Probability Theory
Lecture Notes

Hanbaek Lyu

Department of Mathematics, University of Wisconsin - Madison, WI 53706

Email address: hlyu@math.wisc.edu

WWW.HANBAEKLYU.COM

Contents

CHAPTER 1

Foundations of probability theory

Many things in life are uncertain. Can we 'measure' and compare such uncertainty so that it helps us to make more informed decision? Probability theory provides a systematic way of doing so.

1.1. Probability measure and probability space

1.1.1. Axiomatic definition of probability spaces. A probability space is an idealized mathematical world where we can precisely measure uncertainty of all possible events. A *probability space* is defined to be a triple $(\Omega, \mathscr{F}, \mathbb{P})$ of sample space Ω, set of events $\mathscr{F}$, and probability measure $\mathbb{P}$:

(1) *Sample space* Ω: Set of all possible outcomes ω in a random experiment

(2) *Set of events* $\mathscr{F}$: A collection of subsets (events) E of the sample space Ω. If $A \subseteq \Omega$, then we can only measure the 'probability' or 'size' of A if and only if $A \in \mathscr{F}$. In this case, we say A is an 'event' and it is 'measurable'.[1] In a formal definition, we require $\mathscr{F}$ to be what is called the 'σ-algebra' (see Definition 1.1.10).

(3) *Probability measure* $\mathbb{P}$: A set function[2] $\mathbb{P} : \mathscr{F} \to [0,1]$ that maps each event $E \in \mathscr{F}$ to a numerical value $\mathbb{P}(\mathbb{E})$ between 0 and 1. We call $\mathbb{P}(E)$ the 'probability' of E. If $A \subseteq \Omega$ but $A \notin \mathscr{F}$, then $\mathbb{P}(A)$ is not defined. See Definition 1.1.2 for a formal definition.

A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is said to be *discrete* if Ω is countable (i.e., finite or countably infinite).

Before we give any abstract discussion, we start with the most elementary example of pobability spaces: The coin flip. Recall that if Ω is a set, then we denote by 2^Ω the set of all subsets of Ω, which is called the *power set* of Ω (note that $\emptyset \in 2^\Omega$).

Example 1.1.1 (coin flip)**.** Consider a random experiment of flipping a coin, which comes up heads 'H' or tails 'T'. Hence our sample space of all possible outcomes is $\Omega = \{H, T\}$. Next, let $2^\Omega = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$, which consists of all subsets of Ω. For instance, the subset $\{H\}$ corresponds to the event that a random coin flip comes up heads; the subset $\{H, T\}$ corresponds to the event that a random coin flip comes up heads or tails; and the subset $\emptyset$ corresponds to the event that a random coin flip comes up with noting. We would like to be able to measure the probability of all such events. Hence we set the σ-algebra (or the set of events) to be 2^Ω. Lastly, for the probability measure, fix a parameter $p \in [0,1]$, and define a function $\mathbb{P}_p : 2^\Omega \to [0,1]$ by $\mathbb{P}_p(\emptyset) = 0$, $\mathbb{P}_p(\{H\}) = p$, $\mathbb{P}_p(\{T\}) = 1 - p$, $\mathbb{P}_p(\{H, T\}) = 1$. The resulting probability space $(\Omega = \{H, T\}, \mathscr{F} = 2^\Omega, \mathbb{P} = \mathbb{P}_p)$ is a probability model for the random experiment of flipping a 'probability-p coin'. ▲

Recall that the probability measure $\mathbb{P}$ is a certain function that assigns a numerical value between 0 and 1 to every event, which we regard as some sort of 'size' of that event. Hence it is natural to require that the empty set is assigned with zero probability, and that the probability of the union of disjoint events is the sum of the probabilities of the individual events. This leads us to the following formal definition of 'measures'.

[1] When Ω is uncountable, not all subsets of Ω are measurable necessarily.

[2] Here a 'set function' is a function with sets as input.

Definition 1.1.2 (Measure)**.** Let $\mathscr{A}$ be a collection of subsets of Ω (i.e., $\mathscr{A} \subseteq 2^\Omega$). A set function $\mu : \mathscr{A} \to [-\infty, \infty]$ is called a *measure*[3] if it is nonnegative and countably additive:

(i) (Nonnegativity) $\mu(A) \geq \mu(\emptyset) = 0$;
(ii) (Countable additivity) If $A_i \in \mathscr{A}$ for $i \in \mathbb{N}$ are disjoint, then $\mu\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu(A_i)$.

Furthermore, if $\mu(\Omega) = 1$, then μ is called a *probability measure*.

Note that in Definition 1.1.2, we did not require any particular condition for the collection $\mathscr{A}$ of subsets of Ω. If Ω is countable (i.e., finite or countably infinite), then we usually take $\mathscr{A} = 2^\Omega$. In general, if $\mathscr{A}$ is too large, then it could be impossible to define a measure on $\mathscr{A}$. The right choice of $\mathscr{A}$ on which a measure can be defined is what is called the 'σ-algebra', which we will discuss later.

Exercise 1.1.3. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $A \subseteq \Omega$ be an event. Show that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Example 1.1.4 (Counting measure)**.** Suppose Ω is countable and $\mathscr{F} = 2^\Omega$. Define a set function $\gamma : 2^\Omega \to [-\infty, \infty]$ by

$$\gamma(A) = \begin{cases} |A| = \text{number of elements in } A & \text{if } A \text{ is finite} \\ \infty & \text{if } A \text{ is infinite.} \end{cases}$$

Then γ is a measure on 2^Ω and is called the *counting measure* on Ω. ▲

Example 1.1.5 (Uniform probability measure)**.** Let $\Omega = \{1, 2, \cdots, m\}$ be a finite sample space. The *uniform probability measure* on Ω is the set function $\mathbb{P} : 2^\Omega \to [0, 1]$ such that

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \qquad \text{for all } A \subseteq \Omega,$$

where $|A|$ denotes the cardinality of A (in this case the number of elements in A, so $|\Omega| = m$). Then $(\Omega, 2^\Omega, \mathbb{P})$ is a probability space. According to the definition,

$$\mathbb{P}(\{x\}) = 1/m \qquad \forall x \in \Omega.$$

That is, the probability of all singleton events is identically $1/m$. ▲

Example 1.1.6 (Roll of two dice)**.** Suppose we roll two dice and let X and Y be the outcome of each die, where the outcome of each die is one of the six integers $1, 2 \ldots, 6$. Then the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}^2 = \{(i, j) \,|\, 1 \leq i, j \leq 6\}$. We can visualize the sample space as the 6 by 6 square grid and each node represents a unique outcome (x, y) of the roll (see Figure 1.1.1). For the set of events, we take the power set 2^Ω. For the probability measure, we take the uniform probability measure $\mathbb{P}$. Since $|\Omega| = 36$, for any event $A \subseteq \Omega$, we have $\mathbb{P}(A) = |A|/36$. In particular,

$$\mathbb{P}\big((X, Y) = (x, y)\big) = \mathbb{P}(\{x, y\}) = 1/36 \qquad \forall 1 \leq x, y \leq 6.$$

We can compute various probabilities for this experiment. For example,

$$\begin{aligned} \mathbb{P}(\text{at least one die is even}) &= 1 - \mathbb{P}(\text{both dice are odd}) \\ &= 1 - \mathbb{P}(\{1, 3, 5\} \times \{1, 3, 5\}) = 1 - \frac{9}{36} = 3/4, \end{aligned}$$

where for the first equality we have used the complementary probability in Exercise 1.1.3.

Now think about the following question: *What is the most probable value for the sum $X + Y$?* By considering diagonal lines $x + y = k$ in the 2-dimensional plane for different values of k, we find

$$\mathbb{P}(X + Y = k) = \frac{\#\text{ of intersections between the line } x + y = k \text{ and } \Omega}{36}.$$

[3] To be precise, the set function μ is a *pre-measure* on $\mathscr{A}$ in general, and it is called a *measure* only if $\mathscr{A}$ is a σ-algebra (see Definition 1.1.10). However, we will not distinguish pre-measures and measures in this note.

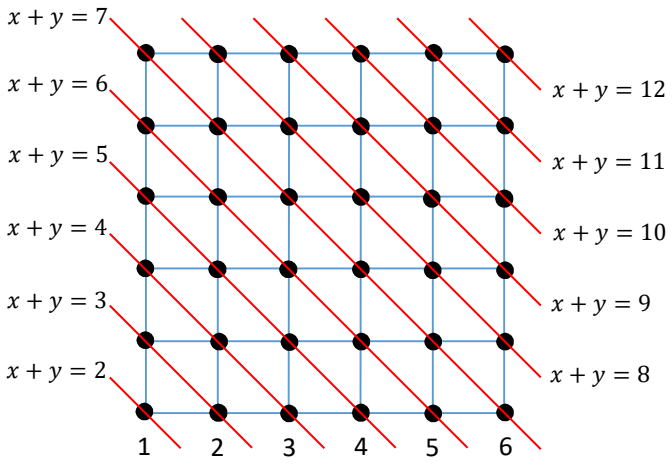


FIGURE 1.1.1. Sample space representation for roll of two independent fair dice and and events of fixed sum of two dice.

Here, we wrote $\{X+Y=k\}$ as a shorthand of the following event

$$\{\omega \in \Omega \,|\, X(\omega)+Y(\omega)=k\},$$

where for each $\omega=(x,y)\in\Omega$, $X(\omega)=x$ and $Y(\omega)=y$. From example, $\mathbb{P}(X+Y=2)=1/36$ and $\mathbb{P}(X+Y=7)=6/36=1/6$. Moreover, from Figure 1.1.1, it is clear that the number of intersections is maximized when the diagonal line $x+y=k$ passes through the extreme points $(1,6)$ and $(6,1)$. Hence 7 is the most probable value for $X+Y$ with the probability being $1/6$. ▲

Exercise 1.1.7 (Roll of three dice)**.** Suppose we roll three dice and all possible joint outcomes are equally likely. Model this experiment as the probability space $(\Omega, 2^{\Omega}, \mathbb{P})$ where $\Omega=\{1,2,3,4,5,6\}^3$ and $\mathbb{P}$ is the uniform probability measure on Ω. Identify the sample space Ω as the $(6\times6\times6)$ 3-dimensional integer lattice, and let X, Y, and Z denote the outcome of each die.

(i) Write down the probability distribution on Ω.

(ii) For each $k\geq1$, show that

$$\mathbb{P}(X+Y+Z=k)=\frac{\#\text{ of intersections between the plane } x+y+z=k \text{ and } \Omega}{6^3}.$$

What are the minimum and maximum possible values for $X+Y+Z$?

(iii) Draw a cube for Ω and planes $x+y+z=k$ for $k=3,5,10,11,16,18$. Argue that the intersection gets larger as k increases from 3 to 10 and smaller as k goes from 11 to 18. Conclude that 10 and 11 are the most probable values for $X+Y+Z$.

(iv) Consider the following identity

$$\begin{aligned}&(x+x^2+x^3+x^4+x^5+x^6)^3\\ &\quad=x^{18}+3x^{17}+6x^{16}+10x^{15}+15x^{14}+21x^{13}+25x^{12}+27x^{11}+27x^{10}\\ &\qquad+25x^9+21x^8+15x^7+10x^6+6x^5+3x^4+x^3\end{aligned}$$

Show that the coefficient of x^k in the right hand side equals the size of the intersection between Ω and the plane $x+y+z=k$. Conclude that

$$\mathbb{P}(X+Y+Z=10)=\mathbb{P}(X+Y+Z=11)=\frac{27}{6^3}=\frac{1}{8}.$$

(This way of calculating probabilities is called the generating function method.)

Example 1.1.8 (Probability measure on countable sample space)**.** Let $\Omega = \{x_1, x_2, \ldots\}$ be a countable sample space. Take the power set 2^Ω as the set of events. A typical way of constructing a probability measure is to specify how likely it is to see each individual element in Ω. Namely, let $f : \Omega \to [0,1]$ be a function that sums up to 1, i.e., $\sum_{x\in\Omega} f(x) = 1$. Define a function $\mathbb{P} : 2^\Omega \to [0,1]$ by

$$\mathbb{P}(E) := \sum_{\omega\in E} f(\omega). \tag{1}$$

Then one can easily check that $\mathbb{P}$ is a probability measure on 2^Ω and f is called *probability mass function* (PMF) of $\mathbb{P}$. For instance, the PMF of the probability measure $\mathbb{P}_p$ in the coin flip example in Exercise 1.1.1 is given by $f(H) = p$ and $f(T) = 1 - p$. ▲

Exercise 1.1.9. Show that the function $\mathbb{P} : 2^\Omega \to [0,1]$ defined in (1) is a probability measure on Ω. Conversely, show that every probability measure in a discrete probability space can be defined in this way (in this case, identify the PMF).

Next, we discuss the requirement for a collection of subsets $\mathscr{F}$ of the sample space Ω to be considered as the set of events. In all examples we have seen above, our sample space Ω was countable and we took the power set 2^Ω to be the set of events. However, when Ω is uncountable (e.g., $\Omega = [0,1]$ or $\Omega = \mathbb{R}$), then the power set 2^Ω is 'too large' to be considered as the set of events. It turns out that $\mathscr{F}$ needs to be a system of subsets of Ω that satisfy certain properties. Namely, it is natural to expect that the union, intersection, and complements of events are also events. In other words, the set of events $\mathscr{F}$, which is a system of subsets of Ω, should be 'closed under' elementary set operations such as taking union, intersection, and complements. Furthermore, we would like to make $\mathscr{F}$ rich enough so that we can perform such operations countably many times. This motivates the following definition:

Definition 1.1.10 (σ-algebra)**.** Let Ω be a set and let $\mathscr{F}$ be a set of subsets of Ω. We call $\mathscr{F}$ a *σ-algebra* on Ω if it satisfies the following properties:

(i) (contains the whole set) $\Omega \in \mathscr{F}$;
(ii) (closed under complements) $A \in \mathscr{F} \implies A^c = \Omega \setminus A \in \mathscr{F}$;
(iii) (closed under countable union) $A_i \in \mathscr{F}$ for $i \in \mathbb{N} \implies \bigcup_{i=1}^\infty A_i \in \mathscr{F}$.

Note that a σ-algebra contains the empty set $\emptyset$, being the complement of the whole set Ω. While the condition **(iii)** only states that the union of countably infinite collection of events is again an event, since A_i's can be the empty set, it also entails closedness under finite union.

Exercise 1.1.11. Let $\mathscr{F}$ be a σ-algebra on a set Ω. Show that $\mathscr{F}$ is closed under countable intersection. That is, show that if $A_i \in \mathscr{F}$ for $i \in \mathbb{N}$, then $\bigcap_{i=1}^\infty A_i \in \mathscr{F}$.

Exercise 1.1.12 (Power set is a σ-algebra)**.** Let Ω be a set and let 2^Ω denote the set of all subsets of Ω, which is called the *power set* of Ω (note that $\emptyset \in 2^\Omega$). Show that 2^Ω is a σ-algebra on Ω.

Exercise 1.1.13 (σ-algebra generated by a partition)**.** Let Ω be a sample space and let $\Omega = \bigcup_{k\geq 1} A_k$ for some disjoint subsets $A_1, A_2, \ldots$ of Ω. Then show that for each $B \in \sigma(\{A_1, A_2, \ldots\})$, there exists an index set $I \subseteq \mathbb{N}$ such that $B = \bigcup_{k\in I} B_k$.

Exercise 1.1.14 (σ-algebra on a countable set)**.** Let Ω be a countable set. Let $\mathscr{F}$ be an arbitrary σ-algebra on Ω containing all singleton sets: $\{x\} \in \mathscr{F}$ for all $x \in \Omega$. Show that $\mathscr{F} = 2^\Omega$.

Now we can give a formal definition of a probability space.

Definition 1.1.15 (Probability spaces)**.** Let Ω be a set, not necessarily countable. Let $\mathscr{F}$ be any σ-algebra on Ω and let $\mathbb{P}$ be any probability measure on $\mathscr{F}$. Then the triple $(\Omega, \mathscr{F}, \mathbb{P})$ defines a *probability space*. If Ω is countable, then $(\Omega, \mathscr{F}, \mathbb{P})$ is called a *discrete probability space*.

The following are important properties of a measure.

Theorem 1.1.16. *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. These followings hold:*

(i) *(Monotonicity) For any events $A \subseteq B$, $\mu(A) \leq \mu(B)$.*

(ii) *(Subadditivity) For $A \subseteq \bigcup_{i=1}^{\infty} A_i$, $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.*

(iii) *(Continuity from below) If $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$ (denote $A_n \nearrow A$), then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.*

(iv) *(Continuity from above) If $A_1 \supseteq A_2 \supseteq \cdots$, $A = \bigcap_{i=1}^{\infty} A_i$ (denote $A_n \searrow A$), and $\mu(A_1) < \infty$, then $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.*

PROOF. (i) Since $A \subseteq B$, write $B = A \cup (B \setminus A)$. Note that A and $B \setminus A$ are disjoint. Hence by the second axiom of probability measure, we get

$$\mu(B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \geq \mu(A),$$

where the last inequality uses the fact that $\mu(B \setminus A) \geq 0$.

(ii) The events A_i's are not necessarily disjoint, but we can cook up a collection of disjoint events B_i's so that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Namely, we define

$$\begin{aligned} B_1 &= A_1 \subseteq A_1 \\ B_2 &= (A_1 \cup A_2) \setminus A_1 \subseteq A_2 \\ B_3 &= (A_1 \cup A_2 \cup A_3) \setminus (A_1 \cup A_2) \subseteq A_3, \end{aligned}$$

and so on. Then clearly B_i's are disjoint and their union is the same as the union of A_i's. Now by part (i) and countable additivity of probability measure, we get

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

(iii) Define a collection of disjoint subsets B_i's similarly as in (ii). In this case, they will be

$$\begin{aligned} B_1 &= A_1 \subseteq A_1 \\ B_2 &= A_2 \setminus A_1 \subseteq A_2 \\ B_3 &= A_3 \setminus A_2 \subseteq A_3, \end{aligned}$$

and so on. Then $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ and $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$. Hence we get

$$\begin{aligned} \mu(A) &= \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mu(B_i) \\ &= \lim_{k\to\infty} \sum_{i=1}^{n} \mu(B_i) = \lim_{n\to\infty} \mu(B_1 \cup \cdots \cup B_n) = \lim_{n\to\infty} \mu(A_n). \end{aligned}$$

(iv) Since $A_1 \setminus A_n \nearrow A_1 \setminus A$, by **(iii)** we get $\mu(A_1 \setminus A_n) \nearrow \mu(A_1 \setminus A)$. Also, since $A \subseteq A_1$, we have $\mu(A_1 \setminus A) = \mu(A_1) - \mu(A)$. Hence

$$\mu(A_n) = \mu(A_1) - \mu(A_1 \setminus A_n) \searrow \mu(A_1) - \mu(A_1 \setminus A) = \mu(A).$$

□

Some immediate consequences of Theorem 1.1.16 (ii) are given in the following exercise.

Exercise 1.1.17 (Union bound)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.

(i) For any $A, B \subseteq \Omega$ such that $A \subseteq B$, show that

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

(ii) For any $A, B \subseteq \Omega$, show that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

(iii) Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. By an induction on k, show that

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathbb{P}(A_i).$$

(iv) (Countable subadditivity) Let $A_1, A_2, \cdots \subseteq \Omega$ be a countable collection of events. Show that

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

Exercise 1.1.18 (Inclusion-exclusion)**.** Let $(\Omega, \mathbb{P})$ be a probability space. Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. Show the following.

(i) For any $A, B \subseteq \Omega$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

(ii) For any $A, B, C \subseteq \Omega$,

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(B \cap C) - \mathbb{P}(C \cap A) + \mathbb{P}(A \cap B \cap C).$$

(iii)* Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. Use an induction on k to show that

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i_1=1}^{k} \mathbb{P}(A_{i_1}) - \sum_{1 \leq i_1 < i_2 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{k+1} \mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right).$$

Remark 1.1.19. Later, we will show the general inclusion-exclusion in a much easier way using random variables and expectation.

1.1.2. Constructing σ-algebras bottom-up. Note that Definition 1.1.10 gives an 'axiomatic definition' of a σ-algebra, but it does not tell us how to construct such an object. For instance, suppose that $\mathscr{F}$ is a σ-algebra on $\Omega = \mathbb{R}$. Suppose that $\mathscr{F}$ contains $\mathscr{C}$, which is the collection of all open intervals of the form (a, b) for $a, b \in \mathbb{R}$, $a \leq b$. Note that the set $A = (1,2) \cup (3,4)$ is itself not an open interval so it is not contained in $\mathscr{C}$. But $A \in \mathscr{F}$ since it is made of taking the union of two sets $(1,2)$ and $(3,4)$ that are members of the σ-algebra $\mathscr{F}$. In this way, one can build new sets from $\mathscr{C}$ by taking the complement, countable union, and countable intersection, until the collection of subsets becomes a σ-algebra. By only including the sets that can be made of this way, one obtains a σ-algebra 'generated by $\mathscr{C}$', which is denoted as $\sigma(\mathscr{C})$. This is an important concept of constructing a σ-algebra from a small collection of subsets.

Definition 1.1.20 (σ-algebra generated by a collection of subsets)**.** Let Ω be a set and let $\mathscr{C}$ be a set of subsets of Ω (i.e., $\mathscr{C} \subseteq 2^{\Omega}$). The *$\sigma$-algebra generated by $\mathscr{C}$* is the smallest σ-algebra on Ω that contains $\mathscr{C}$.

Exercise 1.1.21. Let Ω be a set and let $\mathfrak{F}$ denote an arbitrary non-empty collection of σ-algebras on Ω. Let $\mathscr{A}$ denote the intersection of all σ-algebras in $\mathfrak{F}$:

$$\mathscr{A} := \bigcap_{\mathscr{F} \in \mathfrak{F}} \mathscr{F} = \{B \subseteq \Omega \mid B \in \mathscr{F} \text{ for all } \mathscr{F} \in \mathfrak{F}\}.$$

Then show that $\mathscr{A}$ is a σ-algebra on Ω.

Proposition 1.1.22. *Let Ω be a set and let $\mathscr{C}$ be a set of subsets of Ω. Then $\sigma(\mathscr{C})$ exists and is unique. More precisely,*

$$\sigma(\mathscr{C}) = \bigcap \{\mathscr{F} \mid \mathscr{F} \text{ is a } \sigma\text{-algebra on } \Omega \text{ containing } \mathscr{C}\}.$$

PROOF. (Uniqueness): Let $\mathscr{A}$ and $\mathscr{A}'$ be two smallest σ-algebra on Ω that contains $\mathscr{C}$. Then by definition, $\mathscr{A} \subseteq \mathscr{A}'$ and $\mathscr{A}' \subseteq \mathscr{A}$. Hence $\mathscr{A} = \mathscr{A}'$.

(Existence): Let $\mathfrak{F}$ denote the collection of all σ-algebras on Ω that contains $\mathscr{C}$. Recall that by Exercise 1.1.12, the power set 2^Ω is a σ-algebra on Ω and it contains $\mathscr{C}$, so $2^\Omega \in \mathfrak{F}$. Since $\mathfrak{F}$ is non-empty, if we let $\mathscr{A}$ be the intersection of all σ-algebras in $\mathscr{F}$, then it is a σ-algebra by Exercise 1.1.21 and it contains $\mathscr{C}$. By definition, $\mathscr{A}$ is a smalleset σ-algebra containing $\mathscr{C}$. Indeed, if $\mathscr{A}'$ is an arbitrary σ-algebra that contains $\mathscr{C}$, then by definition $\mathscr{A}' \in \mathfrak{F}$, so by definition $\mathscr{A} \subseteq \mathscr{A}'$. Hence $\mathscr{A}$ is a smallest σ-algebra that contains $\mathscr{C}$. By the uniqueness, we must then have $\mathscr{A} = \sigma(\mathscr{C})$. exists. □

Exercise 1.1.23. Let $\mathscr{F}$ be a σ-algebra on a sample space Ω. Let $\mathscr{C}$ be a collection of subsets of Ω such that $\mathscr{C} \subseteq \mathscr{F}$. Then show that $\sigma(\mathscr{C}) \subseteq \mathscr{F}$.

Definition 1.1.24 (Borel σ-algebra on $\mathbb{R}^d$)**.** Let $\Omega = \mathbb{R}^d$ be the d-dimensional Euclidean space. An *open ball* in $\mathbb{R}^d$ is a subset of $\mathbb{R}^d$ given by $B(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^d \mid \|\mathbf{x} - \mathbf{y}\| < d\}$, where $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$. Let $\mathscr{C}$ denote the set of all open balls in $\mathbb{R}^d$. Then the σ-algebra generated by all open balls in $\mathbb{R}^d$, $\sigma(\mathscr{C})$, is called the *Borel σ-algebra* on $\mathbb{R}^d$, which is typically denoted as $\mathscr{B}^n$. When $n = 1$, we denote $\mathscr{B}^1 = \mathscr{B}$.

Exercise 1.1.25. Let $\mathscr{B}$ be the Borel σ-algebra on $\mathbb{R}$. Show that all closed intervals and half-open intervals are contained in $\mathscr{B}$.

Example 1.1.26 (The Cantor set)**.** Starting from the unit interval $[0,1]$, take away the middle third interval iteratively. The limit of this process defines a subset C of $[0,1]$, which is known as the Cantor set (see Figure 1.1.2). A more precise way to construct the Cantor set is as follows. Define a sequence of subsets $\mathsf{C}_0, \mathsf{C}_1, \ldots$ of $[0,1]$ recursively as $\mathsf{C}_0 := [0,1]$ and

$$\mathsf{C}_{n+1} = \frac{C_n}{3} \cup \left(\frac{2}{3} + \frac{C_n}{3}\right) \quad n = 1, 2, \ldots.$$

Now we define the Cantor set C as

$$\mathsf{C} := \bigcap_{n=1}^{\infty} \mathsf{C}_n.$$

FIGURE 1.1.2. Construction of Cantor set. Starting from the unit interval $[0,1]$, take away the middle third interval iteratively. The limit of this process defines a subset C of $[0,1]$, which is known as the Cantor set. (Figure credit: https://en.wikipedia.org/wiki/Cantor_set)

Observe the following facts:

(i) $\mathsf{C}_0 \supseteq \mathsf{C}_1 \supseteq \mathsf{C}_2 \supseteq \cdots$.

(ii) For each $n \geq 0$, C_n is the disjoint union of some closed intervals.

(iii) The 'Lebesgue measure' of C_n, which equals the sum of lengths of all closed intervals consisting of C_n, equals $(2/3)^n$.

From **(ii)** above, the Cantor set is defined as the countable intersection of C_n, which itself is a countable union of closed intervals. Now letting $\mathscr{B}$ be the Borel σ-algebra on $\mathbb{R}$, it follows that the Cantor set is Borel measurable, that is, $\mathsf{C} \in \mathscr{B}$. Furthermore, **(i)** and **(iii)** imply that the Lebesgue measure (see Example 1.1.39) of C is zero. ▲

1.1.3. Constructing measures bottom-up. In Subsection 1.1.1, we defined measures on a collection of subsets of the sample space by specifying the properties that they need to satisfy. While such an axiomatic definition gives a clean and fast shortcut, it is not clear why certain properties are required as part of the definition. More importantly, we do not know whether the abstract object satisfying all the required axioms *exists*. In this section, we present a 'constructive' approach of defining (probability) measures, following Carathéodory's foundational work on measure theory.

The outline of the construction is as follow. First, we start with a sample space Ω and a collection of its subsets $\mathscr{C} \subseteq 2^\Omega$ containing the empty set. (For example, $\Omega = \mathbb{R}$ and $\mathscr{C}$ the collection of all intervals). We are given with an 'a priori measure' $\alpha : \mathscr{C} \to [0,\infty]$, with only the requirement that $\alpha(\emptyset) = 0$.

(1) Extend $\alpha : \mathscr{C} \to [0,\infty]$ to a set function $\overline{\alpha} : 2^\Omega \to [0,\infty]$ defined on arbitrary subsets of Ω via the outer measure extension (see Definition 1.1.29).

(2) Let $\mathscr{F}$ be the collection of all $\overline{\alpha}$-measurable subsets of Ω (see Definition 1.1.32). Show that $\mathscr{F}$ is a σ-algebra (see Lemma 1.1.33).

(3) Show that $\overline{\alpha}$ restricted on $\mathscr{F}$ is a measure.

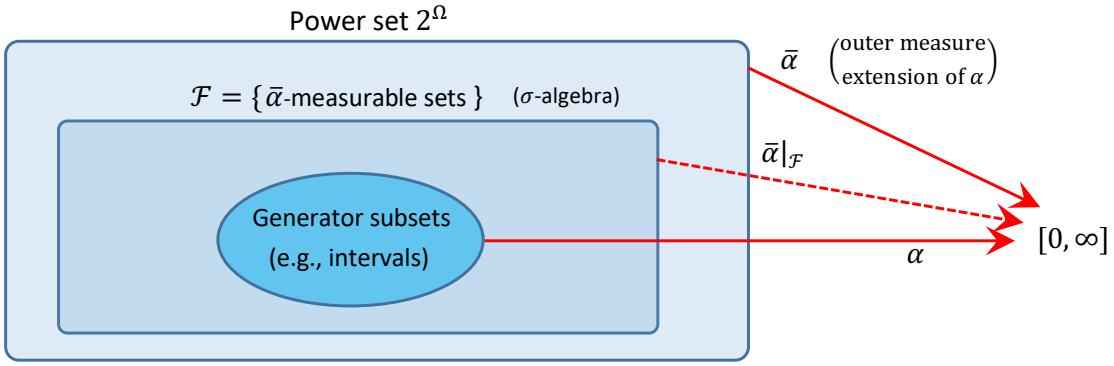


FIGURE 1.1.3. Bottom-up construction of a measure by outer measure extension and restriction onto measurable sets.

Definition 1.1.27 (Outer measures)**.** Let Ω be a set. A set function $\rho : 2^\Omega \to [0,\infty]$ is an *outer measure* on Ω if the following hold:

(i) (null empty set) $\rho(\emptyset) = 0$;

(ii) (countable subadditivity) For subsets $A, B_1, B_2 \cdots \subseteq \Omega$,

$$A \subseteq \bigcup_{i=1}^{\infty} B_i \qquad \Longrightarrow \qquad \rho(A) \leq \sum_{i=1}^{\infty} \rho(B_i).$$

Exercise 1.1.28. Show that outer measures have monotonicity. That is if ρ is an outer measure on Ω and if $A, B \subseteq \Omega$ such that $A \subseteq B$, then $\rho(A) \leq \rho(B)$.

Definition 1.1.29 (Outer measure extension)**.** Let Ω be a set and $\mathscr{C}$ be a collection of subsets of Ω (i.e., $\mathscr{C} \subseteq 2^\Omega$) containing the empty set. Let $\alpha : \mathscr{C} \to [0,\infty]$ be a set function such that $\alpha(\emptyset) = 0$. Define a set function $\alpha^* : 2^\Omega \to [0,\infty]$ by

$$\alpha^*(E) := \inf\left\{ \sum_{i=1}^{\infty} \alpha(C_i) \,\middle|\, C_1, C_2, \cdots \in \mathscr{C} \text{ and } E \subseteq \bigcup_{i=1}^{\infty} C_i \right\}. \tag{2}$$

We say α^* is an *outer measure extension* of α. (Take $\inf(\emptyset) = \infty$.)

Exercise 1.1.30. Let $\mathscr{C} \subseteq 2^\Omega$ contains $\emptyset$ and covers Ω (i.e., $\Omega = \bigcup \mathscr{C}$). Fix an arbitrary set function $\alpha : \mathscr{C} \to [0,\infty]$ such that $\alpha(\emptyset) = 0$. Show that the outer measure extension $\overline{\alpha}$ (see (2)) is indeed an outer measure on Ω.

Example 1.1.31. Let $\Omega = \mathbb{R}$ and $\mathscr{C}$ be the collection of all half-open intervals of the form $(a,b]$, $a,b \in \mathbb{R}$. Let $\alpha : \mathscr{C} \to [0,\infty]$, $\alpha((a,b]) = b-a$. Note that α gives the usual 'length' $b-a$ of the interval $(a,b]$. For now α is only defined on half-open intervals. Hence one cannot evaluate $\alpha((1,2]\cup(3,4])$, for example. Let $\overline{\alpha}$ be the outer measure constructed from α as in (2). Then note that

$$\overline{\alpha}((1,2]\cup(3,4]) = \alpha((1,2]) + \alpha((3,4]) = (2-1)+(4-3) = 2.$$

For another example, consider the open interval $(0,1)$. We first write $(0,1)$ as the following disjoint union of half-open intervals:

$$(0,1) = (0,1-2^{-1}]\cup(1-2^{-1},1-3^{-1}]\cup(1-3^{-1},1-4^{-1}]\cup\dots.$$

Then invoking the definition of the outer measure extension $\overline{\alpha}$ of α in (2),

$$\begin{aligned}\overline{\alpha}((0,1)) &= \alpha((0,1-2^{-1}]) + \alpha((1-2^{-1},1-3^{-1}]) + \alpha((1-3^{-1},1-4^{-1}]) + \dots \\ &= \left(1-2^{-1}\right) + \left(2^{-1}-3^{-1}\right) + \left(3^{-1}-4^{-1}\right) + \dots \\ &= 1.\end{aligned}$$

In general, one can show that $\overline{\alpha}$ is countably additive on $\mathscr{C}$: If $I_1, I_2, \dots$ are disjoint elements of $\mathscr{C}$, then

$$\overline{\alpha}\left(\bigcup_{i=1}^{\infty} I_i\right) = \sum_{i=1}^{\infty} \alpha(I_i).$$

▲

Let $A, E \subseteq \Omega$ be arbitrary. Then by using subadditivity of ρ and since $A \subseteq (A\cap E)\cup(A\setminus E)$, we have

$$\rho(A) \le \rho(A\cap E) + \rho(A\setminus E).$$

If the above inequality becomes an equality for all $A \subseteq \Omega$, then we say E is ρ-measurable. See below.

Definition 1.1.32 (Measurability w.r.t. an outer measure)**.** Let ρ be an outer measure on a set Ω. A subset $E \subseteq \Omega$ is said to be *ρ-measurable* (or Carathéory-measruable relative to ρ) if

$$\rho(A) = \rho(A\cap E) + \rho(A\setminus E) \quad \text{for all } A \subseteq \Omega.$$

Lemma 1.1.33 (Carathéory's lemma)**.** *Let ρ be an outer measure on a set Ω. Let $\mathscr{F}$ be the collection of all ρ-measurable subsets of Ω.*

(i) *$\mathscr{F}$ is a σ-algebra.*

(ii) *ρ restricted to $\mathscr{F}$ is a measure on $\mathscr{F}$.*

PROOF. We first show **(i)**. We will take some steps.

($\mathscr{F}$ is closed under complementation) Let $E \in \mathscr{F}$. This means $\rho(A) = \rho(A\cap E) + \rho(A\setminus E)$ for all $A \subseteq \Omega$. Note that since $A\cap E^c = A\setminus E$ and $A\setminus E = A\cap E^c$, this implies $\rho(A) = \rho(A\cap E^c) + \rho(A\setminus E^c)$ for all $A \subseteq \Omega$. This shows $E^c \in \mathscr{F}$. Since $E \in \mathscr{F}$ was arbitrary, this shows that $\mathscr{F}$ is closed under complementation.

($\mathscr{F}$ is closed under intersection and union) Let $E_1, E_2 \in \mathscr{F}$ and let $A \subseteq \Omega$ be arbitrary. Then by ρ-measurability of E_1 and E_2, we have

$$\begin{cases} \rho(A) & = \rho(A\cap E_1) + \rho(A\setminus E_1) \\ \rho(A\cap E_1) & = \rho(A\cap(E_1\cap E_2)) + \rho((A\cap E_1)\setminus E_2) \\ \rho(A\setminus(E_1\cap E_2)) & = \rho([A\setminus(E_1\cap E_2)]\cap E_1) + \rho([A\setminus(E_1\cap E_2)]\setminus E_1) \\ & = \rho((A\cap E_1)\setminus E_2) + \rho(A\setminus E_1). \end{cases}$$

By combining the above identities, we have

$$\begin{aligned}\rho(A) &= \left(\rho(A\cap(E_1\cap E_2)) + \rho((A\cap E_1)\setminus E_2)\right) + \rho(A\setminus E_1) \\ &= \rho(A\cap(E_1\cap E_2)) + \rho(A\setminus(E_1\cap E_2)).\end{aligned}$$

This verifies the ρ-measurability of $E_1 \cap E_2$. Hence $\mathscr{F}$ is closed under intersection. It follows that $\mathscr{F}$ is also closed under union.

($\mathscr{F}$ is closed under countable disjoint union) Let $(B_i)_{i\geq 1}$ be a sequence of disjoint ρ-measurable subsets of Ω. Let $C_n := \bigcup_{i=1}^n B_i$ and $C_\infty := \bigcup_{i=1}^\infty B_i$. We wish to show $C_\infty \in \mathscr{F}$. Then since $B_n \in \mathscr{F}$, for each $A \subseteq \Omega$,

$$\begin{aligned}\rho(A \cap C_n) &= \rho((A \cap C_n) \cap B_n) + \rho((A \cap C_n) \setminus B_n) \\ &= \rho(A \cap B_n) + \rho(A \cap C_{n-1}).\end{aligned}$$

By iterating the above argument, it follows that

$$\phi(A \cap C_n) = \sum_{i=1}^n \rho(A \cap B_i).$$

Since $B_i \in \mathscr{F}$ for $i \geq 1$ and $\mathscr{F}$ is closed under finite union, we have

$$\begin{aligned}\rho(A) &= \rho(A \cap C_n) + \rho(A \setminus C_n) \\ &\geq \left(\sum_{i=1}^n \rho(A \cap B_i)\right) + \rho(A \setminus C_\infty),\end{aligned}$$

where for the inequality above, we used the subadditivity of the outer measure ϕ and that $A \setminus C_n \supseteq A \setminus C_\infty$. Taking $n \to \infty$ and using subadditivity of ρ with $\bigcup_{i=1}^\infty (A \cap B_i) \supseteq A \cap C_\infty$,

$$\rho(A) \geq \left(\sum_{i=1}^\infty \rho(A \cap B_i)\right) + \rho(A \setminus C_\infty) \geq \rho(A \cap C_\infty) + \rho(A \setminus C_\infty) \geq \rho(A).$$

Thus above inequalities hold as equalities. This shows $C_\infty \in \mathscr{F}$.

($\mathscr{F}$ is closed under countable union) Let $(E_i)_{i\geq 1}$ be a sequence of elements in $\mathscr{F}$. We wish to show that $E := \bigcup_{i\geq 1} E_i \in \mathscr{F}$. We use the standard trick in measure theory to break up the union into the disjoint union. Namely, define sequence of subsets $(B_i)_{i\geq 1}$ by

$$\begin{aligned}B_1 &= E_1 \subseteq E_1 \\ B_2 &= E_2 \setminus E_1 \subseteq E_2 \\ B_3 &= E_3 \setminus E_2 \subseteq E_3,\end{aligned}$$

and so on. In this way, B_i's are disjoint and $\bigcup_{i=1}^n B_i = \bigcup_{i=1}^n E_i$ for all $n \geq 1$. Since we have shown that $\mathscr{F}$ is closed under complementation and intersection, it follows that $B_i \in \mathscr{F}$ for all $i \geq 1$. Since we have also shown that $\mathscr{F}$ is closed under countable disjoint union, it follows that $\bigcup_{i=1}^\infty E_i = \bigcup_{i=1}^n B_i \in \mathscr{F}$.

Now we show **(ii)**. It suffices to show that ρ is countably additive on $\mathscr{F}$. Let $(E_i)_{i\geq 1}$ be a sequence of disjoint elements in $\mathscr{F}$. We wish to show that $\rho(\bigcup_{i=1}^\infty E_i) = \sum_{i=1}^\infty \rho(E_i)$. First note that, since E_1 is ρ-measurable,

$$\begin{aligned}\rho(E_1 \cup E_2) &= \rho((E_1 \cup E_2) \cap E_1) + \rho((E_1 \cup E_2) \setminus E_1) \\ &= \rho(E_1) + \rho(E_2).\end{aligned}$$

By using a similar argument, we can verify that ρ is additive on $\mathscr{F}$: $\sum_{i=1}^n \rho(E_i) = \rho\left(\bigcup_{i=1}^n E_i\right)$. Now by countable subadditivity of ρ, we have

$$\sum_{i=1}^n \rho(E_i) = \rho\left(\bigcup_{i=1}^n E_i\right) \leq \rho\left(\bigcup_{i=1}^\infty E_i\right) \leq \sum_{i=1}^\infty \rho(E_i).$$

Then taking $n \to \infty$ finishes the proof. □

Definition 1.1.34 (Semi-ring)**.** Let Ω be a set and let $\mathscr{R} \subseteq 2^\Omega$. We call $\mathscr{R}$ a *semi-ring* if the following hold:

(i) $\emptyset \in \mathscr{R}$;

(ii) $A, B \in \mathscr{R} \implies A \cap B \in \mathscr{R}$.
(ii) $A, B \in \mathscr{R} \implies A \setminus B = \bigcup_{i=1}^{\infty} C_i$ for disjoint $C_i \in \mathscr{R}$, $i \geq 1$.

A prime example of semi-ring (and the only example you would need to know) is the set of half-open intervals. For instance, letting $A = (0,3]$ and $B = (1,2]$, we have

$$A \setminus B = (0,1] \cup (2,3],$$

which is not a half-open interval of the form $(a,b]$, but is indeed the disjoint union of two such intervals.

Exercise 1.1.35 (Set of half-open intervals is a semi-ring)**.** Let $\Omega = \mathbb{R}$ and let $\mathscr{R}$ be the set of all half-open intervals of the form $(a,b]$ for $a \leq b$, $a, b \in \mathbb{R}$. Show that $\mathscr{R}$ is a semi-ring.

Theorem 1.1.36 (Carathéory extension theorem)**.** *Let Ω be a set and let $\mathscr{R} \subseteq 2^{\Omega}$. Fix a set function $\alpha : \mathscr{R} \to [0,\infty]$. Suppose the following holds:*

(A1) *$\mathscr{R}$ is a semi-ring;*
(A2) *For each $A \in \mathscr{R}$ such that $A = \bigcup_{i=1}^{\infty} B_i$ for some disjoint $B_1, B_2, \cdots \in \mathscr{R}$, $\alpha\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \alpha(B_i)$.*

Then the followings hold:

(i) *There exists a measure μ on $\sigma(\mathscr{R})$ that extends α on $\mathscr{R}$ (i.e., $\mu|_{\mathscr{R}} = \alpha$).*
(ii) *If α is σ-finite, that is, $\Omega = \bigcup_{i=1}^{\infty} A_i$ for some $A_i \in \mathscr{R}$ for $i \geq 1$ such that $\alpha(A_i) < \infty$, then the measure μ in **(i)** is unique.*

SKETCH OF PROOF. Let $\overline{\alpha} : 2^{\Omega} \to [0,\infty]$ be the outer measure extension of $\alpha : \mathscr{R} \to [0,\infty]$. Let $\mathscr{F}$ denote the set of all $\overline{\alpha}$-measurable subsets of Ω. By Lemma 1.1.33, we know that $\mathscr{F}$ is a σ-algebra and $\overline{\alpha}$ restricted on $\mathscr{F}$ is a measure on $\mathscr{F}$.

Now with the additional hypothesis that $\mathscr{R}$ is a semi-ring and α is countably additive on $\mathscr{R}$, one can show that $\mathscr{R} \subseteq \mathscr{F}$. Recall that $\mathscr{F}$ is itself a σ-algebra. Since $\sigma(\mathscr{R})$ is the smallest σ-algebra that contains $\mathscr{R}$, it follows that $\sigma(\mathscr{R}) \subseteq \mathscr{F}$. Now restricting $\overline{\alpha}$ on $\sigma(\mathscr{R})$ shows the existence of a measure μ on $\sigma(\mathscr{R})$ that extends α on $\mathscr{R}$.

For the uniqueness, suppose μ' is another measure on $\sigma(\mathscr{R})$ that extends α. Then by Lemma 1.1.38, one has to have $\mu = \mu'$ over $\sigma(\mathscr{R})$. □

Theorem 1.1.37 (Dynkin's $\pi - \lambda$ theorem)**.** *Let Ω be a set. Let $\mathscr{P} \subseteq 2^{\Omega}$ be a π-system (i.e., closed under intersection) and let $\mathscr{Q} \subseteq 2^{\Omega}$ be a λ-system (i.e., $\Omega \in \mathscr{Q}$; $A, B \in \mathscr{Q}$ with $A \subseteq B \Rightarrow B \setminus A \in \mathscr{Q}$; $\mathscr{Q}$ is closed under ascending union) Then $\sigma(\mathscr{P}) \subseteq \mathscr{Q}$.*

PROOF. The proof of the above theorem is short but non-trivial (see [Dur19, Appendix A]). □

A typical application of Dynkin's $\pi - \lambda$ theorem is the following uniqueness of measure agreeing on a π-system.

Lemma 1.1.38 (Uniqueness of measure agreeing on a π-system)**.** *Let Ω be a sample space and let $\mathscr{R} \subseteq 2^{\Omega}$ be a π-system. Let μ and μ' be two measures on $\sigma(\mathscr{R})$ such that $\mu(A) = \mu'(A)$ for all $A \in \mathscr{R}$. Then $\mu(A) = \mu'(A)$ for all $A \in \sigma(\mathscr{R})$.*

PROOF. Define

$$\mathscr{Q} := \{B \in \sigma(\mathscr{R}) \mid \mu(B) = \mu'(B)\}.$$

Then one can yerify that $\mathscr{Q}$ is a λ-system containing $\mathscr{R}$. Since $\mathscr{R}$ is semi-ring, it is a π-system. Hence by the $\pi - \lambda$ theorem, we have $\sigma(\mathscr{R}) \subseteq \mathscr{Q}$. This implies that $\mu = \mu'$ over $\sigma(\mathscr{R})$, as desired. □

1.1.4. Stieltjes and Lebesgue measure on $\mathbb{R}^d$. Now we are ready to define Stieltjes measures on the Borel σ-algebra on $\mathbb{R}$. A special case of such measures is the Lebesgue measure.

Definition 1.1.39 (Stieltjes measure on $(\mathbb{R}, \mathscr{B})$)**.** Let $\Omega = \mathbb{R}$ and let $\mathscr{B}$ denote the Borel σ-algebra on $\mathbb{R}$ (see Definition 1.1.24). Recall that $\mathscr{B} = \sigma(\mathscr{R})$, where $\mathscr{R}$ is the collection of all half-open intervals $(a,b]$ in $\mathbb{R}$. A function $F : \mathbb{R} \to \mathbb{R}$ is called a *Stieltjes measure function* if

(i) F is non-degreasing; and
(ii) F is right-continuous, that is, $\lim_{y \searrow x} F(y) = F(x)$.

Fix a Stieltjes measure function $F : \mathbb{R} \to \mathbb{R}$. According to Exercise 1.1.35, we know that $\mathscr{R}$ is a semi-ring. Define a set function $\alpha : \mathscr{R} \to [0, \infty]$ as

$$\alpha((a, b]) := F(b) - F(a) \quad \text{for } a, b \in \mathbb{R} \text{ with } a \leq b.$$

Then by Theorem 1.1.36, there exists a unique measure μ on $\mathscr{B}$ that extends α. We call μ the *Stieltjes measure* associate with F. In particular, if $F(x) = x$, then the corresponding Stieltjes measure μ is called the *Lebesgue measure*. ▲

Exercise 1.1.40. Let μ denote the Lebesgue measure on the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}$. Show (using definition) that the following sets have Lebesgue measure zero:

(i) Singletons (i.e., $\{x\}$ for $x \in \mathbb{R}$)
(ii) Countable subsets of $\mathbb{R}$
(iii) The Cantor set (see Example 1.1.26).

Following the approach in Definition 1.1.39, we can also define Stieltjes measures on the Borel σ-algebra on $\mathbb{R}^d$. Again, a special case is the Lebesgue measure on the Borel sets in $\mathbb{R}^d$. For this discussion, we introduce some notation first. Let $\mathbf{a} = (a_1, \ldots, a_d), \mathbf{b} = (b_1, \ldots, b_d) \in \mathbb{R}^d$. We denote

$$\mathbf{a} \leq \mathbf{b} \quad \overset{\text{def}}{\iff} \quad a_i \leq b_i \quad \text{for all } 1 \leq i \leq d.$$

Similarly, define $\mathbf{a} < \mathbf{b}$. Also, define the 'half-open rectangle' $(\mathbf{a}, \mathbf{b}] \subseteq \mathbb{R}^d$ by

$$\begin{aligned} (\mathbf{a}, \mathbf{b}] &:= \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{a} < \mathbf{x} \leq \mathbf{b}\} \\ &= (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_d, b_d]. \end{aligned} \tag{3}$$

If $F : \mathbb{R}^d \to \mathbb{R}$ is a function, then define

$$\Delta_{(\mathbf{a},\mathbf{b}]} F := \sum_{c_i \in \{a_i, b_i\},\, i = 1 \ldots, d} (-1)^{(\#\, i\text{'s s.t. } c_i = a_i)} F(c_1, c_2, \ldots, c_d). \tag{4}$$

We can think of the above quantity as the 'volumn' of the rectangle $(\mathbf{a}, \mathbf{b}]$ induced by the 'potential function' F.

Exercise 1.1.41 (Borel σ-algebra is generated by rectangles)**.** Recall that the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}^d$ is generated by the set of all open balls in $\mathbb{R}^d$. Let $\mathscr{C}$ denote the collection of 'half-open rectangles' of the form $(\mathbf{a}, \mathbf{b}]$ in (3). Show that $\mathscr{B} = \sigma(\mathscr{C})$. (Hint: Show that any open ball is the countable union of some open balls with rational centers and rational radii; Then show every open ball with rational center and rational radii is the countable union of some half-open rectangles in (3) with rational endpoints.)

Exercise 1.1.42. This is a higher-dimensional analog of Exercise 1.1.35, which shows that the collection of half-open intervals form a semi-ring (see Definition 1.1.34). Let $\mathscr{R}$ be the collection of all half-open rectangles $(\mathbf{a}, \mathbf{b}] \subseteq \mathbb{R}^d$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ with $\mathbf{a} \leq \mathbf{b}$. Show that $\mathscr{R}$ is a semi-ring on $\mathbb{R}^d$.

Definition 1.1.43 (Stieltjes measures on $(\mathbb{R}^d, \mathscr{B})$)**.** Let $\Omega = \mathbb{R}^d$ and let $\mathscr{B}$ denote the Borel σ-algebra on $\mathbb{R}$ (see Definition 1.1.24). Recall that $\mathscr{B} = \sigma(\mathscr{R})$, where $\mathscr{R}$ is the collection of all half-open rectanbles $(\mathbf{a}, \mathbf{b}]$ in $\mathbb{R}^d$ (see Exercise 1.1.41). A function $F : \mathbb{R}^d \to \mathbb{R}$ is called a *Stieltjes measure function* if

(i) F is non-degreasing; (i.e., $F(\mathbf{a}) \leq F(\mathbf{b})$ if $\mathbf{a} \leq \mathbf{b}$)
(ii) F is right-continuous, that is, $\lim_{\mathbf{y} \searrow \mathbf{x}} F(\mathbf{y}) = F(\mathbf{x})$;
(iii) If $\mathbf{x}_n \to [-\infty, \ldots, -\infty]$ as $n \to \infty$, then $F(\mathbf{x}_n) \to 0$; If $\mathbf{x}_n \to [\infty, \ldots, \infty]$ as $n \to \infty$, then $F(\mathbf{x}_n) \to 1$;
(iv) For all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ with $\mathbf{a} \leq \mathbf{b}$, we have $\Delta_{(\mathbf{a},\mathbf{b}]} F \geq 0$ (see (4)).

Fix a Stieltjes measure function $F : \mathbb{R}^d \to \mathbb{R}$. Define a set function $\alpha : \mathscr{R} \to [0, \infty]$ as

$$\alpha((\mathbf{a}, \mathbf{b}]) := \Delta_{(\mathbf{a},\mathbf{b}]} F \quad \text{for } \mathbf{a}, \mathbf{b} \in \mathbb{R}^d \text{ with } \mathbf{a} \leq \mathbf{b}.$$

According to Exercise 1.1.42, we know that $\mathscr{R}$ is a semi-ring. Then by Theorem 1.1.36, there exists a unique measure μ on $\mathscr{B}$ that extends α. We call μ the *Stieltjes measure* associate with F. In particular, if $F(\mathbf{x}) = \prod_{i=1}^{d} F_i(x_i)$ for $\mathbf{x} = (x_1, \dots, x_d)$, where $F_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \dots, d$, then we can write

$$\alpha((\mathbf{a}, \mathbf{b}]) = \prod_{i=1}^{d} (F_i(b_i) - F_i(a_i)).$$

In particular, if $F_i(x) = x$, then

$$\alpha((\mathbf{a}, \mathbf{b}]) = \prod_{i=1}^{d} (b_i - a_i),$$

where the right-hand-side is the usual volumn of the rectangle $(\mathbf{a}, \mathbf{b}]$. In this case, the corresponding Stieltjes measure μ is called the *Lebesgue measure* on $\mathbb{R}^d$. ▲

1.2. Random variables and distributions

A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ gives a rigorous mathematical description of a random experiment. For the same random experiment, various different quantities can be observed. Random variables describe particular quantities to be observed from a given random experiment.[4]

1.2.1. Random variables. Random variables are special instances of measurable functions, which are the analogue of continuous functions in the setting of measure theory. (See Remark 1.2.7.)

Definition 1.2.1 (Measurable spaces)**.** A *measurable space* is a pair $(\Omega, \mathscr{F})$ of sample space Ω and a σ-algebra $\mathscr{F}$ on Ω.

Definition 1.2.2 (Measurable functions)**.** Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be measure spaces. Then a function (or a map) $f : \Omega \to \Omega'$ is said to be *($\mathscr{F}$-$\mathscr{F}'$)-measurable* (or *measurable*) if $f^{-1}(A') \in \mathscr{F}$ for all $A' \in \mathscr{F}'$. Here, the inverse image $f^{-1}(A')$ of $A' \in \mathscr{F}'$ under f is defined as

$$f^{-1}(A') := \{\omega \in \Omega \mid f(\omega) \in A'\}.$$

Hence, the function $f : \Omega \to \Omega'$ is measurable if the inverse image of all measurable sets are measurable. In this case, we denote $f : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$.

Definition 1.2.3 (Random variables)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $\mathscr{B}$ be the Borel σ-algebra on $\mathbb{R}^d$. A function $X : \Omega \to \mathbb{R}^d$ is a *random vector* if it is $(\mathscr{F} - \mathscr{B})$-measurable. If $d = 1$, we call X a *random variable* (RV). We also allow RVs taking values from the extended real line $\mathbb{R}^* := [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$. That is, let $\mathscr{B}^*$ denote the Borel σ-algebra on $\mathbb{R}^*$ generated by intervals of the form $(\infty, a]$, $[b, \infty)$, $[-\infty, a]$, and $[b, \infty]$. Then a RV is a measurable function $X : (\Omega, \mathscr{F}) \to (\mathbb{R}^*, \mathscr{B}^*)$.

Example 1.2.4 (Indicator function)**.** Let $(\Omega, \mathscr{F})$ be a measurable space. Fix a subset $A \subseteq \Omega$. Then the *indicator function* of A, $\mathbf{1}_A : \Omega \to \mathbb{R}$, is defined as

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

Note that $\mathbf{1}_A$ is measurable (w.r.t. $\mathscr{F}$ and the Borel σ-algebra) if and only if $A \in \mathscr{F}$. Indeed, observe that for any Borel set $B \subseteq \mathbb{R}$, there following four possible inverse images as below:

$$\mathbf{1}_A^{-1}(B) = \begin{cases} \emptyset & \text{if } 0 \notin B \text{ and } 1 \notin B \\ A & \text{if } 1 \in B \text{ and } 0 \notin B \\ A^c & \text{if } 0 \in B \text{ and } 1 \notin B \\ \Omega & \text{if } 0 \in B \text{ and } 1 \in B, \end{cases}$$

[4] Quoting Elliot Paquette (a probabilist at McGill): "As long as we can compute probabilities involving random variables, we don't really care about what probability spaces they are coming from."

Since $\emptyset, \Omega \in \mathscr{F}$ and since $\mathscr{F}$ is closed under complementation $\mathbf{1}_A$ is measurable if and only if $A \in \mathscr{F}$. ▲

Example 1.2.5 (Roll of two dice, Example 1.1.6 continued)**.** Consider the random experiment of rolling two dice in Example 1.1.6. The corresponding probability space is $(\Omega, \mathscr{F}, \mathbb{P})$ where $\Omega = \{1,2,\dots,6\}^2$, $\mathscr{F} = 2^\Omega$, and $\mathbb{P}$ is the uniform probability measure on Ω. Consider the following functions $X, Y, Z : \Omega \to \mathbb{R}$, where for each outcome $\omega = (\omega_1, \omega_2) \in \Omega$,

$$\begin{aligned} X(\omega_1,\omega_2) &:= \omega_1 \qquad \text{(outcome of the first die)} \\ Y(\omega_1,\omega_2) &:= \omega_2 \qquad \text{(outcome of the second die)} \\ Z(\omega_1,\omega_2) &:= \omega_1 + \omega_2 \qquad \text{(sum of the outcomes of two dice).} \end{aligned}$$

Then these functions are random variables. In order to see this for X, let A be any Borel subset of $\mathbb{R}$. Then $X^{-1}(A) = \{(\omega_1,\omega_2) \in \Omega \,|\, \omega_1 \in A\}$ is a subset of Ω, and hence it is an element of the power set 2^Ω. In other words, $X^{-1}(A)$ is $(2^\Omega - \mathscr{B})$-measurable, where $\mathscr{B}$ denotes the Borel σ-algebra on $\mathbb{R}$. Likewise, Y and Z are also random variables.

Now since Z pulls back Borel subsets of $\mathbb{R}$ to a measurable subset of Ω, we can compute the probability of such inverse image. For instance, recall that $\{7\} \in \mathscr{B}$. Consider the following event

$$\begin{aligned} \{Z = 7\} = Z^{-1}(\{7\}) &= \{\omega \in \Omega \,|\, Z(\omega) = 7\} \\ &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}. \end{aligned}$$

Hence

$$\mathbb{P}(Z=7) = \mathbb{P}(Z^{-1}(\{7\})) = \frac{|Z^{-1}(\{7\})|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

▲

Example 1.2.6 (Uniform RV on $[0,1]$)**.** Consider the probability space $([0,1], \mathscr{B}, \mu)$, where $\mathscr{B}$ is the Borel σ-algebra on $[0,1]$ (i.e., smallest σ-algebra generated by interavls $(a,b]$ for $0 \le a \le b \le 1$) and μ is the Lebesgue measure. Let $U : [0,1] \to \mathbb{R}$ be the identity function, $U(x) = x$. Then U is a random variable on $([0,1], \mathscr{B}, \mu)$.

In order to show $U : ([0,1], \mathscr{B}) \to (\mathbb{R}, \mathscr{B}_\mathbb{R})$ is measurable, we need to show that for any Borel subset $B \in \mathscr{B}_\mathbb{R}$, $U^{-1}(B)$ is Borel in $[0,1]$. But because the Borel σ-algebra on $\mathbb{R}$ is generated by the intervals $(a,b]$ for $a, b \in \mathbb{R}$, we only need to check this for $B = (a,b]$ for all $a, b \in \mathbb{R}$. Now in general, it holds that

$$U^{-1}(B) = B \cap [0,1].$$

So if $B = (a,b]$, then

$$B \cap [0,1] = \begin{cases} (a,b] & \text{if } 0 \le a, b \le 1 \\ [0,b] & \text{if } a < 0,\ b \in [0,1] \\ (a,1] & \text{if } a \in [0,1],\ b > 1 \\ \emptyset & \text{if } a, b < 0 \text{ or } a, b > 1. \end{cases}$$

Hence in all cases, $B \cap [0,1]$ is Borel in $[0,1]$. This shows $U : ([0,1], \mathscr{R}) \to (\mathbb{R}, \mathscr{B}_\mathbb{R})$ is measurable. Hence U is a RV.

Also, the CDF of U is the identity on $[0,1]$, as

$$\mu(U \le x) = \mu(\{y \in [0,1] \,|\, U(y) \le x\}) = \mu(\{y \in [0,1] \,|\, y \le x\}) = \mu([0,x]) = x.$$

In this case, we denote $U \sim \text{Uniform}([0,1])$. ▲

It is helpful to compare the definition of measurable functions to that of continuous functions, as we discuss in the following remark.

Remark 1.2.7 (σ-algebra and topology)**.** Fix a set Ω. A σ-algebra $\mathscr{F}$ on Ω is a system of subsets of Ω that gives rise to the notion of 'measurability'. Another well-known system of subsets in mathematics is called the 'topology', which gives rise to the notion of 'open subsets'. It is helpful to compare the definition of σ-algebra and topology on Ω. Recall that a collection $\mathscr{T}$ of subsets of Ω is a *topology* on Ω if it satisfies the following conditions:

(i) $\emptyset, \Omega \in \mathscr{T}$;
(ii) $\mathscr{T}$ is closed under arbitrary union (not necessarily only the countable ones);
(iii) $\mathscr{T}$ is closed under finite intersection.

Note that the above properties are abstracted from the collection of open balls in the Euclidean space $\mathbb{R}^n$. A pair $(\Omega, \mathscr{T})$ of sample space Ω and a topology is called a *topological space*. If we have two topological spaces $(\Omega, \mathscr{T})$ and $(\Omega', \mathscr{T}')$, then a function $f : \Omega \to \Omega'$ is said to be *continuous* if $f^{-1}(A') \in \mathscr{T}$ for all $A' \in \mathscr{T}$ (i.e., the inverse image of open subsets is open).

Proposition 1.2.8 (Check measurability on basis)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(S, \mathscr{R})$ be a measurable space. A function $X : \Omega \to S$ is $(\mathscr{F} - \mathscr{R})$-measurable if there exists a collection $\mathscr{A}$ of subsets of S such that $\mathscr{R} = \sigma(\mathscr{A})$ and $X^{-1}(A) \in \mathscr{F}$ for all $A \in \mathscr{A}$.*

PROOF. Let $\mathscr{Q} := \{B \in \mathscr{R} \mid X^{-1}(B) \in \mathscr{F}\}$, which is the collection of $\mathscr{R}$-measurable subsets of S that is pulled back by X to a $\mathscr{F}$-measurable set. It is easy to see that $\mathscr{Q}$ is a σ-algebra. Indeed, if $B \in \mathscr{Q}$, then $X^{-1}(B^c) = \Omega \setminus X^{-1}(B) \in \mathscr{F}$, so $\mathscr{Q}$ is closed under complementation. Also, if $B_1, B_2, \cdots \in \mathscr{R}$, then

$$X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty} X^{-1}(B_i) \in \mathscr{F}, \qquad \text{(why?)}$$

so it verifies $\bigcup_{i=1}^{\infty} B_i \in \mathscr{Q}$. Now by the hypothesis, $\mathscr{A} \subseteq \mathscr{Q}$. It follows that $\mathscr{R} = \sigma(\mathscr{A}) \subseteq \mathscr{Q} \subseteq \mathscr{R}$. Hence $\mathscr{Q} = \mathscr{R}$, as desired. □

Example 1.2.9 (Monotone functions are Borel measurable)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. For instance, we assume f is non-decreasing, that is, $f(x) \le f(y)$ if $x \le y$. Then $f : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$, that is, f is Borel measurable. Indeed, note that

$$f^{-1}((-\infty, a]) = \{x : f(x) \le a\} = (-\infty, b],$$

where $b := \sup\{y : f(y) \le a\}$. Now $b \in [-\infty, \infty)$. Hence the inverse image of intervals under f^{-1} is an interval, which is a Borel set. Hence f is Borel measurable by Proposition 1.2.8. ▲

Proposition 1.2.10 (Building RVs from other RVs)**.** *These followings hold:*

(i) *(Composition I) If $X : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ and $f : (\Omega', \mathscr{F}') \to (\Omega'', \mathscr{F}'')$ are measurable maps, then $f(X) = f \circ X$ is a measruable map $(\Omega, \mathscr{F}) \to (\Omega'', \mathscr{F}'')$.*

(ii) *(Composition II) If $X_1, \ldots, X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ are RVs and if $f : (\mathbb{R}^n, \mathscr{B}^n) \to (\mathbb{R}, \mathscr{B})$ is measurable, then $f(X_1, \ldots, X_n)$ is a RV.*

(iii) *(Sum and product) If $X_1, \ldots, X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ are RVs, then $\sum_{i=1}^n X_i$ and $\prod_{i=1}^n X_i$ are also RVs.*

(iv) *(One-sided imits) If $X_1, X_2 \cdots : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ are RVs, then the following are also RVs:*

$$\inf_{n \ge 1} X_n, \qquad \sup_{n \ge 1} X_n, \qquad \liminf_{n \to \infty} X_n, \qquad \limsup_{n \to \infty} X_n,$$

PROOF. (i) Let $B'' \in \mathscr{F}''$ be arbitrary. Then $f^{-1}(B'') \in \mathscr{F}'$ since $f : (\Omega', \mathscr{F}') \to (\Omega'', \mathscr{F}'')$ is measurable, and also $X^{-1}(f^{-1}(B'')) \in \mathscr{F}$ since $X : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ is measurable.

