \frac{|X_n|}{n} \leqslant 1$ almost surely.

(2) Show that $\{X_n\}$ converges in probability to 0 if and only if for certain $r > 0$,

$$\mathbb{E}\left[\frac{|X_n|^r}{1+|X_n|^r}\right] \to 0.$$

Exercise 3.4.10 (YCMC2014)**.** *Suppose X_n is n-dimensional standard Gaussian random vector and denote by $\|X_n\|$ its Euclidean norm. Show that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left[1-\epsilon \leqslant \frac{\|X_n\|}{\sqrt{n}} \leqslant 1+\epsilon\right] = 1.$$

Exercise 3.4.11 (YCMC2016)**.** *Let $\{X_n\}$ be i.i.d. such that $\mathbb{E}[X_1] = -1$. Let $S_n = X_1 + \cdots + X_n$ for all $n \geqslant 1$, and let T be the total number of $n \geqslant 1$ satisfying $S_n \geqslant 0$. Compute $\mathbb{P}[T = \infty]$.*

4 Central Limit Theorem

4.1 Characteristic function

Definition 4.1.1. *For any random variable X with distribution μ, its characteristic function is defined to be:*

$$f : \mathbb{R} \to \mathbb{C}, \quad f(t) := \mathbb{E}\left[e^{itX}\right] = \int e^{itx}\mu[dx].$$

The first equal sign is the definition, the second equal sign is due to Theorem 1.3.3.

In analysis, the characteristic function is known as the Fourier transform of μ.

Lemma 4.1.2. *The characteristic function has the following basic properties.*

(1) $f(0) = 1$, $|f(t)| \leqslant 1$, *and* $f(-t) = \overline{f(t)}$.

(2) f *is uniformly continuous.*

(3) We write f_X *for the characteristic function of* X*, then for any real numbers* a *and* b*, we have*

$$f_{aX+b}(t) = f_X(at)e^{ibt}, \quad f_{-X}(t) = f_X(-t).$$

(4) If $\{f_n\}$ *are characteristic functions and* $\{\lambda_n\}$ *are positive numbers such that* $\sum_n \lambda_n = 1$*, then*

$$\sum_n \lambda_n f_n$$

is a characteristic function. Briefly: convex combination of characteristic functions is a characteristic function.

(5) Suppose X *and* Y *are independent, then the characteristic function of* $X + Y$ *is* $f_X \times f_Y$.

(6) If f *is a characteristic function, so is* $|f|^2$.

Proof of (2). We have

$$|f(t+\epsilon) - f(t)| = |\int e^{itx}(e^{i\epsilon x} - 1)\mu[dx]| \leqslant \int |e^{i\epsilon x} - 1|\mu[dx].$$

By bounded convergence theorem, the last term goes to zero as $\epsilon \to 0$. □

Proof of (4). If $\{\mu_n\}$ are the corresponding probability measures, then $\sum_n \lambda_n\mu_n$ is a probability measure. □

Proof of (5).

$$\mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}].$$

□

Proof of (6). Suppose X and Y are i.i.d., then

$$\mathbb{E}[e^{it(X-Y)}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{-itY}] = f(t)f(-t) = |f(t)|^2.$$

□

We emphasize that the introduction of characteristic function simplifies the calculation for sum of independent random variables.

Lemma 4.1.3. *Suppose X and Y are independent.*

- *The distribution function of $X+Y$ is $F_X * F_Y$.*
- *The characteristic function of $X+Y$ is $f_X \times f_Y$.*

Example 4.1.4. *We list characteristic functions of a few well-known probability measures.*

- *Dirac mass at a:*
$$e^{iat}.$$
- *Bernoulli distribution on $\{-1,+1\}$ with $p=1/2$:*
$$\cos(t).$$
- *Uniform distribution on $[-a,a]$:*
$$\frac{\sin at}{at}.$$
- *Exponential distribution with density $\lambda e^{-\lambda x}$:*
$$\frac{\lambda}{\lambda - it}.$$
- *Normal distribution $\mathcal{N}(m,\sigma^2)$:*
$$\exp\left(imt - \frac{\sigma^2 t^2}{2}\right).$$
- *Poisson distribution with parameter $\lambda > 0$:*
$$\exp\left(\lambda(e^{it}-1)\right).$$
- *Geometric distribution with probability $p \in (0,1)$:*
$$\frac{pe^{it}}{1-(1-p)e^{it}}.$$

Exercise. *Calculate examples in Section 1.6 using characteristic functions.*

4.2 Uniqueness and inversion

We have defined characteristic function for each probability measure. Then the following question comes: given a characteristic function, how can we find the corresponding probability measure? The formula for doing is called inversion formula.

Theorem 4.2.1. *Suppose f is the characteristic function for the probability measure μ. For $x < y$, we have*
$$\mu[(x,y)] + \frac{1}{2}\mu[\{x\}] + \frac{1}{2}\mu[\{y\}] = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-itx} - e^{-ity}}{it} f(t)dt.$$

Observe that the integrand in the right hand side is bounded by $O(|t|^{-1})$ as $|t| \to \infty$; yet we cannot assert the "infinite integral" exists (in the Lebesgue sense). Indeed, it does not in general. The fact that the limit in the right hand side does exist is part of the assertion.

Proof. We claim that

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx}-e^{-ity}}{it}\left(\int e^{itz}\mu[dz]\right)dt=\int\mu[dz]\int_{-T}^{T}\frac{e^{it(z-x)}-e^{it(z-y)}}{2\pi it}dt. \tag{4.2.1}$$

This is true due to Fubini's theorem, because the integrand in the right hand side is bounded by integrable function with respect to $\mu[dz]dt$ on $\mathbb{R}\times[-T,T]$:

$$\left|\frac{e^{it(z-x)}-e^{it(z-y)}}{it}\right|\leqslant|x-y|.$$

This proves (4.2.1). Define

$$I(T,z;x,y):=\int_{-T}^{T}\frac{e^{it(z-x)}-e^{it(z-y)}}{2\pi it}dt.$$

It is clear that

$$I(T,z;x,y)=\int_{0}^{T}\frac{\sin(t(z-x))}{\pi t}dt-\int_{0}^{T}\frac{\sin(t(z-y))}{\pi t}dt.$$

The quantity I is bounded in T, because for any $w\geqslant 0$

$$0\leqslant sgn(\alpha)\int_{0}^{w}\frac{\sin(\alpha t)}{t}dt\leqslant\int_{0}^{\pi}\frac{\sin t}{t}dt. \tag{4.2.2}$$

Therefore, we can interchange the limit and the integral. It remains to derive the limit of I as $T\to\infty$. Note that

$$\lim_{T\to\infty}\int_{0}^{T}\frac{\sin(\alpha t)}{t}dt=\frac{\pi}{2}sgn(\alpha). \tag{4.2.3}$$

As a consequence, we have

$$\lim_{T\to\infty}I(T,z;x,y)=\begin{cases}0, & \text{if } z<x<y \text{ or } x<y<z,\\ 1/2, & \text{if } z=x \text{ or } z=y,\\ 1, & \text{if } x<z<y.\end{cases}$$

Therefore,

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx}-e^{-ity}}{it}f(t)dt&=\int\mu[dz]\lim_{T\to\infty}I(T,z;x,y)\\&=\mu[(x,y)]+\frac{1}{2}\mu[\{x\}]+\frac{1}{2}\mu[\{y\}].\end{aligned}$$

□

Corollary 4.2.2. *If two probability measures μ and ν have the same characteristic function, then $\mu=\nu$.*

Proof. Suppose A_μ is the set of atoms of μ and A_ν is the set of atoms of ν. From the inversion formula, we have

$$\mu[(a,b)]=\nu[(a,b)],\quad \text{for any } a,b\in\mathbb{R}\backslash(A_\mu\cup A_\nu).$$

Note that A_μ and A_ν are countable, thus $\mathbb{R}\backslash(A_\mu\cup A_\nu)$ is dense. The intervals $\{(a,b):a,b\in\mathbb{R}\backslash(A_\mu\cup A_\nu)\}$ generates $\mathcal{B}$. Thus $\mu=\nu$. □

Corollary 4.2.3. *Suppose f is the characteristic function for the distribution function F. If $f \in L^1(\mathbb{R})$, then F is differentiable and*

$$p(x) = F'(x) = \frac{1}{2\pi}\int e^{-ixt} f(t)dt.$$

In other words, when $f \in L^1(\mathbb{R})$, we have

$$p(x) = \frac{1}{2\pi}\int e^{-ixt} f(t)dt, \quad f(t) = \int e^{itx} p(x)dx.$$

Proof. Write the inversion formula in terms of F:

$$\frac{F(y)+F(y-)}{2} - \frac{F(x)+F(x-)}{2} = \frac{1}{2\pi}\int \frac{e^{-itx}-e^{-ity}}{it} f(t)dt.$$

Let $x \to y-$, the right hand side goes to zero (we are allowed to interchange the limit and the integral due to the hypothesis on f). Thus F is continuous. The above formula then writes:

$$F(y) - F(x) = \frac{1}{2\pi}\int \frac{e^{-itx}-e^{-ity}}{it} f(t)dt.$$

Divide both sides by $y-x$ and then let $y \to x+$, we obtain the conclusion (we are allowed to interchange the limit and the integral due to the hypothesis on f). $\square$

Corollary 4.2.4. *For each x, we have*

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} e^{-itx} f(t)dt = \mu[\{x\}]. \tag{4.2.4}$$

Moreover, we have

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} |f(t)|^2 dt = \sum_{x\in\mathbb{R}} \mu[\{x\}]^2. \tag{4.2.5}$$

Proof. For (4.2.4), we have

$$\begin{aligned}\frac{1}{2T}\int_{-T}^{T} e^{-itx} f(t)dt &= \int \mu[dz]\frac{1}{2T}\int_{-T}^{T} e^{it(z-x)}dt \\ &= \int \mu[dz]\frac{1}{T}\int_{0}^{T} \cos(t(z-x))dt \\ &= \int_{\mathbb{R}\setminus\{x\}} \mu[dz]\frac{\sin(T(z-x))}{T(z-x)} + \mu[\{x\}].\end{aligned}$$

Then, Eq. (4.2.4) holds by the following observation:

$$\lim_{T\to\infty}\int_{\mathbb{R}\setminus\{x\}} \mu[dz]\frac{\sin(T(z-x))}{T(z-x)} = 0.$$

For (4.2.5), recall that $|f|^2$ is the characteristic function of $X-Y$ where X, Y are i.i.d. The law of $X-Y$ is $\mu * \mu'$ where $\mu'[B] = \mu[-B]$ for all $B \in \mathcal{B}$. Applying (4.2.4) with $x=0$ and $X-Y$, we have

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} |f(t)|^2 dt = \mu * \mu'[\{0\}].$$

Note that

$$\mu * \mu'[\{0\}] = \int \mu'(\{-y\})\mu[dy] = \int \mu(\{y\})\mu[dy] = \sum_y \mu[\{y\}]^2.$$

$\square$

Definition 4.2.5. *The random variable X is symmetric if X and $-X$ have the same law.*

Lemma 4.2.6. *The random variable X is symmetric if and only if its characteristic function is real-valued for all t.*

Proof. Suppose the characteristic function of X is f. Then the characteristic function of $-X$ is $\overline{f}$. By Corollary 4.2.2, the random variables X and $-X$ have the same law if and only if $f = \overline{f}$. □

4.3 Characteristic function and convergence

The goal of this section is the following convergence theorem.

Theorem 4.3.1. *Let $\{\mu_n\}$ be a sequence of probability measures with characteristic functions $\{f_n\}$. Suppose that*

(a) f_n converges everywhere in $\mathbb{R}$ and defines the limiting function f;

(b) f is continuous at $t = 0$.

Then we have

(1) $\mu_n \Longrightarrow \mu$ where μ is a probability measure;

(2) the characteristic function of μ is f.

We first discuss the converse direction of this theorem.

Lemma 4.3.2. *Let $\{\mu_n\}$ and μ be a sequence of probability measures with characteristic functions $\{f_n\}$ and f respectively. If μ_n converges weakly to μ, then f_n converges to f uniformly in every finite interval. Furthermore, the family $\{f_n\}$ is equicontinuous on $\mathbb{R}$.*

Proof. Since the real part and the imaginary part of e^{itx} are bounded continuous functions, the weak convergence implies $f_n \to f$ pointwise.

We first show the equicontinuity, i.e. we show that, for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f_n(t+h) - f_n(t)| \leqslant \epsilon$ as long as $|h| \leqslant \delta$. For any $t \in \mathbb{R}$ and $h \in \mathbb{R}$, we have

$$\begin{aligned}|f_n(t+h) - f_n(t)| &\leqslant \int |e^{ihx} - 1| \mu_n[dx] \leqslant \int_{\{|x| \leqslant A\}} |hx| \mu_n[dx] + \int_{\{|x|>A\}} 2\mu_n[dx] \\ &\leqslant |h|A + 2\int_{\{|x|>A\}} \mu_n[dx].\end{aligned}$$

For any $\epsilon > 0$, there exists $n_0 = n_0(A, \epsilon)$ such that

$$|f_n(t+h) - f_n(t)| \leqslant |h|A + 2\int_{\{|x|>A\}} \mu[dx] + \epsilon/4.$$

This gives the equicontinuity for $\{f_n\}$: for any $\epsilon > 0$, we choose A large enough such that $\mu[\{|x| > A\}] \leqslant \epsilon/4$, then for any h such that $|h| \leqslant \delta := \epsilon/(4A)$, we have

$$|f_n(t+h) - f_n(t)| \leqslant \epsilon, \quad \forall t \in \mathbb{R}, \quad \forall n \geqslant n_0,$$

as desired.

Next, we show the uniform convergence on compact interval I, i.e. we show that, for any $\epsilon > 0$, there exists $n_0 = n_0(I, \epsilon)$ such that $|f_n(t) - f(t)| \leqslant \epsilon$ for $n \geqslant n_0$. From the equicontinuity, there exists δ such that $|f_n(t) - f_n(s)| \leqslant \epsilon$ as long as $|t-s| \leqslant \delta$. Choose finite sequence of points $\{a_1, \ldots, a_{m_0}\} \subset I$ such that

$\cup_{j=1}^{m_0}(a_j-\delta, a_j+\delta)$ is a cover of I. By the pointwise convergence, we have $f_n(a_j) \to f(a_j)$ for $1 \leqslant j \leqslant m_0$. Thus there exists $n_0 = n_0(I, \epsilon)$ such that

$$|f_n(a_j) - f(a_j)| \leqslant \epsilon, \quad \forall 1 \leqslant j \leqslant m_0, \quad \forall n \geqslant n_0.$$

For any $t \in I$, there exists j such that $t \in (a_j - \delta, a_j + \delta)$, thus

$$|f_n(t) - f(t)| \leqslant |f_n(t) - f_n(a_j)| + |f_n(a_j) - f(a_j)| + |f(a_j) - f(t)| \leqslant 3\epsilon, \quad \forall n \geqslant n_0.$$

This gives the uniform convergence. □

By Proposition 2.3.2, we know that $\{\mu_n\}$ contains convergent subsequence: $\mu_{n_k} \Longrightarrow \mu$ where μ is a subprobability measure. To show Theorem 4.3.1, we need to argue that all subsequential limit are the same which is given by f. To this end, we first argue that, under the assumption of the theorem, any subsequential limit is indeed a probability measure, which needs the following lemma.

Lemma 4.3.3. *Suppose f is the characteristic function of the probability measure μ. For each $A > 0$, we have*

$$\mu([-2A, 2A]) \geqslant A \left| \int_{-A^{-1}}^{A^{-1}} f(t)dt \right| - 1.$$

Proof. By the proof of Theorem 4.2.1, we have

$$\frac{1}{2T} \int_{-T}^{T} f(t)dt = \int \frac{\sin(Tx)}{Tx} \mu[dx].$$

Thus

$$\frac{1}{2T} \left| \int_{-T}^{T} f(t)dt \right| \leqslant \mu([-2A, 2A]) + \frac{1}{2TA} (1 - \mu([-2A, 2A])) = \left(1 - \frac{1}{2TA}\right) \mu([-2A, 2A]) + \frac{1}{2TA}.$$

Set $T = A^{-1}$, we obtain the conclusion. □

Proof of Theorem 4.3.1. Suppose μ_{n_k} is a convergent subsequence of $\{\mu_n\}$ and denote the limit by μ. First, we argue that μ is a probability measure. By the above lemma, we have, (when $\pm 2\delta^{-1} \in \mathcal{C}_\mu$)

$$\mu(\mathbb{R}) \geqslant \mu([-2\delta^{-1}, 2\delta^{-1}]) = \lim_k \mu_{n_k}([-2\delta^{-1}, 2\delta^{-1}]) \geqslant \limsup_k \frac{1}{\delta} \left| \int_{-\delta}^{\delta} f_{n_k}(t)dt \right| - 1.$$

Since $f_{n_k} \to f$ everywhere, and by bounded convergence theorem, we have

$$\limsup_k \frac{1}{\delta} \left| \int_{-\delta}^{\delta} f_{n_k}(t)dt \right| = \frac{1}{\delta} \left| \int_{-\delta}^{\delta} f(t)dt \right|.$$

Since f is continuous at zero, we have

$$\lim_{\delta \to 0} \frac{1}{2\delta} \left| \int_{-\delta}^{\delta} f(t)dt \right| = 1.$$

Thus, for any $\epsilon > 0$, there exists $\delta_0 > 0$ such that, for any $0 < \delta \leqslant \delta_0$,

$$\frac{1}{\delta} \left| \int_{-\delta}^{\delta} f(t)dt \right| \geqslant 2 - \epsilon.$$

Therefore,

$$\mu(\mathbb{R}) \geqslant 1 - \epsilon.$$

This holds for any $\epsilon > 0$. Thus $\mu(\mathbb{R}) = 1$ and μ is indeed a probability measure.

Let g be the characteristic function of μ. By Lemma 4.3.2, we know that $f_{n_k} \to g$ everywhere. By the hypothesis, $g \equiv f$. We see that any subsequential limit has the characteristic function f, and hence, by Theorem 4.2.1, any subsequential limit has the same probability measure whose characteristic function is given by f. □

Corollary 4.3.4. *Suppose $\{\mu_n\}$ and μ are probability measures whose characteristic functions are $\{f_n\}$ and f respectively. Then*

$$\mu_n \Longrightarrow \mu \quad \Longleftrightarrow \quad f_n \to f \text{ uniformly on every finite interval.}$$

Proof. The direction $\Longrightarrow$ is guaranteed by Lemma 4.3.2. The direction $\Longleftarrow$ is due to Theorem 4.3.1 and the fact that the characteristic function f is uniformly continuous. □

The following examples show the necessity of the assumption in Lemma 4.3.2 and Theorem 4.3.1.

Example 4.3.5. *Let μ_n be the probability measure such that it has mass $1/2$ at $\{0\}$ and $1/2$ at $\{n\}$. Then $\mu_n \Longrightarrow \mu$ where μ has mass $1/2$ at $\{0\}$ and zero elsewhere. For the characteristic functions, we have*

$$f_n(t) = \frac{1}{2} + \frac{1}{2}e^{itn}.$$

They do not converge.

Example 4.3.6. *Let μ_n be the uniform distribution on $[-n, n]$. Then $\mu_n \Longrightarrow \mu$ where μ is identically zero. For the characteristic functions, we have*

$$f_n(t) = \begin{cases} \frac{\sin(nt)}{nt}, & \text{if } t \neq 0; \\ 1, & \text{if } t = 0. \end{cases}$$

They converge and the limiting function is

$$f(t) = \begin{cases} 0, & \text{if } t \neq 0; \\ 1, & \text{if } t = 0. \end{cases}$$

The limiting function is not continuous at zero.

The following examples are applications of the theorem.

Example 4.3.7. *Suppose $\{\mu_n\}, \{\nu_n\}$ and μ, ν are probability measures, and $\mu_n \Longrightarrow \mu$ and $\nu_n \Longrightarrow \nu$. Then $\mu_n * \nu_n \Longrightarrow \mu * \nu$.*

Proposition 4.3.8. *Suppose μ is a probability measure and f is its characteristic function.*

(1) If μ has a finite moment of order $k \geqslant 1$, then f has a bounded continuous derivative of order k given by

$$f^{(k)}(t) = \int (ix)^k e^{itx} \mu[dx].$$

(2) If f has a finite derivative of even order k at $t = 0$, then μ has a finite moment of order k.[1]

Proof. The first conclusion is clear. For the second conclusion, we start with $k = 2$ and suppose $f''(0)$ exists and is finite. We have

$$f''(0) = \lim_{\epsilon \to 0} \frac{f(\epsilon) - 2f(0) + f(-\epsilon)}{\epsilon^2}.$$

For the right hand side, we have

$$\begin{aligned} \frac{f(\epsilon) - 2f(0) + f(-\epsilon)}{\epsilon^2} &= \int \frac{e^{i\epsilon x} - 2 + e^{-i\epsilon x}}{\epsilon^2} \mu[dx] \\ &= -2 \int \frac{1 - \cos(\epsilon x)}{\epsilon^2} \mu[dx] \end{aligned}$$

[1] Bonus. Suppose f has finite (first order) derivative, can you conclude that μ has finite expectation?

By Fatou's lemma, we have

$$\int x^2\mu[dx] = 2\int \lim_{\epsilon\to 0}\frac{1-\cos(\epsilon x)}{\epsilon^2}\mu[dx] \leqslant 2\liminf_{\epsilon\to 0}\int \frac{1-\cos(\epsilon x)}{\epsilon^2}\mu[dx] = -f''(0).$$

Therefore, μ has finite second moment.

For general $2k$, suppose the conclusion holds for $2k-2$ and $f^{(2k)}(0)$ exists and is finite. By the induction hypothesis, we have

$$f^{(2k-2)}(0) = \int (ix)^{2k-2}e^{itx}\mu[dx].$$

Define

$$G(x) = \int_{-\infty}^{x} y^{2k-2}\mu[dy], \quad \forall x \in \mathbb{R}.$$

If $G(\infty) > 0$, then $G(\cdot)/G(\infty)$ is a distribution function. For its corresponding probability measure, its characteristic function is given by

$$g(t) = \frac{1}{G(\infty)}\int e^{itx}x^{2k-2}\mu[dx] = \frac{(-1)^{k-1}f^{(2k-2)}(t)}{G(\infty)}.$$

By the induction hypothesis, $g''(0)$ exists and is finite. By the proof of the case with $k=2$, we know that G has finite moment of order 2, and thus μ has finite moment of order $2k$ as desired.

If $G(\infty) = 0$, then $\mu = \delta_0$. Then $f \equiv 1$ and μ has finite moment of order $2k$. □

Corollary 4.3.9. *If μ has a finite moment of order $k \geqslant 1$, then f has the following expansion in the neighborhood of $t = 0$:*

$$f(t) = \sum_{j=0}^{k}\frac{\mu[(ix)^j]}{j!}t^j + o(|t|^k).$$

Moreover, we have

$$\left| f(t) - \sum_{j=0}^{k}\mathbb{E}\left[\frac{(itX)^j}{j!}\right]\right| \leqslant \mathbb{E}\left[|tX|^{k+1}\wedge 2|tX|^k\right].$$

Proof. The first conclusion is due to Taylor expansion. The second one is due to the following control:

$$\left|e^{ix} - \sum_{j=0}^{k}\frac{(ix)^j}{j!}\right| \leqslant \frac{|x|^{k+1}}{(k+1)!}\wedge\frac{2|x|^k}{k!}.$$

□

Example 4.3.10. *Suppose $\{X_n\}$ are i.i.d. and denote by $S_n = \sum_{j=1}^n X_j$. If $\mathbb{E}[|X_1|] < \infty$, then*

$$\frac{S_n}{n} \to \mathbb{E}[X_1], \quad \textit{in probability.}$$

Proof. Suppose f is the characteristic function of X_1 and denote by $m = \mathbb{E}[X_1]$. Then the characteristic function of S_n/n is given by

$$g_n(t) = \mathbb{E}\left[e^{itS_n/n}\right] = f(t/n)^n.$$

By Corollary 4.3.9, we have

$$f(t/n) = 1 + im\frac{t}{n} + o(t/n).$$

Therefore,

$$g_n(t) = \left(1 + im\frac{t}{n} + o(t/n)\right)^n \to e^{imt}, \quad n\to\infty.$$

In other words, the characteristic functions g_n converge to the characteristic function of the dirac measure δ_m. This gives the convergence in probability. □

Theorem 4.3.11. *Let $\{X_n\}$ be i.i.d. with $\mathbb{E}[X_1] = m$ and $\operatorname{var}(X_1) = \sigma^2 \in (0, \infty)$. Define $S_n = \sum_{j=1}^n X_j$. Then*

$$\frac{S_n - nm}{\sigma\sqrt{n}} \Longrightarrow \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is the standard normal distribution.

Proof. We may assume $m = 0$. Let us calculate the characteristic function of $S_n/\sqrt{n}$:

$$\mathbb{E}\left[\exp(itS_n/(\sigma\sqrt{n}))\right] = f(t/(\sigma\sqrt{n}))^n.$$

By Corollary 4.3.9, we have the expansion

$$f(t/(\sigma\sqrt{n})) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{\sigma^2 n}\right).$$

Therefore,

$$\mathbb{E}\left[\exp(itS_n/(\sigma\sqrt{n}))\right] \to e^{-\frac{t^2}{2}}, \quad n \to \infty.$$

This gives the weak convergence. $\square$

Compare Theorem 4.3.11 with Cramér's theorem in Section 1.6.