(ii) According to **(i)**, it suffices to show that $(X_1, \ldots, X_n) : (\Omega, \sigma(\mathscr{F})) \to (\mathbb{R}^n, \mathscr{B}^n)$ is measurable. Recall that the Borel σ-algebra $\mathscr{B}^n$ on $\mathbb{R}^n$ is generated by measurable rectangles (see Exercise 1.1.41). Now for $A_1, \ldots, A_n$ intervals in $\mathbb{R}$, note that

$$(X_1, \ldots, X_n)^{-1}(A_1 \times \cdots \times A_n) = \{(X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n\} = \bigcap_{i=1}^n X_i^{-1}(A_i) \in \mathscr{F}.$$

Hence by Proposition 1.2.8, $(X_1, \ldots, X_n) : (\Omega, \sigma(\mathscr{F})) \to (\mathbb{R}^n, \mathscr{B}^n)$ is indeed measurable.

(iii) According to **(ii)**, it suffices to verify that the functions $f:(x_1,\dots,x_n)\mapsto x_1+\cdots+x_n$ and $g:(x_1,\dots,x_n)\mapsto x_1\cdots x_n$ are measurable. By Proposition 1.2.8, one only needs to check if the inverse image of open intervals under these maps are Borel measurable. Indeed,

$$\{(x_1,\dots,x_n)\in\mathbb{R}^n \mid x_1+\cdots+x_n<a\}=f^{-1}((-\infty,a))$$

is open in $\mathbb{R}^n$ so it is Borel. Thus f Borel measurable. Also, g is a continuous function so the inverse image of an open set is open. Hence $g^{-1}((-\infty,a))$ is open in $\mathbb{R}^d$ so it is Borel. Thus g is Borel measurable.

(iv) Note that for each $a\in\mathbb{R}$,

$$\left\{\inf_{n\geq1}X_n<a\right\}=\bigcup_{n=1}^{\infty}\{X_n<a\},\qquad\left\{\sup_{n\geq1}X_n<a\right\}=\bigcap_{n=1}^{\infty}\{X_n<a\}.$$

Hence $\inf_{n\geq1}X_n$ and $\sup_{n\geq1}X_n$ are RVs. Next, recall that $\liminf_{n\to\infty}X_n$ and $\limsup_{n\to\infty}X_n$ are the smallest and the largest subsequential limits of X_n, respectively. Hence we can write

$$\liminf_{n\to\infty}X_n=\sup_{n\geq1}\left(\inf_{m\geq n}X_m\right),\qquad\limsup_{n\to\infty}X_n=\inf_{n\geq1}\left(\sup_{m\geq n}X_m\right).$$

Since we have verified that $\inf_{n\geq1}X_n$ and $\sup_{n\geq1}X_n$ are RVs, their supremum and infimum are also RVs. Hene the above are also RVs.

□

Suppose we have sequence of RVs $X_1,X_2,\dots$ on a probability space $(\Omega,\mathscr{R},\mathbb{P})$. From Proposition 1.2.10, it follows that

$$\Omega_0:=\left\{\omega\in\Omega\,\middle|\,\lim_{n\to\infty}X_n(\omega)\text{ exists}\right\}=\left\{\omega\in\Omega\,\middle|\,\liminf_{n\to\infty}X_n(\omega)=\limsup_{n\to\infty}X_n(\omega)\right\}\in\mathscr{F}.$$

If $\mathbb{P}(\Omega)=1$, then we say X_n converges *almost surely* as $n\to\infty$. Note that the limiting RV $X_\infty:=\lim_{n\to\infty}X_n$ has well-defined values only on Ω_0. That is, $X_\infty(\omega)\in[-\infty,\infty]$ if $\omega\in\Omega_0$ and $X_\infty(\omega)$ does not exist.

1.2.2. Distributions. Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and let $X:\Omega\to\mathbb{R}$ be a random variable. Note that if $A\subseteq\mathbb{R}$ is a Borel set, then since X is a random variable, we have $X^{-1}(A)\in\mathscr{F}$. Hence we can measure the 'size' of $X^{-1}(A)$ using the probability measure $\mathbb{P}$ on $\mathscr{F}$.

Definition 1.2.11 (Distribution)**.** Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and let $X:\Omega\to\mathbb{R}$ be a random variable. The *distribution* of X is the probability measure μ on $(\mathbb{R},\mathscr{B})$, where $\mathscr{B}=$ Borel σ-algebra on $\mathbb{R}$, defined as

$$\mu(A):=\mathbb{P}(X^{-1}(A))=\mathbb{P}(X\in A)\qquad\textit{for all Borel subset } A\subseteq\mathbb{R}.$$

The *(cumulative) distribution function* (CDF) of X is the function $F:\mathbb{R}\to[0,1]$ defined by $F(x):=\mathbb{P}(X\leq x)$ for $x\in\mathbb{R}$. If the CDF of X has the integral representation[5]

$$\mathbb{P}(X\leq x)=\int_{-\infty}^{x}f_X(x)\,dx\quad\forall\,x\in\mathbb{R}$$

for some function $f_X:\mathbb{R}\to\mathbb{R}$, then we call the function f_X the *probability density function* of X and say X is a *continuous RV*. If there exists a countable subset $S\subset\mathbb{R}$ such that $\mathbb{P}(X\in S)=1$, then X is a *discrete RV*. In this case, if we write $S=\{x_1,x_2,\dots,\}$, then we can write $X=\sum_{i=1}^{\infty}x_i\mathbf{1}(X=x_i)$[6].

[5] We have not actually defined Lebesgue integral yet. For now regard integrals in the sense of Riemann integral.

[6] Equivalently, X is a disrete RV iff the CDF of X is a stepfunction with countably many jumps.

The set function $\mu : \mathscr{B} \to [0,\infty]$ by $\mu(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A)$ defined in is indeed a probability measure on $\mathscr{B}$. In order to see this, first note that $\mu(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = 1$, so $\mu(\emptyset) = 1 - \mu(\mathbb{R}) = 0$; second, if $A_1, A_2, \cdots \in \mathscr{B}$ are disjoint, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(X \in \bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} X^{-1}(A_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(X^{-1}(A_i)\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Proposition 1.2.12. *Let $F : \mathbb{R} \to \mathbb{R}$ be the CDF of a random variable X. It satisfies the following properties.*

(i) *F is non-decreasing;*
(ii) $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$*;*
(iii) *F is right-continuous, that is,* $\lim_{y\searrow x} F(y) = F(x)$*;*
(iv) *Let* $F(x-) := \lim_{y\nearrow x} F(y)$*. Then* $F(x-) = \mathbb{P}(X < x)$*;*
(v) $\mathbb{P}(X = x) = F(x) - F(x-)$.

PROOF. We assume the RV X is from some probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

(i) Follows from the monotonicity of probability measures and $\{X \le x\} \subseteq \{X \le y\}$ if $x \ge y$.

(ii) By continuity of measures, $\lim_{x\to\infty} \mathbb{P}(X \le x) = \mathbb{P}(X^{-1}((-\infty,\infty))) = \mathbb{P}(\Omega) = 1$. Then $\lim_{x\to-\infty} \mathbb{P}(X \le x) = \mathbb{P}(X^{-1}(\emptyset)) = \mathbb{P}(\emptyset) = 0$.

(iii) Note that if a sequence $a_n \searrow 0$ as $n \to \infty$, then $\bigcap_{n=1}^{\infty} \{X \le x + a_n\} = \{X \le x\}$. So by continuity of measure, we get $\lim_{n\to\infty} \mathbb{P}(X \le x + a_n) = \mathbb{P}(X \le x)$. Since this holds for any sequence $a_n \searrow 0$, it holds that $\lim_{y\searrow x} F(y) = F(x)$.

(iv) Fix a sequence $a_n \searrow 0$ as $n \to \infty$. Then $\bigcup_{n=1}^{\infty} \{X \le x - a_n\} = \{X < x\}$. So by continuity of measure, we get $\lim_{n\to\infty} \mathbb{P}(X \le x - a_n) = \mathbb{P}(X < x)$. Since this holds for any sequence $a_n \searrow 0$, it holds that $\lim_{y\nearrow x} F(y) = F(x-)$.

(v) By **(iv)**, $F(x) - F(x-) = \mathbb{P}(X \le x) - \mathbb{P}(X < x) = \mathbb{P}(X = x)$.

□

Proposition 1.2.13 (Constructing RV from its CDF)**.** *If a function $F : \mathbb{R} \to \mathbb{R}$ satisfies Proposition 1.2.12* **(i)**-**(iii)**, *then F is the CDF of some random variable X.*

PROOF. Consider the probability space $(\Omega, \mathscr{B}, \mu)$, where $\Omega = [0,1]$, $\mathscr{B}$ is the Borel σ-algebra on $[0,1]$, and μ is the Lebesgue measure. Now define a function $X : [0,1] \to \mathbb{R}$ by

$$X(\omega) := \sup\{y : F(y) < \omega\} = \text{“}F^{-1}(\omega)\text{”}.$$

(Note that If F is 1-1, then $X(\omega) = F^{-1}(\omega)$.) We need to verify two points: (1) X is a random variable; (2) The CDF of X is F. To this end, we claim that

$$\{\omega : X(\omega) \le x\} = \{\omega : \omega \le F(x)\}. \tag{5}$$

Note that this yields $X^{-1}((-\infty, x]) = [0, F(x)]$. That is, the inverse image of the interval $[0,x]$ under X is the interval $[0,F(x)]$, which is a Borel set. This also yields that $X^{-1}([a,b]) = (F(a), F(b)]$, which is also a Borel set. Hence by Proposition 1.2.8, X is indeed a random variable. Furthermore, assuming (5), we can easily verify that the CDF of X is indeed F:

$$\mu(X \le x) = \mu(\omega \le F(x)) = \mu([0, F(x)]) = F(x).$$

It remains to verify (5). Indeed, on the one hand, suppose $X(\omega) \le x$. Suppose for contradiction that $F(x) < \omega$. Then by using the right-continuity of F, there exists $\varepsilon > 0$ such that $F(x+\varepsilon) < \omega$. It follows that $x + \varepsilon \le X(\omega) \le x$, which is a contradiction. On the other hand, suppose $\omega \le F(x)$. Suppose for contradiction that $X(\omega) > x$. This implies that there exists $x < y \le X(\omega)$ such that $F(y) < \omega$. But since F is non-decreasing and $x < y$, $F(y) \ge F(x) \ge \omega$, which is a contradiction. Thus $X(\omega) \le x$. □

Remark 1.2.14. If the function F in Proposition 1.2.13 is strictly increassing so that it has an inverse function, then the RV X we constructed in the proof of Proposition 1.2.13 is simply

$$X = F^{-1}(U),$$

where U is the uniform RV on $[0,1]$ in Example 1.2.6. Note that X is a RV since F^{-1} is monotone and hence measurable (see Example 1.2.9). Also, indeed its CDF is F:

$$\mathbb{P}(X \le x) = \mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)) = F(x).$$

This is in fact how random variables are generated numerically in computer softwares. Namely, it first genreates a uniform RV U and then outputs $F^{-1}(U)$.

Exercise 1.2.15. Using your favorate programing languate (e.g., `R`, `python`, `C++`), sample i.i.d. uniform RVs $U_1, \dots, U_n \sim \text{Uniform}([0,1])$. Compute the values of $X_1, \dots, X_n$, where $X_i := g(U_i)$, where $g(x) = -\lambda^{-1}\log(1-x)$ for some fixed $\lambda > 0$. Numerically compute and plot the empirical CDF of the samples $X_1, \dots, X_n$. What is the distribution of $X := g(U)$, $U \sim \text{Uniform}([0,1])$?

Example 1.2.16 (Uniform distribution on $[0,1]$)**.** Let $f(x) := \mathbf{1}_{[0,1]}(x)$ and let $F(x) := \int_{-\infty}^{x} f(x)\,dx$. Then

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x \le 1 \\ 1 & \text{if } x > 1 \end{cases}$$

is called the *uniform distrubution* on $[0,1]$. ▲

Example 1.2.17 (Exponential distribution with rate λ)**.** Fix a constant $\lambda > 0$ and let $f(x) := \lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x)$ and let $F(x) := \int_{-\infty}^{x} f(x)\,dx$. Then

$$F(x) = \left(1 - e^{-\lambda x}\right)\mathbf{1}_{[0,\infty)}(x)$$

is called the *exponential distrubution with rate* λ. ▲

Example 1.2.18 (Standard normal distribution on $[0,1]$)**.** Let $f(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Then $F(x) := \int_{-\infty}^{x} f(x)\,dx$ is called the *standard normal distrubution*. There is no closed-form expression for F in this case. ▲

Exercise 1.2.19. Let $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx$ be the standard normal distribution. Show that for all $x > 0$,

$$(x^{-1} - x^{-3})\exp(-x^2/2) \le 1 - F(x) \le x^{-1}\exp(-x^2/2).$$

Definition 1.2.20 (Equivalence in distribution)**.** Let X, Y be RVs, possibly defined on different probability spaces. We say X and Y are *equal in distribution* and write $X \overset{d}{=} Y$ if they have the same CDF (equivalently, distribution):[7]

$$X \overset{d}{=} Y \quad \overset{\text{def}}{\iff} \quad \text{CDF of } X = \text{CDF of } Y \quad \iff \quad \mathbb{P}(X \le x) = \mathbb{P}(Y \le x)\ \forall x \in \mathbb{R}. \tag{6}$$

Exercise 1.2.21. Construct an example of two random variables X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ where $X \overset{d}{=} Y$ but $\mathbb{P}(X = -Y) = 1$.

Definition 1.2.22 (Absolute continuity)**.** Let $(\Omega, \mathscr{F})$ be a measurable space and let μ, ν be two measures on it. We say ν is *absolutely continuous* with respect to μ and write $\nu \ll \mu$ if

$$\mu(A) = 0 \quad \Longrightarrow \quad \nu(A) \quad \forall A \in \mathscr{F}.$$

A measure ρ on $(\mathbb{R}, \mathscr{B})$ is *absolutely continuous* if it is absolutely continuous w.r.t. the Lebesgue measure; otherwise ρ is said to be *singular*. A measure ξ on $(\Omega, \mathscr{F})$ is *discrete* if there exists a countable set $S \subseteq \Omega$

Example 1.2.23 (Point mass)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a probability space and suppose that $\mu(\{y\}) = 1$ for some $y \in \Omega$. Then μ is called the *point mass at* y and denoted as δ_y. If $\Omega = \mathbb{R}$, then the point mass at y corresponds to the distribution function $F(x) = \mathbf{1}(y \ge x)$. Clearly point masses are discete probability measures. ▲

[7] If $X : (\Omega, \mathscr{F}, \mathbb{P}_X) \to (\mathbb{R}, \mathscr{B})$ and $Y : (\Omega, \mathscr{F}, \mathbb{P}_Y) \to (\mathbb{R}, \mathscr{B})$, then the last expression in (6) precisely means $\mathbb{P}_X(X \le x) = \mathbb{P}_Y(Y \le x)\ \forall x \in \mathbb{R}$. However, we usually use generic notation $\mathbb{P}$ for probability measures and interpret it as the probability measure on the appropriate probability space.

Example 1.2.24 (Uniform distribution on the Cantor set)**.** Recall the Cantor set C defined in Example 1.1.26, which is Borel-measurable with Lebesgue measure zero. We are going to consider the 'uniform probability distribution' on C. Namely, let C_n be the set obtained from $[0,1]$ by repeating the procedure of omitting the middle third of every remaining interval n times. Since C_n is a disjoint union of 2^n intervals, one can consider the uniform distribution on C_n with distribution denoted as F_n. Namely,

$$F_n(x) := \frac{1}{\mu(\mathsf{C}_n)} \int_0^x \mathbf{1}(x \in \mathsf{C}_n)\, dx = (3/2)^n \int_0^x \mathbf{1}(x \in \mathsf{C}_n)\, dx.$$

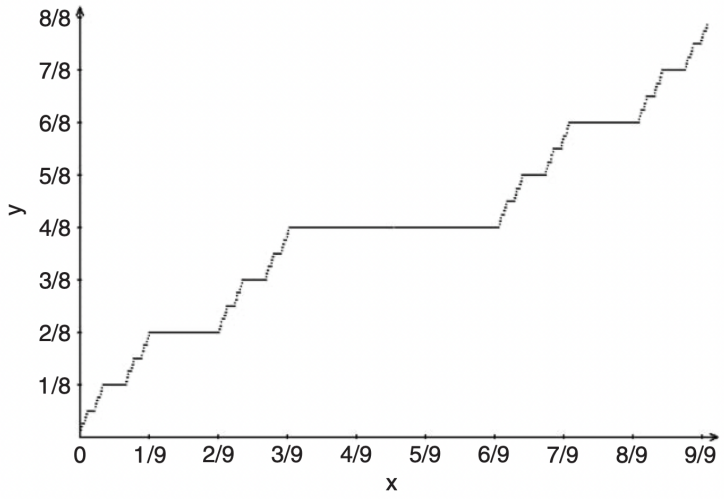


FIGURE 1.2.1. CDF of the Cantor distribution, which is also known as the "devil's staircase". Figure excerpted from [DMRV06].

Note that F_n is a piecewise linear function with slopes either 0 or $3/2$. It can be shown that F_n converges to some limiting function F_∞ pointwise. Since F_n is monotone, the convergence $F_n \to F$ is uniform, so the limiting function F is also continuous. Note that F_∞ cannot have a density function, since $F_\infty = 0$ on C^c and $\mu(\mathsf{C}) = 1$. This means any density function (if exists) of F_∞ has to be zero on the set C^c of measure 1, so it integrates over $[0,1]$ to zero, not one. A distribution function not having a density function is in fact equivalent to the associated Stieltjes measure being singular. This uses the famous Radon-Nikodym theorem (pointer will be added). ▲

1.3. Integration

1.3.1. Definition of Lebesgue integral and basic properties. In this section, we develop the theory of Lebesgue integration. One should be familar with the notion of Riemann integral, where one partitions the domain of the function into intervals or rectangles and approximates the area or volume under the graph of function by the sum of rectangles. On the contrary, the key point in Lebesgue integral is to *partition the range of the function*, rather than its domain (see Figure 1.3.1).

An important building block of Lebesgue integral is simple function, which will play the role of rectangles in Riemann integral.

Definition 1.3.1 (Simple functions)**.** Let $(\Omega, \mathscr{F}, \mu)$ is a measure space. A function $\varphi : \Omega \to \mathbb{R}$ is *simple* if it takes only finitely many values on sets of finite measures. That is, there exists disjoint measurable sets $A_1, \dots, A_n \in \mathscr{F}$ with $\mu(A_i) < \infty$ for $i = 1, \dots, n$ and $a_1, \dots, a_n \in \mathbb{R}$ such that

$$\varphi = \sum_{i=1}^m a_i \mathbf{1}_{A_i}.$$

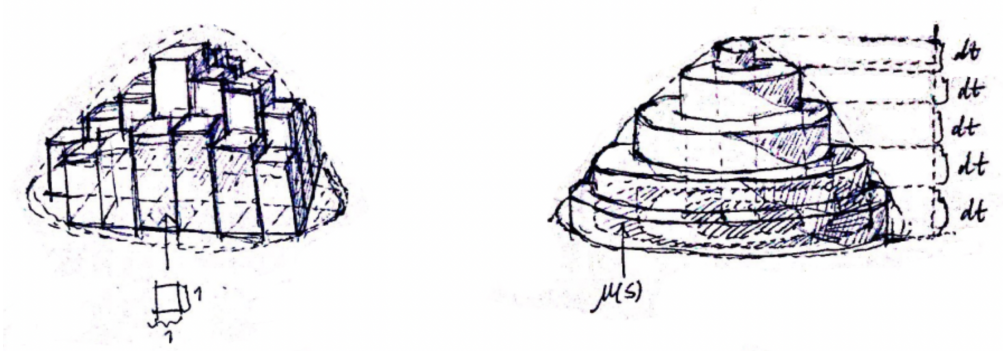


FIGURE 1.3.1. Riemann integral (left) partitions the domain of the function, whereas the Lebesgue integral partitions the range of the function. Figure by Xichu Zhang.

Definition 1.3.2 (μ-almost everywhere)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f, g : \Omega \to \mathbb{R}$ be measurable functions. We say $f \geq g$ μ-almost everywhere (μ-a.e.) if $\mu\big(\{\omega \mid f(\omega) < g(\omega)\}\big) = 0$ (i.e., the exceptional set has μ-measure zero).

Definition 1.3.3 (Lebesgue integral)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Let $f : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ be a measurable function. We will define Lebuesgue integral $\int f \, d\mu$ in four steps:

(I) (*Simple functions*); Given a simple function $\varphi = \sum_{i=1}^{m} a_i \mathbf{1}_{A_i}$,

$$\int \varphi \, d\mu := \sum_{i=1}^{n} a_i \mu(A_i).$$

(II) (*Bounded functions*); Let $f : \Omega \to \mathbb{R}$ be a bounded measurable function vanishing outside of some set $E \in \mathscr{F}$ with $\mu(E) < \infty$. Then

$$\int f \, \mu := \sup_{\varphi \leq f} \int \varphi \, d\mu,$$

where the supermum runs over all simple functions $\varphi \leq f$.

(III) (*Nonnegative functions*); If $f \geq 0$ μ-a.e., then

$$\int f \, \mu := \sup_{0 \leq g \leq f} \int g \, d\mu,$$

where the supermum runs over all bounded functions $0 \leq g \leq f$ such that $\mu(\{\omega \mid g(\omega) > 0\}) < \infty$.

(IV) (*General measurable functions*) Let $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$ for $a, b \in \mathbb{R}$. For a general measurable function $f : \Omega \to \mathbb{R}$, define two nonnegative functions $f^+, f^- : \Omega \to \mathbb{R}$ by[8]

$$f^+(\omega) := f(\omega) \vee 0, \qquad f^{-1}(\omega) := -\big(f(\omega) \wedge 0\big).$$

Then we have

$$f = f^+ - f^- \qquad \text{and} \qquad |f| = f^+ + f^-.$$

We say f is *(Lebesgue) integrable*[9] if f is measurable and $\int |f| \, d\mu < \infty$, or equivalently, $\int f^+ \, d\mu < \infty$ and $\int f^- \, d\mu < \infty$. In this case, we define

$$\int f \, d\mu := \int f^+ \, d\mu - \int f^- \, d\mu.$$

Otherwise, $\int f \, d\mu$ is undefined and we say $\int f \, d\mu$ *does not exist.*

[8] The ReLU function $x \mapsto x \vee 0$ is monotone and hence measurable. Hence f^+ is measurable being the composition of two measurable functions. f^- is also measurable for a similar reason.

[9] $\int f \, d\mu$ is *defined* iff $\int f \, d\mu \in [-\infty, \infty]$ iff [either $\int f^+ \, d\mu < \infty$ or $\int f^- \, d\mu < \infty$].

Remark 1.3.4 (Integral notations)**.** Here we introduce some notations for Lebesgue integrals for measure space $(\Omega, \mathscr{F}, \mu)$ and a measurable function $f : \Omega \to \mathbb{R}$.

(i) For an event $A \subseteq \Omega$, we write

$$\int_A f \, d\mu := \int f \mathbf{1}_A \, d\mu,$$

where we call the RHS above s the 'integral of f over A'.

(ii) When $(\Omega, \mathscr{F}, \mu) = (\mathbb{R}^d, \mathscr{B}^d, \lambda)$, where λ = Lebesgue measure, we write $\int f \, d\mu = \int f(\mathbf{x}) \, d\mathbf{x}$.

(iii) When $(\Omega, \mathscr{F}, \mu) = (\mathbb{R}, \mathscr{B}, \lambda)$, where λ = Lebesgue measure, we write $\int_{[a,b]} f \, d\mu = \int_a^b f(x) \, dx$.

(iv) When $(\Omega, \mathscr{F}, \mu) = (\mathbb{R}, \mathscr{B}, \mu)$ and $\mu((a,b]) = G(b) - G(a)$ for $a < b$[10], then we write $\int f \, d\mu = \int f(x) \, dG(x)$.

(v) If Ω is countable, $\mathscr{F} = 2^\Omega$, and μ is the counting measure[11], then we write $\int f \, d\mu = \sum_{x \in \Omega} f(x)$.

Exercise 1.3.5. Let $f : (\Omega, \mathscr{F}, \mu) \to (\mathbb{R}, \mathscr{B})$ be a measurable function such that $f = 0$ μ-a.e.. Show that $\int f \, d\mu = 0$. Deduce that, if $A \in \mathscr{F}$ with $\mu(A) = 0$ and f is arbitrary measurable function, then $\int_A f \, d\mu = 0$. Conversely, if $\int |f| \, d\mu = 0$, then show that $f = 0$ μ-a.e..

One should get familar with this order of development. This is how we usually prove certain identities (e.g., linearity of integral, change of variables formula, Fubini's theorem) for Lebesgue integral.

Proposition 1.3.6 (Upper and lower Lebesgue sum)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be a bounded measurable function such that $\mu(\{f \neq 0\}) < \infty$. Then*

$$\sup_{\varphi \text{ simple} \le f} \int \varphi \, d\mu = \inf_{\phi \text{ simple} \ge f} \int \varphi \, d\mu.$$

PROOF. First note that a simple function $\varphi \le f$ has to vanish outside E since $f \equiv 0$ on E. Also, without loss of generality, we may assume that the simple functions $\phi \ge f$ in the statement vanishes outside E. We first show

$$\sup_{\varphi \text{ simple} \le f} \int \varphi \, d\mu \le \inf_{\phi \text{ simple} \ge f} \int \varphi \, d\mu. \tag{7}$$

Fix simple functions $\varphi \le f$ and $\phi \ge f$. Write $\varphi = \sum a_i \mathbf{1}_{A_i}$ and $\phi = \sum b_j \mathbf{1}_{B_j}$. The support of $\varphi + \phi$ is contained in $\bigcup_i A_i \cup \bigcup_j B_j$. We can rewrite this union as the disjoint union of the sets $B_j \setminus \bigcup_i A_i$, $A_i \setminus \bigcup_j B_j$, and $A_i \cup B_j$, for $i, j \ge 1$. Enumerate all these sets as C_k for $k = 1, \dots, n$ for some $n \ge 1$. Then we can write $\varphi = \sum_{k=1}^n x_k \mathbf{1}_{C_k}$ and $\phi = \sum_{k=1}^n y_k \mathbf{1}_{C_k}$. Since $\varphi \le f \le \phi$, we have $x_k \le y_k$ for $k = 1, \dots, n$. Then

$$\int \varphi \, d\mu = \sum_{k=1}^n x_k \mathbf{1}_{C_k} \le \sum_{k=1}^n y_k \mathbf{1}_{C_k} = \int \phi \, d\mu.$$

This holds for all simple functions $\varphi \le f$ and $\phi \ge f$. Hence taking supremum on the left hand side and then taking infimum on the right hand side, we obtain (7).

To prove the other inequality, we suppose $|f| < M$ for some $M > 0$. For each $m \ge 1$, define

$$E_{\ell,m} := \left\{ \omega \in E \,\middle|\, \frac{\ell M}{m} \le f(\omega) < \frac{(\ell+1)M}{m} \right\} = f^{-1}\big([\ell M/m, (\ell+1)M/m)\big) \quad \text{for } -m \le \ell \le m,$$

$$\varphi_m := \sum_{\ell=-m}^{m} \frac{\ell M}{m} \mathbf{1}_{E_k}, \qquad \phi_m := \sum_{\ell=-m}^{m} \frac{(\ell+1)M}{m} \mathbf{1}_{E_k}.$$

By definition $\phi_m - \varphi_m = (M/m)\mathbf{1}_E$, so we have

$$\begin{aligned} \sup_{\varphi \text{ simple} \le f} \int \varphi \, d\mu \ge \int \varphi_m \, d\mu &= \int \phi_m \, d\mu - (M/m)\mu(E) \\ &\ge \left(\inf_{\phi \text{ simple} \ge f} \int \varphi \, d\mu \right) - (M/m)\mu(E). \end{aligned}$$

[10] G is a Stieltjes measure function

[11] See Example 1.1.4.

Now since $\mu(E) < \infty$, we can take $m \to \infty$ and obtain the desired inequality. □

The main goal of this section is to establish the following basic properties of Lebesgue integral, which should be familar from Riemann integrals.

Theorem 1.3.7 (Basic properties of integral)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f, g : \Omega \to \mathbb{R}$ be integrable functions such that their supports $\{f \neq 0\}$ and $(\{g \neq 0\})$ are σ-finite*[12]*. Then the following holds:*

(i) *If $f \geq 0$ μ-a.e., then $\int f \, d\mu \geq 0$;*

(ii) *For all $a \in \mathbb{R}$, $\int af \, d\mu = a \int f \, d\mu$;*

(iii) $\int f + g \, d\mu = \int f \, d\mu + \int g \, d\mu$;

(iv) *If $g \leq f$ μ-a.e., then $\int g \, d\mu \leq \int f \, d\mu$;*

(v) *If $\left|\int f \, d\mu\right| \leq \int |f| \, d\mu$.*

PROOF OF THEOREM 1.3.7 FOR BOUNDED OR NONNEGATIVE FUNCTIONS. We will first verify each statements for simple functions and then move on to bounded functions and then to nonnegative functions.

(i) It holds clearly if f is a nonnegative simple function. If f is a bounded function that vanishes outside some set E with $\mu(E) < \infty$, then it also holds by definition since we can take the zero simple function $\varphi \equiv 0$. Lastly, if $f \geq 0$ μ-a.e., then for any bounded measurable function g such that $0 \leq g \leq f$, we have $\mu(g > 0) \leq \mu(f > 0) < \infty$, so $\int g \, d\mu \geq 0$ by the previous case. Taking supremum over all such g, this shows $\int f \, d\mu \geq 0$.

(ii) It holds clearly if f is a simple function. Also, one can easily verify it when $a = 0$ case by case, so we may assume $a \neq 0$. Next, suppose f is a bounded function that vanishes outside some set E with $\mu(E) < \infty$. Suppose $a > 0$. Then φ is a simple function with $\varphi \leq f$ if and only if $a\varphi$ is a simple function with $a\varphi \leq af$. Hence

$$a \int f \, d\mu = a \sup_{\varphi \leq f} \int \varphi \, d\mu = \sup_{\varphi \leq f} a \int \varphi \, d\mu = \sup_{\varphi \leq f} \int a\varphi \, d\mu = \sup_{a\varphi \leq af} \int a\varphi \, d\mu = \int af \, d\mu.$$

If $a < 0$, then noting that φ is a simple function with $\varphi \leq f$ if and only if $a\varphi$ is a simple function with $a\varphi \geq af$, using Proposition 1.3.6, we get

$$a \int f \, d\mu = a \sup_{\varphi \leq f} \int \varphi \, d\mu = \inf_{\varphi \leq f} a \int \varphi \, d\mu = \inf_{\varphi \leq f} \int a\varphi \, d\mu = \inf_{a\varphi \geq af} \int a\varphi \, d\mu = \int af \, d\mu.$$

Next, suppose $f \geq 0$ μ-a.e.. In this case, we will verify **(ii)** for only the case that $a > 0$ (the case $a < 0$ will be handle later). Note that h is a bounded measurable function with $\mu(\{h \neq 0\}) < \infty$ with $0 \leq h \leq f$ if and only if $0 \leq ah \leq af$ and $\mu(\{ah \neq 0\}) < \infty$. Hence

$$a \int f \, d\mu = a \sup_{0 \leq h \leq f} \int \varphi \, d\mu = \sup_{0 \leq h \leq f} a \int h \, d\mu = \sup_{0 \leq h \leq f} \int ah \, d\mu = \sup_{ah \leq af} \int ah \, d\mu = \int af \, d\mu.$$

(iii) For simple functions, write $\varphi = \sum a_i \mathbf{1}_{A_i}$ and $\phi = \sum b_j \mathbf{1}_{B_j}$. The support of $\varphi + \phi$ is contained in $\bigcup_i A_i \cup \bigcup_j B_j$. We can rewrite this union as the disjoint union of the sets $B_j \setminus \bigcup_i A_i$, $A_i \setminus \bigcup_j B_j$, and $A_i \cup B_j$, for $i, j \geq 1$. Enumerate all these sets as C_k for $k = 1, \ldots, n$ for some $n \geq 1$. Then we can write $\varphi = \sum_{k=1}^n x_k \mathbf{1}_{C_k}$ and $\phi = \sum_{k=1}^n y_k \mathbf{1}_{C_k}$. Then $\varphi + \phi = \sum_{i=1}^n (x_k + y_k) \mathbf{1}_{E_k}$. Hence

$$\int \varphi + \phi \, d\mu = \sum_{k=1}^n (x_k + y_k) \mu(E_k) = \sum_{k=1}^n x_k \mu(E_k) + \sum_{k=1}^n y_k \mu(E_k) = \int \varphi \, d\mu + \int \phi \, d\mu.$$

Hence the assertion holds for simple functions.

[12] Given a measure space $(\Omega, \mathscr{F}, \mu)$, a set $A \subseteq \Omega$ is *σ-finite* if there exists countable events $E_n \in \mathscr{F}$, $n \geq 1$ with $\mu(E_n) < \infty$ and $A \subseteq \bigcup_{n \geq 1} E_n$. If μ is a finite measure (i.e., $\mu(\Omega) < \infty$) or Ω itself is σ-finite, then all subsets of Ω is σ-finite.

Next, assume f, g are bounded measurable functions with finite measure support. Then if $\varphi \le f$ and $\phi \le g$ are simple functions, then $\varphi + \phi$ is a simple function $\le f + g$, so

$$\int f + g \, d\mu \ge \int \varphi + \phi \, d\mu = \int \varphi \, d\mu + \int \phi \, d\mu.$$

Then taking $\sup_{\varphi \le f}$ and $\sup_{\phi \le g}$ on the right hand side, we get

$$\int f + g \, d\mu \ge \int f \, d\mu + \int g \, d\mu.$$

Conversely, choose simple functions $\varphi \ge f$ and $\phi \ge g$. Then $\varphi + \phi \ge f + g$ is a simple function. By Proposition 1.3.6,

$$\int f + g \, d\mu \le \int \varphi + \phi \, d\mu = \int \varphi \, d\mu + \int \phi \, d\mu.$$

Then taking $\inf_{\varphi \ge f}$ and $\inf \phi \ge g$ on the right hand side and again using Proposition 1.3.6, we get

$$\int f + g \, d\mu \le \int f \, d\mu + \int g \, d\mu.$$

Lastly, suppose $f, g \ge 0$. Fix bounded functions $0 \le h \le f$ and $0 \le r \le g$ with $\mu(h \ne 0) < \infty$ and $\mu(r \ne 0) < \infty$. Then using **(iii)** for bounded functions, we have $\int h \, d\mu + \int r \, d\mu = \int h + r \, d\mu$. Taking supremum and noting that $h + r$ is a bounded function such that $0 \le h + r \le f + h$ and $\mu(h + r \ne 0) = \mu(h \ne 0 \text{ or } r \ne 0) \le \mu(h \ne 0) + \mu(r \ne 0) < \infty$, we get

$$\int f \, d\mu + \int g \, d\mu = \sup_{0 \le h \le f} \int h \, d\mu + \sup_{0 \le r \le g} \int h \, d\mu = \sup_{0 \le h \le f, 0 \le r \le g} \int h + r \, d\mu \le \int f + g \, d\mu.$$

To show the other direction, first note that since $f, g \ge 0$ and the supports of f and g are σ-finite, the support of $f + g$ is also σ-finite. Hence there exists events $E_n \in \mathscr{F}$, $n \ge 1$ such that $E_n \nearrow \{f + g \ne 0\}$ and $\mu(E_n) < \infty$. Then for each $n \ge 1$, we have

$$[(f + g) \wedge n] \mathbf{1}_{E_n} \le (f \wedge n) \mathbf{1}_{E_n} + (g \wedge n) \mathbf{1}_{E_n}.$$

Then by using **(iii)**-**(iv)** for bounded functions, we get

$$\int_{E_n} (f + g) \wedge n \, d\mu \le \int_{E_n} \left[(f \wedge n) + (g \wedge n) \right] d\mu = \int_{E_n} (f \wedge n) \, d\mu + \int_{E_n} (g \wedge n) \, d\mu.$$

Then by using Lemma 1.3.8, as $n \to \infty$, the first and the last expression above converges to $\int f + g \, d\mu$ and $\int f \, d\mu + \int g \, d\mu$, respectively. This shows the desired inequality.

(iv) It follows immediately from applying **(i)** to $h := f - g \ge 0$ and using **(iii)**.

(v) Note that $f \le |f|$ so $\int f \, d\mu \le \int |f| \, d\mu$ by **(iv)**. Taking absolute value then gives the assertion.

□

The following lemma shows that the integral of a possibly unbounded function f can be achieved by the integrals of its truncations $f \wedge n$ for $n \ge 1$. In other words, the supremum in the definition of $\int f \, d\mu$ in Definition 1.3.3) **(III)** could be taken only over the bounded functions $g := (f \wedge n) \mathbf{1}_{E_n}$.

Lemma 1.3.8 (Trunctation of integrals)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and fix a nonnegative measurable function $f : \Omega \to \mathbb{R}$. Suppose there exists events $E_n \in \mathscr{F}$, $n \ge 1$ such that $\mu(E_n) < \infty$ and $E_n \nearrow \{f > 0\}$ (i.e., $E_1 \subseteq E_2 \subseteq \dots$ and $\bigcup_{n \ge 1} E_n = \{f > 0\}$). Then*

$$\lim_{n \to \infty} \int_{E_n} f \wedge n \, d\mu = \int f \, d\mu. \tag{8}$$

PROOF. Note that $f \wedge n$ is the function that maps ω to $f(\omega) \wedge n$. Hence $(f \wedge n)\mathbf{1}_{E_n}$ is bounded above by n, non-incerasing in n, and has support contained in E_n, which has finite μ-measure by the hypothesis. Hence by Theorem 1.3.7 **(iv)** for the case of bounded functions, we see that the integrals in the LHS of (8) is non-decreasing in n. Moreover, since $g := (f \wedge n)\mathbf{1}_{E_n}$ is itself a bounded measurable function such that $0 \le g \le f$ and $\mu(g > 0) \le \mu(E_n) < \infty$, by definition of $\int f \, d\mu$, we have

$$\lim_{n\to\infty} \int_{E_n} f \wedge n \, d\mu \le \int f \, d\mu.$$

To show that the limit equals the integral in the RHS, we use the definition of the integral in the RHS. Namely, fix a bounded measurable function h such that $0 \le h \le f$ and $\mu(h > 0) < \infty$. Suppose $\sup_\omega h(\omega) < M$ for some constnat $M > 0$. Then for $n > 0$, $h \le f \wedge n \le f$ Hence again using Theorem 1.3.7 **(iv)** for the case of bounded functions,

$$\begin{aligned}
\int_{E_n} f \wedge n \, d\mu \ge \int_{E_n} h \, d\mu &= \int h \, d\mu - \int_{E_n^c} h \, d\mu \\
&= \int h \, d\mu - \int_{\{h>0\}\setminus E_n} h \, d\mu \\
&\ge \int h \, d\mu - M\mu(\{h > 0\} \setminus E_n).
\end{aligned}$$

Since $E_n \nearrow \{f > 0\}$, $\{h > 0\} \subseteq \{f > 0\}$, and $\mu(h > 0) < \infty$, we have $\mu(\{h > 0\} \setminus E_n) \to 0$ as $n \to \infty$. So

$$\int h \, d\mu \le \lim_{n\to\infty} \int_{E_n} f \wedge n \, d\mu.$$

Taking supremum over all bounded measurable functions h such that $0 \le h \le f$ and $\mu(h > 0) < \infty$, by using the definition of $\int f \, d\mu$, we obtain

$$\lim_{n\to\infty} \int_{E_n} f \wedge n \, d\mu \ge \int f \, d\mu.$$

This shows the assertion. □

Lemma 1.3.9. *Suppose $\int f \, d\mu$ exists. Let $f = f_1 - f_2$ where $f_1, f_2 \ge 0$ and $\int f_i \, d\mu < \infty$ for some $i \in \{1, 2\}$. Assume all functions are assumped to be integrable and have σ-finite support. Then*

$$\int f \, d\mu = \int f_1 \, d\mu - \int f_2 \, d\mu.$$

PROOF. Since $f = f^+ - f^-$, we have $f^+ + f_2 = f^- + f_1$. Then since all four involved functions are nonnegative with σ-finite support, by Theorem 1.3.7 for such functions, we have

$$\int f^+ \, d\mu + \int f_2 \, d\mu = \int f^- \, d\mu + \int f_1 \, d\mu. \tag{9}$$

Since $\int f \, d\mu$ exists, one of f^+ and f^- has finite integral. First suppose $\int f^+ \, d\mu = \infty$. Then $\int f^- \, d\mu < \infty$, and the above identity and the hypothesis implies $\int f_1 \, d\mu = \infty$ and $\int f_2 \, d\mu < \infty$. Hence we can subtract the finite numbers $\int f^- \, d\mu < \infty$ and $\int f_2 \, d\mu < \infty$ from both sides to get

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu = \int f_1 \, d\mu - \int f_2 \, d\mu = \infty.$$

A similar argument works by symmetry if $\int f^- \, d\mu = \infty$. Lastly, if both $f^\pm$ have finite integral, then (9) yields that f_1 and f_2 both have finite integrals. Again subtract the finite numbers $\int f^- \, d\mu < \infty$ and $\int f_2 \, d\mu < \infty$ from both sides of (9) and using Definition 1.3.3 **(IV)**, the assertion follows. □

Now we finish the proof of Theorem 1.3.7 for the general case.

PROOF OF THEOREM 1.3.7 FOR INTEGRABLE FUNCTIONS. All functions here are assumed to have σ-finite support.

(i) We have already shown this for nonnegative measurable functions.

(ii) Suppose $a > 0$. Note that $af^+ = (af)^+$ and $af^- = (af)^-$. Then by using Definition 1.3.3 **(IV)** and Theorem 1.3.7 **(ii)** for nonnegative functions and nonnegative scalar,

$$\begin{aligned}\int af\,d\mu &= \int (af)^+\,d\mu - \int (af)^-\,d\mu = \int af^+\,d\mu - \int af^-\,d\mu \\ &= a\int f^+\,d\mu - a\int f^-\,d\mu = a\left(\int f^+\,d\mu - \int f^-\,d\mu\right) = a\int f\,d\mu.\end{aligned}$$

Next, suppose $a < 0$. Then $(af)^+ = -af^-$ and $a(af)^- = -af^+$. Then by using Definition 1.3.3 **(IV)** and Theorem 1.3.7 **(ii)** for nonnegative functions and nonnegative scalar,

$$\begin{aligned}\int af\,d\mu &= \int (af)^+\,d\mu - \int (af)^-\,d\mu = \int (-a)f^-\,d\mu - \int (-a)f^+\,d\mu \\ &= (-a)\int f^-\,d\mu + a\int f^+\,d\mu = a\left(\int f^+\,d\mu - \int f^-\,d\mu\right) = a\int f\,d\mu.\end{aligned}$$

(iii) Write $f+g = (f^+ + g^+) - (f^- + g^-)$. By the hypothesis, either $f^+ + g^+$ or $f^- + g^-$ have finite integral, since otherwise either f or g is not integrable or $\int f\,d\mu$ and $\int g\,\mu$ are both infinity with the opposite sign. Then the assertion follows from Lemma 1.3.9.

(iv) It follows immediately from applying **(i)** to $h := f - g \geq 0$ and using **(iii)**.

(v) Note that $f \leq |f|$ so $\int f\,d\mu \leq \int |f|\,d\mu$ by **(iv)**. Taking absolute value then gives the assertion.

□

1.3.2. Inequalities and limit theorems for integral.

Proposition 1.3.10 (Jensen's inequality)**.** *Suppose a function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, that is, $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$. If μ is a probability measure and if f and $\varphi(f)$ are integrable and if $\int f\,d\mu < \infty$, then*

$$\varphi\left(\int f\,d\mu\right) \leq \int \varphi(f)\,d\mu.$$

PROOF. Denote $c := \int f\,d\mu < \infty$. Definition of convexity implies existence of a tangent line at $(c, \varphi(c))$. That is, there exists $a, b \in \mathbb{R}$ such that the line $\ell(x) = ax + b$ lower bounds $\varphi(x)$ for all $x \in \mathbb{R}$ and $\ell(c) = \varphi(c)$ (Why?). Since $\ell \leq \varphi$, Theorem 1.3.7 implies

$$\int \varphi(f)\,d\mu \geq \int af + b\,d\mu = a\int f\,d\mu + b = ac + b = \varphi(c) = \varphi\left(\int f\,d\mu\right),$$

where the first equality uses the fact that μ is a probability measure so $\int b\,d\mu = b\int d\mu = b$. □

Remark 1.3.11. Convexity of φ means

$$\varphi(\text{avg of two points}) \leq \text{avg. of } \varphi \text{ at two points.}$$

Integral against a probability measure can be throught of as taking an average in a general sense. (In fact, we will define this to be taking 'expectation'.) Hence Jensen's inequality generalizes the above to the general weighted average on a probability space where the 'weight' is given by the function f.

Definition 1.3.12. Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f : \Omega \to \mathbb{R}$ be a measurable function. Fix $p \in (0, \infty)$. We define the *p-norm* of f as [13]

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

Proposition 1.3.13 (Hölder's inequality)**.** *If $p, q \in (1,\infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\int |fg|\,d\mu \leq \|f\|_p \|g\|_q. \tag{10}$$

[13] The function $x \mapsto x^p$ is continuous, so it is Borel measurable. Hence if f is Borel, then f^p is also Borel.

PROOF. If $\|f\|_p = 0$ or $\|g\|_p = 0$, then $f = 0$ or $g = 0$ μ-a.e., so $|fg| = 0$ μ-a.e.. (see Exercise 1.3.5). Hence in this case there is nothing to show. We may assume $\|f\|_p, \|g\|_p > 0$. Then we may divide both sides of (10) by $\|f\|_p\|g\|_p > 0$, use Theorem 1.3.7 **(ii)**, and rename $f/\|f\|_p$ as f and $g/\|g\|_p$ as g. Hence it suffices to show the assertion for the case when $\|f\|_p = \|g\|_q = 1$.

The rest of the proof is based on the following inequality (Justify it by showing the function $\phi_y : x \mapsto \frac{x^p}{p} + \frac{y^q}{q} - xy$ is minimized at $x = y^{1/(p-1)}$ with minimum value zero.)

$$xy \le \frac{x^p}{p} + \frac{y^q}{q} \quad \forall x, y \in \mathbb{R}\,.$$

Using this inequality, replacing $x = f$ and $y = g$ and integrating,

$$\int |fg|\,d\mu = \frac{1}{p} + \frac{1}{q} = 1 = \|f\|_p\,\|g\|_q.$$

This shows the assertion. □

Next, we will study when we can interchange limit and integral. Namely, if we have a sequence of functions f_n such that $f_n \to f$ in some appropriate sense, we would like to know sufficient conditions under which

$$\lim_{n\to} \int f_n\,d\mu = \int \lim_{n\to} f_n\,d\mu = \int f\,d\mu.$$

Definition 1.3.14 (Convergence in measure and almost everywhere convergence)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f, f_n : \Omega \to \mathbb{R}$ for $n \ge 1$ be measurable functions. We say $f_n \to f$ as $n \to \infty$ *in measure* if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu\big(\{\omega : |f_n(\omega) - f(\omega)| > \varepsilon\}\big) = 0.$$

We say $f_n \to f$ as $n \to \infty$ μ*-a.e.* (almost everywhere) if

$$\mu\left(\left\{\omega : \left|\lim_{n\to\infty} f_n(\omega) - f(\omega)\right| \ne 0\right\}\right) = 0.$$

If μ is a probability measure, then we say convergence *in probability* instead of convergence in measure, and *almost sure convergence* instead of almost everywhere convergence.

Exercise 1.3.15 (Convergence μ-a.e. implies convergence in measure)**.** Suppose μ is a finite measure and assume $f_n \to f$ μ-a.e.. Show that $f_n \to f$ in measure.

Theorem 1.3.16 (Bounded Convergence Theorem)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f, f_n : \Omega \to \mathbb{R}$ for $n \ge 1$ be measurable functions such that $f_n \to f$ in measure. Assume that there exists an event $E \in \mathscr{F}$ such that $\mu(E) < \infty$ and $f_n \equiv 0$ on E^c. Further assume $|f_n| < M$ for all $n \ge 1$ for some constant $M > 0$. Then*

$$\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu.$$

(By Exercise 1.3.15, the statement also holds if $f_n \to f$ μ-a.e..)

PROOF. The idea is to decompose integrals on sets G_n where $|f_n - f|$ is small and on E_n where $|f_n - f|$ is large. The difference of the integrals is small either because $f_n \approx f$ or the underlying set has small measure. (The proof is much simpler if we also assume $f \equiv 0$ on E^c and $|f| < M$, but we argue for the general case below.)

We first claim that

$$\mu\big(\{f \ne 0\} \cap E^c\big) = 0 \quad \text{and} \quad \mu(\{|f| > M\}) = 0. \tag{11}$$

Indeed, since $f_n \equiv 0$ on E^c and $f_n \to f$ in measure, for each fixed $\varepsilon > 0$,

$$\mu\big(\{|f| > \varepsilon\} \cap E^c\big) = \mu\big(\{|f_n - f| > \varepsilon\} \cap E^c\big) \to 0 \quad \text{as } n \to \infty.$$

Since the left hand side does not depend on n, we have $\mu\big(\{|f|>\varepsilon\}\cap E^c\big)=0$. Since this holds for arbitrary $\varepsilon>0$, using that $\{f\neq 0\}=\bigcup_{n=1}^{\infty}\{|f|>n^{-1}\}$, we get (by using union bound, or subadditivity of measure, see Thm 1.1.16)

$$\mu\big(\{f\neq 0\}\cap E^c\big)=\mu\left(\bigcup_{n=1}^{\infty}\{|f|>n^{-1}\}\cap E^c\right)\leq\sum_{n=1}^{\infty}\mu\big(\{|f|>n^{-1}\}\cap E^c\big)=0,$$

as desired. Similarly, for each $\varepsilon>0$,

$$\mu\big(|f|\geq M+\varepsilon\big)\leq\mu\big(|f-f_n|\geq\varepsilon\big)\to 0\quad\text{as } n\to\infty.$$

Hence the left hand side equals 0. This holds for all $\varepsilon>0$, so we deduce $\mu(|f|>M)=0$.

Now fix $\varepsilon>0$. Let $A_n:=\{|f_n-f|>\varepsilon\}$, $B_n:=\{|f_n-f|\leq\varepsilon\}$, and $C:=\{|f|\leq M\}$. Then we have

$$\begin{aligned}\left|\int f_n\,d\mu-\int f\,d\mu\right|&=\left|\int f_n-f\,d\mu\right|\quad(\because\text{Thm 1.3.7 \textbf{(iii)}})\\&\leq\int|f_n-f|\,d\mu\quad(\because\text{Thm 1.3.7 \textbf{(v)}})\\&=\int_{A_n}|f_n-f|\,d\mu+\int_{B_n}|f_n-f|\,d\mu\quad(\because\text{Thm 1.3.7 \textbf{(v)}}).\end{aligned}$$

We will show each of the integrals above vanishes as $n\to\infty$.

Note that $\mu(C^c)=0$ by (11) so $\int_{A_n\setminus C}|f_n-f|\,d\mu=0$ (see Exercise 1.3.5). Hence

$$\begin{aligned}\int_{A_n}|f_n-f|\,d\mu&=\int_{A_n\cap C}|f_n-f|\,d\mu+\int_{A_n\setminus C}|f_n-f|\,d\mu\quad(\because\text{Thm 1.3.7 \textbf{(v)}})\\&=\int_{A_n\cap C}|f_n-f|\,d\mu\quad(\because\mu(C^c)=0)\\&\leq 2M\mu(A_n)\to 0\quad\text{as } n\to\infty.\end{aligned}$$

Also, note that $\int_{B_n\setminus E}|f_n-f|\,d\mu=\int_{B_n\setminus E}|f|\,d\mu=\int|f|\mathbf{1}_{E^c}\mathbf{1}_B\,d\mu=0$ since $f\mathbf{1}_{E^c}=0$ μ-a.e. due to (11). Thus

$$\begin{aligned}\int_{B_n}|f_n-f|\,d\mu&=\int_{B_n\cap E}|f_n-f|\,d\mu+\int_{B_n\setminus E}|f_n-f|\,d\mu\quad(\because\text{Thm 1.3.7 \textbf{(v)}})\\&=\int_{B_n\cap E}|f_n-f|\,d\mu\quad(\because f\mathbf{1}_{E^c}=0\ \mu\text{-a.e.})\\&\leq\varepsilon\mu(E).\end{aligned}$$

In summary, we have shown that

$$\left|\int f_n\,d\mu-\int f\,d\mu\right|\leq\overbrace{2M\mu(A_n)}^{o(1)}+\varepsilon\mu(E).$$

Since $\varepsilon>0$ is arbitrary and $\mu(E)<\infty$, the left hand side above must equal to zero. □

Example 1.3.17 (BCT does not hold without finite support)**.** Consider $(\mathbb{R},\mathscr{B},\mu)$, where μ =Lebesgue measure. Let $f_n:=n^{-1}\mathbf{1}_{[0,n]}$. Then $f_n\leq 1$ for $n\geq 1$ and $\int f_n\,d\mu=1$ for all $n\geq 1$. However, $f_n\to 0$ in measure:

$$\mu(|f_n-0|>\varepsilon)=\mu(|f_n|>\varepsilon)=\mu(\emptyset)=0\quad\text{if } n^{-1}<\varepsilon.$$

Hence BCT does not hold in this case. In fact, f_n violates the uniformly finite support assumption in BCT, since $\{f_n\neq 0\}=[0,n]$ and $\mu([0,n])\to\infty$. ▲.

Theorem 1.3.18 (Fatou's lemma)**.** *Let $(\Omega,\mathscr{F},\mu)$ be a measure space and let $f_n:\Omega\to\mathbb{R}$ for $n\geq 1$ be nonnegative measurable functions. Suppose that for each $n\geq 1$, there exists events $E_m\nearrow\{f_n>0\}$ as $m\to\infty$ and $\mu(E_m)<\infty$. Then*

$$\liminf_{n\to\infty}\int f_n\,d\mu\geq\int\liminf_{n\to\infty}f_n\,d\mu.$$

PROOF. (The proof is simpler if there exists $E_n \nearrow \Omega$ with $\mu(E_n) < \infty$.) Denote $g_n := \inf_{m \geq n} f_m$. Then $f_n \geq g_n \nearrow g := \liminf f_n$ as $n \to \infty$. Hence $\liminf_{n\to\infty} f_n \geq g$, so it suffices to show that

$$\liminf_{n\to\infty} \int g_n \, d\mu \geq \int g \, d\mu.$$

To this end, we claim that there exists finite μ-measure events $C_m \nearrow \{g > 0\}$. For this, we can use a 'diagonal argument'. For each $n \geq 1$, let $E_{n,m} \nearrow \{f_n > 0\}$ as $m \to \infty$ and $\mu(E_{n,m}) < \infty$ for $m \geq 1$. Arrange these sets as below:

$$\begin{array}{lllll} f_1: & E_{1,1} & E_{1,2} & E_{1,3} & \dots \\ f_2: & E_{2,1} & E_{2,2} & E_{2,3} & \dots \\ f_3: & E_{3,1} & E_{3,2} & E_{3,3} & \dots \\ \vdots & & & & \end{array}$$

We can the enumerate these sets in diagonal fashion, namely, $E_{1,1}, E_{1,2}, E_{2,1}, E_{1,3}, E_{2,2}, E_{3,1}, \dots$. Let B_m denote the union of the first m sets in this enumeration. Then $\mu(B_m) < \infty$ for $m \geq 1$ and $\bigcup_{n\geq 1}\{f_n > 0\} \subseteq \bigcup_{m\geq 1} B_m$. In particular, this implies $\{g > 0\} \subseteq \bigcup_{m\geq 1} B_m$. Then setting $C_m := B_m \cap \{g > 0\}$ proves the claim.

Now, by using the truncation argument (Lemma 1.3.8), we have

$$\int g \, d\mu = \lim_{m\to\infty} \int_{C_m} g \wedge m \, d\mu.$$

On the other hand, fix $m \geq 1$ and observe that $h_n := (g_n \wedge m)\mathbf{1}_{C_m}$ for $n \geq 1$ is a sequence of bounded measurable functions with supports all contained in C_m that has finite μ-measure. Moreover, $h_n \to h := (g \wedge m)\mathbf{1}_{C_m}$ μ-a.e. as $n \to \infty$. Hence by BCT (Theorem 1.3.16),

$$\int_{C_m} g \wedge m \, d\mu = \lim_{n\to\infty} \int_{C_m} g_n \wedge m \, d\mu.$$

Then we have

$$\begin{aligned} \liminf_{n\to\infty} \int g_n \, d\mu &\geq \liminf_{n\to\infty} \int g_n \wedge m \, d\mu \\ &\geq \liminf_{n\to\infty} \int_{C_m} g_n \wedge m \, d\mu \quad (\because g_n \geq 0) \\ &= \int_{C_m} g \wedge m \, d\mu \overset{m\to\infty}{\longrightarrow} \int g \, d\mu. \end{aligned}$$

This shows the assertion. □

Theorem 1.3.19 (Monotone Convergence Theorem)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $f_n : \Omega \to \mathbb{R}$ for $n \geq 1$ be measurable functions. Suppose $f_1 \leq f_2 \leq \dots$. Then*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int \lim_{n\to\infty} f_n \, d\mu.$$

PROOF. The existence of limits $\lim_{n\to\infty} \int f_n \, d\mu$ and $\lim_{n\to\infty} f_n$ follows from the monotonicity assumption. Note that "$\geq$" holds by Fatou's lemma (Theorem 1.3.18). On the other hand, by the monotonicity, $f_m \leq \lim_{n\to\infty} f_n$, so $\int f_m \, d\mu \leq \int \lim_{n\to\infty} f_n \, d\mu$. Taking $\lim_{m\to\infty}$ then shows "$\leq$". □

Theorem 1.3.20 (Dominated Convergence Theorem)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and let $g, f, f_n : \Omega \to \mathbb{R}$ for $n \geq 1$ be measurable functions. Suppose $f_n \to f$ μ-a.e., $|f_n| \leq g$ for all $n \geq 1$, and g is integrable. Then*

$$\lim_{n\to\infty} \int f_n \, d\mu = \int \lim_{n\to\infty} f_n \, d\mu.$$

PROOF. Note that $f_n + g \geq 0$ for $n \geq 1$. So by Fatou's lemma (Theorem 1.3.18),

$$\liminf_{n\to\infty} \int f_n \, d\mu + \int g \, d\mu = \liminf_{n\to\infty} \int f_n + g \, d\mu \geq \int f + g \, d\mu = \int f \, d\mu + \int g \, d\mu.$$

Since g is integrable and $g \geq 0$, we have $\int g \, d\mu < \infty$. Hence subtracting this quantity from both sides gives

$$\liminf_{n\to\infty} \int f_n \, d\mu \geq \int f \, d\mu.$$

Applying the above result for $-f_n$ gives

$$\limsup_{n\to\infty} \int f_n \, d\mu \leq \int f \, d\mu.$$

Combining the above inequalities gives the assertion. □

1.4. Product measures and Fubini's theorem

Let $(X, \mathscr{A}, \mu_1)$ and $(Y, \mathscr{B}, \mu_2)$ be two σ-finite measure spaces. We would like to define the 'product measure space' with the sample space Ω being the Cartesian product $X \times Y$:

$$(\Omega := X \times Y, \, \sigma\text{-alg} = ?, \, \text{measure} = ?).$$

Namely, we need to define a natural σ-algbegra on $X \times Y$ and a measure on that σ-algebra. A simple choice for the σ-algebra would be the Cartesian product of σ-algebras $\mathscr{A} \times \mathscr{B}$. The sets in $\mathscr{A} \times \mathscr{B}$ takes the form of $A \times B$ for $A \in \mathscr{A}$ and $B \in \mathscr{B}$, which is called a (measurable) *rectangle*. However, $\mathscr{A} \times \mathscr{B}$ is not necessarily a σ-algebra. Hence, we consider the smallest σ-algebra generated by all rectangles, namely, $\sigma(\mathscr{A} \times \mathscr{B})$. This will be our cannonical choice of σ-algebra on the product space $\Omega = X \times Y$. It then remains to define a measure on this σ-algebra. This is provided by the following result.

Theorem 1.4.1 (Construction of product measure)**.** *Let $(\Omega_1, \mathscr{A}, \mu_1)$ and $(\Omega_2, \mathscr{B}, \mu_2)$ be two σ-finite measure spaces. Consider the product measurable space $(\Omega_1 \times \Omega_2, \sigma(\mathscr{A} \times \mathscr{B}))$. Then there exists a unique measure μ on $\sigma(\mathscr{A} \times \mathscr{B})$ such that*

$$\mu(A \times B) = \mu_1(A)\, \mu_2(B) \qquad \forall \; A \in \mathscr{A}, \, B \in \mathscr{B}.$$

We call μ the product measure and denote $\mu := \mu_1 \otimes \mu_2$.

PROOF. Denote $\mathscr{S} := \mathscr{A} \times \mathscr{B}$, the set of all measurable rectangles. Define a premeasure $\alpha : \mathscr{S} \to [0, \infty]$ by $\alpha(A \times B) := \mu_1(A)\, \mu_2(B)$. Notice that $\mathscr{S}$ form a semi-ring: For $A, C \in \mathscr{A}$ and $B, D \in \mathscr{B}$,

$$\begin{aligned}
(A \times B) \cap (C \times D) &= (A \cap C) \times (B \cap D) \\
(A \times B) \setminus (C \times D) &= (A \times B) \cap (C \times D)^c \\
&= (A \times B) \cap \big((C^c \times D) \sqcup (C \times D^c) \sqcup (C^c \times D^c)\big) \\
&= ((A \cap C^c) \times (B \cap D)) \sqcup ((A \cap C) \times (B \cap D^c)) \sqcup ((A \cap C^c) \times (B \cap D^c)).
\end{aligned}$$

It remains to verify σ-additivity of α on $\mathscr{S}$, since then Carathéory's extension theorem (Thm. 1.1.36) will imply the desired result.