Example 4.3.12. *Let $\{X_n\}$ be i.i.d. Bernoulli random variables with $\mathbb{P}[X_n = 0] = \mathbb{P}[X_n = 1] = 1/2$. Set $S_n = \sum_{j=1}^n X_j$. Let us estimate $\mathbb{P}[S_{16} = 8]$. The exact probability is given by*

$$\mathbb{P}[S_{16} = 8] = \binom{16}{8} 2^{-16} = 0.1964.$$

The approximation given by central limit theorem is the following:

$$\mathbb{P}\left[S_{16} \in (7.5, 8.5)\right] = \mathbb{P}\left[\frac{S_{16} - 8}{2} \in (-0.25, 0.25)\right] \approx \mathbb{P}[\chi \in (-0.25, 0.25)] = 0.1974.$$

Example 4.3.13. *Let Z_λ have a Poisson distribution with mean λ. Show that*

$$\frac{Z_\lambda - \lambda}{\sqrt{\lambda}} \Longrightarrow \mathcal{N}(0,1), \quad \text{as } \lambda \to \infty.$$

Proof. Suppose $\{X_n\}$ are i.i.d. with Poisson distribution with mean one, then $\mathbb{E}[X_n] = 1$ and $\operatorname{var}(X_n) = 1$. Set $S_n = \sum_{j=1}^n X_j$, then S_n has the law of Poisson distribution with mean n. Central limit theorem implies

$$\frac{S_n - n}{\sqrt{n}} \Longrightarrow \mathcal{N}(0,1).$$

This gives the conclusion for $\lambda = n \to \infty$. For general λ, we have $S_{\lfloor\lambda\rfloor} \leqslant_{st} Z_\lambda \leqslant_{st} S_{\lfloor\lambda\rfloor+1}$. Thus for any x,

$$\mathbb{P}\left[\frac{S_{\lfloor\lambda\rfloor+1} - \lambda}{\sqrt{\lambda}} \leqslant x\right] \leqslant \mathbb{P}\left[\frac{Z_\lambda - \lambda}{\sqrt{\lambda}} \leqslant x\right] \leqslant \mathbb{P}\left[\frac{S_{\lfloor\lambda\rfloor} - \lambda}{\sqrt{\lambda}} \leqslant x\right].$$

As $\lambda \to \infty$, we have

$$\mathbb{P}\left[\frac{Z_\lambda - \lambda}{\sqrt{\lambda}} \leqslant x\right] \to \mathbb{P}[\chi \leqslant x] \text{ where } \chi \sim \mathcal{N}(0,1).$$

$\square$

Example 4.3.14. *Let X and Y are i.i.d. with mean zero and variance one. If $X + Y$ and $X - Y$ are independent, then the common law of X and Y is $\mathcal{N}(0,1)$.*

Proof. Let f be the characteristic function of X. By the hypothesis, we have $f'(0) = 0$ and $f''(0) = -1$. The characteristic function of $X+Y$ is $f(t)^2$ and the characteristic function of $X-Y$ is $f(t)f(-t)$. Since they are independent, the characteristic function of $2X = (X+Y)+(X-Y)$ is $f(t)^3 f(-t)$. In other words,

$$f(2t) = f(t)^3 f(-t). \tag{4.3.1}$$

First, we argue that f never vanishes. If $f(t_0) = 0$ for some t_0. By (4.3.1), we have $f(t_0/2) = 0$ or $f(-t_0/2) = 0$. By induction, we have either $f(t_0/2^n) = 0$ or $f(-t_0/2^n) = 0$. Since f is continuous at zero, this implies $f(0) = 0$, contradiction.

Next, we argue that $f(t) = f(-t)$. Define $g(t) = f(t)/f(-t)$. Then g has finite second derivative. Eq. (4.3.1) gives $g(2t) = g(t)^2$. By iteration, we have

$$g(t) = g(t/2^n)^{2^n} = (1 + o(t/2^n))^{2^n} \to 1, \quad n \to \infty.$$

Thus $g \equiv 1$. This gives $f(t) = f(-t)$. Then (4.3.1) comes

$$f(2t) = f(t)^4.$$

By iteration, we have

$$f(t) = f(t/2^n)^{4^n} = \left(1 - \frac{t^2}{2 \times 4^n} + o(t^2/4^n)\right)^{4^n} \to e^{-t^2/2}, \quad n \to \infty.$$

This gives $f(t) = e^{-t^2/2}$ as desired. □

4.4 The Lindeberg-Feller Theorem

The goal of this section is the following generalization of central limit theorem.

Theorem 4.4.1. *For each n, let $\{X_{n,m} : 1 \leqslant m \leqslant n\}$ be independent random variables with $\mathbb{E}[X_{n,m}] = 0$. Suppose*

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2] = \sigma^2 \in (0,\infty); \tag{4.4.1}$$

and, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right] = 0. \tag{4.4.2}$$

Then

$$S_n^{\#} := \sum_{m=1}^{n} X_{n,m} \Longrightarrow \mathcal{N}(0,\sigma^2).$$

Theorem 4.4.1 does imply Theorem 4.3.11: suppose $\{Y_n\}$ are i.i.d. with $\mathbb{E}[Y_1] = 0$ and $\mathrm{var}(Y_1) = \sigma^2 \in (0,\infty)$. Denote by $S_n = \sum_{j=1}^n Y_j$. Set $X_{n,m} = Y_m/\sqrt{n}$, then we have

$$\sum_{m=1}^{n} \mathbb{E}[X_{n,m}^2] = \sigma^2; \quad \sum_{m=1}^{n} \mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right] = \mathbb{E}\left[Y_1^2 \mathbb{1}_{\{|Y_1|>\epsilon\sqrt{n}\}}\right] \to 0, \quad \text{as } n \to \infty.$$

Thus $S_n/\sqrt{n} = S_n^{\#} \Longrightarrow \mathcal{N}(0,\sigma^2)$.

Proof. Define $f_{n,m}(t) = \mathbb{E}[\exp(itX_{n,m})]$, then the goal is to show that

$$\lim_n \prod_{m=1}^{n} f_{n,m}(t) = \exp(-t^2\sigma^2/2).$$

Define $\sigma_{n,m}^2 = \mathbb{E}[X_{n,m}^2]$, then we have

$$\sigma_{n,m}^2 \leqslant \epsilon^2 + \mathbb{E}[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}].$$

From the hypotheses, we have

$$\lim_n \max_m \sigma_{n,m}^2 = 0.$$

Define $g_{n,m}(t) = 1 - t^2\sigma_{n,m}^2/2$. Then $g_{n,m}$ approximates $f_{n,m}$:

$$\begin{aligned} |f_{n,m}(t) - g_{n,m}(t)| &\leqslant \mathbb{E}\left[|tX_{n,m}|^3 \wedge 2|tX_{n,m}|^2\right] && \text{(by Corollary 4.3.9)} \\ &\leqslant \mathbb{E}\left[|tX_{n,m}|^3 \mathbb{1}_{\{|X_{n,m}|\leqslant\epsilon\}}\right] + \mathbb{E}\left[2|tX_{n,m}|^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right] \\ &\leqslant t^3\epsilon\mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|\leqslant\epsilon\}}\right] + 2t^2\mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right]. \end{aligned}$$

Summing over m from 1 to n, from the hypotheses, we have

$$\limsup_n \sum_{m=1}^n |f_{n,m}(t) - g_{n,m}(t)| \leqslant \epsilon t^3\sigma^2.$$

Let $\epsilon \to 0$, we have

$$\lim_n \sum_{m=1}^n |f_{n,m}(t) - g_{n,m}(t)| = 0.$$

Since $|f_{n,m}(t)| \leqslant 1$ and $|g_{n,m}(t)| \leqslant 1$ for n large enough, thus

$$\left|\prod_{m=1}^n f_{n,m}(t) - \prod_{m=1}^n g_{n,m}(t)\right| \leqslant \sum_{m=1}^n |f_{n,m}(t) - g_{n,m}(t)| \to 0, \quad \text{as } n \to \infty.$$

It remains to show

$$\lim_n \prod_{m=1}^n g_{n,m}(t) = \exp(-t^2\sigma^2/2), \quad \text{or} \quad \lim_n \sum_{m=1}^n \log g_{n,m}(t) = -t^2\sigma^2/2.$$

For any $\rho > 1$, we have $-\rho x \leqslant \log(1-x) \leqslant -x$ for $x > 0$ small enough. Since $\lim_n \max_m \sigma_{n,m}^2 = 0$, we have, for n large enough,

$$-\rho t^2\sigma_{n,m}^2/2 \leqslant \log g_{n,m}(t) \leqslant -t^2\sigma_{n,m}^2/2.$$

Summing over m and let $n \to \infty$, we have

$$-\rho t^2\sigma^2/2 \leqslant \liminf_n \sum_{m=1}^n \log g_{n,m}(t) \leqslant \limsup_n \sum_{m=1}^n \log g_{n,m}(t) \leqslant -t^2\sigma^2/2.$$

Let $\rho \to 1+$, we obtain the conclusion. □

Example 4.4.2. *Suppose $\{Y_n\}$ are independent Bernoulli random variables with $\mathbb{P}[Y_n = 1] = 1/n$ and $\mathbb{P}[Y_n = 0] = 1 - 1/n$. Define $S_n = \sum_{j=1}^n Y_j$. Then we have*

$$\frac{S_n - \log n}{\sqrt{\log n}} \Longrightarrow \mathcal{N}(0,1).$$

Proof. Since $\mathbb{E}[Y_n] = 1/n$ and $\mathrm{var}(Y_n) = 1/n - 1/n^2$, we find $\mathbb{E}[S_n] \sim \log n$ and $\mathrm{var}(S_n) \sim \log n$. For $1 \leqslant m \leqslant n$, define

$$X_{n,m} = \frac{Y_m - 1/m}{\sqrt{\log n}}.$$

Then we have $\mathbb{E}[X_{n,m}] = 0$, and $\sum_{m=1}^n \mathbb{E}[X_{n,m}^2] \to 1$, and for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{m=1}^n \mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right] = 0,$$

because $|X_{n,m}| \leqslant 1/\sqrt{\log n}$. Then by Theorem 4.4.1, we obtain the conclusion. $\square$

Example 4.4.3 (Lyapunov's Theorem)**.** *Suppose $\{Y_n\}$ are independent and define $S_n = \sum_{j=1}^n Y_j$. Define $\alpha_n = \sqrt{\mathrm{var}(S_n)}$. If there is $\delta > 0$ such that*

$$\lim_{n\to\infty} \alpha_n^{-2-\delta} \sum_{m=1}^n \mathbb{E}[|Y_m - \mathbb{E}[Y_m]|^{2+\delta}] = 0,$$

then we have

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{var}(S_n)}} \Longrightarrow \mathcal{N}(0,1).$$

Proof. For $1 \leqslant m \leqslant n$, set

$$X_{n,m} = \frac{Y_m - \mathbb{E}[Y_m]}{\alpha_n}.$$

Then we have $\mathbb{E}[X_{n,m}] = 0$, and $\sum_{m=1}^n \mathbb{E}[X_{n,m}^2] = 1$. For any $\epsilon > 0$, we have

$$\mathbb{E}\left[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}\right] = \frac{1}{\alpha_n^2}\mathbb{E}\left[(Y_m - \mathbb{E}[Y_m])^2 \mathbb{1}_{\{|Y_m - \mathbb{E}[Y_m]|>\epsilon\alpha_n\}}\right] \leqslant \frac{1}{\epsilon^\delta \alpha_n^{2+\delta}}\mathbb{E}[|Y_m - \mathbb{E}[Y_m]|^{2+\delta}].$$

Thus we can apply Theorem 4.4.1 to $\{X_{n,m}\}$ and we obtain the conclusion. $\square$

Example 4.4.4. *Suppose $\{Y_n\}$ are i.i.d. with the common law given by $\mathbb{P}[Z > x] = \mathbb{P}[Z < -x]$ and $\mathbb{P}[|Z| > x] = x^{-2}$ for $x \geqslant 1$. Define $S_n = \sum_{j=1}^n Y_j$. Then we have*

$$\frac{S_n}{\sqrt{n \log n}} \Longrightarrow \mathcal{N}(0,1).$$

Proof. Note that $\mathbb{E}[Z^2] = \infty$. But this example tells us that, when we renormalize correctly, we still have the convergence of S_n to the normal distribution.

We truncate the random variables at the level c_n whose value is to be decided later: define for $1 \leqslant m \leqslant n$,

$$Y_{n,m} = Y_m \mathbb{1}_{\{|Y_m| \leqslant c_n\}}, \quad \tilde{S}_n = \sum_{m=1}^n Y_{n,m}.$$

A good choice of c_n should satisfy $\mathbb{P}[S_n \neq \tilde{S}_n] \to 0$ as $n \to \infty$ and it is close to the lowest possible level. Since

$$\mathbb{P}[\tilde{S}_n \neq S_n] \leqslant n\mathbb{P}[|Z| > c_n],$$

we need to choose c_n so that $n/c_n^2 \to 0$ as $n \to \infty$.

Consider $\{Y_{n,m}\}$, we have $\mathbb{E}[Y_{n,m}] = 0$, and $\sum_{m=1}^n \mathbb{E}[Y_{n,m}^2] = 2n \log c_n$. This indicates that we should define

$$X_{n,m} = \frac{Y_{n,m}}{\sqrt{2n \log c_n}}, \quad \tilde{S}_n^\# = \sum_{m=1}^n X_{n,m}.$$

Then we have $\mathbb{E}[X_{n,m}] = 0$ and $\sum_{m=1}^n \mathbb{E}[X_{n,m}^2] \to 1$. For any $\epsilon > 0$, we need

$$\sum_{m=1}^n \mathbb{E}[X_{n,m}^2 \mathbb{1}_{\{|X_{n,m}|>\epsilon\}}] = \frac{1}{\log c_n}\left(\log c_n - \log(\epsilon\sqrt{2n \log c_n})\right) \to 0.$$

Combining all the requirements, we need to choose $c_n \uparrow \infty$ such that

$$\frac{n}{c_n^2} \to 0, \quad \frac{1}{\log c_n}\left(\log c_n - \frac{1}{2}(\log n + \log\log c_n)\right) \to 0.$$

We can choose $c_n = \sqrt{n \log n}$. Then we can apply Theorem 4.4.1 to $\tilde{S}_n^{\#}$ and we have

$$\frac{\tilde{S}_n}{\sqrt{n \log n}} \Longrightarrow \mathcal{N}(0,1).$$

Finally, for any $x \in \mathbb{R}$, we have

$$\left|\mathbb{P}\left[\frac{S_n}{\sqrt{n\log n}} \leqslant x\right] - \mathbb{P}\left[\frac{\tilde{S}_n}{\sqrt{n\log n}} \leqslant x\right]\right| \leqslant \mathbb{P}[S_n \neq \tilde{S}_n] \to 0.$$

Thus

$$\frac{S_n}{\sqrt{n \log n}} \Longrightarrow \mathcal{N}(0,1).$$

□

4.5 Poisson convergence

The goal of this section is the following "weak law of small numbers" or the "law of rare events".

Theorem 4.5.1. *For each n, let $\{X_{n,m} : 1 \leqslant m \leqslant n\}$ be independent variables with $\mathbb{P}[X_{n,m} = 1] = p_{n,m}$ and $\mathbb{P}[X_{n,m} = 0] = 1 - p_{n,m}$. Suppose*

$$\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty), \quad \text{and} \quad \max_{1\leqslant m\leqslant n} p_{n,m} \to 0.$$

Then

$$S_n^{\#} = \sum_{m=1}^{n} X_{n,m} \Longrightarrow \text{Poisson}(\lambda).$$

Proof. Let us calculate the characteristic functions of $X_{n,m}$ and S_n:

$$\mathbb{E}[\exp(itX_{n,m})] = 1 + p_{n,m}(e^{it} - 1), \quad \mathbb{E}[\exp(itS_n)] = \prod_{m=1}^{n}\left(1 + p_{n,m}(e^{it}-1)\right).$$

Then the goal is the following:

$$\sum_{m=1}^{n} \log\left(1 + p_{n,m}(e^{it}-1)\right) \to \lambda(e^{it}-1).$$

Since $\sum_{m=1}^{n} p_{n,m} \to \lambda$, it is sufficient to show

$$\left|\sum_{m=1}^{n} \log\left(1 + p_{n,m}(e^{it}-1)\right) - \sum_{m=1}^{n} p_{n,m}(e^{it}-1)\right| \to 0.$$

For $z \in \mathbb{C}$ such that $|z| \leqslant 1/2$, we have $|\log(1+z) - z| \leqslant C|z|^2$. Thus

$$\begin{aligned}\left|\sum_{m=1}^{n} \log\left(1 + p_{n,m}(e^{it}-1)\right) - \sum_{m=1}^{n} p_{n,m}(e^{it}-1)\right| &\leqslant C\sum_{m=1}^{n} 4p_{n,m}^2 \\ &\leqslant 4C \max_{1\leqslant m\leqslant n} p_{n,m} \sum_{m=1}^{n} p_{n,m} \to 0.\end{aligned}$$

□

Example 4.5.2. *In a class with 400 students, the number of students who have their birthday on the day of the final exam has approximately a Poisson distribution with mean* $400/365 = 1.096$*.*

We will give a second proof which involves the notion of total variation distance, a concept which is of great interest in "Markov chain".

Definition 4.5.3. *Suppose S is a countable set. The total variation distance between two probability measures on S is given by*

$$\|\mu - \nu\| := \frac{1}{2} \sum_{z \in S} |\mu(z) - \nu(z)|.$$

Lemma 4.5.4. *We have*

$$\|\mu - \nu\| = \sup_{A \subset S} |\mu[A] - \nu[A]|.$$

Proof. On the one hand, for any $A \subset S$, we have

$$2\|\mu - \nu\| = \sum_{z \in S} |\mu[z] - \nu[z]| \geqslant |\mu[A] - \nu[A]| + |\mu[A^c] - \nu[A^c]| = 2|\mu[A] - \nu[A]|.$$

On the other hand, set $A_0 = \{z : \mu[z] \geqslant \nu[z]\}$. Then we have

$$\mu[A_0] - \nu[A_0] = \sum_{z \in A_0} (\mu[z] - \nu[z]), \quad \nu[A_0^c] - \mu[A_0^c] = \sum_{z \in A_0^c} (\nu[z] - \mu[z]).$$

Thus

$$2(\mu[A_0] - \nu[A_0]) = \sum_{z} |\mu[z] - \nu[z]| = 2\|\mu - \nu\|.$$

□

Lemma 4.5.5. *Denote by $\mathcal{P}_{\mathbb{Z}}$ the collection of probability measures on $\mathbb{Z}$.*

(1) The total variation distance $\|\mu - \nu\|$ defines a metric on $\mathcal{P}_{\mathbb{Z}}$.

(2) $\|\mu_n - \mu\| \to 0$ if and only if $\mu_n[x] \to \mu[x]$ for all $x \in \mathbb{Z}$. In particular, $\|\mu_n - \mu\| \to 0$ if and only if $\mu_n \Longrightarrow \mu$.

Proof. To show the total variation distance is a metric, we only need to check the triangle inequality: suppose $\mu, \nu, \eta \in \mathcal{P}_{\mathbb{Z}}$, for any $A \subset S$, we have

$$|\mu[A] - \nu[A]| \leqslant |\mu[A] - \eta[A]| + |\nu[A] - \eta[A]| \leqslant \|\mu - \eta\| + \|\nu - \eta\|.$$

Thus

$$\|\mu - \nu\| \leqslant \|\mu - \eta\| + \|\nu - \eta\|.$$

□

Lemma 4.5.6.

$$\|\mu - \nu\| = \inf\{\mathbb{P}[X \neq Y] : (X, Y) \textit{ is a coupling of } \mu, \nu\}.$$

We call (X, Y) the optimal coupling if $\mathbb{P}[X \neq Y] = \|\mu - \nu\|$.

Proof. There are two steps: first, show that $\|\mu - \nu\| \leqslant \mathbb{P}[X \neq Y]$ for any coupling (X, Y); next, construct a coupling (X, Y) such that $\|\mu - \nu\| = \mathbb{P}[X \neq Y]$.

First step. Suppose that (X, Y) is any coupling of μ and ν and that A is any subset of S. Note that

$$\begin{aligned}\mu[A] - \nu[A] &= \mathbb{P}[X \in A] - \mathbb{P}[Y \in A] \\ &= \mathbb{P}[X \in A, Y \notin A] - \mathbb{P}[Y \in A, X \notin A] \\ &\leqslant \mathbb{P}[X \in A, Y \notin A] \leqslant \mathbb{P}[X \neq Y].\end{aligned}$$

Thus $\|\mu - \nu\| \leqslant \mathbb{P}[X \neq Y]$.

Second step. We need to construct a coupling (X, Y) so that $X = Y$ as often as possible. Define three probability measures

$$\gamma_1(x) = \frac{\mu(x) - \nu(x)}{\|\mu - \nu\|} \mathbb{1}_{\{\mu(x) > \nu(x)\}}, \quad \gamma_2(x) = \frac{\nu(x) - \mu(x)}{\|\mu - \nu\|} \mathbb{1}_{\{\nu(x) > \mu(x)\}}, \quad \gamma_3(x) = \frac{\mu(x) \wedge \nu(x)}{p},$$

where $p = 1 - \|\mu - \nu\|$. We construct the coupling in the following way: Flip a coin with probability of heads equal to p.

- If head, choose a value Z according to γ_3, and set $X = Y = Z$.
- If tail, choose X according to γ_1, and independently choose Y according to γ_2. Since γ_1 and γ_2 are singular, we have $X \neq Y a.s.$ in this case.

Now we have a pair (X, Y), and let us check the marginal laws.

- The marginal law of X: $p\gamma_3 + (1 - p)\gamma_1 = \mu$.
- The marginal law of Y: $p\gamma_3 + (1 - p)\gamma_2 = \nu$.

Moreover, we have that

$$\mathbb{P}[X \neq Y] = \mathbb{P}[\text{tail}, X \neq Y] = 1 - p = \|\mu - \nu\|.$$

□

Lemma 4.5.7. *Suppose $\mu_1, \mu_2, \nu_1, \nu_2$ are probability measures on $\mathbb{Z}$, then*

$$\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| \leqslant \|\mu_1 - \nu_1\| + \|\mu_2 - \nu_2\|.$$

Proof. We have

$$\begin{aligned}2\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| &= \sum_{x,y} |\mu_1(x)\mu_2(y) - \nu_1(x)\nu_2(y)| \\ &\leqslant \sum_{x,y} |\mu_1(x)\mu_2(y) - \nu_1(x)\mu_2(y)| + \sum_{x,y} |\nu_1(x)\mu_2(y) - \nu_1(x)\mu_2(y)| \\ &= \sum_{x,y} |\mu_1(x) - \nu_1(x)|\mu_2(y) + \sum_{x,y} \nu_1(x)|\mu_2(y) - \mu_2(y)| \\ &= 2\|\mu_1 - \nu_1\| + 2\|\mu_2 - \nu_2\|.\end{aligned}$$

□

Lemma 4.5.8. *Suppose $\mu_1, \mu_2, \nu_1, \nu_2$ are probability measures on $\mathbb{Z}$, then*

$$\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| \leqslant \|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|.$$

Proof. We have

$$\begin{aligned} 2\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| &\leqslant \sum_{x,y} |\mu_1(x-y)\mu_2(y) - \nu_1(x-y)\nu_2(y)| \\ &= 2\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|. \end{aligned}$$

□

Second Proof of Theorem 4.5.1. Denote by $\mu_{n,m}$ the law of $X_{n,m}$ and by $\nu_{n,m}$ the law of Poisson distribution with mean $p_{n,m}$. Denote by μ_n the law of $S_n^{\#}$, by ν_n the law of Poisson distribution with mean $\sum_{m=1}^n p_{n,m}$, and by ν the law of Poisson distribution with mean λ. Then we have

$$\|\mu_n - \nu_n\| \leqslant \sum_{m=1}^n \|\mu_{n,m} - \nu_{n,m}\|.$$

By direct calculation, we have

$$\|\mu_{n,m} - \nu_{n,m}\| = p_{n,m}(1 - e^{-p_{n,m}}) \leqslant p_{n,m}^2.$$

Thus

$$\|\mu_n - \nu_n\| \leqslant \sum_{m=1}^n p_{n,m}^2 \to 0.$$

□

4.6 Representation theorems

A characteristic function is the Fourier transform of a probability measure. In this section, we will give a characterization for the characteristic function.

Definition 4.6.1. *A complex-valued function f defined on $\mathbb{R}$ is called positive definite if for any finite set of real numbers t_j and complex numbers z_j, we have*

$$\sum_{j,k=1}^n f(t_j - t_k) z_j \overline{z}_k \geqslant 0.$$

The goal of this section is the following theorem.

Theorem 4.6.2. *f is a characteristic function if and only if it is positive definite and continuous at zero with $f(0) = 1$.*

To prove Theorem 4.6.2, we first derive some properties of positive definite functions.

Lemma 4.6.3. *If f is positive definite, then for each $t \in \mathbb{R}$:*

$$f(-t) = \overline{f(t)}, \quad |f(t)| \leqslant f(0). \tag{4.6.1}$$

If we assume further that f is continuous at $t = 0$, then it is uniformly continuous in $\mathbb{R}$; furthermore, for every continuous complex-valued function ξ on $\mathbb{R}$ and every $T > 0$, we have

$$\int_0^T \int_0^T f(s-t)\xi(s)\overline{\xi(t)} ds dt \geqslant 0. \tag{4.6.2}$$

Proof. In Definition 4.6.1, taking $n = 1$, $t_1 = 0$ and $z_1 = 1$, we have $f(0) \geqslant 0$.

Taking $n = 2$, $t_1 = 0, t_2 = t$, $z_1 = z$ and $z_2 = w$, we have

$$f(0)|z|^2 + f(0)|w|^2 + f(t)\overline{z}w + f(-t)z\overline{w} \geqslant 0. \tag{4.6.3}$$

Taking $z = w = 1$ in (4.6.3), we have

$$2f(0) + f(t) + f(-t) \geqslant 0.$$

Taking $z = 1$ and $w = i$ in (4.6.3), we have

$$2f(0) + f(t)i - f(-t)i \geqslant 0.$$

These imply that $f(t) + f(-t)$ is real and $f(t) - f(-t)$ is pure imaginary. Thus $\overline{f(t)} = f(-t)$. We can rewrite (4.6.3) as follows:

$$\begin{pmatrix} z & w \end{pmatrix} \begin{pmatrix} f(0) & f(-t) \\ f(t) & f(0) \end{pmatrix} \begin{pmatrix} \overline{z} \\ \overline{w} \end{pmatrix}.$$

The 2×2 matrix is self adjoint[2]. Thus the determinant is positive: $f(0)^2 \geqslant |f(t)|^2$ which completes the proof of (4.6.1).

Next, we assume f is continuous at zero with $f(0) = 1$. Taking $n = 3$ and $t_1 = 0, t_2 = t, t_3 = t + h$, we have

$$\begin{pmatrix} z_1 & z_2 & z_3 \end{pmatrix} \begin{pmatrix} f(0) & f(-t) & f(-t-h) \\ f(t) & f(0) & f(-h) \\ f(t+h) & f(h) & f(0) \end{pmatrix} \begin{pmatrix} \overline{z_1} \\ \overline{z_2} \\ \overline{z_3} \end{pmatrix}.$$

The determinant of the 3×3 matrix has to be positive:

$$1 - |f(h)|^2 - |f(t)|^2 - |f(t+h)|^2 + 2\Re f(t)f(h)\overline{f(t+h)} \geqslant 0.$$

Therefore,

$$\begin{aligned} |f(t+h) - f(t)|^2 &\leqslant |f(t)|^2 + |f(t+h)|^2 - 2\Re f(t)\overline{f(t+h)} \\ &\leqslant 1 - |f(h)|^2 + 2\Re f(t)(f(h) - 1)\overline{f(t+h)} \\ &\leqslant 1 - |f(h)|^2 + 2|1 - f(h)| \leqslant 4|1 - f(h)|. \end{aligned}$$

This gives the uniform continuity. Finally, since the integrand in (4.6.2) is continuous, the integral is the limit of Riemann sums which is positive. ☐

Proof of Theorem 4.6.2. First, we assume that f is the characteristic function of the probability measure μ, then we have

$$\begin{aligned} \sum_{j,k=1}^{n} f(t_j - t_k)z_j\overline{z}_k &= \int \sum_{j,k=1}^{n} e^{ix(t_j - t_k)} z_j\overline{z}_k \mu[dx] \\ &= \int \left| \sum_{j=1}^{n} e^{ixt_j} z_j \right|^2 \mu[dx] \geqslant 0. \end{aligned}$$

Next, we assume that f is positive definite and is continuous at zero with $f(0) = 1$. Taking $\xi(t) = e^{-itx}$ in (4.6.2), we have

$$p_T(x) := \frac{1}{2\pi T} \int_0^T \int_0^T f(s-t)e^{-i(s-t)x} ds dt \geqslant 0.$$

[2]Any $n \times n$ Hermitian matrix H can be diagonalized by a unitary matrix. All eigenvalues of H are real, and H has n linearly independent eigenvectors.

By a change of variables, we have

$$p_T(x) = \frac{1}{2\pi}\int_{-T}^{T} f_T(u)e^{-iux}du, \quad \text{where } f_T(u) = \left(1 - \frac{|u|}{T}\right) f(u)\, \mathbb{1}_{\{|u|\leqslant T\}}\,.$$

First, we argue that p_T is a probability density function. To this end, we calculate

$$\begin{aligned}
\frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} p_T(x)dx &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} e^{-iux}dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{1}{\alpha}\int_0^\alpha d\beta \frac{2\sin(\beta u)}{u} \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{(1-\cos(\alpha u))}{\alpha u^2} = \frac{1}{\pi}\int_{-\infty}^{\infty} f_T\left(\frac{t}{\alpha}\right)\frac{(1-\cos t)}{t^2}dt.
\end{aligned}$$

Note that $|f_T| \leqslant |f| \leqslant 1$, dominated convergence theorem shows that

$$\lim_{\alpha\to\infty}\frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} p_T(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(1-\cos t)}{t^2}dt = 1.$$

The last equation is due to the following fact:

$$\int_0^\infty \frac{1-\cos(\alpha x)}{x^2}dx = \frac{\pi}{2}|\alpha|. \tag{4.6.4}$$

Since $p_T(x) \geqslant 0$ and $\int_{-\beta}^{\beta} p_T(x)dx$ is increasing in β, we have

$$\int_{-\infty}^{\infty} p_T(x)dx = \lim_{\beta\to\infty}\int_{-\beta}^{\beta} p_T(x)dx = 1.$$

Thus $p_T(x)$ is a probability density function.

Second, we argue that f_T is the characteristic function of p_T. To this end, we calculate the following: for any $t \in \mathbb{R}$,

$$\begin{aligned}
\frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} e^{itx}p_T(x)dx &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} e^{-i(u-t)x}dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{1}{\alpha}\int_0^\alpha d\beta \frac{2\sin(\beta(u-t))}{u-t} \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} f_T(u)du \frac{(1-\cos(\alpha(u-t)))}{\alpha(u-t)^2} = \frac{1}{\pi}\int_{-\infty}^{\infty} f_T\left(t + \frac{v}{\alpha}\right)\frac{(1-\cos v)}{v^2}dv.
\end{aligned}$$

Dominated convergence theorem gives

$$\lim_{\alpha\to\infty}\frac{1}{\alpha}\int_0^\alpha d\beta \int_{-\beta}^{\beta} e^{itx}p_T(x)dx = f_T(t).$$

Since $\int p_T(x)dx = 1$, the following limit exists and hence equals $f_T(t)$:

$$\int_{-\infty}^{\infty} e^{itx}p_T(x)dx = \lim_{\beta\to\infty}\int_{-\beta}^{\beta} e^{itx}p_T(x)dx = f_T(t).$$

This shows that f_T is the characteristic function of $p_T(x)dx$.

Finally, we see that $f_T \to f$ as $T \to \infty$ and f is continuous at zero with $f(0) = 1$ by the hypothesis, combining with Theorem 4.3.1, we see that f is the characteristic function of a probability measure. $\square$

Example 4.6.4. *If f is a characteristic function, then so is $e^{\lambda(f-1)}$ for $\lambda \geqslant 0$.*

Proof. For each $\lambda \geqslant 0$, as soon as $n \geqslant \lambda$, the function

$$1 + \frac{\lambda(f-1)}{n}$$

is also a characteristic function, hence so is its nth power. As $n \to \infty$,

$$\left(1 + \frac{\lambda(f-1)}{n}\right)^n \to e^{\lambda(f-1)},$$

and the limit is clearly continuous at zero with value one. Hence it is a characteristic function. ☐

In this example, if we take $f(t) = e^{it}$ which is the characteristic function of the dirac mass at one, we have

$$e^{\lambda(e^{it}-1)} = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} e^{itn}$$

which is the characteristic function of the Poisson distribution.

4.7 Exercises

Exercise 4.7.1. *If the sequence of characteristic functions $\{f_n\}$ converges uniformly in a neighborhood of the origin, then $\{f_n\}$ is equicontinuous, and there exists a subsequence that converges to a characteristic function f.*

Exercise 4.7.2. *Let $\{X_n\}$ be i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathrm{var}(X_1) = 1$. Set $S_n = \sum_{j=1}^n X_j$. Show that*

$$\limsup_n \frac{S_n}{\sqrt{n}} = \infty, \quad a.s.; \quad \textit{and} \quad \frac{S_n}{\sqrt{n}} \Longrightarrow \mathcal{N}(0,1).$$

Exercise 4.7.3. *Let $\{X_n\}$ be i.i.d. with mean zero and variance $\sigma^2 \in (0, \infty)$, then*

$$\lim_n \mathbb{E}\left[\frac{|S_n|}{\sqrt{n}}\right] = 2 \lim_n \mathbb{E}\left[\frac{S_n^+}{\sqrt{n}}\right] = \sqrt{\frac{2}{\pi}}\sigma.$$

Exercise 4.7.4. *Suppose $\{X_n\}$ are i.i.d. with mean zero and variance one. Prove that both*

$$\frac{\sum_{j=1}^n X_j}{\sqrt{\sum_{j=1}^n X_j^2}}, \quad \textit{and} \quad \frac{\sqrt{n}\sum_{j=1}^n X_j}{\sum_{j=1}^n X_j^2}$$

converges in distribution to $\mathcal{N}(0,1)$.

Exercise 4.7.5. *Let $\{X_n\}$ be i.i.d. with mean zero and set $S_n = \sum_{j=1}^n X_j$. Assume that $S_n/\sqrt{n}$ converges in distribution, prove that $\mathbb{E}[X_1^2] < \infty$.*

Exercise 4.7.6. *Use Theorem 4.4.1 to give a second proof of three series theorem.*

Exercise 4.7.7. *Suppose $\{X_n\}$ are independent and*

$$\mathbb{P}[X_n = -n] = \mathbb{P}[X_n = n] = \frac{1}{2n^2}, \quad \mathbb{P}[X_n = -1] = \mathbb{P}[X_n = 1] = \frac{1}{2} - \frac{1}{2n^2}.$$

Define $S_n = \sum_{j=1}^n X_j$. Show that

$$\frac{\mathrm{var}(S_n)}{n} \to 2, \quad \frac{S_n}{\sqrt{n}} \Longrightarrow \mathcal{N}(0,1).$$

Figure out why the conclusion violates Theorem 4.4.1.

Exercise 4.7.8. *Suppose $f(t,u)$ is a function on $\mathbb{R}^2$ such that for each u, the function $f(\cdot,u)$ is a characteristic function; and for each t, the function $f(t,\cdot)$ is continuous. Then, for any probability measure ν, the following function is also a characteristic function:*

$$\exp\left(\int_{-\infty}^{\infty}(f(t,u)-1)\nu[du]\right).$$

Exercise 4.7.9 (YCMC2013)**.** *Let $\{X_n\}$ be i.i.d. with $\mathbb{E}[X]=\mu$, $\sigma^2 := \mathrm{var}(X) < \infty$ and characteristic function $\varphi_X(t)$. Let N be a non-negative integer-valued random variable with $\mathbb{E}[N]=\nu$, $\eta^2=\mathrm{var}(N)$and characteristic function $\varphi_N(t)$. Suppose $\{X_n\}$ and N are independent. Let $Y=\sum_{k=1}^N X_k$.*

(1) What is the characteristic function of Y?

(2) Evaluate the variance of Y.

Exercise 4.7.10 (YCMC2013)**.** *Suppose that X and Y are two independent random variables and X has a density. Does $X+Y$ also have a density?*

Exercise 4.7.11 (YCMC2013)**.** *Suppose that N is a random variable such that $\mathbb{P}[N=i]=1/3$ for $i\in\{1,2,3\}$ and X_1,X_2,X_3 are i.i.d with standard normal distribution $\mathcal{N}(0,1)$. Is $X=\sum_{i=1}^N X_i$ also normal?*

Exercise 4.7.12 (YCMC2013)**.** *Suppose that X and Y are independent and the law of the two-dimensional random vector $Z:=(X,Y)$ is rotationally invariant: for any orthogonal matrix O (i.e. $O^tO=I_2$), OZ has the same law as Z as a random vector. Show that both X and Y have the law of a centered normal distribution $\mathcal{N}(0,\sigma^2)$ for some $\sigma^2>0$.*

Exercise 4.7.13 (YCMC2013)**.** *Show that $f_1(t)=(\cos t)^2$ is a characteristic function and $f_2(t)=|\cos t|$ is not a characteristic function.*

Exercise 4.7.14 (YCMC2015)**.** *Suppose that X and Z are jointly normal with mean zero and standard deviation one. Show that, for a strictly monotone function $f(\cdot)$,*

$$\mathrm{cov}(X,Z)=0, \quad \textit{if and only if} \quad \mathrm{cov}(X,f(Z))=0,$$

provided the latter covariance exists. Hint: Z can be expressed as $Z=\rho X+\sqrt{1-\rho^2}Y$ where X and Y are i.i.d. with $\mathcal{N}(0,1)$.

Exercise 4.7.15 (YCMC2016)**.** *Let X,Y be two real-valued random variables such that $X-Y$ and X are independent, and that $X-Y$ are Y are independent. Show that $X-Y$ is almost surely constant.*

Exercise 4.7.16 (YCMC2016)**.** *Let X_λ be a Poisson random variable with parameter λ. What is the limiting distribution of $\sqrt{X_\lambda}-\sqrt{\lambda}$ as $\lambda\to\infty$?*

5 Martingales

5.1 Conditional expectation

Theorem 5.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that X is a random variable in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathcal{A}$ is a sub-σ-field of $\mathcal{F}$. Then there exists a random variable Y such that*

(1) Y is $\mathcal{A}$-measurable;

(2) $\mathbb{E}[|Y|] < \infty$;

(3) for any $A \in \mathcal{A}$, we have $\mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A]$.

Moreover, if $\tilde{Y}$ is another random variable satisfying the above three properties, then $\tilde{Y} = Y$ a.s.

Definition 5.1.2. *A random variable Y with the three properties in Theorem 5.1.1 is called the conditional expectation of X given $\mathcal{A}$, denoted by $\mathbb{E}[X \,|\, \mathcal{A}]$.*

We can easily check that the conditional expectation has the following basic properties. Suppose that $X, X_1, X_2 \in L^1$ and $\mathcal{A}$ is a sub-σ-field.

- If $\mathcal{A} = \{\varnothing, \Omega\}$, then $\mathbb{E}[X \,|\, \mathcal{A}] = \mathbb{E}[X]$.
- If X is $\mathcal{A}$-measurable, then $\mathbb{E}[X \,|\, \mathcal{A}] = X$.
- If $Y = \mathbb{E}[X \,|\, \mathcal{A}]$, then $\mathbb{E}[Y] = \mathbb{E}[X]$.
- (Linearity). $\mathbb{E}[a_1 X_1 + a_2 X_2 \,|\, \mathcal{A}] = a_1 \mathbb{E}[X_1 \,|\, \mathcal{A}] + a_2 \mathbb{E}[X_2 \,|\, \mathcal{A}]$ for constants a_1, a_2.
- (Positivity). If $X \geqslant 0$, then $\mathbb{E}[X \,|\, \mathcal{A}] \geqslant 0$.

Proof of Positivity. Denote $\mathbb{E}[X \,|\, \mathcal{A}]$ by Y. By the definition, we know that

$$\mathbb{E}[Y \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A] \geqslant 0, \quad \forall A \in \mathcal{A}. \tag{5.1.1}$$

For $n \geqslant 1$, define $A_n = \{Y \leqslant -1/n\}$. Since Y is $\mathcal{A}$-measurable, we know that $A_n \in \mathcal{A}$. Thus, by (5.1.1), we have

$$\frac{-1}{n} \mathbb{P}[A_n] \geqslant \mathbb{E}[Y \mathbb{1}_{A_n}] \geqslant 0.$$

Thus $\mathbb{P}[A_n] = 0$, and $\mathbb{P}[\cup_n A_n] = 0$. Therefore, $\mathbb{P}[Y < 0] = 0$. ☐

Proof of Theorem 5.1.1. First, we show the existence of $\mathbb{E}[X \,|\, \mathcal{A}]$ for $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Consider the subspace $L^2(\Omega, \mathcal{A}, \mathbb{P})$ of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, it is complete under L^2-norm. Thus there exists $Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ such that

$$\mathbb{E}\left[(X - Y)^2\right] = \inf\{\mathbb{E}\left[(X - Z)^2\right] : Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})\}.$$

Moreover, we know that, for any $Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$,

$$\mathbb{E}[(X - Y)Z] = 0.$$

Therefore, for any $A \in \mathcal{A}$, we have

$$\mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A].$$

Thus, we can choose $\mathbb{E}[X \,|\, \mathcal{A}] = Y$.

Second, we show the existence of $\mathbb{E}[X \,|\, \mathcal{A}]$ for $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Since X can be written as the difference of two non-negative L^1 random variables, it is sufficient to show the conclusion for non-negative $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. There exists a sequence of bounded variables $\{X_n\}$ such that $0 \leqslant X_n \uparrow X$ a.s. Since $X_n \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, there exists some $Y_n = \mathbb{E}[X_n \,|\, \mathcal{A}]$. By "Positivity", we have that $0 \leqslant Y_n \uparrow$ a.s. Define $Y = \lim Y_n$. We will check that Y satisfies all the requirements.

- Since $\{Y_n\}$ are $\mathcal{A}$-measurable, the limit Y is also $\mathcal{A}$-measurable.
- For any $A \in \mathcal{A}$, by Monotone Convergence Theorem, we have

$$\mathbb{E}[Y \mathbb{1}_A] = \lim_n \mathbb{E}[Y_n \mathbb{1}_A] = \lim_n \mathbb{E}[X_n \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A].$$

  In particular $\mathbb{E}[Y] = \mathbb{E}[X]$ and $Y \in L^1$.

Thus Y satisfies all the requirements and we can choose $\mathbb{E}[X \,|\, \mathcal{A}] = Y$.

Finally, we show the uniqueness. If $\tilde{Y}$ is another random variable satisfying the three properties, then by a similar proof for "Positivity", we have that

$$\mathbb{P}[Y > \tilde{Y}] = 0, \quad \mathbb{P}[Y < \tilde{Y}] = 0.$$

□

Proposition 5.1.3. *Suppose that X and $\{X_n\}$ are random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathcal{A}$ is a sub-σ-field. We have the following properties.*

(1) (Monotone Convergence Theorem). If $0 \leqslant X_n \uparrow X$ a.s., then $\mathbb{E}[X_n \,|\, \mathcal{A}] \uparrow \mathbb{E}[X \,|\, \mathcal{A}]$ a.s.

(2) (Fatou's Lemma). If $X_n \geqslant 0$ for all n, then $\mathbb{E}[\liminf_n X_n \,|\, \mathcal{A}] \leqslant \liminf_n \mathbb{E}[X_n \,|\, \mathcal{A}]$ a.s.

(3) (Dominated Convergence Theorem). If $X_n \to X$ a.s. and there is $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $|X_n| \leqslant Z$ for all n, then $\mathbb{E}[X_n \,|\, \mathcal{A}] \to \mathbb{E}[X \,|\, \mathcal{A}]$ a.s.

(4) (Jensen's Inequality). If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[|\varphi(X)|] < \infty$, then $\mathbb{E}[\varphi(X) \,|\, \mathcal{A}] \geqslant \varphi(\mathbb{E}[X \,|\, \mathcal{A}])$.

Proof of Item (4). For convex function φ, there exists a sequence of pairs of reals $\{(a_n, b_n), n \geqslant 0\}$ such that

$$\varphi(x) = \sup_n (a_n x + b_n), \quad \forall x \in \mathbb{R}.$$

Since $\varphi(X) \geqslant a_n X + b_n$ a.s., we have

$$\mathbb{E}[\varphi(X) \,|\, \mathcal{A}] \geqslant a_n \mathbb{E}[X \,|\, \mathcal{A}] + b_n \quad a.s.$$

Therefore,

$$\mathbb{E}[\varphi(X) \,|\, \mathcal{A}] \geqslant \sup_n (a_n \mathbb{E}[X \,|\, \mathcal{A}] + b_n) = \varphi(\mathbb{E}[X \,|\, \mathcal{A}]) \quad a.s.$$

□

Proposition 5.1.4. *Suppose that X and $\{X_n\}$ are random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathcal{A}$ is a sub-σ-field. We have the following properties.*

(1) (Tower property). Suppose that $\mathcal{B}$ is a sub-σ-field of $\mathcal{A}$. Then

$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{A}] \,|\, \mathcal{B}] = \mathbb{E}[X \,|\, \mathcal{B}].$$

(2) ("Taking out what is known"). Suppose that Z is $\mathcal{A}$-measurable and $\mathbb{E}[|XZ|] < \infty$, then

$$\mathbb{E}[ZX \,|\, \mathcal{A}] = Z\mathbb{E}[X \,|\, \mathcal{A}].$$

(3) (Independence). If $\mathcal{B}$ is independent of $\sigma(X, \mathcal{A})$, then

$$\mathbb{E}[X \,|\, \sigma(\mathcal{A}, \mathcal{B})] = \mathbb{E}[X \,|\, \mathcal{A}] \quad a.s.$$

In particular, if X is independent of $\mathcal{B}$, then $\mathbb{E}[X \,|\, \mathcal{B}] = \mathbb{E}[X]$ a.s.

Proof of Item (2). The relation is true when Z is an indicator function (by the definition of the conditional expectation). Hence, by Linearity, it is true for linear combinations of the indicator functions. Finally, by Dominated Convergence Theorem, it is true for general Z with $\mathbb{E}[|XZ|] < \infty$. □

Proof of Item (3). Denote $\mathbb{E}[X \mid \mathcal{A}]$ by Y. We need to show that

$$\mathbb{E}[X \mathbb{1}_C] = \mathbb{E}[Y \mathbb{1}_C], \quad \forall C \in \sigma(\mathcal{A}, \mathcal{B}). \tag{5.1.2}$$

We first argue that (5.1.2) holds for $C = A \cap B$ where $A \in \mathcal{A}$ and $B \in \mathcal{B}$. In other words, we first show that $\mathbb{E}[X \mathbb{1}_{A \cap B}] = \mathbb{E}[Y \mathbb{1}_{A \cap B}]$. We have the following observations:

- Since $\mathcal{B}$ is independent of $\sigma(X, \mathcal{A})$, we have $\mathbb{E}[X \mathbb{1}_A \mathbb{1}_B] = \mathbb{E}[X \mathbb{1}_A]\mathbb{P}[B]$.
- Since Y is $\mathcal{A}$-measurable which is independent of $\mathcal{B}$, we have $\mathbb{E}[Y \mathbb{1}_A \mathbb{1}_B] = \mathbb{E}[Y \mathbb{1}_A]\mathbb{P}[B]$.
- Since $Y = \mathbb{E}[X \mid \mathcal{A}]$, we know that $\mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A]$.

Combining the above three facts, we obtain (5.1.2) for $C = A \cap B$.

Denote by $\mathcal{C}$ the collection of sets $C \in \sigma(\mathcal{A}, \mathcal{B})$ such that (5.1.2) holds. Denote by $\mathcal{C}_0$ the field generated by $A \cap B$ where $A \in \mathcal{A}, B \in \mathcal{B}$. We can check that $\mathcal{C}_0 \subset \mathcal{C}$. It is clear that $\mathcal{C}$ is a monotone class. Lemma 1.1.4 implies that $\sigma(\mathcal{C}_0) \subset \mathcal{C}$ which gives the conclusion. □

Example 5.1.5. *Suppose X and Y are independent. Let φ be a measurable function on $\mathbb{R}^2$ such that $\mathbb{E}[|\varphi(X,Y)|] < \infty$ and let $g(x) = \mathbb{E}[\varphi(x,Y)]$. Then we have*

$$\mathbb{E}[\varphi(X,Y) \mid \sigma(X)] = g(X) \quad a.s.$$

The conclusion in this example looks intuitive, but let us emphasize that the conclusion can not be true without the assumption that X and Y are independent. For instantce, without such assumption, we may take $Y = X$, then we have

$$\mathbb{E}\left[\varphi(X,Y) \mid \sigma(X)\right] = \varphi(X,X) \quad a.s.$$

Proof. It is clear that $g(X) \in \sigma(X)$, it remains to show

$$\mathbb{E}[\varphi(X,Y) \mathbb{1}_A] = \mathbb{E}[g(X) \mathbb{1}_A], \quad \forall A \in \sigma(X).$$

This is equivalent to showing, for any measurable function f,

$$\mathbb{E}[\varphi(X,Y)f(X)] = \mathbb{E}[g(X)f(X)].$$

Denote by $\mu = \mathcal{L}(X)$ and $\nu = \mathcal{L}(Y)$. Then we have

$$\begin{aligned}
\mathbb{E}[\varphi(X,Y)f(X)] &= \iint \varphi(x,y)f(x)\mu(dx)\nu(dy) && \text{(since } X, Y \text{ are independent)} \\
&= \int \mu(dx)f(x) \int \varphi(x,y)\nu(dy) && \text{(by Fubini's theorem)} \\
&= \int \mu(dx)f(x)g(x) && \text{(by the definition of } g\text{)} \\
&= \mathbb{E}[g(X)f(X)].
\end{aligned}$$

□

Example 5.1.6. *Suppose that $\{X_n\}$ are i.i.d. with finite expectation. Let $S_n = X_1 + \cdots + X_n$, and define*

$$\mathcal{A}_n = \sigma(S_n, S_{n+1}, ...) = \sigma(S_n, X_{n+1}, ...).$$

Then, for all $n \geqslant 1$,

$$\mathbb{E}[X_1 \,|\, \mathcal{A}_n] = S_n/n.$$

Proof. Since X_1, S_n are independent of $X_{n+1}, ...$, we have that

$$\mathbb{E}[X_1 \,|\, \mathcal{A}_n] = \mathbb{E}[X_1 \,|\, \sigma(S_n)].$$

We will argue that, for all $1 \leqslant k \leqslant n$,

$$\mathbb{E}[X_k \,|\, \sigma(S_n)] = \mathbb{E}[X_1 \,|\, \sigma(S_n)] \tag{5.1.3}$$

This implies the conclusion. To show (5.1.3), we only need to show

$$\mathbb{E}[X_k \,\mathbb{1}_A] = \mathbb{E}[X_1 \,\mathbb{1}_A], \quad \forall A \in \sigma(S_n).$$

This is equivalent to the following

$$\mathbb{E}[X_k f(S_n)] = \mathbb{E}[X_1 f(S_n)], \quad \forall \text{ bounded measurable function } f.$$

This is true since (X_k, S_n) has the same law as (X_1, S_n). □

Example 5.1.7 (Conditional probability)**.** *Suppose A, B are events such that $\mathbb{P}[B] > 0$ and $\mathcal{G}$ is a σ-field. We define conditional probability as follows:*

$$\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A \,|\, \mathcal{G}], \quad \mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

In particular, if we have $0 < \mathbb{P}[B] < 1$ and $\mathcal{G} = \{\varnothing, \Omega, B, B^c\}$, we obtain

$$\mathbb{P}[A \,|\, \mathcal{G}] = \mathbb{P}[A \,|\, B] \,\mathbb{1}_B + \mathbb{P}[A \,|\, B^c] \,\mathbb{1}_{B^c}.$$

5.2 Martingales

Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

- A *filtration* $\{\mathcal{F}_n\}$ is an increasing family of sub-σ-fields of $\mathcal{F}$.
- A sequence of random variables $\{X_n\}$ is measurable (adapted) with respect to the filtration $\{\mathcal{F}_n\}$ if, for all n, the random variable X_n is $\mathcal{F}_n$-measurable.
- The natural filtration associated to $\{X_n\}$ is given by $\mathcal{F}_n = \sigma(X_1, ..., X_n)$ for $n \geqslant 1$.
- A sequence of random variables $\{X_n\}$ is integrable if $X_n \in L^1$ for all n.