To this end, fix a rectangle $A \times B$ and suppose $A \times B = \bigsqcup_{i=1}^{\infty} A_i \times B_i$ for disjoint rectangles $(A_i \times B_i)$ for $i \geq 1$. We wish to show that

$$\mu_1(A)\, \mu_2(B) = \sum_{i=1}^{\infty} \mu_1(A_i)\, \mu_2(B_i). \tag{12}$$

To show the above, fix $(x,y) \in A \times B$. Since $A \times B = \bigsqcup_{i=1}^{\infty} A_i \times B_i$, we have $\mathbf{1}(x \in A)\mathbf{1}(y \in B) = \sum_{i=1}^{\infty} \mathbf{1}(x \in A_i)\mathbf{1}(y \in B_i)$. Integrating with respect to y and using MCT, we get

$$\begin{aligned}\mathbf{1}(x \in A)\mu_2(B) &= \int \sum_{i=1}^{\infty} \mathbf{1}(x \in A_i)\,\mathbf{1}(y \in B_i)\,d\mu_2(y) \\ &= \sum_{i=1}^{\infty} \int \mathbf{1}(x \in A_i)\,\mathbf{1}(y \in B_i)\,d\mu_2(y) \\ &= \sum_{i=1}^{\infty} \mathbf{1}(x \in A_i) \int \mathbf{1}(y \in B_i)\,d\mu_2(y) \\ &= \sum_{i=1}^{\infty} \mathbf{1}(x \in A_i)\mu_2(B).\end{aligned}$$

Then integrating with respect to x and using MCT again shows (12). □

Remark 1.4.2 (Lebesgue measure on $\mathbb{R}^d$)**.** By using Thoerem 1.4.1 and an induction on the number of products, we can construct product measures on arbitrary product space with finitely many products. Namely, if $(\Omega_i, \mathscr{F}, \mu_i)$, $i \geq 1$ are σ-finite measure spaces, then for any $n \geq 1$, we have the product measure space

$$\left(\prod_{i=1}^{n} \Omega_i,\ \sigma\left(\prod_{i=1}^{n} \mathscr{F}_i\right),\ \mu_1 \otimes \cdots \otimes \mu_n\right),$$

where $\mu_1 \otimes \cdots \otimes \mu_n$ is the unique measure on $\sigma\left(\prod_{i=1}^{n} \mathscr{F}_i\right)$ such that

$$(\mu_1 \otimes \cdots \otimes \mu_n)(A_1 \times \cdots \times A_n) = \prod_{i=1}^{n} \mu_i(A_i).$$

In particular, if $(\Omega_i, \mathscr{F}_i, \mu) = (\mathbb{R}, \mathscr{B}, \lambda)$, where $\mathscr{B}$ is the Borel σ-algebra and λ is the Lebesgue measure on $\mathbb{R}$, then $\lambda \otimes \cdots \otimes \lambda = \bigotimes_{i=1}^{n} \lambda$ is the *Lebesgue measure* on the Borel σ-algebra on $\mathbb{R}^n$.

Theorem 1.4.3 (Fubini's theorem)**.** *Let $(\Omega_1, \mathscr{F}_1, \mu_1)$ and $(\Omega_2, \mathscr{B}, \mu_2)$ be two σ-finite measure spaces. Let $(\Omega_1 \times \Omega_2, \sigma(\mathscr{A} \times \mathscr{F}_2), \mu_1 \otimes \mu_2)$ be the product measure space. If a Borel measurable function $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ satisfies $f \geq 0$ or $\int |f|\,d\mu < \infty$, then*

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f\,d\mu_1 \otimes \mu_2 &= \int_{\Omega_2} \int_{\Omega_1} f(x,y)\,d\mu_1(x)\,d\mu_2(y) \\ &= \int_{\Omega_1} \int_{\Omega_2} f(x,y)\,d\mu_2(y)\,d\mu_1(x).\end{aligned} \tag{13}$$

There are two technical points we need to address before we can prove Fubini's theorem. In order to make sense of the second expression in (13), we need to make sure that

(1) For each fixed $x \in \Omega_1$, the function $f_x : \Omega_2 \to \mathbb{R}$, $y \mapsto f(x,y)$ is measurable;

(2) The function $g : \Omega_1 \to \mathbb{R}$, $x \mapsto \int_{\Omega_2} f(x,y)\,d\mu_2(y)$ is measurable.

We will assume the above two points for the following proof.

PROOF. We will only prove the first equality in (13). The second equality follows from symmetry.

(Indicator functions) Let $f = \mathbf{1}_E$ for some $E \in \sigma(\mathscr{F}_1 \times \mathscr{F}_2)$. Further assume E is a rectangle $E_1 \times E_2$. Then $\mathbf{1}((x,y) \in E) = \mathbf{1}(x \in E_1)\mathbf{1}(y \in E_2)$, so we can verify the assertion as

$$\int_{\Omega_1 \times \Omega_2} f\,d\mu_1 \otimes \mu_2 = \mu_1 \otimes \mu_2(E) = \mu_1(E_1)\,\mu_2(E_2),$$

$$\int_{\Omega_2} \int_{\Omega_1} f(x,y)\,d\mu_1(x)\,d\mu_2(y) = \int_{\Omega_2} \mathbf{1}(x \in E_1)\mu_2(E_2)\,d\mu_1(x) = \mu_1(E_1)\,\mu_2(E_2).$$

What if E is not a rectangle? Then we proceed as follows. Let $\mathscr{E}$ denote the set of all measurable sets in $\sigma(\mathscr{F}_1 \times \mathscr{F}_2)$ for which the assertion holds. Then one can show that $\mathscr{E}$ is a λ-system. (i.e., $\Omega \in \mathscr{E}$; $A, B \in \mathscr{E}$ with $A \subseteq B \Rightarrow B \setminus A \in \mathscr{E}$; $\mathscr{E}$ is closed under ascending union). The measurable rectangles form a π-system (i.e., closed under intersection) and we just showed that they belong to $\mathscr{E}$. Hence by Dynkin's $\pi - \lambda$ theorem (see Theorem 1.1.37), $\mathscr{E}$ should contain the σ-algebra $\sigma(\mathscr{F}_1 \times \mathscr{F}_2)$ generated by the rectangles.

(Simple functions) Follows from linearity of integral and the assertion for characteristic functions.

(Nonnegative functions) Suppose $f \geq 0$. Note that $(\mu_1 \otimes \mu_2)(\Omega_1 \times \Omega_2) = \mu_1(\Omega_1)\mu_2(\Omega_2)$ by the construction of the product measure. By the hypothesis on σ-finiteness, it follows that $\mu_1 \otimes \mu_2$ is a σ-finite measure. Choose measurable sets $A_n \nearrow \Omega_1 \times \Omega_2$ such that $(\mu_1 \otimes \mu_2)(A_n) < \infty$. Define $f_n : \Omega_1 \times \Omega_2 \to \mathbb{R}$, $f_n(x,y) := [2^n f(x,y)\mathbf{1}((x,y) \in A_n)]/2^n \wedge n$, where $[\cdot]$ denotes the largest smaller integer than the input $\cdot$. Note that f_n is a simple function: It takes at most $n2^n$ distinct values on sets of finite measure as $(\mu_1 \otimes \mu_2)(f_n \neq 0) \leq (\mu_1 \otimes \mu_2)(A_n) < \infty$. Also, $f_n \nearrow f$ $(\mu_1 \otimes \mu_2)$-a.e.. Then the assertion follows from the previous case and MCT:

$$\begin{aligned}
\int_{\Omega_1 \times \Omega_2} f \, d\mu_1 \otimes \mu_2 &= \int_{\Omega_1 \times \Omega_2} \lim_{n\to\infty} f_n \, d\mu_1 \otimes \mu_2 \\
&= \lim_{n\to\infty} \int_{\Omega_1 \times \Omega_2} f_n \, d\mu_1 \otimes \mu_2 \\
&= \lim_{n\to\infty} \int_{\Omega_2} \int_{\Omega_1} f_n(x,y) \, d\mu_1(x) \, d\mu_2(y) \\
&= \int_{\Omega_2} \lim_{n\to\infty} \int_{\Omega_1} f_n(x,y) \, d\mu_1(x) \, d\mu_2(y) \\
&= \int_{\Omega_2} \int_{\Omega_1} \lim_{n\to\infty} f_n(x,y) \, d\mu_1(x) \, d\mu_2(y) = \int_{\Omega_2} \int_{\Omega_1} f(x,y) \, d\mu_1(x) \, d\mu_2(y),
\end{aligned}$$

where brining the limit inside the double integrals uses MCT for the increasing sequence of functions $y \mapsto \int_{\Omega_1} f_n(x,y) \, d\mu_1(x) \nearrow \int_{\Omega_1} f(x,y) \, d\mu_1(x)$ and $f_n \nearrow f$.

(Integrable functions) Write $f = f^+ - f^-$. The general case follows from applying the previous case to the nonnegative functions $f^+, f^-, |f|$ and using linearity of integral.

□

Example 1.4.4. Let $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ be a function. Then when do we have the following infinite double sum interchangable?

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} f(n,m) \stackrel{?}{=} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(n,m).$$

Applying Fubini's theorem for the counting measure, we know that the above equality holds if $f \geq 0$ or if f is absolutely summable: $\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |f(n,m)| < \infty$.

Exercise 1.4.5. Construct a counterexample to the Fubini's theorem, where the infinite double sums $\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} f(n,m)$ and $\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(n,m)$ does not equal to each other.

1.5. Expectation

1.5.1. Measure-theoretic definition of expectation. The following is the most general definition of the expectation of a RV in terms of a Lebesgue integral on the probability space.

Definition 1.5.1 (Expectaion)**.** Let X be a random variable from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ into the real line $(\mathbb{R}, \mathscr{B}, \mu)$, where $\mathscr{B}$ is the Borel σ-algebra on $\mathbb{R}$ and μ denotes the Lebesgue measure. The *expectation*

of X in this general setting is defined as the following Lebesgue integeral

$$\mathbb{E}[X] := \int_\Omega X d\mathbb{P}.$$

We define the random variable X is *integrable* if $\mathbb{E}[|X|] < \infty$.

An advantage of defining the expectation of a RV as a Lebesgue integral is that all properties and results we have established for Lebesgue integral (e.g., linearity, various inequalities, and limit theorems) readily apply to expectation. For instance, from Theorem 1.3.7, if we have two integrable random variables X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. then for any $a, b \in \mathbb{R}$, $aX + bY$ is integrable and

$$\text{(Linearity of expectation)} \qquad \mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{14}$$

To see the measure-theoretic definition of expectation of a RV in Definition 1.5.1 agrees with our undergraduate-level definition of expectation using PMFs and PDFs, we need a change of variables formula ('u-substitution'). But for the case of discrete RVs, a direct computation can show a direct connection between these two definitions of expectation. Consider the following computation: If we have a discrete RV $X = \sum_{i=1}^\infty x_i \mathbf{1}_{A_i}$, then

$$\begin{aligned}
\mathbb{E}[X] &= \int_\Omega \sum_{i=1}^\infty x_i \mathbf{1}(\omega \in A_i)\, d\mathbb{P}(\omega) \\
&\overset{?}{=} \sum_{i=1}^\infty \int_\Omega x_i \mathbf{1}(\omega \in A_i)\, d\mathbb{P}(\omega) \\
&= \sum_{i=1}^\infty x_i \int_\Omega \mathbf{1}(\omega \in A_i)\, d\mathbb{P}(\omega) \\
&= \sum_{i=1}^\infty x_i \mathbb{P}(A_i) \\
&= \sum_{i=1}^\infty x_i \mathbb{P}(X = x_i).
\end{aligned}$$

Note that the last expression is the usual formula for the expectation of the discrete RV X. The only caveat in this computation is the second equality, where we swapped the integral on Ω and the infinite sum (or integral against the counting measure). This is allowed if the sum over i was finite, by using linearity of integral in Theorem 1.3.7. In general, Fubini's theorem gives sufficient conditions under which we can interchange the order of two integrals, which we will study in the following section. Here, we give a direct justifiction of the above computation using MCT.

Proposition 1.5.2 (Expectation of discrete RV)**.** *Suppose X is a discrete random variable taking constant values x_i on disjoint measurable sets A_i for $i \geq 1$. If X is integrable, then*

$$\mathbb{E}[X] = \sum_{i=1}^\infty x_i\, \mathbb{P}(X = x_i).$$

PROOF. We first assume $X \geq 0$. Let $X_n := \sum_{i=1}^n x_i \mathbf{1}_{A_i}$. Then $X_n \nearrow X$. So we have

$$\begin{aligned}
\mathbb{E}[X] &= \lim_{n\to\infty} \mathbb{E}[X_n] \qquad (\because \text{MCT, Theorem 1.3.19}) \\
&= \lim_{n\to\infty} \sum_{i=1}^n x_i \mathbb{P}(X = x_i) \qquad (\because \text{Definition of } \mathbb{E}) \\
&= \sum_{i=1}^\infty x_i \mathbb{P}(X = x_i) \quad (\because \text{Def. of infinite sum}).
\end{aligned}$$

Hence the assertion holds for nonnegative RVs. For the general case, write $X = X^+ - X^-$. Since X is integrable, $\mathbb{E}[X^+], \mathbb{E}[X^-] < \infty$. Note that $X^+ = \sum_{i=1}^\infty (x_i \vee 0)\mathbf{1}_{A_i}$ and $X^- = \sum_{i=1}^\infty (-x_i \vee 0)\mathbf{1}_{A_i}$. By the assertion

for nonnegative RVs,

$$\begin{aligned}\mathbb{E}[X] = \mathbb{E}[X^+ - X^-] = \mathbb{E}[X^+] - \mathbb{E}[X^-] &= \sum_{i=1}^{\infty} (x_i \vee 0)\mathbb{P}(X = x_i) - \sum_{i=1}^{\infty} (-x_i \vee 0)\mathbb{P}(X = x_i) \\ &= \sum_{i=1}^{\infty} ((x_i \vee 0) - (-x_i \vee 0))\,\mathbb{P}(X = x_i) \\ &= \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i),\end{aligned}$$

where the second to the last equality uses linearity of integral against the counting measure. □

Theorem 1.5.3 (Change of variables)**.** *Let X be measurable map from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Omega', \mathscr{F}')$*[14]*. Let μ denote the measure on $\mathscr{F}'$ induced by X, that is, $\mu(A) := \mathbb{P}(X \in A)$ for $A \in \mathscr{F}'$*[15]*. If $f : \Omega' \to \mathbb{R}$ is Borel measurable, $f \geq 0$, and $\mathbb{E}[|f(X)|] < \infty$, then*

$$\mathbb{E}[f(X)] = \int_{\Omega'} f(u)\, d\mu(u).$$

PROOF. We verify the formula following the usual pipeline by establishing the assertion for f = characteristic function, simple function, nonnegative function, and then general integrable function.

(Indicator functions) Let $B \in \mathscr{F}'$ and $f := \mathbf{1}_B$. Then

$$\mathbb{E}[f(X)] = \mathbb{E}[\mathbf{1}_B(X)] = \mathbb{E}[\mathbf{1}(X \in B)] = \mathbb{P}(X \in B) = \mu(B) = \int_{\Omega'} f\, d\mu.$$

(Simple functions) Let $f := \sum_{i=1}^m a_i \mathbf{1}_{A_i}$ for disjoint events $A_1, \ldots, A_m \in \mathscr{F}'$. Then by linearity of expectation (14) and Lebesgue integral (see Theorem 1.3.7) and the assertion for indicator functions,

$$\begin{aligned}\mathbb{E}\big[f(X)\big] = \mathbb{E}\left[\sum_{i=1}^m a_i \mathbf{1}(X \in A_i)\right] = \sum_{i=1}^m a_i \mathbb{E}[\mathbf{1}(X \in A_i)] &= \sum_{i=1}^m a_i \int_{\Omega'} \mathbf{1}_{A_i}\, d\mu \\ &= \int_{\Omega'} \sum_{i=1}^m a_i \mathbf{1}_{A_i}\, d\mu = \int_{\Omega'} f\, d\mu.\end{aligned}$$

(Nonnegative functions) Let $f \geq 0$. Note that μ is a finite measure since $\mu(\Omega') = \mathbb{P}(X \in \Omega') = 1$. Define $f_n : \Omega' \to \mathbb{R}$, $f_n(x) := [2^n f(x)]/2^n \wedge n$, where $[\cdot]$ denotes the largest smaller integer than the input $\cdot$. Note that f_n is a simple function, since it takes at most $n2^n$ distinct values on sets of finite measure(note that $\mu(f_n \neq 0) < \infty$ since μ is a finite measure). Also, $f_n \nearrow f$ μ-a.e.. Furthermore, $f_n(X)$ is also a simple function and $f_n(X) \nearrow f(X)$ $\mathbb{P}$-a.e.. Hence by using the monotone convergence theorem for the two sequences $f_n(X)$ and f_n, with also using the assertion for simple functions,

$$\mathbb{E}[f(X)] = \lim_{n\to\infty} \mathbb{E}[f_n(X)] = \lim_{n\to\infty} \int_{\Omega'} f_n\, d\mu = \int_{\Omega'} \lim_{n\to\infty} f_n\, d\mu = \int_{\Omega'} f\, d\mu.$$

(Integrable functions) Write $f = f^+ - f^-$. The hypothesis $\mathbb{E}[|f(X)|] < \infty$ ensures that $\mathbb{E}[f^+(X)], \mathbb{E}[f^-(X)] < \infty$. Then using the assertion for nonnegative functions and by linearity of expectation and Lebesgue integral,

$$\mathbb{E}[f(X)] = \mathbb{E}[f^+(X)] - \mathbb{E}[f^-(X)] = \int_{\Omega'} f^+\, d\mu - \int_{\Omega'} f^-\, d\mu = \int_{\Omega'} f^+ - f^-\, d\mu = \int_{\Omega'} f\, d\mu.$$

□

[14] X is called a *random element* in Ω'. In the special case $\Omega' = \mathbb{R}$, X is called a random variable.

[15] Such measure μ is called the *pushforward* of $\mathbb{P}$ under X, which is denoted as $X_*(\mathbb{P})$. When $\Omega' = \mathbb{R}$ and $\mathscr{F} = \mathscr{B}$ Borel, the pushforword measure is called as the distribution of X (see Def. 1.2.11.)

Remark 1.5.4. The above "change of variables" formula is the measure theoretic version of "u-substitution". It allows us to transfer integral on one domain Ω to another Ω'. Usually we use it with the new domain $\Omega' = \mathbb{R}$.

To make such a connection, writing $h := X$ and $\mu := \mathbb{P} \circ h^{-1}$, Theorem 1.5.3 states that

$$\int f(h(\omega))\, d\mathbb{P}(\omega) = \int_{\Omega'} f(u)\, d\mathbb{P} \circ h^{-1}(u).$$

In the LHS integral, the variable ω lives in the original sample space Ω. In the RHS integral, the variable y lives in the new space Ω'. Here we are performing the 'u-substitution' where $u := h(\omega)$.

Exercise 1.5.5. Suppose X is a discrete random variable taking values x_i on a disjoint measurable sets A_i for $i \geq 1$. Suppose X is integrable. Then by Theroem 1.5.3 with $\Omega' = \mathbb{R}$ and $f(x) = x$,

$$\mathbb{E}[X] = \int_{\mathbb{R}} u\, d\,\mathbb{P} \circ X^{-1}(u).$$

From this, derive the formula $\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i\, \mathbb{P}(X = x_i)$. (Hint: Note that $\mathbb{P} \circ X^{-1}$ is a probability measure on $\mathbb{R}$ that puts mass $\mathbb{P}(A_i)$ on $\{x_i\}$ for $i \geq 1$.)

Exercise 1.5.6 (Stieltjes integral and Lebesgue integral)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space with μ being σ-finite. Let $\nu(A) = \int \rho \mathbf{1}_A\, d\mu$ for some Borel measurable function $\rho : \Omega \to \mathbb{R}$ (that is, $\nu \ll \mu$ with Radon-Nikodym derivative $\frac{d\nu}{d\mu} = \rho$) such that $\int \rho\, d\mu < \infty$. Following the standard procesure, show that for an integrable function $f : \Omega \to \mathbb{R}$,

$$\int f\, d\nu = \int f \rho\, d\mu.$$

Now we can connect the measure theoretic definition of expectation with the usual definition of expectation for continuous RVs.

Proposition 1.5.7 (Expectation of a continuous RV)**.** *Suppose X is a contiuous random variable with PDF f, that is, $\mathbb{P}(X \in B) = \int_B f\, d\mu$ for each Borel set B. Then*

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

PROOF. Let $\nu = \mathbb{P} \circ X^{-1}$ denote the probability measure on $\mathbb{R}$ defined by $\nu(B) = \mathbb{P}(X \in B)$. By the hypothesis, $\mathbb{P}(X \in B) = \int_B f\, d\mu$, so $\nu(B) = \int_B f\, d\mu$, where μ denotes the Lebesgue measure on $\mathbb{R}$. Then by Theroem 1.5.3 with $\Omega' = \mathbb{R}$ and $f(x) = x$ and Exercise 1.5.6,

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

□

Exercise 1.5.8 (Tail sum formula for expectation of discrete RVs)**.** Let X be a RV taking values on positive integers.

(i) For any x, show that

$$\mathbb{P}(X \geq x) = \sum_{y=x}^{\infty} \mathbb{P}(X = y).$$

(ii) Use (i) and Fubini's theorem (see Theorem 1.4.3) to show

$$\sum_{x=1}^{\infty} \mathbb{P}(X \geq x) = \sum_{y=1}^{\infty} \sum_{x=1}^{y} \mathbb{P}(X = y)$$

(iii) From (ii), conclude that

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

Exercise 1.5.9 (Tail sum formula for expectation of continuous RVs)**.** Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. Use Fubini's theorem to show that

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

The following statement generalizes the tail-sum formulas in Exercises 1.5.8 and 1.5.9.

Proposition 1.5.10 (Tail sum formula for expectation for nonnegative RVs)**.** *Let $X \geq 0$ be a RV on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then for each $p \geq 1$,*

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

PROOF. The statement for general $p \geq 1$ follows from that for $p = 1$ and a change of variable. Namely, since $|X|^p$ itself is a nonnegative RV,

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

by the change of variable $x^{1/p} = t$. Below we will show the statement for $p = 1$.

Let λ denote the Lebesgue measure on $\mathbb{R}$. Consider the joint measure space $(\Omega \times \mathbb{R}, \sigma(\mathscr{F} \times \mathscr{B}), \mathbb{P} \times \lambda)$, where $\mathscr{B}$ denotes the Borel σ-algebra on $\mathbb{R}$. Denote $A := \{(\omega, x) \,|\, 0 \leq x < X(\omega)\}$. We claim that $A \in \sigma(\mathscr{F} \times \mathscr{B})$. Indeed, define a function $G : \Omega \times \mathbb{R} \to \mathbb{R}$ by $G(\omega, x) := X(\omega) - x$. Note that the functions $(\omega, x) \mapsto X(\omega)$ and $(\omega, x) \mapsto x$ are measurable by considering their inverse images of the sets $(-\infty, a]$, and G is the difference of these functions, so it is measurable. Hence

$$G^{-1}((-\infty, 0)) = \{(\omega, x) : G(\omega, x) < 0\} = \{(\omega, x) : x < X(\omega)\} \in \sigma(\mathscr{F} \times \mathscr{B}).$$

It follows that

$$A = G^{-1}((-\infty, 0)) \cap (\Omega \times [0, \infty)) \in \sigma(\mathscr{F} \times \mathscr{B}).$$

Thus $\mathbf{1}_A : \Omega \times \mathbb{R} \to \mathbb{R}$ is measurable. Then $\mathbf{1}_A$ is also measurable and is nonnegative, so by Fubini's theorem,

$$\int_\Omega \int_\mathbb{R} \mathbf{1}_A\, d\lambda(x)\, d\mathbb{P}(\omega) = \int_\mathbb{R} \int_\Omega p x^{p-1} \mathbf{1}_A\, d\mathbb{P}(\omega)\, d\lambda(x).$$

Note that the LHS equals

$$\begin{aligned}
\int_\Omega \int_\mathbb{R} \mathbf{1}_A\, d\lambda(x)\, d\mathbb{P}(\omega) &= \int_\Omega \int_\mathbb{R} \mathbf{1}(0 \leq x < X(\omega))\, d\lambda(x)\, d\mathbb{P}(\omega) \\
&= \int_\Omega \int_0^{X(\omega)} d\lambda(x)\, d\mathbb{P}(\omega) \\
&= \int_\Omega X(\omega)\, d\mathbb{P}(\omega).
\end{aligned}$$

The RHS equals

$$\begin{aligned}
\int_\mathbb{R} \int_\Omega \mathbf{1}_A\, d\mathbb{P}(\omega)\, d\lambda(x) &= \int_\mathbb{R} \int_\Omega \mathbf{1}(0 \leq x < X(\omega))\, d\mathbb{P}(\omega)\, d\lambda(x) \\
&= \int_0^\infty \int_\Omega \mathbf{1}(x < X(\omega))\, d\mathbb{P}(\omega)\, d\lambda(x) \\
&= \int_0^\infty \mathbb{P}(X > x)\, d\lambda(x).
\end{aligned}$$

This shows the desired assertion. □

Next, we will see a nice application of linearity of expectation.

Exercise 1.5.11 (Inclusion-exclusion)**.** Let $(\Omega, \mathbb{P})$ be a probability space. Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. We will show the following inclusion-exclusion principle:

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i_1=1}^{k} \mathbb{P}(A_{i_1}) - \sum_{1\le i_1<i_2\le k} \mathbb{P}(A_{i_1}\cap A_{i_2}) + \sum_{1\le i_1<i_2<i_3\le k} \mathbb{P}(A_{i_1}\cap A_{i_2}\cap A_{i_3}) - \cdots + (-1)^k \mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right).$$

The method is so-called the 'indicator trick'.

For each $1 \le i \le k$, let $X_i = \mathbf{1}(A_i)$ be the indicator variable for the event A_i. Consider the following RV

$$Y = (1-X_1)(1-X_2)\cdots(1-X_k).$$

(i) By expanding the product and using linearity of expectation, show that

$$\mathbb{E}[Y] = 1 - \sum_{i_1=1}^{k} \mathbb{E}[X_{i_1}] + \sum_{1\le i_1<i_2\le k} \mathbb{E}[X_{i_1}X_{i_2}] - \sum_{1\le i_1<i_2<i_3\le k} \mathbb{E}[X_{i_1}X_{i_2}X_{i_3}] + \cdots - (-1)^k \mathbb{E}[X_1X_2\cdots X_k].$$

(ii) Show that Y is the indicator variable of the event $\bigcap_{i=1}^{k} A_i^c$. Conclude that

$$\mathbb{E}[Y] = \mathbb{P}\left(\bigcap_{i=1}^{k} A_i^c\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right).$$

(iii) From (i) and (ii), deduce the inclusion-exclusion principle.

Example 1.5.12 (Doubling strategy)**.** Suppose we bet $\$x$ on a game where we have to predict whether a fair coin clip comes up heads. We win $\$x$ on heads and lose $\$x$ on tails. Suppose we are playing the 'doubling strategy'. Namely, we start betting \$1, and until the first time we win, we double our bet; upon the first win, we quit the game. Let X be the random variable giving the net gain of the overall game. How can we evaluate this strategy?

Let N be the random number of coin flips we have to encounter until we see the first head. For instance,

$$\mathbb{P}(N=1) = \mathbb{P}(\{H\}) = 1/2,$$
$$\mathbb{P}(N=2) = \mathbb{P}(\{(T,H)\}) = 1/2^2$$
$$\mathbb{P}(N=3) = \mathbb{P}(\{(T,T,H)\}) = 1/2^3.$$

In general, we have

$$\mathbb{P}(N=k) = 1/2^k.$$

Note that on the event that $N=k$ (i.e., we bet k times), the net gain $X|\{N=k\}$ is

$$X|\{N=k\} = (-1) + (-2) + (-2^2) + (-2^3) + \cdots + (-2^{k-1}) + 2^k = -(2^k-1) + 2^k = 1.$$

Since the last expression do not depend on k, we conclude that

$$\mathbb{P}(X=1) = \sum_{k=1}^{\infty} \mathbb{P}(X=1 \mid N=k)\mathbb{P}(N=k) = \sum_{k=1}^{\infty} \mathbb{P}(N=k) = 1.$$

Hence our net gain is \$1 with probability 1. In particular, the expected net gain is also 1:

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) = 1.$$

What if we use tripling strategy? ▲

Exercise 1.5.13 (Jensen's inequality for multivaraite convex functions)**.** Suppose a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is convex, that is, $\varphi(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda\varphi(\mathbf{x}) + (1-\lambda)\varphi(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in [0,1]$. Let $X_1, \dots, X_d$ be RVs such that $\mathbb{E}[|X_i|] < \infty$ for $i = 1, \dots, d$ and $\mathbb{E}[|\varphi(X_1, \dots, X_d)|] < \infty$. Then show that

$$\mathbb{E}\left[\varphi(X_1, \dots, X_d)\right] \ge \varphi\left(\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]\right).$$

1.5.2. Variance and moments. Say you play two different games where in the first game, you win or lose \$1 depending on a fair coin flip, and in the second game, you win or lose \$10. In both games, your expected winning is 0. But the two games are different in how much the outcome fluctuates around the mean. This notion if fluctuation is captured by the following quantity called 'variance'.

Definition 1.5.14 (Covariance and variance)**.** Let X, Y be RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We define the *covariance* of X and Y as $\mathrm{Cov}(X,Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. The *variance* of X is defined by $\mathrm{Var}(X) := \mathrm{Cov}(X,X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

Exercise 1.5.15 (Covariance is symmetric and bilinear)**.** Let X, Y be RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and fix constants $a, b \in \mathbb{R}$. Show the following.

(i) $\mathrm{Cov}(aX + b, Y) = a\mathrm{Cov}(X,Y)$.
(ii) $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y)$.
(iii) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

Linearity of expectation implies

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

Here, $\mathbb{E}[X^2]$ is called the *second moment* of X. In general,

Definition 1.5.16 (Moments of RVs)**.** Let X be a RV. For each $k \ge 1$, the kth moment of X is defined as $\mathbb{E}[X^k]$.

Exercise 1.5.17 (Finite second moment implies finite first moment)**.** Let X be a RV with $\mathbb{E}[X^2] < \infty$.

(i) Show that

$$\mathbb{E}[|X|] = \mathbb{E}[|X|\mathbf{1}(|X| \le 1)] + \mathbb{E}[|X|\mathbf{1}(|X| > 1)] \le 1 + \mathbb{E}[X^2\mathbf{1}(|X| > 1)] \le 1 + \mathbb{E}[X^2] < \infty.$$

Hence X has finite first moment and is integrable.
(ii) Deduce that $\mathrm{Var}(X) < \infty$.

Example 1.5.18 (Higher moments)**.** Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. We will show that for any real number $\alpha > 0$,

$$\mathbb{E}[X^\alpha] = \int_0^\infty x^\alpha f_X(x)\, dx.$$

First, Use Exercise 1.5.9 and to write

$$\begin{aligned}\mathbb{E}[X^\alpha] &= \int_0^\infty \mathbb{P}(X^\alpha \ge x)\, dx \\ &= \int_0^\infty \mathbb{P}(X \ge x^{1/\alpha})\, dx \\ &= \int_0^\infty \int_{x^{1/\alpha}}^\infty f_X(t)\, dt\, dx.\end{aligned}$$

We then use Fubini's theorem to change the order of integral. This gives

$$\mathbb{E}[X] = \int_0^\infty \int_{x^{1/\alpha}}^\infty f_X(t)\,dt\,dx = \int_0^\infty \int_0^{t^\alpha} f_X(t)\,dx\,dt = \int_0^\infty t^\alpha f_X(t)\,dt,$$

as desired. ▲

Exercise 1.5.19. Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function. Use Fubini's theorem and tail sum formula for expectation to show

$$\mathbb{E}[g(X)] = \int_0^\infty g(x) f_X(x)\,dx.$$

Furthermore, assume X is a general continuous RV and g is a measurable function with $\mathbb{E}[|g(X)|] < \infty$. Show that

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

by using Theorem 1.5.3 and Exercise 1.5.6.

The following exercise ties the expectation and the variance of a RV into a problem of finding a point estimator that minimizes the mean squared error.

Exercise 1.5.20 (Variance as minimum MSE)**.** Let X be a RV. Let $\hat{x} \in \mathbb{R}$ be a number, which we consider as a 'guess' (or 'estimator' in statistics) of X. Let $\mathbb{E}[(X - \hat{x})^2]$ be the *mean squared error* (MSE) of this estimation.

(i) Show that

$$\mathbb{E}[(X - \hat{x})^2] = (\hat{x} - \mathbb{E}[X])^2 + \mathrm{Var}(X).$$

(ii) Conclude that the MSE is minimized when $\hat{x} = \mathbb{E}[X]$ and the global minimum is $\mathrm{Var}(X)$. In this sense, $\mathbb{E}[X]$ is the 'best guess' for X and $\mathrm{Var}(X)$ is the corresponding MSE.

Example 1.5.21 (Linear transform)**.** In this example, we argue that $\mathrm{Cov}(X, Y)$ measures the 'linear tendency' between X and Y. Let X be a RV, and define another RV Y by $Y = aX + b$ for some constants $a, b \in \mathbb{R}$. Let's compute their covariance using linearity of expectation.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, aX + b) \\
&= \mathbb{E}(aX^2 + bX) - \mathbb{E}(X)\mathbb{E}(aX + b) \\
&= a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(X)(a\mathbb{E}(X) + b) \\
&= a[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\
&= a\,\mathrm{Var}(X).
\end{aligned}$$

Thus, $\mathrm{Cov}(X, aX + b) > 0$ if $a > 0$ and $\mathrm{Cov}(X, aX + b) < 0$ if $a < 0$. In other words, if $\mathrm{Cov}(X, Y) > 0$, then X and Y tend to be large at the same time; if $\mathrm{Cov}(X, Y) > 0$, then Y tends to be small if X tends to be large. ▲

Next, let's say four RVs X, Y, Z, and W are given. Suppose that $\mathrm{Cov}(X, Y) > \mathrm{Cov}(Z, W) > 0$. Can we say that 'the positive linear relation' between X and Y is stronger than that between Z and W? Not quite.

Example 1.5.22. Suppose X is a RV. Let $Y = 2X$, $Z = 2X$, and $W = 4X$. Then

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, 2X) = 2\,\mathrm{Var}(X),$$

and

$$\mathrm{Cov}(Z, W) = \mathrm{Cov}(2X, 4X) = 8\,\mathrm{Var}(X).$$

But $Y = 2X$ and $W = 2Z$, so the linear relation between the two pairs should be same. ▲

So to compare the magnitude of covariance, we first need to properly normalize covariance so that the effect of fluctuation (variance) of each coordinate is not counted: then only the correlation between the two coordinates will contribute. This is captured by the following quantity.

Definition 1.5.23 (Correlation coefficient)**.** Given two RVs X and Y, their *correlation coefficient* $\rho(X,Y)$ is defined by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}\,Y}}.$$

Example 1.5.24. Suppose X is a RV and fix constants $a, b \in \mathbb{R}$. Then

$$\rho(X, aX+b) = \frac{a\mathrm{Cov}(X,X)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(aX+b)}} = \frac{a\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)}\sqrt{a^2\mathrm{Var}(X)}} = \frac{a}{|a|} = \mathrm{sign}(a).$$

▲

Exercise 1.5.25 (Cauchy-Schwarz inequality)**.** Let X, Y are RVs. Suppose $\mathbb{E}(Y^2) > 0$. We will show that the 'inner product' of X and Y is at most the product of their 'magnitudes'

(i) For any $t \in \mathbb{R}$, show that

$$\mathbb{E}[(X-tY)^2] = \mathbb{E}[Y^2]\left(t - \frac{\mathbb{E}[XY]}{\mathbb{E}(Y^2)}\right)^2 + \frac{\mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}.$$

Conclude that

$$0 \le \mathbb{E}\left[\left(X - \frac{\mathbb{E}(XY)}{\mathbb{E}[Y^2]}Y\right)^2\right] = \frac{\mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}.$$

(ii) Show that a RV Z satisfies $\mathbb{E}[Z^2] = 0$ if and only if $\mathbb{P}[Z=0] = 1$.
(iii) Show that

$$\mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]},$$

where the equality holds if and only if

$$\mathbb{P}\left(X = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y\right) = 1.$$

Exercise 1.5.26. Let X, Y are RVs such that $\mathrm{Var}(Y) > 0$. Let $\bar{X} = X - \mathbb{E}[X]$ and $\bar{Y} = Y - \mathbb{E}[Y]$.

(i) Use (1.5.25) to show that

$$0 \le \mathbb{E}\left[\left(\bar{X} - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\bar{Y}\right)^2\right] = \mathrm{Var}(X)\left(1 - \rho(X,Y)^2\right).$$

(ii) Show that $|\rho(X,Y)| \le 1$.
(iii) Show that $|\rho(X,Y)| = 1$ if and only if $\bar{X} = a\bar{Y}$ for some constant $a \neq 0$.

1.6. Examples

1.6.1. Discrete RVs.

Example 1.6.1 (Bernoulli RV)**.** A RV X is a *Bernoulli* variable with (success) probability $p \in [0,1]$ if it takes value 1 with probability p and 0 with probability $1-p$. In this case we write $X \sim \mathrm{Bernoulli}(p)$. Then $\mathbb{E}(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.

Example 1.6.2 (Indicator variables)**.** Let $(\Omega, \mathbb{P})$ be a probability space and let $E \subseteq \Omega$ be an event. The *indicator variable* of the event E, which is denoted by $\mathbf{1}_E$, is the RV such that $\mathbf{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbf{1}_E(\omega) = 0$ if $\omega \in E^c$. Then $\mathbf{1}_E$ is a Bernoulli variable with success probability $p = \mathbb{P}(E)$.

Example 1.6.3 (Binomial RV)**.** Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed Bernoulli p variables. Let $X = X_1 + \cdots + X_n$. One can think of flipping the same probability p coin n times. Then X is the total number of heads. Note that X has the following PMF

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for k nonnegative integer, and $\mathbb{P}(X = k) = 0$ otherwise. We say X follows the Binomial distribution with parameters n and p, and write $X \sim \text{Binomial}(n, p)$.

We can compute the mean and variance of X using the above PMF directly, but it is much easier to break it up into Bernoulli variables and use linearity. Recall that $X_i \sim \text{Bernoulli}(p)$ and we have $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$ for each $1 \le i \le n$ (from Example 1.6.1). So by linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$

On the other hand, since X_i's are independent, variance of X is the sum of variance of X_i's, so

$$\text{Var}(X) = \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = np(1-p).$$

Example 1.6.4 (Geometric RV)**.** Suppose we flip a probability p coin until it lands heads. Let X be the total number of trials until the first time we see heads. Then in order for $X = k$, the first $k-1$ flips must land on tails and the kth flip should land on heads. Since the flips are independent with each other,

$$\mathbb{P}(X = k) = \mathbb{P}(\{T, T, \cdots, T, H\}) = (1-p)^{k-1} p.$$

This is valid for k positive integer, and $\mathbb{P}(X = k) = 0$ otherwise. Such a RV is called a *Geometric RV* with (success) parameter p, and we write $X \sim \text{Geom}(p)$.

The mean and variance of X can be easily computed using its moment generating function, which we will learn soon in this course. For their direct computation, note that

$$\begin{aligned}
\mathbb{E}(X) - (1-p)\mathbb{E}(X) &= (1-p)^0 p + 2(1-p)^1 p + 3(1-p)^2 p + 4(1-p)^3 p \cdots \\
&\qquad - \left[(1-p)^1 p + 2(1-p)^2 p + 3(1-p)^3 p + \cdots\right] \\
&= (1-p)^0 p + (1-p)^1 p + (1-p)^2 p + (1-p)^3 p \cdots \\
&= \frac{p}{1-(1-p)} = 1,
\end{aligned}$$

where we recognized the series after the second equality as a geometric series. This gives

$$\mathbb{E}(X) = 1/p.$$

(In fact, one can apply Exercise 1.5.8 and quickly compute the expection of a Geometric RV.)

Exercise 1.6.5. Let $X \sim \text{Geom}(p)$. Use a similar computation as we had in Example 1.6.4 to show $\mathbb{E}(X^2) = (2-p)/p^2$. Using the fact that $\mathbb{E}(X) = 1/p$, conclude that $\text{Var}(X) = (1-p)/p^2$.

Example 1.6.6 (Poisson RV)**.** A RV X is a *Poisson RV* with rate $\lambda > 0$ if

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{15}$$

for all nonnegative integers $k \ge 0$. We write $X \sim \text{Poisson}(\lambda)$.

Poisson distribution is obtained as a limit of the Binomial distribution as the number n of trials tend to infinity while the mean np is kept at constant λ. Namely, let $Y \sim \text{Binomial}(n, p)$ and suppose $np = \lambda$. This means that we expect to see λ successes out of n trials. Then what is the probability that we see,

say, k successes out of n trials, when n is large? Since the mean is λ, this probability should be very small when k is large compared to λ. Indeed, we can rewrite the Binomial PMF as

$$\begin{aligned}\mathbb{P}(Y=k) &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}p^k(1-p)^{n-k}\\ &= \frac{n}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{(np)^k}{k!}(1-p)^{n-k}\\ &= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}.\end{aligned}$$

As n tends to infinity, the limit of the last expression is precisely the right hand side of (15). [16]

Exercise 1.6.7. Let $X \sim \text{Poisson}(\lambda)$. Show that $\mathbb{E}[X] = \text{Var}(X) = \lambda$.

1.6.2. Continuous RVs. In this section, we introduce three important continuous RVs.

Example 1.6.8 (Uniform RV)**.** X is a *uniform RV* on the interval $[a,b]$ (denoted by $X \sim \text{Uniform}([a,b])$) if it has PDF

$$f_X(x) = \frac{1}{b-a}\mathbf{1}(a \le x \le b).$$

An easy computation gives its CDF:

$$\mathbb{P}(X \le x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ 1 & x > b. \end{cases}$$

▲

Exercise 1.6.9. Let $X \sim \text{Uniform}([a,b])$. Show that

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \text{Var}(E) = \frac{(b-a)^2}{12}.$$

Example 1.6.10 (Exponential RV)**.** X is an *exponential* RV with rate λ (denoted by $X \sim \text{Exp}(\lambda)$) if it has PDF

$$f_X(x) = \lambda e^{-\lambda x}\mathbf{1}(x \ge 0).$$

Integrating the PDF gives its CDF

$$\mathbb{P}(X \le x) = (1-e^{-\lambda x})\mathbf{1}(x \ge 0).$$

Using Exercise 1.5.9, we can compute

$$\mathbb{E}(X) = \int_0^\infty e^{-\lambda t}\,dt = \left[-\frac{e^{-\lambda t}}{\lambda}\right]_0^\infty = 1/\lambda.$$

▲

Exercise 1.6.11. Let $X \sim \text{Exp}(\lambda)$. Show that $\mathbb{E}[X] = 1/\lambda$ directly using definition. Also show that $\text{Var}(X) = 1/\lambda^2$.

Example 1.6.12 (Normal RV)**.** X is a *normal* RV with mean μ and variance σ^2 (denoted by $X \sim N(\mu,\sigma^2)$) if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

[16] Later, we will interpret the value of a Poisson variable $X \sim \text{Poisson}(\lambda)$ as the number of customers arriving during a unit time interval, where the waiting time between consecutive customers is distributed as an independent exponential distribution with mean $1/\lambda$. Such an arrival process is called the Poisson process.

If $\mu = 0$ and $\sigma^2 = 1$, then X is called a standard normal RV. Note that if $X \sim N(\mu, \sigma^2)$, then $Y := X - \mu$ has PDF

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

Since this is an even function, it follows that $\mathbb{E}(Y) = 0$. Hence $\mathbb{E}(X) = \mu$. ▲

Exercise 1.6.13 (Gaussian integral)**.** In this exercise, we will show $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$.

(i) Show that

$$\int x e^{-x^2}\,dx = -\frac{1}{2} e^{-x^2} + C.$$

(ii) Let $I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$. Show that

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dx\,dy.$$

(iii) Use polar coordinate (r, θ) to rewrite

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\, r\,dr\,d\theta = 2\pi \int_0^{\infty} r e^{-r^2}\,dr.$$

Then use (i) to deduce $I^2 = \pi$. Conclude $I = \sqrt{\pi}$.

Exercise 1.6.14. Let $X \sim N(\mu, \sigma^2)$. In this exercise, we will show $\mathrm{Var}(X) = \sigma^2$.

(i) Show that $\mathrm{Var}(X) = \mathrm{Var}(X - \mu)$.

(ii) Use integration by parts and Exercise 1.6.13 to show that

$$\int_0^{\infty} x^2 e^{-x^2}\,dx = \left[x\left(-\frac{1}{2}e^{-x^2}\right)\right]_0^{\infty} + \int_0^{\infty} \frac{1}{2} e^{-x^2}\,dx = \frac{\sqrt{\pi}}{4}.$$

(iii) Use change of variable $x = \sqrt{2}\sigma t$ and (ii) to show

$$\int_0^{\infty} \frac{x^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} t^2 e^{-t^2}\,dt = \frac{\sigma^2}{2}.$$

Use (i) to conclude $\mathrm{Var}(X) = \sigma^2$.

Proposition 1.6.15 (Linear transform)**.** *Let X be a RV with PDF f_X. Fix constants $a, b \in \mathbb{R}$ with $a > 0$, and define a new RV $Y = aX + b$. Then*

$$f_{aX+b}(y) = \frac{1}{|a|} f_X((y-b)/a).$$

PROOF. First suppose $a > 0$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \le (y-b)/a) = \int_{-\infty}^{(y-b)/a} f_X(t)\,dt.$$

By differentiating the last integral by y, we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a} f_X((y-b)/a).$$

For $a < 0$, a similar calculation shows

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \ge (y-b)/a) = \int_{(y-b)/a}^{\infty} f_X(t)\,dt,$$

so we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{1}{a} f_X((y-b)/a).$$

This shows the assertion. □

Example 1.6.16 (Linear transform of normal RV is normal)**.** Let $X \sim N(\mu, \sigma^2)$ and fix constants $a, b \in \mathbb{R}$ with $a \neq 0$. Define a new RV $Y = aX + b$. Then since

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

by Proposition 1.6.15, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp\left(-\frac{(y-b-a\mu)^2}{2(a\sigma)^2}\right).$$

Notice that this is the PDF of a normal RV with mean $a\mu + b$ and variance $(a\sigma)^2$. In particular, if we take $a = 1/\sigma$ and $b = \mu/\sigma$, then $Y = (X-\mu)/\sigma \sim N(0,1)$, the standard normal RV. This is called *standardization* of normal RV.

Proposition 1.6.17 (Sum of ind. normal RVs is normal)**.** *Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs. Then*

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

PROOF. Details omitted. One can use the convolution formula or moment generating functions. □

Exercise 1.6.18. Compute the following probabilities using the standard normal table in Table **??**.
(i) $\mathbb{P}(-1 \leq X \leq 2)$ where $X \sim N(0, 3^2)$.
(ii) $\mathbb{P}(X^2 + X - 1 \geq 0)$ where $X \sim N(1,1)$.
(iii) $\mathbb{P}(\exp(2X + 2Y) - 3\exp(X+Y) + 2 \leq 0)$ where $X \sim N(0,1)$ and $Y \sim N(-2,3)$.

Exercise 1.6.19 (Beta distribution)**.** A random varible X taking values from $[0,1]$ has Beta distribution of parameters α and β, which we denote by $\text{Beta}(\alpha, \beta)$, if it has PDF

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $\Gamma(z)$ is the Euler Gamma function defined by

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

(i) Use integration by parts to show the following recursion

$$\Gamma(z+1) = z\Gamma(z).$$

Deduce that $\Gamma(n) = (n-1)!$ for all integers $n \geq 1$.
(ii) Write $A_{n,k} = \int_0^1 y^k (1-y)^{n-k}\, dy$. Use integration by parts and show that

$$A_{n,k} = \frac{k}{n-k+1} A_{n,k-1}.$$

for all $1 \leq k \leq n$. Conclude that for all $0 \leq k \leq n$,

$$A_{n,k} = \frac{1}{\binom{n}{k}} \frac{1}{n+1}.$$

(iii) Let $X \sim \text{Beta}(k+1, n-k+1)$. Use (i) to show that

$$f_X(x) = \frac{n!(n+1)}{k!(n-k)!} x^k (1-x)^{n-k} = \frac{x^k(1-x)^{n-k}}{1/\binom{n}{k}(n+1)}.$$

Use **(ii)** to verify that the above function is indeed a PDF (i.e., it integrates to 1).
(iv) Show that if $X \sim \text{Beta}(\alpha, \beta)$, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

CHAPTER 2

Independence

According to Durrett [Dur19], "*Measure theory ends and probability theory begins with the definition of independence.*" Also, Kac [Kac87] says "*Independence is the central concept of probability theory and few would believe today that understanding what it meant was ever a problem.*"

2.1. Definition of Independence

Let X, Y be two random variable on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We can think of some random experiment modeled by this probability space, and X, Y being two different quantities that we observe about this experiment. For instance, the experiment could be charing an electromagnetic field to accerate an electron, and X and Y could be the position and the momentum of the accelerated electron. While it is reasonable to believe in this case that knowing the value of X can yield some non-trivial information on Y and vice versa, it is possible that there is no information-theoretic relation between the two random variables. Independence is a mathematical notion that captures this intuition into our measuer-theoretic foundation of probability theory.

Definition 2.1.1 (Independence between two objects)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let X, Y be random variables on it.

(i) Two events $A, B \in \mathscr{F}$ are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

(ii) The RVs X, Y are *independent* if $\mathbb{P}(X \in C,\ Y \in D) = \mathbb{P}(X \in C)\mathbb{P}(Y \in D)$ for all Borel subsets C, D of $\mathbb{R}$. Equivalentely, X, Y are independent if the events $X^{-1}(C)$ and $Y^{-1}(D)$ are independent for all Borel subsets C, D of $\mathbb{R}$.

(iii) Let $\mathscr{G}, \mathscr{H} \subseteq \mathscr{F}$ be two σ-algebras on Ω contained in $\mathscr{F}$. Then $\mathscr{G}, \mathscr{H}$ are *independent* if two events A, B are independent for all $A \in \mathscr{G}$ and $B \in \mathscr{H}$.

Definition 2.1.2 (σ-algebra generated by a function)**.** Let Ω be a sample space and let $f : \Omega \to \mathbb{R}$ be an arbitrary function. The *σ-algebra generated by f*, denoted by $\sigma(f)$, is defined as

$$\sigma(f) := \{f^{-1}(B) \mid \forall \text{ Borel subset } B \subseteq \mathbb{R}\}.$$

Exercise 2.1.3 (σ-algebra generated by functions)**.** Let Ω be a sample space and let $f : \Omega \to \mathbb{R}$ be an arbitrary function. Show the following:

(i) Show that $\sigma(f)$ is indeed a σ-algebra.

(ii) Show that $\sigma(f)$ is the smallest σ-algebra that makes f Borel measurable. That is, show that, denoting $\mathscr{B}$ = Borel σ-algebra on $\mathbb{R}$,

$$\sigma(f) = \bigcap \left\{ \mathscr{F} \subseteq 2^{\Omega} \mid \mathscr{F} \text{ is a } \sigma\text{-algebra such that } f \text{ is } (\mathscr{F} - \mathscr{B})\text{-measurable} \right\}.$$

(iii) Let X be a RV on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Show that $\sigma(X) \subseteq \mathscr{F}$.

The following proposition shows that independence between RVs is equivalent to independence between the σ-algebras generated by them.

Proposition 2.1.4. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let X, Y be random variables on it. The following hold:*

(i) *If X, Y are independent, then $\sigma(X), \sigma(Y)$ are independent.*

(ii) *If $\mathscr{G}, \mathscr{H} \subseteq \mathscr{F}$ are independent σ-algebras on Ω and if functions $X, Y : \Omega \to \mathbb{R}$ are $(\mathscr{G}-\mathscr{B})$- and $(\mathscr{H}-\mathscr{B})$-measureable, respectively*[1]*, then X, Y are independent. In particular, if $\sigma(X), \sigma(Y)$ are independent, then X, Y are independent.*

PROOF. (i) First recall that by Exercise 2.1.3, $\sigma(X), \sigma(Y)$ are σ-algebras on Ω contained in $\mathscr{F}$. Fix $A \in \sigma(X)$ and $B \in \sigma(Y)$. By definition of σ-algebra generated by a function, there exists Borel sets $C, D \subseteq \mathbb{R}$ such that $A = X^{-1}(C)$, $B = Y^{-1}(D)$. Then

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(X^{-1}(C) \cap Y^{-1}(D)) = \mathbb{P}(X \in C,\, Y \in D) = \mathbb{P}(X \in C)\,\mathbb{P}(Y \in D) \\ &= \mathbb{P}(X^{-1}(C))\,\mathbb{P}(Y^{-1}(D)) = \mathbb{P}(A)\,\mathbb{P}(B).\end{aligned}$$

So A, B are independent. Since A, B were arbitrary, $\sigma(X), \sigma(Y)$ are independent.

(ii) Let C, D be Borel subsets of $\mathbb{R}$. We wish to show that $\mathbb{P}(X \in C,\, Y \in D) = \mathbb{P}(X \in C)\mathbb{P}(Y \in D)$. But since $X^{-1}(C) \in \mathscr{G}$, $Y^{-1}(D) \in \mathscr{H}$ and since $\mathscr{G}, \mathscr{H}$ are independent, we have $\mathbb{P}(X^{-1}(C) \cap Y^{-1}(D)) = \mathbb{P}(X^{-1}(C))\,\mathbb{P}(Y^{-1}(D))$. The asesrtion then follows.

□

The following proposition shows that independence between events is equivalent to the independence between the corresponding indicator RVs.

Proposition 2.1.5. *$(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Fix two events $A, B \in \mathscr{F}$.*

(i) *If A and B are independent, then so are A^c and B, A and B^c, and A^c and B^c.*

(ii) *A, B are independent if and only if the indicator RVs $\mathbf{1}_A, \mathbf{1}_B$ are independent.*

PROOF. (i) Subtracting $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ from $\mathbb{P}(B) = \mathbb{P}(B)$ shows $\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c)\mathbb{P}(B)$. By symmetry, we also get $\mathbb{P}(B^c \cap A) = \mathbb{P}(B^c)\mathbb{P}(A)$. Subtracting the former from $\mathbb{P}(A^c) = \mathbb{P}(A^c)$ yields $\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\mathbb{P}(B^c)$.

(ii) Recall that $\sigma(\mathbf{1}_E) = \{\emptyset, E, E^c, \Omega\}$ for any event $E \in \mathscr{F}$ (see Example 1.2.4). Hence part (i) shows that A, B are independent if and only if $\sigma(\mathbf{1}_A)$ and $\sigma(\mathbf{1}_B)$ are independent. This is if and only if $\mathbf{1}_A$ and $\mathbf{1}_B$ are independent by Proposition 2.1.7.

□

Next, we introduce independence between several objects.

Definition 2.1.6 (Independence between finitely many objects)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.

(i) Events $A_1, \ldots, A_n \in \mathscr{F}$ are *independent* if for all subset $I \subseteq \{1, \ldots, n\}$,

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

(ii) RVs $X_1, \ldots, X_n$ on $(\Omega, \mathscr{F}, \mathbb{P})$ are *independent* if for all Borel subsets $B_1, \ldots, B_n \subseteq \mathbb{R}$,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in B_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(X \in B_i).$$

(iii) σ-algebras $\mathscr{G}_1, \ldots, \mathscr{G}_n \subseteq \mathscr{F}$ are independent if for all events $A_1 \in \mathscr{G}_1, \ldots, A_n \in \mathscr{G}_n$, $A_1, \ldots, A_n$ are indepdnent.

Proposition 2.1.7. *$(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.*

(i) *If $X_1, \ldots, X_n$ are RVs on $(\Omega, \mathscr{F}, \mathbb{P})$, then they are independent if and only if $\sigma(X_1), \ldots, \sigma(X_n) \subseteq \mathscr{F}$ are independent.*

(ii) *If $A_1, \ldots, A_n \in \mathscr{F}$ are independent, then so are $A_1^c, A_2, \ldots, A_n$.*

(iii) *$A_1, \ldots, A_n \in \mathscr{F}$ are independent if and only if the indicator RVs $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ are independent.*

[1] In this case we write, as shorthand, $X \in \mathscr{G}$ and $Y \in \mathscr{H}$.

PROOF. (i) Suppose $X_1,\dots,X_n$ are independent. Fix events $A_i \in \sigma(X_i)$ for $i = 1,\dots,n$. Then there exists Borel subsets $B_i \subseteq \mathbb{R}$ with $A_i = X^{-1}(B_i)$ for $i = 1,\dots,n$. Then

$$\mathbb{P}\left(\bigcap_{i=1}^n A_n\right) = \mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^n \mathbb{P}(X \in B_i) = \prod_{i=1}^n \mathbb{P}(A_i).$$

Conversely, suppose $\sigma(X_1),\dots,\sigma(X_n) \subseteq \mathscr{F}$ are independent. Fix Borel subsets $B_i \subseteq \mathbb{R}$ for $i = 1,\dots,n$. Then

$$\mathbb{P}\left(\bigcap_{i=1}^n \{X_i \in B_i\}\right) = \mathbb{P}\left(\bigcap_{i=1}^n X^{-1}(B_i)\right) = \prod_{i=1}^n \mathbb{P}\left(X^{-1}(B_i)\right) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i).$$

(ii) In order to show the independence of $A_1^c, A_2, \dots, A_n$, we need to show that

$$\mathbb{P}\left(\bigcap_{i\in I} A_i'\right) = \prod_{i\in I} \mathbb{P}(A_i') \qquad \forall I \subseteq \{1,\dots,n\},$$

where we set $A_i' = A_i^c$ for $i = 1$ and $A_i' = A_i$ for $i \neq 1$. If I does not contain 1, the the equality above follows from the definition of independence of $A_1,\dots,A_n$ with the same index set I. Hence we may assume $1 \in I$. By permuting the indices if necessary, we may assume that $I = \{1,2,\dots,r\}$ for some $1 \leq r \leq n$. Then by applying the definition of independence between events $A_1,\dots,A_n$ with index sets $J = \{1,2,\dots,r\}$ and $J' = \{2,\dots,r\}$, we have $\mathbb{P}\left(\bigcap_{i=1}^r A_i\right) = \prod_{i=1}^r \mathbb{P}(A_i)$ and $\mathbb{P}\left(\bigcap_{i=2}^r A_i\right) = \prod_{i=2}^r \mathbb{P}(A_i)$. Subtracting the former from the latter, we get

$$\mathbb{P}\left(A_1^c \cap A_2 \cap \cdots \cap A_r\right) = \mathbb{P}(A_1^c)\,\mathbb{P}(A_2)\cdots\mathbb{P}(A_r),$$

as desired.

(iii) Repeating the same argument in the proof of (i), we can show that if $A_1,\dots,A_n$ are independent, then for any subset $J \in \{1,\dots,n\}$, the events $A_1',\dots,A_n'$ are independent, where $A_i' = A_i^c$ if $i \in J$ and $A_i' = A_i$ if $i \notin J$. Recall that $\sigma(\mathbf{1}_E) = \{\emptyset, E, E^c, \Omega\}$ for any event $E \in \mathscr{F}$ (see Example 1.2.4). Hence $\sigma(\mathbf{1}_{A_1}),\dots,\sigma(\mathbf{1}_{A_n})$ are independent if and only if for any subset $J \in \{1,\dots,n\}$, the events $A_1',\dots,A_n'$ are independent, where $A_i' = A_i^c$ if $i \in J$ and $A_i' = A_i$ if $i \notin J$. The latter holds if and only if $A_1,\dots,A_n$ are independent by definition and the observation we made at the beginning of this paragraph. To conclude, note that $\mathbf{1}_{A_1},\dots,\mathbf{1}_{A_n}$ are independent if and only if $\sigma(\mathbf{1}_{A_1}),\dots,\sigma(\mathbf{1}_{A_n})$ are independent by part (i).

□

Example 2.1.8 (Pairwise independnece does not imply independence)**.** Events $A_1,\dots,A_n$ are said to by *pairwise independent* if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$, $i,j \in \{1,\dots,n\}$. notice that this is weaker than the definition of independence between events $A_1,\dots,A_n$. In fact, pairwise independence is a strictly weaker notion than independence. To see this, let X_1, X_2, X_3 be independent RVs such that $X_i \sim$ Bernoulli(1/2) for $i = 1,2,3$. Define events

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

These events are pairwise independent since for each $i \neq j$,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(X_1 = X_2 = X_3) = 1/4 = \mathbb{P}(A_i)\,\mathbb{P}(A_j),$$

but they are not independent since

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1 \cap A_2) = 1/4 \neq 1/8 = \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\mathbb{P}(A_3).$$

Intuitively speaking, knowing two of the three events A_1, A_2, A_3 completely determins the remaining event. However, knowing only one of these events does not give any information on any other single event. ▲

2.2. Sufficient condition for independence

Definition 2.2.1 (Independence between collections of events)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $\mathscr{A}_1, \dots, \mathscr{A}_n \subseteq \mathscr{F}$ such that $\Omega \in \mathscr{A}_i$ for all $i = 1, \dots, n$. Then we say $\mathscr{A}_1, \dots, \mathscr{A}_n$ are independent if $A_1, \dots, A_n$ are independent for all $A_1 \in \mathscr{A}_1, \dots, A_n \in \mathscr{A}_n$.

The following proposition shows that independence between σ-algebras generated by some collections of subsets can be checked by only checking the independence between the generating sets.

Proposition 2.2.2 (Independence of generating sets imply independence of σ-algebras)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $\mathscr{A}_1, \dots, \mathscr{A}_n \subseteq \mathscr{F}$ such that $\Omega \in \mathscr{A}_i$ for all $i = 1, \dots, n$. Further assume that each $\mathscr{A}_i$ is a π-system. If $\mathscr{A}_1, \dots, \mathscr{A}_n$ are independent, then $\sigma(\mathscr{A}_1), \dots, \sigma(\mathscr{A}_n)$ are independent.*

PROOF. Let $A_2, \dots, A_n$ be arbitrary sets with $A_i \in \mathscr{A}_i$ for $i = 2, \dots, n$. Denote $F := A_2 \cap \cdots \cap A_n$. Let

$$\mathscr{L} := \{A \in \mathscr{F} \mid \mathbb{P}(A \cap F) = \mathbb{P}(A)\,\mathbb{P}(F)\}.$$

That is, $\mathscr{L}$ is the collection of all events that are independent from F. Since $\mathscr{A}_1, \dots, \mathscr{A}_n$ are independent, it follows that $\mathscr{A}_1 \subseteq \mathscr{L}$. We wish to show that $\sigma(\mathscr{A}_1) \subseteq \mathscr{L}$. By using Dynkin's $\pi - \lambda$ theorem (see Theorem 1.1.37), we only need to verify that $\mathscr{L}$ is a λ-system.

(i) ($\Omega \in \mathscr{L}$): $\mathbb{P}(\Omega \cap F) = \mathbb{P}(F) = 1 \cdot \mathbb{P}(F) = \mathbb{P}(\Omega)\,\mathbb{P}(F)$.

(ii) ($A, B \in \mathscr{L}$ with $A \subseteq B \Rightarrow B \setminus A \in \mathscr{L}$): Since $A \subseteq B$, we have $(A \setminus B) \cap F = (A \cap F) \setminus (B \cap F)$. Hence

$$\mathbb{P}((A \setminus B) \cap F) = \mathbb{P}(A \cap F) - \mathbb{P}(B \cap F) = \mathbb{P}(A)\,\mathbb{P}(F) - \mathbb{P}(B)\,\mathbb{P}(F) = (\mathbb{P}(A) - \mathbb{P}(B))\,\mathbb{P}(F) = \mathbb{P}(A \setminus B)\,\mathbb{P}(F).$$

This shows $A \setminus B \in \mathscr{L}$.

(iii) ($A_1, A_2, \cdots \in \mathscr{L}$, $A_1 \subseteq A_2 \subseteq \cdots \Rightarrow A := \bigcup_{i \geq 1} A_i \in \mathscr{L}$): Note that $A_i \cap F \nearrow A \cap F$. Hence by using continuity of measure from below,

$$\mathbb{P}(A \cap F) = \lim_{n \to \infty} \mathbb{P}(A_n \cap F) = \lim_{n \to \infty} \mathbb{P}(A_n)\,\mathbb{P}(F) = \mathbb{P}(F) \lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}(F)\,\mathbb{P}(A).$$

This shows $A \in \mathscr{L}$.

The above three items verify that $\mathscr{L}$ is a λ-system. Thus we conclude $\sigma(\mathscr{A}_1) \subseteq \mathscr{L}$.

The above discussion shows that, under the hypothesis, $\sigma(A_1), \mathscr{A}_2, \dots, \mathscr{A}_n$ are independent. Repeating the same arguments, one shows that $\sigma(\mathscr{A}_1), \sigma(\mathscr{A}_2), \dots, \sigma(\mathscr{A}_n)$ are independent. □

As a corollary, we deduce that RVs are independent if their joint CDF factorizes into the product of marginal CDFs.