Definition 5.2.1. *Let $X = \{X_n\}$ be an integrable process adapted to the filtration $\{\mathcal{F}_n, n \geqslant 0\}$.*

(1) X is a martingale if $\mathbb{E}[X_n \,|\, \mathcal{F}_m] = X_m$ a.s. for all $n \geqslant m$.

(2) X is a supermartingale if $\mathbb{E}[X_n \,|\, \mathcal{F}_m] \leqslant X_m$ a.s. for all $n \geqslant m$.

(3) X is a submartingale if $\mathbb{E}[X_n \,|\, \mathcal{F}_m] \geqslant X_m$ a.s. for all $n \geqslant m$.

Example 5.2.2. *(1) Let $\{\xi_i\}_{i \geqslant 1}$ be a sequence of independent random variables in L^1 with $\mathbb{E}[\xi_i] = 0$. Then $\{X_n := \sum_1^n \xi_i\}_{n \geqslant 1}$ is a martingale.*

(2) Let $\{\xi_i\}_{i\geqslant 1}$ be a sequence of independent random variables in L^1 with $\mathbb{E}[\xi_i] = 1$. Then $\{X_n := \prod_1^n \xi_i\}_{n\geqslant 1}$ is a martingale.

(3) Consider biased random walk on $\mathbb{Z}$: at each step, the walker goes to the right with probability p and goes to the left with probability $(1-p)$. Let X_n be the location of the walker at time n.

- *If $p = 1/2$, then $\{X_n\}$ is a martingale.*
- *If $p < 1/2$, then $\{X_n\}$ is a supermartingale. But $\{X_n - n(2p-1)\}$ is a martingale.*

The martingales have the following basic properties.

- If $\{X_n\}$ is a martingale, then $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ for all n.
- If $\{X_n\}$ is a supermartingale (resp. submartingale), then $\mathbb{E}[X_n]$ is decreasing (resp. increasing).
- If $\{X_n\}$ is a martingale and φ is a convex function, then $\{\varphi(X_n)\}$ is a submartingale. In particular, $\{|X_n|\}$ is a submartingale.

Definition 5.2.3. *Suppose that $\{\mathcal{F}_n\}$ is a filtration. A stopping time $T : \Omega \to \mathbb{N}^* = \{0, 1, 2, ..., \infty\}$ is a random variable such that*

$$\{T = n\} \in \mathcal{F}_n, \quad \forall n.$$

Lemma 5.2.4. *The following statements are equivalent.*

(1) $\{T = n\} \in \mathcal{F}_n$ for all n.

(2) $\{T \leqslant n\} \in \mathcal{F}_n$ for all n.

(3) $\{T > n\} \in \mathcal{F}_n$ for all n.

(4) $\{T \geqslant n\} \in \mathcal{F}_{n-1}$ for all n.

Lemma 5.2.5. *If S, T, T_j are stopping times for $j \geqslant 1$. The following random variables are also stopping times:*

$$S \vee T, \quad S \wedge T, \quad \inf_j T_j, \quad \sup_j T_j, \quad \liminf_j T_j, \quad \limsup_j T_j.$$

Proof. First, for the random variable $S \vee T$, we have that

$$\{S \vee T \leqslant n\} = \{S \leqslant n\} \cap \{T \leqslant n\} \in \mathcal{F}_n.$$

Next, for the random variable $\inf_j T_j$, we have that

$$\{\inf_j T_j \leqslant n\} = \cup_j \{T_j \leqslant n\} \in \mathcal{F}_n.$$

Finally, for the random variable $\liminf_j T_j$, we have that

$$\{\liminf_j T_j \leqslant n\} = \cap_m \cup_{j\geqslant m} \{T_j \leqslant n\} \in \mathcal{F}_n.$$

□

Definition 5.2.6. *Suppose that $\{\mathcal{F}_n\}$ is a filtration and that T is a stopping time. Define*

$$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \leqslant n\} \in \mathcal{F}_n, \forall n\}.$$

Intuitively, $\mathcal{F}_T$ is the information available by time T.

- If $T = n_0$, then $\mathcal{F}_T = \mathcal{F}_{n_0}$.

- The random variable $X_T \mathbb{1}_{\{T<\infty\}}$ is measurable with respect to $\mathcal{F}_T$.
- Suppose S and T are stopping times that $S \leqslant T$, then $\mathcal{F}_S \subset \mathcal{F}_T$.
- Suppose $\{X_n\}$ is a process adapted to $\{\mathcal{F}_n\}$, define $X_n^T := X_{T\wedge n}$. Then $\{X_n^T\}$ is also adapted to $\{\mathcal{F}_n\}$.

Proof. We leave the first and the second items as exericese. We first show the third item. For any $A \in \mathcal{F}_S$, in order to show that $A \in \mathcal{F}_T$, it is sufficient to show that $A \cap \{T \leqslant n\} \in \mathcal{F}_n$ for all n.

Since $S \leqslant T$, the event $A \cap \{S \leqslant n\}$ is the disjoint union of $A \cap \{T \leqslant n\}$ and $A \cap \{T > n \geqslant S\}$. Since $A \in \mathcal{F}_S$, we have that

$$A \cap \{S \leqslant n\} \in \mathcal{F}_n, \quad A \cap \{T > n \geqslant S\} = (A \cap \{S \leqslant n\}) \cap \{T > n\} \in \mathcal{F}_n.$$

This implies that $A \cap \{T \leqslant n\} \in \mathcal{F}_n$.

Next, we show the last item. It is equivalent to showing that $\{X_{T\wedge n} \in B\} \in \mathcal{F}_n$ for any Borel set B and for any n. Note that

$$\{X_{T\wedge n} \in B\} = \{X_n \in B, T > n\} \cup \{X_T \in B, T \leqslant n\}.$$

The first part $\{X_n \in B, T > n\} = \{X_n \in B\} \cap \{T > n\} \in \mathcal{F}_n$. The second part $\{X_T \in B, T \leqslant n\} = \cup_{j=1}^n \{X_j \in B, T = j\} \in \mathcal{F}_n$. ☐

Theorem 5.2.7 (Optional Stopping Theorem)**.** *Let $\{X_n\}$ be a martingale.*

(1) If T is a stopping time, then $\{X_n^T\}$ is also a martingale. In particular, $\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_0]$.

(2) If T is a stopping time bounded by a constant N, then $\mathbb{E}[X_N \mid \mathcal{F}_T] = X_T$ a.s. Furthermore, if S is stopping time such that $S \leqslant T$ a.s., we have $\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S$ a.s. In particular, $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

(3) Suppose that there is a random variable $Y \in L^1$ such that $|X_n| \leqslant Y$ for all n and that T is a stopping time which is a.s. finite. Then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

(4) Suppose that X has bounded increments, i.e. there is $M < \infty$ such that $|X_{n+1} - X_n| \leqslant M$ for all n, and that T is a stopping time with $\mathbb{E}[T] < \infty$. Then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

Proof of Item (1). We first show that $\{X_n^T\}$ is integrable. Since $\{X_n\}$ is integrable, we have

$$\mathbb{E}\left[|X_n^T|\right] \leqslant \mathbb{E}\left[\max_{i\leqslant n} |X_i|\right] \leqslant \sum_{i=1}^n \mathbb{E}[|X_i|] < \infty.$$

We already see that $\{X_n^T\}$ is adapted to $\{\mathcal{F}_n\}$. It remains to check the conditional expectation. For every $n \geqslant 1$,

$$\begin{aligned} \mathbb{E}[X_n^T \mid \mathcal{F}_{n-1}] &= \mathbb{E}[X_{n-1}^T + (X_n - X_{n-1}) \mathbb{1}_{\{T>n-1\}} \mid \mathcal{F}_{n-1}] \\ &= X_{n-1}^T + \mathbb{1}_{\{T>n-1\}} \mathbb{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] = X_{n-1}^T. \end{aligned}$$

By Tower Property, we could conclude that $\{X_n^T\}$ is a martingale. ☐

Proof of Item (2). For $A \in \mathcal{F}_T$,

$$\begin{aligned} \mathbb{E}[X_N \mathbb{1}_A] &= \sum_{i=1}^N \mathbb{E}[X_N \mathbb{1}_A \mathbb{1}_{\{T=i\}}] \\ &= \sum_{i=1}^N \mathbb{E}\left[\mathbb{E}[X_N \mid \mathcal{F}_i] \mathbb{1}_A \mathbb{1}_{\{T=i\}}\right] \\ &= \sum_{i=1}^N \mathbb{E}\left[X_i \mathbb{1}_A \mathbb{1}_{\{T=i\}}\right] = \mathbb{E}[X_T \mathbb{1}_A]. \end{aligned}$$

Therefore $\mathbb{E}[X_N \mid \mathcal{F}_T] = X_T$. Similarly, we have $\mathbb{E}[X_N \mid \mathcal{F}_S] = X_S$. The tower property gives

$$X_S = \mathbb{E}[X_N \mid \mathcal{F}_S] = \mathbb{E}[\mathbb{E}[X_N \mid \mathcal{F}_T] \mid \mathcal{F}_S] = \mathbb{E}[X_T \mid \mathcal{F}_S].$$

□

Proof of Item (3). Since $|X_n| \leqslant Y$ for all n and T is finite a.s., we have $|X_{n\wedge T}| \leqslant Y$. Then Dominated Convergence Theorem implies that

$$\lim_{n\to\infty} \mathbb{E}[X_{n\wedge T}] = \mathbb{E}[\lim_{n\to\infty} X_{n\wedge T}] = \mathbb{E}[X_T].$$

As $n \wedge T$ is a bounded stopping time, Item (2) implies that $\mathbb{E}[X_{n\wedge T}] = \mathbb{E}[X_0]$. Hence we conclude that $\mathbb{E}[X_T] = \mathbb{E}[X_0]$. □

Proof of Item (4). We can write $\mathbb{E}[X_T] = \mathbb{E}[X_0] + \mathbb{E}\left[\sum_{i=1}^{T}(X_i - X_{i-1})\right]$, so it suffices to show that the last term is zero. Note that

$$|\sum_{i=1}^{T}(X_i - X_{i-1})| \leqslant \sum_{i=1}^{T} |X_i - X_{i-1}| \leqslant MT \in L^1.$$

Then Dominated Convergence Theorem implies that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{T}(X_i - X_{i-1})\right] &= \mathbb{E}\left[\sum_{i=1}^{\infty}(X_i - X_{i-1})\,\mathbb{1}_{\{T\geqslant i\}}\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[(X_i - X_{i-1})\,\mathbb{1}_{\{T\geqslant i\}}] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[\mathbb{E}[(X_i - X_{i-1})\,\mathbb{1}_{\{T\geqslant i\}} \mid \mathcal{F}_{i-1}]] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[\mathbb{1}_{\{T\geqslant i\}}\,\mathbb{E}[(X_i - X_{i-1}) \mid \mathcal{F}_{i-1}]] = 0,
\end{aligned}$$

where we used that $\{T \geqslant i\} = \{T \leqslant i-1\}^c \in \mathcal{F}_{i-1}$ as T is a stopping time. □

Example 5.2.8. *Let $\{X_n\}$ be a simple random walk on $\mathbb{Z}$ starting from $k \in \{0, 1, \ldots, N\}$. Define $\tau = \min\{n : X_n = 0 \text{ or } N\}$. Then*

$$\mathbb{P}[X_\tau = N] = k/N.$$

First proof. We denote the probability measure for the simple random walk starting from k by $\mathbb{P}_k$ and set

$$p(k) = \mathbb{P}_k[X_\tau = N].$$

For $1 \leqslant k \leqslant N-1$,

$$p(k) = \mathbb{P}_k[X_\tau = N, X_1 = k+1] + \mathbb{P}_k[X_\tau = N, X_1 = k-1] = \frac{1}{2}p(k+1) + \frac{1}{2}p(k-1).$$

Thus p is a harmonic function on $\{0, 1, \ldots, N\}$ and it has boundary values $p(0) = 0$ and $p(N) = 1$, thus it is uniquely determined: $p(k) = k/N$. □

Second proof. The process $\{X_n\}$ is a martingale:

$$\mathbb{E}[X_n \mid \sigma(X_1, \ldots, X_{n-1})] = \frac{1}{2}(X_{n-1} + 1) + \frac{1}{2}(X_{n-1} - 1) = X_{n-1}.$$

By Optional Stopping Theorem, we have $\mathbb{E}[X_{\tau \wedge n}] = k$. Note that $|X_{\tau \wedge n}| \leqslant N$. Then Bounded Convergence Theorem gives

$$\mathbb{E}[X_\tau] = \lim_n \mathbb{E}[X_{\tau \wedge n}] = k.$$

Thus

$$N\mathbb{P}[X_\tau = N] = k$$

which gives the conclusion. □

Theorem 5.2.9 (Optional Stopping Theorem for supermartingale)**.** *Let $\{X_n\}$ be a supermartingale.*

(1) If T is a stopping time, then $\{X_n^T\}$ is also a supermartingale. In particular, $\mathbb{E}[X_{T \wedge n}] \leqslant \mathbb{E}[X_0]$.

(2) If T is a stopping time bounded by a constant N, then $\mathbb{E}[X_N \mid \mathcal{F}_T] \leqslant X_T$ a.s. Furthermore, if S is a stopping time such that $S \leqslant T$, we have $\mathbb{E}[X_T \mid \mathcal{F}_S] \leqslant X_S$ a.s. In particular, $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$.

(3) Suppose that there is a random variable $Y \in L^1$ such that $|X_n| \leqslant Y$ for all n and that T is a stopping time which is a.s. finite. Then $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$.

(4) Suppose that X has bounded increments, i.e. there is $M < \infty$ such that $|X_{n+1} - X_n| \leqslant M$ for all n, and that T is a stopping time with $\mathbb{E}[T] < \infty$. Then $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$.

(5) Suppose that X is a non-negative supermartingale. Then for any stopping time T which is finite a.s., we have $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$.

Proof of Item (2). We only show $\mathbb{E}[X_T \mid \mathcal{F}_S] \leqslant X_S$. To this end, it suffices to show $\mathbb{E}[X_T \mathbb{1}_A] \leqslant \mathbb{E}[X_S \mathbb{1}_A]$ for any $A \in \mathcal{F}_S$. By Item (1), we have $\mathbb{E}[X_{T \wedge n} \mid \mathcal{F}_m] \leqslant X_{T \wedge m}$ for $m \leqslant n$. In particular, we have

$$\mathbb{E}[X_T \mid \mathcal{F}_m] \leqslant X_{T \wedge m}, \quad \forall m \leqslant N. \tag{5.2.1}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[X_T \mathbb{1}_A] &= \sum_{k=0}^{N} \mathbb{E}[X_T \mathbb{1}_{A \cap \{S=k\}}] \\
&= \sum_{k=0}^{N} \mathbb{E}[\mathbb{1}_{A \cap \{S=k\}} \mathbb{E}[X_T \mid \mathcal{F}_k]] \\
&\leqslant \sum_{k=0}^{N} \mathbb{E}[\mathbb{1}_{A \cap \{S=k\}} X_{T \wedge k}] && \text{(by (5.2.1))} \\
&= \sum_{k=0}^{N} \mathbb{E}[\mathbb{1}_{A \cap \{S=k\}} X_k] = \mathbb{E}[X_S \mathbb{1}_A].
\end{aligned}$$

□

Proof of Item (5). By Item (1), we have that $\mathbb{E}[X_{T \wedge n}] \leqslant \mathbb{E}[X_0]$. By Fatou's Lemma, we have

$$\mathbb{E}[X_T] = \mathbb{E}[\liminf_n X_{T \wedge n}] \leqslant \liminf \mathbb{E}[X_{T \wedge n}] \leqslant \mathbb{E}[X_0].$$

□

5.3 Martingale convergence theorem

In this section, we will discuss three different notions of convergence of the martingales: almost sure convergence, convergence in L^p for $p > 1$, and convergence in L^1.

Almost sure convergence

Theorem 5.3.1 (Almost Sure Martingale Convergence)**.** *Let $\{X_n\}$ be a supermartingale which is bounded in L^1, i.e. $\sup_n \mathbb{E}[|X_n|] < \infty$. Then, for some $X_\infty \in L^1$,*

$$X_n \to X_\infty, \text{ almost surely, as } n \to \infty.$$

Let $x = \{x_n\}$ be a sequence of real numbers. Let $a < b$ be two real numbers. We define $T_0(x) = 0$ and inductively, for $k \geqslant 0$,

$$S_{k+1}(x) = \inf\{n \geqslant T_k(x) : x_n \leqslant a\}, \quad T_{k+1}(x) = \inf\{n \geqslant S_{k+1}(x) : x_n \geqslant b\},$$

with the usual convention that $\inf \varnothing = \infty$.

Define the number of upcrossings of $[a, b]$ by $x = \{x_n\}$ by time n to be

$$N_n([a, b], x) = \sup\{k \geqslant 0 : T_k(x) \leqslant n\}.$$

As $n \uparrow \infty$, we have

$$N_n([a, b], x) \uparrow N([a, b], x) = \sup\{k \geqslant 0 : T_k(x) < \infty\},$$

which is the total number of upcrossings of $[a, b]$ by x.

Lemma 5.3.2. *A sequence of real numbers x converges in $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ if and only if*

$$N([a, b], x) < \infty \quad \text{for all rationals } a < b.$$

Lemma 5.3.3 (Doob's upcrossing inequality)**.** *Let $X = \{X_n\}$ be a supermartingale and $a < b$ be two real numbers. Then, for all n,*

$$(b - a)\mathbb{E}[N_n([a, b], X)] \leqslant \mathbb{E}[(a - X_n)^+].$$

Proof. To simplify the notations, we write

$$T_k = T_k(X), \quad S_k = S_k(X), \quad N = N_n([a, b], X).$$

On the one hand, by the definition of $\{T_k\}$ and $\{S_k\}$, we have that, for all $k \geqslant 1$,

$$X_{T_k} - X_{S_k} \geqslant b - a. \tag{5.3.1}$$

On the other hand, we have

$$\begin{aligned}
&\sum_{k=1}^{n} (X_{T_k \wedge n} - X_{S_k \wedge n}) \\
&= \sum_{k=1}^{N} (X_{T_k} - X_{S_k}) + \sum_{k=N+1}^{n} (X_n - X_{S_k \wedge n}) \\
&= \sum_{k=1}^{N} (X_{T_k} - X_{S_k}) + \left(X_n - X_{S_{N+1}}\right) \mathbb{1}_{\{S_{N+1} \leqslant n\}}. \quad \text{(Note that } T_N \leqslant n, S_{N+1} < T_{N+1} < S_{N+2}\text{)}.
\end{aligned}$$

Since $\{T_k\}$ and $\{S_k\}$ are stopping times, we have that $S_k \wedge n \leqslant T_k \wedge n$ are bounded stopping times. Optional stopping theorem implies

$$\mathbb{E}[X_{S_k \wedge n}] \geqslant \mathbb{E}[X_{T_k \wedge n}], \quad \forall k.$$

Combining with (5.3.1), we have

$$0 \geqslant \mathbb{E}\left[\sum_{k=1}^{n} (X_{T_k \wedge n} - X_{S_k \wedge n})\right] \geqslant (b-a)\mathbb{E}[N] - \mathbb{E}[(a - X_n)^+],$$

since $\left(X_n - X_{S_{N+1}}\right) \mathbb{1}_{\{S_{N+1} \leqslant n\}} \geqslant -(a - X_n)^+$. This implies the conclusion. □

Proof of Theorem 5.3.1. Let $a < b$ be rationals. By Lemma 5.3.3, we have that

$$\mathbb{E}[N_n([a,b], X)] \leqslant \frac{\mathbb{E}[(a - X_n)^+]}{b-a} \leqslant \frac{\mathbb{E}[|X_n|] + |a|}{b-a}.$$

Monotone convergence theorem implies

$$\mathbb{E}[N([a,b], X)] \leqslant \frac{\sup_n \mathbb{E}[|X_n|] + |a|}{b-a} < \infty.$$

Therefore, we have almost surely that $N([a,b], X) < \infty$. Write

$$\Omega_0 = \bigcap_{a<b: a,b \in \mathbb{Q}} \{N([a,b], X) < \infty\}.$$

Then $\mathbb{P}[\Omega_0] = 1$. By Lemma 5.3.2 on Ω_0, we have that X converges to a possibly infinite limit. Set

$$X_\infty = \begin{cases} \lim_n X_n, & \text{on } \Omega_0, \\ 0 & \text{on } \Omega \backslash \Omega_0. \end{cases}$$

Then X_∞ is $\mathcal{F}_\infty$-measurable and by Fatou's lemma, we have

$$\mathbb{E}[|X_\infty|] \leqslant \mathbb{E}[\liminf_n |X_n|] \leqslant \sup_n \mathbb{E}[|X_n|] < \infty.$$

Therefore $X_\infty \in L^1$. □

Corollary 5.3.4. *Let $\{X_n\}$ be a non-negative supermartingale. Then X_n converges a.s. to some a.s. finite limit.*

Example 5.3.5. *Let $\{\xi_i\}_{i \geqslant 1}$ be a sequence of non-negative independent random variables in L^1 with $\mathbb{E}[\xi_i] = 1$. Then $\{X_n := \Pi_1^n \xi_i\}_{n \geqslant 1}$ is a non-negative martingale and X_n converges a.s. to some limit $X_\infty \in L^1$.*

When we have almost sure convergence, a natural question is: Do we have $\mathbb{E}[X_\infty] = \mathbb{E}[X_0]$? Answer: It is true when we have convergence in L^1. We will discuss convergence in L^p for $p > 1$ (which implies convergence in L^1) and convergence in L^1 separately.

Convergence in L^p

Theorem 5.3.6 (Doob's Maximal Inequality). *Let $\{X_n\}$ be a non-negative submartingale. Define $X_n^* = \max_{k \leqslant n} X_k$. Then*

$$\lambda \mathbb{P}[X_n^* \geqslant \lambda] \leqslant \mathbb{E}\left[X_n \mathbb{1}_{\{X_n^* \geqslant \lambda\}}\right] \leqslant \mathbb{E}[X_n].$$

Proof. Define the stopping time $T = \min\{n : X_n \geqslant \lambda\}$. Then we have $\{X_n^* \geqslant \lambda\} = \{T \leqslant n\}$. Moreover,

$$\mathbb{E}[X_n \mathbb{1}_{\{T \leqslant n\}}] = \sum_0^n \mathbb{E}[X_n \mathbb{1}_{\{T=j\}}] \geqslant \sum_0^n \mathbb{E}[X_j \mathbb{1}_{\{T=j\}}] \geqslant \sum_0^n \lambda \mathbb{P}[T = j] = \lambda \mathbb{P}[T \leqslant n].$$

□

Theorem 5.3.7 (Doob's Maximal Inequality). *Let $\{X_n\}$ be a non-negative submartingale. Define $X_n^* = \max_{k \leqslant n} X_k$. Then, for all $p > 1$, we have*

$$||X_n^*||_p \leqslant \frac{p}{p-1} ||X_n||_p.$$

Proof.

$$\begin{aligned}
\mathbb{E}[(X_n^*)^p] &= \mathbb{E}\left[\int_0^\infty p x^{p-1} \mathbb{1}_{\{X_n^* \geqslant x\}}\, dx\right] \\
&= \int_0^\infty p x^{p-1} \mathbb{P}[X_n^* \geqslant x] dx && \text{(by Monotone Cvg Thm)} \\
&\leqslant \int_0^\infty p x^{p-2} \mathbb{E}[X_n \mathbb{1}_{\{X_n^* \geqslant x\}}] dx \\
&= \mathbb{E}\left[\int_0^\infty p x^{p-2} X_n \mathbb{1}_{\{X_n^* \geqslant x\}}\, dx\right] && \text{(by Monotone Cvg Thm)} \\
&= \frac{p}{p-1} \mathbb{E}[X_n (X_n^*)^{p-1}] \leqslant \frac{p}{p-1} ||X_n||_p \mathbb{E}[(X_n^*)^p]^{1-1/p}.
\end{aligned}$$

□

Theorem 5.3.8. *Let $\{X_n\}$ be a martingale and $p > 1$, then the following statements are equivalent.*

(1) $\{X_n\}$ is bounded in L^p: $\sup_n ||X_n||_p < \infty$

(2) X_n converges a.s and in L^p to a random variable X_∞.

(3) There exists a random variable $Z \in L^p$ such that

$$X_n = \mathbb{E}[Z \mid \mathcal{F}_n] \quad a.s. \quad \forall n.$$

Proof of Item (1) to (2). Assume that $\sup_n \mathbb{E}[|X_n|^p] \leqslant C$. This implies that $\{X_n\}$ is bounded in L^1, thus we know that X_n converges a.s. to some limit X_∞.

Define $X_n^* = \max_{k \leqslant n} |X_k|$. By Doob's Maximal Inequality, we have that $\mathbb{E}[(X_n^*)^p] \leqslant (p/(p-1))^p C$. Note that the sequence $\{X_n^*\}$ is increasing in n, thus we may define $X_\infty^* = \lim_n X_n^*$ and $X_\infty^* \in L^p$.

For the sequence $\{X_n\}$, we have $|X_n| \leqslant X_\infty^* \in L^p$. Therefore, by Dominated Convergence Theorem, we have that X_n converges to X_∞ in L^p. □

Proof of Item (2) to (3). We may take $Z = X_\infty$. We have that,

$$|X_n - \mathbb{E}[X_\infty \mid \mathcal{F}_n]| = |\mathbb{E}[X_m \mid \mathcal{F}_n] - \mathbb{E}[X_\infty \mid \mathcal{F}_n]| \leqslant \mathbb{E}[|X_m - X_\infty| \mid \mathcal{F}_n] \to 0, \text{ as } m \to \infty.$$

□

Proof of Item (3) to (1). By Jensen's Inequality, we have

$$|X_n|^p \leqslant \mathbb{E}[|Z|^p \,|\, \mathcal{F}_n].$$

This implies the conclusion. $\square$

Corollary 5.3.9. *Let $Z \in L^p$ for $p > 1$. Then*

$$\mathbb{E}[Z \,|\, \mathcal{F}_n] \to \mathbb{E}[Z \,|\, \mathcal{F}_\infty], \quad a.s. and\ in\ L^p.$$

Convergence in L^1

Next, we discuss convergence in L^1, to this end, we need the notion of "uniform integrable (UI)" introduced in Definition 2.5.1.

Lemma 5.3.10. *If $Z \in L^1$, then the family*

$$\{\mathbb{E}[Z \,|\, \mathcal{A}] : \mathcal{A} \text{ sub-}\sigma\text{-field of } \mathcal{F}\}$$

is UI.

Proof. First, recall from (2.5.1) that, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\mathbb{E}[|Z|\,\mathbb{1}_A] \leqslant \epsilon \quad \text{as long as} \quad \mathbb{P}[A] \leqslant \delta.$$

Next, we show the conclusion. It is sufficient to show that, for any $\epsilon > 0$, there exists α such that

$$\mathbb{E}\left[|\mathbb{E}[Z \,|\, \mathcal{A}]|\,\mathbb{1}_{\{|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha\}}\right] \leqslant \epsilon.$$

Note that the event $\{|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha\}$ is $\mathcal{A}$-measurable, thus

$$\mathbb{E}\left[|\mathbb{E}[Z \,|\, \mathcal{A}]|\,\mathbb{1}_{\{|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha\}}\right] \leqslant \mathbb{E}[|Z|\,\mathbb{1}_{\{|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha\}}].$$

By (2.5.1), it is sufficient to show that, there exists α such that

$$\mathbb{P}[|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha] \leqslant \delta.$$

Note that,

$$\mathbb{P}[|\mathbb{E}[Z \,|\, \mathcal{A}]| > \alpha] \leqslant \mathbb{E}[|\mathbb{E}[Z \,|\, \mathcal{A}]|]/\alpha \leqslant \mathbb{E}[|Z|]/\alpha.$$

This implies the conclusion. $\square$

Lemma 5.3.11. *Suppose that $X_n, X \in L^1$ and $X_n \to X$ a.s. Then*

$$X_n \to X \text{ in } L^1 \quad \text{if and only if} \quad \{X_n\} \text{ is UI}.$$

Proof. See Proposition 2.5.4. $\square$

Theorem 5.3.12. *Let $\{X_n\}$ be a martingale. The following statements are equivalent.*

(1) $\{X_n\}$ *is* UI.

(2) X_n *converges to* X_∞ *a.s. and in* L^1.

(3) There exists $Z \in L^1$ such that

$$X_n = \mathbb{E}[Z \,|\, \mathcal{F}_n] \quad a.s. \quad \forall n.$$

Proof of Item (1) to (2). Combining Lemma 5.3.11 and Theorem 5.3.1. $\square$

Proof of Item (2) to (3). The same as the proof of Theorem 5.3.8. ☐

Proof of Item (3) to (1). Lemma 5.3.10. ☐

Example 5.3.13. *Let $\{\xi_i\}_{i\geqslant 1}$ be a sequence of non-negative independent random variables in L^1 with $\mathbb{E}[\xi_i] = 1$. Then $\{X_n := \Pi_1^n \xi_i\}_{n\geqslant 1}$ is a non-negative martingale and X_n converges a.s. to some limit $X_\infty \in L^1$. Define, for $i \geqslant 1$,*

$$a_i = \mathbb{E}[\sqrt{\xi_i}].$$

(1) If $\Pi_i a_i > 0$, then X_n converges in L^1 and $\mathbb{E}[X_\infty] = 1$.

(2) If $\Pi_i a_i = 0$, then $X_\infty = 0$ a.s.

Proof of Case (1). Define $Y_n = \Pi_{i=1}^n \sqrt{\xi_i}/a_i$. Then $\{Y_n\}$ is a martingale and the relation with $\{X_n\}$ is $Y_n^2 = X_n/\Pi_{i=1}^n a_i^2$. Thus,

$$\mathbb{E}[Y_n^2] = 1/\Pi_{i=1}^n a_i^2, \quad \sup_n \mathbb{E}[Y_n^2] \leqslant 1/\Pi_{i=1}^\infty a_i^2 < \infty.$$

Define $Y_n^* = \max_{k\leqslant n} Y_k$, by Doob's Maximal Inequality, we have that

$$Y_n^* \uparrow Y_\infty^*, \quad \mathbb{E}[(Y_\infty^*)^2] \leqslant 2/\Pi_{i=1}^\infty a_i^2 < \infty.$$

Note that

$$X_n = Y_n^2 \Pi_{i=1}^n a_i \leqslant (Y_\infty^*)^2 \in L^1.$$

By Dominated Convergence Theorem, we have that X_n converges to X_∞ in L^1. ☐

Proof of Case (2). Define Y_n in the same way. Since $\{Y_n\}$ is a non-negative martingale, it converges a.s. to some limit $Y_\infty \in L^1$. Therefore, a.s.

$$X_n = Y_n^2 \times \Pi_{i=1}^n a_i^2 \to Y_\infty^2 \times 0 = 0.$$

☐

Example 5.3.14 (Kolmogorov's 0-1 Law, Another proof)**.** *Let $\{\xi_i\}_{i\geqslant 1}$ be a sequence of independent random variables. Let $\mathcal{G}_n = \sigma(\xi_k, k \geqslant n)$ and $\mathcal{G}_\infty = \cap_{n\geqslant 1}\mathcal{G}_n$. Then $\mathcal{G}_\infty$ is trivial, i.e. for any $A \in \mathcal{G}_\infty$, we have*

$$\mathbb{P}[A] = 0 \text{ or } 1.$$

Proof. Define $\{\mathcal{F}_n\}$ to be the natural filtration:

$$\mathcal{F}_n = \sigma(\xi_k, k \leqslant n).$$

Fix an event $A \in \mathcal{G}_\infty$.

- On the one hand, the sequence $\{\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n]\}$ is a UI martingale, thus

  $$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] \to \mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_\infty] = \mathbb{1}_A, \quad a.s.$$

  where $\mathcal{F}_\infty = \sigma(\mathcal{F}_n, n \geqslant 1)$.

- On the other hand, since $A \in \mathcal{G}_\infty \subset \mathcal{G}_{n+1}$ which is independent of $\mathcal{F}_n$, thus

  $$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] = \mathbb{P}[A].$$

Combining these two facts, we have that $\mathbb{P}[A] = \mathbb{1}_A$ *a.s.* Therefore $\mathbb{P}[A]$ is 0 or 1. ☐

Theorem 5.3.15 (Optional Stopping Theorem for UI Martingale)**.** *Let $X = \{X_n\}$ be a UI martingale. If $S \leqslant T$ are stopping times, then $\mathbb{E}[X_T \,|\, \mathcal{F}_S] = X_S, a.s.$ In particular, $\mathbb{E}[X_T] = \mathbb{E}[X_S]$.*

Compare with Optional Stopping Theorem for Martingale: Let $X = \{X_n\}$ be a martingale. If $S \leqslant T$ are bounded stopping times, then $\mathbb{E}[X_T \,|\, \mathcal{F}_S] = X_S, a.s.$ In particular, $\mathbb{E}[X_T] = \mathbb{E}[X_S]$.

Proof. It is sufficient to show that $X_T = \mathbb{E}[X_\infty \,|\, \mathcal{F}_T]$. Assume this is true, then we have that

$$\mathbb{E}[X_T \,|\, \mathcal{F}_S] = \mathbb{E}[\mathbb{E}[X_\infty \,|\, \mathcal{F}_T] \,|\, \mathcal{F}_S] = \mathbb{E}[X_\infty \,|\, \mathcal{F}_S] = X_S, \quad a.s.$$

To show $X_T = \mathbb{E}[X_\infty \,|\, \mathcal{F}_T]$, we have two steps.

First, we show $X_T \in L^1$. Since $\{X_n\}$ is UI, we have that $X_n = \mathbb{E}[X_\infty \,|\, \mathcal{F}_n]$, thus $|X_n| \leqslant \mathbb{E}[|X_\infty| \,|\, \mathcal{F}_n]$. Therefore,

$$\mathbb{E}[|X_T|] \leqslant \sum_{n=0}^{\infty} \mathbb{E}[|X_n| \mathbb{1}_{\{T=n\}}] + \mathbb{E}[|X_\infty| \mathbb{1}_{\{T=\infty\}}] \leqslant \sum_{n=0}^{\infty} \mathbb{E}[|X_\infty| \mathbb{1}_{\{T=n\}}] + \mathbb{E}[|X_\infty| \mathbb{1}_{\{T=\infty\}}] = \mathbb{E}[|X_\infty|].$$

Next, for any $A \in \mathcal{F}_T$, we have

$$\mathbb{E}[X_T \mathbb{1}_A] = \sum_{n \in \mathbb{N} \cup \{\infty\}} \mathbb{E}[X_n \mathbb{1}_A \mathbb{1}_{\{T=n\}}] = \sum_{n \in \mathbb{N} \cup \{\infty\}} \mathbb{E}[X_\infty \mathbb{1}_A \mathbb{1}_{\{T=n\}}] = \mathbb{E}[X_\infty \mathbb{1}_A].$$

This completes the proof. □

5.4 Applications: Galton-Watson Tree

A tree is a connected graph with no cycles. A rooted tree has a distinguished vertex v_0, called the root. The depth of a vertex v is its graph distance to the root. A leaf is a vertex with degree one.

Consider a regular rooted tree:

- Each vertex has a fixed number (say m) of offspring;
- Let Z_n be the number vertices in the n-th generation, then $Z_n = m^n$.

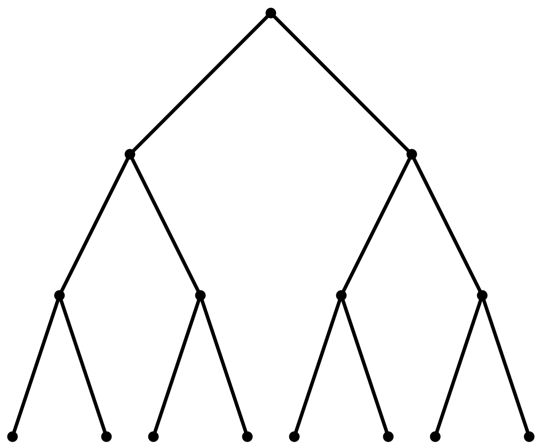
In real life, we often encounter trees where the number of offspring of a vertex is random. In this section, we will talk about the simplest random tree—Galton-Watson Tree:

- It starts with one initial ancestor;
- It produces a certain number of offspring according to some distribution μ;
- The new particles form the first generation;
- Each of the new particles produces offspring according to μ, independently of each other;
- The system regenerates.

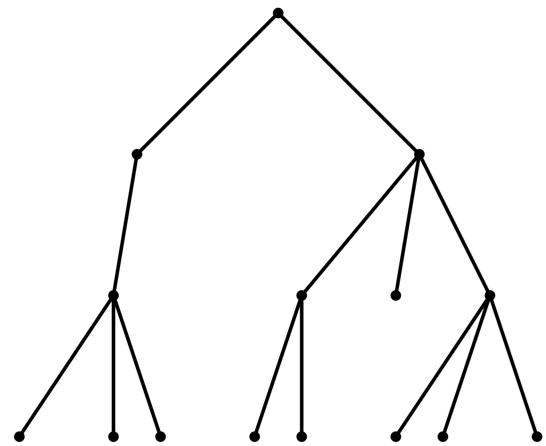
Let Z_n be the number of particles in n-th generation. Note that, if $Z_n = 0$ for some n, then $Z_m = 0$ for all $m \geqslant n$, and this is the situation that the family become extinct. The natural question here is whether or not the family become extinct, and what is the extinction probability

$$q = \mathbb{P}[Z_n = 0 \text{ eventually}]?$$

Of course, the answer to these questions depend on the reproduction law μ. For the measure μ, let p_k denote the probability that a particle has k children for $k \geqslant 0$. Clearly, $\sum_k p_k = 1$. To avoid trivial situation, throughout this section, we are always under the assumption that $p_0 + p_1 < 1$.



(a) A binary tree.



(b) A GW tree.

Figure 5.1

Define the mean of the measure

$$m := \mathbb{E}[Z_1] = \sum_{k \geqslant 0} k p_k.$$

Define the generating function of μ:

$$f(s) = \mathbb{E}\left[s^{Z_1}\right] = \sum_{k \geqslant 0} s^k p_k,$$

with the convention that $0^0 = 1$. Note that

$$f(0) = p_0, \quad f(1) = 1, \quad f'(1) = m.$$

Theorem 5.4.1. *The extinction probability q is the smallest root of $f(s) = s$ for $s \in [0,1]$. In particular, $q = 1$ if $m \leqslant 1$, and $q < 1$ if $m > 1$.*

Proof. First, we will give a mathematical description of the model. The tree starts with one ancestor: $Z_0 = 1$. The ancestor has Z_1 (with the law μ) children which form the 1st generation. For the particles in the 1st generation, they have $\xi_j^{(1)}$ children for $j = 1, ..., Z_1$. The random variables $\xi_j^{(1)}$ are i.i.d. with the common law μ. The number of particles in 2nd generation is then

$$Z_2 = \sum_{j=1}^{Z_1} \xi_j^{(1)}.$$

Generally, given Z_n, the particles in n-th generation have $\xi_j^{(n)}$ children for $j = 1, ..., Z_n$. Given Z_n, these random variables are i.i.d. with the common law μ. The number of particles in $(n+1)$-th generation is then

$$Z_{n+1} = \sum_{j=1}^{Z_n} \xi_j^{(n)}.$$

It is clear that, we have $\mathbb{E}[Z_n] = m^n$ for all $n \geqslant 1$.

Second, we will prove by induction that the generating function of Z_n is the n-th fold composition of f (denoted by f_n), i.e.

$$\mathbb{E}[s^{Z_n}] = f_n(s), \quad \forall s \in [0,1], \quad n \geqslant 1.$$

This is true for $n = 1$. Assume it holds for n, and consider $n+1$, we have

$$\mathbb{E}\left[s^{Z_{n+1}} \,|\, \sigma(Z_n)\right] = f(s)^{Z_n}, \quad \mathbb{E}[s^{Z_{n+1}}] = \mathbb{E}\left[f(s)^{Z_n}\right] = f_n(f(s)) = f_{n+1}(s).$$

Third, we will relate the extinction probability to the generating function. On the one hand, we have $\mathbb{P}[Z_n = 0] = f_n(0)$. On the other hand, the events $\{Z_n = 0\}$ is increasing in n, i.e.

$$\{Z_n = 0\} \subset \{Z_{n+1} = 0\}.$$

Thus, the probabilities $\mathbb{P}[Z_n = 0]$ is increasing, and we have

$$q = \mathbb{P}[\cup_n \{Z_n = 0\}] = \lim_n \mathbb{P}[Z_n = 0] = \lim_n f_n(0).$$

Finally, let us find the limit $\lim_n f_n(0)$. Consider the generating function f:

$$f(s) = p_0 + sp_1 + s^2 p_2 + \cdots, \quad f(0) = p_0, \quad f(1) = 1, \quad f'(1) = m.$$

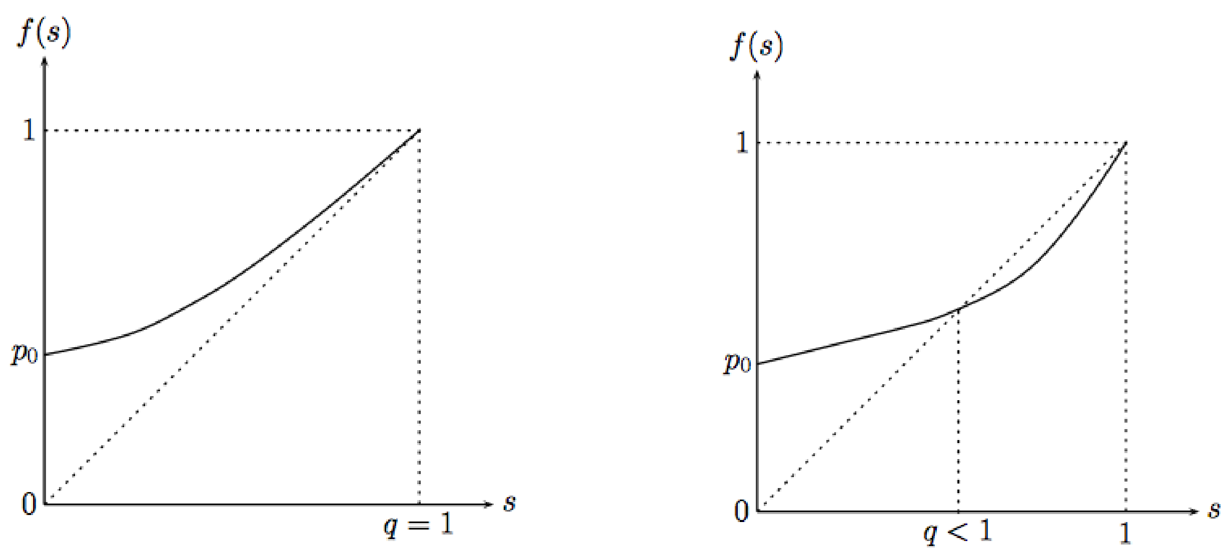
It is clear that the function is strictly increasing ($f'(s) > 0$) and is strictly convex ($f''(s) > 0$). Therefore, it has at most two fixed points.[3] The sequence

$$f(0) = p_0, \quad f_1(0) = f(p_0), \cdots$$

is increasing, and converges to some fixed point of the function f.

If $f'(1) = m \leqslant 1$, then $p_0 > 0$ and $f(s) > s$ for all $s \in [0,1)$. Thus $f_n(0) \to 1$, and $q = 1$ which is the unique root of $f(s) = s$.

If $f'(1) = m > 1$, then the function f has exactly two fixed points, and $f_n(0)$ converges to the root of $f(s) = s$ for $s \in [0,1)$. In particular, $q < 1$. $\square$



(a) When $f'(1) \leqslant 1$, the sequence $f_n(0)$ converges to 1 which is the unique fixed point.

(b) When $f'(1) > 1$, the sequence $f_n(0)$ converges to the smaller fixed point.

Figure 5.2: Figure from Zhan Shi.

From this theorem, we have that

[3] A fixed point of function f is a point x such that $f(x) = x$.

- In the subcritical case ($m < 1$), the GW tree dies out with probability 1.
- In the critical case ($m = 1$), the GW tree dies out with probability 1.
- In the supercritical case ($m > 1$), the GW tree survives with strictly positive probability $1 - q$.

In the supercritical case $m > 1$, we know that the tree survives with positive probability $1-q$. Our next question is, conditioned on survival, how fast does the population Z_n grow? We know that $\mathbb{E}[Z_n] = m^n$, and whether do we have that Z_n grows like m^n?

Assume the supercritical case $m \in (1, \infty)$. Define $W_n = Z_n/m^n$. Then $\{W_n\}$ is a non-negative martingale, therefore W_n converges a.s. to some limit W. By Fatou's Lemma, we have that $\mathbb{E}[W] \leqslant 1$. Note that, if $W > 0$, then $Z_n \sim m^n$; and if $W = 0$, then $Z_n << m^n$. Thus, in order to see whether or not Z_n grows like m^n, we need to examine whether W is strictly positive or not.

Theorem 5.4.2 (Kesten-Stigum Theorem)**.**

$$\mathbb{E}[W] = 1 \quad \Leftrightarrow \quad \mathbb{P}[W > 0 \,|\, \textit{survival}] = 1 \quad \Leftrightarrow \quad \mathbb{E}[Z_1 \log^+ Z_1] < \infty.$$[4]

Proof. Bonus. □

From Theorem 5.4.2, if $\mathbb{E}[Z_1 \log^+ Z_1] < \infty$, then $W > 0$ almost surely on survival. In particular, we know that Z_n grows like m^n as $n \to \infty$ on survival. The next question is the following: If $\mathbb{E}[Z_1 \log^+ Z_1] = \infty$, we have $W = 0$ almost surely, thus $Z_n << m^n$. This implies that m^n is not the correct normalization. It is natural to ask whether there exists c_n such that $c_n Z_n$ converges to non-trivial limit. This is the so-called Seneta-Hedye norming problem.

We do not plan to give a proof of Theorem 5.4.2 in this note, instead we give a weaker version of this theorem.

Lemma 5.4.3. *The probability $\mathbb{P}[W = 0]$ is either q or 1.*

Proof. Given Z_1, for $n \geqslant 1$, we have that

$$Z_{n+1} \overset{d}{=} \sum_{j=1}^{Z_1} Z_n^{(j)},$$

where $Z_n^{(j)}$ are independent copies of Z_n. Rearranging, we have

$$m \frac{Z_{n+1}}{m^{n+1}} \overset{d}{=} \sum_{j=1}^{Z_1} \frac{Z_n^{(j)}}{m^n}.$$

Note that, LHS will converges to mW a.s. and RHS will converge a.s. to

$$\sum_{1}^{Z_1} W^{(j)},$$

where $W^{(j)}$ are independent copies of W. Denote by p the probability $\mathbb{P}[W = 0]$, then

$$p = \mathbb{P}[W = 0] = \mathbb{P}[W^{(j)} = 0, j = 1, ..., Z_1] = \mathbb{E}[p^{Z_1}] = f(p).$$

Thus p is a fixed point of f. □

Theorem 5.4.4. *If $\mathbb{E}[Z_1^2] < \infty$, then $\mathbb{E}[W] = 1$ and $\mathbb{P}[W > 0 \,|\, \textit{survival}] = 1$.*

[4] Notation $\log^+ x$ means $\log(1 \vee x)$.

Proof. We will show that the martingale $\{W_n\}$ is bounded in L^2. Assume this is true, then the martingale converges in L^1 and therefore

$$\mathbb{E}[W] = 1, \quad \mathbb{P}[W = 0] = q.$$

To show $\{W_n\}$ is bounded in L^2, we need to calculate $\mathbb{E}[Z_n^2]$. Define

$$\sigma^2 = \mathrm{var}(Z_1) = \mathbb{E}[Z_1^2] - m^2.$$

Then

$$\mathbb{E}[Z_{n+1}^2 \,|\, \sigma(Z_n)] = Z_n\sigma^2 + Z_n^2 m^2, \quad \mathbb{E}[Z_{n+1}^2] = \sigma^2 m^n + m^2\mathbb{E}[Z_n^2].$$

Thus,

$$\mathbb{E}[W_{n+1}^2] = \frac{\sigma^2}{m^{n+1}} + \mathbb{E}[W_n^2].$$

Note that $\mathbb{E}[W_1^2] = \sigma^2/m^2 + 1$, therefore

$$\mathbb{E}[W_{n+1}^2] = \frac{\sigma^2}{m^{n+2}} + \frac{\sigma^2}{m^{n+1}} + \cdots + \frac{\sigma^2}{m^2} + 1.$$

This implies that $\{W_n\}$ is bounded in L^2. $\square$

5.5 Applications: continued

Polya's Urn

Example 5.5.1. *An urn contains r red balls and g green balls. At each time, we draw a ball out, put it back, and add c more balls of the same color. Let X_n be the fraction of green balls after the nth draw. Then $\{X_n\}$ is a martingale and thus $X_n \to X_\infty$ a.s.*

- *When $g = r = c = 1$, the limit X_∞ is uniform in $(0, 1)$.*
- *When $g = 2, r = c = 1$, the limit X_∞ has density function $p(x) = 2x\,\mathbb{1}_{\{x\in(0,1)\}}$.*
- *Derive the distribution of X_∞ in general case. Bonus.*

Proof. We first check that $\{X_n\}$ is a martingale. Given all the information by time n, let us consider X_{n+1}: suppose there are j green balls and i red balls by time n (after the nth draw is completed and the new balls have been added), i.e. $X_n = j/(i + j)$. Then we have

$$X_{n+1} = \begin{cases} \frac{j+c}{i+j+c}, & \text{with probability } \frac{j}{i+j}, \\ \frac{j}{i+j+c}, & \text{with probability } \frac{i}{i+j}. \end{cases}$$

Thus

$$\mathbb{E}[X_{n+1} \,|\, \mathcal{F}_n] = \frac{j+c}{i+j+c} \times \frac{j}{i+j} + \frac{j}{i+j+c} \times \frac{i}{i+j} = \frac{j}{i+j} = X_n.$$

This confirms that $\{X_n\}$ is a martingale. As $0 \leqslant X_n \leqslant 1$, Theorem 5.3.1 gives $X_n \to X_\infty$ a.s. It remains to derive the distribution of X_∞.

We have the following two observations.