Proposition 2.2.3. *RVs $X_1, \dots, X_n$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ are independent if*

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i) \quad \text{for all } x_1, \dots, x_n \in (-\infty, \infty].$$

PROOF. By Proposition 2.1.4, $X_1, \dots, X_n$ are independent if and only if $\sigma(X_1), \dots, \sigma(X_n)$ are independent. Let $\mathscr{A}$ denote the set of all intervals of the form $(-\infty, a]$ for $a \in \mathbb{R}$ and $\mathbb{R}$ itself. Then $\sigma(\mathscr{A}) = \mathscr{B}$ =Borel σ-algebra on $\mathbb{R}$. Hence for each $i = 1, \dots, n$, $\sigma(X_i)$ is generated by the sets $X_i^{-1}(A)$ for $A \in \mathscr{A}$. The condition in the assertion exactly states that the collections $\mathscr{A}_i := \{X_i^{-1}(A) : A \in \mathscr{A}\}$ for $i = 1, \dots, n$ are independent. Also note that each $\mathscr{A}_i$ is a π-system: For $X^{-1}((-\infty, a]) \cap X^{-1}((-\infty, b]) = X^{-1}((\infty, a \wedge b])$. Hence the assertion follows from Proposition 2.2.2. □

Proposition 2.2.4. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $\mathscr{F}_{i,j}$ for $1 \leq i, j \leq n$ be independent σ-algebras on Ω. For each $1 \leq i \leq n$, let $\mathscr{G}_i := \sigma\left(\bigcup_{1 \leq j \leq n} \mathscr{F}_{i,j}\right)$. Then $\mathscr{G}_1, \dots, \mathscr{G}_n$ are independent.*

PROOF. Let $\mathscr{A}_i$ be the collections of sets of the form $\bigcap_{1\le j\le n} A_{i,j}$ where $A_{i,j}\in\mathscr{F}_{i,j}$. Then each $\mathscr{A}_i$ contains Ω and $\sigma(\mathscr{A}_i)\supseteq\bigcup_{1\le j\le n}\mathscr{F}_{i,j}$. Furthermore, $\mathscr{A}_1,\dots,\mathscr{A}_n$ are independent since, by using the independence of $\mathscr{F}_{i,j}$,

$$\mathbb{P}\left(\bigcap_i\bigcap_j A_{i,j}\right)=\prod_{i,j}\mathbb{P}(A_{i,j})=\prod_i\left(\prod_j\mathbb{P}(A_{i,j})\right)=\prod_i\left(\mathbb{P}\left(\bigcap_j A_{i,j}\right)\right).$$

Hence by Proposition 2.2.2, it follows that $\sigma(\mathscr{A}_1),\dots,\sigma(\mathscr{A}_n)$ are independent. Since $\sigma(\mathscr{A}_i)\supseteq\bigcup_{1\le j\le n}\mathscr{F}_{i,j}$, we also have $\sigma(\mathscr{A}_i)\supseteq\mathscr{G}_i$. Hence $\mathscr{G}_1,\dots,\mathscr{G}_n$ are independent. □

Remark 2.2.5. In the proposition above, we can take $\mathscr{F}_{i,j}=2^\Omega$ for some j's if one wants to apply the statement for different numbers of $\mathscr{F}_{i,j}$'s for each i.

An important consequence of the previous proposition is that functions of disjoint sets of independent RVs are independent.

Proposition 2.2.6. *Let $X_{i,j}$ for $1\le i,j\le n$ be independent RVs on a probability space. $(\Omega,\mathscr{F},\mathbb{P})$. Let $f_i:\mathbb{R}^n\to\mathbb{R}$ be Borel measurable for each $i=1,\dots,n$. Define $Y_i:=f_i(X_{i,1},\dots,X_{i,n})$ for $i=1,\dots,n$. Then $Y_1,\dots,Y_n$ are independent.*

PROOF. By Proposition 2.1.4, $Y_1,\dots,Y_n$ are independent if and only if $\sigma(Y_1),\dots,\sigma(Y_n)$ are independent. Let $\mathscr{F}_{i,j}:=\sigma(X_{i,j})$ and $\mathscr{G}_i:=\sigma(\bigcup_j\mathscr{F}_{i,j})$. Then $\sigma(Y_i)\subseteq\mathscr{G}_i$ for $i=1,\dots,n$. Since $X_{i,j}$'s are independent, $\mathscr{F}_{i,j}$'s are independent. Then by Prooposition 2.2.4, $\mathscr{G}_i$'s are independent. It follows that $\sigma(Y_i)$"s are also independent. □

2.3. Independence, distribution, and expectation

In Section 1.4, we constructed product measures on the product measurable space. The following lemma connects it with the distribution of joint random vectors consisting of independent RVs.

Lemma 2.3.1 (Independence and product measure)**.** *Let $X_1,\dots,X_n$ be independent RVs on a probability space $(\Omega,\mathscr{F},\mathbb{P})$. Let $\mu_i:=\mathbb{P}\circ X_i^{-1}$ denote the distribution of X_i for $i=1,\dots,n$. Then the random vector $(X_1,\dots,X_n):(\Omega,\mathscr{F},\mathbb{P})\to(\mathbb{R}^n,\mathscr{B}^n)$ has distribution $\mu_1\otimes\cdots\otimes\mu_n$.*

PROOF. Fix Borel subsets $A_1,\dots,A_n\subset\mathbb{R}$. Note that

$$\mathbb{P}(X_1\in A_1,\dots,X_n\in A_n)=\prod_{i=1}^n\mathbb{P}(X_i\in A_i)=\prod_{i=1}^n\mu_i(A_i)=(\mu_1\otimes\cdots\otimes\mu_n)(A_1\times\cdots\times A_n),$$

where the three equalities above follow from independence, definition of distribution, and definition of the product measure. This shows that the distribution of $(X_1,\dots,X_n)$ and the product measure $\mu_1\otimes\cdots\otimes\mu_n$ which are both probability measures on $\mathscr{B}^n$, agree on all measurable rectangles, which form a π-system. Two measures agreeing on a π-system must be identical on the σ-algebra generated by that π-system (see Lemma 1.1.38). Hence they agree on the Borel σ-algebra $\mathscr{B}^n$. □

Lemma 2.3.2 (Expectation of a function of independent RVs)**.** *Let X,Y be independent RVs on a probability space $(\Omega,\mathscr{F},\mathbb{P})$ with distributions $\mu:=\mathbb{P}\circ X^{-1}$ and $\nu:=\mathbb{P}\circ Y^{-1}$. If $h:\mathbb{R}^2\to\mathbb{R}$ is a measurable function with either $h\ge 0$ or $\mathbb{E}[|h(X,Y)|]<\infty$, then*

$$\mathbb{E}[h(X,Y)]=\iint h(x,y)\,d\mu(x)\,d\nu(y).$$

In particular, if $h(x,y)=f(x)g(y)$ for $(x,y)\in\mathbb{R}^2$, then

$$\mathbb{E}[f(X)g(Y)]=\mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$

PROOF. Viewing $h(X,Y):\Omega\to\mathbb{R}^2\to\mathbb{R}$,

$$\begin{aligned}\mathbb{E}[h(X,Y)] &= \int_{\mathbb{R}^2} h(\mathbf{x})\, d(\mathbb{P}\circ(X,Y)^{-1}) \qquad (\because \text{change of variables, see Proposition 1.5.3})\\ &= \int_{\mathbb{R}^2} h(\mathbf{x})\, d\mu\otimes\nu \qquad (\because \text{Lemma 2.3.1})\\ &= \iint h(x,y)\, d\mu(x)\, d\nu(y) \qquad (\because \text{Fubini's theorem, see Theorem 1.4.3}).\end{aligned}$$

This show the first assertion. For the second assertion, assume $h(x,y)=f(x)h(y)$ for $(x,y)\in\mathbb{R}^2$. From the first assertion,

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \mathbb{E}[h(X,Y)]\\ &= \iint f(x)g(y)\, d\mu(x)\, d\nu(y)\\ &= \int g(y)\left(\int f(x)\, d\mu(x)\right) d\mu(y)\\ &= \int g(y)\,\mathbb{E}[f(X)]\, d\mu(y) \qquad (\because \text{change of variables, see Proposition 1.5.3})\\ &= \mathbb{E}[f(X)]\int g(y)\, d\mu(y)\\ &= \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] \qquad (\because \text{change of variables, see Proposition 1.5.3}).\end{aligned}$$

This shows the assertion. □

Exercise 2.3.3 (Expectation of product of independent RVs)**.** Let $X_1,\dots,X_n$ be independent RVs on a probability space $(\Omega,\mathscr{F},\mathbb{P})$. Use Lemma 2.3.2 to deduce that if either $X_i\ge 0$ for $i=1,\dots,n$ or $\mathbb{E}[|X_i|]<\infty$ for $i=1,\dots,n$, then

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

Example 2.3.4 (Uncorrelated but dependent)**.** Two random variables X,Y are said to be *uncorrelated* if $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$. Due to Exercise 2.3.3, we know that independence implies uncorrelation. However, the converse is not true. That is, two random variables can be uncorrelated but still be dependent.

Let (X,Y) be a uniformly sampled point from the unit circle in the 2-dimensional plane. Parameterize the unit circle by $S^1=\{(\cos\theta,\sin\theta)\,|\,0\le\theta<2\pi\}$. Then we can first sample a uniform angle $\Theta\sim\text{Uniform}([0,2\pi))$, and then define $(X,Y)=(\cos\Theta,\sin\Theta)$. Recall from your old memory that

$$\sin 2t = 2\cos t\sin t.$$

Now

$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}(\cos\Theta\sin\Theta)\\ &= \frac{1}{2}\mathbb{E}(\sin 2\Theta)\\ &= \frac{1}{2}\int_0^{2\pi}\sin 2t\, dt\\ &= \frac{1}{2}\left[-\frac{1}{2}\cos 2t\right]_0^{2\pi} = 0.\end{aligned}$$

On the other hand,

$$\mathbb{E}(X) = \mathbb{E}(\cos\Theta) = \int_0^{2\pi}\cos t\, dt = 0$$

and likewise $\mathbb{E}(Y) = 0$. This shows $\mathrm{Cov}(X, Y) = 0$, so X and Y are uncorrelated. However, they satisfy the following deterministic relation

$$X^2 + Y^2 = 1,$$

so clearly they cannot be independent. It is clear that why the x- and y-coordinates of a uniformly sampled point from the unit circle are uncorrelated – they have no linear relation. ▲

2.4. Sums of independent RVs – Convolution

When two RVs X and Y are independent and if the new random variable Z is their sum $X + Y$, then the distribution Z is given by the *convolution* of PMFs (or PDFs) of each RV. The idea should be clear from the following baby example.

Example 2.4.1 (Two dice)**.** Roll two dice independently and let their outcome be recorded by RVs X and Y. Note that both X and Y are uniformly distributed over $\{1,2,3,4,5,6\}$. So the pair (X, Y) is uniformly distributed over the (6×6) integer grid $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. In other words,

$$\mathbb{P}((X, Y) = (x, y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

Now, what is the distribution of the sum $Z = X + Y$? Since each point (x, y) in the grid is equality probable, we just need to count the number of such points on the line $x + y = z$, for each value of z. In other words,

$$\mathbb{P}(X + Y = z) = \sum_{x=1}^{6} \mathbb{P}(X = x)\mathbb{P}(Y = z - x).$$

This is easy to compute from the following picture: For example, $\mathbb{P}(X + Y = 7) = 6/36 = 1/6$. ▲

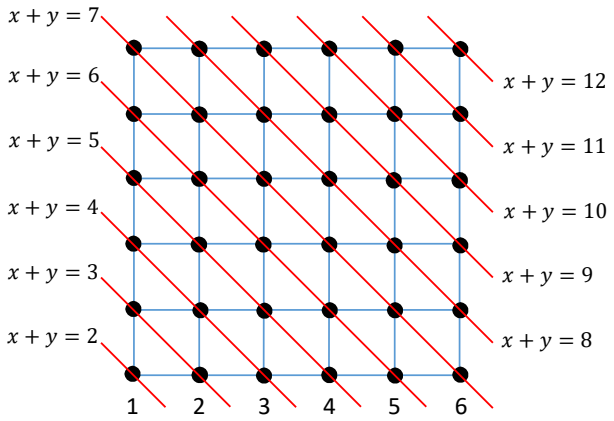


FIGURE 2.4.1. Probability space for two dice and lines on which sum of the two are constant.

Proposition 2.4.2 (Convolution of PMFs)**.** *Let X, Y be two independent integer-valued RVs. Let $Z = X + Y$. Then*

$$\mathbb{P}(Z = z) = \sum_{x} \mathbb{P}(X = x)\mathbb{P}(Y = z - x).$$

PROOF. Note that the pair (X, Y) is distributed over $\mathbb{Z}^2$ according to the distribution

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y),$$

since X and Y are independent. Hence in order to get $\mathbb{P}(Z = z) = \mathbb{P}(X + Y = z)$, we need to add up all probabilities of the pairs (x, y) over the line $x + y = z$. If we first fix the values of x, then y should take value $z - x$. Varying the range of x, we get (2.4.2). □

Exercise 2.4.3 (Sum of ind. Poisson RVs is Poisson)**.** Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent Poisson RVs. Show that $X+Y \sim \text{Poisson}(\lambda_1+\lambda_2)$.

For the continuous case, a similar observation should hold as well. Namely, we should be integrating all the probabilities of the pair (X,Y) at points (x,y) along the line $x+y=z$ in order to get the probability density $f_{X+Y}(z)$. We first derive a general convolution formula for CDFs in Proposition 2.4.4 and then deduce convolution formula for PDFs in Proposition 2.4.5.

Proposition 2.4.4 (Convolution of CDFs)**.** *If X,Y are independent RVs on a probability space $(\Omega,\mathscr{F},\mathbb{P})$, then letting $F(x) := \mathbb{P}(X \le x)$ and $G(y) := \mathbb{P}(Y \le y)$,*

$$\mathbb{P}(X+Y \le z) = \int F(z-y)\,dG(y) =: (F*G)(z),$$

where $\int \cdot\, dG$ is the shorthand of integrating with resepct to the measure ν with distribution function G. Here $F*G$ is called the convolution of F and G.

PROOF. Let μ and ν denote the probability measure on $\mathbb{R}$ with distribution functions F and G, respectively (i.e., $\mu((-\infty,x]) = F(x)$ and so on). Note that

$$\begin{aligned}
\mathbb{P}(X+Y \le z) &= \mathbb{E}[\mathbf{1}(X+Y \le z)] \\
&= \int_{\mathbb{R}^2} \mathbf{1}(x+y \le z)\, d(\mathbb{P}\circ(X,Y)^{-1}) && (\because \text{change of variables, see Proposition 1.5.3}) \\
&= \int_{\mathbb{R}^2} \mathbf{1}(x+y \le z)\, d\mu\otimes\nu(x,y) && (\because \text{Lemma 2.3.1}) \\
&= \iint \mathbf{1}(x \le z-y)\, d\mu(x)\, d\nu(y) && (\because \text{Fubini's theorem, see Theorem 1.4.3}) \\
&= \int F(z-y)\, d\nu(y) && (\because \text{def of } F) \\
&= \int F(z-y)\, dG(y) && (\because \text{def of } dG).
\end{aligned}$$

□

Proposition 2.4.5 (Convolution of PDFs)**.** *Let If X,Y are independent RVs on a probability space $(\Omega,\mathscr{F},\mathbb{P})$. Let $Z := X+Y$.*

(i) *If X is continuous with PDF f_X, then Z has PDF*

$$f_Z(z) := \int_{-\infty}^{\infty} f(z-y)\, dG(y).$$

(ii) *If furthermore Y is continuous with PDF f_Y, then Z has PDF*

$$f_Z(z) := \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy.$$

PROOF. We begin with computing the CDF of Z. By Proposition 2.4.4,

$$\begin{aligned}
\mathbb{P}(Z \le z) &= \int F(z-y)\, dG(y) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(x)\, dx\, dG(y) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_X(x-y)\, dx\, dG(y) \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_X(x-y)\, dG(y)\, dx.
\end{aligned}$$

The last expression shows that Z has PDF $\int_{-\infty}^{\infty} f_X(z-y)\, dG(y)$ as asserted in **(i)**. Then **(ii)** follows from the definition of $\int dG(y)$ and Exercise 1.5.6. □

Example 2.4.6. Let $X, Y \sim N(0,1)$ be independent standard normal RVs. Let $Z = X + Y$. We will show that $Z \sim N(0,2)$ using the convolution formula. Recall that X and Y have the following PDFs:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

By taking convolution of the above PDFs, we have

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-x)^2}{2}\right)\right) dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x)^2}{2}\right) dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-x^2 + xz - \frac{z^2}{2}\right) dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\left(x - \frac{z}{2}\right)^2 - \frac{z^2}{4}\right) dx \\
&= \frac{1}{\sqrt{4\pi}} e^{-z^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\left(x - \frac{z}{2}\right)^2\right) dx = \frac{1}{\sqrt{4\pi}} e^{-z^2/4},
\end{aligned}$$

where we have recognized the integrand in the line as the PDF of $N(-z/2, 1/2)$ so that the integral is 1. Since the last expression is the PDF of $N(0,2)$, it follows that $Z \sim N(0,2)$. ▲.

The following example generalizes the observation we made in the previous example.

Example 2.4.7 (Sum of ind. normal RVs is normal)**.** Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs. We will see that $Z = X+Y$ is again a normal random variable with distribution $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. The usual convolution computation for this is pretty messy (c.f., wikipedia article). Instead let's save some work by using the fact that normal distributions are preserved under linear transform (Exercise 1.6.16). So instead of X and Y, we may consider $X' := (X-\mu_1)/\sigma_1$ and $Y' := (Y-\mu_2)/\sigma_2$ (It is important to note that we must use the same linear transform here for X and Y). Then $X' \sim N(0,1)$, and $Y' \sim N(\mu, \sigma^2)$ where $\mu = (\mu_2 - \mu_1)/\sigma_1$ and $\sigma = \sigma_2/\sigma_1$. Now it suffices to show that $Z' := X' + Y' \sim N(\mu, 1+\sigma^2)$ (see the following exercise for details).

To compute the convolution of the corresponding normal PDFs:

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-x-\mu)^2}{2\sigma^2}\right)\right) dx \\
&= \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x-\mu)^2}{2\sigma^2}\right) dx.
\end{aligned}$$

At this point, we need to 'complete the square' for x for the bracket inside the exponential as below:

$$\begin{aligned}
\frac{x^2}{2} + \frac{(z-x-\mu)^2}{2\sigma^2} &= \frac{1}{2\sigma^2}\left(\sigma^2 x^2 + (x+\mu-z)^2\right) \\
&= \frac{1+\sigma^2}{2\sigma^2}\left(x^2 + \frac{2(\mu-z)x}{1+\sigma^2} + \frac{(\mu-z)^2}{1+\sigma^2}\right) \\
&= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(\mu-z)^2}{1+\sigma^2} - \frac{(\mu-z)^2}{(1+\sigma^2)^2}\right] \\
&= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{1+\sigma^2}\frac{\sigma^2}{1+\sigma^2}\right] \\
&= \frac{1+\sigma^2}{2\sigma^2}\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{2(1+\sigma^2)}.
\end{aligned}$$

Now rewriting (2.4.7),

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right)\exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(x+\frac{(\mu-z)}{1+\sigma^2}\right)^2\right)dx \\ &= \frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{\sigma^2}{1+\sigma^2}}}\exp\left(-\frac{\left(x+\frac{(\mu-z)}{1+\sigma^2}\right)^2}{\frac{2\sigma^2}{1+\sigma^2}}\right)dx \\ &= \frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right), \end{aligned}$$

where we have recognized the integral after second equality as that of the PDF of a normal RV with mean $\frac{z-\mu}{1+\sigma^2}$ and variance $\frac{\sigma^2}{1+\sigma^2}$. Hence $Z' \sim N(\mu, 1+\sigma^2)$, as desired. ▲

Exercise 2.4.8. Let X, Y be independent RVs and fix constants $a > 0$ and $b \in \mathbb{R}$.

(i) Show that $X+Y$ is a normal RV if and only if $(aX+b)+(aY+b)$ is so.

(ii) Show that $X+Y$ is a normal RV, then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$, where $\mu_1 = \mathbb{E}(X)$, $\mu_2 = \mathbb{E}(Y)$, $\sigma_1 = \mathrm{Var}(X)$, and $\sigma_1^2 = \mathrm{Var}(Y)$.

Exercise 2.4.9 (Sum of i.i.d. Exp is Gamma)**.** Let $X_1, X_2, \cdots, X_n \sim \mathrm{Exp}(\lambda)$ be independent exponential RVs.

(i) Show that $f_{X_1+X_2}(z) = \lambda^2 z e^{-\lambda z}\mathbf{1}(z \geq 0)$.

(ii) Show that $f_{X_1+X_2+X_3}(z) = 2^{-1}\lambda^3 z^2 e^{-\lambda z}\mathbf{1}(z \geq 0)$.

(iii) Let $S_n = X_1 + X_2 + \cdots + X_n$. Use induction to show that $S_n \sim \mathrm{Gamma}(n, \lambda)$, that is,

$$f_{S_n}(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!}.$$

CHAPTER 3

Laws of Large Numbers

3.1. Overview of limit theorems

The primary subject in this note is the sequence of i.i.d. RVs and their partial sums. Namely, let $X_1, X_2, \cdots$ be an (infinite) sequence of i.i.d. RVs, and define their nth partial sum $S_n = X_1 + X_2 + \cdots + X_n$ for all $n \geq 1$. If we call X_i the ith step size or *increment*, then the sequence of RVs $(S_n)_{n\geq 1}$ is called a *random walk*, where we usually set $S_0 = 0$. Think of X_i as the gain or loss after betting once in a casino. Then S_n is the net gain of fortune after betting n times. Of course there are ups and downs in the short term, but what we want to analyze using probability theory is the long-term behavior of the random walk $(S_n)_{n\geq 1}$. Results of this type is called limit theorems.

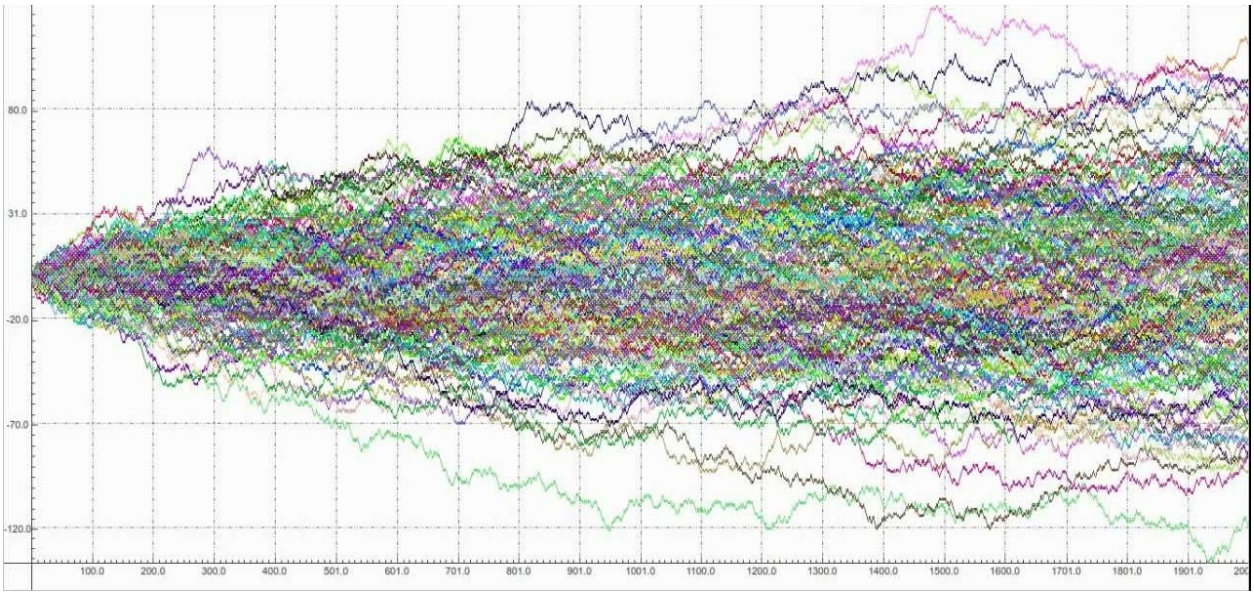


FIGURE 3.1.1. Simulation of simple random walks

Suppose each increment X_k has a finite mean μ. Then by linearity of expectation and independence of the increments, we have

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{\mathbb{E}[S_n]}{n} = \mu,$$
$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\mathrm{Var}(X_1)}{n}.$$

So the sample mean S_n/n has constant expectation and shrinking variance. Hence it makes sense to guess that it should behave as the constant μ, without taking the expectation. That is,

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu.$$

But this expression is shaky, since the left hand side is a limit of RVs while the right hand side is a constant. In what sense the random sample means converge to μ? This is the content of the *law of large numbers*, for which we will prove a weak and a strong versions.

The first limit theorem we will encounter is called the Weak Law of Large Numbers (WLLN), which is stated below:

Theorem 3.1.1 (WLLN, preliminary ver.)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with mean $\mu < \infty$ and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Then for any positive constant $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

In words, the probability that the sample mean S_n/n is *not* within ε distance from its expectation μ decays to zero as n tends to infinity. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *in probability*.

The second version of law of large numbers is call the *strong law of large numbers* (SLLN), which is available if the increments have finite variance.

Theorem 3.1.2 (SLLN, preliminary ver.)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Suppose $\mathbb{E}[X_1] = \mu < \infty$ and $\mathbb{E}[X_1^2] < \infty$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

To make sense out of this, notice that the limit of sample mean $\lim_{n\to} S_n/n$ is itself a RV. Then SLLN says that this RV is well defined and its value is μ with probability 1. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *with probability 1* or *almost surely*.

Perhaps one of the most celebrated theorems in probability theory is the *central limit theorem* (CLT), which tells about how the sample mean S_n/n "fluctuates" around its mean μ. From 3.1, if we denote $\sigma^2 = \mathrm{Var}(X_1) < \infty$, we know that $\mathrm{Var}(S_n/n) = \sigma^2/n \to 0$ as $n \to \infty$. So the fluctuation decays as we add up more increments. To see the effect of fluctuation, we first center the sample mean by subtracting its expectation and "zoom in" by dividing by the standard deviation $\sigma/\sqrt{n}$. This is where the name 'central limit' comes from: it describes the limit of centered random walks.

Theorem 3.1.3 (CLT, preliminary ver.)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Suppose $\mathbb{E}[X_1] = \mu < \infty$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}.$$

Then for all $z \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx.$$

In words, the centered and rescaled RV Z_n is asymptotically distributed as a standard normal RV $Z \sim N(0,1)$. In this case, we say Z_n converges to Z as $n \to \infty$ *in distribution*. This is a remarkable result since as long as the increments X_k have finite mean and variance, it does not matter which distribution that they follow: the 'central limit' always looks like a standard normal distribution. Later in this section, we will prove this result by using the MGF of S_n and Taylor-expanding it up to the second order term.

3.2. Bounding tail probabilities

In this subsection, we introduce two general inequalities called the Markov's and Chebyshef's inequalities. They are useful in bounding tail probabilities of the form $\mathbb{P}(X \geq x)$ using the expectation $\mathbb{E}[X]$ and variance $\mathrm{Var}(X)$, respectively. Their proofs are quite simple but they have lots of nice applications and implications.

Proposition 3.2.1 (Markov's inequality)**.** *Let $X \geq 0$ be a nonnegative RV with finite expectation. Then for any $a > 0$, we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

PROOF. Consider an auxiliary RV Y define as follows:

$$Y = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{if } X < a. \end{cases}$$

Note that we always have $Y \leq X$. Hence we should have $\mathbb{E}[Y] \leq \mathbb{E}[X]$. But since $\mathbb{E}[Y] = a\mathbb{P}(X \geq a)$, we have

$$\lambda\mathbb{P}(X \geq a) \leq \mathbb{E}[X].$$

Dividing both sides by $a > 0$ gives the assertion. □

Example 3.2.2. We will show that, for any RV Z, $\mathbb{E}[Z^2] = 0$ implies $\mathbb{P}(Z = 0) = 1$. Indeed, Markov's inequality gives that for any $a > 0$,

$$\mathbb{P}(Z^2 \geq a) \leq \frac{\mathbb{E}[Z^2]}{a} = 0.$$

By continuity of measure, it follows that $\mathbb{P}(Z^2 = 0) = \lim_{\varepsilon \searrow 0} \mathbb{P}(Z^2 < \varepsilon) = 1$, so $\mathbb{P}(Z = 0) = 1$. ▲

Proposition 3.2.3 (Chebyshef's inequality)**.** *Let X be any RV with $\mathbb{E}[X] = \mu < \infty$ and $\mathrm{Var}(X) < \infty$. Then for any $a > 0$, we have*

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

PROOF. Applying Markov's inequality for the nonnegative RV $(X - \mu)^2$, we get

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}((X - \mu)^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

□

Example 3.2.4. Let $X \sim \mathrm{Exp}(\lambda)$. Since $\mathbb{E}[X] = 1/\lambda$, for any $a > 0$, the Markov's inequality gives

$$\mathbb{P}(X \geq a) \leq \frac{1}{a\lambda},$$

while the true probability is

$$\mathbb{P}(X \geq a) = e^{-\lambda a}.$$

On the other hand, $\mathrm{Var}(X) = 1/\lambda^2$ so Chebyshef's inequality gives

$$\mathbb{P}(|X - 1/\lambda| \geq a) = \frac{1}{a^2\lambda^2}.$$

If $1/\lambda \leq a$, the true probability is

$$\begin{aligned} \mathbb{P}(|X - 1/\lambda| \geq a) &= \mathbb{P}(X \geq a + 1/\lambda) + \mathbb{P}(X \leq -a + 1/\lambda) \\ &= \mathbb{P}(X \geq a + 1/\lambda) = e^{-\lambda(a+1/\lambda)} = e^{-1-\lambda a}. \end{aligned}$$

As we can see, both Markov's and Chebyshef's inequalities give loose estimates, but the latter gives a slightly stronger bound. ▲

Example 3.2.5 (Chebyshef's inequality for bounded RVs)**.** Let X be a RV taking values from the interval $[a,b]$. Suppose we don't know anything else about X. Can we say anything useful about tail probability $\mathbb{P}(X \geq \lambda)$? If we were to use Markov's inequality, then certainly $a \leq \mathbb{E}[X] \leq b$ and in the worst case $\mathbb{E}[X] = b$. Hence we can at least conclude

$$\mathbb{P}(X \geq \lambda) \leq \frac{b}{\lambda}.$$

On the other hand, let's get a bound on $\mathrm{Var}(X)$ and use Chebyshef's inequality instead. We claim that

$$\mathrm{Var}(X) \leq \frac{(b-a)^4}{4},$$

which would yield by Chebyshef's inequality that

$$\mathbb{P}(|X - \mathbb{E}[X]| \leq \lambda) \leq \frac{(b-a)^2}{4\lambda^2}.$$

Intuitively speaking, $\mathrm{Var}(X)$ is the largest when the value of X is as much spread out as possible at the two extreme values, a and b. Hence the largest variance will be achieved when X takes a and b with equal probabilities. In this case, $\mathbb{E}[X] = (a+b)/2$ so

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2+b^2}{2} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{4}.$$

▲

Exercise 3.2.6. Let X be a RV taking values from the interval $[a,b]$.

(i) Use the usual 'completing squares' trick for a second moment to show that

$$0 \leq \mathbb{E}[(X-t)^2] = (t - \mathbb{E}[X])^2 + \mathrm{Var}(X) \quad \forall t \in \mathbb{R}.$$

(ii) Conclude that $\mathbb{E}[(X-t)^2]$ is minimized when $t = \mathbb{E}[X]$ and the minimum is $\mathrm{Var}(X)$.

(iii) By plugging in $t = (a+b)/2$ in (3.2.6), show that

$$\mathrm{Var}(X) = \mathbb{E}[(X-a)(X-b)] + \frac{(b-a)^2}{4} - \left(\mathbb{E}[X] - \frac{a+b}{2}\right)^2.$$

(iv) Show that $\mathbb{E}[(X-a)(X-b)] \leq 0$.

(v) Conclude that $\mathrm{Var}(X) \leq (b-a)^2/4$, where the equality holds if and only if X takes the extreme values a and b with equal probabilities.

Exercise 3.2.7 (Paley-Zigmund inequality)**.** Let X be a nonnegative RV with $\mathbb{E}[X^2] < \infty$. Fix a constant $\theta \geq 0$. We prove the Paley-Zigmund inequality, which gives a lower bound on the tail probabilities and also implies the so-called 'second moment method'.

(i) Write $X = X\mathbf{1}(X > \theta\mathbb{E}[X]) + X\mathbf{1}(X \leq \theta\mathbb{E}[X])$. Show that

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbf{1}(X \leq \theta\mathbb{E}[X])] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])] \\ &\leq \theta\mathbb{E}[X] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])].\end{aligned}$$

(ii) Use Cauchy-Schwartz inequality (Exercise 1.5.25) to show

$$\begin{aligned}(\mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])])^2 &\leq \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])] \\ &= \mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X).\end{aligned}$$

(iii) From (i) and (ii), derive

$$\mathbb{E}[X] \leq \theta\mathbb{E}[X] + \sqrt{\mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X)}.$$

(Note that since $X \geq 0$, the above inequality is meaningful only when $\theta \leq 1$.) Conclude that, for $\theta \in [0,1]$,

$$\mathbb{P}(X > \theta \mathbb{E}X) \geq \frac{(1-\theta)^2 \mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

(iv) (Second moment method) From (iii), conclude that

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Alternatively, use the Cauchy-Schwarz inequality (Exercise 1.5.25) to the RV $X\mathbf{1}(X > 0)$ to deduce the above result.

Proposition 3.2.8 (Chernoff bound)**.** *Let $X_n := \xi_1 + \cdots + \xi_n$, where $\xi_1, \ldots, \xi_n$ are i.i.d. RVs such that $\mathbb{E}[\exp(\theta \xi_1)] < \infty$ for $\theta \in [0, c)$. Denote the log moment generating function of ξ_1 as $\varphi(\theta) = \log \mathbb{E}[\exp(\theta \xi_1)]$. Then*

$$\mathbb{P}(X_n \geq t) \leq \exp\left(- \sup_{\theta \in [0,c)} \theta t - n\varphi(\theta) \right). \tag{16}$$

Furthermore, assume that $\mathbb{E}[\xi_1] = 0$. Then $L := \sup_{0 \leq \theta \leq c/2} |\varphi''(\theta)| < \infty$ and

$$\mathbb{P}(X_n \geq t) \leq \begin{cases} \exp\left(-\frac{Lt^2}{2}\right) & \text{if } t \leq cL/2 \quad (\textit{Gaussian tail}) \\ \exp\left(-\frac{ct}{2L} + \frac{c^2}{8L}\right) & \text{if } t > cL/2 \quad (\textit{Exponential tail}). \end{cases}$$

PROOF. For any parameter $\theta \in [0, c)$, by exponentiating and taking Markov's inequality,

$$\begin{aligned} \mathbb{P}(X_n \geq t) &\leq \mathbb{P}(\exp(\theta X_n) \geq \exp(\theta t)) \\ &\leq \exp(-\theta t)\, \mathbb{E}[\exp(\theta X_n)] \\ &\leq \exp(-\theta t)\, \mathbb{E}[\exp(\theta \xi_1) \cdots \exp(\theta \xi_n)] \\ &= \exp\left(-\theta t + n\varphi(\theta)\right), \end{aligned} \tag{17}$$

where the identity above uses the independence between the increments ξ_i. This holds for all $\theta \in [0, c)$. This shows (16).

Next, denote the moment generating function of ξ_1 as $\psi(\theta) = \mathbb{E}[\exp(\theta \xi_1)]$. Recall that it is a power series with radius of convergence $\geq c$ by the hypothesis. Hence we can differentiate it term-by-term when $\theta \in (-c, c)$. In particular, this gives $\psi(0) = 1$, $\psi'(0) = \mathbb{E}[\xi_1] = 0$, and $\psi''(0) = \mathbb{E}[\xi_1^2] \geq 0$. It follows that $\varphi(0) = 0$, $\varphi'(0) = \psi'(0) = 0$, and $\varphi''(0) = \psi''(0) \geq 0$. Then by Taylor's theorem,

$$\varphi(\theta) \leq \varphi(0) + \varphi'(0)\theta + \frac{L}{2}\theta^2 = \frac{L}{2}\theta^2 \quad \text{for all } \theta \in [0, c/2].$$

This and (16) gives

$$\mathbb{P}(X_n \geq t) \leq \exp\left(- \sup_{\theta \in [0, c/2]} \theta t - \frac{L}{2}\theta^2 \right).$$

The quadratic function $\theta \mapsto \theta t - \frac{L}{2}\theta^2$ is globally minimized at $\theta = t/L$ with minimum value $\frac{L\theta^2}{2}$. Hence when $t < cL/2$, the supremum in the above bound is attained at $\theta = t/L$. If $t \geq c/2L$, the same quadratic function is increasing in $[0, c/2L]$ so the constrained maximization is solved at $\theta = c/2L$. □

Example 3.2.9 (Chernoff bound for Gamma distribution)**.** Let $\xi_1, \ldots, \xi_n$ be i.i.d. Exp(c) RVs (with mean $1/c$). Let $X_n := \xi_1 + \cdots + \xi_n \sim \text{Gamma}(n, c)$. Recall that

$$\psi(\theta) = \mathbb{E}[\exp(\theta \xi_1)] = \frac{c}{c - \theta} \qquad \text{for } \theta < c.$$

By Prop. 3.2.8, for $t \geq 0$,

$$\mathbb{P}(X_n \geq t) \leq \exp\left(- \sup_{\theta \in [0,c)} \theta t - n \log \frac{c}{c-\theta} \right).$$

Let $g(\theta) := \theta t - n \log \frac{c}{c-\theta}$. Then $g'(\theta) = t - \frac{n}{c-\theta}$ and $g''(\theta) = -\frac{n}{(c-\theta)^2} \leq 0$. Thus $\theta = c - (n/t)$ is the global maximizer of g. Using this choice,

$$\mathbb{P}(X_n \geq t) \leq \exp\left(-(ct - n) + n \log \frac{ct}{n} \right).$$

Note that the above tail bound is only useful when $t \gg n/c = \mathbb{E}[X_n]$. ▲

3.3. Weak Law of Large Numbers

3.3.1. L^2 weak law and examples. In this section, we state and prove various versions of the weak law of large numbers (Theorem 3.1.1).

Estimating the variance of the sum of n RVs is of central interest. In general, when we try to write down the variance of a sum of RVs, in addition to the variance of each RV, we have extra contribution of covariance between each pairs of distinct RVs.

Definition 3.3.1. Let $X_1, \ldots, X_n$ be RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We say they are *uncorrelated* if

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] \qquad \text{for all } 1 \leq i < j \leq n.$$

Note that independent RVs are uncorrelated by Exercise 2.3.3.

Exercise 3.3.2. In this exercise, we will see how we can express the variance of sums of RVs.

(i) Show that

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

(ii) Use induction to show that for RVs $X_1, X_2 \cdots, X_n$

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

(iii) Show that if $X_1, \cdots, X_n$ are uncorrelated RVs, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

Assuming finite variance for each increment, the weak law is an easy consequence of Chebyshef's inequality.

Theorem 3.3.3 (L^2 weak law)**.** *Let $(X_k)_{k \geq 1}$ be uncorrelated RVs with finite mean $\mu < \infty$ and uniformly bounded finite variance. Let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Then for any positive constant $\varepsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{S_n}{n} - \mu \right| > \varepsilon \right) = 0.$$

PROOF. Let $C > 0$ be a constant such that $\mathrm{Var}(X_k) < C$ for $k \geq 1$. By Chebyshef's inequality, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left| \frac{S_n}{n} - \mu \right| > \varepsilon \right) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{C}{n\varepsilon^2},$$

where for the last equality we used Exercise (3.3.2). The last expression converges to 0 as $n \to \infty$. □

The weak law of large numbers is the first time that we encounter the notion of 'convergence in probability'. We say a sequence of RVs converge to a constant in probability if the the probability of staying away from that constant goes to zero:

Definition 3.3.4. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let $\mu \in \mathbb{R}$ be a constant. We say X_n converges to μ *in probability* if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|X_n - \mu\right| > \varepsilon\right) = 0.$$

Exercise 3.3.5. Let $X_n \to x$ and $Y_n \to y$ in probability as $n \to \infty$.

(i) Show that for every $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n + Y_n - x - y| > \varepsilon\right) \leq \mathbb{P}\left(|X_n - x| > \varepsilon/2\right) + \mathbb{P}\left(|Y_n - y| > \varepsilon/2\right).$$

Conclude that $X_n + Y_n \to x + y$ in probability as $n \to \infty$.

(ii) Show that for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(|X_nY_n - xy| > \varepsilon\right) &= \mathbb{P}\left(|X_nY_n - X_ny + X_ny - xy| > \varepsilon\right)\\ &\leq \mathbb{P}\left(|X_n||Y_n - y| + |y||X_n - x| > \varepsilon\right)\\ &\leq \mathbb{P}\left(|X_n||Y_n - y| > \varepsilon/2\right) + \mathbb{P}\left(|y||X_n - x| > \varepsilon/2\right).\end{aligned}$$

Conclude that $X_nY_n \to xy$ in probability as $n \to \infty$.

(iii) Suppose $x \neq 0$ and $\mathbb{P}(X_n \neq 0) = 1$ for all $n \geq 1$. Show that for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{Y_n}{X_n} - \frac{y}{x}\right| > \varepsilon\right) &= \mathbb{P}\left(\left|\frac{Y_n}{X_n} - \frac{y}{X_n} + \frac{y}{X_n} - \frac{y}{x}\right| > \varepsilon\right)\\ &\leq \mathbb{P}\left(\frac{1}{|X_n|}\left|Y_n - y\right| + |y|\frac{|X_n - x|}{|X_nx|} > \varepsilon\right)\\ &\leq \mathbb{P}\left(\frac{1}{|X_n|}\left|Y_n - y\right| > \varepsilon/2\right) + \mathbb{P}\left(|y|\frac{|X_n - x|}{|X_nx|} > \varepsilon/2\right).\end{aligned}$$

Conclude that $Y_n/X_n \to y/x$ in probability as $n \to \infty$.

Before we proceed further, let us take a moment and think about the definition of convergence in probability. Recall that a sequence of real numbers $(x_n)_{n\geq 0}$ *converges* to x if for each 'error level' $\varepsilon > 0$, there exists a large integer $N(\varepsilon) > 0$ such that

$$|x_n - x| < \varepsilon \quad \forall n \geq N(\varepsilon).$$

If we would like to say that a sequence of RVs $(X_n)_{\geq 0}$ 'converges' to some real number x, how should we formulate this? Since X_n is an RV, $\{|X_n - x| < \varepsilon\}$ is an event. On the other hand, we can also view each x_n as an RV, even though it is a real number. Then we can rewrite (3.3.1) as

$$\mathbb{P}(|x_n - x| < \varepsilon) = 1 \quad \forall n \geq N(\varepsilon).$$

For general RVs, requiring $\mathbb{P}(|X_n - x| < \varepsilon) = 1$ for any large n might not be possible. But we can fix any desired level of 'confidence', $\delta > 0$, and require

$$\mathbb{P}(|x_n - x| < \varepsilon) \geq 1 - \delta$$

for sufficiently large n. This is precisely (3.3.4).

Example 3.3.6 (Empirical frequency)**.** Let A be an event of interest. We would like to estimate the unknown probability $p = \mathbb{P}(A)$ by observing a sequence of indpendent experiments. namely, let $(X_k)_{k\geq 0}$ be a sequence of i.i.d. RVs where $X_k = \mathbf{1}(A)$ is the indicator variable of the event A for each $k \geq 1$. Let $\hat{p}_n := (X_1 + \cdots + X_n)/n$. Since $\mathbb{E}[X_1] = \mathbb{P}(A) = p$, by WLLN we conclude that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

▲

Example 3.3.7 (Polling)**.** Let E_A be the event that a randomly select voter supports candidate A. Using a poll, we would like to estimate $p = \mathbb{P}(E_A)$, which can be understood as the proportion of supporters of candidate A. As before, we observe a sequence of i.i.d. indicator variables $X_k = \mathbf{1}(E_A)$. Let $\hat{p}_n := S_n/n$ be the empirical proportion of supporters of A out of n samples. We know by WLLN that $\hat{p}_n$ converges to p in probability. But if we want to guarantee a certain confidence level α for an error bound ε, how many samples should be take?

By Chebyshef's inequality, we get the following estimate:

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \le \frac{\mathrm{Var}(\hat{p}_n)}{\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon^2} \le \frac{1}{4n\varepsilon^2}.$$

Note that for the last inequality, we noticed that $X_1 \in [0,1]$ and used Exercise 3.2.6 (or you can use that for $Y \sim \text{Bernoulli}(p)$, $\mathrm{Var}(Y) = p(1-p) \le 1/4$). Hence, for instance, if $\varepsilon = 0.01$ and $\alpha = 0.95$, then we would need to set n large enough so that

$$\mathbb{P}\left(|\hat{p}_n - p| > 0.01\right) \le \frac{10000}{4n} \le 0.05.$$

This yields $n \ge 50{,}000$. In other words, if we survey at least $n = 50{,}000$ independent voters, then the empirical frequency $\hat{p}_n$ is between $p-0.01$ and $p+0.01$ with probability at least 0.95. Still in other words, the true frequency p is between $\hat{p}_n - 0.01$ and $\hat{p}_n + 0.01$ with probability at least 0.95 if $n \ge 50{,}000$. We don't actually need this many samples. We will improve this result later using central limit theorem. ▲

Example 3.3.8 (Bernstein's polynomial approximation)**.** Let $f : [0,1] \to \mathbb{R}$ be a continuous function. For each $n \ge 1$, define the *Bernstein polynomial* f_n of degree n by

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

We claim that

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

PROOF. While the above statement is completely analytic, we will use probability theory to prove it. Fix $p \in [0,1]$ and let $X_1, X_2, \dots$ be i.i.d. RVs with distribution Bernoulli(p). Then $\mathbb{E}[X_i] = p$ and $\mathrm{Var}(X_i) = p(1-p)$ for $i \ge 1$. Let $S_n := X_1 + \cdots + X_n$. Note that

$$\begin{aligned}\mathbb{E}[f(S_n/n)] &= \sum_{m=0}^{n} f(m/n)\,\mathbb{P}(S_n = m) \\ &= \sum_{m=0}^{n} f(m/n) \binom{n}{m} p^m (1-p)^{n-m} = f_n(p).\end{aligned}$$

For each $\delta > 0$, by using Chebyshef's inequality and the fact that $p(1-p) \le 1/4$ for $p \in [0,1]$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \delta\right) \le \frac{\mathrm{Var}(S_n/n)}{\delta^2} = \frac{p(1-p)}{n\delta^2} \le \frac{1}{4n\delta^2} \quad \forall\, p \in [0,1]. \tag{18}$$

Now since f is continuous on the compact interval $[0,1]$ it is uniformly continuous. That is, for each $\varepsilon > 0$, there exists $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \delta$. Fix $\varepsilon > 0$, and let $M := \sup_{x\in[0,1]} f(x)$. Then by Jensen's inequality and using (18),

$$\begin{aligned}\left|f(p) - f_n(p)\right| &= \left|f(p) - \mathbb{E}\left[f(S_n/n)\right]\right| \\ &\le \mathbb{E}\left[\left|f(S_n/n) - f(p)\right|\right] \\ &= \mathbb{E}\left[\left|f(S_n/n) - f_n(p)\right| \mathbf{1}\left(\left|\frac{S_n}{n} - p\right| \le \delta\right)\right] + \mathbb{E}\left[\left|f(S_n/n) - f_n(p)\right| \mathbf{1}\left(\left|\frac{S_n}{n} - p\right| > \delta\right)\right] \\ &\le \varepsilon + 2M\,\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \delta\right) \le \varepsilon + \frac{M}{n\delta^2}.\end{aligned}$$

This shows that

$$\limsup_{n\to\infty} \sup_{x\in[0,1]} \left|f(x) - f_n(x)\right| \le \limsup_{n\to\infty} \left(\varepsilon + \frac{M}{n\delta^2}\right) = \varepsilon.$$

But since $\varepsilon > 0$ was arbitrary, this shows that the above limsup equals zero, as desired. □

▲

Exercise 3.3.9 (Monte Carlo integration)**.** This exercise introduces a probabilistic technique to approximate a complicated intergral, called the Monte Carlo integration, which is based on Weak Law of Large Numbers and Chebyshef's inequality.

Let $(X_k)_{k\ge1}$ be i.i.d. Uniform($[0,1]$) RVs and let $f : [0,1] \to \mathbb{R}$ be a continuous function. Let

$$I_n = \frac{1}{n}\left(f(X_1) + f(X_2) + \cdots + f(X_n)\right)$$

for each $n \ge 1$.

(i) Suppose $\int_0^1 |f(x)|\,dx < \infty$. Show that $I_n \to I := \int_0^1 f(x)\,dx$ in probability. Thus I_n serves as a probabilistic estimation of the unknown integral I.

(ii) Further assume that $\int_0^1 |f(x)|^2\,dx < \infty$. Use Chebyshef's inequality to show that

$$\mathbb{P}\left(|I_n - I| \ge a/\sqrt{n}\right) \le \frac{\mathrm{Var}(f(X_1))}{a^2} = \frac{1}{a^2}\left(\int_0^1 f(x)^2\,dx - I^2\right).$$

Exercise 3.3.10. Let $(X_n)_{n\ge0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$.

(i) For each $\varepsilon > 0$, show that $\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda\varepsilon n}$.

(ii) Conclude that $Y_n \to 0$ in probability as $n \to \infty$.

Example 3.3.11. For each integer $n \ge 1$, define a RV X_n by

$$X_n = \begin{cases} n & \text{with prob. } 1/n \\ 1/n & \text{with prob. } 1 - 1/n. \end{cases}$$

Then $X_n \to 0$ in probability as $n \to \infty$. Indeed, for each $\varepsilon > 0$,

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = 1/n$$

for all $n > 1/\varepsilon$. Hence $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = 0$. However, note that

$$\mathbb{E}[X_n] = 1 + n^{-1} - n^{-2} \to 1 \quad \text{as } n \to \infty.$$

This example indicates that convergence in probability only cares about probability of the event $\mathbb{P}(|X_n - \mathbb{E}[X_n]| > \varepsilon)$ but not the actual value of X_n when that 'bad' event occurs. ▲

Example 3.3.12 (Coupon collector's problem)**.** Let $(X_t)_{\ge t}$ be a sequence of i.i.d. Uniform($\{1, 2, \cdots, n\}$) variables. Think of the value of X_t as the label of the coupon you collect at tth trial. We are interested in how many times we need to reveal a new random coupon to collect a full set of n distinct coupons. That is, let

$$\tau^n = \min\{r \ge 1 \,|\, \#\{X_1, X_2, \cdots, X_r\} = n\}.$$

Because of the possible overlap, we expect n reveals should not get us the full set of n coupons. Indeed,

$$\mathbb{P}(\tau^n = n) = \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{1}{n} = \frac{n!}{n^n}.$$

Certainly this probability rapidly goes to zero as $n \to \infty$. So we need to reveal more than n coupons. But how many? The answer turns out to be $\tau^n \approx n \log n$. More precisely,

$$\frac{\tau^n}{n\log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.} \tag{19}$$

A change of perspective might help us. Instead of waiting to collect all n coupons, let's progressively collect k distinct coupons for $k = 1$ to n. Namely, for each $1 \le k \le n$, define

$$\tau^k = \min\{r \ge 1 \,|\, \#\{X_1, X_2, \cdots, X_r\} = k\}.$$

So τ^k is the first time that we collect k distinct coupons.

Now consider what has to happen to collect $k+1$ distinct coupons from k distinct coupons? Here is an example. Say at time τ^2 we have coupons $\{1,3\}$. τ^3 is the first time that we pick up a new coupon from except 1 and 3. This happens with probability $(n-2)/n$ and since each draw is i.i.d.,

$$\tau^3 - \tau^2 \sim \text{Geom}\left(\frac{n-2}{n}\right).$$

A similar reasoning shows

$$\tau^{k+1} - \tau^k \sim \text{Geom}\left(\frac{n-k}{n}\right).$$

So starting from the first coupon, we wait a $\text{Geom}(1/n)$ time to get a new coupon, and wait a $\text{Geom}(2/n)$ time to get another new coupon, and so on. Note that these geometric waiting times are all independent. So we can decompose τ^n into a sum of independent geometric RVs:

$$\tau^n = \sum_{k=1}^{n-1} (\tau^{k+1} - \tau^k).$$

Then using the estimates in Exercise 3.3.13, it is straightforward to show that

$$\mathbb{E}[\tau^n] \approx n \log n, \quad \text{Var}(\tau^n) \le n^2.$$

In Exercise 3.3.14, we will show (19) using Chebyshef's inequality. ▲

Exercise 3.3.13. In this exercise, we estimate some partial sums using integral comparison.

(i) For any integer $d \ge 1$, show that

$$\sum_{k=2}^{n} \frac{1}{k^d} \le \int_1^n \frac{1}{x^d}\, dx \le \sum_{k=1}^{n-1} \frac{1}{k^d}$$

by considering the upper and lower sum for the Riemann integral $\int_1^n x^{-d}\, dx$.

(ii) Show that

$$\log n \le \sum_{k=1}^{n-1} \frac{1}{k} \le 1 + \log(n-1).$$

(iii) Show that for all $d \ge 2$,

$$\sum_{k=1}^{n-1} \frac{1}{k^d} \le \sum_{k=1}^{\infty} \frac{1}{k^d} \le 1 + \int_1^{\infty} \frac{1}{x^d}\, dx \le 2.$$

Exercise 3.3.14. For each $n \ge 1$, let $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$ be a sequence of independent geometric RVs where $X_{k,n} \sim \text{Geom}((n-k)/n)$. Define $\tau^n = X_{1,n} + X_{2,n} + \cdots + X_{n,n}$.

(i) Show that $\mathbb{E}[\tau^n] = n \sum_{k=1}^{n-1} k^{-1}$. Using Exercise 3.3.13 (ii), deduce that

$$n \log n \le \mathbb{E}[\tau^n] \le n \log(n-1) + n.$$

(ii) Using $\text{Var}(\text{Geom}(p)) = (1-p)/p^2 \le p^{-2}$ and Exercise 3.3.13 (iii), show that

$$\text{Var}(\tau^n) \le n^2 \sum_{k=1}^{n-1} k^{-2} \le 2n^2.$$

(iii) By Chebyshef's inequality, show that for each $\varepsilon > 0$,

$$\mathbb{P}\left(|\tau^n - \mathbb{E}[\tau^n]| > \varepsilon n \log n\right) \le \frac{\mathrm{Var}(\tau^n)}{\varepsilon^2 n^2 \log^2 n} \le \frac{2}{\varepsilon^2 \log^2 n}.$$

Conclude that

$$\frac{\tau^n - \mathbb{E}[\tau^n]}{n \log n} \to 0 \quad \text{as } n \to \infty \text{ in probability.}$$

(iv) By using part (i), conclude that

$$\frac{\tau^n}{n \log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.}$$

Exercise 3.3.15 (Tail bound on the coupon collecting time)**.** Let τ denote the first time to collect n distinct types of coupons, where each draw picks one of the n types independently uniformly at random. Whow that for each $c > 0$,

$$\mathbb{P}(\tau \ge \lceil n \log n + cn \rceil) \le e^{-c}.$$

(*Hint:* Let A_i denote the event that the ith coupon is not selected among the first $\lceil n \log n + cn \rceil$ draws. Then

$$\mathbb{P}(\tau \ge \lceil n \log n + cn \rceil) \le \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \le \sum_{i=1}^n \mathbb{P}(A_i) \le \sum_{i=1}^n \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} \le n \exp\left(-\log n - c\right) = e^{-c}.$$

)

3.3.2. Weak law without finite second moment. In this section, we will state and prove more general versions of the weak law of large numbers than Theorem 3.3.3, which assumed finite second moments. The goal of this section is to establish the following general WLLN:

Theorem 3.3.16 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. RVs. Suppose that*

$$\lim_{x \to \infty} x \mathbb{P}(|X_1| > x) = 0. \tag{20}$$

Let $S_n := X_1 + \cdots + X_n$ and $\mu_n := \mathbb{E}[X_1 \mathbf{1}(|X_1| \le n)]$. Then

$$\frac{S_n - n\mu_n}{n} \to 0 \quad \text{as } n \to \infty \text{ in probability.}$$

Note that in Theorem 3.3.16, we do not even assume finite first moment condition for X_i's. In this case, we cannot use Chebyshef's inequality directly as we did in the proof of Theorem 3.3.3. An immediate consequence of the above result is the following familiar version of WLLN assuming finite mean:

Corollary 3.3.17 (Weak Law of Large Numbers with finite mean)**.** *Let $X_1, X_2, \ldots$ be i.i.d. RVs. Suppose that $\mathbb{E}[|X_1|] < \infty$. Let $S_n := X_1 + \cdots + X_n$ and $\mu_n := \mathbb{E}[X_1]$. Then*

$$\frac{S_n}{n} \to \mu \quad \text{as } n \to \infty \text{ in probability.}$$

PROOF. Note that

$$x\mathbb{P}(|X_1| > x) \le \mathbb{E}\left[|X_1| \mathbf{1}(|X_1| > x)\right] = \mathbb{E}[|X_1|] - \mathbb{E}\left[|X_1| \mathbf{1}(|X_1| \le x)\right] \to 0 \quad \text{as } n \to \infty,$$

where for the limit we have applied MCT (Theorem 1.3.19) for the increasing sequence of RVs $|X_1|\mathbf{1}(|X_1| \le x) \nearrow |X_1|$ and the fact that $\mathbb{E}[|X_1|] < \infty$. Hence (20) holds, and by Theorem 3.3.16, we have $\frac{S_n - n\mu_n}{n} \to 0$ in probability, where $\mu_n = \mathbb{E}[X_1 \mathbf{1}(|X_1| \le n)]$. Also note that $\mu_n \to \mu$ as $n \to \infty$ by DCT (Theorem 1.3.20), noting that $X_1 \mathbf{1}(|X_1| \le n) \to X_1$ a.s., $|X_1|\mathbf{1}(|X_1| \le n) \le |X_1|$ and $\mathbb{E}[|X_1|] < \infty$. Then $\frac{S_n}{n} \to \mu$ in probability since for each $\varepsilon > 0$, by taking n large enough so that $|\mu_n - \mu| < \varepsilon/2$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n}{n} - \mu_n\right| + |\mu_n - \mu| > \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n}{n} - \mu_n\right| > \varepsilon/2\right) \to 0 \quad \text{as } n \to \infty,$$

where the limit follows since $\frac{S_n - n\mu_n}{n} \to 0$ in probability. □

The key idea to extend WLLN to the "heavy-tail" case is to use truncation and then use Chebyshef. In order to prove Theorem 3.3.16, we prove a lemma that concerns WLLN for triangular array of RVs. (A special case is WLLN for a sequence of RVs as usual)

$$\begin{aligned} &X_{1;1} && \leftarrow \text{independent} \\ &X_{1;2},\, X_{2;2} && \leftarrow \text{independent} \\ &X_{1;3},\, X_{2;3},\, X_{3;3} && \leftarrow \text{independent} \\ &\vdots \end{aligned}$$

Namely, all RVs that appear in the above triangular array are independent, and for each $k \geq 1$, the k^{th} row consists of k RVs $X_{1;k}, \dots, X_{k;k}$.

Lemma 3.3.18 (WLLN for triangular array)**.** *For each $n \geq 1$, let $X_{n;1}, \dots, X_{n;n}$ be independent. Let $b_n > 0$ with $b_n \to \infty$ as $n \to \infty$. Let $\overline{X}_{n;k} := X_{n;k}\mathbf{1}(|X_{n;k}| \leq b_n)$. Suppose that*

(a) $\sum_{k=1}^{n} \mathbb{P}(|X_{n;k}| > b_n) = o(1)$*; and*

(b) $b_n^{-2} \sum_{k=1}^{n} \mathbb{E}\left[\overline{X}_{n;k}^2\right] = o(1)$.

Let $S_n := X_{n;1} + \cdots + X_{n;n}$ and $a_n := \sum_{k=1}^{n} \mathbb{E}\left[\overline{X}_{n;k}\right]$. Then

$$\frac{S_n - a_n}{b_n} \to 0 \quad \text{as } n \to \infty \text{ in probability.}$$

PROOF. Denote $\overline{S}_n := \overline{X}_{n;1} + \cdots + \overline{X}_{n;n}$. Fix $\varepsilon > 0$. We start by noting that

$$\mathbb{P}\left(\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right) \leq \mathbb{P}\left(\overline{S}_n \neq S_n\right) + \mathbb{P}\left(\left|\frac{\overline{S}_n - a_n}{b_n}\right| > \varepsilon\right).$$

In order to bound the first term, we use union bound and **(a)** to get

$$\mathbb{P}\left(\overline{S}_n \neq S_n\right) \leq \mathbb{P}\left(\bigcup_{1 \leq k \leq n} \{|X_{n;k}| > b_n\}\right) \leq \sum_{k=1}^{n} \mathbb{P}(|X_{n;k}| > b_n) = o(1).$$

In order to bound the second term, note that $\mathbb{E}[\overline{S}_n] = a_n$ and $|\overline{S}_n| \leq \sum_{k=1}^{n} b_k$, so by Chebyshef's inequality,

$$\mathbb{P}\left(\left|\frac{\overline{S}_n - a_n}{b_n}\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(\overline{S}_n)}{b_n^2 \varepsilon^2} = \frac{1}{b_n^2 \varepsilon^2} \sum_{k=1}^{n} \mathrm{Var}(\overline{X}_{n;k}) \leq \frac{1}{b_n^2 \varepsilon^2} \sum_{k=1}^{n} \mathbb{E}\left[\overline{X}_{n;k}^2\right] = o(1),$$

where the last estimate uses **(b)**. Then the conclusion follows. □

PROOF OF THEOREM 3.3.16. According to Lemma 3.3.18 with $b_n = n$, it suffices to verify conditions **(a)**-**(b)** in Lemma 3.3.18. For **(a)**, note that by (20),

$$\sum_{k=1}^{n} \mathbb{P}(|X_{n;k}| > b_n) = n\mathbb{P}(|X_n| > n) = o(1).$$

It remains to verify **(b)**. For this, we need to show that

$$n^{-1}\mathbb{E}\left[(X_1^2 \mathbf{1}(|X_1| \leq n))\right] = o(1).$$

To this end, let $X := X_1\mathbf{1}(|X_1| \le n)$. Use Exercise 1.5.9 and to write

$$\begin{aligned}\mathbb{E}[|X|^2] &= \int_0^\infty \mathbb{P}(X^2 \ge x)\,dx \\ &= \int_0^\infty \mathbb{P}(|X| \ge x^{1/2})\,dx \\ &= \int_0^\infty 2u\mathbb{P}(|X| \ge u)\,du \\ &= \int_0^n 2u\mathbb{P}(|X| \ge u)\,du \\ &\le \int_0^n 2u\mathbb{P}(|X_1| \ge u)\,du. \qquad (\because |X| \le |X_1|)\end{aligned}$$

Thus, by making a change of variable $u/n = t$ and denoting $g(u) = 2u\mathbb{P}(|X_1| \ge u)$,

$$n^{-1}\mathbb{E}\left[(X_1^2\mathbf{1}(|X_1| \le n))\right] = \frac{1}{n}\int_0^n g(u)\,du = \int_0^1 g(nt)\,du.$$

The last expression vanishes as $n \to \infty$ since $g(nt) \to 0$ as $n \to \infty$ for each $t > 0$. To give the details, note that since $0 \le g(u) \le 2u$ and $g(u) = o(1)$, g is bounded by some constant, say, $M > 0$. Fix $\varepsilon > 0$. Since $g(u) = o(1)$, there exists $N > 0$ such that $g(y) \le \varepsilon$ for all $y \le N$. Then for all $n \ge N/\varepsilon$,

$$0 \le \int_0^1 g(nt)\,du = \int_0^\varepsilon g(nt)\,du + \int_\varepsilon^1 g(nt)\,du \le \varepsilon M + (1-\varepsilon)\varepsilon.$$

Then letting $\varepsilon \searrow 0$ shows that $\int_0^1 g(nt)\,du = o(1)$, as desired. □

The following exercise demonstrates a direct route to prove Corollary 3.3.17 without using triangular arrays.

Exercise 3.3.19 (WLLN for RVs with finite mean)**.** In this exercise, we will prove the WLLN for RVs with infinite variance by using a 'truncation argument'. (Note that we cannot use Chebyshef here.)

Let $(X_k)_{k\ge1}$ be i.i.d. RVs such that $\mathbb{E}[|X_1|] < \infty$ and $\mathrm{Var}(X_1) \in [0,\infty]$. Let $\mu = \mathbb{E}[X_1]$ and $S_n = \sum_{k=1}^n X_i$, $n \ge 1$. We will show that for any positive constant $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0. \tag{21}$$

(i) Fix $M \ge 1$. Let $S_n^{\le M} := \sum_{i=1}^n X_i\mathbf{1}(|X_i| \le M)$ and $\mu^{\le M} := \mathbb{E}[X_1\mathbf{1}(|X_1| \le M)]$. Show that $n^{-1}S_n^{\le M} \to \mu^{\le M}$ as $n \to \infty$ in probability.

(ii) Show that $\mu^{\le M} = \mathbb{E}[X_1\mathbf{1}(|X_1| \le M)] \to \mu$ as $M \to \infty$. (Hint: Use dominated convergence theorem.)

(iii) Let $S_n^{>M} := \sum_{i=1}^n X_i\mathbf{1}(|X_i| > M)$. Use Markov's inequality and (ii) to show that, for any $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{S_n^{>M}}{n}\right| \ge \delta\right) \le \delta^{-1}\mathbb{E}\left[|X_1|\mathbf{1}(|X_1| > M)\right] \to 0 \quad \text{as } M \to \infty.$$

(iv) Fix $\varepsilon, \delta > 0$, and show the following inequality

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \ge \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n^{\le M}}{n} - \mu^{\le M}\right| \ge \varepsilon/3\right) + \mathbb{P}\left(\left|\frac{S_n^{>M}}{n}\right| \ge \varepsilon/3\right) + \mathbf{1}(|\mu^{\le M} - \mu| \ge \varepsilon/3).$$

Use (ii)-(iii) to deduce that there exists a large $M' \ge 1$ such that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \ge \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n^{\le M'}}{n} - \mu^{\le M'}\right| \ge \varepsilon/3\right) + \delta/2.$$

Finally, use (i) to show that there exists a large $n' \ge 1$ such that

$$\mathbb{P}\left(\left|\frac{S_{n'}}{n'} - \mu\right| \ge \varepsilon\right) \le \delta.$$

Conclude (21).

3.4. Borel-Cantelli Lemmas

In this section, we study almost sure convergence systematically.

Definition 3.4.1 (Almost sure convergence)**.** Let $X, (X_n)_{n\geq 1}$ be a RVs on the same probability space. We say that X_n converges to *a almost surely* (or *with probability* 1) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Definition 3.4.2 (Infinitely often)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(A_n)_{n\geq 1}$ be a sequence of events $A_n \in \mathscr{F}$. We say A_n occurs *infinitely often* (i.o. for short) if the following event occurs:

$$\{A_n \text{ i.o.}\} := \{\omega \in \Omega : \omega \in A_n \text{ i.o.}\} := \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\text{'s}\}.$$

Exercise 3.4.3. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(A_n)_{n\geq 1}$ be a sequence of events $A_n \in \mathscr{F}$. Recall that

$$\limsup_{n\to\infty} A_n := \lim_{m\to\infty} \bigcup_{n=m}^{\infty} A_n, \qquad \liminf_{n\to\infty} A_n := \lim_{m\to\infty} \bigcap_{n=m}^{\infty} A_n.$$

(i) Show that $\limsup_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}(\limsup_{n\to\infty} A_n)$ and $\liminf_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}(\liminf_{n\to\infty} A_n)$.

(ii) Show that $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) \geq \limsup_{n\to\infty} \mathbb{P}(A_n)$ and $\mathbb{P}\left(\liminf_{n\to\infty} A_n\right) \leq \liminf_{n\to\infty} \mathbb{P}(A_n)$.

(iii) Let X_n, $n \geq 1$ be RVs on $(\Omega, \mathscr{F}, \mathbb{P})$ and let $x \in \mathbb{R}$. Show that $X_n \to 0$ a.s. as $n \to \infty$ if and only if for all $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon \text{ i.o.}) = 0$.

Example 3.4.4 (a.s. convergence is strictly stronger than in probability convergence)**.** In this example, we will see that convergence in probability does not necessarily imply convergence with probability 1. Define a sequence of RVs $(T_n)_{n\geq 1}$ as follows. Let $T_1 = 1$, and $T_2 \sim \text{Uniform}(\{2,3\})$, $T_3 \sim \text{Uniform}(\{4,5,6\})$, and so on. In general, $X_k \sim \text{Uniform}\{(k-1)k/2, \cdots, k(k+1)/2\}$ for all $k \geq 2$. Let $X_n = \mathbf{1}(\text{some } T_k \text{ takes value } n)$. Think of

$$T_n = n\text{th arrival time of customers}$$
$$X_n = \mathbf{1}(\text{some customer arrives at time } n).$$

Then note that

$$\mathbb{P}(X_1 = 1) = 1,$$
$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_3 = 1) = 1/2,$$
$$\mathbb{P}(X_4 = 1) = \mathbb{P}(X_5 = 1) = \mathbb{P}(X_6 = 1) = 1/3,$$

and so on. Hence it is clear that $\lim_{n\to\infty} \mathbb{P}(X_n = 1) = 0$. Since X_n is an indicator variable, this yields that $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(X_n = 1) = 0$ for all $\varepsilon > 0$, that is, X_n converges to 0 in probability. On the other hand, $X_n \to 0$ a.s. means $\mathbb{P}(\lim_{n\to\infty} X_n = 0)$, which implies that $X_n = 0$ for sufficiently large $n \geq 1$. However, $X_n = 1$ for infinitely many n's since customer always arrive after any large time N. Hence X_n cannot converge to 0 almost surely. ▲

Exercise 3.4.5 (a.s. convergence implies in probability convergence)**.** Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let a be a real number. Suppose X_n converges to a with probability 1.

(i) Show that

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \leq \varepsilon\right) = 1 \quad \forall \varepsilon > 0.$$

(ii) Fix $\varepsilon > 0$. Let A_k be the event that $|X_n - a| \le \varepsilon$ for all $n \ge k$. Show that $A_1 \subseteq A_2 \subseteq \cdots$ and

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \le \varepsilon/2\right) \le \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right).$$

(iii) Justify the following: that for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - a| \le \varepsilon) \ge \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \ge \mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \le \varepsilon/2\right) = 1.$$

Conclude that $X_n \to a$ in probability.

A typical tool for proving convergence with probability 1 is the following.

Lemma 3.4.6 (Borel-Cantelli lemma)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(A_n)_{n\ge1}$ be a sequence of events $A_n \in \mathscr{F}$.*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty. \qquad \text{(i.e., } \mathbb{P}(A_n)\text{'s are summable)}$$

Then we have

$$\mathbb{P}(A_n \text{ not i.o.}) = \mathbb{P}\big(A_n \text{ occurs only for finitely many } n\text{'s}\big) = 1.$$

PROOF. Let $N = \sum_{n=1}^{\infty} \mathbf{1}(A_n)$, which is the number of n's such that A_n occurs. By Fubini's theorem (or MCT, Theorem 1.3.19),

$$\mathbb{E}[N] = \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}(A_n)\right] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbf{1}(A_n)] = \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty.$$

Since $N \ge 0$, it follows that $\mathbb{P}(N < \infty) = 1$ since otherwise $\mathbb{E}[N] = \infty$. Deduce that the RV N must not take ∞ with positive probability. Thus $\mathbb{P}(A_n \text{ not i.o.}) = \mathbb{P}(N < \infty) = 1$. □

Exercise 3.4.7 (BC Lemma and a.s. convergence)**.** Let $(X_n)_{n\ge1}$ be a sequence of RVs and fix $x \in \mathbb{R}$. We will show that $X_n \to x$ a.s. if the tail probabilities are 'summable'. (This is the typical application of the Borel-Cantelli lemma.)

(i) Fix $\varepsilon > 0$. Suppose $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$. Use Borel-Cantelli lemma to deduce that $|X_n - x| > \varepsilon$ for only finitely many n's.

(ii) Conclude that, if $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$ for all $\varepsilon > 0$, then $X_n \to x$ a.s.

Example 3.4.8. Let $(X_n)_{n\ge0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$. Recall that in Exercise 3.3.10, we have shown that

$$\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda\varepsilon n}.$$

for all $\varepsilon > 0$ and that $Y_n \to 0$ in probability as $n \to \infty$. In fact, $Y_n \to 0$ with probability 1. To see this, we note that, for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|Y_n - 0| > \varepsilon) = \sum_{n=1}^{\infty} e^{-\lambda\varepsilon n} = \frac{e^{-\lambda\varepsilon}}{1 - e^{\lambda\varepsilon}} < \infty.$$

By Borel-Cantelli lemma (or Exercise 3.4.7), we conclude that $Y_n \to 0$ a.s. ▲

Example 3.4.9. Let $(X_n)_{n\ge0}$ be a sequence of i.i.d. Uniform($[0,1]$) RVs.

(i) We show that $X_n^{1/n}$ converges to 1 almost surely, as $n \to \infty$. Fix any $\epsilon > 0$. Since $X_n \ge 1$,

$$\begin{aligned}\mathbb{P}(|(X_n)^{1/n} - 1| > \epsilon) &= \mathbb{P}\big((X_n)^{1/n} > (1+\epsilon) \text{ or } (X_n)^{1/n} < (1-\epsilon)\big) = \mathbb{P}((X_n)^{1/n} < (1-\epsilon)) \\ &= \mathbb{P}\big(X_n < (1-\epsilon)^n\big) = (1-\epsilon)^n,\end{aligned}$$

which goes to 0 as $n \to \infty$. Therefore, $X_n^{1/n}$ converges to 1 in probability. However, since $\sum_{n=1}^{\infty}(1-\epsilon)^n < \infty$, we have that $X_n^{1/n}$ also converges to 1 almost surely.

(ii) Define $U_n = \max\{X_1, X_2^2, X_3^3, \dots, X_{n-1}^{n-1}, X_n^n\}$. We show that the sequence U_n converges in probability to 1. For an $\epsilon < 1$ fixed

$$\begin{aligned}\mathbb{P}(|U_n - 1| \geq \epsilon) = \mathbb{P}(U_n \leq 1-\epsilon) &= \mathbb{P}(X_1 \leq 1-\epsilon, X_2^2 \leq 1-\epsilon, \dots, X_n^n \leq 1-\epsilon) \\ &= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2^2 \leq 1-\epsilon)\cdots\mathbb{P}(X_n^n \leq 1-\epsilon) \\ &= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2 \leq (1-\epsilon)^{1/2})\cdots\mathbb{P}(X_n \leq (1-\epsilon)^{1/n}) \\ &= (1-\epsilon)\cdot(1-\epsilon)^{1/2}\cdots(1-\epsilon)^{1/n} = (1-\epsilon)^{1+1/2+\cdots+1/n} \to 0,\end{aligned}$$

since $1 + 1/2 + \cdots + 1/n \to \infty$, as $n \to \infty$.

(iii) Define $V_n = \max\{X_1, X_2^{2^2}, X_3^{3^2}, \dots, X_{n-1}^{(n-1)^2}, X_n^{n^2}\}$. Does the sequence V_n converge in probability to 1? Similarly as in the previous part, for a fixed $\epsilon < 1$ we have

$$\mathbb{P}(|U_n - 1| \geq \epsilon) = (1-\epsilon)^{1+1/2^2+\cdots+1/n^2},$$

which doesn't converge to zero (but to a positive number), since $1 + 1/2^2 + \cdots + 1/n^2$ is a convergent series. Therefore, V_n doesn't converge to zero in probability. Hence it also doesn't converge to 0 a.s..

▲

The following proposition is often used to 'upgrade' in-probability convergence to a.s. convergence.

Proposition 3.4.10 (Subsequential a.s. conv. = in prob. conv.)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of RVs and fix $x \in \mathbb{R}$. Then $X_n \to x$ as $n \to \infty$ in probability if and only if for every subsequence $X_{n(k)}$, $k \geq 1$, there is a further subsequence $X_{n(k(m))}$, $m \geq 1$ such that $X_{n(k(m))} \to x$ a.s. as $m \to \infty$.*

PROOF. Suppose $X_n \to x$ in probability. Fix $\varepsilon > 0$. Then for each $n \geq 1$, there exists $n(k) \geq 1$ such that $\mathbb{P}(|X_{n(k)} - x| > \varepsilon) \leq 2^{-k}$ for $k \geq 1$. Then $\sum_{k\geq 1}\mathbb{P}(|X_{n(k)} - x| > \varepsilon) \leq \sum_{k\geq 1} 2^{-k} = 1 < \infty$, so by Borel-Cantelli Lemma 3.4.6 (or Exercise 3.4.7), it holds that $|X_{n(k)} - x| > \varepsilon$ for only finitely many k's with probability one. This means $X_{n(k)} \to x$ a.s. as $k \to \infty$.

Conversely, suppose that for each subsequence $X_{n(k)}$, $k \geq 1$ of $(X_n)_{n\geq 1}$, there is a further subsequence $X_{n(k(m))}$, $m \geq 1$ such that $X_{n(k(m))} \to x$ a.s. as $m \to \infty$. Fix $\varepsilon > 0$. Let $y_n := \mathbb{P}(|X_n - x| > \varepsilon)$. We wish to show that $y_n \to 0$ as $n \to \infty$. Since a.s. convergence implies in probability convergence, the hypothesis implies that for each subsequence $y_{n(k)}$ of y_n, there exists a further subsequence $n(k(m))$ such that $y_{n(k(m))} \to 0$. This implies that y_n cannot have a limit point other than zero. If $y_n \not\to 0$, then there is an open neighborhood U of 0 and a subsequence $y_{n(k)}$ such that $y_{n(k)} \notin U$ for all $k \geq 1$. But then no subsequence $y_{n(k(m))}$ of $y_{n(k)}$ can converge to 0 since they are all outside of U. Hence y_n must converge to zero. □

Example 3.4.11 (Fatou's lemma with in-probability convergence)**.** Let $X_n \geq 0$ and $X_n \to X$ in probability. Then the conclusion of Fatou's lemma (Thm. 1.3.18) still holds despite the in-probability convergence:

$$\liminf_{n\to\infty} \mathbb{E}[X_n] \geq \mathbb{E}[X].$$

To see this, let $n(k)$ be a subsequence such that $\mathbb{E}[X_{n(k)}] \to \liminf_{n\to\infty} \mathbb{E}[X_n]$ as $k \to \infty$. By Prop. 3.4.10, there exists a further subsequence $n(k(m))$, $m \geq 1$ such that $X_{n(k(m))} \to X$ almost surely. Then we apply Fatou's lemma along this sub-subsequence to get

$$\liminf_{n\to\infty} \mathbb{E}[X_n] = \liminf_{k\to\infty} \mathbb{E}[X_{n(k)}] = \liminf_{m\to\infty} \mathbb{E}[X_{n(k(m))}] \geq \mathbb{E}[X].$$

▲

Proposition 3.4.12 (In prob. conv. implies weak conv.)**.** *Let $X, (X_n)_{n\geq 1}$ be a RVs on the same probability space such that $X_n \to X$ in probability (i.e., $X_n - X \to 0$ in prob.). If $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $f(X_n) \to f(X)$ in probability. Furthermore, if f is continuous and bounded, then $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$.*

PROOF. Fix a subsequence $n(k)$, $k \geq 1$. By Proposition 3.4.10, there exists a further subsequence $n(k(m))$, $m \geq 1$ such that $X_{n(k(m))} \to X$ a.s..Then $f(X_{n(k(m))}) \to f(X)$ a.s.. [1] By Proposition 3.4.10, it follows that $f(X_n) \to f(X)$ in probability.

Further assume that f is bounded. Recall that Then $f(X_{n(k(m))}) \to f(X)$ a.s.. Then by BCT (Theorem 1.3.16), $\mathbb{E}[f(X_{n(k(m))})] \to \mathbb{E}[f(X)]$ as $m \to \infty$. Since every subsequential limit of $\mathbb{E}[f(X_n)]$ equals $\mathbb{E}[f(X)]$, it follows that $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$. □

Now we prove a special case of the strong law of large numbers. The proof of full statement (Theorem 3.1.2) with finite second moment assumption has extra technicality, so here we prove the result under a stronger assumption of finite fourth moment.

Theorem 3.4.13 (SLLN with fourth moment)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. RVs such that $\mathbb{E}[X_n^4] < \infty$. Let $S_n = X_1 + \cdots + X_n$ for all $n \geq 1$. Then S_n/n converges to $\mathbb{E}[X_1]$ with probability 1.*

PROOF. Our aim is to show that

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] < \infty.$$

Then by Borel-Cantelli lemma, $(S_n/n)^4$ converges to 0 with probability 1. Hence S_n/n converges to 0 with probability 1, as desired.

For a preparation, we first verify that we have finite first and second moments for X_1. It is easy to verify the inequality $|x| \leq 1 + x^4$ for all $x \in \mathbb{R}$, so we have

$$\mathbb{E}[|X_1|] \leq 1 + \mathbb{E}[X_1^4] < \infty.$$

Hence $\mathbb{E}[X_1]$ exists. By shifting, we may assume that $\mathbb{E}[X_1] = 0$. Similarly, it holds that $x^2 \leq c + x^4$ for all $x \in \mathbb{R}$ if $c > 0$ is large enough. Hence $\mathbb{E}[X_1^2] < \infty$.

Note that

$$\mathbb{E}\left[S_n^4\right] = \mathbb{E}\left[\left(\sum_{k=1}^{n} X_k\right)^4\right] = \mathbb{E}\left[\sum_{1\leq i,j,k,\ell\leq n} X_i X_j X_k X_\ell\right] = \sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right].$$

Note that by independence and the assumption that $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_i X_j X_k X_\ell] = 0$ if at least one of the four indices does not repeat. For instance,

$$\mathbb{E}[X_1 X_2^3] = \mathbb{E}[X_1]\mathbb{E}[X_2^3] = 0,$$
$$\mathbb{E}[X_1 X_2^2 X_3] = \mathbb{E}[X_1]\mathbb{E}[X_2^2]\mathbb{E}[X_3] = 0.$$

Hence by collecting terms based on number of overlaps, we have

$$\begin{aligned}\sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right] &= \sum_{i=1}^{n} \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1\leq i<j\leq n} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] \\ &= n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2.\end{aligned}$$

Thus for all $n \geq 1$,

$$\mathbb{E}\left[(S_n/n)^4\right] = \frac{n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2}{n^4} \leq \frac{n^2\mathbb{E}[X_1^4] + 3n^2\mathbb{E}[X_1^2]^2}{n^4} = \frac{\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]}{n^2}.$$

Summing over all n, this gives

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] \leq (\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]) \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Hence by Borell-Cantelli lemma, we conclude that $(S_n/n)^4$ converges to 0 with probability 1. The same conclusion holds for S_n/n. This shows the assertion. □

[1] To see this, assume $Y_n \to Y$ a.s., fix $\omega \in \Omega$ s.t. $Y_n(\omega) \to Y(\omega)$. Then by continuity of f, $f(Y_n(\omega)) \to f(Y(\omega))$.

The following example shows that the converse of the Borel-Cantelli lemma is false.

Example 3.4.14 (The converse of BC lemma is false)**.** Let $([0,1],\mathscr{B},\mu)$ be the probability space on the unit interval $[0,1]$ with $\mathscr{B}$ =Borel σ-algebra on $[0,1]$ and μ =Lebesgue measure on $[0,1]$. Fix a sequence $(a_n)_{n\geq 1}$ and let $A_n := (0,a_n)$. Suppose $a_n = o(1)$. Then

$$\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n = \emptyset.$$

However, if $a_n \geq 1/n$, then

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

Hence in this case, the probabilities of A_n are not summable but still A_n does not occur infinitely often. ▲

Lemma 3.4.15 (The second Borel-Cantelli lemma)**.** *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and let $(A_n)_{n\geq 1}$ be independent events in $\mathscr{F}$. Then we have the following implication:*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \qquad \Longrightarrow \qquad \mathbb{P}(A_n \text{ occurs i.o.}) = 1.$$

PROOF. Let $N = \sum_{n=1}^{\infty} \mathbf{1}(A_n)$, which is the number of n's such that A_n occurs. Fix an integer $M \geq 1$. Then by using independence and $1-x \leq \exp(-x)$ for all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\sup_{M\leq n\leq N} \mathbf{1}_{A_n} = 0\right) = \mathbb{P}\left(\bigcap_{n=M}^{N} A_n^c\right) = \prod_{n=M}^{N} (1-\mathbb{P}(A_n)) \leq \prod_{n=M}^{N} \exp(-\mathbb{P}(A_n)) = \exp\left(-\sum_{n=M}^{N} \mathbb{P}(A_n)\right) \overset{N\to\infty}{\longrightarrow} 0.$$

It follows that $\mathbb{P}\left(\sup_{n\geq M} \mathbf{1}_{A_n} = 0\right) = 0$ for all $M \geq 1$.Noting that $\sup_{n\geq M} \mathbf{1}_{A_n}$ takes values from $\{0,1\}$, we have $\sup_{n\geq M} \mathbf{1}_{A_n} = 1$ a.s.. Hence

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{E}[\mathbf{1}(A_n \text{ i.o.})] = \mathbb{E}\left[\limsup_{n\to\infty} \mathbf{1}_{A_n}\right] = \mathbb{E}\left[\lim_{M\to\infty} \sup_{n\geq M} \mathbf{1}_{A_n}\right] = \mathbb{E}\left[\lim_{M\to\infty} 1\right] = 1.$$

□

The following is a typical application of the second BC Lemma.

Proposition 3.4.16 (SLLN does not hold without first moment)**.** *Let $(X_n)_{n\geq 1}$ be i.i.d. RVs on the same probability space with $\mathbb{E}[|X_n|] = \infty$. Then $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$. Furthermore, for $S_n = X_1 + \cdots + X_n$,*

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} \text{ exists in } (-\infty,\infty)\right) = 0.$$

PROOF. By the tail-sum formula (Prop. 1.5.10) and since $x \mapsto \mathbb{P}(|X_1| > x)$ is non-increasing,

$$\infty = \mathbb{E}[|X_1|] = \int_0^{\infty} \mathbb{P}(|X_1| > x)\, dx = \int_0^{\infty} \mathbb{P}(|X_n| > x)\, dx \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq n).$$

Since the events $|X_n| \geq n$ are independent, by the second BC Lemma (see Lemma 3.4.15), $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$. To show the second part, write

$$A := \left\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} \text{ exists in } (-\infty,\infty)\right\}, \quad B := \{\omega : |X_n(\omega)| \geq n \text{ i.o.}\}.$$

Since we have shown that $\mathbb{P}(B) = 1$, $\mathbb{P}(A \setminus B) = 0$ since $0 \leq \mathbb{P}(A \setminus B) \leq \mathbb{P}(B^c) = 0$. Hence

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

We would like to show the RHS equals zero. For this, it suffices to show that $A \cap B = \emptyset$. To this end, suppose there exists $\omega \in A \cap B$. Write

$$\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}.$$

One the one hand, since $\omega \in A$, the LHS must converge to zero. On the other hand, since $\omega \in A$, we have $\frac{S_n(\omega)}{n(n+1)} \to 0$. Since $\omega \in B$ at the same time, $\frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1} < -1/2$ i.o.. This is a contradiciton. Thus $A \cap B = \emptyset$, as desired. □

From Proposition 3.4.16, we now know that $\mathbb{E}[|X_1|] < \infty$ is necessary order for SLLN to hold. In fact, this condition is also sufficient for SLLN to hold, as we will prove later.

It will be useful for later examples to strengthen the second BC Lemma as the following version:

Lemma 3.4.17 (A quantitative version of the 2nd BC Lemma)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \dots$ be pairwise independent events. Suppose that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Then*

$$\frac{\#\text{ of occurance of } A_1, \dots, A_n}{\text{expected } \#\text{ of occurance of } A_1, \dots, A_n} = \frac{\sum_{m=1}^{n} \mathbf{1}(A_n)}{\sum_{m=1}^{n} \mathbb{P}(A_n)} \to 1 \qquad \text{as } n \to \infty.$$

PROOF. Let $X_m := \mathbf{1}(A_m)$ and $S_n = X_1 + \cdots + X_n$. We wish to show that $S_n/\mathbb{E}[S_n] \to 1$ a.s.. We first show that the convergence holds in probability. First, since the events $A_1, A_2, \dots$ are pairwise independent, so are the indicators $\mathbf{1}(A_1), \mathbf{1}(A_2), \dots$, so

$$\begin{aligned} \operatorname{Var}(S_n) &= \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n) \\ &\le \mathbb{E}[X_1^2] + \cdots + \mathbb{E}[X_n^2] \\ &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n], \end{aligned}$$

where the last equality uses the fact that X_m's are 0-1 RVs. Now by Chebyshef's inequality and using the hypothesis that $\mathbb{E}[S_n] \to \infty$, for each $\delta > 0$,

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| > \delta \mathbb{E}[S_n]\right) \le \frac{\operatorname{Var}(S_n)}{\delta^2 \mathbb{E}[S_n]^2} \le \frac{1}{\delta^2 \mathbb{E}[S_n]} \to 0 \quad \text{as } n \to \infty.$$

This shows that $S_n/\mathbb{E}[S_n] \to 1$ in probability.

In order to upgrade the above convergence to almost sure convergence, we use a subsequence and monotonicity argument. Denote $W_n := S_n/\mathbb{E}[S_n]$. We wish to show that $W_n \to 1$ a.s.. Recall that we have shown the following: For all $\delta > 0$,

$$\mathbb{P}(|W_n - 1| > \delta) \le \frac{1}{\delta^2 \mathbb{E}[S_n]} \quad \text{as } n \to \infty.$$

Let $n(k) := \inf\{m \ge 1 \,|\, \mathbb{E}[S_m] \ge k^2\}$. Then

$$\sum_{k=1}^{\infty} \mathbb{P}(|W_{n(k)} - 1| > \delta) \le \sum_{k=1}^{\infty} \frac{1}{\delta^2 \mathbb{E}[S_{n(k)}]} \le \sum_{k=1}^{\infty} \frac{1}{\delta^2 k^2} < \infty. \tag{22}$$

This holds for all $\delta > 0$, so by the BC Lemma (see Lemma 3.4.6) that $W_{n(k)} \to 1$ a.s. as $k \to \infty$.

Now we have shown the desired a.s. convergence along the subsequence $n(k)$. It remains to 'interpolate' between these indices using the fact that S_n is non-decreasing. Namely, for each $m \ge 1$ and choose $k = k(m) \ge 1$ so that $n(k) \le m < n(k+1)$. Then

$$S_{n(k)} \le S_m \le S_{n(k+1)}.$$

Dividing using the inequality $\mathbb{E}[S_{n(k)}] \le \mathbb{E}[S_m] \le \mathbb{E}[S_{n(k+1)}]$,

$$\frac{\mathbb{E}[S_{n(k)}]}{\mathbb{E}[S_{n(k+1)}]} \frac{S_{n(k)}}{\mathbb{E}[S_{n(k)}]} \le \frac{S_m}{\mathbb{E}[S_m]} \le \frac{S_{n(k+1)}}{\mathbb{E}[S_{n(k+1)}]} \frac{\mathbb{E}[S_{n(k+1)}]}{\mathbb{E}[S_{n(k)}]}.$$

Recall that by the choice of $n(k)$, we have $k^2 \le \mathbb{E}[S_{n(k)}] < (k+1)^2 \le \mathbb{E}[S_{n(k+1)}] < (k+2)^2$. Hence

$$\frac{k^2}{(k+2)^2} \le \frac{\mathbb{E}[S_{n(k+1)}]}{\mathbb{E}[S_{n(k)}]} \le \frac{(k+2)^2}{k^2}.$$

This shows that $\mathbb{E}[S_{n(k+1)}]/\mathbb{E}[S_{n(k)}] \to 1$ as $k \to \infty$. Since $W_{n(k)} = S_{n(k)}/\mathbb{E}[S_{n(k)}] \to 1$ a.s., (22) implies that $W_m = S_m/\mathbb{E}[S_m] \to 1$ as $n \to \infty$. This completes the proof. □

The key point of the proof above is the following: In order to show $X_n/c_n \to 1$ a.s., where $c_n \to \infty$ and $X_n \geq 0$, $X_n \nearrow$, it suffices to show that $X_{n(k)}/c_{n(k)} \to 1$ and $c_{n(k+1)}/c_{n(k)} \to 1$ a.s. for some subsequence $n(k) \to \infty$.

Example 3.4.18 (Record values)**.** Suppose $X_1, X_2, \ldots$ are i.i.d. RVs on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Let $A_n := \{X_n > \sup_{1\leq k<n} X_k\}$ denote the event that X_n exceeds all preceeding RVs, i.e., a record is made at the nth draw. Let $R_n := \sum_{k=1}^n \mathbf{1}(A_k)$ denote the number of records up to time n. We will show that

$$\frac{R_n}{\log n} \to 1 \quad \text{a.s. as } n \to \infty.$$

In order to show this, we claim the following:

$(*)$ $A_1, A_2, \ldots$ are mutually independent and $\mathbb{P}(A_n) = 1/n$.

Once we have the above claim, noting that $\mathbb{E}[R_n] = \sum_{k=1}^n n^{-1} \sim \log n$, we can deduce $R_n/\log n \to 1$ a.s. by Lemma 3.4.17.

It remains to justify the claim (*) above. To do so, fix $n \geq 1$ and we consider the order statistics $Y_{1;n} \geq Y_{2;n} \geq \cdots \geq Y_{n;n}$ of $X_1, \ldots, X_n$. For each sample $\omega \in \Omega$, there exists a permutation $\sigma(\omega)$ on $\{1, \ldots, n\}$ such that $Y_{i;n}(\omega) = X_{\sigma(\omega)(i)}(\omega)$. Omitting dependence on ω, we obtain a random permutation σ on $\{1, \ldots, n\}$. Since $X_1, \ldots, X_n$ are i.i.d., the distribution of the order statistics is invariant under permuting the order of $X_1, \ldots, X_n$. It follows that σ is uniformly distributed among all possible $n!$ permuations on $\{1, \ldots, n\}$. From here, we deduce

$$\mathbb{P}(A_n) = \mathbb{P}(\sigma(n) = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

To show mutual independence, notice that, for $1 \leq m < n$,

$$\begin{aligned}
\mathbb{P}(A_m \cap A_n) &= \mathbb{P}(A_m \cap \{X_n > X_{n-1}, \ldots, X_1\}) \\
&= \mathbb{P}(\{\sigma(m) < \sigma(1), \ldots, \sigma(m-1)\} \cap \{\sigma(n) = 1\}) \\
&= \binom{n-1}{m} \frac{(m-1)!(n-1-m)!}{n!} = \frac{1}{m}\frac{1}{n} = \mathbb{P}(A_m)\,\mathbb{P}(A_n).
\end{aligned}$$

This justifies the claim $(*)$ above. ▲

Example 3.4.19 (Head runs)**.** Consider the bi-infinite sequence of RVs $(X_n)_{n\geq\mathbb{Z}}$, where $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = 1/2$ for each $n \in \mathbb{Z}$. This gives a random bi-infinite sequnece of $+1$'s and -1's. In this example, we are interested in the statistics of runs of $+1$'s (i.e., a consecutive block of $+1$'s). Namey, let $\ell_n = \max\{m : X_{n-m+1} = \cdots = X_n = 1\}$ be the length of run of $+1$'s at time n. Let $L_n := \max_{1\leq m\leq n} \ell_m$ denote the maximum length of runs of $+1$'s among times 1 through n. We claim that

$$\frac{L_n}{\log n} \to 1 \quad \text{a.s. as } n \to \infty.$$

To justify the above, first note that for any $x \geq 0$,

$$\mathbb{P}(\ell_n \geq x) = \mathbb{P}(\ell_n \geq \lceil x \rceil) = \mathbb{P}(X_{n-\lceil x\rceil+1} = \cdots = X_n = 1) = 2^{-\lceil x\rceil} \leq 2^{-x}.$$

Hence for each $\varepsilon > 0$,

$$\sum_{n\geq 1} \mathbb{P}(\ell_n \geq (1+\varepsilon)\log_2 n) \leq \sum_{n\geq 1} n^{-(1+\varepsilon)} < \infty.$$

By the BC Lemma (see Lemma 3.4.6), it follows that there exists $N_\varepsilon > 0$ such that $\ell_n \leq (1+\varepsilon)\log_2 n$ for all $n \geq N_\varepsilon$ almost surely. Then for all $n \geq N_\varepsilon$, since $x \mapsto \log_2 x$ is increasing,

$$\max(\ell_{N_\varepsilon}, \ldots, \ell_n) \leq (1+\varepsilon)\max(\log_2 N_\varepsilon, \ldots, \log_2 n) \leq (1+\varepsilon)\log_2 n.$$

Then almost surely,

$$\limsup_{n\to\infty} \frac{L_n}{\log_2 n} \le \limsup_{n\to\infty} \frac{\max(\ell_1,\dots,\ell_{N_\varepsilon})}{\log_2 n} + (1+\varepsilon) = 1+\varepsilon.$$

Here, we have used the fact that, for any almost surely finite RV X, $|X|/\log n \to 0$ a.s.. Since $\varepsilon > 0$ was arbitrary, this shows $\limsup_{n\to\infty} \frac{L_n}{\log_2 n} \le 1$ a.s..

For the other direction, we fix $n \ge 1$ and partition $\{1,\dots,n\}$ into disjoint blocks of size $R := [1 + (1-\varepsilon)\log_2 n]$. Namely, Consider the interval partition $(0,R],(R,2R]$, and so on. We consider the indicators $\mathbf{1}(\ell_{kR} < R)$, which are determined by sets of non-overlapping RVs $\{X_{(k-1)R+1},\dots,X_{kR}\}$. Hence these indicator RVs are independent. Furthermore, $\mathbb{P}(\ell_{kR} < R) = 1 - 2^{-R} = 1 - n^{-1+\varepsilon}$. Then

$$\mathbb{P}(L_n \le (1-\varepsilon)\log_2 n) \le \mathbb{P}(\ell_R, \ell_{2R},\dots,\ell_{kR} \le (1-\varepsilon)\log_2 n) = \left(1 - \frac{1}{n^{1-\varepsilon}}\right)^{n/R} \le \exp\left(-\frac{n^{-\varepsilon}}{2\log_2 n}\right).$$

Since the last term is summable, by BC Lemma, $L_n > (1-\varepsilon)\log_2 n$ for all sufficiently large n almost surely. This yields $\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \le 1-\varepsilon$ a.s.. Then letting $\varepsilon \searrow 0$ shows the claim. ▲

3.5. Strong Law of Large Numbers

In this section, we finally establish the strong law of large numbers only assuming the exisistence of first moment of increments. Recall that this was a necessary condition for SLLN to hold, see Proposition 3.4.16.

Theorem 3.5.1 (SLLN with first moment)**.** *Let $(X_n)_{n\ge1}$ be a sequence of i.i.d. RVs such that $\mathbb{E}[|X_n|] < \infty$. Let $S_n = X_1 + \cdots + X_n$ for all $n \ge 1$. Then S_n/n converges to $\mathbb{E}[X_1]$ with probability 1.*

PROOF. (This proof is based on Etemadi's original proof (1981) in [Ete81].) Since $\mathbb{E}[|X_1|] < \infty$, we have $\mu := \mathbb{E}[X_1] \in (-\infty,\infty)$. We first note that since $X_n = X_n^+ - X_n^-$ and the sequences of RVs $(X_n^+)_{n\ge1}$ and $(X_n^-)_{n\ge1}$ also satisfy the hypothesis, by linearity, it suffices to show the assertion for these two sequences of nonnegative RVs. Without loss of generality, we may simply assume that $X_n \ge 0$. Now we show the assertion in three steps.

(Truncation) Let $\overline{X}_k := X_k \mathbf{1}(|X_k| \le k)$ for $k \ge 1$. Denote $\overline{S}_k := \overline{X}_1 + \cdots + \overline{X}_n$. We claim that it is enough to show that

$$\frac{\overline{S}_n}{n} \to \mu \quad \text{a.s. as } n \to \infty. \tag{23}$$

To see this, let $N := \sum_{k=1}^\infty \mathbf{1}(X_k \ne \overline{X}_k)$. Note that by using MCT (or Fubini's theorem),

$$\begin{aligned}\mathbb{E}[N] = \mathbb{E}\left[\sum_{k=1}^\infty \mathbf{1}(X_k \ne \overline{X}_k)\right] = \sum_{k=1}^\infty \mathbb{E}\left[\mathbf{1}(X_k \ne \overline{X}_k)\right] &= \sum_{k=1}^n \mathbb{P}(|X_k| \ge k) \\ &\le \int_0^\infty \mathbb{P}(|X_k| > t)\,dt \\ &\le \int_0^\infty \mathbb{P}(|X_1| > t)\,dt = \mathbb{E}[|X_1|] < \infty.\end{aligned}$$

Since $N \ge 0$, it follows that $N < \infty$ almost surely. Fix $\omega \in \Omega$ such that $N(\omega) < \infty$. Then

$$\left|n^{-1}(S_n(\omega) - \overline{S}_n(\omega))\right| \le n^{-1} \sum_{k=1}^{N(\omega)} \left|X_k(\omega) - \overline{X}_k(\omega)\right| \to 0 \quad \text{as } n \to \infty.$$

This shows that

$$\mathbb{P}\left(\lim_{n\to\infty} \left|n^{-1}(S_n - \overline{S}_n)\right| = 0\right) \ge \mathbb{P}(N < \infty) = 1.$$

Thus $n^{-1}\overline{S}_n \to \mu$ a.s. implies $n^{-1}S_n \to \mu$ a.s..

(a.s. convergence along a subsequence) Fix $\alpha > 1$. We use an exponential subsequence $n(k) := [\alpha^k]$, $k \geq 1$. We will show (23) along the subsequence $n(k)$, $k \geq 1$. We start as usual by estimating the tail probabilities using Chebyshef: For each $\varepsilon > 0$,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\overline{S}_{n(k)} - \mathbb{E}\left[\overline{S}_{n(k)}\right]\right| > \varepsilon\, n(k)\right) \leq \varepsilon^{-2} \sum_{k=1}^{\infty} \frac{\mathrm{Var}(\overline{S}_{n(k)})}{n(k)^2} = \sum_{k=1}^{\infty} \sum_{m=1}^{n(k)} \frac{\mathrm{Var}(\overline{X}_m)}{\varepsilon^2 n(k)^2}$$
$$= \varepsilon^{-2} \sum_{m=1}^{\infty} \mathrm{Var}(\overline{X}_m) \sum_{k:n(k)\geq m} n(k)^{-2},$$

where the last equality uses Fubini's theorem for interchanging double sum of nonnegative terms.

Now noting that $n(k) = [\alpha^k] \geq \alpha^k/2$ for $k \geq 1$,

$$\sum_{k:n(k)\geq m} n(k)^{-2} \leq 4 \sum_{k:\alpha^k \geq m} \alpha^{-2k} \leq \frac{4}{1-\alpha^{-2}} m^{-2},$$

where we have used the fact that α^{-2k} for $k \geq 1$ s.t. $\alpha^k \geq m$ is a geometric sequence with ratio α^{-2} with first term at most m^{-2}. This and Lemma 3.5.2 give

$$\sum_{m=1}^{\infty} \mathrm{Var}(\overline{X}_m) \sum_{k:n(k)\geq m} n(k)^{-2} \leq \frac{4}{1-\alpha^{-2}} \sum_{m=1}^{\infty} \frac{\mathrm{Var}(\overline{X}_m)}{m^2} \leq \frac{16\mathbb{E}[|X_1|]}{1-\alpha^{-2}} < \infty.$$

Combining with the first displayed inequality in this case, we have shown that

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\overline{S}_{n(k)} - \mathbb{E}\left[\overline{S}_{n(k)}\right]\right| > \varepsilon\, n(k)\right) < \infty.$$

Since this holds for all $\varepsilon > 0$, by the Borel-Cantelli lemma, it follows that

$$\frac{\overline{S}_{n(k)} - \mathbb{E}\left[\overline{S}_{n(k)}\right]}{n(k)} \to 0 \quad \text{a.s. as } k \to \infty.$$

To complete the proof of the current case, it remains to show

$$n(k)^{-1}\mathbb{E}\left[\overline{S}_{n(k)}\right] \to \mu \quad \text{as } k \to \infty.$$

Note that by DCT, we have that $\mathbb{E}[\overline{X}_n] \to \mathbb{E}[X_1] = \mu$ as $n \to \infty$. Hence the above is the mean of quantities that each converge to μ, so it should also converge to μ. To do a more rigorous justification, note that $\mathbb{E}[|X_k|\mathbf{1}(|X_k| > k)] = \mathbb{E}[|X_k|] - \mathbb{E}[|X_k|\mathbf{1}(|X_k| \leq k)] \to 0$ by MCT. Fix $\varepsilon > 0$. Then there exists $N = N(\varepsilon) \geq 1$ such that $\mathbb{E}[|X_k|\mathbf{1}(|X_k| > k)] \leq \varepsilon$. Then by Jensen's inequality,

$$\left|\mathbb{E}[S_n] - \mathbb{E}[\overline{S_n}]\right| \leq \left|\sum_{k=1}^{n} \mathbb{E}[X_k \mathbf{1}(|X_k| > k)]\right| \leq \sum_{k=1}^{n} \mathbb{E}[|X_k|\mathbf{1}(|X_k| > k)] \leq n\varepsilon + \sum_{k=1}^{N(\varepsilon)} \mathbb{E}[|X_k|].$$

Dividing both sides by n and letting $n \to \infty$ shows $\limsup_{n\to\infty} n^{-1}|\mathbb{E}[S_n] - \mathbb{E}[\overline{S_n}]| \leq \varepsilon$. Since $\varepsilon > 0$ was arbitrary, this shows $\limsup_{n\to\infty} n^{-1}|\mathbb{E}[S_n] - \mathbb{E}[\overline{S_n}]| = 0$, which is enough to conclude.

(Interpolation) For each $m \geq 1$, let $k = k(m)$ be the unique integer such that $n(k) \leq m < n(k+1)$. Since we are assuming $X_n \geq 0$ for all $n \geq 1$, we have

$$S_{n(k)} \leq S_m \leq S_{n(k+1)}.$$

This yields

$$\frac{n(k)}{n(k+1)} \frac{S_{n(k)}}{n(k)} \leq \frac{S_m}{m} \leq \frac{S_{n(k+1)}}{n(k+1)} \frac{n(k+1)}{n(k)}.$$

Recall that by the choice of $n(k)$, we have $n(k+1)/n(k) \to \alpha$ as $k \to \infty$. Using the previous part, we conclude, almost surely,

$$\alpha^{-1}\mu \le \liminf_{m\to\infty} \frac{S_m}{m} \le \limsup_{m\to\infty} \frac{S_m}{m} \le \alpha\mu.$$

Since $\alpha > 1$ was arbitrary, letting $\alpha \searrow 1$ then shows the assertion.

□

The following lemma was used in the proof of SLLN above.

Lemma 3.5.2 (An estimate used in the proof of SLLN)**.** *Let $(X_n)_{n\ge1}$ be a sequence of i.i.d. RVs such that $\mathbb{E}[|X_n|] < \infty$. Let $\overline{X}_n := X_n \mathbf{1}(|X_n| \le n)$. Then*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(\overline{X}_n)}{n^2} \le 4\mathbb{E}[|X_1|] < \infty.$$

PROOF. We first show two claims. Note that for any RV Y with $\mathbb{E}[Y^2] < \infty$,

$$\mathbb{E}[Y^2] = \int_0^\infty \mathbb{P}(Y^2 > y)\,dy = \int_0^\infty \mathbb{P}(|Y| > \sqrt{y})\,dy = \int_0^\infty 2t\mathbb{P}(|Y| > t)\,dt,$$

where we made a change of variable $\sqrt{y} = t$. Moreover, for each $y \ge 0$, we will show that

$$y\sum_{n=1}^{\infty} n^{-2}\mathbf{1}(y \le n) \le 2.$$

Indeed, note that for $y > 1$,

$$y\sum_{n=1}^{\infty} n^{-2}\mathbf{1}(y \le n) = y\sum_{n=1}^{\infty} n^{-2}\mathbf{1}(\lceil y\rceil \le n) = y\int_{\lceil y\rceil-1}^{\infty} x^{-2}\,dx = \frac{y}{\lceil y\rceil - 1} \le 2.$$

For $0 \le y < 1$, we have

$$y\sum_{n=1}^{\infty} n^{-2}\mathbf{1}(y \le n) = \sum_{n=1}^{\infty} n^{-2} = 1 + \sum_{n=2}^{\infty} n^{-2} \le 1 + \int_1^\infty x^{-2}\,dx = 2.$$

This shows the two claims above.

Now observe that, since $\mathbb{P}(|\overline{X}_n| > n) = 0$ and $|\overline{X}_n| \le |X_n|$,

$$\mathrm{Var}(\overline{X}_n) \le \mathbb{E}\left[\overline{X}_n^2\right] = \int_0^\infty 2y\mathbb{P}(|\overline{X}_n| > y)\,dy \le \int_0^\infty \mathbf{1}(y \le n)2y\mathbb{P}(|X_1| > y)\,dy,$$

where we have also used the fact that $\mathbb{P}(|X_1| > y) = \mathbb{P}(|X_n| > y)$. Now using Fubini's theorem,

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{\mathrm{Var}(\overline{X}_n)}{n^2} &\le \sum_{n=1}^{\infty}\int_0^n n^{-2}\mathbf{1}(y \le n)2y\mathbb{P}(|X_1| > y)\,dy \\ &= \int_0^\infty 2\,\mathbb{P}(|X_1| > y)\left(y\sum_{n=1}^{\infty} n^{-2}\mathbf{1}(y \le n)\right)dy \\ &\le 4\int_0^\infty \mathbb{P}(|X_1| > y) = 4\mathbb{E}[|X_1|] < \infty.\end{aligned}$$

□

Next, we establish a 'uniform' version of SLLN.

Definition 3.5.3 (Empirical distribution function)**.** Let $X_1, X_2, \dots$ be a sequence of RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. For each $n \ge 1$, define the *empirical distribution function* F_n by

$$F_n(x) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}(X_k \le x).$$

That is, $F_n(x)$ is the frequency of X_k's among the first n RVs that are at most x.

Now suppose $X_1, X_2, \dots$ are i.i.d. with distribution function F. For each fixed $x \in \mathbb{R}$, the RVs $Y_k := \mathbf{1}(X_k \le x)$ are i.i.d., so by SLLN, as $n \to \infty$, we must have

$$F_n(x) = n^{-1} \sum_{k=1}^{n} Y_k \xrightarrow{a.s.} \mathbb{E}[Y_1] = \mathbb{P}(X_1 \le x) = F(x).$$

This shows $F_n \to F$ a.s. pointwise. In fact, a stronger statement holds. That is, the worst-case error $\sup_x |F_n(x) - F(x)|$ also converges to zero almost surely. This is the content of the following Glivenko-Cantelli theorem.

Theorem 3.5.4 (Glivenko-Cantelli theorem)**.** *Let $X_1, X_2, \dots$ be i.i.d. RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0 \right) = 1.$$

PROOF. Fix $m > 0$. For each $j = 1, \dots, m-1$, define $x_{m,j} := \inf\{y : F(y) \ge j/m\}$. Since F is non-decreasing, $x_{m,1} \le x_{m,2} \le \cdots \le x_{m,m-1}$. Denote $x_{0,m} = -\infty$ and $x_{m,m} = \infty$. Then[2]

$$F(x_{m,j}-) - F(x_{m,j-1}) \le j/m - (j-1)/m = m^{-1}.$$

Fix $x \in \mathbb{R}$. Then there exists $k \in \{1, \dots, m-1\}$ such that $x_{m,k-1} \le x < x_{m,k}$. Then

$$\begin{aligned} F_n(x) - F(x) &= \underbrace{F_n(x) - F_n(x_{m,k}-)}_{\le 0} + F_n(x_{m,k}-) - F(x_{m,k}-) + F(x_{m,k}-) - F(x) \\ &\le F_n(x_{m,k}-) - F(x_{m,k}-) + F(x_{m,k}-) - F(x_{m,k-1}) \\ &\le F_n(x_{m,k}-) - F(x_{m,k}-) + m^{-1} \\ &\le m^{-1} + \sum_{k=1}^{m} |F_n(x_{m,k}-) - F(x_{m,k}-)|. \end{aligned}$$

Similarly,

$$\begin{aligned} F_n(x) - F(x) &= \underbrace{F_n(x) - F_n(x_{m,k-1})}_{\ge 0} + F_n(x_{m,k-1}) - F(x_{m,k-1}) + F(x_{m,k-1}) - F(x) \\ &\ge F_n(x_{m,k-1}) - F(x_{m,k-1}) + F(x_{m,k-1}) - F(x_{m,k}-) \\ &\ge F_n(x_{m,k-1}) - F(x_{m,k-1}) - m^{-1} \\ &\ge -m^{-1} - \sum_{k=1}^{m} |F_n(x_{m,k}) - F(x_{m,k})|. \end{aligned}$$

Thus we have

$$|F_n(x) - F(x)| \le m^{-1} + \sum_{k=1}^{m} |F_n(x_{m,k}) - F(x_{m,k})| + |F_n(x_{m,k}-) - F(x_{m,k}-)|.$$

The RHS above does not depend on x so we may take the supremum over x on the LHS. Also, by SLLN (see Theorem 3.5.1), for each fixed $y \in \mathbb{R}$, $F_n(y) \to F(y)$ almost surely as $n \to \infty$. Furthermore, $F_n(y-) = n^{-1} \sum_{k=1}^{n} \mathbb{P}(X_k < y) \to \mathbb{P}(X_1 < y) = F(y-)$ almost surely as $n \to \infty$. Hence

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le m^{-1} \quad \text{a.s..}$$

Than taking $m \to \infty$ shows the assertion. □

[2] If F is continuous, then $F(x_{m,j}) = j/m$ and $F(x_{m,j}) - F(x_{m,j-1}) = m^{-1}$ However, F may have jumps in general so $F(x_{m,j}) - F(x_{m,j-1})$ could be as large as 1.

Example 3.5.5 (Shannon's theorem)**.** Let $X_1, X_2, \dots$ be i.i.d. RVs on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ taking values from $\mathscr{A} := \{1, \dots, r\}$ with PMF $p(k) := \mathbb{P}(X_1 = k) > 0$ for all $k \in \{1, \dots, r\}$. Think of $\mathscr{A}$ as the set of alphabet and the sequence of RVs $X_1, \dots, X_n$ as a random string of length n with these alphabet. For each $n \geq 1$, define a RV Σ_n as

$$\Sigma_n := p(X_1)\, p(X_2) \dots p(X_n).$$

Then by SLLN, almost surely as $n \to \infty$,

$$-n^{-1} \log \Sigma_n = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \longrightarrow \mathbb{E}\big[\log p(X_i)\big] = -\sum_{k=1}^{r} p(k) \log p(k) =: \mathbf{H}.$$

The quantity $\mathbf{H}$ is called the *entropy* of the PMF p, and it measures how random it is. The above result implies that $-n^{-1} \log \Sigma_n \to \mathbf{H}$ in probability, so for each $\varepsilon > 0$,

$$\mathbb{P}\big(\exp(-n(\mathbf{H}+\varepsilon) \leq \Sigma_n \leq \exp(-n(\mathbf{H}-\varepsilon)\big) = \mathbb{P}\big(\big|\mathbf{H} + n^{-1} \log \Sigma_n\big| \leq \varepsilon\big) \to 1 \quad \text{as } n \to \infty.$$

This is called the *asymptotic equipartition property*. ▲

3.6. Renewal processes and Renewal SLLN

An *arrival* process is a sequence of strictly increasing RVs $0 < T_1 < T_2 < \cdots$ with the convention of setting $T_0 = 0$. For each integer $k \geq 1$, its kth *inter-arrival time* is defined by $\tau_k = T_k - T_{k-1}$. For a given arrival process $(T_k)_{k\geq 1}$, the associated *counting process* $(N(t))_{t\geq 0}$ is defined by

$$N(t) = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t) = \#(\text{arrivals up to time } t).$$

Note that these three processes (arrival times, inter-arrival times, and counting) determine each other:

$$(T_k)_{k\geq 1} \Longleftrightarrow (\tau_k)_{k\geq 1} \Longleftrightarrow (N(t))_{t\geq 0}.$$

Exercise 3.6.1. Let $(T_k)_{k\geq 1}$ be any arrival process and let $(N(t))_{t\geq 0}$ be its associated counting process. Show that these two processes determine each other by the following relation

$$\{T_n \leq t\} = \{N(t) \geq n\}.$$

In words, nth customer arrives by time t if and only if at least n customers arrive up to time t.

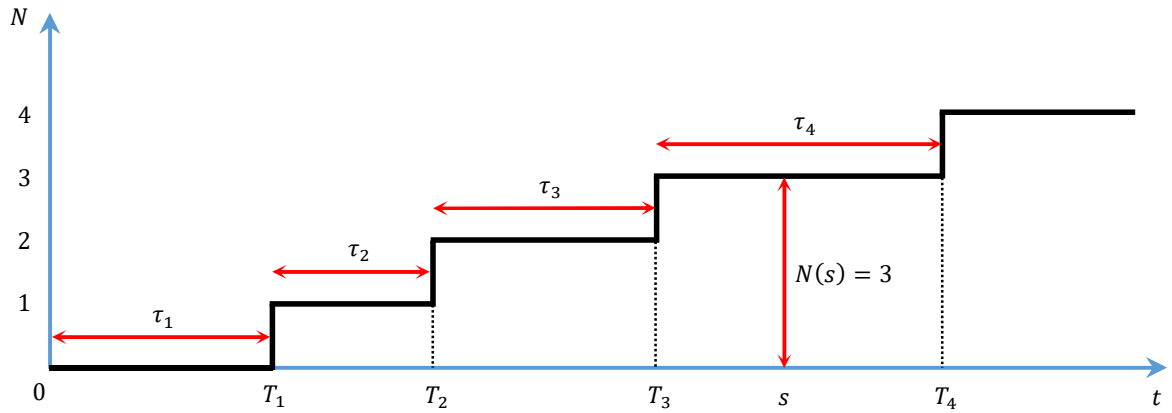


FIGURE 3.6.1. Illustration of a continuous-time arrival process $(T_k)_{k\geq 1}$ and its associated counting process $(N(t))_{t\geq 0}$. τ_k's denote inter-arrival times. $N(t) \equiv 3$ for $T_3 < t \leq T_4$.

Exercise 3.6.2 (Well-definedness of the counting process)**.** Let $(N(t))_{t\geq 0}$ denote the counting process of a renewal process with inter-arrival times $(\tau_k)_{k\geq 1}$ such that there exists $\varepsilon > 0$ for which $\tau_k > \varepsilon$ infinitely often (but not necessarily independent or identically distributed). Recall that $N(t)$ and τ_k's are associated as

$$N(t) = \sum_{n=1}^{\infty} \mathbf{1}(\tau_1 + \cdots + \tau_n \leq t).$$

(i) Show that $T_n \geq \sum_{k=1}^{n} \varepsilon \mathbf{1}(\tau_k \geq \varepsilon)$ for $n \geq 1$.

(ii) Fix $t \geq 0$. Justify the following steps:

$$\begin{aligned}\mathbb{P}(N(t) = \infty) = \mathbb{P}\left(\sup_{n\geq 1} T_n < t\right) \leq \mathbb{P}\left(\sum_{k=1}^{\infty} \varepsilon \mathbf{1}(\tau_k \geq \varepsilon) < t\right) &= \mathbb{P}\left(\sum_{k=1}^{\infty} \mathbf{1}(\tau_k \geq \varepsilon) < t/\varepsilon\right) \\ &= \mathbb{P}\left(\exists N \geq 1 \text{ s.t. } \tau_k < \varepsilon \text{ for all } k \geq N\right) \\ &\leq \sum_{N=1}^{\infty} \mathbb{P}\left(\tau_k < \varepsilon \text{ for all } k \geq N\right) = 0.\end{aligned}$$

Deduce that $\mathbb{P}(N(t) < \infty) = 1$ for all $t \geq 0$.

Definition 3.6.3 (Renewal process)**.** A counting process $(N(t))_{t\geq 0}$ is called a *renewal process* if its inter-arrival times $\tau_1, \tau_2, \cdots$ are i.i.d. with $\mathbb{E}[\tau_1] < \infty$.

A cornerstone in the theory of renewal processes is the following strong law of large numbers for renewal processes.

Theorem 3.6.4 (Renewal SLLN)**.** *Let $(T_k)_{k\geq 0}$ be a renewal process and let $(\tau_k)_{k\geq 0}$ and $(N(t))_{t\geq 0}$ be the associated inter-arrival times and counting process, respectively. Let $\mathbb{E}[\tau_k] = \mu$ be the mean inter-arrival time. If $0 < \mu < \infty$, then*

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mu}\right) = 1.$$

PROOF. First, write $T_k = \tau_1 + \tau_2 + \cdots + \tau_k$. Since the inter-arrival times are i.i.d. with mean $\mu < \infty$, the strong law of large numbers imply

$$\mathbb{P}\left(\lim_{k\to\infty} \frac{T_k}{k} = \mu\right) = 1. \tag{24}$$

Next, fix $t \geq 0$ and let $N(t) = n$, so that there are total n arrivals up to time t. Then the nth arrival time T_n must occur by time t, whereas the $n+1$st arrival time T_{n+1} must occur after time t. Hence $T_n \leq t < T_{n+1}$. In general, we have

$$T_{N(t)} \leq t < T_{N(t)+1}.$$

Dividing by $N(t)$, we get

$$\frac{T_{N(t)}}{N(t)} \leq \frac{t}{N(t)} < \frac{T_{N(t)+1}}{N(t)+1} \frac{N(t)+1}{N(t)}. \tag{25}$$

To take the limit as $t \to \infty$, we note that $\mathbb{P}(T_k < \infty) = 1$ for all k since $\mathbb{P}(T_k = \infty) \leq \sum_{i=1}^{k} \mathbb{P}(\tau_i = \infty) = 0$ (since $\mathbb{P}(\tau_i = \infty) > 0$ implies $\mathbb{E}[\tau_i] = \infty$, which is a contradiction). This yields $N(t) \geq k$ for some large enough t. Since k was arbitrary, this yields $N(t) \nearrow \infty$ as $t \to \infty$ with probability 1. Therefore, according to (24) and SLLN (see Theorem 3.5.1), we get

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{T_{N(t)}}{N(t)} = \mu\right) = \mathbb{P}\left(\lim_{t\to\infty} \frac{T_{N(t)+1}}{N(t)+1} = \mu\right) = \mathbb{P}\left(\lim_{t\to\infty} \frac{N(t)+1}{N(t)} = 1\right) = 1.$$

Hence (25) gives

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{t}{N(t)} = \mu\right) = 1.$$

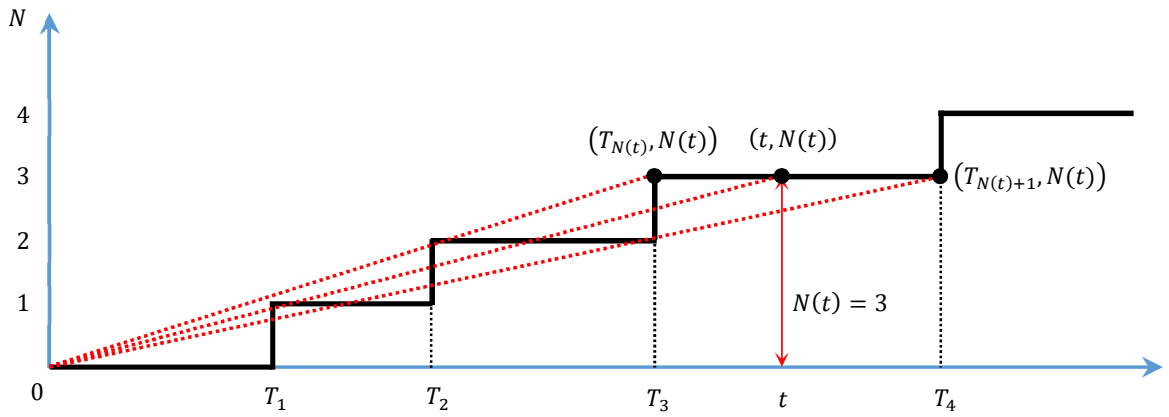


FIGURE 3.6.2. Illustration of the inequalities (25).

Since $\mu > 0$, we can take the reciprocal inside the above probability. This shows the assertion. □

Example 3.6.5 (Poisson process)**.** Suppose $(N(t))_{t\geq 0}$ is a renewal process with its inter-arrival times $(\tau_k)_{k\geq 1}$ as $\mathrm{Exp}(\lambda)$ distribution with some $\lambda > 0$. In this case, we call $(N(t))_{t\geq 0}$ a "Poisson process with arrival rate λ". Note that the mean inter-arrival time is $\mathbb{E}[\tau_1] = 1/\lambda$, so the renewal SLLN yields

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{N(t)}{t} = \lambda\right) = 1.$$

Namely, with probability 1, we tend to see about λt arrivals during $[0, t]$ as $t \to \infty$. In other words, we tend to see λ arrivals during an interval of unit length. Hence it makes sense to call the parameter λ as the 'arrival rate' of the process. ▲

Next, we consider a renewal process together with rewards, which are given for each inter-arrival times. This simple extension of the renewal processes will greatly improve the applicability of our theory.

Let $(T_k)_{k\geq 0}$ be a renewal process and let $(\tau_k)_{k\geq 0}$ and $(N(t))_{t\geq 0}$ be the associated inter-arrival times and counting process, respectively. Suppose we have a sequence of *rewards* $(Y_k)_{k\geq 1}$, so we regard a reward of Y_k is given at the kth arrival time T_k. We define the *reward process* $(R(t))_{t\geq 0}$ associated with the sequence $(\tau_k, Y_k)_{k\geq 1}$ as

$$R(t) = \sum_{k=1}^{N(t)} Y_k.$$

Namely, upon the kth arrival at time $T_k = \tau_1 + \cdots + \tau_k$, we receive a reward of Y_k. Then $R(t)$ is the total reward up to time t.

As we looked at the average number of arrivals $N(t)/t$ as $t \to \infty$, a natural quantity to look at for the reward process is the 'average reward' $R(t)/t$ as $t \to \infty$. Intuitively, since everything refreshes upon new arrivals, we should expect

$$\frac{R(t)}{t} \to \frac{\text{expected reward during one 'cycle'}}{\text{expected duration of one 'cycle'}}$$

as $t \to \infty$ almost surely. This is made precise by the following result.

Theorem 3.6.6 (Renewal reward SLLN)**.** *Let $(R(t))_{t\geq 0}$ be the reward process associated with an i.i.d. sequence of inter-arrival times $(\tau_k)_{k\geq 1}$ and an i.i.d. sequence of rewards $(Y_k)_{k\geq 1}$. Suppose $\mathbb{E}[Y_k] < \infty$ and $\mathbb{E}[\tau_k] \in (0, \infty)$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{R(t)}{t} = \frac{\mathbb{E}[Y_1]}{\mathbb{E}[\tau_1]}\right) = 1.$$

PROOF. Let $(\tau_k)_{k\geq 0}$ denote the inter-arrival times for the renewal process $(T_k)_{k\geq 0}$. Note that

$$\frac{R(t)}{t} = \left(\frac{1}{N(t)}\sum_{k=1}^{N(t)} Y_k\right)\frac{N(t)}{t}.$$

Hence by SLLN, the 'average reward' up to time t in the bracket converges to $\mathbb{E}[Y_1]$ almost surely. Moreover, the average number of arrivals $N(t)/t$ converges to $1/\mathbb{E}[\tau_1]$ by Theorem 3.6.4. Hence the assertion follows. □

Remark 3.6.7. Theorem 3.6.4 can be obtained as a special case of the above reward version of SLLN, simply by choosing $Y_k = \tau_k$ for $k \geq 1$ so that $R(t) = N(t)$.