- The probability of getting green balls on the first m draws then red on the next $\ell = n - m$ draws is
$$\frac{g}{g+r} \cdot \frac{g+c}{g+r+c} \cdots \frac{g+(m-1)c}{g+r+(m-1)c} \cdot \frac{r}{g+r+mc} \cdots \frac{r+(\ell-1)c}{g+r+(n-1)c}.$$
- Any other outcome of the first n draws with m green balls and ℓ red balls has the same probability.

Combining the two facts, we obtain the probability for $X_n = (g + mc)/(g + r + nc)$.

When $g = r = c = 1$, we have

$$\mathbb{P}\left[X_n = \frac{m+1}{n+2}\right] = \binom{n}{m}\frac{m!(n-m)!}{(n+1)!} = \frac{1}{n+1}.$$

Thus X_∞ is uniform in $(0, 1)$.

When $g = 2, r = c = 1$, we have

$$\mathbb{P}\left[X_n = \frac{m+2}{n+3}\right] = \binom{n}{m}\frac{(m+1)!(n-m)!}{(n+2)!/2} = \frac{2(m+1)}{(n+2)(n+1)}.$$

Thus X_∞ has density function $p(x) = 2x\,\mathbb{1}_{\{x\in(0,1)\}}$. □

Backwards martingales

Definition 5.5.2. *Let $\cdots \subset \mathcal{G}_{-2} \subset \mathcal{G}_{-1} \subset \mathcal{G}_0$ be a sequence of sub-σ-fields indexed by $\mathbb{Z}_-$. Given such a filtration, a process $\{X_n, n \leqslant 0\}$ is called a backwards martingale, if it is adapted to the filtration, $X_0 \in L^1$, and for all $n \leqslant -1$, we have*

$$\mathbb{E}[X_{n+1} \,|\, \mathcal{G}_n] = X_n \quad a.s.$$

Suppose that $\{X_n, n \leqslant 0\}$ is a backwards martingale. By Tower Property, we have that

$$\mathbb{E}[X_0 \,|\, \mathcal{G}_n] = X_n \quad a.s.$$

Thus the backwards martingale is automatically UI, and all conclusions for (forward) martingales convergence also hold for backwards martingales.

Theorem 5.5.3. *Let $\{X_n, n \leqslant 0\}$ be a backwards martingale, with $X_0 \in L^p$ for some $p \in [1, \infty)$. Then*

$$X_n \to X_{-\infty} := \mathbb{E}[X_0 \,|\, \mathcal{G}_{-\infty}], \quad a.s. \text{ and in } L^p,$$

where $\mathcal{G}_{-\infty} = \cap_{n\leqslant 0}\mathcal{G}_n$.

We leave the proof of this theorem as an exercise. By the convergence theorem of backwards martingales, we could give a new proof of Strong Law of Large Numbers.

Corollary 5.5.4 (Strong Law of Large Numbers, Another proof)**.** *Let $\{\xi_i\}_{i\geqslant 1}$ be i.i.d with $\mathbb{E}[\xi_i] = m$. Let $S_n = \sum_{i=1}^n \xi_i$. Then*

$$S_n/n \to m, \quad a.s. \text{ and in } L^1.$$

Proof. For $n \geqslant 1$, define

$$\mathcal{G}_{-n} = \sigma(S_n, S_{n+1}, ...), \quad X_{-n} = S_n/n.$$

We will prove that $\{X_n, n \leqslant 0\}$ is a backwards martingale with respect to $\{\mathcal{G}_n, n \leqslant 0\}$. Assume this is true, then by Theorem 5.5.3, we have that S_n/n converges a.s. and in L^1 to some random variable, denoted by Y. Note that, for any k,

$$Y = \lim_n \frac{1}{n}\sum_{i=k}^{k+n} \xi_i.$$

Thus Y is $\sigma(\xi_i, i \geqslant k)$-measurable, for all k. By Kolmogorov 0-1 Law, we conclude that there exists a constant $c \in \mathbb{R}$ such that $\mathbb{P}[Y = c] = 1$. At the same time, since we also have convergence in L^1, we derive that

$$c = \mathbb{E}[Y] = \lim_n \mathbb{E}[S_n/n] = m.$$

In order to show that $\{X_n, n \leqslant 0\}$ is a backwards martingale with respect to $\{\mathcal{G}_n, n \leqslant 0\}$, we need to calculate $\mathbb{E}[X_{-n+1} \,|\, \mathcal{G}_{-n}]$. Recall Example 5.1.6, we have that

$$\begin{aligned}
\mathbb{E}[X_{-n+1} \,|\, \mathcal{G}_{-n}] &= \mathbb{E}\left[\frac{S_{n-1}}{n-1} \,|\, \sigma(S_n, X_{n+1}, ...)\right] \\
&= \mathbb{E}\left[\frac{S_{n-1}}{n-1} \,|\, \sigma(S_n)\right] \\
&= \frac{1}{n-1}\left(S_n - \mathbb{E}[X_n \,|\, \sigma(S_n)]\right) \\
&= \frac{1}{n-1}\left(S_n - \frac{S_n}{n}\right) = \frac{S_n}{n} = X_{-n},
\end{aligned}$$

as desired. □

Law of the Iterated Logarithm

Example 5.5.5. *Let $\{X_n\}$ be i.i.d. with the common law $\mathcal{N}(0,1)$. Define $S_n = \sum_1^n X_j$. Then, almost surely,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1, \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1.$$

Define

$$h(n) = \sqrt{2n\log\log n}.$$

It is sufficient to show that, for any $\rho > 1 > \nu$, we have almost surely,

$$\limsup_n \frac{S_n}{h(n)} \leqslant \rho, \quad \limsup_n \frac{S_n}{h(n)} \geqslant \nu.$$

Proof. Upper bound. We know that $\{S_n\}$ is a martingale. For $\theta > 0$, the function $x \mapsto e^{\theta x}$ is convex, thus $\{\exp(\theta S_n)\}$ is a submartingale. By Doob's Maximal Inequality Theorem 5.3.6, we have that, for any $c > 0$,

$$\mathbb{P}\left[\max_{k\leqslant n} S_k \geqslant c\right] = \mathbb{P}\left[\max_{k\leqslant n} \exp(\theta S_k) \geqslant \exp(\theta c)\right] \leqslant \exp(-\theta c)\mathbb{E}[\exp(\theta S_n)] = \exp(-\theta c + \theta^2 n/2).$$

Pick $\theta = c/n$, we have that, for any $c > 0$,

$$\mathbb{P}\left[\max_{k\leqslant n} S_k \geqslant c\right] \leqslant \exp(-c^2/(2n)).$$

Thus,

$$\mathbb{P}\left[\max_{k\leqslant n} S_k \geqslant \rho h(n)\right] \leqslant \exp(-\rho^2 h(n)^2/(2n)) = \exp(-\rho^2 \log\log n) = (\log n)^{-\rho^2}.$$

Fix some $N > 1$, then we have

$$\mathbb{P}\left[\max_{k\leqslant N^m} S_k \geqslant \rho h(N^m)\right] \leqslant (m\log N)^{-\rho^2}.$$

Thus

$$\sum_m \mathbb{P}\left[\max_{k\leqslant N^m} S_k \geqslant \rho h(N^m)\right] < \infty.$$

By Borel-Cantelli Lemma, we have almost surely

$$\max_{k\leqslant N^m} S_k \leqslant \rho h(N^m), \quad \text{for } m \text{ large enough.} \tag{5.5.1}$$

On this event, for $N^m \leqslant n \leqslant N^{m+1}$, we have

$$S_n \leqslant \max_{k \leqslant N^{m+1}} S_k \leqslant \rho h(N^{m+1}) \leqslant \rho h(Nn).$$

Therefore, almost surely,

$$\limsup_n \frac{S_n}{h(n)} \leqslant \rho \lim_n \frac{h(Nn)}{h(n)} = \rho\sqrt{N}.$$

This holds for any $N > 1, \rho > 1$. Let $N \to 1, \rho \to 1$, we have almost surely,

$$\limsup_n \frac{S_n}{h(n)} \leqslant 1.$$

$\square$

Proof. Lower bound. Note that S_n is Gaussian with mean zero and variance n. Thus

$$\mathbb{P}[S_n \geqslant \nu h(n)] = \int_{\nu\sqrt{2\log\log n}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

It is known for standard normal distribution that, for x large,

$$\int_x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \approx x^{-1} \exp(-x^2/2).$$

Therefore

$$\mathbb{P}[S_n \geqslant \nu h(n)] \approx \nu^{-1}(2\log\log n)^{-1/2}(\log n)^{-\nu^2}.$$

Fix some $N > 1$, then we have

$$\mathbb{P}[S(N^{m+1}) - S(N^m) \geqslant \nu h(N^{m+1} - N^m)] \approx \nu^{-1}(2\log(m\log N))^{-1/2}(m\log N)^{-\nu^2}.$$

Therefore,

$$\sum_m \mathbb{P}[S(N^{m+1}) - S(N^m) \geqslant \nu h(N^{m+1} - N^m)] = \infty.$$

By Borel-Cantelli Lemma, we have almost surely

$$S(N^{m+1}) - S(N^m) \geqslant \nu h(N^{m+1} - N^m), \quad i.o.$$

By (5.5.1), we have almost surely

$$S(N^m) \geqslant -\rho h(N^m), \quad \text{for } m \text{ large}.$$

Combining these two, we have almost surely,

$$S(N^{m+1}) \geqslant \nu h(N^{m+1} - N^m) - \rho h(N^m), \quad i.o.$$

Therefore, almost surely

$$\limsup_n \frac{S_n}{h(n)} \geqslant \limsup_m \frac{S(N^m)}{h(N^m)} \geqslant \nu\sqrt{\frac{N-1}{N}} - \frac{\rho}{\sqrt{N}}.$$

This holds for any $N > 1$ and $\nu < 1$. Let $N \to \infty, \nu \to 1$, we have almost surely,

$$\limsup_n \frac{S_n}{h(n)} \geqslant 1.$$

$\square$

The above example is a particular case of the following theorem.

Theorem (Law of the iterated logarithm)**.** *Let $\{X_n\}$ be i.i.d. random variables with mean zero and unit variance. Define $S_n = \sum_{j=1}^n X_j$. Then, almost surely,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1, \quad \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1.$$

Proof. Bonus. □

Let us summarize LLN, CLT and Law of the iterated logarithm here: Suppose $\{X_n\}$ are i.i.d. random variables with zero mean and unit variance. Then

$$\frac{S_n}{n} \to 0, \quad a.s.$$

$$\frac{S_n}{\sqrt{n}} \Longrightarrow \mathcal{N}(0,1)$$

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1, \quad a.s.$$

5.6 Exercises

Exercise 5.6.1.

- *Chebshev's inequality. For $a > 0$, we have*

$$\mathbb{P}[|X| > a \,|\, \mathcal{A}] \leqslant a^{-2}\mathbb{E}[|X|^2 \,|\, \mathcal{A}].$$

- *Cauchy-Schwarz inequality.*

$$\mathbb{E}[XY \,|\, \mathcal{A}]^2 \leqslant \mathbb{E}[X^2 \,|\, \mathcal{A}] \times \mathbb{E}[Y^2 \,|\, \mathcal{A}].$$

Exercise 5.6.2. *Show that if X and Y are random variables with $\mathbb{E}[Y \,|\, \mathcal{A}] = X$ and $\mathbb{E}[X^2] = \mathbb{E}[Y^2] < \infty$, then $X = Y$ a.s.*

Exercise 5.6.3. *Let X, Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{A} \subset \mathcal{F}$ be a sub-σ-field. The random variables X and Y are said to be independent conditionally on $\mathcal{A}$ if for all non-negative measurable functions f and g, we have*

$$\mathbb{E}[f(X)g(Y) \,|\, \mathcal{A}] = \mathbb{E}[f(X) \,|\, \mathcal{A}] \times \mathbb{E}[g(Y) \,|\, \mathcal{A}] \quad a.s.$$

Show that X, Y are independent conditionally on $\mathcal{A}$ if and only if for every non-negative $\mathcal{A}$-measurable random variable Z, and all non-negative measurable functions f and g, we have

$$\mathbb{E}[f(X)g(Y)Z] = \mathbb{E}[f(X)Z\mathbb{E}[g(Y) \,|\, \mathcal{A}]].$$

Exercise 5.6.4. *Let $\{\xi_i\}$ be i.i.d non-negative random variables with $\mathbb{E}[\xi_1] = 1$ and $\mathbb{P}[\xi_1 = 1] < 1$. Set $X_n = \prod_{i=1}^n \xi_i$.*

- *Show that $X_n \to 0$ a.s.*
- *Show that $\frac{1}{n}\log X_n \to c$ a.s. where $c < 0$ is a constant.*

Exercise 5.6.5. *Suppose X and Y are two random variables which are integrable. Assume that $\mathbb{E}[X \,|\, \sigma(Y)] = Y$ and $\mathbb{E}[Y \,|\, \sigma(X)] = X$. Show that $X = Y$ almost surely.*

Exercise 5.6.6. *Let $\{X_n\}$ be a martingale in L^2.*

(1) Show that its increments $\{X_{n+1} - X_n\}$ are pairwise orthogonal, i.e. for all $n \neq m$, we have

$$\mathbb{E}[(X_{n+1} - X_n)(X_{m+1} - X_m)] = 0.$$

(2) Show that $\{X_n\}$ is bounded in L^2 if and only if

$$\sum_n \mathbb{E}[(X_{n+1} - X_n)^2] < \infty.$$

Exercise 5.6.7. *Let $\{X_n\}$ be a simple random walk on $\mathbb{Z}$ starting from $k \in \{0, 1, \ldots, N\}$. Define $\tau = \min\{n : X_n = 0 \text{ or } N\}$.*

(1) Show that $\{X_n\}$ is a martingale and prove that

$$\mathbb{P}[X_\tau = N] = k/N.$$

(2) Show that $\{X_n^2 - n\}$ is a martingale and prove that

$$\mathbb{E}[\tau] = k(N - k).$$

(3) Show that $\{X_n^3 - 3nX_n\}$ is a martingale and prove that

$$\mathbb{E}[\tau | X_\tau = N] = \frac{1}{3}(N^2 - k^2).$$

(4) Compute $\mathbb{E}[\tau^2]$.

Exercise 5.6.8 (YCMC2012)**.** *Suppose that $\{\xi_k : k = 1, 2, ..., n\}$ are i.i.d. random variables with uniform distribution on the interval $[0, 1]$. Let $Y = \max\{\xi_k : 1 \leqslant k \leqslant n\}$.*

(1) What is the joint distribution of (ξ_1, Y)

(2) Evaluate the probability $\mathbb{P}[\xi_1 = Y]$.

(3) Evaluate the conditional expectation $\mathbb{E}[\xi_1 \,|\, \sigma(Y)]$.

Exercise 5.6.9 (YCMC2013)**.** *Let X be an integrable random variable, $\mathcal{G}$ a σ-algebra, and $Y = \mathbb{E}[X \,|\, \mathcal{G}]$. Assume that X and Y have the same distribution.*

(1) Prove that if X is square-integrable, then $X = Y$ a.s. (i.e. X must be $\mathcal{G}$-measurable) ;

(2) Using (1) to prove that for any pair of real numbers a, b with $a < b$, we have $\min\{\max\{X, a\}, b\} = \min\{\max\{Y, a\}, b\}$, and consequently, $X = Y$ a.s.

Exercise 5.6.10 (YCMC2013)**.** *Let X and Y be independent $\mathcal{N}(0, 1)$ random variables.*

(1) Find $\mathbb{E}[X + Y \,|\, X \geqslant 0, Y \geqslant 0]$;

(2) Find the distribution function of $X + Y$ given that $X \geqslant 0$ and $Y \geqslant 0$. (Hint: Using the fact that $U = (X + Y)/\sqrt{2}$ and $V = (X - Y)/\sqrt{2}$ are independent and $\mathcal{N}(0, 1)$ distributed.)

Exercise 5.6.11 (YCMC2014)**.** *Suppose that (X, Y) is a two-dimensional Gaussian with mean $(0, 0)$, variance (σ^2, τ^2) and correlation coefficient ρ. Determine $\mathbb{E}[X \,|\, \sigma(X + Y)]$.*

Exercise 5.6.12 (YCMC2014)**.** *Given two independent random variables X and Y such that X has the uniform law on $[0, 1]$ and $\mathbb{P}[Y = 0] = \mathbb{P}[Y = 1/2] = 1/2$. Show that $W := X + 1/2Y$ has the uniform law on $[0, 1]$ and compute $\mathbb{E}[Y \,|\, \sigma(W)]$.*

Exercise 5.6.13 (YCMC2015). *(a) Let X and Y be two random variables with zero means, unit variances, and correlation ρ. Prove that*

$$\mathbb{E}\left[\max\{X^2, Y^2\}\right] \leqslant 1 + \sqrt{1-\rho^2}.$$

(b) Let X and Y be a two-dimensional Gaussian with means zero, variances σ^2 and τ^2, and correlation ρ. Find the conditional expectation $\mathbb{E}[X \mid \sigma(Y)]$.

Exercise 5.6.14 (YCMC2016). *Suppose $\{X_n\}$ are i.i.d. random variables in L^1. Define $S_n = \sum_{j=1}^{n} X_j$. What is the conditional expectation of S_{n-1} given $\sigma(S_n)$?*

Exercise 5.6.15 (YCMC2017). *Let $\{X_n\}$ be a sequence of non-negative random variables. Let $\{\mathcal{F}_n\}$ be a sequence of increasing σ-algebras. Assume that $\mathbb{E}[X_n \mid \mathcal{F}_n] \to 0$ in probability. Show that $X_n \to 0$ in probability. Is it true reversely? If yes, prove it; if not, give a counterexample.*

Exercise 5.6.16 (YCMC2018). *Suppose that a random vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is distributed as n-dimensional Gaussian $\mathcal{N}(0, \Sigma)$ where Σ is an $n \times n$ positive definite matrix. Let the (i, j) element of Σ^{-1} be ω_{ij} with $1 \leqslant i, j \leqslant n$. For $1 \leqslant i \neq j \leqslant n$, show that, if $\omega_{ij} = 0$, then x_i and x_j are conditionally independent when the other elements of x are given.*

6 Markov chain: finite state space

6.1 Finite Markov chains: introduction

A finite Markov chain is a process which moves among the vertices of a finite set $\mathcal{S}$ which is called state space. The law of the Markov chain is characterized by the transition matrix P of size $|\mathcal{S}| \times |\mathcal{S}|$.

Definition 6.1.1. *A sequence of random variables $\{X_n\}$ is a Markov chain with state space $\mathcal{S}$ and transition matrix P if for all $n \geqslant 0$, and all sequences $(x_0, x_1, ..., x_n, x_{n+1})$ with $x_i \in \mathcal{S}$, we have that*

$$\mathbb{P}[X_{n+1} = x_{n+1} \,|\, X_0 = x_0, ..., X_n = x_n] = \mathbb{P}[X_{n+1} = x_{n+1} \,|\, X_n = x_n] = P(x_n, x_{n+1}).$$

In the above definition,

- the conditional probability of jumping from x to y is the same $P(x, y)$, no matter what sequence $x_0, ..., x_{n-1}$ of states proceeds the current state x;
- the transition matrix P is stochastic:
  - $P(x, y) \geqslant 0$ for all x, y;
  - $\sum_y P(x, y) = 1$ for all x.

  It is clear that the largest eigenvalue of P is one.

Example 6.1.2 (Gambler's ruin)**.** *Consider a gambler betting on the outcome of a sequence of independent fair coin tosses. If head, he gains one dollar. If tail, he loses one dollar. If he reaches a fortune of N dollars, he stops. If his purse is ever empty, he stops. Questions:*

- *What are the probabilities of the two possible fates?*
- *How long will it take for the gambler to arrive at one of the two possible fates?*

The gambler's situation can be modeled by a Markov chain on the state space $\mathcal{S} = \{0, 1, ..., N\}$:

- X_0 : initial money in purse; X_n : the gambler's fortune at time n.
- $\mathbb{P}[X_{n+1} = X_n + 1 \,|\, X_n] = 1/2$ and $\mathbb{P}[X_{n+1} = X_n - 1 \,|\, X_n] = 1/2$.
- The states 0 and N are absorbing.
- τ : the time that the gambler stops.

Following is the transition matrix when $N = 4$:

$$\left[\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1/2 & 0 & 1/2 & 0 & 0 \\ 2 & 0 & 1/2 & 0 & 1/2 & 0 \\ 3 & 0 & 0 & 1/2 & 0 & 1/2 \\ 4 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Example (Example 6.1.2 continued)**.** *Let X_n be the gambler's fortune at time n and let τ be the time required to be absorbed at either 0 or N. Assume that $x_0 = k$ for some $0 \leqslant k \leqslant n$. Then*

$$\mathbb{P}[X_\tau = N] = k/N, \tag{6.1.1}$$

$$\mathbb{E}[\tau] = k(N - k). \tag{6.1.2}$$

Proof of (6.1.1). Let p_k be the probability $\mathbb{P}[X_\tau = N \mid X_0 = k]$. Then we have that, $p_0 = 0, p_N = 1$, and, for $1 \leqslant k \leqslant N-1$,

$$\begin{aligned} p_k &= \mathbb{P}[X_\tau = N \mid X_0 = k] \\ &= \mathbb{P}[X_1 = k-1, X_\tau = N \mid X_0 = k] + \mathbb{P}[X_1 = k+1, X_\tau = N \mid X_0 = k] \\ &= \frac{1}{2}\mathbb{P}[X_\tau = N \mid X_1 = k-1] + \frac{1}{2}\mathbb{P}[X_\tau = N \mid X_1 = k+1] \\ &= \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1}. \end{aligned}$$

There exists a unique solution of p with $p_0 = 0, p_N = 1, p_k = (p_{k-1} + p_{k+1})/2$:

$$p_k = k/N, \quad 0 \leqslant k \leqslant N.$$

□

Proof of (6.1.2). Let m_k be the expectation $\mathbb{E}[\tau \mid X_0 = k]$. Then we have that, $m_0 = 0, m_N = 0$, and, for $1 \leqslant k \leqslant N-1$,

$$m_k = \frac{1}{2}\mathbb{E}[\tau \mid X_1 = k-1] + \frac{1}{2}\mathbb{E}[\tau \mid X_1 = k+1] = \frac{1}{2}(m_{k-1} + 1) + \frac{1}{2}(m_{k+1} + 1).$$

There exists a unique solution of m with $m_0 = 0, m_N = 0, m_k = (m_{k-1} + m_{k+1} + 2)/2$:

$$m_k = k(N-k), \quad 0 \leqslant k \leqslant N.$$

□

Example 6.1.3 (Coupon collecting)**.** *A company issues N different types of coupons. A collector desires a complete set. Question: How many coupons must he obtain so that his collection contains all N types. Assumption: each coupon is equally likely to be each of the N types.*

The collector's situation can be modeled by a Markov chain on the state space $\mathcal{S} = \{0, 1, ..., N\}$:

- $X_0 = 0$; X_n : the number of different types among the collector's first n coupons.
- $\mathbb{P}[X_{n+1} = k+1 \mid X_n = k] = (N-k)/N$ and $\mathbb{P}[X_{n+1} = k \mid X_n = k] = k/N$.
- τ : the first time that the collector obtains all N types.

Following is the transition matrix when $N = 4$:

$$\left[\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1/4 & 3/4 & 0 & 0 \\ 2 & 0 & 0 & 2/4 & 2/4 & 0 \\ 3 & 0 & 0 & 0 & 3/4 & 1/4 \\ 4 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Example (Example 6.1.3 continued)**.** *We have*

$$\mathbb{E}[\tau] = N \sum_{k=1}^{N} \frac{1}{k} \approx N \log N. \tag{6.1.3}$$

Moreover, for any $c > 0$, we have that

$$\mathbb{P}[\tau > N \log N + cN] \leqslant e^{-c+1}. \tag{6.1.4}$$

Proof of (6.1.3). For $1 \leqslant k \leqslant N$, let τ_k be the first time that the collector has k different types. Then $(\tau_{k+1} - \tau_k)$ satisfies geometric distribution:

$$\mathbb{P}[\tau_{k+1} - \tau_k > n] = (k/N)^n.$$

Therefore

$$\mathbb{E}[\tau_{k+1} - \tau_k] = N/(N-k).$$

Thus,

$$\mathbb{E}[\tau] = \mathbb{E}[\tau_N] = \sum_{k=0}^{N-1} \mathbb{E}[\tau_{k+1} - \tau_k] = \sum_{k=1}^{N} N/k.$$

□

Proof of (6.1.4). For $1 \leqslant i \leqslant N$, let A_i be the event that ith type does not appear in the first $N \log N + cN$ coupons. Then

$$\mathbb{P}[\tau > N \log N + cN] = \mathbb{P}[\cup_1^N A_i] \leqslant \sum_1^N \mathbb{P}[A_i] = N\mathbb{P}[A_1].$$

Then let us evaluate $\mathbb{P}[A_1]$:

$$\begin{aligned}
\mathbb{P}[A_1] &\leqslant \left(1 - \frac{1}{N}\right)^{N \log N + cN - 1} \\
&= \exp\left((N \log N + cN - 1)\log(1 - 1/N)\right) \\
&\leqslant \exp\left((N \log N + cN - 1)(-1/N)\right) \leqslant \frac{1}{N} e^{-c+1}.
\end{aligned}$$

Thus,

$$\mathbb{P}[\tau > N \log N + cN] \leqslant e^{-c+1}.$$

□

Random mapping representation

Definition 6.1.4. *A random mapping representation of a transition matrix P on state space $\mathcal{S}$ is a function $f : \mathcal{S} \times \Lambda \to \mathcal{S}$, along with a $\Lambda-$valued random variable Z, satisfying*

$$\mathbb{P}[f(x, Z) = y] = P(x, y).$$

Question: How is it related to Markov chain? Suppose (f, Z) is a random mapping representation, let $\{Z_n\}$ be i.i.d. with common law the same as Z and be independent of X_0. Define $X_n = f(X_{n-1}, Z_n)$ for $n \geqslant 1$. Then $\{X_n\}$ is a Markov chain with transition matrix P.