Example 3.6.8 (Long run car costs)**.** This example is excerpted from [DD99]. Mr. White do not drive the same car more than t^* years, where $t^* > 0$ is some fixed number. He changes to a new car when the old one breaks down or reaches t^* years. Let X_k be the life time of the kth car that Mr. White drives, which are i.i.d. with finite expectation. Let τ_k be the duration of his kth car. According to his policy, we have

$$\tau_k = \min(X_k, t^*).$$

Let $T_k = \tau_1 + \cdots + \tau_k$ be the time that Mr. White is done with the kth car. Then $(T_k)_{k\geq 0}$ is a renewal process. Note that the expected running time for the kth car is

$$\begin{aligned}\mathbb{E}[\tau_k] &= \mathbb{E}[\tau_k \,|\, X_k < t^*]\mathbb{P}(X_k < t^*) + \mathbb{E}[\tau_k \,|\, X_k \geq t^*]\mathbb{P}(X_k \geq t^*)\\ &= \mathbb{E}[X_k \,|\, X_k < t^*]\mathbb{P}(X_k < t^*) + t^*\mathbb{P}(X_k \geq t^*).\end{aligned}$$

Suppose that the car cost g during each cycle is given by

$$g(t) = \begin{cases} A+B & \text{if } t < t^* \\ A & \text{if } t \geq t^*. \end{cases}$$

Namely, if the car breaks down by t^* years, then Mr. White has to pay $A+B$ dolors; otherwise, the cost is only A dolors. Then the expected cost for one cycle is

$$\mathbb{E}[g(\tau_k)] = A + B\mathbb{P}(\tau_k < t^*) = A + B\mathbb{P}(X_k < t^*).$$

Thus by Theorem 3.6.6, the long-run car cost of Mr. White is

$$\lim_{t\to\infty}\frac{R(t)}{t} = \frac{\mathbb{E}[g(\tau_k)]}{\mathbb{E}[\tau_k]} = \frac{A + B\mathbb{P}(X_k < t^*).}{\mathbb{E}[X_k \,|\, X_k < t^*]\mathbb{P}(X_k < t^*) + t^*\mathbb{P}(X_k \geq t^*)}.$$

For more concrete example, let $X_k \sim \text{Uniform}([0,10])$ and let $A = 10$ and $B = 3$. Then

$$\mathbb{E}[g(\tau_k)] = 10 + 3t^*/10.$$

On other other hand,

$$\begin{aligned}\mathbb{E}[\tau_k] &= \mathbb{E}[X_k \,|\, X_k < t^*]\mathbb{P}(X_k < t^*) + t^*\mathbb{P}(X_k \geq t^*)\\ &= \frac{t^*}{2}\frac{t^*}{10} + t^*\frac{10-t^*}{10} = t^* - (t^*)^2/20.\end{aligned}$$

Note that for $\mathbb{E}[X_k \,|\, X_k < t^*] = t^*/2$, observe that a uniform RV over $[0,10]$ conditioned on being $[0,t^*]$ is uniformly distributed over $[0,t^*]$. This yields

$$\frac{\mathbb{E}[g(\tau_k)]}{\mathbb{E}[\tau_k]} = \frac{10 + 0.3t^*}{t^*(1 - t^*/20)}.$$

Lastly, in order to minimize the above long-run cost, we differentiate it in t^* and find global minimum. A straightforward computation shows that the long-run cost is minimized at

$$t^* = \frac{-1+\sqrt{1.6}}{0.03} \approx 8.83.$$

Thus the optimal strategy for Mr. White in this situation is to drive each car up to 8.83 years. ▲

3.7. Convergence of random series

So far, our treatment of the laws of large numbers takes the perspective of analyzing the ample mean S_n/n. In this section, we introduce another perspective of analyzing the partial sums, or random series, S_n, directly. It has an important advantage to yield a near-optimal rate of convergence for the laws of large numbers.

Our starting point is the Kolmogorov's maximal inequality, which is demonstrated in the following exercise.

Exercise 3.7.1 (Kolmogorov's maximal inequality)**.** Let $X_1, X_2, \cdots$ be independent RVs with $\mathbb{E}[X_i] = 0$. Denote $S_n = X_1 + \cdots + X_n$ and $S_0 = 0$. In this exercise, we will show that

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| \ge t\right) \le t^{-2}\mathrm{Var}(S_n). \tag{26}$$

(i) Let $\tau = \inf\{k \ge 0 : |S_k| \ge t\}$ denote the first time that $|S_k|$ exceeds t. Show that

$$\mathbb{E}[S_n^2] \ge \sum_{k=1}^{n} \mathbb{E}[S_n^2 \mathbf{1}(\tau = k)].$$

(ii) For each $1 \le k \le n$, note that $S_k\mathbf{1}(\tau = k)$ and $S_n - S_k = X_{k+1} + \cdots + X_n$ are independent. Deduce that

$$\mathbb{E}[S_k\mathbf{1}(\tau = k)(S_n - S_k)] = 0.$$

(iii) For each $1 \le k \le n$, write $S_n^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2$. Using (ii), show that

$$\begin{aligned}\mathbb{E}[S_n^2\mathbf{1}(\tau = k)] &\ge \mathbb{E}[(S_k^2 + 2S_k(S_n - S_k))\mathbf{1}(\tau = k)] \\ &= \mathbb{E}[(S_k^2\mathbf{1}(\tau = k)] + 2\mathbb{E}[S_k\mathbf{1}(\tau = k)(S_n - S_k)] \\ &= \mathbb{E}[(S_k^2\mathbf{1}(\tau = k)] \ge t^2\mathbb{E}[\mathbf{1}(\tau = k)].\end{aligned}$$

(iv) From (i)-(iii), deduce that

$$\mathbb{E}[S_n^2] \ge t^2 \sum_{k=1}^{n} \mathbb{E}[\mathbf{1}(\tau = k)] = t^2\mathbb{E}\left[\sum_{k=1}^{n} \mathbf{1}(\tau = k)\right] = t^2\mathbb{P}(\tau \le n) = t^2\mathbb{P}\left(\max_{1\le k\le n} |S_k| \ge t\right).$$

Conclude Kolmogorov's maximal inequality (26).

An immediate application of Kolmogorov's maximal inequality is the following lemma.

Lemma 3.7.2 (Summability of variance implies convergence of sums)**.** *Let $X_1, X_2, \ldots$ be independent RVs on the same probability space with zero mean. Then the following implication holds:*

$$\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty \quad \Longrightarrow \quad \sum_{n=1}^{\infty} X_n \text{ converges almost surely.}$$

PROOF. Denote $S_n = X_1 + \cdots + X_n$. Fix integers $M > N \ge 1$ and $\varepsilon > 0$. By Exercise 3.7.1,

$$\mathbb{P}\left(\max_{M\le m\le N} |S_m - S_M| > \varepsilon\right) \le \varepsilon^{-2}\mathrm{Var}(S_N - S_M) = \varepsilon^{-2} \sum_{m=M+1}^{N} \mathrm{Var}(X_m).$$

Taking $N \to \infty$ and using continuity of measure and the hypothesis,

$$\mathbb{P}\left(\max_{M\le m} |S_m - S_M| > \varepsilon\right) \le \varepsilon^{-2} \sum_{m=M+1}^{\infty} \mathrm{Var}(X_m) \to 0 \quad \text{as } M \to \infty.$$

By using triangle inequality,

$$\begin{aligned}\mathbb{P}\left(\max_{m,n\ge M} |S_m - S_n| > 2\varepsilon\right) &\le \mathbb{P}\left(\max_{m\ge M} |S_m - S_M| + \max_{n\ge M} |S_n - S_M| > 2\varepsilon\right) \\ &= \mathbb{P}\left(\max_{m\ge M} |S_m - S_M| > \varepsilon\right) \to 0 \quad \text{as } M \to \infty.\end{aligned}$$

Denote $W_M := \max_{m,n\geq M} |S_m - S_n|$. Then W_M is non-increasing in M. So $\{\sup_{n\geq M} W_n > \varepsilon\}$ is a decreasing event. Hence by continuity of probability measure,

$$\mathbb{P}\left(\limsup_{M\to\infty} W_M > 2\varepsilon\right) = \mathbb{P}\left(\lim_{M\to\infty} \sup_{n\geq M} W_n > 2\varepsilon\right) = \lim_{M\to\infty} \mathbb{P}\left(\sup_{n\geq M} W_n > 2\varepsilon\right)$$
$$= \lim_{M\to\infty} \mathbb{P}(W_M > 2\varepsilon) = 0.$$

Since $\varepsilon > 0$ was arbitrary, this shows $\limsup_{M\to\infty} W_M = 0$ almost surely. Hence $W_M \to 0$ a.s. as $M \to \infty$. Therefore, the random sequence S_n, $n \geq 1$, forms a Cauchy sequence almost surely. Hence S_n converges almost surely. □

Lemma 3.7.3 (Kolomogorov's three-series theorem)**.** *Let $X_1, X_2, \ldots$ be independent RVs on the same probability space. Fix $A > 0$ and let $Y_n := X_n \mathbf{1}(|X_n| < A)$. Then $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if*

(i) $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$*; and*

(ii) $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$ *converges a.s..; and*

(iii) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$.

PROOF. We first show the "if" part. Lemma 3.7.2 and **(iii)** shows that $\sum_{n=1}^{\infty}(Y_n - \mathbb{E}[Y_n])$ converges almost surely. Using **(ii)**, it follows that $\sum_{n=1}^{\infty} Y_n$ converges almost surely. Let $N = \sum_{n=1}^{\infty} \mathbf{1}(X_n \neq Y_n)$. Then by MCT or Fubini and using **(i)**,

$$\mathbb{E}[N] = \sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty.$$

Since $N \geq 0$, it follows that $N < \infty$ a.s.. Hence

$$\mathbb{P}\left(\sum_{n=1}^{\infty} X_n \text{ does not converge}\right) = \mathbb{P}\left(\left\{\sum_{n=1}^{\infty} X_n \text{ does not converge}\right\} \cap \left\{\sum_{n=1}^{\infty} Y_n \text{ converges}\right\}\right)$$
$$\leq \mathbb{P}(N = \infty) = 0.$$

Thus $\sum_{n=1}^{\infty} X_n$ converges a.s..

Next, we show the "only if " part. For this direction, we will crucially use Lindeberg-Feller CLT (see Thm. 4.4.8) from later sections. Suppose $\sum_{n=1}^{\infty} X_n$ converges a.s.. We wish to show that all three conditions in the statement must hold.

First suppose **(i)** is violoated, that is, $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) = \infty$. By independence of the RVs X_n, the second BC lemma (see Lem. 3.4.15) implies that $|X_n| > A$ i.o. with probability one, so $\sum_{k=1}^{\infty} X_n$ does not exists with probability one (i.e., $\mathbb{P}(\liminf_{n\to\infty} \sum_{k=1}^{n} X_k \neq \limsup_{n\to\infty} \sum_{k=1}^{n} X_k) = 1$).

Second, suppose **(i)** holds but **(iii)** is violated. Take

$$c_n := \sum_{k=1}^{n} \mathrm{Var}(Y_k), \qquad X_{n;m} := \frac{Y_m - \mathbb{E}[Y_m]}{\sqrt{c_n}}.$$

Then note that $\mathbb{E}[X_{n;m}] \equiv 0$ and $\mathrm{Var}(X_{n;m}) \equiv 1$. Also, since $|Y_n| < A$, $\mathbb{E}[|Y_n|] < A$ so $|X_{n;m}| < 2A/\sqrt{c_n}$ a.s.. Hence for each fixed $\varepsilon > 0$, since $c_n \to \infty$ as $n \to \infty$, $2A/\sqrt{c_n} < \varepsilon$ for n large enough. Hence $|X_{n;m}| < \varepsilon$ for n large enough, so

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{E}[X_{n;m}^2 \mathbf{1}(|X_{n;m}| > \varepsilon)] = 0.$$

By Theorem 4.4.8, it follows that $S_n := X_{n;1} + \cdots + X_{n;n} \Rightarrow Z \sim N(0,1)$. Furthermore, that $\sum_{k=1}^{\infty} X_n$ converges a.s. and **(i)** imply $\sum_{k=1}^{\infty} Y_n$ converges a.s. by using a similar argument as in the proof of the "if" part. Consequently, $T_n := \sum_{k=1}^{n} Y_n/\sqrt{c_n} \Rightarrow 0$ (note that $\mathbb{P}(|T_n| > \varepsilon) = \mathbb{P}(|\sum_{k=1}^{n} Y_n| > \varepsilon\sqrt{c_n}) \to 0$). By Slutsky's theorem (see Exc. 4.2.20), $S_n - T_n \Rightarrow Z \sim N(0,1)$. Thus

$$\sum_{k=1}^{n} \mathbb{E}[Y_n]/\sqrt{c_n} = (S_n - T_n) \Rightarrow Z \sim N(0,1),$$

but this is a contradiction since the LHS above is a sequence of real numbers so it cannot converge weakly to a normal RV.

Lastly, suppose **(i)** and **(iii)** hold. Then by Lemma 3.7.2, $\sum_{m=1}^{n} Y_n - \mathbb{E}[Y_n]$ converges a.s.. We have seen that $\sum_{k=1}^{\infty} X_n$ converges a.s. and **(i)** imply $\sum_{k=1}^{\infty} Y_n$ converges a.s., so in our case $\sum_{k=1}^{\infty} Y_n$ converges a.s.. Now $\sum_{m=1}^{n} \mathbb{E}[Y_m] = \left(\sum_{m=1}^{n} Y_m\right) - \left(\sum_{m=1}^{n} Y_m - \mathbb{E}[Y_m]\right)$ is the difference of two a.s. converging series, so it converges a.s.. This shows **(ii)** holds. □

Exercise 3.7.4. Let $X_1, X_2, \ldots$ be independent RVs on the same probability space such that $\mathbb{P}(X_n = n^{-\alpha}) = \mathbb{P}(X_n = -n^{-\alpha}) = 1/2$ for $n \geq 1$. Show that $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if $\alpha > 1/2$. (Note that $\sum_{n=1}^{\infty} |X_n|$ converges if and only if $\alpha > 1$ as $|X_n| = n^{\alpha}$ a.s.. Due to the cancelation, convergence of $\sum_{n=1}^{\infty} X_n$ requires slower decay rate.)

Exercise 3.7.5 (Kronecker's lemma)**.** If $a_n \nearrow \infty$ and $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$ converges, then show that $\frac{1}{a_n} \sum_{n=1}^{\infty} x_n \to 0$.

Theorem 3.7.6 (SLLN with a rate of convergence)**.** *Let $X_1, X_2, \ldots$ be independent RVs on the same probability space. Suppose that $\mathbb{E}[X_n] = 0$ and $\mathbb{E}[X_n^2] = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for any $\varepsilon > 0$,*

$$\frac{S_n}{\sqrt{n}(\log n)^{(1/2)+\varepsilon}} \to 0 \quad \text{a.s. as } n \to \infty.$$

PROOF. Let $a_n := \sqrt{n}(\log n)^{(1/2)+\varepsilon}$ for $n \geq 2$ and set $a_1 = 1$. Then

$$\sum_{n=1}^{\infty} \mathrm{Var}(X_n/a_n) = \sigma^2 \left(1 + \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{1+2\varepsilon}}\right) < \infty.$$

By Lemma 3.7.2, it follows that $\sum_{n=1}^{\infty} X_n/a_n$ converges almost surely. Then by Kroneker's lemma (see Exercise 3.7.5), $a_n^{-1} S_n \to 0$ almost surely, as desired. □

Remark 3.7.7. The correct rate of convergence for SLLN is given by the law of the iterated logarithm:

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}\sqrt{\log\log n}} \to \frac{\sigma}{\sqrt{2}} \quad \text{a.s. as } n \to \infty.$$

The rate given by Theorem 3.7.6 is not too far from the above optimal rate.

In this section, we have studied necessary and sufficient conditions for the condition

$$\mathbb{P}\left(\sum_{n=1}^{\infty} X_n \text{ converges}\right) = 1$$

to hold. But what if we only want some positive probability that the random series converges? Is there some weaker condition than those stated in Theorem 3.7.3 that imply

$$\mathbb{P}\left(\sum_{n=1}^{\infty} X_n \text{ converges}\right) > 0?$$

The answer is no. Surprisingly, the only possible value for $\mathbb{P}\left(\sum_{n=1}^{\infty} X_n \text{ converges}\right)$ is 0 or 1, so if it is positive, it has to be one. This type of result is known as '0-1' law. See Exercise 3.7.8 for details.

Exercise 3.7.8 (Kolmogorov's 0-1 law)**.** Let $X_1, X_2, \ldots$ be RVs on the same probability space. For each $n \geq 1$, let $\mathcal{F}_n := \sigma(X_n, X_{n+1}, \ldots)$ and let $\mathcal{T} := \bigcap_{n=1}^{\infty} \mathcal{F}_n$. Here $\mathcal{T}$ is called the *tail σ-algebra* for the sequence of RVs $(X_n)_{n\geq 1}$[3] If $A \in \mathcal{T}$, then A is called a *tail event*.

Let $X_1, X_2, \ldots$ be independent RVs on the same probability space and let $\mathcal{T}$ denote the corresponding tail σ-algebra. Let $A \in \mathcal{T}$ be any tail event. In this exercise, we will show that $\mathbb{P}(A) \in \{0, 1\}$. This is called Kolmogorov's 0-1 law.

[3]Note that $A \in \mathcal{T}$ if and only if the occurance of A does not depend on changing the values of any finite number of RVs X_n's — A depends on the 'tail' of the sequence of $(X_n)_{n\geq 1}$.

(i) If $B \in \sigma(X_1, \dots, X_m)$ and $C \in \sigma(X_{m+1}, X_{m+2}, \dots)$, then show that B and C are independent. (Hint: *if $C \in \sigma(X_{m+1}, \dots, X_{m+k})$ for some $k \geq 1$ use Prop. 2.2.4. For the general case, the latter σ-algebra is generated by $\bigcup_{k\geq 1} \sigma(X_{m+1}, \dots, X_{m+k})$, which is independent from $\sigma(X_1, \dots, X_m)$. Then use Prop. 2.2.2 to conclude.*)

(ii) If $B \in \sigma(X_1, X_2, \dots)$ and $C \in \mathscr{T}$, then B and C are independent. (Hint: *if $B \in \sigma(X_1, \dots, X_k)$ for some $k \geq 1$ use Prop. 2.2.4. For the general case, argue similarly as before.*)

(iii) From **(i)**-**(ii)**, dedcue that A is independent from itself. Hence $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$, so $\mathbb{P}(A) \in \{0, 1\}$.

Example 3.7.9. If B_n is a Borel subset of $\mathbb{R}$, then $\{X_n \in B_n \text{ i.o.}\}$ is a tail event. If $X_n = \mathbf{1}(A_n)$ and $B_n = \{1\}$, then it follows that $\{A_n \text{ i.o.}\}$ is also a tail event. Also, $\{\lim_{n\to\infty} X_n \text{ exists}\}$ is a tail event. By Exercise 3.7.8, these events all have probability 0 or 1.

Exercise 3.7.10 (Exchangeable sequence and Hewitt-Savage 0-1 law)**.** Let $X_1, X_x, \dots$ be a sequence of RVs on a probability spcae $(\Omega, \mathscr{F}, \mathbb{P})$. An event $A \in \mathscr{F}$ is called *exchangeable* if there exists a Borel set $B \subseteq \mathbb{R}^{\mathbb{N}}$ such that

$$\{(X_1, X_2, \dots) \in B\} = \{(X_{\sigma(1)}, X_{\sigma(2)}, \dots) \in B\}$$

for all finite permutations $\sigma : \mathbb{N} \to \mathbb{N}$ (here σ only permutes the first n entries for some finite n).

Now further assume that $X_1, X_2, \dots$ are i.i.d. and $\mathscr{F} = \sigma(X_1, X_2, \dots)$. Let $A \in \mathscr{F}$ be an exchangeable event. In this exercise, we will show that $\mathbb{P}(A) \in \{0, 1\}$. This is called Hewitt-Savage 0-1 law.

(i) Let $\mathscr{F}_n := \sigma(X_1, \dots, X_n)$. Show that there is a sequence of events $A_n \in \mathscr{F}_n$, $n \geq 1$, such that $\mathbb{P}(A_n \triangle A) \to 0$ as $n \to \infty$ (here $C \triangle D := (C \setminus D) \cup (D \setminus C)$ denotes the symmetric difference). We will write $A_n = \{(X_1, \dots, X_n) \in B_n\}$ for some Borel $B_n \subseteq \mathbb{R}^n$.

(ii) Let $\tilde{A}_n := \{X_{n+1}, \dots, X_{2n} \in B_n\}$. Show that $\mathbb{P}(\tilde{A}_n \triangle A) = \mathbb{P}(A_n \triangle A) = o(1)$. (*Hint:* Let σ be the permuation on $\mathbb{N}$ that maps $\{1, \dots, n\}$ to $\{n+1, \dots, 2n\}$ and leaves all other integers fixed. Then σ maps A_n to $\tilde{A}_n$ but fixes A. Since the RVs are i.i.d., applying σ to the indices does not change the probability.)

(iii) Deduce that $\mathbb{P}(A_n \cap \tilde{A}_n) \to \mathbb{P}(A)$ and $\mathbb{P}(A_n \cap \tilde{A}_n) = \mathbb{P}(A_n)\mathbb{P}(\tilde{A}_n) = \mathbb{P}(A_n)^2 \to \mathbb{P}(A)^2$ as $n \to \infty$. Conclude that $\mathbb{P}(A) \in \{0, 1\}$. (*Hint:* Use **(ii)** and that the RVs are i.i.d..)

Remark 3.7.11. Given a stochastic process with i.i.d. increments, the event that a state is visited infinitely often is in the tail space of the values of the process, however it is not in the tail space of the increments, so Kolmogorov's 0-1 does not apply. It is however an exchangeable event, and so occurs with probability 0 or 1 by Hewitt-Savage.

CHAPTER 4

Central Limit Theorems

In this chapter, we will study the celebrated central limit theorem (CLT) that we briefly touched upon in Theorem 3.1.3.

4.1. The De Moivre-Laplace CLT

In this section, we prove the following De Moivre-Laplace Theorem, which is a special case of the Central Limit Theorem (Theorem 3.1.3) we stated for the more general situation with i.i.d. increments with finite second moments. The argument is lementary and is based on explicit asymptotic computation using Stirling's formula.

Theorem 4.1.1 (De Moivre-Laplace CLT)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with*

$$\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = 1/2.$$

Let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Let $Z \sim N(0,1)$ be a standard normal RV and define

$$Z_n = \frac{S_n}{\sqrt{n}}.$$

Then Z_n converges to Z as $n \to \infty$ in distribution, namely,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z).$$

The proof is based on the asymptotic expression of simple symmetric random walk PMF (Lemma 4.1.3) based on Stirling's approximation (Exercise 4.1.4).

Proposition 4.1.2. *If $c_j \to 0$, $a_j \to \infty$, and $c_j a_j \to \lambda$ as $j \to \infty$, then*

$$\lim_{n\to\infty} (1 + c_j)^{a_j} = e^{\lambda}$$

PROOF. Let $y_j = (1 + c_j)^{a_j}$. Then $\log y_j = a_j \log(1 + c_j)$. Note that

$$\log y_j = a_j c_j \frac{\log(1 + c_j)}{c_j}.$$

Recall that $\log(1 + x)/x \to 1$ as $x \to 0$ (Use L'Hospital). Hence the above expression converges to $\lambda \cdot 1$ as $j \to \infty$. It follows that $y_j \to e^{\lambda}$ as $j \to \infty$. □

Lemma 4.1.3. *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with*

$$\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = 1/2.$$

Let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. If $\lim_{n\to\infty} 2k/\sqrt{2n} = x$, then

$$\mathbb{P}(S_{2n} = 2k) \sim (\pi n)^{-1/2} e^{-x^2/2},$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$.

PROOF. Note that $S_{2n} = 2k$ if and only if $n+k$ X_i's are $+1$ and $n-k$ X_i's are -1. Hence

$$\mathbb{P}(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n} = \frac{(2n)!2^{-2n}}{(n+k)!(n-k)!}.$$

By Stirling's approximation (27),

$$\begin{aligned}
\frac{(2n)!2^{-2n}}{(n+k)!(n-k)!} &\sim 2^{-2n} \frac{(2n/e)^{2n}\sqrt{4\pi n}}{((n+k)/e)^{n+k}\sqrt{2\pi(n+k)}\cdot((n-k)/e)^{n-k}\sqrt{2\pi(n-k)}} \\
&= \frac{1}{\sqrt{\pi}} \frac{n^{2n}\sqrt{n}}{(n+k)^{n+k}\sqrt{n+k}\cdot(n-k)^{n-k}\sqrt{n-k}} \\
&= \frac{1}{\sqrt{\pi n}} \left(1+\frac{k}{n}\right)^{-n-k} \left(1-\frac{k}{n}\right)^{-n+k} \left(1+\frac{k}{n}\right)^{-1/2} \left(1-\frac{k}{n}\right)^{-1/2} \\
&= \frac{1}{\sqrt{\pi n}} \left(1-\frac{k^2}{n^2}\right)^{-n} \left(1+\frac{k}{n}\right)^{-k} \left(1-\frac{k}{n}\right)^{k} \left(1+\frac{k}{n}\right)^{-1/2} \left(1-\frac{k}{n}\right)^{-1/2}.
\end{aligned}$$

Note that using Proposition 4.1.2, we have

$$\begin{aligned}
\lim_{n\to\infty} \left(1-\frac{k^2}{n^2}\right)^{-n} &= \exp\left(\lim_{n\to\infty} \frac{k^2}{n}\right) = \exp(x^2/2), \\
\lim_{n\to\infty} \left(1+\frac{k}{n}\right)^{-k} &= \exp\left(-\lim_{n\to\infty} \frac{k^2}{n}\right) = \exp(-x^2/2), \\
\lim_{n\to\infty} \left(1-\frac{k}{n}\right)^{k} &= \exp\left(-\lim_{n\to\infty} \frac{k^2}{n}\right) = \exp(-x^2/2).
\end{aligned}$$

Moreover, $k/n \to 0$ so $(1 \pm (k/n))^{-1/2} \to 1$. Combining the above equations, we get

$$\sqrt{\pi n}\mathbb{P}(S_{2n} = 2k) \to \exp(-x^2/2).$$

This shows the assertion. □

Now we prove the De Moivre-Laplace CLT.

PROOF OF THEOREM 4.1.1. Let $2\mathbb{Z}$ denote the set of all even integers. Then

$$\begin{aligned}
\mathbb{P}(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}) &= \sum_{m\in[a\sqrt{2n},b\sqrt{2n}]\cap 2\mathbb{Z}} \mathbb{P}(S_{2n} = m) \\
&= \sum_{x\in[a,b]\cap 2\mathbb{Z}/\sqrt{2n}} \mathbb{P}(S_{2n} = x\sqrt{2n}),
\end{aligned}$$

where we used change of variable $x = m/\sqrt{2n}$ and $2\mathbb{Z}/\sqrt{2n}$ denotes the set of all even integers divided by $\sqrt{2n}$. Then by Lemma 4.1.3,

$$\begin{aligned}
\sum_{x\in[a,b]\cap 2\mathbb{Z}/\sqrt{2n}} \mathbb{P}(S_{2n} = x\sqrt{2n}) &\approx \sum_{x\in[a,b]\cap 2\mathbb{Z}/\sqrt{2n}} \frac{1}{\sqrt{\pi n}} e^{-x^2/2} \\
&= \sum_{x\in[a,b]\cap 2\mathbb{Z}/\sqrt{2n}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{n/2}}.
\end{aligned}$$

We recognize the last summation as a Riemann sum for the definite integral $\int_a^b (2\pi)^{-1/2} e^{-x^2/2}\,dx$, since we are summing over all rectangles of width $1/\sqrt{n/2}$ within the interval $[a,b]$. This gives

$$\sum_{x\in[a,b]\cap 2\mathbb{Z}/\sqrt{2n}} \mathbb{P}(S_{2n} = x\sqrt{2n}) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx.$$

The integral above equals $\mathbb{P}(a \le Z \le b)$, where $Z \in N(0,1)$. This shows the assertion. □

Exercise 4.1.4 (Stirling's approximation)**.** In this exercise, we will show that

$$n! \sim (n/e)^n \sqrt{2\pi n}, \tag{27}$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$.

(i) Compare the summation $\sum_{k=1}^n \log k$ with Riemann integeral of $\log x$ to show that

$$\int_1^n \log x\, dx \le \log n! \le \int_1^{n+1} \log x\, dx.$$

(ii) Use the fact that $\int \log x\, dx = x\log x - x + C$ to deduce that

$$n\log n - n \le \log n! \le (n+1)\log(n+1) - n.$$

(iii) Define

$$d_n = \log n! - \left(n + \frac{1}{2}\right)\log(n) + n.$$

Show that

$$d_n - d_{n+1} = \left(n + \frac{1}{2}\right)\log\left(1 + \frac{1}{n}\right) - 1.$$

(iv) Use (iii) and the Taylor expansion

$$\log\left(1 + \frac{1}{n}\right) = \frac{1}{n} - \frac{1}{2n^2} + \frac{1}{3n^3} - \frac{1}{4n^4} + \cdots$$

to snow that

$$d_n - d_{n+1} = \sum_{k=2}^{\infty} \frac{(-1)^k (k-1)}{2k(k+1)n^k} = \frac{1}{2\cdot 2\cdot 3n^2} - \frac{2}{2\cdot 3\cdot 4n^3} + \frac{3}{2\cdot 4\cdot 5n^4} - \frac{4}{2\cdot 5\cdot 6n^5} + \cdots.$$

Use facts about alternating series to deduce

$$0 < d_n - d_{n+1} < \frac{1}{n^2}.$$

(v) From (iv), deduce that d_n is a decreasing sequence and

$$d_1 - d_{n+1} = \sum_{k=1}^n (d_k - d_{k+1}) < \sum_{k=1}^n \frac{1}{k^2} < \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty.$$

Conclude that $\lim_{n\to\infty} d_n = C$ for some constant C.

(vi) Using (v) and (ii), deduce

$$\lim_{n\to\infty} \frac{n!}{n^{n+1/2} e^n} = e^C.$$

According to Wallis formula (see, e.g., ref), $C = \log\sqrt{2\pi}$. Thus we obtain Stirling's approximation (27).

Exercise 4.1.5 (Convergence of product)**.** This exericise generalizes Proposition 4.1.2 to triangular arrays. Consider a trianglular array of sequence $c_{1,n}, c_{2,n}, \ldots, c_{n,n}$ for $n \ge 1$. Suppose

$$\max_{1\le k\le n} |c_{k,n}| = o(1), \quad \lim_{n\to\infty} \sum_{k=1}^n c_{k,n} = \lambda, \quad \sup_{n\ge 1} \sum_{k=1}^n |c_{k,n}| < \infty.$$

Then show that

$$\lim_{n\to\infty} \prod_{k=1}^n \left(1 + c_{k,n}\right) = \exp(\lambda).$$

4.2. Weak convergence

So far we have studied two modes of convergence of RVs: (1) convergnece in probability that shows up in the weak laws of large numbers (see, e.g., Theorem 3.5.1) and (2) the almost sure convergence that shows up in the strong laws of large numbers (see, e.g., Theorem 3.5.1). In this section, we introduce another mode of convergence that appears in CLTs and give it a systematic treatment.

Definition 4.2.1 (Weak convergence)**.** A sequence of distribution functions F_n is said to *converge weakly* (or *in distribution*) to another distribution function F, in which case we denote $F_n \Rightarrow F$, if

$$\lim_{n\to\infty} F_n(y) = F(y) \quad \text{for all } y \in \mathbb{R} \text{ s.t. } F \text{ is continuous at } y.$$

If X, X_n, $n \geq 1$ are RVs (not necessarily defined on the same probability space), then we say X_n *converges weakly* (or *in distribution*) to X (and write $X_n \Rightarrow X$) if their distribution functions $F_n(x) := \mathbb{P}(X_n \leq x)$ converge weakly. If μ_n, μ for $n \geq 1$ are probability measures on $\mathbb{R}$, then we say μ_n converge to μ weakly (and write $\mu_n \Rightarrow \mu$) if the associated distribution functions $x \mapsto \mu_n((-\infty, x])$ converge weakly to $x \mapsto \mu((-\infty, x])$.

Exercise 4.2.2. (Countable discontinuities of CDF) Let F denote the distribution function of some random variable X. Show that F has at most countably many discontinuity points. Furthermore, show that the set of continuity points of F is dense in $\mathbb{R}$. (*Hint: Let D = set of discontinuity points of F. For each $x \in D$, $F(x^-) < F(x)$, so $I_x := (F(x^{-1}), F(x))$ is a non-empty open interval. Argue that if $x, y \in D$ are distinct, then I_x and I_y are disjoint. Noting that each I_x contain at least one rational number, deduce that D is countable.*)

Exercise 4.2.3 (Uniqueness of weak limit)**.** Suppose $F_n \Rightarrow F$ and $F_n \Rightarrow G$ for distribution functions F_n, F, and G. Show that $\lambda(F \neq G) = 0$, where λ denotes the Lebesgue measure on $\mathbb{R}$.

4.2.1. Examples of weak convergence.

Example 4.2.4. Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = 1/2$. Let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Let $Z \sim N(0,1)$ be a standard normal RV and define

$$Z_n = \frac{S_n}{\sqrt{n}}.$$

Then Theorem 4.1.1 shows that $Z_n \Rightarrow Z$ as $n \to \infty$. ▲

Example 4.2.5 (Weak convergence of empirical distributions)**.** Let $X_1, X_2, \ldots$ be a sequence of RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with common distribution function F. By the Glivenko-Cantelli theorem (see Theorem 3.5.4),

$$F_n(x) := \frac{1}{n}\sum_{k=1}^n \mathbf{1}(X_k \leq x) \quad \longrightarrow \quad F(x) \quad \text{for all } x \in \mathbb{R}.$$

In particular, this implies $F_n \Rightarrow F$ as $n \to \infty$. In fact Glivenko-Cantelli shows that the convergence holds for all points x, but weak convergence only requires the convergence only at the continuity points of F. ▲

The following example shows that it is possible to have $F_n(x) \to F(x)$ precisely only at the continuity points of F.

Example 4.2.6. Let a RV X has distribution function F. Let $X_n := X + n^{-1}$, which has distribution function F_n given by

$$F_n(x) = \mathbb{P}(X_n \leq x) = \mathbb{P}(X + n^{-1} \leq x) = \mathbb{P}(X \leq x - n^{-1}) = F(x - n^{-1}).$$

Then $\lim_{n\to\infty} F_n(x) = \lim_{y \nearrow x} F(y) = F(x^-)$. Since F is right-continuous, it follows that $\lim_{n\to\infty} F_n(x) = F(x)$ if and only if x is a continuity point of F. ▲

Example 4.2.7 (Birthday problem)**.** Let $X_1, X_2, \dots$ be i.i.d. RVs with uniform distribution on $\{1, \dots, N\}$. For each $N \geq 1$, define $T_N := \max\{n \geq 1 : X_1, \dots, X_n \text{ are all distinct}\}$. Note that

$$\mathbb{P}(T_N \geq n) = \prod_{m=2}^{n} \left(1 - \frac{m-1}{N}\right). \tag{28}$$

When $N = 365$, the above equals the probability of not having the same birthday in a class of n students. This quantity is surprisingly small for modest values of n. To see this, let $c_{m,N} := \frac{m-1}{N}$ and consider the triangular array $c_{1,N}, \dots, c_{\lfloor x\sqrt{N}\rfloor, N}$ for $N \geq 1$. Then

$$\max\left(\frac{1-1}{N}, \dots, \frac{\lfloor x\sqrt{N}\rfloor - 1}{N}\right) = \frac{\lfloor x\sqrt{N}\rfloor - 1}{N} = o(1),$$

$$\sum_{k=1}^{\lfloor x\sqrt{N}\rfloor} \frac{m-1}{N} = \frac{1}{N} \frac{\lfloor x\sqrt{N}\rfloor(\lfloor x\sqrt{N}\rfloor - 1)}{2} \to \frac{x^2}{2},$$

$$\sup_{N\geq 1} \sum_{k=1}^{\lfloor x\sqrt{N}\rfloor} \frac{m-1}{N} = \sup_{N\geq 1} \frac{1}{N} \frac{\lfloor x\sqrt{N}\rfloor(\lfloor x\sqrt{N}\rfloor - 1)}{2} < \infty.$$

Hence by Exercise 4.1.5, we have

$$\lim_{N\to\infty} \mathbb{P}(T_N/\sqrt{N} \geq x) = \lim_{N\to\infty} \mathbb{P}(T_N \geq \lfloor x\sqrt{N}\rfloor) = \exp(-x^2/2) \quad \text{for } x \geq 0.$$

Take $N = 365$ and noting that $22/\sqrt{365} = 1.1515$ and $(1.1515)^2/2 = 0.6630$, this yields

$$\mathbb{P}(T_{365} > 22) \approx \exp(-0.6630) \approx 0.515.$$

The above approximation is within 2% from the true value $\mathbb{P}(T_{365} > 22) \approx 0.524$ that can be computed from the exact formula (28). Thus, in a class of at least 22 students, there is at least 47% chance to have two students of the exactly same birthday. ▲

4.2.2. Theory on weak convergence. In this subsection, we give some theoretical results concerning weak convergence.

The following result shows that one can construct a sequence of RVs converging almost surely to realize a weak convergence of distribution functions.

Lemma 4.2.8 (Weak convergence realized as a.s. convergence)**.** *If $F_n \Rightarrow F$, then there are random variables Y, Y_n, $n \geq 1$, with distribution function F_n so that $Y_n \to Y$ a.s.*

PROOF. Following Proposition 1.2.13, we define RVs on the 'standard probability space' $(\Omega, \mathscr{B}, \mathbb{P})$, where $\Omega = (0,1)$, $\mathscr{B}$ is the Borel σ-algebra on $(0,1)$ and $\mathbb{P}$ denotes the Lebesgue measure on $(0,1)$. Let $Y_n : (0,1) \to \mathbb{R}$ defined by $Y_n(\omega) := \sup\{y : F_n(y) < \omega\}$ and $Y(\omega) := \sup\{y : F(y) < \omega\}$. For convenience, we denote $Y_n(\omega) = F_n^{-1}(\omega)$ and $Y(\omega) = F^{-1}(\omega)$. In Proposition 1.2.13, we have shown that F^{-1}, F_n^{-1}, $n \geq 1$ are RVs with prescribed distribution functions. It remains to show that $F_n^{-1} \to F^{-1}$ a.s.. To this end, we make the following claims:

(a) For each $\omega \in (0,1)$, define $a_\omega := \sup\{y : F(y) < \omega\}$ and $b_\omega := \inf\{y : F(y) > \omega\}$. Let $\Omega_0 = \{\omega : a_\omega = b_\omega\}$. Then $\Omega \setminus \Omega_0$ is countable.
(b) $F_n^{-1}(\omega) \to F^{-1}(\omega)$ for all $\omega \in \Omega_0$.

Given the above claims, $\mathbb{P}(\lim_{n\to\infty} Y_n = Y) \geq \mathbb{P}(\Omega_0) = 1 - \mathbb{P}(\Omega \setminus \Omega_0) = 1$, since $\mathbb{P}$ is the Lebesgue measure on $(0,1)$ and $\mathbb{P}(\Omega \setminus \Omega_0) = 0$ since $\Omega \setminus \Omega_0$ is countable.

To justify (a), first note that $a_\omega \leq b_\omega$ by the definition. Also from the definition, nonempty open intervals (a_ω, b_ω) for $\omega \in \Omega \setminus \Omega_0$ are disjoint. Hence these open intervals contain distinct rational number, so the cardinality of $\Omega \setminus \Omega_0$ is at most the cardinality of the rational numbers. It follows that $\Omega \setminus \Omega_0$ is countable.

To justify (b), first observe that

$$y < F^{-1}(\omega) \Rightarrow F(y) < \omega \quad \text{for } \omega \in \Omega \tag{29}$$
$$y > F^{-1}(\omega) \Rightarrow F(y) > \omega \quad \text{for } \omega \in \Omega_0.$$

Indeed, the first implication follows from the definition of F^{-1}. The second implication follows since for $\omega \in \Omega_0$, $F^{-1}(\omega) = a_\omega = b_\omega$, and $y > b_\omega$ implies $F(y) > \omega$ by definition of b_ω.

Next, fix $\omega \in \Omega$ and we show

$$\liminf_{n\to\infty} F_n^{-1}(\omega) \geq F^{-1}(\omega). \tag{30}$$

Suppose for contradiction that the strict inequality $<$ holds above. By Exercise 4.2.2, we can choose a continuity point y of F such that $\liminf_{n\to\infty} F_n^{-1}(\omega) < y < F^{-1}(\omega)$. If we choose a subsequence $n(k)$ such that $\lim_{k\to\infty} F_{n(k)}^{-1}(\omega) = \liminf_{n\to\infty} F_n^{-1}(\omega)$, then for all sufficiently large $k \geq 1$,

$$F_{n(k)}^{-1}(\omega) < y < F^{-1}(\omega).$$

Applying $F_{n(k)}$ and F to the first and the second inequalies with (29) and using the fact that they are non-decreasing,

$$\omega \leq F_{n(k)}(y), \quad F(y) < \omega \quad \text{for all } k \gg 1.$$

But since $F_n \Rightarrow F$ and since F is continuous at y, we have $F_n(y) \to F(y)$ as $n \to \infty$. Hence by taking $k \to \infty$,

$$\omega = \lim_{k\to\infty} \omega \leq \lim_{k\to\infty} F_{n(k)}(y) = F(y) < \omega,$$

which is a contradiction. This shows (30).

Lastly, fix $\omega \in \Omega_0$ and we show

$$\limsup_{n\to\infty} F_n^{-1}(\omega) \leq F^{-1}(\omega). \tag{31}$$

Suppose for contradiction that the strict inequality $>$ holds above. By Exercise 4.2.2, we can choose a continuity point y of F such that $\liminf_{n\to\infty} F_n^{-1}(\omega) > y > F^{-1}(\omega)$. If we choose a subsequence $n(k)$ such that $\lim_{k\to\infty} F_{n(k)}^{-1}(\omega) = \limsup_{n\to\infty} F_n^{-1}(\omega)$, then for all sufficiently large $k \geq 1$,

$$F_{n(k)}^{-1}(\omega) > y > F^{-1}(\omega).$$

Applying $F_{n(k)}$ and F to the first and the second inequalies with (29) (here we need to use the hypothesis that $\omega \in \Omega_0$) and using the fact that they are non-decreasing,

$$\omega \geq F_{n(k)}(y), \quad F(y) > \omega \quad \text{for all } k \gg 1.$$

But since $F_n \Rightarrow F$ and since F is continuous at y, we have $F_n(y) \to F(y)$ as $n \to \infty$. Hence by taking $k \to \infty$,

$$\omega = \lim_{k\to\infty} \omega \geq \lim_{k\to\infty} F_{n(k)}(y) = F(y) > \omega,$$

which is a contradiction. This shows (31). This completes the proof. □

An immediate and important application of Lemma 4.2.8 is the following characterization of weak convergence, which explains the naming of 'weak convergence'.

Proposition 4.2.9 (Characterization of weak convergence)**.** *Let X, X_n, $n \geq 1$ be a sequence of RVs not necessarily defined on the same probability space. Then $X_n \Rightarrow X$ if and only if for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$.*

Remark 4.2.10. If $F_n(x) = \mathbb{P}(X_n \leq x)$, $F(x) = \mathbb{P}(X \leq x)$, then Prop. 4.2.9 shows that $F_n \Rightarrow F$ iff $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$. But by change of variables,

$$\int_{\mathbb{R}} g(x)\, dF_n(x) \to \int_{\mathbb{R}} g(x)\, dF(x).$$

Similarly, if $\mu_n \Rightarrow \mu$ is given, then by Prop. 1.2.13, we can cook up sequence of RVs $X_n \Rightarrow X$ with $\mu_n = \mathbb{P} \circ X^{-1}$ and $\mu = \mathbb{P} \circ X^{-1}$ so $\mu_n \Rightarrow \mu$ iff $F_n \Rightarrow F$ iff $X_n \Rightarrow X$ iff $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ for all test functions g iff

$$\int_{\mathbb{R}} g(x)\, d\mu_n(x) \to \int_{\mathbb{R}} g(x)\, d\mu(x)$$

for all test functions, by change of variables.

Proof of Proposition 4.2.9. According to Lemma 4.2.8, we can choose a sequence of RVs Y, Y_n, $n \geq 1$ such that $X_n \overset{d}{=} Y_n$, $X \overset{d}{=} Y$, and that $Y_n \to Y$ a.s.. Then by BCT (see Theorem 1.3.16),

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \lim_{n\to\infty} \mathbb{E}[g(Y_n)] = \mathbb{E}[g(Y)] = \mathbb{E}[g(X)].$$

For the converse direction, suppose that for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$. We wish to show that $X_n \Rightarrow X$. This would be immediate if we can take g to be the indicator function $g_x(y) = \mathbf{1}(y \leq x)$, but such indicator 'test function' is not allowed since we only use bounded and continuous test functions. Instead, we approximate g_x by a piecewise linear function. That is, fix $x \in \mathbb{R}$ and $\varepsilon > 0$, and define a bounded and continuous function $g_{x,\varepsilon}$ by

$$g_{x,\varepsilon}(y) = \begin{cases} 1 & \text{if } y < x \\ 0 & \text{if } y > x + \varepsilon \\ 1 - \frac{1}{\varepsilon}(y - x) & \text{if } x \leq y \leq x + \varepsilon. \end{cases}$$

Then $\mathbf{1}(y \leq x) \leq g_{x,\varepsilon}(y) \leq \mathbf{1}(y \leq x + \varepsilon)$, so

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}(X_n \leq x) &= \limsup_{n\to\infty} \mathbb{E}[\mathbf{1}(X_n \leq x)] \\ &\leq \limsup_{n\to\infty} \mathbb{E}[g_{x,\varepsilon}(X_n)] \\ &= \mathbb{E}[g_{x,\varepsilon}(X)] \\ &\leq \mathbb{E}[\mathbf{1}(X \leq x + \varepsilon)] = \mathbb{P}(X \leq x + \varepsilon). \end{aligned}$$

Letting $\varepsilon \searrow 0$ and using the right continuity of CDF (see Prop. 1.2.12 **(iii)**), we deduce

$$\limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x) \quad \text{for all } x \in \mathbb{R}.$$

For the other direction, fix $x \in \mathbb{R}$ and $\varepsilon > 0$, and define a bounded and continuous function $h_{x,\varepsilon}$ by

$$h_{x,\varepsilon}(y) = \begin{cases} 1 & \text{if } y < x \\ 0 & \text{if } y > x - \varepsilon \\ 1 + \frac{1}{\varepsilon}(y - x) & \text{if } x - \varepsilon \leq y \leq x. \end{cases}$$

Then $\mathbf{1}(y \leq x - \varepsilon) \leq h_{x,\varepsilon}(y) \leq \mathbf{1}(y \leq x)$, so

$$\begin{aligned} \liminf_{n\to\infty} \mathbb{P}(X_n \leq x) &= \limsup_{n\to\infty} \mathbb{E}[\mathbf{1}(X_n \leq x)] \\ &\geq \liminf_{n\to\infty} \mathbb{E}[h_{x,\varepsilon}(X_n)] \\ &= \mathbb{E}[h_{x,\varepsilon}(X)] \\ &\geq \mathbb{E}[\mathbf{1}(X \leq x - \varepsilon)] = \mathbb{P}(X \leq x - \varepsilon). \end{aligned}$$

Letting $\varepsilon \searrow 0$ and using continuity of probability measure from below (see Prop. 1.2.12 **(iv)**), we deduce

$$\liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X < x) \quad \text{for all } x \in \mathbb{R}.$$

To finish, now let x be any continuity point of F. Then

$$\liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X < x) = \mathbb{P}(X \leq x) \geq \limsup_{n\to\infty} \mathbb{P}(X_n \leq x).$$

This shows $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$, as desired. □

Remark 4.2.11 (Test functions)**.** At a high level, the role of a 'test function' is testing some aspect of an unknown object. Think of slicing a high-dimensional and complicated shape by some two-dimensional plane. You can look at the section (now it's 2D) and study properties of it. That's certainly not enough to tell you the whole object, but still gives you some information. You can also take sections of multiple other planes and study the corresponding sections. Technically if you study enough collection of such sections, sometimes that gives you complete understanding of the object.

In our situation, a RV X is the object we want to know, and a function $f:\mathbb{R}\to\mathbb{R}$ gives some "section" of X by the expectation $\mathbb{E}[f(X)]$. If we take f to be of the form $f(x)=\mathbf{1}(x\le A)$ for A an interval, then knowing $\mathbb{E}[f(X)]$ for all such indicator test functions is enough to determine the distribution of X. Also, if we want to know if $X_n\Rightarrow X$, then it is enough to know if $\mathbb{E}[f(X_n)]\to\mathbb{E}[f(X)]$ for all f bounded and continuous functions (Prop. 4.2.9).

As for another example in modern graph theory, there is a notion of a sequence of graphs G_n converging to a limiting objected U called a 'graphon' in a topology induced by the 'cut metric'. A well-known theorem by Lováz states that $G_n\to U$ if and only if $f(G_n)\to f(U)$ for all f, which takes a graph/graphon and gives the 'homomorphism density' of an arbitrary finite graph.

Theorem 4.2.12 (Continuous mapping theorem)**.** *Let g be a measurable function and let D_g denote the set of discontinuity of g. If $X_n\Rightarrow X$ and $\mathbb{P}(X\in D_g)=0$ then $g(X_n)\Rightarrow g(X)$. If in addition g is bounded then $\mathbb{E}[g(X_n)]\to\mathbb{E}[g(X)]$.*

PROOF. First note that D_g is a Borel set. By Lemma 4.2.8, we may choose RVs $Y_n=_d X_n$ with $Y_n\to Y$ a.s. and $X=_d Y$. If f is continuous then $D_{f\circ g}\subseteq D_g$ so $\mathbb{P}(Y\in D_{f\circ g})\le\mathbb{P}(Y\in D_g)=\mathbb{P}(X\in D_g)=0$ so $\mathbb{P}(Y\in D_{f\circ g})=0$. It follows that

$$\begin{aligned}\mathbb{P}\left(f(g(Y_n))\to f(g(Y))\right)&=\mathbb{P}\left(f(g(Y_n))\to f(g(Y)) \text{ and } Y\notin D_{f\circ g}\right)\\&=\mathbb{P}\left(\left\{\omega:\lim_{n\to\infty}f\circ g(Y_n(\omega))=f\circ g(Y(\omega))\right\}\cap\{\omega:Y(\omega)\notin D_{f\circ g}\}\right)\\&\ge\mathbb{P}\left(\{\omega:f\circ g(Y(\omega))=f\circ g(Y(\omega))\}\cap\{\omega:Y(\omega)\notin D_{f\circ g}\}\right)\\&=\mathbb{P}\left(\{\omega:Y(\omega)\notin D_{f\circ g}\}\right)=1.\end{aligned}$$

Hence $f(g(Y_n))\to f(g(Y))$ a.s.. If, in addition, f is bounded then the bounded convergence theorem implies $\mathbb{E}[f(g(Y_n))]\to\mathbb{E}[f(g(Y))]$. Since this holds for all bounded continuous functions f, it follows from Proposition 4.2.9 that $g(X_n)\Rightarrow g(X)$. For the second part of the assertion, note that $Y_n\to Y$ a.s. and $\mathbb{P}(Y\in D_g)$. So by a similar argument,

$$\begin{aligned}\mathbb{P}\left(g(Y_n)\to g(Y)\right)&=\mathbb{P}\left(\{\omega:g(Y_n(\omega))\to g(Y(\omega))\}\cap\{\omega:Y(\omega)\notin D_g\}\right)\\&\ge\mathbb{P}(\{\omega:Y(\omega)\notin D_g\})=1.\end{aligned}$$

Hence $g(Y_n)\to g(Y)$ a.s., and the desired result follows from the bounded convergence theorem. □

Proposition 4.2.13 (Equivalent conditiosn for weak convergence)**.** *These following are all equivalent:*

(i) $X_n\Rightarrow X$*;*

(ii) *For all open sets $G\subseteq\mathbb{R}$, $\liminf_{n\to\infty}\mathbb{P}(X_n\in G)\ge\mathbb{P}(X\in G)$;*

(iii) *For all closed sets $K\subseteq\mathbb{R}$, $\limsup_{n\to\infty}\mathbb{P}(X_n\in K)\le\mathbb{P}(X\in K)$;*

(iv) *For all Borel sets $B\subseteq\mathbb{R}$ with $\mathbb{P}(X\in\partial B)=0$, $\lim_{n\to\infty}\mathbb{P}(X_n\in B)=\mathbb{P}(X\in B)$.*

PROOF. (i) ⇒ (ii) : By Lemma 4.2.8, we may choose RVs $Y_n=_d X_n$ with $Y_n\to Y$ a.s. and $X=_d Y$, where Y,Y_n are defined on the same probability space $(\Omega,\mathscr{F},\mathbb{P})$. Then since G is open, almost surely,

$$\liminf_{n\to\infty}\mathbf{1}(Y_n\in G)\ge\mathbf{1}(Y\in G).$$

To see this, let $A:=\{\omega:Y_n(\omega)\to Y(\omega)\}$. Then by the hypothesis, $\mathbb{P}(A)=1$. Fix $\omega\in A$. Choose a subsequence $n(k)$ so that $\lim_{k\to\infty}\mathbf{1}(Y_{n(k)}(\omega)\in G)=0$. This means for all large enough $k\ge1$,

$Y_{n(k)}(\omega) \in G^c$. Since G^c is closed, any limit point of the sequence $Y_{n(k)}(\omega)$ belongs to G^c. Since $\omega \in A$, $Y_{n(k)}(\omega) \to Y(\omega) \in G^c$.

$$\begin{aligned}\mathbb{P}\left(\liminf_{n\to\infty} \mathbf{1}(Y_n \in G) = 0\right) &= \mathbb{P}\left(\left\{\liminf_{n\to\infty} \mathbf{1}(Y_n \in G) = 0\right\} \cap A\right) \\ &\geq \mathbb{P}\left(Y \in G^c\right) = \mathbb{P}(\mathbf{1}(Y \in G) = 0).\end{aligned}$$

Since $\liminf_{n\to\infty} \mathbf{1}(Y_n \in G)$ and $\mathbf{1}(Y \in G)$ are RVs taking values from $\{0,1\}$, this shows the desired conclusion. Now applying Fatou's lemma (see Thm. 1.3.18) gives

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}(X_n \in G) &= \liminf_{n\to\infty} \mathbb{P}(Y_n \in G) \\ &\geq \mathbb{P}(\liminf_{n\to\infty} Y_n \in G) \\ &\geq \mathbb{P}(Y \in G) \\ &= \mathbb{P}(X \in G).\end{aligned}$$

(ii)⇒(iii) : Apply the previous implication by noting that K^c is open.

(ii), (iii) ⇒ (iv) : Let $\overline{B}$ and B° denote the closure and the intrerior if B, respectively. Note taht $\overline{B}$ and B° are closed and open in $\mathbb{R}$, respectively. The boundary $\partial B := \overline{B} \setminus B^\circ$. Then

$$\mathbb{P}(X \in \overline{B}) = \mathbb{P}(X \in B^\circ \sqcup \partial B) = \mathbb{P}(X \in B^\circ) + \mathbb{P}(X \in \partial B) = \mathbb{P}(X \in B^\circ).$$

Since $\overline{B} \subseteq B \subseteq B^\circ$, we get

$$\mathbb{P}(X \in \overline{B}) = \mathbb{P}(X \in B) = \mathbb{P}(X \in B^\circ).$$

Thus by using (ii) and (iii),

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in B) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \in \overline{B}) \leq \mathbb{P}(X \in \overline{B}) = \mathbb{P}(X \in B), \\ \liminf_{n\to\infty} \mathbb{P}(X_n \in B) \geq \liminf_{n\to\infty} \mathbb{P}(X_n \in B^\circ) \leq \mathbb{P}(X \in B^\circ) = \mathbb{P}(X \in B).\end{aligned}$$

Hence the conclusion follows.

(iv)⇒ (i) : Let x be any continuity point of F, the CDF of X. Let $B := (-\infty, x]$. Then $\mathbb{P}(X \in \partial B) = \mathbb{P}(X = x) = 0$. Hence by (iv), $\lim_{n\to\infty} \mathbb{P}(X_n \in x) = \mathbb{P}(X \leq x)$. Thus $X_n \Rightarrow X$.

□

Theorem 4.2.14 (Helly's selection theorem)**.** *For every sequence F_n of distribution functions, there is a subsequence $F_{n(k)}$ and a right continuous nondecreasing function F so that $\lim_{k\to\infty} F_{n(k)}(y) = F(y)$ at all continuity points y of F.*

PROOF. We will first construct a candidate limiting function G on the rationals. Let $q_1, q_2, \ldots$ be an enumeration of the rationals. Since $F_n(q_1)$ is an infinite sequence in the compact set $[0,1]$, there exists a subsequence $n(m_1(k))$ such that $F_{n(m_1(k))}(q_1)$ converges to some point in $[0,1]$, which we will denote as $G(q_1)$. Next, $F_{n(m_1(k))}(q_2)$ is an infinite sequence in $[0,1]$, so we can choose a further subsequence $n(m_2(k))$ of $n(m_1(k))$ such that $F_{n(m_2(k))}(q_2)$ converges to some point in $[0,1]$, which we will denote as $G(q_2)$. Repeating this process, we can consturct a nested subsequences $(n_1(k))_{k\geq 1} \supseteq (n_2(k))_{k\geq 1} \supseteq \cdots$ and a function $G : \mathbb{Q} \to [0,1]$ such that

$$\lim_{k\to\infty} F_{n(m_i(k))}(q_i) = G(q_i) \quad \text{for all } i \geq 1.$$

Now we take the 'diagonal subsequence', $n_k := n(m_k(k))$, $k \geq 1$. Then by construction,

$$\lim_{k\to\infty} F_{n_k}(q_i) = G(q_i) \quad \text{for all } i \geq 1. \tag{32}$$

Note that G is non-decreasing over $\mathbb{Q}$ since if $q, q' \in \mathbb{Q}$ and $q < q'$, then $F_{n_k}(q) \leq F_{n_k}(q')$ for all $k \geq 1$ so taking $k \to \infty$ gives $G(q) \leq G(q')$.

Next, we define a function $F:\mathbb{R}\to[0,1]$ by extending G to all reals by

$$F(x):=\inf\{G(q):q\in\mathbb{Q},\,q>x\}.$$

From the definition it is easy to see that $F(x)\le F(y)$ if $x<y$. Also, F is right-continuous. To see this, let $x_n\searrow x$. Then $F(x_n)\searrow F(x)$ so $\lim_{n\to\infty}F(x_n)=\inf_{n\ge1}F(x_n)$. Hence

$$\begin{aligned}\lim_{x_n\searrow x}F(x_n)&=\inf_{n\ge1}\inf\{G(q):q\in\mathbb{Q},\,q>x_n\}\\&=\inf\{G(q):q\in\mathbb{Q},\,q>x_n\text{ for some }n\ge1\}\\&=\inf\{G(q):q\in\mathbb{Q},\,q>x\}=F(x).\end{aligned}$$

This shows that F is a distribution function.

To complete the proof, we will show $F_{n_k}\Rightarrow F$. Let x be a continuity point of F so that $\lim_{y\to x}F(y)=F(x)$. Fix $\varepsilon>0$ and pick rationals a,b,c such that $a<b<x<c$ and

$$F(x)-\varepsilon<F(a)\le F(b)\le F(x)\le F(c)<F(x)+\varepsilon. \tag{33}$$

By definition of F and since G is non-decreasing over $\mathbb{Q}$,

$$F(a)\le G(b)\le G(c)\le F(c).$$

From (33), this yields

$$F(x)-\varepsilon<G(b)\le G(c)<F(x)+\varepsilon.$$

Now using (32) and $F_{n_k}(a)\le F_{n_k}(x)\le F_{n_k}(b)$, we get

$$F(x)-\varepsilon<\lim_{k\to\infty}F_{n_k}(b)\le\liminf_{k\to\infty}F_{n_k}(x)\le\limsup_{k\to\infty}F_{n_k}(x)\le\lim_{k\to\infty}F_{n_k}(c)<F(x)+\varepsilon.$$

Then taking $\varepsilon\nearrow0$ shows $\lim_{k\to\infty}F_{n_k}(x)=F(x)$, as desired. □

Remark 4.2.15. The limit F in Theorem 4.2.14 might not be a distribution function. For instance, let $a,b,c>0$ s.t. $a+b+c=1$ and let G be a distribution function. Define F_n by

$$F_n(x)=a\mathbf{1}(x\ge-n)+b\mathbf{1}(x\ge n)+cG(x)$$

Then F_n is a distribution function, as

$$F_n(x)=\begin{cases}cG(x) & \text{if } x<-n\\ a+cG(x) & \text{if } -n\le x<n\\ a+b+cG(x) & \text{if } n\le x.\end{cases}$$

Now for any fixed $x\in\mathbb{R}$, $\lim_{n\to\infty}F_n(x)=a+cG(x)$, so $F_n\Rightarrow F$ where $F=a+cG$. But F is not a distribution function since $\lim_{x\to\infty}F(x)=a+c=1-b$ and $\lim_{x\to-\infty}F(x)=a$. In other words, mass b 'escapes to $+\infty$ 'and mass b 'escapes to $-\infty$'. This type of 'improper weak' convergence is called the *vague convergence*.

Definition 4.2.16 (Tightness)**.** A sequence of RVs X_n is *tight* if for each $\varepsilon>0$, there exists $M_\varepsilon>0$ such that

$$\limsup_{n\to\infty}\mathbb{P}(|X_n|>M_\varepsilon)\le\varepsilon.$$

(in words, 'no mass excapes to $\pm\infty$'). A sequence of distribution functions F_n, $n\ge1$ is *tight* if for each $\varepsilon>0$, there exists $M_\varepsilon>0$ such that

$$\limsup_{n\to\infty}F_n(-M_\varepsilon)+1-F_n(M_\varepsilon)\le\varepsilon.$$

Exercise 4.2.17 (Tightness and weak convergence)**.** Show the following statement: Let F_n be a sequence of distribution functions. Then every subsequential limit of F_n is a distribution function if and only if $(F_n)_{n\ge1}$ is tight.

Exercise 4.2.18 (A sufficient condition for tightness)**.** Let F_n be a sequence of distribution functions. Show that $(F_n)_{n\geq 1}$ is tight if there exists a function $\varphi \geq 0$ such that $\varphi(x) \to 0$ as $|x| \to \infty$ and

$$\sup_{n\geq 1} \int \varphi(x)\, dF_n(x) < \infty.$$

Exercise 4.2.19 (The Lévy metric)**.** Let $\mathscr{D}$ denote the space of all distribution functions. Define a function $\rho : \mathscr{D} \times \mathscr{D} \to \mathbb{R}$ by

$$\rho(F,G) = \inf\{\varepsilon : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x \in \mathbb{R}\}.$$

Show that ρ defines a metric on $\mathscr{D}$. Furthermore, show that if $F, F_n, n \geq 1$ are in $\mathscr{D}$, then $F_n \Rightarrow F$ if and only if $\rho(F_n, F) \to 0$.

Exercise 4.2.20 (Slutzky's theorem)**.** Show that if $X_n \Rightarrow X$ and $Y_n \to y \in \mathbb{R}$, then $X_n + Y_n \Rightarrow X + y$.

4.3. Characteristic functions

In this section, we introduce the theory of characteristic functions. Roughly speaking, one considers a function φ that 'represent' a random variable X in a way that study of RVs can be transfered to a study of functions. In this section, we let $i := \sqrt{-1}$ and let $\mathbb{C} := \{a + ib : a, b \in \mathbb{R}\}$ denote the set of all complex numbers. For a complex number $z = a + ib$, we denote $\mathrm{Re}(z) := a$ (the real part of z), $\mathrm{Im}(z) := b$ (the imaginary part of z), $\bar{z} := a - ib$ (the complex conjugate of z), and $|z| = \sqrt{a^2 + b^2} = z\bar{z}$ (the modulus of z).

4.3.1. Definition and the inversion formula. One can define a complex-valued RVs in a straightforward mannaer. That is, let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We can put the Borel σ-algebra $\mathscr{B}$ on $\mathbb{C}$ the usual Borel σ-algebra we put on $\mathbb{R}^2$ by identifying $a + ib$ with (a, b). Then a complex-valued RV is simply a measurable function $Z : (\Omega, \mathscr{F}) \to (\mathbb{C}, \mathscr{B})$.

Definition 4.3.1. Let X be a RV. We define its *characteristic function* $\varphi_X : \mathbb{R} \to \mathbb{C}$ by

$$\varphi_X(t) := \mathbb{E}\left[\exp(-itX)\right] = \mathbb{E}\left[\cos(tX)\right] + i\mathbb{E}\left[\sin(tX)\right].$$

In general, if μ is a probability measure on a $(\mathbb{R}, \mathscr{B})$, then the characteristic function φ_μ of μ is defined by

$$\varphi_\mu(t) := \int_{\mathbb{R}} \exp(-itx)\, \mu(dx) = \int_{\mathbb{R}} \cos(tx)\, \mu(dx) + i \int_{\mathbb{R}} \sin(tx)\, \mu(dx).$$

Note that if μ is the distribution of a RV X (i.e., $\mu = \mathbb{P} \circ X^{-1}$), then the characteristic functions of X and μ as defined above agree by the change of variables formula (see Thm. 1.5.3).