Theorem 6.1.5. *Every transition matrix on a finite state space has a random mapping representation.*

Proof. Let P be the transition matrix of a Markov chain with state space $\mathcal{S} = \{x_1, x_2, ..., x_N\}$. Take $\Lambda = [0, 1]$, and let Z be a uniform random variable on $[0, 1]$. Set, for $1 \leqslant n, k \leqslant N$,

$$F_{n,k} = \sum_{j=1}^{k} P(x_n, x_j), \quad \text{for } 1 \leqslant n, k \leqslant N;$$

$$f(x_n, z) = x_k, \quad \text{if } F_{n,k-1} < z \leqslant F_{n,k}.$$

Then

$$\mathbb{P}[f(x_n, Z) = x_k] = F_{n,k} - F_{n,k-1} = P(x_n, x_k).$$

□

6.2 Irreducible, aperiodic, stationary, reversible

Irreducible and aperiodic

Definition 6.2.1. *A transition matrix P is called irreducible, if for any $x, y \in \mathcal{S}$, there exists a number n (possibly depending on x, y) such that $P^n(x,y) > 0$.*

Definition 6.2.2. *For any $x \in \mathcal{S}$, define $T(x) = \{n \geqslant 1 : P^n(x,x) > 0\}$. The period of state x is the greatest common divisor of $T(x)$, denoted by $\gcd(T(x))$.*

Lemma 6.2.3. *If P is irreducible, then $\gcd(T(x)) = \gcd(T(y))$ for all $x, y \in \mathcal{S}$. We define this common number to be the period of the chain.*

Proof. Fix two states $x, y \in \mathcal{S}$. Since P is irreducible, there exist integers n, m such that $P^n(x,y) > 0$ and $P^m(y,x) > 0$. Then

$$P^{n+m}(x,x) \geqslant P^n(x,y)P^m(y,x) > 0 \quad \Rightarrow \quad n+m \in T(x);$$

$$P^{n+m}(y,y) \geqslant P^m(y,x)P^n(x,y) > 0 \quad \Rightarrow \quad n+m \in T(y).$$

For any $u \in T(x)$, we have that $P^u(x,x) > 0$; moreover

$$P^{n+m+u}(y,y) \geqslant P^m(y,x)P^u(x,x)P^n(x,y) > 0.$$

Combining the facts that $n+m \in T(y)$ and that $n+m+u \in T(y)$, we see $\gcd(T(y))$ divides u. This holds for any $u \in T(x)$, therefore $\gcd(T(y))$ divides $\gcd(T(x))$. Symmetrically, $\gcd(T(x))$ divides $\gcd(T(y))$. Thus $\gcd(T(y)) = \gcd(T(x))$. $\square$

Definition 6.2.4. *For an irreducible chain, the chain is aperiodic if all states have period 1.*

Example 6.2.5 (Simple Random Walk on Cycles)**.** *Consider a simple random walk on N-cycle.*

- *The walk is irreducible.*
- *When N is odd, the walk is aperiodic.*
- *When N is even, the walk is not aperiodic.*

Proof. On N-cycle, each vertex has two neighbors. At each step, the walk jumps to the left vertex with probability 1/2 and jumps to the right vertex with probability 1/2.

For vertex i and j on the cycle, define $r = |i - j|$, then

$$P^r(i,j) \geqslant (1/2)^r > 0.$$

Thus the walk is irreducible.

For any vertex i, we have

$$P^2(i,i) \geqslant P(i,i+1)P(i+1,i) = 1/4 > 0, \quad \Rightarrow \quad 2 \in T(i);$$

$$P^N(i,i) \geqslant (1/2)^N > 0, \quad \Rightarrow \quad N \in T(i).$$

Thus $\gcd(T(i)) = 1$ when N is odd.

When N is even, the walk always needs even number of steps to come back to the starting position, thus the period is two. $\square$

Theorem 6.2.6. *If P is irreducible and aperiodic, then there exists an integer r such that*

$$P^n(x,y) > 0, \quad \forall x, y \in \mathcal{S}, \forall n \geqslant r.$$

Proof. Recall a standard fact in number theory:

Fact. Any set of non-negative integers, which is closed under addition and which has gcd 1, must contain all but finitely many of the non-negative integers.

First, we show that, for $x \in \mathcal{S}$, there exist $n(x)$ such that the set $T(x)$ contains all $n \geqslant n(x)$. This is true by combining the following two facts and the above fact.

- $T(x) = \{n \geqslant 1 : P^n(x,x) > 0\}$ is closed under addition. For any $n, m \in T(x)$, we have

$$P^{n+m}(x,x) \geqslant P^n(x,x)P^m(x,x) > 0, \quad \Rightarrow \quad n+m \in T(x).$$

- $T(x)$ has gcd 1, since this is an aperiodic chain.

Second, we show that, for $x \in \mathcal{S}$, there exists $n'(x)$ such that $P^n(x,y) > 0$ for all $y \in \mathcal{S}$ and all $n \geqslant n'(x)$. We have the following two observations.

- By the first step, there exists $n(x)$ such that $P^n(x,x) > 0$ for all $n \geqslant n(x)$.
- Since the chain is irreducible, for any $y \in \mathcal{S}$, there exists $r = r(x,y)$ such that $P^r(x,y) > 0$.

Combining these two facts, we have that, for any $m \geqslant n(x) + r$,

$$P^m(x,y) \geqslant P^{m-r}(x,x)P^r(x,y) > 0.$$

Define

$$n'(x) = n(x) + \max_y r(x,y),$$

and it satisfies the desired property.

Finally, define

$$N = \max_x n'(x),$$

and it satisfies the property in the conclusion. □

Stationary distribution

Consider a Markov chain with state space $\mathcal{S}$ and transition matrix P. Recall that

$$\mathbb{P}[X_{n+1} = y \,|\, X_n = x] = P(x,y).$$

We introduce the following notations:

- μ_0 : the distribution of X_0;
- μ_n : the distribution of X_n.

Then we have that

$$\mu_{n+1} = \mu_n P, \quad \mu_n = \mu_0 P^n, \quad \mathbb{E}[f(X_n)] = \mu_0 P^n f.$$

Definition 6.2.7. *We call a probability measure π stationary if $\pi = \pi P$.*

If π is stationary and the initial measure μ_0 equals π, then $\mu_n = \pi$, for all n.

Example 6.2.8 (Simple Random Walk on Graph)**.** *A graph $G = (V, E)$ consists of a vertex set V and an edge set E:*

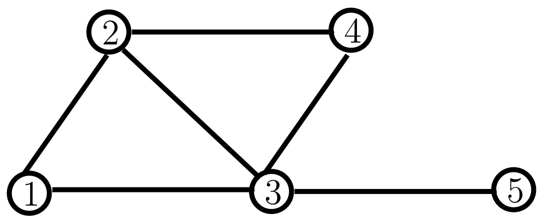
- *V : set of vertices*
- *E : set of pairs of vertices*

- *When* $(x, y) \in E$, *we write* $x \sim y : x$ *and* y *are joined by an edge. We say* y *is a neighbor of* x.
- *For* $x \in V$, $\deg(x)$: *the number of neighbors of* x.

Given a graph $G = (V, E)$, *we define simple random walk on* G *to be the Markov chain with state space* V *and transition matrix:*

$$P(x, y) = \begin{cases} 1/\deg(x), & \textit{if } y \sim x; \\ 0, & \textit{else.} \end{cases}$$

For the random walk on graph, it is the Markov chain such that, when the chain is at vertex x, it examines all its neighbors, picks one uniformly at random, and jumps to the chosen vertex. For the following graph, we have that $V = \{1, 2, 3, 4, 5\}$ and $E = \{(1, 2), (1, 3), (2, 3), (2, 4), (3, 4), (3, 5)\}$.



The corresponding transition matrix for the simple random walk on this graph is the following:

$$\left[\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 1/2 & 1/2 & 0 & 0 \\ 2 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 3 & 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 4 & 0 & 1/2 & 1/2 & 0 & 0 \\ 5 & 0 & 0 & 1 & 0 & 0 \end{array}\right].$$

Example (Example 6.2.8 continued)**.** *Define*

$$\pi(x) = \frac{\deg(x)}{2|E|}, \quad \forall x \in V.$$

Then π *is a stationary distribution for the simple random walk on graph* G.

Proof. First, π is a probability measure:

$$\sum_x \deg(x) = 2|E|,$$

since each edge is counted twice in the LHS.

Next, π is stationary:

$$(\pi P)(x) = \sum_{z \in V} \pi(z) P(z, x) = \sum_{z \in V} \frac{\deg(z)}{2|E|} \frac{1}{\deg(z)} \mathbb{1}_{\{z \sim x\}} = \frac{\deg(x)}{2|E|}.$$

□

Time-reversal of Markov chains

Definition 6.2.9. *We say that a probability measure* π *on* $\mathcal{S}$ *satisfies detailed balance equations if*

$$\pi(x) P(x, y) = \pi(y) P(y, x), \quad \forall x, y \in \mathcal{S}.$$

Note that any distribution π satisfying the detailed balance equations is stationary for P.

Suppose that a probability measure π satisfies the detailed balance equations. Then, for any sequence $x_0, ..., x_n$, we have

$$\pi(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) = \pi(x_n)P(x_n, x_{n-1}) \cdots P(x_1, x_0).$$

Or equivalently,

$$\mathbb{P}_\pi[X_0 = x_0, X_1 = x_1, ..., X_n = x_n] = \mathbb{P}_\pi[X_0 = x_n, X_1 = x_{n-1}, ..., X_n = x_0].$$

In other words, if the Markov chain has initial distribution π, then the distribution of $(X_0, X_1, ..., X_n)$ is the same as its time-reversal $(X_n, X_{n-1}, ..., X_0)$. For this reason, a chain satisfying detailed balance equations is called reversible.

Example 6.2.10 (Birth-and-Death Chain)**.** *A birth-and-death chain has state space $\mathcal{S} = \{0, 1, ..., N\}$. In one step the state can increase or decrease by at most one. The current state can be thought of as the size of some population; in a single step of the chain, there can be at most one birth or death. Then transition probabilities can be specified by $\{(p_k, r_k, q_k), 0 \leqslant k \leqslant N\}$ where $p_k + r_k + q_k = 1$ for each k and*

- *p_k is the probability of moving from k to $k+1$ when $0 \leqslant k < N$; $p_N = 0$;*
- *q_k is the probability of moving from k to $k-1$ when $0 < k \leqslant N$; $q_0 = 0$;*
- *r_k is the probability of remaining at k when $0 \leqslant k \leqslant N$.*

Every birth-and-death chain is reversible.

Proof. A measure $\tilde{\pi}$ on $\mathcal{S}$ satisfies detailed balance equations if and only if, for all $1 \leqslant k \leqslant N$

$$\tilde{\pi}(k)P(k, k-1) = \tilde{\pi}(k-1)P(k-1, k), \quad \text{or} \quad \frac{\tilde{\pi}(k)}{\tilde{\pi}(k-1)} = \frac{p_{k-1}}{q_k}.$$

Define

$$\tilde{\pi}(0) = 1, \quad \tilde{\pi}(k) = \Pi_{n=1}^k p_{n-1}/q_n, \quad 1 \leqslant k \leqslant N.$$

Set

$$\pi(k) = \frac{\tilde{\pi}(k)}{\sum_0^N \tilde{\pi}(n)}, \quad 0 \leqslant k \leqslant N.$$

Then π is a probability measure on $\mathcal{S}$ satisfying detailed balance equations. ☐

Proposition 6.2.11. *Let $\{X_n\}$ be an irreducible Markov chain with transition matrix P and stationary distribution π. Define $\hat{P}$ to be*

$$\hat{P}(x, y) = \pi(y)P(y, x)/\pi(x).$$

Then $\hat{P}$ is a stochastic matrix and π is stationary for $\hat{P}$. Let $\{\hat{X}_n\}$ be a Markov chain with transition matrix $\hat{P}$, then, for any $x_0, x_1, ..., x_n$,

$$\mathbb{P}_\pi[X_0 = x_0, X_1 = x_1, ..., X_n = x_n] = \mathbb{P}_\pi[\hat{X}_0 = x_n, \hat{X}_1 = x_1, ..., \hat{X}_n = x_0].$$

For this reason, we call $\hat{X}$ the time-reversal of X.

Proof. First, we show that $\hat{P}$ is stochastic. For $x \in \mathcal{S}$,

$$\sum_y \hat{P}(x, y) = \sum_y \pi(y)P(y, x)/\pi(x) = 1. \qquad \text{(Since } \pi \text{ is stationary for } P\text{)}$$

Second, we show that π is stationary for $\hat{P}$.

$$\sum_y \pi(y)\hat{P}(y,x) = \sum_y \pi(x)P(x,y) = \pi(x). \qquad \text{(Since } P \text{ is stochastic)}$$

Finally,

$$\begin{aligned}\mathbb{P}_\pi[X_0 = x_0, X_1 = x_1, ..., X_n = x_n] &= \pi(x_0)P(x_0,x_1)\cdots P(x_{n-1},x_n)\\ &= \pi(x_n)\hat{P}(x_n,x_{n-1})\cdots \hat{P}(x_1,x_0)\\ &= \mathbb{P}_\pi[\hat{X}_0 = x_n, \hat{X}_1 = x_1, ..., \hat{X}_n = x_0].\end{aligned}$$

□

If a Markov chain with transition matrix P is reversible, then $\hat{P} = P$ and $\hat{X}$ has the same law as X.

6.3 Stationary measure

Definition 6.3.1. *Let $\{X_n\}$ be a Markov chain on $\mathcal{S}$. For $x \in \mathcal{S}$, define*

$$\tau_x = \inf\{n \geqslant 0 : X_n = x\}, \quad \tau_x^+ = \inf\{n \geqslant 1 : X_n = x\}.$$

We call τ_x the hitting time for x, and τ_x^+ the first return time when $X_0 = x$.

The goal of this section is the following existence and uniqueness of stationary measure of irreducible Markov chains.

Theorem 6.3.2. *Suppose P is irreducible, then there exists a unique probability measure π such that $\pi = \pi P$. Moreover, for all $x \in \mathcal{S}$,*

$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x^+]} > 0.$$

Lemma 6.3.3. *Suppose that P is irreducible. Then, for any $x, y \in \mathcal{S}$, we have*

$$\mathbb{E}_x[\tau_y^+] < \infty.$$

Proof. Since P is irreducible, for any $x, y \in \mathcal{S}$, there exists $r(x,y)$ such that $P^{r(x,y)}(x,y) > 0$. Define

$$\epsilon = \min\{P^{r(x,y)}(x,y) : x, y \in \mathcal{S}\}, \quad R = \max\{r(x,y) : x, y \in \mathcal{S}\}.$$

Then, for any value of X_n, the probability of hitting state y at a time between n and $n+R$ is at least ϵ. Thus,

$$\mathbb{P}_x[\tau_y^+ > R] \leqslant 1 - \epsilon, \quad \mathbb{P}_x[\tau_y^+ > (k+1)R] \leqslant (1-\epsilon)\mathbb{P}_x[\tau_y^+ > kR].$$

Repeating this inequality, we have that, for any $k \geqslant 1$,

$$\mathbb{P}_x[\tau_y^+ > kR] \leqslant (1-\epsilon)^k.$$

Therefore,

$$\mathbb{E}_x[\tau_y^+] = \sum_n \mathbb{P}_x[\tau_y^+ > n] \leqslant R\sum_k \mathbb{P}_x[\tau_y^+ > kR] \leqslant R\sum_k (1-\epsilon)^k \leqslant R/\epsilon < \infty.$$

□

Proof of Theorem 6.3.2–existence. Fix a state $z \in \mathcal{S}$, we will examine the time that the chain spends at each state in between two visits to z. Define, for $x \in \mathcal{S}$,

$$\tilde{\pi}(x) = \mathbb{E}_z[\#\text{visits to } x \text{ before returning to } z].$$

Note that, we can write $\tilde{\pi}(x)$ in the following way

$$\tilde{\pi}(x) = \sum_{n=0}^{\infty} \mathbb{P}_z[X_n = x, n < \tau_z^+].$$

We claim that $\tilde{\pi} = \tilde{\pi}P$. For any $x \in \mathcal{S}$,

$$\begin{aligned}
\tilde{\pi}P(x) &= \sum_{w \in \mathcal{S}} \tilde{\pi}(w)P(w, x) \\
&= \sum_{w \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}_z[X_n = w, n < \tau_z^+]P(w, x) \\
&= \sum_{w \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}_z[X_n = w, X_{n+1} = x, n+1 \leqslant \tau_z^+] \\
&= \sum_{n=0}^{\infty} \mathbb{P}_z[X_{n+1} = x, n+1 \leqslant \tau_z^+] = \sum_{n=1}^{\infty} \mathbb{P}_z[X_n = x, n \leqslant \tau_z^+].
\end{aligned}$$

In short, we need to compare the following two terms

$$\tilde{\pi}P(x) = \sum_{n=1}^{\infty} \mathbb{P}_z[X_n = x, n \leqslant \tau_z^+], \quad \tilde{\pi}(x) = \sum_{n=0}^{\infty} \mathbb{P}_z[X_n = x, n < \tau_z^+].$$

There are two different cases: $x \neq z$ or $x = z$. If $x \neq z$, we have that

$$\tilde{\pi}P(x) = \sum_{n=1}^{\infty} \mathbb{P}_z[X_n = x, n < \tau_z^+] = \sum_{n=0}^{\infty} \mathbb{P}_z[X_n = x, n < \tau_z^+] = \tilde{\pi}(x).$$

If $x = z$, we have that

$$\tilde{\pi}(x) = \sum_{n=1}^{\infty} \mathbb{P}_z[\tau_z^+ = n] = 1 = \tilde{\pi}(z).$$

In any case, we have $\tilde{\pi}P(x) = \tilde{\pi}(x)$.

To make $\tilde{\pi}$ a probability measure, we need to normalize it by its total mass:

$$\sum_x \tilde{\pi}(x) = \mathbb{E}_z[\tau_z^+].$$

Define, for any $x \in \mathcal{S}$,

$$\pi(x) = \tilde{\pi}(x)/\mathbb{E}_z[\tau_z^+].$$

Then the probability measure π is a stationary distribution. □

From the proof of Theorem 6.3.2–existence, we do not know whether the measure π depends on the choice of state z. By the construction of π, we have that

$$\pi(z) = 1/\mathbb{E}_z[\tau_z^+].$$

We will show that there is a unique stationary measure. After we show the uniqueness, we could conclude that, for all $x \in \mathcal{S}$,

$$\pi(x) = 1/\mathbb{E}_x[\tau_x^+].$$

Recall that a measure μ on $\mathcal{S}$ is stationary if $\mu P = \mu$. The corresponding notion for functions on $\mathcal{S}$ is harmonic.

Definition 6.3.4. *A function f on $\mathcal{S}$ is harmonic if $f = Pf$.*

Lemma 6.3.5. *Suppose P is irreducible. Then any harmonic function f on $\mathcal{S}$ has to be constant.*

Proof. Since $\mathcal{S}$ is finite, there must be a state x_0 such that $f(x_0) = \max\{f(x) : x \in \mathcal{S}\}$. Denote the value $f(x_0)$ by M.

For any state z such that $P(x_0, z) > 0$, if $f(z) < M$, we have

$$f(x_0) = \sum_x P(x_0, x)f(x) < \sum_x P(x_0, x)M = M,$$

which is a contradiction. Thus, we must have $f(z) = M$ provided $P(x_0, z) > 0$.

For any state z, since P is irreducible, there exists a sequence $x_0, x_1, ..., x_n = z$ such that $P(x_j, x_{j+1}) > 0$ for $j = 0, ..., n-1$. Repeating the same argument as above tells us that

$$f(x_0) = f(x_1) = \cdots = f(x_n) = f(z) = M.$$

Therefore f is constant. ☐

Proof of Theorem 6.3.2–uniqueness. From Lemma 6.3.5, we know that the kernel of $P - I$ has dimension one. Therefore, the row-vector equation $\mu = \mu P$ also has a one-dimensional space of solutions; and this space contains only one vector whose entries sum to one.

From the proof of existence, we know that $\pi = \tilde{\pi}/\mathbb{E}_z[\tau_z^+]$ is stationary. Note that the definition of $\tilde{\pi}$ depends on z, but since there is a unique stationary distribution, the measure π does not depend on the choice of z. In particular, for all x, we have $\pi(x) = 1/\mathbb{E}_x[\tau_x^+]$. ☐

Theorem 6.3.6 (Ergodic Theorem)**.** *Let f be a real-valued function defined on $\mathcal{S}$. If $\{X_n\}$ is an irreducible Markov chain with stationary distribution π, then for any starting distribution μ, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n} f(X_j) = \pi(f), \quad \mathbb{P}_\mu - a.s.$$

In particular,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n} \mathbb{1}_{\{X_j = x\}} = \pi(x), \quad \mathbb{P}_\mu - a.s.$$

Proof. Since any probability measure μ is a linear combination of Dirac masses $\mu = \sum_x \mu(x)\delta_x$, it suffices to show the conclusion for $\mu = \delta_x$. In other words, the chain starts at x. Define the sequence of return times: for $k \geqslant 1$

$$\tau_0 = 0, \quad \tau_1 = \tau_x^+, \quad \tau_{k+1} = \min\{n > \tau_k : X_n = x\}.$$

Consider one block $(X_{\tau_k}, X_{\tau_k+1}, ..., X_{\tau_{k+1}-1})$, define

$$Y_k = \sum_{j=\tau_k}^{\tau_{k+1}-1} f(X_j).$$

By Markov property, the random variables $\{Y_k\}$ are i.i.d. By Strong Law of Large Numbers, we have

$$\frac{1}{n}\sum_{0}^{n-1} Y_k \to \mathbb{E}_x[Y_0], \quad \mathbb{P}_x - a.s.$$

Since $\tau_n = \sum_0^{n=1}(\tau_{j+1} - \tau_j)$, by Strong Law of Large Numbers again, we have

$$\frac{1}{n}\tau_n \to \mathbb{E}_x[\tau_x^+], \quad \mathbb{P}_x - a.s.$$

Therefore, we have

$$\frac{\sum_0^{\tau_n-1} f(X_j)}{\tau_n} \to \frac{\mathbb{E}_x[Y_0]}{\mathbb{E}_x[\tau_x^+]}, \quad \mathbb{P}_x - a.s.$$

Note that

$$\mathbb{E}_x[Y_0] = \mathbb{E}_x\left[\sum_{j=0}^{\tau_1-1} f(X_j)\right] = \sum_y f(y)\mathbb{E}_x\left[\sum_{j=0}^{\tau_1-1} \mathbb{1}_{\{X_j=y\}}\right] = \sum_y f(y)\tilde{\pi}(y),$$

where $\tilde{\pi}(y) = \mathbb{E}_x[\#\text{visits to } y \text{ before } \tau_x^+]$. Since $\tilde{\pi}(y) = \pi(y)\mathbb{E}_x[\tau_x^+]$, we have that

$$\mathbb{E}_x[Y_0] = \sum_y f(y)\pi(y)\mathbb{E}_x[\tau_x^+] = \pi(f)\mathbb{E}_x[\tau_x^+].$$

Therefore,

$$\frac{\sum_0^{\tau_n-1} f(X_j)}{\tau_n} \to \pi(f), \quad \mathbb{P}_x - a.s. \tag{6.3.1}$$

The goal is to show the following

$$\frac{\sum_0^{m-1} f(X_j)}{m} \to \pi(f), \quad \mathbb{P}_x - a.s. \tag{6.3.2}$$

It remains to derive from (6.3.1) to (6.3.2). Denote by $M = \max_x |f(x)|$ and we may assume $\pi(f) = 0$. For large m, suppose $\tau_n \leqslant m < \tau_{n+1}$. Then we have

$$\frac{|\sum_0^{m-1} f(X_j)|}{m} \leqslant \frac{\left|\sum_0^{\tau_n-1} f(X_j)\right| + (m-\tau_n)M}{m} \leqslant \frac{\left|\sum_0^{\tau_n-1} f(X_j)\right|}{\tau_n} + \left(\frac{\tau_{n+1}}{\tau_n} - 1\right)M.$$

Plugging in (6.3.1) and $\tau_{n+1}/\tau_n \to 1$, we obtain (6.3.2). $\square$

6.4 The convergence theorem

Theorem 6.4.1. *Suppose that P is irreducible, aperiodic, with stationary distribution π. Then there exist constants $\alpha \in (0,1)$ and $C > 0$ such that*

$$\max_{x\in\mathcal{S}} \|P^n(x,\cdot) - \pi\|_{TV} \leqslant C\alpha^n \quad \forall n \geqslant 1.$$

Proof. Define Π to be the matrix with $|\mathcal{S}|$ rows, each of which is the row vector π. It is clear that $\Pi P = P\Pi = \Pi$, and $\Pi^2 = \Pi$. We only need to show that

$$\|P^n - \Pi\| \leqslant C\alpha^n.$$ [5]

Since P is irreducible and aperiodic, there exists r such that

$$P^r(x,y) > 0, \quad \forall x,y \in \mathcal{S}.$$

Thus, for sufficiently small $\delta > 0$ we have

$$P^r(x,y) \geqslant \delta\pi(y), \quad \forall x,y \in \mathcal{S}.$$

Define matrix Q such that $P^r = \delta\Pi + (1-\delta)Q$. Then we can check that

$$Q \text{ is stochastic}, \quad Q\Pi = \Pi Q = \Pi.$$

[5] Here we use distance between matrices: $\|A - B\| = \sum_{i,j} |A(i,j) - B(i,j)|$. Note that $\|\mu - \nu\| = 2\|\mu - \nu\|_{TV}$.