Proposition 4.3.2 (Basic properties of characteristic functions)**.** *Let X be any RV and let φ be its characteristic function. Then the following hold:*

(i) $\varphi(0) = 1$;

(ii) $\varphi(-t) = \overline{\varphi(t)}$;

(iii) $|\varphi(t)| = |\mathbb{E}[\exp(itX)]| \leq \mathbb{E}[|\exp(itX)|] = 1$;

(iv) $|\varphi(t+h) - \varphi(t)| \leq \mathbb{E}\left[|\exp(itX) - 1|\right]$, *so $\varphi(t)$ is uniformly continuous on* $(-\infty, \infty)$;

(v) $\varphi_{aX+b}(t) = \exp(itb)\varphi_X(at)$.

PROOF. (i) Trivial.

(ii) $\varphi(-t) = \mathbb{E}\left[\cos(-tX)\right] + i\mathbb{E}\left[\sin(-tX)\right] = \mathbb{E}\left[\cos(tX)\right] - i\mathbb{E}\left[\sin(tX)\right] = \overline{\varphi(t)}$.

(iii) Let $f(x, y) = \sqrt{x^2 + y^2}$, which is a convex function. By multivariate Jensen's inequality (see Exercise 1.5.13),

$$|\mathbb{E}[\exp(itX)]| = f\left(\mathbb{E}\left[\cos(tX)\right], \mathbb{E}\left[\sin(tX)\right]\right) \leq \mathbb{E}\left[f\left(\cos(tX), \sin(tX)\right)\right] = \mathbb{E}[1] = 1.$$

(iv) Note that

$$|\varphi(t+h)-\varphi(t)| = \left|\mathbb{E}\left[(e^{ihX}-1)e^{itX}\right]\right| \le \mathbb{E}\left[\left|(e^{ihX}-1)e^{itX}\right|\right]$$
$$= \mathbb{E}\left[\left|(e^{ihX}-1)\right|\left|e^{itX}\right|\right] \le \mathbb{E}\left[\left|e^{ihX}-1\right|\right].$$

Since $\left|e^{ihX}-1\right| \le 2$ and $\left|e^{ihX}-1\right| \to 0$ as $h \to 0$, we have $\mathbb{E}\left[\left|e^{ihX}-1\right|\right] \to 0$ as $h \to \infty$ by BCT (see Theorem 1.3.16). Since this convergence does not depend on t, uniform continuity of φ follows.

(e) $\mathbb{E}[\exp(it(aX+b))] = \exp(itb)\mathbb{E}[\exp(iatX)] = \exp(itb)\varphi(at)$.

□

The following factorization property of the characteristic functions is a key to analyzing sums of independent RVs.

Proposition 4.3.3 (Factorization property of the characteristic functions)**.** *Let X, Y be independent RVs on the same probability space. Then $\varphi_{X+Y}(t) = \varphi_X(t)\,\varphi_Y(t)$ for all t.*

PROOF. Note that $\exp(itX)$ and $\exp(itY)$ are independent since X and Y are independent (see Prop. 2.2.6 and Lem. 2.3.2). Hence

$$\varphi_{X+Y}(t) = \mathbb{E}\left[\exp(it(X+Y))\right] = \mathbb{E}\left[\exp(itX)\exp(itY))\right]$$
$$= \mathbb{E}\left[\exp(itX)\right]\mathbb{E}\left[\exp(itY))\right] = \varphi_X(t)\varphi_Y(t).$$

□

Example 4.3.4 (Coin flips)**.** Let X be an RV with $\mathbb{P}(X=1) = \mathbb{P}(X=-1) = 1/2$. Then

$$\varphi_X(t) = \frac{e^{it}+e^{-it}}{2} = \cos(t).$$

▲

Example 4.3.5 (Poisson)**.** Let $X \sim \text{Poisson}(\lambda)$, so $\mathbb{P}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, \ldots$. Then

$$\varphi_X(t) = \sum_{k=0}^{\infty} e^{itk}\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.$$

▲

Example 4.3.6 (Normal)**.** Let $X \sim N(0,1)$, so X has PDF f_X with $f_X(t) = (2\pi)^{-1/2}e^{-t^2/2}$. Then

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{itx}e^{-x^2/2}\,dx$$
$$= \exp(-t^2/2)\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-it)^2}{2}\right)dx.$$

The integral in the last expression should be 1 since it is the integral of the PDF of the normal distribution with 'mean it' and variance 0. This will show

$$\varphi_X(t) = \exp(-t^2/2).$$

(See [Dur19, Ex 3.3.5] for a more rigorous justification.) ▲

Example 4.3.7 (Uniform)**.** Let $X \sim \text{Uniform}((a,b))$, so X has PDF f_X with $f_X(t) = \frac{1}{b-a}\mathbf{1}(x \in (a,b))$. Then

$$\varphi_X(t) = \frac{1}{b-a}\int_a^b e^{itx}\,dx = \frac{e^{itb}-e^{ita}}{it(b-a)}.$$

In particular, if $a = -b$, then $\varphi_X(t) = \frac{e^{itb}-e^{-itb}}{2ib} = \sin(tb)/(bt)$. ▲

Example 4.3.8 (Exponential)**.** Let $X \sim \text{Exp}(\lambda)$, so X has PDF f_X with $f_X(t) = \lambda e^{-\lambda x}\mathbf{1}(x \geq 0)$. Then

$$\varphi_X(t) = \lambda \int_0^\infty e^{itx} e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(it-\lambda)x}\, dx = \lambda \left[\frac{e^{(it-\lambda x)}}{it-\lambda} \right]_0^\infty = \frac{\lambda}{it-\lambda}.$$

In particular, if $a = -b$, then $\varphi_X(t) = \frac{e^{itb}-e^{-itb}}{2ib} = \sin(tb)/(bt)$. ▲

In the following result is a characteristic function version of the well-known Fourier inversion formula. An important corollary of it is that the characteristic function uniquely determines the distribution (see Corollary 4.3.11).

Theorem 4.3.9 (Inversion formula)**.** *Let $\varphi(t) := \int e^{itx}\, \mu(dx)$ where μ is a probability measure. If $a < b$, then*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt = \mu((a,b)) + \frac{\mu(\{a,b\})}{2}.$$

PROOF. Denote

$$I_T := \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt = \int_{-T}^{T} \int \overbrace{\frac{e^{-ita} - e^{-itb}}{it} e^{itx}}^{=:g(t,x)} \mu(dx)\, dt. \tag{34}$$

We would like to apply Fubini's theorem to interchange the integrals in the last expression. For this, observe that

$$\frac{e^{-ita} - e^{-itb}}{it} = \int_a^b e^{-itx}\, dx, \tag{35}$$

so by the multivariate Jensen's inequality (see Exercise 1.5.13),

$$\left| \frac{e^{-ita} - e^{-itb}}{it} \right| \leq \int_a^b |e^{-itx}|\, dx = b - a. \tag{36}$$

Since $|e^{itx}| \leq 1$, it follows that the integrand $g(t,x)$ in the last expression in (34) is uniformly bounded by $b-a$. Since μ is a probability measure, it follows that $\int_{-T}^{T} \int |g(t,x)|\, \mu(dx)dt \leq 2T(b-a) < \infty$. Hence we can apply Fubini's theorem and use the fact that $\cos(t)/t$ is an odd function to get

$$\begin{aligned} I_T &= \int_{\mathbb{R}} \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt\, \mu(dx) \\ &= \int_{\mathbb{R}} \left(\int_{-T}^{T} \frac{\sin(t(x-a))}{t}\, dt - \int_{-T}^{T} \frac{\sin(t(x-b))}{t}\, dt \right) \mu(dx). \end{aligned} \tag{37}$$

Now we introduce the following notations

$$R(\theta, T) := \int_{-T}^{T} \frac{\sin \theta t}{t}\, dt, \qquad S(T) := \int_0^T \frac{\sin t}{t}\, dt, \qquad \text{sgn}(x) = \mathbf{1}(x > 0) - \mathbf{1}(x < 0).$$

Then we have

$$R(\theta, T) = 2\int_0^T \frac{\sin \theta t}{t}\, dt = 2\int_0^{\theta T} \frac{\sin(u)}{u}\, du = 2S(\theta T) = 2\text{sgn}(\theta)\, S(|\theta| T).$$

Note that $\lim_{T\to\infty} S(T) = \pi/2$ by Exercise 4.3.10, so we get $\lim_{T\to\infty} R(\theta, T) = \pi\, \text{sgn}(\theta)$. It follows that

$$\lim_{T\to\infty} R(x-a, T) - R(x-b, T) = R_{a,b}(x) := \begin{cases} 0 & \text{if } x < a \text{ or } x > b \\ 2\pi & \text{if } a < x < b \\ \pi & \text{if } x \in \{a, b\}. \end{cases}$$

Now (37) and BCT yield the desired result as

$$\lim_{T\to\infty} I_T = \int R_{a,b}(x)\, \mu(dx) = 2\pi\mu((a,b)) + \pi\mu(\{a,b\}).$$

□

Exercise 4.3.10. Show that $(x,y) \mapsto e^{-xy} \sin x$ is integrable on $(0,a)\times(0,\infty)$ with respect to the Lebuesgue measure on $\mathbb{R}^2$. Use Fubini's theorem to deduce

$$\int_0^a \frac{\sin x}{x}\,dx = \arctan(a) - (\cos a)\int_0^\infty \frac{e^{-ay}}{1+y^2}\,dy - (\sin a)\int_0^\infty \frac{ye^{-ay}}{1+y^2}\,dy$$

By using the bound $1+y^2 \le 1$, deduce that

$$\left|\int_0^a \frac{\sin x}{x}\,dx - \arctan(a)\right| \le \frac{2}{a}.$$

Further, show that $\lim_{a\to\infty}\int_0^a \frac{\sin x}{x}\,dx = \lim_{a\to\infty}\arctan(a) = \frac{\pi}{2}$.

Corollary 4.3.11. *Let X and Y be RVs and suppose they have the identical characteristic functions. Then $X =_d Y$.*

PROOF. Let μ and ν denote the distributionx of X and Y, respectively. By Theorem 4.3.9,

$$\mu((a,b)) + \frac{\mu(\{a,b\})}{2} = \nu((a,b)) + \frac{\nu(\{a,b\})}{2} \qquad \text{for all } a,b\in\mathbb{R} \text{ with } a<b. \tag{38}$$

Let $A_X := \{x\in\mathbb{R} : \mu(\{x\}) > 0\}$ and $A_Y := \{x\in\mathbb{R} : \nu(\{x\})$ and $B := \mathbb{R}\setminus(A_X\cup A_Y)$. Note that A_X corresponds to the set of discontinuity points of the distribution function of X, $x\mapsto \mathbb{P}(X\le x)$, so it is countable. Likewise, A_Y is countable. The set B consists of the common continuity points of the distribution functions of X and Y. Note that B is dense in $\mathbb{R}$.

Fix $z\in A_x$. Then we may choose $y<z$ with $y\in B$ so that (38) gives

$$\mu((y,z)) + \frac{\mu(\{z\})}{2} = \nu((y,z)) + \frac{\nu(\{z\})}{2}.$$

Letting $y \nearrow z$ with $y\in B$ (we can do this since B is dense in $\mathbb{R}$), and using continuity of probability measures, we deduce $\mu(\{z\}) = \nu(\{z\})$. This holds for all $z\in A_X$. By symmetry, this shows

$$\mu(\{z\}) = \nu(\{z\}) \quad \text{for all } z\in A_X\cup A_Y. \tag{39}$$

Hence, it follows from (38) and (39),

$$\mu((a,b)) = \nu((a,b)) \qquad \text{for all } a,b\in\mathbb{R} \text{ with } a<b.$$

Since the set of open intervals form a π-system and generate the Borel σ-algebra on $\mathbb{R}$, by Lemma 1.1.38, we conclude $\mu=\nu$. □

Exercise 4.3.12. Let X be a RV. If its chatacteristic function φ_X takes values in $\mathbb{R}$, then show that $X =_d -X$.

Exercise 4.3.13 (Sum of independent normals = Normal)**.** Let X_1 and X_2 be independent normal RVs. Use the inversion formula (Theorem 4.3.9) to deduce that X_1+X_2 has the normal distribution with mean $\mathbb{E}[X_1]+\mathbb{E}[X_2]$ and variance $\mathrm{Var}(X_1)+\mathrm{Var}(X_2)$.

When the characteristic function φ of a probability measure μ is integrable, then μ admits a PDF which is given by the inverse Fourier transform of φ.

Proposition 4.3.14 (Inverse Frourier transform of integrable characteristic ft.)**.** *Let μ be a probability measure on $(\mathbb{R},\mathscr{B})$ and let φ be its characteristic function. Suppose $\int|\varphi(x)|\,dx<\infty$. Then μ has a bounded and continuous density function*

$$f(y) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ity}\varphi(t)\,dt.$$

PROOF. We first note that, by the inversion formula (see Thm. 4.3.9), for $a < b$,

$$\mu((a,b)) + \frac{\mu(\{a,b\})}{2} = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt.$$

Let $f_T(x) := \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)\mathbf{1}(-T \le x \le T)$. Then $|f_T(x)| \le (b-a)|\varphi(x)|$ (see (36)), so f_T is dominated by $(b-a)\varphi$, which is integrable by the hypothesis. Hence by DCT, we have

$$\mu((a,b)) + \frac{\mu(\{a,b\})}{2} = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt. \tag{40}$$

Note that the right hand side above is

$$\le \lim_{T\to\infty} \frac{b-a}{2\pi} \int_{-T}^{T} |\varphi(t)|\, dt \le \frac{b-a}{2\pi} \int_{-\infty}^{\infty} |\varphi(t)|\, dt,$$

where the first inequality above uses (36) and the second inequality uses DCT with functions $f_T(x) := |\varphi(x)|\mathbf{1}(-T \le x \le T)$ and the integrability hypothesis of φ. It follows that μ is a continuous measure (i.e., has no point mass, $\mu(\{a\}) = 0$ for all $a \in \mathbb{R}$). Hence from (40), (35), and Fubini's theorem,

$$\begin{aligned}\mu((a,b)) &= \frac{1}{2\pi} \int_{\mathbb{R}} \left(\int_a^b e^{-ity}\, dy \right) \varphi(t)\, dt \\ &= \int_a^b \left(\frac{1}{2\pi} \int_{\mathbb{R}} e^{-ity} \varphi(t)\, dt \right) dy.\end{aligned}$$

The above holds for all open interavals (a,b). Hence by Lemma 1.1.38, it follows that f in the statement is indeed a density function for μ. That f is bounded follows easily from the hypothesis. The continuity of f follows from DCT. Indeed, noting that

$$\begin{aligned}\lim_{h\searrow 0} |f(x+h) - f(x)| &= \lim_{h\searrow 0} \left| \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx}(e^{-ith} - 1)\varphi(t)\, dt \right| \\ &= \lim_{h\searrow 0} \frac{1}{2\pi} \int_{\mathbb{R}} \left| (e^{-ith} - 1)\varphi(t) \right| dt \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \lim_{h\searrow 0} \left| (e^{-ith} - 1)\varphi(t) \right| dt = 0,\end{aligned}$$

where for the second equality, we used DCT for $g_h(x) := \left|(e^{-ith} - 1)\varphi(t)\right| \to 0$ a.s. and $g_h(x) \le 2|\varphi(x)|$. □

4.4. Proof of Central Limit Theorem

Lemma 4.4.1 (Continuity theorem)**.** *Let μ_n, $1 \le n \le \infty$ be probability mesaures on $(\mathbb{R}, \mathscr{B})$ with characteristic functions φ_n.*

(i) *If $\mu_n \Longrightarrow \mu$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all $t \in \mathbb{R}$.*

(ii) *If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.*

PROOF. (i) Let X_n, $1 \le n \le \infty$ be RVs with distribution function $x \mapsto \mu_n((-\infty, x])$ (see Proposition 1.2.13). Fix $t \in \mathbb{R}$. Since $x \mapsto e^{itx}$ is bounded and continuous function, by Proposition 4.2.9,

$$\varphi_n(t) = \mathbb{E}[\exp(itX_n)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[\exp(itX_\infty)] = \varphi_\infty(t).$$

(ii). In order to show tightness of $(\mu_n)_{n\ge 1}$, we first note that by Exercise 4.4.2, the tail of μ_n can be controlled by its characterstic function φ_n as

$$\mu_n(\{x : |x| > 2/u\}) \le u^{-1} \int_{[-u,u]} (1 - \varphi_n(t))\, dt, \quad \text{for all } u > 0. \tag{41}$$

Fix $\varepsilon > 0$. First we observe $\varphi(0) = \lim_{n\to\infty} \varphi_n(0) = 1$. By the hypothesis φ is continuous at 0, so we get $\lim_{t\to\infty} \varphi(t) = 1$. It follows that

$$\lim_{u\to 0} u^{-1} \int_{[-u,u]} (1 - \varphi(t))\, dt = 0.$$

Hence we may choose $u(\varepsilon) > 0$ small enough so that the integral in the LHS above is bounded by $\varepsilon/2$ for all $u < u(\varepsilon)$. Also, note that by BCT, we have

$$\lim_{n\to\infty} u^{-1} \int_{[-u,u]} (1 - \varphi_n(t))\, dt = u^{-1} \int_{[-u,u]} (1 - \varphi(t))\, dt \le \varepsilon/2.$$

Thus we may choose $N(\varepsilon) \ge 1$ such that for all $n \ge N(\varepsilon)$, the integral in the LHS is at most ε. Henec by (41),

$$\mu_n(\{x : |x| > 2/u\}) \le \varepsilon \quad \text{for all } u < u(\varepsilon) \text{ and } n \ge N(\varepsilon).$$

This shows that $(\mu_n)_{n\ge 1}$ is tight.

Now we conclude **(ii)**. Let F_n denote the distribution function of μ for $n \ge 1$. Choose an arbitrary diverging subsequence $(n(k))_{k\ge 1}$. By Helly's selection (see Thm. 4.2.14), there exists a further subsequence $n(k(m))$ such that $F_{n(k(m))} \Rightarrow F$ as $m \to \infty$ for some right-continuous and increasing function F. By tightness and Exercise 4.2.17, F must be a probability distribution function. Let μ denote the probability measure whose distribution function is F (such μ is unique by Lem. 1.1.38). By **(i)**, the characteristic function of μ is φ. Since characteristic function uniquely determines the associated probability measure (see Cor. 4.3.11), it follows that μ does not depend on the subsequence $n(k)$ and the further subsequence $n(k(m))$.

To finish the proof, let $f : \mathbb{R} \to \mathbb{R}$ be a bounded and continuous function. Consider the sequence of real numbers $a_n := \int f\, d\mu_n$. From what we just proved in the above paragraph, we have that every subsequence of a_n has a further subsequence that converges to $a := \int f\, d\mu$. It follows that $a_n \to a$ as $n \to \infty$. Since this holds for all bounded continuous test functions f, by Prop. 4.2.9, we conclude that $\mu_n \Rightarrow \mu$.

□

Exercise 4.4.2 (Tail bound and characteristic function)**.** Let X be a RV and let φ be its characteristic function. Show that for each $u > 0$,

$$\mathbb{P}(|X| \ge 2/u) \le \frac{1}{u} \int_{-u}^{u} (1 - \varphi(t))\, dt.$$

Exercise 4.4.3 (Tail bound of Taylor expansion of complex exponential)**.**

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^n}{m!} \right| \le \frac{|x|^{n+1}}{(n+1)!} \wedge \frac{2|x|^n}{n!}$$

Lemma 4.4.4 (Moments and characteristic function)**.** *Let X be a RV. Then the following hold:*

(i) $\left| \varphi_X(t) - \sum_{m=0}^{n} \frac{\mathbb{E}[(itX)^n]}{m!} \right| \le \mathbb{E}\left[|tX|^{n+1} \wedge 2|tX|^n \right]$

(ii) *If $\mathbb{E}[X^2] < \infty$, then $\varphi_X(t) = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + o(t^2)$.*

PROOF. (i) From Exercise 4.4.3, we have

$$\left| e^{itX} - \sum_{m=0}^{n} \frac{(itX)^n}{m!} \right| \le \frac{|tX|^{n+1}}{(n+1)!} \wedge \frac{2|tX|^n}{n!} \le |tX|^{n+1} \wedge 2|tX|^n$$

almost surely. Taking expectation and using Jensen's inequality then shows the assertion.

(ii) From **(i)**, we have

$$\left| \varphi_X(t) - \left(1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] \right) \right| \le t^2 \mathbb{E}\left[|t||X|^3 \wedge 2|X|^2 \right].$$

Let Y_t denote the RV in the expectation in the RHS above. Then $Y_t \to 0$ a.s. as $t \nearrow 0$, and $|Y_t| \le 2|X|^2$. Since $E[X^2] < \infty$, by DCT, it follows that $\mathbb{E}[Y_t] = o(1)$.

□

Now we are finally ready to prove the classical central limit theorem for i.i.d. increments under second moment assumption.

Theorem 4.4.5 (CLT)**.** *Let $(X_k)_{k\ge1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \ge 1$. Suppose $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 - \mu}{\sigma\sqrt{n}} + \cdots + \frac{X_n - \mu}{\sigma\sqrt{n}}$$

Then $Z_n \Rightarrow Z$ as $n \to \infty$.

PROOF. First notice that we can assume $\mathbb{E}[X_1] = 0$ without loss of generality. Then $\sigma^2 = \mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$. Since the increments X_i's are i.i.d., we have

$$\varphi\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}]\cdots\mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n.$$

By Lemma 4.4.4 and Prop 4.3.3, we have

$$\varphi_{S_n}(t) = \varphi_{X_1}(t)^n = \left(1 - \frac{t^2\sigma^2}{2} + o(t^2)\right)^n.$$

This and Prop. 4.3.2 **(v)** yield

$$\varphi_{Z_n}(t) = \varphi_{X_1}(t)^n = \left(1 - \frac{t^2}{2n} + o(t^2/n^{-1})\right)^n.$$

It is easy to extend Prop. 4.1.2 for sequences of complex numbres. Hence the RHS above converges to $\exp(-t^2/2)$ for all $t \in \mathbb{R}$ as $n \to \infty$. With Example 4.3.6, we conclude that

$$\varphi_{Z_n}(t) \to \exp(-t^2/2) = \varphi_Z(t) \quad \text{for all } t \in \mathbb{R}.$$

Then by the continuity theorem (Lemma 4.4.1), it follows that $Z_n \Rightarrow Z$. □

As a typical application of CLT, we can approximate Binomial(n,p) variables by normal RVs.

Exercise 4.4.6. Let $(X_n)_{n\ge1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - np)/\sqrt{np(1-p)}$. Use CLT to deduce that, as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim$ Binomial(n,p), then

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \Rightarrow Z \sim N(0,1).$$

(iii) From (ii), deduce that have the following approximation

$$\mathbb{P}(Y_n \le x) \approx \mathbb{P}\left(Z \le \frac{x - np}{\sqrt{np(1-p)}}\right),$$

which becomes more accurate as $n \to \infty$.

Exercise 4.4.7. Let $(X_n)_{n\ge1}$ be a sequence of i.i.d. Poisson(λ) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - n\lambda)/\sqrt{n\lambda}$. Show that as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim$ Poisson$(n\lambda)$, then

$$\frac{Y_n - n\lambda}{\sqrt{n\lambda}} \Rightarrow Z \sim N(0,1).$$

(iii) From (ii) deduce that we have the following approximation

$$\mathbb{P}(Y_n \le x) \approx \mathbb{P}\left(Z \le \frac{x - n\lambda}{\sqrt{n\lambda}}\right),$$

which becomes more accurate as $n \to \infty$.

4.4.1. CLT for triangular arrays. The standard CLT (Theorem 4.4.5) applies only for i.i.d. RVs. However, there are some instances where we will have to use CLT for independent but not identically distributed RVs. (e.g., CLT for the difference of sample means $\bar{X} - \bar{Y}$ in Example 4.4.13[1]. In order to handle this, we introduce extentions of CLT for trianglular array of independent RVs.

Consider the following triangular array of RVs:

$$\begin{array}{ll} X_{1;1} & \leftarrow \text{independent} \\ X_{2;1},\, X_{2;2} & \leftarrow \text{independent} \\ X_{3;1},\, X_{3;2},\, X_{3;3} & \leftarrow \text{independent} \\ \vdots & \end{array}$$

Namely, all RVs that appear in the above triangular array are independent, and for each $k \ge 1$, the k^{th} row consists of k RVs $X_{k;1}, \dots, X_{k;k}$ with finite variance. We are interested in the following statement

$$\frac{\left(\sum_{i=1}^n X_{n;i}\right) - \mathbb{E}\left[\sum_{i=1}^n X_{n;i}\right]}{\sqrt{\mathrm{Var}\left(\sum_{i=1}^n X_{n;i}\right)}} \quad \Longrightarrow \quad N(0,1) \quad \text{as } n \to \infty.$$

The above statement is not alway true. A sufficient condition for this is provbided by the following Lindeberg-Feller CLT.

Theorem 4.4.8 (Lindberg-Feller CLT)**.** *For each $n \ge 1$, let $X_{n;1}, \dots, X_{n;m}$ be independent RVs with mean zero. Suppose the following hold:*

(i) *(Limiting variance) There exists $\sigma > 0$ s.t. $\sum_{m=1}^n \mathbb{E}[X_{n;m}^2] \to \sigma^2$ as $n \to \infty$.*

(ii) *(Tightness) For all $\varepsilon > 0$, $\sum_{m=1}^n \mathbb{E}[|X_{n;m}|^2 \mathbf{1}(|X_{n;m}| > \varepsilon)] \to 0$ as $n \to \infty$.*

Let $Z \sim N(0,1)$. Then $X_{n;1} + \cdots + X_{n;n} \Rightarrow \sigma Z$ as $n \to \infty$.

PROOF. Denote $Z_n := X_{n;1} + \cdots + X_{n;n}$. Since the increments $X_{n;i}$'s are independent., we have

$$\varphi_{Z_n}(t) = \prod_{m=1}^n \varphi_{X_{n;m}}(t).$$

By the continuity theorem (Lemma 4.4.1), it suffies to show that

$$\prod_{m=1}^n \varphi_{X_{n;m}}(t) \to \exp(-t^2\sigma^2/2) \quad \text{for all } t \in \mathbb{R}.$$

Denote $\alpha_{n;m} := \varphi_{X_{n;m}}(t)$ and $\beta_{n;m} := 1 - \frac{t^2\mathbb{E}[X_{n;m}^2]}{2}$. Fix $\varepsilon > 0$. Then by Lemma 4.4.4,

$$\begin{aligned} \left|\alpha_{n;m} - \beta_{n;m}\right| &\le \mathbb{E}\left[|tX_{n;m}|^3 \wedge 2|tX_{n;m}|^2\right] \\ &\le \mathbb{E}\left[|tX_{n;m}|^3 \mathbf{1}(|X_{n;m} \le \varepsilon|)\right] + \mathbb{E}\left[2|tX_{n;m}|^2 \mathbf{1}(X_{n;m} > \varepsilon)\right] \\ &\le \varepsilon t^3 \mathbb{E}\left[|X_{n;m}|^2 \mathbf{1}(|X_{n;m} \le \varepsilon|)\right] + 2t^2\mathbb{E}\left[|X_{n;m}|^2 \mathbf{1}(X_{n;m} > \varepsilon)\right] \\ &\le \varepsilon t^3 \mathbb{E}\left[X_{n;m}^2\right] + 2t^2\mathbb{E}\left[|X_{n;m}|^2 \mathbf{1}(X_{n;m} > \varepsilon)\right]. \end{aligned}$$

Summing over $m = 1, \dots, n$ and using hypothesis **(ii)**, we get

$$\limsup_{n\to\infty} \sum_{m=1}^n |\alpha_{n;m} - \beta_{n;m}| \le \varepsilon t^3 \sigma^2.$$

[1] Also in Wilcoxon's signed-rank test in statistics

Then by letting $\varepsilon \searrow 0$, we see that the limsup in the LSH above is zero for each $t \in \mathbb{R}$.

Next, recall that for each fixed $t \in \mathbb{R}$ $|\varphi_{Z_n}(t)| \le 1$ (see Prop 4.3.2). Also, hypothesis **(i)** yields that for large enough $n \ge 1$, $|1 - \frac{t^2\mathbb{E}[X_{n;m}^2]}{2}| \le 1$. Hence by Prop. 4.4.9 with $\theta = 1$, we deduce

$$\lim_{n\to\infty} \left| \varphi_{Z_n}(t) - \prod_{m=1}^{n} \left(1 - \frac{t^2\mathbb{E}[X_{n;m}^2]}{2}\right) \right| \le \lim_{n\to\infty} \sum_{m=1}^{n} |\alpha_{n;m} - \alpha_{n;m}| = 0.$$

Thus, in order to conclude, it is enough to show that

$$\lim_{n\to\infty} \prod_{m=1}^{n} \left(1 - \frac{t^2\mathbb{E}[X_{n;m}^2]}{2}\right) = \exp(-t^2\sigma^2/2). \tag{42}$$

Let $c_{n;m} := -t^2\mathbb{E}[X_{n;m}^2]/2$. By the hypothesis **(i)**, we have $\lim_{n\to\infty} \sum_{m=1}^{n} c_{n;m} = \sigma^2$. Hence by Exercise 4.1.5, we will have (42), provided that we check the following conditions:

$$\max_{1\le k\le n} |c_{k,n}| = o(1), \quad \sup_{n\ge 1} \sum_{k=1}^{n} |c_{k,n}| < \infty. \tag{43}$$

The second condition follows directly from the hypothesis **(i)**. For the first condiion, fix $\varepsilon > 0$ and write

$$\begin{aligned}
\mathbb{E}[X_{n;m}^2] &\le \mathbb{E}[X_{n;m}^2 \mathbf{1}(|X_{n;m}| \le \varepsilon)] + \mathbb{E}[X_{n;m}^2 \mathbf{1}(|X_{n;m}| > \varepsilon)] \\
&= \varepsilon^2 + \mathbb{E}[X_{n;m}^2 \mathbf{1}(|X_{n;m}| > \varepsilon)] \\
&= \varepsilon^2 + \sum_{m=1}^{n} \mathbb{E}[X_{n;m}^2 \mathbf{1}(|X_{n;m}| > \varepsilon)].
\end{aligned}$$

According to the hypothesis **(ii)**, this shows

$$\limsup_{n\to\infty} \max_{1\le m\le n} \mathbb{E}[X_{n;m}^2] \le \varepsilon^2.$$

By letting $\varepsilon \searrow 0$, we see that the limsup in the LHS above is zero. This yields the first condition in (43), as desired. □

Proposition 4.4.9 (Comparing product of complex numbers)**.** *Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be complex numbers of modulus at most θ. Then*

$$\left| \prod_{m=1}^{n} a_m - \prod_{m=1}^{n} b_m \right| \le \theta^{n-1} \sum_{m=1}^{n} |a_m - b_m|.$$

PROOF.

$$\begin{aligned}
\left| \prod_{m=1}^{n} a_m - \prod_{m=1}^{n} b_m \right| &\le \left| a_1 \prod_{m=2}^{n} a_m - a_1 \prod_{m=2}^{n} b_m \right| + \left| a_1 \prod_{m=2}^{n} b_m - b_1 \prod_{m=2}^{n} b_m \right| \\
&\le \theta \left| \prod_{m=2}^{n} a_m - \prod_{m=2}^{n} b_m \right| + |a_1 - b_1| \theta^{n-1}.
\end{aligned}$$

Then the assertion follows by an induction. □

Exercise 4.4.10 (Lyapunov's CLT)**.** Suppose we have a triangular array of independent RVs, $X_{n;1}, \dots, X_{n;n}$ for $n \ge 1$. Suppose the *Lyapunov's condition* holds: There exists $\delta > 0$ s.t.

$$\frac{\sum_{i=1}^{n} \mathbb{E}\left[|X_{n;i} - \mathbb{E}[X_{n;i}]|^{2+\delta}\right]}{\sqrt{\sum_{k=1}^{n} \mathrm{Var}(X_{n;i})}^{2+\delta}} \to 0 \quad \text{as } n \to \infty. \tag{44}$$

Then show that the hypothesis of Lindeberg-Feller CLT in Thm. 4.4.8 holds for the normalized triangular array ($S_n = X_{n;1} + \cdots + X_{n;1}$.)

$$\frac{X_{n;1}}{\sqrt{\mathrm{Var}(S_n)}}, \dots, \frac{X_{n;1}}{\sqrt{\mathrm{Var}(S_n)}}.$$

4.4.2. Applications of CLT for confidence intervals. Suppose we have n i.i.d. RVs $X_1, \dots, X_n$ with mean μ and finite second moment. We are interested in the deviation probability $\mathbb{P}(|\bar{X} - \mu| > \varepsilon)$, which will give a confidence interval for the population mean μ using the sample mean $\bar{X} := n^{-1} S_n$ as its estimator. A typical practice in statistics is to "use CLT" as

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = \mathbb{P}\left(|Z_n| > \varepsilon\sqrt{n}\right) \overset{\text{"CLT"}}{\approx} \mathbb{P}\left(|Z| > \varepsilon\sqrt{n}\right),$$

where $Z_n = n^{-1}\sum_{k=1}^{n}(X_k - \mu)$ and $Z \in N(0,1)$. However, the approximation step marked as "CLT" does not make sense, since CLT holds when the sample size $n \to \infty$ but the last expression compares with $|Z|$ with $\varepsilon\sqrt{n} \to \infty$ as $n \to \infty$. A cleaner approach is to start from the CLT statement $\mathbb{P}(|Z_n| \le z) \to \mathbb{P}(Z \le z)$ as $n \to \infty$ and manipulate the events to match up with the desired event involving $\bar{X}$, as discussed in Exercise 4.4.11.

Example 4.4.11 (Confidence interval for the mean with known variance)**.** Let $(X_t)_{t\ge 0}$ be a sequence of i.i.d. RVs with unknown mean μ but known variance σ^2. Here we do not know if X_i's are drawn from a normal distribution. By CLT, we have

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Longrightarrow Z \sim N(0,1),$$

as the sample size $n \to \infty$. In other words, Z_n becomes more and more likely to be a standard normal RV, so for any $z \in \mathbb{R}$

$$\mathbb{P}(Z_n \le z) \approx \mathbb{P}(Z \le z)$$

when n is large enough.

From this we can construct a *confidence interval* for μ similarly as in Example 4.4.13. Namely,

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \approx 1 - \alpha. \tag{45}$$

Rewriting, this gives

$$\mathbb{P}\left(\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) \approx 1 - \alpha.$$

That is, the probability that the random interval $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ containing the population mean μ is approximately $1-\alpha$. In this sense, this random interval is called the $100(1-\alpha)\%$ *confidence interval* for μ. For instance, noting that $z_{0.05/2} = 1.96$, $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is a 95% confidence interval for μ. ▲

Exercise 4.4.12. Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of $\mathrm{Var}(X) = 1296$. If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $\bar{x} = 1478$ hours, what is the corresponding 95% confidence interval for μ?

Example 4.4.13 (Confidence interval for the difference of two means)**.** Let X and Y be independent RVs such that $\mathbb{E}[X] = \mu_X$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathbb{E}[Y] = \mu_Y$, and $\mathrm{Var}(Y) = \sigma_Y^2$. Suppose we do not know if X and Y have normal distribution but we know σ_X and σ_Y. Then by using CLT, all our confidence interval estimates in the previous example hold asymptotically. Namely, by CLT,

$$\bar{X} \Rightarrow N(\mu_X, \sigma_X^2/n), \qquad \bar{Y} \Rightarrow N(\mu_X, \sigma_Y^2/m),$$

as $n, m \to \infty$. As all samples are independent, we should have, as $n, m \to \infty$,

$$\bar{X} - \bar{Y} \Rightarrow N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

(See Exercise 4.4.14 for a rigorous justification of the above). Thus when n and m are large, $\bar{X}-\bar{Y}$ nearly follows normal distribution above. This implies the following $100(1-\alpha)\%$ (approximate) confidence interval for $\mu_X-\mu_Y$:

$$\mathbb{P}\left(\mu_X-\mu_Y \in \left[(\bar{X}-\bar{Y})-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}, (\bar{X}-\bar{Y})+z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right]\right) \approx 1-\alpha.$$

For instance, let $n=60$,$m=40$, $\bar{x}=70.1$,$\bar{y}=75.3$, $\sigma_X^2=60$, $\sigma_Y^2=40$, and $1-\alpha=0.90$. Note that $1-\alpha/2=0.95=\mathbb{P}(Z\le 1.645)$ for $Z\sim N(0,1)$. Hence $z_{\alpha/2}=1.645$, so

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}} = 1.645\sqrt{\frac{60}{60}+\frac{40}{40}} = 1.645\cdot\sqrt{2} = 2.326.$$

Noting that $\bar{x}-\bar{y}=-5.2$, we conclude that $[-5.2-2.326, -5.2+2.326]=[-7.526,-2.87]$ is a 90% confidence interval for $\mu_X-\mu_Y$. In particular, we can claim that, with at least 90% confidence, μ_Y is larger than μ_X by at least 2.87. ▲

Exercise 4.4.14 (Lyapunov's CLT for the Difference of sample means)**.** Let X,Y be two RVs with $\mathbb{E}[X]=\mu_X$, $\mathbb{E}[Y]=\mu_Y$, $\sigma_X^2:=\mathrm{Var}(X)<\infty$, and $\sigma_Y^2:=\mathrm{Var}(Y)<\infty$. Let $X_1,\dots,X_n$ be i.i.d. samples of X and let $Y_1,\dots,Y_m$ be i.i.d. samples of Y. Further assume that X_i and Y_j are all independent. The goal is this exercise is the prove that the difference $\bar{X}-\bar{Y}$ of the sample means satisfy the CLT, that is,

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \quad\Longrightarrow\quad N(0,1) \quad \text{as } n,m\to\infty. \tag{46}$$

The standard CLT for i.i.d. RVs does not apply since X_i's and Y_j's have possibly different distribution. We will use a more general a particular instance of Lindeberg-Feller CLT for triangular arrays using Lyapunov's condition (see Exercise 4.4.10).

(i) Show that $\mathbb{E}[\bar{X}-\bar{Y}]=\mu_X-\mu_Y$ and $\mathrm{Var}(\bar{X}-\bar{Y})=\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}$.

(ii) Fix sequences $(n_N)_{N\ge1}$, $(m_N)_{N\ge1}$ such that $n_N+m_N=N$ for all $N\ge1$. In light of Exercise 4.4.10, consider the following 'triangular array' of N RVs:

$$N\text{th row:} \qquad \underbrace{X_1/n_N}_{W_{1;N}}, \underbrace{X_2/n_N}_{W_{2;N}}, \dots, \underbrace{X_{n_N}/n_N}_{W_{n_N;N}}, \underbrace{Y_1/m_N}_{W_{n_N+1;N}}, \underbrace{Y_2/m_N}_{W_{n_N+2;N}}, \dots \underbrace{Y_{m_N}/m_N}_{W_{N;N}}. \tag{47}$$

Show that, for each $\delta>0$,

$$\begin{aligned}\frac{\sum_{i=1}^N \mathbb{E}[|W_{i;N}-\mathbb{E}[W_{i;N}]|^{2+\delta}]}{\sqrt{\sum_{i=1}^n \mathrm{Var}(W_{i;N})}^{2+\delta}} &= \frac{n_N^{-1-\delta}\,\mathbb{E}[|X-\mu_X|^{2+\delta}]+m_N^{-1-\delta}\,\mathbb{E}[|Y-\mu_Y|^{2+\delta}]}{\sqrt{\frac{\sigma_X^2}{n_N}+\frac{\sigma_Y^2}{m_N}}^{2+\delta}} \\ &= \frac{\mathbb{E}[|X-\mu_X|^{2+\delta}]}{\sqrt{n_N^{\frac{1+\delta}{1+\delta/2}-1}\sigma_X^2+(n_N^{\frac{1+\delta}{1+\delta/2}}/m_N)\sigma_Y^2}^{2+\delta}} + \frac{\mathbb{E}[|Y-\mu_Y|^{2+\delta}]}{\sqrt{m_N^{\frac{1+\delta}{1+\delta/2}-1}\sigma_X^2+(m_N^{\frac{1+\delta}{1+\delta/2}}/n_N)\sigma_Y^2}^{2+\delta}}.\end{aligned}$$

(iii) From **(ii)** deduce that the Lyapunov's condition (44) for the triangular array (47) is satisfied for $\delta>0$, n_N, and m_N whenever

$$\mathbb{E}[|X-\mu_X|^{2+\delta}]<\infty, \quad \mathbb{E}[|Y-\mu_Y|^{2+\delta}]<\infty, \quad \lim_{N\to\infty} n_N=\infty, \quad \lim_{N\to\infty} m_N=\infty. \tag{48}$$

(iv) From **(i)**, **(iii)** and Lyapunov's CLT (see Exercise 4.4.10), deduce that (46) holds whenever the first two moment conditions in (48) hold.

4.5. CLT with rate of convergence: Berry-Esséen theorem

What can we say about the rate of convergence in CLT? Obtaining a non-asymptotic convergence bound is theoretically important, but also this question is practically motivated as well. For instance, what is the error of approximation in (45) when we constructed confidence interval? Given $\varepsilon > 0$, can we find a sufficiently large sample size n such that

$$|\mathbb{P}(Z_n \le z) - \mathbb{P}(Z \le z)| \le \varepsilon \,?$$

If the RVs have finite third moment, the following classical result by Berry and Esséen provides such bound.

Theorem 4.5.1 (Berry-Esséen theorem)**.** *Let $X_1, \dots, X_n$ be i.i.d. RVs with zero mean, unit variance, and finite third moment. Let $Z_n := n^{-1/2}(X_1 + \cdots + X_n)$ and $Z \in N(0,1)$. Then for any $a \in \mathbb{R}$,*

$$\mathbb{P}(Z_n \le a) - \mathbb{P}(Z \le z) = O(n^{-1/2}\mathbb{E}[|X_1|^3]),$$

where the implied constant in $O(\cdot)$ is absolute.[2]

PROOF. See Terrence Tao's blog: link. □

In this section, we will introduce a short and elegant proof of a weak version of Theorem 4.5.1 using the so-called "Lindeberg replacement trick".

Theorem 4.5.2 (Berry-Esséen theorem, weak form)**.** *Let $X_1, \dots, X_n$ be i.i.d. RVs with zero mean, unit variance, and finite third moment. Let $Z_n := n^{-1/2}(X_1 + \cdots + X_n)$ and $Z \in N(0,1)$. Then for any smooth function $f : \mathbb{R} \to \mathbb{R}$ with uniformly bounded derivative up to third order,*

$$\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Z)] = O\left(n^{-1/2}\mathbb{E}[|X_1|^3] \sup_{x \in \mathbb{R}} |f'''(x)| \right),$$

where the implied constant in $O(\cdot)$ does not depend on f.

PROOF. A first key step is to decompose a single standard normal RV Z into the sum of n i.i.d. normal RVs. Let $Y_1, \dots, Y_n$ be i.i.d. standard normal RVs and let $W_n := n^{-1}(Y_1 + \cdots + Y_n)$. Note that $W_n =_d Z \sim N(0,1)$ by Exercise 4.3.13. Next, we consider a 'continuous transform' from Z_n to W_n by replacing X_k with Y_k for $k = 1, \dots, n$: (a.k.a. Lindeberg replacement trick)

$$\begin{aligned}
\sqrt{n} Z_{n;0} &:= X_1 + X_2 + \cdots + X_n \\
\sqrt{n} Z_{n;1} &:= Y_1 + X_2 + \cdots + X_n \\
\sqrt{n} Z_{n;2} &:= Y_1 + Y_2 + \cdots + X_n \\
&\vdots \\
\sqrt{n} Z_{n;n} &:= Y_1 + Y_2 + \cdots + Y_n.
\end{aligned}$$

Using the above sequence of intermediate RVs, we consider the following telescoping sum

$$\mathbb{E}[f(Z_n)] - \mathbb{E}[f(Z)] = \mathbb{E}[f(Z_n)] - \mathbb{E}[f(W_n)] = \sum_{k=1}^{n} \mathbb{E}[f(Z_{n;k-1})] - \mathbb{E}[f(Z_{n;k})], \tag{49}$$

where $Z_{n;k-1}$ and $Z_{n;k}$ only differs at the kth summand. For $i = 2, \dots, n-1$, denote

$$S_{n;i} := n^{-1/2}\left(X_1 + X_2 + \cdots + X_{i-1} + Y_{i+1} + \cdots + Y_n\right).$$

Since

$$Z_{n;k-1} = S_{n;k} + n^{-1/2} X_k, \qquad Z_{n;k} = S_{n;k} + n^{-1/2} Y_k,$$

[2] does not depend on z and the distribution of X_1

using the second order Taylor expansion with remainder, we can write

$$\begin{aligned} f(Z_{n;k-1}) &= f(S_{n;k} + n^{-1/2}X_k) \\ &= f(S_{n;k}) + f'(S_{n;k})n^{-1/2}X_k + \frac{f''(S_{n;k})}{2n}X_k^2 + O\Big(n^{-3/2}|X_k|^3 \sup_{x\in\mathbb{R}}|f'''(x)|\Big), \\ f(Z_{n;k}) &= f(S_{n;k} + n^{-1/2}Y_k) \\ &= f(S_{n;k}) + f'(S_{n;k})n^{-1/2}Y_k + \frac{f''(S_{n;k})}{2n}Y_k^2 + O\Big(n^{-3/2}|Y_k|^3 \sup_{x\in\mathbb{R}}|f'''(x)|\Big). \end{aligned}$$

Here the implied constant in $O(\cdot)$ is an absolute constant.

Now we take the expectation on both sides and take the difference. Notice that X_k and Y_k have the same first and second moments. Also note that $S_{n;k}$ is independent from X_k and Y_k. Using Lemma 2.3.2, we get

$$\mathbb{E}\big[f(Z_{n;k-1})\big] - \mathbb{E}\big[f(Z_{n;k})\big] = O\Big(n^{-3/2}\big(\mathbb{E}|Y_k|^3 + \mathbb{E}|Y_k|^3\big)\sup_{x\in\mathbb{R}}|f'''(x)|\Big).$$

Adding the above up for $k = 1,\dots,n$, using (49) and $\mathbb{E}|Y_k|^3 = 3$,

$$\mathbb{E}\big[f(Z_{n;k-1})\big] - \mathbb{E}\big[f(Z_{n;k})\big] = O\Big(n^{-1/2}\big(3 + \mathbb{E}|X_k|^3\big)\sup_{x\in\mathbb{R}}|f'''(x)|\Big).$$

This shows the assertion. □

Remark 4.5.3. Note that Theorem 4.5.1 implies Theorem 4.5.2 by integration by parts. Conversely, Theorem 4.5.2 implies a weaker version of 4.5.1 by upper approximating the indicator function $\mathbf{1}(x \le z)$ by a smooth three-times differentiable function f with uniformly (both in x and z) bounded third order derivative of order $O(\varepsilon^{-3})$ (see Exercise 4.5.4). From the construction of f, we have $\mathbf{1}(x \le z) \le f(x) \le \mathbf{1}(x \le z+\varepsilon)$. Then

$$\mathbb{P}(Z \le z) \le \mathbb{E}[f(Z)], \qquad \mathbb{P}(Z_n \le z) \le \mathbb{E}[f(Z_n)].$$

Also, since $Z \sim N(0,1)$ has bounded PDF, we have

$$\begin{aligned} \mathbb{P}(Z \le z) &\le \mathbb{P}(Z \le z) + (\mathbb{P}(Z \le z+\varepsilon) - \mathbb{P}(Z \le z)) \\ &\le \mathbb{E}[f(Z)] + (\mathbb{P}(Z \le z+\varepsilon) - \mathbb{P}(Z \le z)) = \mathbb{E}[f(Z)] + O(\varepsilon). \end{aligned}$$

A similar argument using lower approximating $\mathbf{1}(x \le z)$ by third order smooth function shows corresponding lower bounds. This and Theorem 4.5.2 show

$$|\mathbb{P}(Z_n \le z) - \mathbb{P}(Z \le z)| = O(\varepsilon) + O\big(n^{-1/2}\mathbb{E}|X_k|^3\varepsilon^{-3}\big).$$

The right hand side can be upper bounded by $C(\varepsilon + n^{-1/2}\varepsilon^{-3})$ for some constant $C > 0$. Differentiating this by ε, we find it is minized at $\varepsilon = O(n^{-1/8})$ with minimum value of order $O(n^{-1/8})$. Hence we deduce

$$|\mathbb{P}(Z_n \le z) - \mathbb{P}(Z \le z)| = O(n^{-1/8}).$$

The implied constant in $O(\cdot)$ above does not depend on z, so we in fact get

$$\sup_{z\in\mathbb{R}}|\mathbb{P}(Z_n \le z) - \mathbb{P}(Z \le z)| = O(n^{-1/8}).$$

Notice that the bound of $O(n^{-1/8})$ on the CLT rate of convergence is slower than $O(n^{-1/2})$ in the full Berry-Esséen theorem (Theorem 4.5.1).

Exercise 4.5.4 (Third order smooth approximation of step function)**.** Fix $z \in \mathbb{R}$ and $\varepsilon > 0$. We will construct a function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \equiv 1$ for $x \le z$, $f(x) \equiv 0$ for $x \ge z+\varepsilon$, and f''' is uniformly bounded of order $O(\varepsilon^{-3})$ and the implied constant does not depend on z.

(i) Fix $a < b$ and consider $g'(x) = \lambda x(x-a)(x-b)$. Then $g(x) = \frac{1}{3}x^3 - \frac{a+b}{2}x^2 + abx + C$ has local maximum at a, local minimum at b, and strictly decreasing on $[a,b]$. Let $a = z$ and $b = z+\varepsilon$. Find λ and C such that $g(z) = 1$, $g(z+\varepsilon) = 0$. Show that $\lambda = O(\varepsilon^{-3})$.

(ii) Take $f = g$ with $a = z$, $b = z+\varepsilon$ and λ as in **(i)**. Show that such f safistifes all required conditions.

CHAPTER 5

Maringales

5.1. Conditional expectation

5.1.1. Definition of conditional expectation. Let X, Y be discrete RVs. Recall that the expectation $\mathbb{E}[X]$ is the 'best guess' on the value of X when we do not have any prior knowledge on X. But suppose we have observed that some possibly related RV Y takes value y. What should be our best guess on X, leveraging this added information? This is called the *conditional expectation of X given $Y = y$*. In the most elementary case, this is defined by

$$\mathbb{E}[X|Y = y] = \sum_x x\,\mathbb{P}(X = x|Y = y),$$

where we are for the moment using the undergraduate definition of conditional probability $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$[1]. This best guess on X given $Y = y$, of course, depends on y. So it is a function in y. Now if we do not know what value Y might take, then we omit y and $\mathbb{E}[X|Y]$ becomes a RV, which is called the *conditional expectation of X given Y*.

Example 5.1.1. Suppose we have a biased coin whose probability of heads is itself random and is distributed as $Y \sim \text{Uniform}([0,1])$. Let's flip this coin n times and let X be the total number of heads. Given that $Y = y \in [0,1]$, we know that X follows $\text{Binomial}(n, y)$ (in this case we write $X|U \sim \text{Binomial}(n, Y)$). So $\mathbb{E}[X|Y = y] = ny$. Hence as a random variable, $\mathbb{E}[X|Y] = nY \sim \text{Uniform}([0, n])$. So the expectation of $\mathbb{E}[X|Y]$ is the mean of $\text{Uniform}([0, n])$, which is $n/2$. This value should be the true expectation of X. ▲

The above example suggests that if we first compute the conditional expectation of X given $Y = y$, and then average this value over all choice of y, then we should get the actual expectation of X. This fact is known as the *law of iterated expectation*:

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

In fact, for any event $A \in \sigma(Y)$, we should have iterated expectation for $Y\mathbf{1}_A$:

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|Y]\mathbf{1}_A].$$

Here is an intuitive reason why the iterated expectation works. Suppose you want to make the best guess $\mathbb{E}[X]$. Pretending you know Y, you can imporove your guess to be $\mathbb{E}[X\,|\,Y]$. Then you admit that you didn't know anything about Y and average over all values of Y. The result is $\mathbb{E}[\mathbb{E}[X\,|\,Y]]$, and this should be the same best guess on X when we don't know anything about Y.

The above introductory discussion suggest the following observations:

(a) Conditional expectation of X given another RV Y should be a RV depending on Y: $\mathbb{E}[X|Y] = g(Y)$.

(b) Iterated expectation should hold: $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|Y]\mathbf{1}_A]$ for all $A \in \sigma(Y)$.

In the measure-theoretic definition of conditional expectation, we condition on a sub σ-algebra instead of another RV, and require iterated expectation as an axiom.

Definition 5.1.2 (Conditional expectation)**.** Let $(\Omega, \mathscr{F}_0, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a $(\mathscr{F}_0 - \mathscr{B})$-measurable RV with $\mathbb{E}[|X|] < \infty$. Suppose $\mathscr{F} \subseteq \mathscr{F}_0$ is a σ-algebra on Ω. The *conditional expectation of F given $\mathscr{F}$*, denoted as $\mathbb{E}[X\,|\,\mathscr{F}]$, is any RV $Y : \Omega \to \mathbb{R}$ such that

[1] This is undefined if $\mathbb{P}(B) = 0$. This is one of the main reasons that requires more advanced measure-theoretic definition of conditional expectation.

(i) (Measurability) Y is $(\mathscr{F}-\mathscr{B})$-measurable;

(ii) (Iterated expectation) For each $A\in\mathscr{F}$, $\mathbb{E}[X\mathbf{1}_A]=\mathbb{E}[Y\mathbf{1}_A]$ (that is, $\int_A X\,d\mathbb{P}=\int_A Y\,d\mathbb{P}$).

We say any RV Y satisfying the above conditions is a *version of* $\mathbb{E}[X\,|\,\mathscr{F}]$. In particular, if $\mathscr{F}=\sigma(W)$ for some $\mathscr{F}_0$-measurable RV W, then we write $\mathbb{E}[X\,|\,\mathscr{F}]=\mathbb{E}[X\,|\,W]$. If $X=\mathbf{1}_B$ for some $B\in\mathscr{F}_0$, then we write $\mathbb{E}[\mathbf{1}_B\,|\,\mathscr{F}]=\mathbb{P}(B\,|\,\mathscr{F})$.

Whenever we define a mathematical object axiomatically, we need to justify its existence and uniqueness.

Proposition 5.1.3 (Existence, uniqueness, and integrability of conditional expectation)**.** *Conditional expectation $\mathbb{E}[X\,|\,\mathscr{F}]$ in Def. 5.1.2 uniquely exists and is integrable.*

PROOF. (Integrability) Suppose Y is a version of $\mathbb{E}[X\,|\,\mathscr{F}]$. We will show that $\mathbb{E}[|Y|]<\infty$. Indeed, let $A:=\{Y>0\}\in\mathscr{F}$ (using **(i)**). Then by **(ii)**,

$$\mathbb{E}[Y\mathbf{1}_A]=\mathbb{E}[X\mathbf{1}_A]\le\mathbb{E}[|X|\mathbf{1}_A]\le\mathbb{E}[|X|]<\infty.$$

Similarly,

$$\mathbb{E}[-Y\mathbf{1}_{A^c}]=\mathbb{E}[-X\mathbf{1}_{A^c}]\le\mathbb{E}[|X|\mathbf{1}_A]\le\mathbb{E}[|X|]<\infty.$$

It follows that

$$\mathbb{E}[|Y|]=\mathbb{E}[Y\mathbf{1}_A]+\mathbb{E}[-Y\mathbf{1}_{A^c}]\le 2\mathbb{E}[|X|]<\infty.$$

(Uniqueness) Suppose Y,Y' are versions of $\mathbb{E}[X\,|\,\mathscr{F}]$. Fix $\varepsilon>0$. Then $A:=\{Y-Y'>\varepsilon\}\in\mathscr{F}$. By **(ii)**,

$$\begin{aligned}0=\mathbb{E}[(X-X)\mathbf{1}_A]&=\mathbb{E}[X\mathbf{1}_A]-\mathbb{E}[X\mathbf{1}_A]\\&=\mathbb{E}[Y\mathbf{1}_A]-\mathbb{E}[Y'\mathbf{1}_A]\\&=\mathbb{E}[(Y-Y')\mathbf{1}_A]\ge\varepsilon\mathbb{P}(A)\ge 0.\end{aligned}$$

This shows $\mathbb{P}(A)=0$. Since $\varepsilon>0$ was arbitrary, this shows that $Y\ge Y'$ a.s.. By symmetry, $Y'\ge Y$ a.s. as well. This shows $Y=Y'$ a.s..

(Existence) The existence of conditional expectation relies on the famous Radon-Nikodym theorem (see Thm 5.1.4). First suppose $X\ge 0$. Denote $\mu=\mathbb{P}$ and define a set function $\nu:\mathscr{F}\to\mathbb{R}$ by

$$\nu(B):=\mathbb{E}[X\mathbf{1}_B],\quad\text{for all } B\in\mathscr{F}.$$

Then ν is a measure on $\mathscr{F}$. Indeed, $\nu(B)\ge 0$ since $X\ge 0$ and countable additivity follows from DCT and linearity of expectation:

$$\begin{aligned}\nu\left(\bigsqcup_{k=1}^{\infty}B_k\right)=\mathbb{E}\left[X\sum_{k=1}^{\infty}\mathbf{1}_{B_k}\right]&=\lim_{n\to\infty}\mathbb{E}\left[X\sum_{k=1}^{n}\mathbf{1}_{B_k}\right]\\&=\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}\left[X\mathbf{1}_{B_k}\right]=\lim_{n\to\infty}\sum_{k=1}^{n}\nu(B_k)=\sum_{k=1}^{\infty}\nu(B_k).\end{aligned}$$

Furthermore, observe that $\nu\ll\mu$ (recall absolute continuity in Def. 1.2.22): If $\mu(B)=\mathbb{P}(B)=0$, then $\mathbb{P}(X\mathbf{1}_B=0)=1$ so $\nu(B)=\mathbb{E}[X\mathbf{1}_B]=0$. Hence by Radon-Nikodym theorem (Thm. 5.1.4), the Radon-Nikodym derivative $Y:=\frac{d\nu}{d\mu}:\Omega\to\mathbb{R}$ exists and is measurable w.r.t. $\mathscr{F}$. Then

$$\mathbb{E}[X\mathbf{1}_B]=\nu(B)=\mathbb{E}[Y\mathbf{1}_B]\quad\text{for all } B\in\mathscr{F}.$$

Thus, the Radon-Nikodym derivative $Y:=\frac{d\nu}{d\mu}$ is a version of $\mathbb{E}[X\,|\,\mathscr{F}]$.

For the general case, write $X=X^+-X^-$. Let $Y^+:=\mathbb{E}[X^+\,|\,\mathscr{F}]$ and $Y^-:=\mathbb{E}[X^-\,|\,\mathscr{F}]$. Then $Y:=Y^+-Y^-$ is $\mathscr{F}$-measurable and is integrable. Also, for all $B\in\mathscr{F}$, by linearity of expectation,

$$\mathbb{E}[X\mathbf{1}_B]=\mathbb{E}[X^+\mathbf{1}_B]-\mathbb{E}[X^-\mathbf{1}_B]=\mathbb{E}[Y^+\mathbf{1}_B]-\mathbb{E}[Y^-\mathbf{1}_B]=\mathbb{E}[Y\mathbf{1}_B].\tag{50}$$

Hence Y is a version of $\mathbb{E}[X\,|\,\mathscr{F}]$.

□

Theorem 5.1.4 (Radon-Nikodym theorem)**.** *Let ν and μ be σ-finite measures on $(\Omega,\mathscr{F})$. Then $\nu \ll \mu$ if and only if there exists a measurable function $f:(\Omega,\mathscr{F})\to(\mathbb{R},\mathscr{B})$ such that*

$$\nu(B)=\int_B f\,d\mu \quad \forall B\in\mathscr{F}.$$

In this case, the function f is denoted as $\frac{d\nu}{d\mu}$ and is called the Radon-Nikodym derivative of ν w.r.t. μ.

PROOF. The "only if" part is the core, and a proof of it can be found in [Dur19, Thm. A.4.8]. For the "if" part, suppose $B\in\mathscr{F}$ with $\mu(B)=0$. Then $\int_B f\,d\mu=0$ by the definition of Lebesgue integral. □

Example 5.1.5. Let $X:(\Omega,\mathscr{F},\mathbb{P})\to(\mathbb{R},\mathscr{B})$ be a continuous RV with probability density function f_X. Consider the induced probability measure ν on $\mathbb{R}$ as $\nu(B):=\mathbb{P}(X\in B)=\int_B f\,d\mu$ for $B\in\mathscr{B}$, where μ denotes the Lebesgue measure on $\mathbb{R}$. Then $\nu\ll\mu$, In this case, $\frac{d\nu}{d\mu}=f$.

Exercise 5.1.6. Let $(\Omega,\mathscr{F}_0,\mathbb{P})$ be a probability space and let $X,X':\Omega\to\mathbb{R}$ be $(\mathscr{F}_0-\mathscr{B})$-measurable RVs that are absolutely integrable. Suppose that $\mathbb{P}(X\mathbf{1}_B=X'\mathbf{1}_B)=1$ for some $B\in\mathscr{F}$, where $\mathscr{F}\subseteq\mathscr{F}_0$ is a σ-algebra on Ω. Show that $\mathbb{E}[X\,|\,\mathscr{F}]=\mathbb{E}[X'\,|\,\mathscr{F}]$ a.s. on B. (*Hint*: Repeat the uniqueness argument in the proof of Prop. 5.1.3.)

5.1.2. Examples. In this section, we connect the measure-theoretic definition of conditional expectaion in Def. 5.1.2 with more familar versions of conditional expectation in some special cases. It is helpful to recall the following interpretation of involved objects:

$$\begin{aligned}\mathscr{F}_0 &= \text{Baseline information}\\ X:(\Omega,\mathscr{F}_0)\to(\mathbb{R},\mathscr{B}) &= \text{RV of interest}\\ \mathscr{F}(\subseteq\mathscr{F}_0) &= \text{Additional information}\\ \mathbb{E}[X\,|\,\mathscr{F}] &= \text{Best guess of } X \text{ given } \mathscr{F}.\end{aligned}$$

Example 5.1.7 (Conditioning on perfect information)**.** If X is measurable w.r.t. $\mathscr{F}$, then $\mathbb{E}[X\,|\,\mathscr{F}]=X$. That is, the best guess of X knowing X is X. Indeed, note that $\mathbb{E}[X\mathbf{1}_B]=\mathbb{E}[X\mathbf{1}_B]$ for all $B\in\mathscr{F}$. Hence the only thing that would prevent X for being a version of $\mathbb{E}[X\,|\,\mathscr{F}]$ is its measurability w.r.t. $\mathscr{F}$. Hence if X is $\mathscr{F}$-measruable, then $\mathbb{E}[X\,|\,\mathscr{F}]=X$. In particular, if $\mathbb{P}(X=c)=1$ for some constant $c\in\mathbb{R}$, then X is always measurable w.r.t. $\mathscr{F}$, so $\mathbb{E}[c\,|\,\mathscr{F}]=c$. ▲

Example 5.1.8 (Conditinoing on no information)**.** Suppose X is independent from $\mathscr{F}$. That is, $\sigma(X)$ and $\mathscr{F}$ are independent. Then by defintion of independence, $\sigma(X)$ and $\sigma(\mathbf{1}_B)=\{\emptyset,B,B^c,\Omega\}$ are independent for all $B\in\mathscr{F}$. By Prop. 2.1.4, it follows that

$$X\perp\mathbf{1}_B \qquad \text{for all } B\in\mathscr{F}.$$

In this case, knowing $\mathscr{F}$ would not help us to improve our current guess of X, so we should have $\mathbb{E}[X\,|\,\mathscr{F}]=\mathbb{E}[X]$. In order to verify this, first note that the constant RV $\mathbb{E}[X]$ is measurable w.r.t. $\mathscr{F}$. Second, for the iterated expectation, we use the independence assumption above as

$$\mathbb{E}[X\mathbf{1}_B]=\mathbb{E}[X]\,\mathbb{E}[\mathbf{1}_B]=\mathbb{E}[\mathbb{E}[X]\mathbf{1}_B] \quad \text{for all } B\in\mathscr{F}.$$

This shows $\mathbb{E}[X\,|\,\mathscr{F}]=\mathbb{E}[X]$.