Denote $1-\delta$ by θ, then $P^{nr} = (1-\theta^n)\Pi + \theta^n Q^n$. Thus

$$\|P^{nr} - \Pi\| \leqslant \theta^n \|\Pi - Q^n\| \leqslant C\theta^n,$$

where $C = 2|\mathcal{S}| \times |\mathcal{S}|$. For $0 \leqslant j < r$, we have $P^{nr+j} = P^{nr}P^j = (1-\theta^n)\Pi P^j + \theta^n Q^n P^j = (1-\theta^n)\Pi + \theta^n Q^n P^j$. Therefore,

$$\|P^{nr+j} - \Pi\| = \theta^n \|\Pi - Q^n P^j\| \leqslant C\theta^n.$$

□

Example (Example 6.2.5 continued). *Suppose $\{X_n\}$ is a simple random walk on N-cycle. It is clear that the stationary measure π is uniform over the cycle. Recall that the walk is irreducible. By ergodic theorem, we have, for any $x \in \mathcal{S}$:*

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_n = x\}} \to \frac{1}{N}, \quad a.s.; \quad and \quad \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}[X_n = x] \to \frac{1}{N}. \tag{6.4.1}$$

When N is odd, the walk is aperiodic, and thus for any $x \in \mathcal{S}$:

$$\mathbb{P}[X_n = x] \to \frac{1}{N}. \tag{6.4.2}$$

When N is even. Suppose $x_0, x \in \mathcal{S}$ and the distance between these two points is even, then we have

$$\mathbb{P}_{x_0}[X_{2n} = x] \to \frac{2}{N}, \quad \mathbb{P}_{x_0}[X_{2n+1} = x] = 0.$$

Thus, the conclusion (6.4.2) *can not hold in this case; whereas, the conclusion* (6.4.1) *is still true.*

6.5 Exercises

Exercise 6.5.1. *A graph G is* connected *when, for two vertices x and y of G, there exists a sequence of vertices $x_0, x_1, \ldots, x_k$ such that $x_0 = x, x_k = y$, and $x_i \sim x_{i+1}$ for $0 \leqslant i \leqslant k-1$. Show that simple random walk on G is irreducible if and only if G is connected.*

Exercise 6.5.2. *Let P be the transition matrix of a Markov chain with state space $\mathcal{S}$ and let μ and ν be any two distributions on $\mathcal{S}$. Prove that*

$$\|\mu P - \nu P\|_{\mathrm{TV}} \leqslant \|\mu - \nu\|_{\mathrm{TV}}.$$

(This in particular shows that $\|\mu P^{t+1} - \pi\|_{\mathrm{TV}} \leqslant \|\mu P^t - \pi\|_{\mathrm{TV}}$, that is, advancing the chain can only move it closer to stationary.)

Exercise 6.5.3. *A professor has n umbrellas, of which initially $k \in (0, n)$ are at his office and $n-k$ are at his home. Every day, the professor walks to the office in the morning and returns home in the evening. In each trip, he takes an umbrella with him only if it is raining. Assume that in every trip between home and office or back, the chance of rain is $p \in (0, 1)$, independently of other trips.*

(1) Asymptotically, in what fraction of his trips does the professor get wet?

(2) Determine the expected number of trips until all n umbrellas at the same location.

(3) Determine the expected number of trips until the professor gets wet.

Exercise 6.5.4 (YCMC2016). *Consider the numbers 1, 2, ..., 12 written around a ring as they usually are on a clock. A random walker starts at 12 and at each step moves at random to one of its two nearest neighbors (with probability half-half). What is the probability that she will visit all the other numbers before her first returning back to 12?*

Exercise 6.5.5 (YCMC2016). *A boy tries to collect some special tennis cards. There are 100 different types. Each time he put 1 yuan into the card machine, he will randomly get a tennis card. The type of the card is uniformly distributed. Let T be the total money he will spend to collect all different types of cards. What is the expectation and variance of T?*

7 Markov chain: countable state space

7.1 Recurrence and positive recurrence

In this section, we will consider Markov chain on countable state space $\mathcal{S}$. The following notions for Markov chain on finite state space can be generalized to the countable space.

Definition 7.1.1. *A sequence of random variables $\{X_n\}$ is a Markov chain with state space $\mathcal{S}$ and transition matrix P if for all $n \geqslant 0$, and all sequences $(x_0, x_1, ..., x_n, x_{n+1})$ with $x_i \in \mathcal{S}$, we have that*

$$\mathbb{P}[X_{n+1} = x_{n+1} \,|\, X_0 = x_0, ..., X_n = x_n] = \mathbb{P}[X_{n+1} = x_{n+1} \,|\, X_n = x_n] = P(x_n, x_{n+1}).$$

For the Markov chain on $\mathcal{S}$ with transition matrix P, the stationary distribution and irreducibility are defined in the same way as before.

- *A measure π on $\mathcal{S}$ is a stationary distribution if $\pi = \pi P$ and π has unit total mass.*
- *The transition matrix P is irreducible if for any $x, y \in \mathcal{S}$, there exists n such that $P^n(x, y) > 0$.*

Definition 7.1.2. *For the Markov chain $\{X_n\}$, define hitting time and first return time: for $x \in \mathcal{S}$*

$$\tau_x = \min\{n \geqslant 0 : X_n = x\}, \quad \tau_x^+ = \min\{n \geqslant 1 : X_n = x\}.$$

We say a state $x \in \mathcal{S}$ is recurrent if

$$\mathbb{P}_x[\tau_x^+ < \infty] = 1.$$

Otherwise, we say x is transient.

If $\mathcal{S}$ is finite and P is irreducible, every state is recurrent by Lemma 6.3.3. However, when $\mathcal{S}$ is infinite countable, we have two different cases: recurrent or transient.

Lemma 7.1.3. *Suppose that P is irreducible. The following two conditions are equivalent.*

(1) $\mathbb{P}_x[\tau_x^+ < \infty] = 1$ *for some* $x \in \mathcal{S}$

(2) $\mathbb{P}_x[\tau_y^+ < \infty] = 1$ *for all* $x, y \in \mathcal{S}$

Proof. Suppose that $\mathbb{P}_{x_0}[\tau_{x_0}^+ < \infty] = 1$ for some $x_0 \in \mathcal{S}$. First, we show that, for any $y \neq x_0$, we have $\mathbb{P}_{x_0}[\tau_y < \infty] = 1$. By irreducibility, we know that $q := \mathbb{P}_{x_0}[\tau_y < \tau_{x_0}^+ < \infty] > 0$. Thus

$$p := \mathbb{P}_{x_0}[\tau_y < \infty] = \mathbb{P}_{x_0}[\tau_y < \tau_{x_0}^+ < \infty] + \mathbb{P}_{x_0}[\tau_{x_0}^+ < \tau_y < \infty] = q + (1-q)p.$$

Therefore, $q(1-p) = 0$. Since $q > 0$, we have $p = 1$.

Second, we show that, for any $x \neq x_0$, we have $\mathbb{P}_x[\tau_{x_0} < \infty] = 1$. By irreducibility, we know that $q := \mathbb{P}_{x_0}[\tau_x < \tau_{x_0}^+ < \infty] > 0$. Thus, by Markov property,

$$q = \mathbb{P}_{x_0}[\tau_x < \tau_{x_0}^+ < \infty] = q\mathbb{P}_x[\tau_{x_0} < \infty].$$

Since $q > 0$, we have $\mathbb{P}_x[\tau_{x_0} < \infty] = 1$.

Finally, for any $x, y \in \mathcal{S}$, define $\tau_{x_0 y} = \min\{n \geqslant \tau_{x_0} : X_n = y\}$, we have

$$\mathbb{P}_x[\tau_y < \infty] \geqslant \mathbb{P}_x[\tau_{x_0} < \infty, \tau_{x_0 y} < \infty] = \mathbb{P}_x[\tau_{x_0} < \infty]\mathbb{P}_{x_0}[\tau_y < \infty] = 1.$$

□

From this lemma, we know that, for an irreducible chain, a single state is recurrent if and only if all states are recurrent. For this reason, an irreducible Markov chain can be classified as either recurrent or transient.

Definition 7.1.4. *A state x is positive recurrent if $\mathbb{E}_x[\tau_x^+] < \infty$.*

Lemma 7.1.5. *Suppose that P is irreducible. The following two conditions are equivalent.*

(1) $\mathbb{E}_x[\tau_x^+] < \infty$ for some $x \in \mathcal{S}$.

(2) $\mathbb{E}_x[\tau_y^+] < \infty$ for all $x, y \in \mathcal{S}$.

Proof. Suppose that $\mathbb{E}_{x_0}[\tau_{x_0}^+] < \infty$ for some x_0. First, we show that, for $x \neq x_0$, we have $\mathbb{E}_x[\tau_{x_0}] < \infty$.

$$\infty > \mathbb{E}_{x_0}[\tau_{x_0}^+] \geqslant \mathbb{E}_{x_0}\left[\tau_{x_0}^+ \mathbb{1}_{\{\tau_x < \tau_{x_0}^+\}}\right] \geqslant \mathbb{E}_{x_0}\left[\left(\tau_{x_0}^+ - \tau_x\right) \mathbb{1}_{\{\tau_x < \tau_{x_0}^+\}}\right] = \mathbb{P}_{x_0}[\tau_x < \tau_{x_0}^+]\mathbb{E}_x[\tau_{x_0}].$$

Since $\mathbb{P}_{x_0}[\tau_x < \tau_{x_0}^+] > 0$ (by irreducibility), we have $\mathbb{E}_x[\tau_{x_0}] < \infty$.

Second, we show that, for $y \neq x_0$, we have $\mathbb{E}_{x_0}[\tau_y] < \infty$. Define

$$\tau_0 = 0, \quad \tau_1 = \tau_{x_0}^+, \quad \tau_{k+1} = \min\{n > \tau_k : X_n = x_0\}.$$

By irreducibility, we have $q := \mathbb{P}_{x_0}[\tau_1 < \tau_y] < 1$; moreover, $\mathbb{P}_{x_0}[\tau_k < \tau_y] = q^k$. Thus

$$\begin{aligned}
\mathbb{E}_{x_0}[\tau_y] &= \sum_k \mathbb{E}_{x_0}\left[\tau_y \mathbb{1}_{\{\tau_k < \tau_y < \tau_{k+1}\}}\right] \\
&\leqslant \sum_k \mathbb{E}_{x_0}\left[\tau_{k+1} \mathbb{1}_{\{\tau_k < \tau_y < \tau_{k+1}\}}\right] \\
&= \sum_k \mathbb{E}_{x_0}\left[(\tau_{k+1} - \tau_k) \mathbb{1}_{\{\tau_k < \tau_y\}}\right] \\
&= \sum_k \mathbb{E}_{x_0}[\tau_1]\mathbb{P}_{x_0}[\tau_k < \tau_y] \\
&= \sum_k \mathbb{E}_{x_0}[\tau_1] q^k < \infty.
\end{aligned}$$

Finally, for any x, y, we have

$$\mathbb{E}_x[\tau_y^+] \leqslant \mathbb{E}_x[\tau_{x_0} + \tau_{x_0 y}] = \mathbb{E}_x[\tau_{x_0}] + \mathbb{E}_{x_0}[\tau_y] < \infty.$$

□

Therefore, for an irreducible chain, a single state is positive recurrent if and only if all states are positive recurrent. For this reason, an irreducible recurrent Markov chain can be classified as either positive recurrent or else which we call null recurrent.

Example 7.1.6. *Simple random walk on $\mathbb{Z}$ is null recurrent.*

Proof. Denote $\mathbb{E}_1[\tau_0]$ by α. Note that

$$\alpha = \frac{1}{2} + \frac{1}{2}(1 + \mathbb{E}_2[\tau_0]) = 1 + \frac{1}{2}\mathbb{E}_2[\tau_0],$$

where

$$\mathbb{E}_2[\tau_0] = \mathbb{E}_2[\tau_1] + \mathbb{E}_1[\tau_0] = 2\alpha.$$

Therefore $\alpha = 1 + \alpha$ and α has to be infinite. Furthermore, $\mathbb{E}_0[\tau_0^+] = \mathbb{E}_1[\tau_0] = \infty$. See another proof in Section 7.2. □

Theorem 7.1.7. *An irreducible Markov chain is positive recurrent if and only if there exists a probability measure π on $\mathcal{S}$ such that $\pi = \pi P$.*

Proof of Theorem 7.1.7–positive recurrence implies stationary distribution. Fix x_0, define

$$\pi(x) = \mathbb{E}_{x_0}\left[\#\text{visits to } x \text{ before } \tau_{x_0}^+\right] / \mathbb{E}_{x_0}\left[\tau_{x_0}^+\right].$$

We could show that π is a stationary distribution in the same way as in finite state space case. ☐

Proof of Theorem 7.1.7–stationary distribution implies positive recurrence. Suppose that there exists a probability measure π such that $\pi = \pi P$. First, we show that $\pi(x) > 0$ for all $x \in \mathcal{S}$. Assume that $\pi(x_0) = 0$ for some x_0. Since $\pi = \pi P$, we have

$$\pi(x_0) = \sum_y \pi(y) P(y, x_0).$$

Thus, combining $\pi(x_0) = 0$ and $P(y, x_0) > 0$, we obtain that $\pi(y) = 0$.

By irreducibility, for any $z \in \mathcal{S}$, there exists sequence $y_0 = z, y_1, ..., y_n = x_0$ such that $P(y_j, y_{j+1}) > 0$, thus $\pi(z) = 0$ by the above analysis. Therefore, $\pi(z) = 0$ for all z, contradicts with the fact that π is a probability measure.

Second, we show that the chain is recurrent, i.e. we will show that $\mathbb{P}_x[\tau_x^+ < \infty] = 1$ for some fixed x. For $n \geqslant 0$, define

$$\alpha(n) = \mathbb{P}_\pi[X_n = x, X_m \neq x, \forall m > n].$$

On the one hand,

$$\alpha(n) = \mathbb{P}_\pi[X_n = x]\mathbb{P}_x[\tau_x^+ = \infty] = \pi(x)\mathbb{P}_x[\tau_x^+ = \infty].$$

On the other hand, the events $\{X_n = x, X_m \neq x, \forall m > n\}$ are disjoint for different n's. Thus $\sum_n \alpha(n) \leqslant 1$. Combining these two facts, we have $\mathbb{P}_x[\tau_x^+ = \infty] = 0$.

Third, we show that the time reversal of $\{X_n\}$ is recurrent. Let $\{Y_n\}$ be the Markov chain with transition matrix $\hat{P}(x, y) = \pi(y)P(y, x)/\pi(x)$. We know that

$$\pi(x)\mathbb{P}[X_0 = x, X_1 = x_1, ..., X_{n-1} = x_{n-1}, X_n = y] = \pi(y)\mathbb{P}[Y_0 = y, Y_1 = x_{n-1}, ..., Y_{n-1} = x_1, Y_n = x].$$

Moreover, π is also stationary for $\hat{P}$. by the second step, we know that $\{Y_n\}$ is recurrent.

Finally, we show that $\pi(x)\mathbb{E}_x[\tau_x^+] = 1$.

$$\begin{aligned}
\pi(x)\mathbb{E}_x[\tau_x^+] &= \sum_{n \geqslant 1} \pi(x)\mathbb{P}_x[\tau_x^+ \geqslant n] \\
&= \sum_{n \geqslant 1} \sum_y \pi(x)\mathbb{P}_x[X_n = y, \tau_x^+ \geqslant n] \\
&= \sum_{n \geqslant 1} \sum_y \pi(y)\hat{\mathbb{P}}_y[Y_n = x, \tau_x^+ \geqslant n] \\
&= \sum_{n \geqslant 1} \sum_y \pi(y)\hat{\mathbb{P}}_y[\tau_x^+ = n] \\
&= \sum_y \pi(y)\hat{\mathbb{P}}_y[\tau_x^+ < \infty] \\
&= \sum_y \pi(y) \qquad \text{(Since } \{Y_n\} \text{ is recurrent)} \\
&= 1.
\end{aligned}$$

☐

Corollary 7.1.8. *If an irreducible Markov chain is positive recurrent, then*

- *there exists a probability measure π such that $\pi = \pi P$;*

- *$\pi(x) > 0$ for all x. In fact,*

$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

Theorem 7.1.9. *Suppose that the Markov chain is irreducible, aperiodic and positive recurrent, then*

$$\lim_n ||P^n(x,\cdot) - \pi||_{TV} = 0.$$

In particular, for any state y, we have

$$\lim_n P^n(x,y) = \pi(y) > 0.$$

Proof. Idea: construct a coupling $\{(X_n, Y_n)\}$ of two Markov chains such that $X_0 = x$ and $Y_0 \sim \pi$, and that $X_n = Y_n$ as often as possible.

First, construct a transition matrix on $\mathcal{S} \times \mathcal{S}$. Define

$$Q((x,y),(z,w)) = P(x,z)P(y,w), \quad \forall x,y,z,w \in \mathcal{S}.$$

It is clear that the matrix Q is stochastic. We first show that the matrix Q is irreducible. We need to show that, for any x, y, z, w, there exists N such that

$$Q^N((x,y),(z,w)) = P^N(x,y)P^N(z,w) > 0.$$

Define $T(x) = \{n : P^n(x,x) > 0\}$. Note that $T(x)$ is closed under addition and has gcd 1. Thus, there exists $N(x)$ such that $P^n(x,x) > 0$ for all $n \geqslant N(x)$. By irreducibility, there exists $r = r(x,z)$ such that $P^r(x,z) > 0$. Define $M(x,z) := N(x) + r(x,z)$, then $P^n(x,z) > 0$ for all $n \geqslant M(x,z)$. Similarly, there exists $M(y,w)$ such that $P^n(y,w) > 0$ for all $n \geqslant M(y,w)$. Let $N = \max(M(x,z), M(y,w))$, then $Q^N((x,y),(z,w)) = P^N(x,y)P^N(z,w) > 0$.

Second, we show that Q is positive recurrent. Define $\pi \otimes \pi$ on $\mathcal{S} \times \mathcal{S}$ to be

$$\pi \otimes \pi(x,y) = \pi(x)\pi(y).$$

It is clear that $\pi \otimes \pi$ is a probability measure. Moreover, $\pi \otimes \pi$ is stationary for Q:

$$\begin{aligned}
\pi \otimes \pi Q(z,w) &= \sum_{x,y} \pi \otimes \pi(x,y) Q((x,y),(z,w)) \\
&= \sum_{x,y} \pi(x)\pi(y)P(x,z)P(y,w) \\
&= \pi(z)\pi(w) = \pi \otimes \pi(z,w).
\end{aligned}$$

Finally, we construct the coupling. The Markov chain $\{(X_n, Y_n)\}$ starts from $(X_0, Y_0) \sim \delta_x \otimes \pi$ and moves by Q. Define

$$\tau = \min\{n \geqslant 0 : (X_n, Y_n) = (y_0, y_0)\}.$$

Run the chain by Q until time τ. After τ, we keep them together. Note that $X_n \sim P^n(x,\cdot)$ and $Y_n \sim \pi$.

$$\|P^n(x,\cdot) - \pi\|_{TV} \leqslant \mathbb{P}[X_n \neq Y_n] \leqslant \mathbb{P}[\tau > n] = \sum_y \pi(y)\mathbb{Q}_{x,y}[\tau > n].$$

Since Q is recurrent, we have $\mathbb{Q}_{x,y}[\tau > n] \to 0$ as $n \to \infty$. This implies the conclusion. $\square$

7.2 Simple random walk on $\mathbb{Z}^d$

We have shown in Example 7.1.6 that the simple random walk on $\mathbb{Z}$ is null recurrent. In this section, we will show that the simple random walk on $\mathbb{Z}^2$ is recurrent; and the simple random walk on $\mathbb{Z}^3$ is transient, which implies that simple random walk on $\mathbb{Z}^d$ is transient for all $d \geqslant 3$.

Theorem 7.2.1.

- *Simple random walk on $\mathbb{Z}^2$ is recurrent.*
- *Simple random walk on $\mathbb{Z}^3$ is transient.*

Suppose $\{S_n\}$ is a simple random walk in $\mathbb{Z}^d$ starting from the origin. Define the sequence of stopping times that the walk returns to the origin:

$$\tau_0 = 0, \quad \tau_{k+1} = \min\{n > \tau_k : S_n = 0\}, \quad k \geqslant 0.$$

Lemma 7.2.2. *The following three assertions are equivalent:*

- *the walk is recurrent;*
- $\mathbb{P}[\tau_1 < \infty] = 1$*;*
- $\sum_m \mathbb{P}[S_m = 0] = \infty$*.*

Proof. Let us calculate the expectation of the number of visits to the origin. On the one hand,

$$\mathbb{E}[\#\text{visits to } 0] = \sum_n \mathbb{P}[\tau_n < \infty] = \sum_n \mathbb{P}[\tau_1 < \infty]^n.$$

On the other hand,

$$\mathbb{E}[\#\text{visits to } 0] = \sum_m \mathbb{P}[S_m = 0].$$

These give the conclusion. □

From the above lemma, to obtain recurrence, it suffices to determine the asymptotic of the probability $p_d(m) := \mathbb{P}[S_m = 0]$ as $m \to \infty$. Since $p_d(m) = 0$ for odd m, we only need to consider $p_d(m)$ with even m.

Lemma 7.2.3. *When $d = 1$, we have*

$$p_1(2n) \sim \frac{1}{\sqrt{\pi n}}.$$

Proof. It is clear that

$$p_1(2n) = \binom{2n}{n} 2^{-2n}.$$

The conclusion follows by combining with Stirling's formula:

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

□

Lemma 7.2.4. *When $d = 2$, we have*

$$p_2(2n) = p_1(2n)^2 \sim \frac{1}{\pi n}.$$

Proof. In order for $S_{2n} = 0$, there exists $m \in \{0, 1, \ldots, n\}$ so that the walk has m up steps, m down steps, $n - m$ to the left, and $n - m$ to the right. Thus

$$p_2(2n) = 4^{-2n} \sum_m \frac{(2n)!}{(m!(n-m)!)^2} = 4^{-2n} \binom{2n}{n} \sum_m \binom{n}{m}^2 = 4^{-2n} \binom{2n}{n}^2 = p_1(2n)^2.$$

□

Lemma 7.2.5. *When $d = 3$, we have*

$$p_3(2n) \leqslant O(n^{-3/2}).$$

Proof. In order for $S_{2n} = 0$, there exists $j, k \in \{0, 1, \ldots, n\}$ so that the walk has j up steps, j down steps, k to the left, k to the right, $n - j - k$ to the forward, and $n - j - k$ to the back. Thus

$$\begin{aligned} p_3(2n) &= 6^{-2n} \sum_{j,k} \frac{(2n)!}{(j!k!(n-j-k)!)^2} \\ &= 2^{-2n} \binom{2n}{n} \sum_{j,k} \left(\frac{3^{-n} n!}{j!k!(n-j-k)!} \right)^2 = p_1(2n) \sum_{j,k} \left(\frac{3^{-n} n!}{j!k!(n-j-k)!} \right)^2. \end{aligned}$$

Note that

$$\sum_{j,k} \frac{3^{-n} n!}{j!k!(n-j-k)!} = 1,$$

thus

$$p_3(2n) \leqslant p_1(2n) \max_{j,k} \frac{3^{-n} n!}{j!k!(n-j-k)!}.$$

The max will be obtained when j, k are integers close to $n/3$. By Stirling's formula, we have

$$\max_{j,k} \frac{3^{-n} n!}{j!k!(n-j-k)!} = O(n^{-1}).$$

Combining with the asymptotic of $p_1(2n)$, we obtain the conclusion. □

Proof of Theorem 7.2.1. Combining Lemmas 7.2.2 and 7.2.4, we obtain the recurrence when $d = 2$. Combining Lemmas 7.2.2 and 7.2.5, we obtain the transience when $d = 3$. □

7.3 Exercises

Exercise 7.3.1 (YCMC2016). *For a random walk process on the complete infinite binary tree starting from root (i.e. level 0), we assume that the object moves to the neighbor nodes with equal probability. Let X_n denote the level number at time n. Prove that $\mathbb{E}[X_n] \leqslant n/3 + 4/3$.*

Exercise 7.3.2 (YCMC2016). *A random walker moves on the lattice $\mathbb{Z}^2$ according to the following rule: in the first step it moves to one of its neighbors with probability 1/4, and then in step $n > 1$ it moves to one of the neighbors that it didn't visit in the step $n - 1$ with equal probability. Let T be the time when the random walker steps on a site that it already visited. Show that the expectation of T is less than 35.*

Exercise 7.3.3 (YCMC2017). *Let $\{S_n\}$ and $\{S'_n\}$ be two independent simple random walks on $\mathbb{Z}^d$ such that $X_0 = X'_0 = 0$. Define $\mathcal{I} = \{(s,t) : X_s = X'_t\}$. Prove that $|\mathcal{I}| < \infty$ a.s.*

Hint: You can first prove that

$$\mathbb{P}[X_n = 0] = O(n^{-d/2}), \quad n \to \infty.$$

Exercise 7.3.4 (YCMC2017). *Suppose a number $X_0 \in \{1, -1\}$ at the root of a binary tree is propagated away from the root as follows. The root is the node at level 0. After obtaining the 2^h numbers at the nodes at level h, each number at level $h+1$ is obtained from the number adjacent to it (at level h) by flipping its sign with probability $p \in (0, 1/2)$ independently. Let X_h be the average of the 2^h values received at the nodes at level h. Define the signal-to-noise ratio at level h to be*

$$R_h := \frac{(\mathbb{E}[X_h \,|\, X_0 = 1] - \mathbb{E}[X_h \,|\, X_0 = -1])^2}{\operatorname{var}(X_h \,|\, X_0 = 1)}.$$

Find the threshold number p_c such that R_h converges to 0 if $p \in (p_c, 1/2)$ and diverges if $p \in (0, p_c)$, as $h \to \infty$.