Conversely, assuming $\mathbb{E}[X\,|\,\mathscr{F}]=\mathbb{E}[X]$, we have

$$\mathbb{E}[X\mathbf{1}_B]=\mathbb{E}[\mathbb{E}[X\,|\,\mathscr{F}]\mathbf{1}(B)]=\mathbb{E}[\mathbb{E}[X]\mathbf{1}_B]=\mathbb{E}[X]\,\mathbb{E}[\mathbf{1}_B] \quad \text{for all } B\in\mathscr{F}.$$

Thus $X\perp\mathscr{F}$. Therefore, we have shown that $\mathbb{E}[X\,|\,\mathscr{F}]$ if and only if $X\perp\mathscr{F}$. ▲

Example 5.1.9 (Conditioning on discrete outcomes)**.** Suppose $B_1,B_2,\dots$ be disjoint events in $\mathscr{F}_0$ that partitions the whole sample space, i.e., $\Omega=\bigsqcup_{k=1}^{\infty}B_k$. Suppose $\mathbb{P}(B_k)>0$ for each $k\geq 1$. Let $\mathscr{F}:=\sigma(\{B_1,B_2,\dots\})$. Then we claim that

$$\mathbb{E}[X\,|\,\mathscr{F}](\omega)=\sum_{k=1}^{\infty}\frac{\mathbb{E}[X\mathbf{1}_{B_k}]}{\mathbb{P}(B_k)}\mathbf{1}(\omega\in B_k). \tag{51}$$

Due to the partitioning, there exists a unique $k \geq 1$ such that $\omega \in B_k$ for each $\omega \in \Omega$, so only one term in the RHS is possibly nonzero. That is, for each $\omega \in \Omega$, we know the unique $k \geq 1$ such that $\omega \in B_k$. In this case, $\mathbb{E}[X \,|\, \mathscr{F}](\omega) = \frac{\mathbb{E}[X\mathbf{1}_{B_k}]}{\mathbb{P}(B_k)}$. In words, the information $\mathscr{F}$ tells us which B_k in the partition each outcome ω lies, and the best guess of X is the average value of X over such B_k.

To verify (51), denote the RV in the RHS of (51) as Y. First note that Y is measurable w.r.t. $\mathscr{F}$ since it is constant over each B_k. Denoting its value on B_k by β_k, then for each Borel $A \subseteq \mathbb{R}$,

$$Y^{-1}(A) = \bigcup_{k\geq 1,\, \beta_k \in A} B_k \in \mathscr{F} = \sigma(\{B_1, B_2, \ldots\}).$$

Second, for the iterated expectation, fix an event $B \in \mathscr{F}$. Note that we can write $B = \bigcup_{k\in I} B_k$ for some $I \subseteq \{1,2,\ldots\}$ (see Exc. 1.1.13). Then we need to verify

$$\mathbb{E}[X\mathbf{1}_B] = \mathbb{E}[Y\mathbf{1}_B] = \mathbb{E}\left[\sum_{k\in I} \frac{\mathbb{E}[X\mathbf{1}_{B_k}]}{\mathbb{P}(B_k)} \mathbf{1}(B_k)\right].$$

If $X \geq 0$, the above holds by MCT:

$$\mathbb{E}\left[\sum_{k\in I} \frac{\mathbb{E}[X\mathbf{1}_{B_k}]}{\mathbb{P}(B_k)} \mathbf{1}(B_k)\right] = \sum_{k\in I} \mathbb{E}\left[\frac{\mathbb{E}[X\mathbf{1}_{B_k}]}{\mathbb{P}(B_k)} \mathbf{1}(B_k)\right] = \sum_{k\in I} \mathbb{E}\left[X\mathbf{1}_{B_k}\right] = \mathbb{E}\left[X \sum_{k\in I} \mathbf{1}_{B_k}\right] = \mathbb{E}[X\mathbf{1}_B]. \tag{52}$$

For the general case, we write $X = X^+ - X^-$ and use the fact that $\mathbb{E}[X \,|\, \mathscr{F}] = \mathbb{E}[X^+ \,|\, \mathscr{F}] - \mathbb{E}[X^- \,|\, \mathscr{F}]$ (see (50)). A degenerate but notable special case is when $\mathscr{F} = \{\emptyset, \Omega\}$. In this case, (51) yields $\mathbb{E}[X \,|\, \mathscr{F}] = \mathbb{E}[X]$. ▲

Example 5.1.10 (Undergraduate conditional probability)**.** Suppose $X = \mathbf{1}_A$ and consider the partition $\Omega = B \cup B^c$ for some event $B \in \mathscr{F}$ with $\mathbb{P}(B) \in (0,1)$. Then (51) yields

$$\mathbb{P}(A \,|\, \mathbf{1}_B) = \mathbb{E}[\mathbf{1}_A \,|\, \sigma(\mathbf{1}_B)] = \begin{cases} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} & \text{on } B \\ \frac{\mathbb{P}(A\cap B^c)}{\mathbb{P}(B^c)} & \text{on } B^c. \end{cases}$$

Denoting $\mathbb{P}(A|C) := \mathbb{P}(A \cap C)/\mathbb{P}(C)$ for all $A, C \in \mathscr{F}$ with $\mathbb{P}(C) \in (0,1)$, we can rewrite the above as

$$\mathbb{P}(A \,|\, \mathbf{1}_B) = \mathbb{E}[\mathbf{1}_A \,|\, \sigma(\mathbf{1}_B)] = \begin{cases} \mathbb{P}(A|B) & \text{on } B \\ \mathbb{P}(A|B^c) & \text{on } B^c. \end{cases}$$

▲

Example 5.1.11 (Conditioning on discrete RV)**.** Let T be a discrete RV on $(\Omega, \mathscr{F}_0, \mathbb{P})$ taking countably many values from $\{t_1, t_2, \ldots\}$. We denote

$$\mathbb{E}[X \,|\, T = t_k] := \frac{\mathbb{E}[X\mathbf{1}(T = t_k)]}{\mathbb{P}(T = t_k)} = \mathbb{E}\left[X \frac{\mathbf{1}(T = t_k)}{\mathbb{P}(T = t_k)}\right]. \tag{53}$$

Note that if X is also a discrete RV taking values x_i for $i \geq 1$, then $X\mathbf{1}(T = t_k)$ takes value x_i with probability $\mathbb{P}(X = x_i,\, T = t_k)$ for $i \geq 1$ and otherwise zero, so by Prop. 1.5.2,

$$\mathbb{E}[X \,|\, T = t_k] = \sum_{i\geq 1} x_i \frac{\mathbb{P}(X = x_i,\, T = t_k)}{\mathbb{P}(T = t_k)} = \sum_{i\geq 1} x_i \mathbb{P}(X = x_i \,|\, T = t_k).$$

The above agrees with the undergraduate formula for conditional expectation of a discrete RV conditional on another discrete RV.

Now (51) with the notation in (53) gives

$$\begin{aligned} \mathbb{E}[X \,|\, T](\omega) &= \sum_{k=1}^{\infty} \frac{\mathbb{E}[X\mathbf{1}(T = t_k)]}{\mathbb{P}(T = t_k)} \mathbf{1}(T(\omega) = t_k) \\ &= \sum_{k=1}^{\infty} \mathbb{E}[X \,|\, T = t_k] \mathbf{1}(T(\omega) = t_k). \end{aligned}$$

The computation in (52) and applying it to $X = X^+ - X^-$ shows

$$\begin{aligned}\mathbb{E}[X] = \mathbb{E}[E[X \mid T]] &= \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{E}[X \mid T = t_k]\mathbf{1}(T = t_k).\right] \\ &= \sum_{k=1}^{\infty} \mathbb{E}[X \mid T = t_k]\mathbb{P}(T = t_k),\end{aligned}$$

which is the usual iterated expectation formula in undergraudate probability for conditioning on discrete RVs. ▲

Example 5.1.12 (Conditional expectation from continuous joint distribution)**.** Suppose we have two RVs X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with joint density $f(x, y)$, that is,

$$\mathbb{P}((X, Y) \in B) = \int_B f(x, y)\, dx\, dy \quad \text{for } B \in \mathscr{B}^2.$$

Fix a function $g : \mathbb{R} \to \mathbb{R}$ and suppose $\mathbb{E}[|g(X)|] < \infty$. Can we guess what is $\mathbb{E}[g(X) \mid Y] = h(Y)$? A naive compuation suggests the following:

$$\mathbb{E}[g(X) \mid Y] = \int_{\mathbb{R}} g(x) f_{X|Y=y}(x)\, dx,$$

where $f_{X|Y=y}(x)$ is the "conditional probability density function for X given $Y = y$". This must be computed as

$$f_{X|Y=y}(x) = \frac{f(x, y)}{\int_{\mathbb{R}} f(x, y)\, dy}.$$

Now that we have a guess, we formally verify it.

Let $h : \mathbb{R} \to \mathbb{R}$ be any function such that

$$h(y) \int_{\mathbb{R}} f(x, y)\, dx = \int_{\mathbb{R}} g(x) f(x, y)\, dx \qquad \text{for all } y \in \mathbb{R}.$$

In particular, whenever $\int_{\mathbb{R}} f(x, y)\, dx > 0$, we have

$$h(y) = \frac{\int_{\mathbb{R}} g(x) f(x, y)\, dx}{\int_{\mathbb{R}} f(x, y)\, dx}.$$

If $\int_{\mathbb{R}} f(x, y)\, dx = 0$, then $h(\cdot, y) = 0$ a.s., so $\int_{\mathbb{R}} g(x) f(x, y)\, dx = 0$. Hence $h(y)$ can be defined to be any number in $\mathbb{R}$ in this case. We claim that

$$\mathbb{E}[g(X) \mid Y] = h(Y). \tag{54}$$

It is customary to denote

$$\mathbb{E}[g(X) \mid Y = y] := h(y).$$

Then we recover the familar iterated expectation formula

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \int_{\mathbb{R}} h(y)\, f_Y(y)\, dy = \int_{\mathbb{R}} \mathbb{E}[X \mid Y = y]\, f_Y(y)\, dy.$$

Now we show (54). Clearly $h(Y) \in \sigma(Y)$. To check iterated expecation, let $A \in \sigma(Y)$. Then $A = Y^{-1}(B)$ for some $B \in \mathscr{B}$. Then we note the following series of identities:

$$\begin{aligned}
\mathbb{E}[h(Y)\mathbf{1}_A] &= \int_\Omega h(Y(\omega))\mathbf{1}(Y(\omega) \in B)\, d\mathbb{P} \\
&\overset{(a)}{=} \int_{\mathbb{R}^2} h(y)\mathbf{1}(y \in B)\, d\mathbb{P} \circ (X,Y)^{-1}(x,y) \\
&\overset{(b)}{=} \int_{\mathbb{R}^2} h(y)\mathbf{1}(y \in B)\, f(x,y)\, dx\, dy \\
&\overset{(c)}{=} \int_{\mathbb{R}} h(y)\mathbf{1}(y \in B) \left(\int_{\mathbb{R}} f(x,y)\, dx \right) dy \\
&\overset{(d)}{=} \int_{\mathbb{R}} \int_{\mathbb{R}} g(x)\mathbf{1}(y \in B) f(x,y)\, dx\, dy \\
&\overset{(e)}{=} \int_{\mathbb{R}^2} g(x)\mathbf{1}(y \in B) d\mathbb{P} \circ (X,Y)^{-1}(x,y) \\
&\overset{(f)}{=} \mathbb{E}[g(X)\mathbf{1}(Y \in B)] \\
&= \mathbb{E}[g(X)\mathbf{1}_A].
\end{aligned}$$

Here, (a) follows from change of variables (Thm. 1.5.3) for the composite RV

$$\omega \mapsto (X(\omega), Y(\omega)) \mapsto h(Y(\omega))\mathbf{1}(Y(\omega) \in B),$$

(b) follows from Exc. 1.5.6, (c) is Fubini's theorem, (d) uses the definition of h, and (e) follows from Exc. 1.5.6, and (f) is another change of variables. Hence we conclude that $h(Y)$ is a version of $\mathbb{E}[g(X) \mid Y]$. ▲

Example 5.1.13 (A discrete RV conditional on a continuous RV)**.** Suppose we have two RVs X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Suppose X is a discrete RV taking values from $\{x_1, x_2, \ldots\}$ and Y has PDF f_Y. Suppose there exists a function $h(x,y)$ such that

$$\mathbb{P}(X = x,\, Y \in B) = \int_B h(x,y) f_Y(y)\, dy \quad \text{for } x \in \mathbb{R} \text{ and } B \in \mathscr{B}.$$

(In this case, we denote $\mathbb{P}(X = x \mid Y = y) := h(x,y)$.) Note that by Fubini's theorem,

$$\int_B f_Y(y)\, dy = \mathbb{P}(Y \in B) = \sum_{k \geq 1} \mathbb{P}(X = x_k,\, Y \in B) = \sum_{k \geq 1} \int_B h(x_k, y) f_Y(y)\, dy = \int_B f_Y(y) \sum_{k \geq 1} h(x_k, y)\, dy.$$

Since this holds for all Borel sets $B \subseteq \mathbb{R}$, we must have $\sum_{k \geq 1} h(x_k, y) = 1$ for all y except for a set of Lebesgue measure zero.

Now we claim that

$$\mathbb{E}[X \mid Y] = \sum_{k \geq 1} x_k\, h(x_k, Y) =: W.$$

Clearly W is measurable w.r.t. $\sigma(Y)$. To check iterated expecation, let $A \in \sigma(Y)$. Then $A = Y^{-1}(B)$ for some $B \in \mathscr{B}$. Then by change of variables and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}[W\mathbf{1}_A] &= \int_B \left(\sum_{k \geq 1} x_k\, h(x_k, y) \right) f_Y(y)\, dy \\
&= \sum_{k \geq 1} x_k \int_B h(x_k, y) f_Y(y)\, dy \\
&= \sum_{k \geq 1} x_k \mathbb{P}(X = x_k,\, Y \in B) \\
&= \mathbb{E}[X\mathbf{1}_A].
\end{aligned}$$

Hence W is a version of $\mathbb{E}[X|Y]$. Lastly, it is customary to denote

$$\mathbb{E}[X \mid Y = y] = \sum_{k\geq 1} x_k\, h(x_k, y).$$

Then we recover the familar iterated expectation formula

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \int_{\mathbb{R}} \sum_{k\geq 1} x_k\, h(x_k, y) f_Y(y)\, dy = \int_{\mathbb{R}} \mathbb{E}[X \mid Y = y]\, f_Y(y)\, dy.$$

▲

Example 5.1.14 (Conditioning on independent RV)**.** Suppose we have two independent RVs X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Fix a measurable function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and suppose $\mathbb{E}[|\varphi(X,Y)|] < \infty$. Let $g(x) := \mathbb{E}_Y[\varphi(x,Y)]$. Then we claim that

$$\mathbb{E}[\varphi(X,Y) \mid X] = g(X) = \mathbb{E}_Y[\varphi(X,Y)].$$

To vefiry, first note that $g(X) \in \sigma(X)$. Second, fix $A \in \sigma(X)$. Then $A = X^{-1}(B)$ for some Borel $B \subseteq \mathbb{R}$. Let $\mu := \mathbb{P} \circ X^{-1}$ and $\nu := \mathbb{P} \circ Y^{-1}$. Since $X \perp Y$, $\mathbb{P} \circ (X,Y)^{-1} = \mu \otimes \nu$ (see Lem. 2.3.1). By change of variables,

$$\begin{aligned}
\mathbb{E}[g(X)\mathbf{1}_A] = \mathbb{E}[g(X)\mathbf{1}(X \in B)] &= \int_{\mathbb{R}} g(x)\mathbf{1}(x \in B)\, \mu(dx) \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \varphi(x,y)\, \nu(dy) \right) \mathbf{1}(x \in B)\, \mu(dx) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(x,y)\mathbf{1}(x \in B)\, \nu(dy)\, \mu(dx) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \varphi(x,y)\mathbf{1}(x \in B)\, d\mathbb{P} \circ (X,Y)^{-1} \\
&= \int_{\Omega} \varphi(X(\omega), Y(\omega))\mathbf{1}(X(\omega) \in B)\, d\mathbb{P}(\omega) \\
&= \mathbb{E}[\varphi(X,Y)\mathbf{1}(X \in B)] \\
&= \mathbb{E}[\varphi(X,Y)\mathbf{1}_A].
\end{aligned}$$

Hence $g(X)$ is a version of $\mathbb{E}[\varphi(X,Y) \mid X]$. ▲

Example 5.1.15 (Reducing redundant information)**.** Suppose X, Y, Z are independent RVs on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Fix a measurable function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ with $\mathbb{E}[|\varphi(X,Z)|] < \infty$ and let $g(x) := \mathbb{E}[\varphi(x,Z)]$.

$$\mathbb{E}[\varphi(X,Z) \mid X, Y] = g(X) = \mathbb{E}[\varphi(X,Z) \mid X].$$

This should be reasonable since $\varphi(X,Z)$ is independent from Y. The second equality above was shown in Ex. 5.1.14 so we only need to justify the first.

Note that $g(X) \in \sigma(X)$. Fix $A \in \sigma(X,Y)$. Then $A = (X,Y)^{-1}(B)$ for some Borel $B \subseteq \mathbb{R}^2$. Let $\mu := \mathbb{P} \circ X^{-1}$, $\lambda := \mathbb{P} \circ X^{-1}$, and $\nu := \mathbb{P} \circ Z^{-1}$. Since X, Y, Z are independent, $\mathbb{P} \circ (X,Y)^{-1} = \mu \otimes \lambda$ and $\mathbb{P} \circ (X,Y,Z)^{-1} =$

$\mu \otimes \lambda \otimes \nu$ (see Lem. 2.3.1). By change of variables and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}[g(X)\mathbf{1}_A] = \mathbb{E}[g(X)\mathbf{1}((X,Y)\in B)] &= \int_{\mathbb{R}^2} g(x)\mathbf{1}((x,y)\in B)\, d\mathbb{P}\circ(X,Y)^{-1} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} g(x)\mathbf{1}((x,y)\in B)\,\mu(dx)\,\lambda(dy) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\varphi(x,z)\,\nu(dz)\right)\mathbf{1}((x,y)\in B)\,\mu(dx)\,\lambda(dy) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}}\varphi(x,z)\mathbf{1}((x,y)\in B)\, d\mathbb{P}\circ(X,Y,Z)^{-1} \\
&= \mathbb{E}[\varphi(X,Z)\mathbf{1}((X,Y)\in B)] \\
&= \mathbb{E}[\varphi(X,Z)\mathbf{1}_A].
\end{aligned}$$

Hence $g(X)$ is a version of $\mathbb{E}[\varphi(X,Z)\,|\,X,Y]$. ▲

Example 5.1.16 (Tail-sum formula for conditional expectation)**.** Let X be a nonnegative RV on a probability space $(\Omega,\mathscr{F}_0,\mathbb{P})$ and let $\mathscr{F}\subseteq\mathscr{F}_0$ be a sub-σ algebra. We claim that the 'tail-sum' formula for the conditional expectation of X holds:

$$\mathbb{E}[X\,|\,\mathscr{F}] = \int_0^\infty \mathbb{P}(X\geq x\,|\,\mathscr{F})\,dx. \tag{55}$$

We will use the unconditional tail-sum formula (Prop. 1.5.10) and definition of conditional expectation to show this. Fix $A\in\mathscr{F}$ and $x>0$. Clearly the RHS of (55) is $\mathscr{F}$-measurable. Note that

$$\begin{aligned}
\mathbb{P}(X\mathbf{1}(A)\geq x) &= \mathbb{E}[\mathbf{1}(X\mathbf{1}(A)\geq x)] \\
&\overset{(a)}{=} \mathbb{E}[\mathbf{1}(X\geq x)\mathbf{1}(A)] \\
&\overset{(b)}{=} \mathbb{E}\left[\mathbb{E}[\mathbf{1}(X\geq x)\,|\,\mathscr{F}]\mathbf{1}(A)\right],
\end{aligned}$$

where (a) uses $X\geq 0$ and $x>0$ and (b) uses the definition of $\mathbb{E}[\mathbf{1}(X\geq x)\,|\,\mathscr{F}]$. Then by using the unconditional tail-sum formula (Prop. 1.5.10) and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}[X\mathbf{1}(A)] &= \int_0^\infty \mathbb{P}(X\mathbf{1}(A)\geq x)\,dx \\
&= \int_0^\infty \mathbb{E}\left[\mathbb{E}[\mathbf{1}(X\geq x)\,|\,\mathscr{F}]\mathbf{1}(A)\right]dx \\
&= \mathbb{E}\left[\int_0^\infty \mathbb{E}[\mathbf{1}(X\geq x)\,|\,\mathscr{F}]\mathbf{1}(A)\,dx\right] \\
&= \mathbb{E}\left[\left(\int_0^\infty \mathbb{P}(X\geq x\,|\,\mathscr{F})\,dx\right)\mathbf{1}(A)\right].
\end{aligned}$$

Since the above holds for all $A\in\mathscr{F}$, we have shown (55). ▲

Example 5.1.17 (Example 5.1.1 revisited)**.** Let $Y\sim\text{Uniform}([0,1])$ and $X\sim\text{Binomial}(n,Y)$. Then $X|Y=y\sim\text{Binomial}(n,y)$ so $\mathbb{E}[X|Y=y]=ny$. Hence

$$\mathbb{E}[X] = \int_0^1 \mathbb{E}[X\,|\,Y=y]\,f_Y(y)\,dy = \int_0^1 ny\,dy = n/2.$$

▲

Example 5.1.18. Let $X_1\sim\text{Exp}(\lambda_1)$ and $X_2\sim\text{Exp}(\lambda_2)$ be independent exponential RVs. We will show that

$$\mathbb{P}(X_1<X_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

using the iterated expectation. Using iterated expectation for probability,

$$\begin{aligned}\mathbb{P}(X_1 < X_2) &= \int_0^\infty \mathbb{P}(X_1 < X_2 \,|\, X_1 = x_1)\lambda_1 e^{-\lambda_1 x_1}\, dx_1 \\ &= \int_0^\infty \mathbb{P}(X_2 > x_1)\lambda_1 e^{-\lambda_1 x_1}\, dx_1 \\ &= \lambda_1 \int_0^\infty e^{-\lambda_2 x_1} e^{-\lambda_1 x_1}\, dx_1 \\ &= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)x_1}\, dx_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}.\end{aligned}$$

▲

Exercise 5.1.19. Consider a post office with two clerks. Three people, A, B, and C, enter simultaneously. A and B go directly to the clerks, and C waits until either A or B leaves, and then she starts getting serviced. Let X_A be the time that A spends at a register, and define X_B and X_B similarly. Compute the probability $\mathbb{P}(X_A > X_B + X_C)$ that A leaves the post office after B and C do so in the following scenarios:

(a) The service time for each clerk is exactly (nonrandom) ten minutes.
(b) The service times are i, independently with probability $1/3$ for $i \in \{1,2,3\}$.
(c) The service times are independent $\text{Exp}(\lambda)$ RVs. You may use the fact that the PDF of $X_B + X_C$ is

$$f_{X_B+Y_C}(z) = \lambda^2 z e^{-\lambda z}\mathbf{1}(z \geq 0).$$

Exercise 5.1.20. Suppose we have a stick of length L. Break it into two pieces at a uniformly chosen point and let X_1 be the length of the longer piece. Break this longer piece into two pieces at a uniformly chosen point and let X_2 be the length of the longer one. Define $X_3, X_4, \cdots$ in a similar way.

(i) Let $U \sim \text{Uniform}([0,L])$. Show that X_1 takes values from $[L/2, L]$, and that $X_1 = \max(U, L-U)$.
(ii) From (i), deduce that for any $L/2 \leq x \leq L$, we have

$$\mathbb{P}(X_1 \geq x) = \mathbb{P}(U \geq x \text{ or } L-U \geq x) = \mathbb{P}(U \geq x) + \mathbb{P}(U \leq L-x) = \frac{2(L-x)}{L}.$$

Conclude that $X_1 \sim \text{Uniform}([L/2, L])$. What is $\mathbb{E}[X_1]$?
(iii) Show that $X_2 \sim \text{Uniform}([x_1/2, x_1])$ conditional on $X_1 = x_1$. That is,

$$\mathbb{P}(X_2 \geq x \,|\, X_1) = \frac{2(X_1 - x)}{X_1} \quad \text{for } X_1/2 \leq x \leq X_1.$$

(*Hint:* Use the results in Ex. 5.1.12.) Using iterated expectation, show that $\mathbb{E}[X_2] = (3/4)^2 L$.
(iv) In general, show that $X_{n+1} \,|\, X_n \sim \text{Uniform}([X_n/2, X_n])$. Conclude that $\mathbb{E}[X_n] = (3/4)^n L$.

5.1.3. Properties of conditional expectation. Conditional expectation enjoys properties analogous to the usual (unconditional) expectation. There are some properties of conditional expectation for which there is no analogue for the unconditional expectation. Most of the properties of conditional expectation are verified by checking the two defining axioms.

Proposition 5.1.21 (Basic properties of conditional expectation)**.** *Suppose we have two RVs X, Y on the same probability space $(\Omega, \mathscr{F}_0, \mathbb{P})$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be a sub-σ-albegra. Suppose $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|Y|] < \infty$ for the first two properties below.*

(i) *(Linearity)* $\mathbb{E}[aX + Y \,|\, \mathscr{F}] = a\mathbb{E}[X \,|\, \mathscr{F}] + \mathbb{E}[Y \,|\, \mathscr{F}]$.

(ii) *(Monotonicity)* *If $X \leq Y$, then $\mathbb{E}[X \,|\, \mathscr{F}] \leq \mathbb{E}[Y \,|\, \mathscr{F}]$ a.s..*

(iii) *(Continuity from below)* *If $X_n \geq 0$ and $X_n \nearrow X$ with $\mathbb{E}[X] < \infty$, then $\mathbb{E}[X_n \,|\, \mathscr{F}] \nearrow \mathbb{E}[X \,|\, \mathscr{F}]$.*

PROOF. For the proofs of **(i)** and **(ii)**, denote $W := \mathbb{E}[X \,|\, \mathscr{F}]$ and $Z := \mathbb{E}[Y \,|\, \mathscr{F}]$.

(i) Since W and Z are $\mathscr{F}$-measurable, so is their linear combination $aW+Z$. For iterated expectation, fix $A \in \mathscr{F}$. Then by linearity of expectation,

$$\begin{aligned}\mathbb{E}[(aX+Y)\mathbf{1}_A] &= a\mathbb{E}[X\mathbf{1}_A \,|\, \mathscr{F}] + \mathbb{E}[Y\mathbf{1}_A \,|\, \mathscr{F}] \\ &= a\mathbb{E}[W\mathbf{1}_A \,|\, \mathscr{F}] + \mathbb{E}[Z\mathbf{1}_A \,|\, \mathscr{F}] \\ &= \mathbb{E}[(aW+Z)\mathbf{1}_A].\end{aligned}$$

Hence $aW+Z$ is a version of $\mathbb{E}[aX+Y \,|\, \mathscr{F}]$.

(ii) Fix $\varepsilon > 0$ and let $A := \{W - Z \geq \varepsilon\} \in \mathscr{F}$. Then since $0 \geq (X-Y)\mathbf{1}_A$,

$$0 \geq \mathbb{E}[(X-Y)\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] - \mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[W\mathbf{1}_A] - \mathbb{E}[Z\mathbf{1}_A] = \mathbb{E}[(W-Z)\mathbf{1}_A] \geq \varepsilon \mathbb{P}(A).$$

It follows that $\mathbb{P}(W - Z \geq \varepsilon) = \mathbb{P}(A) = 0$. By taking $\varepsilon \searrow 0$, we deduce $\mathbb{P}(W - Z > 0) = 0$. Hence $W \leq Z$ a.s., as desired.

(iii) Let $Y_n := X - X_n$. Then $Y_n \searrow 0$ a.s.. By **(i)**-**(ii)**, it suffices to show that $Z_n := \mathbb{E}[Y_n \,|\, \mathscr{F}] \searrow 0$. Since $Y_n \searrow$, by **(ii)** we also have $Z_n \searrow$. Also, $Y_n \geq 0$ and **(ii)** imply $Z_n \geq 0$. Hence for each $\omega \in \Omega$, $Z_n(\omega)$ is a decreasing sequence of nonnegative reals, so it converges as $n \to \infty$. Hence $Z_n \searrow Z$ for some RV Z measurable with respect to $\mathscr{F}$. So it is enough to show that $Z = 0$ a.s..

For this, fix $A \in \mathscr{F}$ and note that by DCT,

$$\lim_{n\to\infty} \mathbb{E}[Z_n \mathbf{1}_A] = \mathbb{E}[Z\mathbf{1}_A],$$

as $0 \leq Z_n \leq Z_1$ and $\mathbb{E}[|Z_1|] = \mathbb{E}[Z_1] = \mathbb{E}[Y_1] = \mathbb{E}[X] - \mathbb{E}[X_1] \leq \mathbb{E}[X] < \infty$. But since $Y_n \searrow 0$, $0 \leq Y_n \leq Y_1$, and $\mathbb{E}[Y_1] \leq \mathbb{E}[X] < \infty$, by DCT again, we have

$$\lim_{n\to\infty} \mathbb{E}[Z_n \mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[Y_n \mathbf{1}_A] = \mathbb{E}[\lim_{n\to\infty} Y_n \mathbf{1}_A] = \mathbb{E}[0] = 0.$$

Hence $\mathbb{E}[Z\mathbf{1}_A] = 0$ for all $A \in \mathscr{F}$. In particular, letting $A := \{Z > 0\} \in \mathscr{F}$, $\mathbb{E}[Z\mathbf{1}_A = \mathbb{E}[Z^+] = 0$, so $Z^+ = 0$ a.s.. Similarly, $Z^- = 0$ a.s., so $Z = 0$ a.s..

□

Exercise 5.1.22 (Jensen's inequality for conditional expectation)**.** Let X be a RV on $(\Omega, \mathscr{F}_0, \mathbb{P})$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be a sub-σ-algebra. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|\varphi(X)|] < \infty$, then show that

$$\varphi(\mathbb{E}[X \,|\, \mathscr{F}]) \leq \mathbb{E}[\varphi(X) \,|\, \mathscr{F}].$$

Furthermore, deduce that conditional expectation is a contraction in L^p for $p \geq 1$. That is,

$$\|\mathbb{E}[X \,|\, \mathscr{F}]\|_p \leq \|X\|_p,$$

where for a RV Y, $\|Y\|_p := \mathbb{E}[|Y|^p]^{1/p}$.

Proposition 5.1.23 (Conditional expectation and nested σ-algebras)**.** *Fix a probability space $(\Omega, \mathscr{F}_0, \mathbb{P})$, RVs X, Y on it, and sub-σ-algebras $\mathscr{F} \subseteq \mathscr{G} \subseteq \mathscr{F}_0$.*

(i) *If $\mathbb{E}[X \,|\, \mathscr{G}]$ is $\mathscr{F}$-measurable, then $\mathbb{E}[X \,|\, \mathscr{G}] = \mathbb{E}[X \,|\, \mathscr{F}]$.*

(ii) $\mathbb{E}[\mathbb{E}[X \,|\, \mathscr{G}] \,|\, \mathscr{F}] = \mathbb{E}[X \,|\, \mathscr{F}]$.

PROOF. (i) $Y := \mathbb{E}[X \,|\, \mathscr{G}]$ is assumed to be $\mathscr{F}$-measurable. For iterated expectation, fix $A \in \mathscr{F}$. Then $A \in \mathscr{G}$, so $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$. Hence Y is a version of $\mathbb{E}[X \,|\, \mathscr{F}]$.

(ii) Denote $G := \mathbb{E}[X \,|\, \mathscr{G}]$ and $F := \mathbb{E}[X \,|\, \mathscr{F}]$. Then $F \in \mathscr{F}$. For each $A \in \mathscr{F}$, $A \in \mathscr{G}$ as well, so

$$\mathbb{E}[G\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[F\mathbf{1}_A].$$

Hence F is a version of $\mathbb{E}[G \,|\, \mathscr{F}]$.

□

Proposition 5.1.24 (Conditional expectation with known factor)**.** *Fix a probability space* $(\Omega, \mathscr{F}_0, \mathbb{P})$*, RVs* X, Y *on it, and sub-*σ*-algebra* $\mathscr{F} \subseteq \mathscr{F}_0$*. If* $X \in \mathscr{F}$*, then*

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

PROOF. Clearly $X\mathbb{E}[Y \mid \mathscr{F}]$ is $\mathscr{F}$-measurable by the hypothsis. To check iterated expectation, wish to show that for any $A \in \mathscr{F}$,

$$\mathbb{E}[X\mathbb{E}[Y \mid \mathscr{F}]\mathbf{1}_A] = \mathbb{E}[XY\mathbf{1}_A]. \tag{56}$$

We verify the above following the standard pipeline. First assume $X = \mathbf{1}_B$ for some $B \in \mathscr{F}$. Then $A \cap B \in \mathscr{F}$, so

$$\mathbb{E}[\mathbf{1}_B\mathbb{E}[Y \mid \mathscr{F}]\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[Y \mid \mathscr{F}]\mathbf{1}_{A\cap B}] = \mathbb{E}[Y\mathbf{1}_{A\cap B}] = \mathbb{E}[\mathbf{1}_B Y\mathbf{1}_A].$$

By linearity of expectation, (56) also holds for X simple functions. If $X, Y \geq 0$ and if $X_n \nearrow X$, where X_n is a simple function, then by MCT (note that $Y \geq 0$ implies $\mathbb{E}[Y \mid \mathscr{F}] \geq 0$ so $X_n\mathbb{E}[Y \mid \mathscr{F}] \nearrow X\mathbb{E}[Y \mid \mathscr{F}]$),

$$\mathbb{E}[X\mathbb{E}[Y \mid \mathscr{F}]\mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[X_n\mathbb{E}[Y \mid \mathscr{F}]\mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[X_n Y\mathbf{1}_A] = \mathbb{E}[XY\mathbf{1}_A].$$

For the general case, decompose $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ and use linearity of expectation and conditional expectation. □

Proposition 5.1.25 (Conditional expectation as a projection)**.** *Suppose* $\mathbb{E}[X^2] < \infty$ *and fix a sub-*σ*-algebra* $\mathscr{F} \subseteq \mathscr{F}_0$*. Then* $\mathbb{E}[X \mid \mathscr{F}]$ *is the closest* $\mathscr{F}$*-measurable RV to* X *in the sense that*

$$\mathbb{E}[X \mid \mathscr{F}] = \operatorname*{argmin}_{Y:RV, Y\in\mathscr{F}} \mathbb{E}\left[|X - Y|^2\right]$$

PROOF. Suppose Y is a $\mathscr{F}$-measurable RV and let $Z = \mathbb{E}[X \mid \mathscr{F}] - Y$. Then

$$\begin{aligned}
\mathbb{E}\left[(X-Y)^2\right] &= \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}] + Z)^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2 + 2Z(X - \mathbb{E}[X \mid \mathscr{F}]) + Z^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2\right] + 2\mathbb{E}\left[Z(X - \mathbb{E}[X \mid \mathscr{F}])\right] + \mathbb{E}\left[Z^2\right].
\end{aligned}$$

Note that Z is $\mathscr{F}$-measurable, so by Prop. 5.1.24,

$$\begin{aligned}
\mathbb{E}\left[Z(X - \mathbb{E}[X \mid \mathscr{F}])\right] &= \mathbb{E}[ZX] - \mathbb{E}[Z\mathbb{E}[X \mid \mathscr{F}]] \\
&= \mathbb{E}[ZX] - \mathbb{E}[\mathbb{E}[ZX \mid \mathscr{F}]] = \mathbb{E}[ZX] - \mathbb{E}[ZX] = 0.
\end{aligned}$$

It follows that

$$\mathbb{E}\left[(X-Y)^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2\right] + \mathbb{E}\left[Z^2\right].$$

In order to minimize the RHS above, we need to make $Z = 0$ a.s., so it is minimized at $Y = \mathbb{E}[X \mid \mathscr{F}]$. (Note that the minimum value of the RVS equals $\mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2\right]$, which is also called the *conditional variance* $\mathrm{Var}(X \mid \mathscr{F})$.) □

Exercise 5.1.26 (Markov's inequality)**.** Let X be a RV on $(\Omega, \mathscr{F}_0, \mathbb{P})$ with $X \geq 0$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be a sub-σ-algebra. Show that for each $a > 0$,

$$\mathbb{P}(X \geq a \mid \mathscr{F}) \leq a^{-1}\mathbb{E}[X \mid \mathscr{F}].$$

Exercise 5.1.27 (Chebyshev's inequality)**.** Let X be a RV on $(\Omega, \mathscr{F}_0, \mathbb{P})$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be a sub-σ-algebra. Show that for each $a > 0$,

$$\mathbb{P}(|X| \geq a \mid \mathscr{F}) \leq a^{-2}\mathbb{E}[X^2 \mid \mathscr{F}].$$

Exercise 5.1.28 (Cauchy-Schwarz inequality)**.** Let X, Y be RVs on $(\Omega, \mathscr{F}_0, \mathbb{P})$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be a sub-σ-algebra. Show that

$$\mathbb{E}[XY \mid \mathscr{F}]^2 \leq \mathbb{E}[X^2 \mid \mathscr{F}]\mathbb{E}[Y^2 \mid \mathscr{F}].$$

Exercise 5.1.29 (Bias-Variance decomposition)**.** Let X, Y be RVs on $(\Omega, \mathscr{F}_0, \mathbb{P})$ and let $\mathscr{G} \subseteq \mathscr{F} \subseteq \mathscr{F}_0$ be sub-σ-algebras. Show that

$$\mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{G}])^2\right] = \mathbb{E}\left[(\mathbb{E}[X \mid \mathscr{F}] - \mathbb{E}[X \mid \mathscr{G}])^2\right] + \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2\right].$$

In particular, if $\mathscr{G} = \{\varnothing, \Omega\}$, then

$$\begin{aligned} \underbrace{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}_{\text{MSE}} &= \underbrace{\mathbb{E}\left[\left(\mathbb{E}[X \mid \mathscr{F}] - \mathbb{E}[X]\right)^2\right]}_{\text{variance}} + \mathbb{E}\left[\left(\underbrace{X - \mathbb{E}[X \mid \mathscr{F}]}_{\text{bias}}\right)^2\right] \\ &\geq \mathbb{E}\left[\left(X - \mathbb{E}[X \mid \mathscr{F}]\right)^2\right]. \end{aligned} \tag{57}$$

To put in context, suppose we are estimating a random response X by two estimators $\bar{X} = \mathbb{E}[X]$ and $\hat{X} = \mathbb{E}[X \mid \mathscr{F}]$. Since $\hat{X}$ is the best guess given more information $\mathscr{F}$, it should result in smaller error than the uninformed guess $\bar{X}$. Indeed, the variance term in (57) is nonnegative so this speculation is justified.

Exercise 5.1.30 (Law of total variance)**.** Let X be RVs on $(\Omega, \mathscr{F}_0, \mathbb{P})$ with $X \geq 0$ and let $\mathscr{F} \subseteq \mathscr{F}_0$ be sub-σ-algebra. $\mathrm{Var}(X \mid \mathscr{F}) = \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathscr{F}])^2 \mid \mathscr{F}\right]$. Show that

$$\mathrm{Var}(X \mid \mathscr{F}) = \mathbb{E}[X^2 \mid \mathscr{F}] - \mathbb{E}[X \mid \mathscr{F}]^2.$$

Furthermore, show that

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid \mathscr{F})] + \mathrm{Var}(\mathbb{E}[X \mid \mathscr{F}]).$$

5.2. Basics of Martingales

5.2.1. Motivations. Martingale is a class of stochastic processes, whose expected increment conditioned on the past is always zero. If the conditional increment is always nonnegative then the process is called a 'submartingale' and otherwise a 'supermartingale'. Recall that the simple symmetric random walk has this property, since each increment is i.i.d. and has mean zero. Martingales do not assume any kind of independence between increments, but it turns out that we can proceed quite far with just the unbiased conditional increment property.

There are two motivations to study martingales. First, it is valuable for *modeling real-life stochastic procesess in the market*. For instance, let $(X_t)_{t\geq 0}$ be the sequence of observations of the price of a particular stock over time. Suppose that an investor has a strategy to adjust his portfolio $(M_t)_{t\geq 0}$ according to the observation $(X_t)_{t\geq 0}$. Namely,

$$M_t = \text{Net value of portfolio after observing } (X_k)_{0\leq k\leq t}.$$

We are interested in the long-term behavior of the 'portfolio process' $(M_t)_{t\geq 0}$. Martingales provide a very nice framework for this purpose since (1) assuming expected conditional gain being zero is reasonable from the no-arbitrage principle and (2) one cannot expect the increments of the portfolio process are independent.

Second, martingales are exteremely powerful tools for analyzing stochastic processes, for instance showing convergence of certain processes. Recall the elementary fact from real analysis that a monotone decreasing sequence of real numbers bounded from below converges to some limit:

$$a_1 \geq a_2 \geq \cdots > -\infty \quad \Longrightarrow \quad \exists \lim_{n\to\infty} a_n.$$

Can we make sense of a similar monotonicity statement for a sequence of RVs, $(X_n)_{n\geq 1}$?

There are multiple versions that we can think of. The easiest version is to require the above monotonicify for every single sample ω:

$$\forall \omega \in \Omega: \qquad X_1(\omega) \geq X_2(\omega) \geq \cdots > -\infty \quad \Longrightarrow \quad \exists \lim_{n\to\infty} X_n.$$

In this case, the limiting random variable $X = \liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n$ exists almost surely. However, such 'almost sure monotonicity' is not so much different from the real analysis counterpart and hence it is too restrictive to be used in stochastic analysis.

Next, we can many require monotonicity of expectations:

$$\mathbb{E}[X_1] \geq \mathbb{E}[X_2] \geq \cdots > -\infty.$$

In this case, we do have limiting expectation $\lim_{n\to\infty}\mathbb{E}[X_n]$, but not necessarily we also have convergence of the RVs. A simple counterexample would be when $X_n = \pm n$ with equal probabilities. Hence monotonicity of expected values are too weak to draw meaningful conclusions.

It turns out that the sweat spot between flexibility and nice convergence properties is to require monotonicity of conditional expectations:

$$\mathbb{E}[X_2 - X_1 \,|\, \mathscr{F}_1] \leq 0, \quad \mathbb{E}[X_3 - X_2 \,|\, \mathscr{F}_2] \leq 0, \quad \cdots \quad \text{and} \quad \inf_{n\to\infty} X_n > -\infty \qquad \Longrightarrow \quad \exists \lim_{n\to\infty} X_n.$$

Here we need to have an acompanying, growing sequence of σ-algebras $(\mathscr{F}_n)_{n\geq 1}$ called a 'filtration' corresponding (in analogy) to our growning knowledge on the market. In this case, we can ensure that the limiting random variable X_n exists almost surely. This is called the 'martingale convergence theorem', which we will study soon in this section.

5.2.2. Definition and examples. In this section, we will define martingales and their cousins: supermartingales and submartingales, and take the first steps in developing their theory. A *stochastic process*, or a *process* for short, is a sequence $(X_n)_{n\geq 1}$ of RVs. A 'martingale' is a process that models the value of a finantial portfolio at market equilibrium (no arbitragy exists).

Definition 5.2.1 (Martingale)**.** Let $(\mathscr{F}_n)_{n\geq 1}$ be a filtration, i.e., an increasing sequence of σ-fields: $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$. A process $(X_n)_{n\geq 1}$ is said to be *adapted to* $(\mathscr{F}_n)_{n\geq 1}$ if $X_n \in \mathscr{F}_n$ for all n (i.e., X_n is $(\mathscr{F}_n - \mathscr{B})$-measurable). If $X = (X_n)_{n\geq 1}$ is a process with

(i) (*finite expectation*) $\mathbb{E}[|X_n|] < \infty$,
(ii) (*adaptedness*) X_n is adapted to $\mathscr{F}_n$,
(iii) (*conditional increments*) $\mathbb{E}[X_{n+1}|\mathscr{F}_n] = X_n$ for all n,

then X is said to be a *martingale* (w.r.t. $(\mathscr{F}_n)_{n\geq 1}$). X is a *supermartingale* or *submartingale* if the equality in **(iii)** is replaced with $\leq$ or $\geq$, respectively.

Remark 5.2.2 (Harmonic functions and martingales)**.** Let $f : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. It's Laplacian is defined as

$$\Delta(f) := \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2} = \text{tr}(\text{Hessian}(f)).$$

f is called *harmonic* if $\Delta(f) = 0$, *superharmonic* if $\Delta(f) \leq 0$, and *subharmonic* if $\Delta(f) \geq 0$. If f is harmonic, then by mean value property, for every ball $B(x,r) \subseteq \Omega$,

$$f(x) = \frac{1}{\text{Vol}(\partial B(x,r))} \int_{\partial B(x,r)} f(y)\, dy.$$

That is, the function value at the center of a ball equals the average over the sphere. If f is superharmonic, then = above becomes $\geq$; If f is subharmonic, then = above becomes $\leq$. To connect with the definition of martingales, compare the martingale condition

$$X_n = \mathbb{E}[X_{n+1} \,|\, \mathscr{F}_n].$$

Then the equality becomes $\leq$ if $(X_n)_{n\geq 1}$ is a supermartingale, and so on.

Where do the names of superharmonic and subharmonic functions come from? If another function $u : \Omega \to \mathbb{R}$ is harmonic and if $f = u$ on $\partial\Omega$, then $f = u$ by the uniquess of harmonic functions. If f is superharmonic, then $f \geq u$ (f lies above u); if f is subharmonic, then $f \leq u$ (f lies below u). This is easy to see in one dimension. For instance, $f(x) = ax^2$ is superharmonic if $a < 0$ and it is concave up. ▲

Example 5.2.3 (Martingale w.r.t. the filtration generated by a stochastic process)**.** In many cases, we work with filtrations generated by observing a single reference process. Let $(Y_n)_{n\geq 1}$ denote a discrete-time stochastic process on Ω and for each n, let $\mathscr{F}_n = \sigma(Y_1,\dots,Y_n)$ denote the collection of all events that can be determined by the values of $Y_1,\dots,Y_n$. Then $(\mathscr{F}_n)_{n\geq 1}$ is called the *filtration generated by the process* $(Y_n)_{n\geq 1}$.

In order to get familiar with martingales, we provide three examples of martingales related to random walks. Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. increments with $\mathbb{E}[\xi_i] = \mu < \infty$. Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$, where S_0 is a constant. Denote $\mathscr{F}_n := \sigma(\xi_1,\dots,\xi_n)$ and let $\mathscr{F}_0 = \{\emptyset,\Omega\}$. Then $(S_n)_{n\geq 0}$ is called a *random walk*.

Example 5.2.4 (linear martingale from RW)**.** Define a stochastic process $(X_n)_{n\geq 0}$ by

$$X_n := S_n - \mu n.$$

Then $(X_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathscr{F}_n)_{t\geq 0}$. Indeed, note that $X_n \in \mathscr{F}_n$ and

$$\mathbb{E}[|X_n|] = \mathbb{E}[|S_n - \mu t|] \leq \mathbb{E}[|S_n| + |\mu t|] = n\mathbb{E}[|\xi_1|] + \mu t < \infty,$$

and by linearity of conditional expectation,

$$\begin{aligned}
\mathbb{E}[X_{n+1}\,|\,\mathscr{F}_n] &= \mathbb{E}[S_{n+1} - \mu n\,|\,\mathscr{F}_n] \\
&= \mathbb{E}[S_{n+1}\,|\,\mathscr{F}_n] - \mu(n+1) \qquad (\because \mu(n+1) \in \mathscr{F}_n) \\
&= \mathbb{E}[S_n\,|\,\mathscr{F}_n] + \mathbb{E}[\xi_{n+1}\,|\,\mathscr{F}_n] - \mu n \\
&= S_n + \mu n \qquad (\because S_n \in \mathscr{F}_n \text{ and } \mathbb{E}[\xi_{n+1}\,|\,\mathscr{F}_n] = \mathbb{E}[\xi_{n+1}] = \mu) \\
&= X_n.
\end{aligned}$$

From this, we deduce

$$\mathbb{E}[S_{n+1}\,|\,\mathscr{F}_n] = S_n + \mu.$$

Hence if $\mu = 0$, then $(S_n)_{n\geq 0}$ is a martingale; If $\mu \geq 0$, then $(S_n)_{n\geq 0}$ is a submartingale; If $\mu \leq 0$, then $(S_n)_{n\geq 0}$ is a supermartingale. ▲

Example 5.2.5 (Quadratic martingale from RW)**.** Suppose we have the same random walk $(S_n)_{n\geq 0}$ now with increments having mean zero and finite variance $\sigma^2 > 0$. Then we claim that $S_n^2 - \sigma^2 n$ is a martingale w.r.t. the usual filtration $\mathscr{F}_n = \sigma(\xi_1,\dots,\xi_n)$ with $\mathscr{F}_0 = \{\emptyset,\Omega\}$.

Finite expectation and adaptedness are easy to check. For the conditional increment condition, note that since $S_n \in \mathscr{F}_n$ and $\xi_{n+1} \perp \mathscr{F}_n$,

$$\begin{aligned}
\mathbb{E}[S_{n+1}^2\,|\,\mathscr{F}_n] &= \mathbb{E}[S_n^2 + 2S_n\xi_{n+1} + \xi_{n+1}^2\,|\,\mathscr{F}_n] \\
&= \mathbb{E}[S_n^2\,|\,\mathscr{F}_n] + 2S_n\mathbb{E}[\xi_{n+1}\,|\,\mathscr{F}_n] + \mathbb{E}[\xi_{n+1}^2\,|\,\mathscr{F}_n] \\
&= \mathbb{E}[S_n^2\,|\,\mathscr{F}_n] + 2S_n\mathbb{E}[\xi_{n+1}] + \mathbb{E}[\xi_{n+1}^2] \\
&= S_n^2 + \sigma^2.
\end{aligned}$$

Thus S_n^2 is a submartingale with the positive drift $\sigma^2 > 0$[2]. Hence subtracting this linear bias from S_n^2 should make it a martingale. Indeed,

$$\mathbb{E}[S_{n+1}^2 - \sigma^2(n+1)\,|\,\mathscr{F}_n] = S_n^2 - \sigma^2 n.$$

▲

Example 5.2.6 (Product martingale)**.** Let $(\xi_n)_{t\geq 0}$ be a sequence of independent RVs with $\xi_n \geq 0$ and $\mathbb{E}[X_n] = 1$ for all $n \geq 0$. For each $n \geq 0$, let $\mathscr{F}_n = \sigma(\xi_1,\dots,\xi_n)$ and $\mathscr{F}_0 = \{\emptyset,\Omega\}$. Define

$$X_n := X_0\xi_1\xi_2\cdots\xi_n \qquad \text{for } n \geq 1$$

where X_0 is a constant. Then $(X_n)_{n\geq 0}$ is a martingale with respect to $(\mathscr{F}_n)_{n\geq 0}$.

[2]Informally, the variance grows at rate σ^2.

Clearly $X_n \in \mathscr{F}_n$ and by using independence,

$$\mathbb{E}[|X_n|] = \mathbb{E}[X_n] = X_0\mathbb{E}[\xi_1]\cdots\mathbb{E}[\xi_n] = X_0 < \infty.$$

For the martingale condition, since $X_n \in \mathscr{F}_n$ and $\xi_{n+1} \perp \mathscr{F}_n$,

$$\mathbb{E}[X_{n+1} - X_n \,|\, \mathscr{F}_n] = \mathbb{E}[X_n\xi_{n+1} - X_n \,|\, \mathscr{F}_n] = X_n\mathbb{E}[\xi_{n+1} - 1 \,|\, \mathscr{F}_n] = 0.$$

This multiplicative model a reasonable for the stock market since the changes in stock prices are believed to be proportional to the current stock price. Moreover, it also guarantees that the price will stay positive, in comparison to additive models. ▲

Example 5.2.7 (Exponential martingale)**.** Let $(\xi_n)_{t\geq 0}$ be a sequence of i.i.d. RVs such that their moment generating function exists, namely, there exists $\theta > 0$ for which

$$\varphi(\theta) := \mathbb{E}[\exp(\theta\xi_k)] < \infty \qquad \forall k \geq 0.$$

Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$ for S_0 a constant. For each $n \geq 0$, let $\mathscr{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ and $\mathscr{F}_0 = \{\emptyset, \Omega\}$. Define

$$X_n := \exp(\theta S_n)/\varphi(\theta)^n.$$

Then $(X_n)_{n\geq 0}$ is a martingale with respect to filtration $(\mathscr{F}_n)_{n\geq 0}$.

This follows easily from Example 5.2.6. Indeed, let $\eta_n := \exp(\theta\xi_n)/\varphi(\theta)$. These are independent, nonnegative RVs with expectation one. Furthermore,

$$X_n = \exp(\theta(S_0 + \xi_1 + \cdots + \xi_n))/\varphi(\theta)^n = \exp(\theta S_0)\eta_1\cdots\eta_n.$$

So X_n is the product martingale in Example 5.2.6 with multiplicative increments η_n. ▲

The following lemma allows us to provide many examples of martingales from Markov chains (see Def. 6.1.1).

Lemma 5.2.8. *Let $(X_n)_{n\geq 0}$ be a Markov chain on a discrete state space $\mathscr{S}$ with transition matrix P. For each $n \geq 0$, let f_n be a bounded function $\Omega \to \mathbb{R}$ such that*

$$f_n(x) = \sum_{y\in\Omega} P(x,y) f_{n+1}(y) \qquad \forall x \in \Omega.$$

Then $M_n := f_n(X_n)$ defines a martingale with respect to filtration $(\mathscr{F}_n)_{n\geq 0}$, where $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$.

PROOF. Since f_n is bounded, $\mathbb{E}[|f_n(M_n)|] < \infty$. By definition, M_n is given as a function $f_n(X_n)$ of the reference process $(X_n)_{n\geq 0}$, so $M_n \in \mathscr{F}_n$. For the conditional increment property, note that by using the Markov property,

$$\begin{aligned}\mathbb{E}[M_{n+1} - M_n \,|\, \mathscr{F}_n] &= \mathbb{E}[M_{n+1} - M_n \,|\, \mathscr{F}_n] \\ &= \mathbb{E}\left[\left(\sum_{y\in\Omega} P(X_n, y) f_{n+1}(y)\right) - f_n(X_n) \,\middle|\, \mathscr{F}_n\right] = 0.\end{aligned}$$

This shows the assertion. □

Example 5.2.9 (Simple random walk)**.** Let $(\xi_t)_{t\geq 1}$ be a sequence of i.i.d. RVs with

$$\mathbb{P}(\xi_k = 1) = p, \qquad \mathbb{P}(\xi_k = -1) = 1 - p.$$

Let $S_t = S_0 + \xi_1 + \cdots + \xi_t$. Note that $(S_t)_{t\geq 0}$ is a Markov chain on $\mathbb{Z}$. Define

$$M_t = \left(\frac{1-p}{p}\right)^{S_n}.$$

Then $(M_t)_{t\geq 0}$ is a martingale with respect to filtration $\mathscr{F}_n = \sigma(\xi_1, \ldots, \xi_n)$.

In order to see this, define a function $h_n(x) = ((1-p)/p)^x$, which does not depend on n. According to Lemma 5.2.8, it suffice to show that h is a 'harmonic function' with respect to the transition matrix of the RW[3] Namely,

$$\begin{aligned}\sum_{y\in\mathbb{Z}} P(x,y)h_{n+1}(y) &= ph(x+1) + (1-p)h(x-1)\\ &= p\left(\frac{1-p}{p}\right)^{x+1} + (1-p)\left(\frac{1-p}{p}\right)^{x-1}\\ &= (1-p)\left(\frac{1-p}{p}\right)^{x} + p\left(\frac{1-p}{p}\right)^{x} = \left(\frac{1-p}{p}\right)^{x} = h_n(x).\end{aligned}$$

Hence by Lemma 5.2.8, $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathscr{F}_n)_{n\geq 0}$. ▲

Example 5.2.10 (Simple symmetric random walk)**.** Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. RVs with

$$\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = 1/2.$$

Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$. Note that $(S_n)_{n\geq 0}$ is a Markov chain on $\mathbb{Z}$. Define

$$M_n = S_n^2 - n.$$

Then $(M_n)_{n\geq 0}$ is a martingale with respect to $\mathscr{F}_n = \sigma(\xi_1,\ldots,\xi_n)$.

We have already shown this cliam in Example 5.2.5. One can also check the martingale condition for $S_n^2 - n$ using Lemma 5.2.8. Namely, for each $n \geq 0$, define a function $f_n : \mathbb{Z} \to \mathbb{R}$ by $f_n(x) = x^2 - n$. By Lemma 5.2.8, it suffices to check if $f_n(x)$ is the average of $f_{n+1}(y)$ with respect to the transition matrix of S_n. Namely,

$$\begin{aligned}\sum_{y\in\mathbb{Z}} P(x,y)f_{n+1}(y) &= \frac{1}{2}f_{n+1}(x+1) + \frac{1}{2}f_{n+1}(x-1)\\ &= \frac{(x+1)^2-(n+1)}{2} + \frac{(x-1)^2-(n-1)}{2} = x^2 - n = f_n(x).\end{aligned}$$

Hence by Lemma 5.2.8, $(M_n)_{n\geq 0}$ is a martingale with respect to filtration $(\mathscr{F}_n)_{n\geq 0}$. ▲

5.2.3. Basic properties of martingales. If martingale is a fair gambling strategy, then one can think of supermartingale and submartingale as unfavorable and favorable gambling strategies, respectively. For instance, expected winning in gambling on an unfavorable game should be non-increasing in time. This is an immediate consequence of the definition and iterated expectation.

Proposition 5.2.11. *Let $(X_n)_{n\geq 0}$ be a stochastic process adapted to a filtration $(\mathscr{F}_n)_{n\geq 0}$.*

(i) *If $(X_n)_{n\geq 0}$ is a supermartingale w.r.t. $(\mathscr{F}_n)_{n\geq 0}$, then $\mathbb{E}[X_n \mid \mathscr{F}_m] \leq X_m$ for all $n \geq m \geq 0$.*

(ii) *If $(X_n)_{n\geq 0}$ is a submermartingale w.r.t. $(\mathscr{F}_n)_{n\geq 0}$, then $\mathbb{E}[X_n \mid \mathscr{F}_m] \geq X_m$ for all $n \geq m \geq 0$.*

(iii) *If $(X_n)_{n\geq 0}$ is a martingale w.r.t. $(\mathscr{F}_n)_{n\geq 0}$, then $\mathbb{E}[X_n \mid \mathscr{F}_m] = X_m$ for all $n \geq m \geq 0$.*

PROOF. Note that $-X_n$ is a submartingale if X_n is a supermartingale, so (ii) and (iii) follows directly from (i). We will only show (i). Let $(X_n)_{n\geq 0}$ be a supermartingale w.r.t. $(\mathscr{F}_n)_{\geq 0}$, so

$$\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leq X_n.$$

Then by iterated expectation (e.g., Prop. 5.1.23 with $\mathscr{F}_m \subseteq \mathscr{F}_n \subseteq \mathscr{F}_0$),

$$\begin{aligned}\mathbb{E}[X_{m+2} - X_m \mid \mathscr{F}_m] &= \mathbb{E}[\underbrace{\mathbb{E}[X_{m+2} - X_m \mid \mathscr{F}_{m+1}]}_{\leq X_{m+1} - X_m} \mid \mathscr{F}_m]\\ &\leq \mathbb{E}[X_{m+1} - X_m \mid \mathscr{F}_n] \leq 0.\end{aligned}$$

Then the assertion follows by an induction. □

[3] One can also check that M_n is a martingale by Example 5.2.6.

Proposition 5.2.12. *Suppose $(X_n)_{n\geq 0}$ is a martingale w.r.t. filtration $(\mathscr{F}_n)_{n\geq 0}$.*

(i) *If $\phi:\mathbb{R}\to\mathbb{R}$ is a convex function with $\mathbb{E}[|\phi(X_n)|]<\infty$ for all n, then $\phi(X_n)$ is a submartingale w.r.t. $\mathscr{F}_n$.*

(ii) *If $p\geq 1$ and $\mathbb{E}[|X_n|^p]<\infty$ for all n, then $|X_n|^p$ is a submartingale w.r.t. $\mathscr{F}_n$.*

PROOF. By Jensen's inequality (Exc. 5.1.22),

$$\mathbb{E}[\phi(X_{n+1})\,|\,\mathscr{F}_n]\geq \phi\left(\mathbb{E}[X_{n+1}\,|\,\mathscr{F}_n]\right)=\phi(X_n).$$

This shows **(i)**. Noting that $x\mapsto |x|^p$ for $p\geq 1$ is a convex function, **(ii)** follows from **(i)**. □

Proposition 5.2.13. *If X_n is a submartingale w.r.t. $\mathscr{F}_n$ and ϕ is an increasing convex function with $\mathbb{E}[|\phi(X_n)|]<\infty$ for all n, then $\phi(X_n)$ is a submartingale w.r.t. $\mathscr{F}_n$. Consequently: For any $a\in\mathbb{R}$,*

(i) *If X_n is a submartingale, then $(X_n-a)^+$ is a submartingale.*

(ii) *If X_n is a supermartingale, then $X_n\wedge a$ is a supermartingale.*

PROOF. By Jensen's inequality (Exc. 5.1.22) and since ϕ is increasing,

$$\mathbb{E}[\phi(X_{n+1})\,|\,\mathscr{F}_n]\geq \phi\left(\mathbb{E}[X_{n+1}\,|\,\mathscr{F}_n]\right)\geq\phi(X_n).$$

Hence $\phi(X_n)$ is a submartingale. **(i)** is immediate since $x\mapsto (x-a)^+=\max\{0,x-a\}$ is an increasing convex function. Similarly, $x\mapsto (x\wedge a)=\min\{x,a\}$ is a decreasing concave function, so $-(X_n\wedge a)$ is a submartingale. It follows that $(X_n\wedge a)$ is a supermartingale. This shows **(ii)**. □

An important notion in the theory of stochastic processes is that of 'stopping time'.

Definition 5.2.14 (Stopping time)**.** Let $(\mathscr{F}_n)_{n\geq 0}$ be a filtration. A nonnegative integer-valued random variable N is a *stopping time* w.r.t. $(\mathscr{F}_n)_{n\geq 0}$ if for every $n\in\mathbb{Z}_{\geq 0}$, $\{N=n\}\in\mathscr{F}_n$.[4]

Typical examples of stopping times are 'hitting times'.

Example 5.2.15 (First hitting time is a stopping time)**.** Let $(X_n)_{n\geq 0}$ be a stochastic process defined on a measurable space $(\mathscr{S},\mathscr{G})$ adapted to a fltration $(\mathscr{F}_n)_{n\geq 0}$. Let $A\subseteq\mathscr{S}$ be a measurable subset of the state space. We will define τ_A to be the 'first hitting time' of A, which is the first time that the process X_n enters into states in A:

$$\tau_A:=\inf\{n\geq 0\,|\,X_n\in A\}.$$

Here we use the convention of $\inf\emptyset=\infty$ so that $\tau_A=\infty$ if X_n never enters A. In order to see that τ_A is indeed a stopping time, note that for each $n\geq 0$,

$$\{\tau_A=n\}=\{X_0\in A^c,X_1\in A^c,\ldots,X_{n-1}\in A^c,X_n\in A\}\in\mathscr{F}_n.$$

Similarly, define T_A to be the first time after time zero to hit some state in A:

$$T_A:=\inf\{n\geq 1\,|\,X_n\in A\}.$$

As before, T_A is also a stopping time since, for each $n\geq 0$,

$$\{T_A=n\}=\{X_1\in A^c,X_1\in A^c,\ldots,X_{n-1}\in A^c,X_n\in A\}.$$

Hence, in general, first hitting times are stopping times. ▲

Exercise 5.2.16 (kth hitting time)**.** Let $(X_n)_{n\geq 0}$ be a stochastic process defined on a measurable space $(\mathscr{S},\mathscr{G})$ adapted to a fltration $(\mathscr{F}_n)_{n\geq 0}$. Let $A\subseteq\mathscr{S}$ be a measurable subset of the state space. For each $k\geq 1$, let $T_A^{(k)}$ denote the kth time that the process X_n visits some state in A. That is, $T_A^{(0)}=0$ and for $m\geq 1$,

$$T_A^{(m)}=\begin{cases}\inf\{n>T_A^{(m-1)}:X_n\in A\} & \text{if } T_A^{(m-1)}<\infty\\ \infty & \text{otherwise.}\end{cases}$$

Show that $T_A^{(k)}$ is a stopping time for all $k\geq 1$.

[4] That is, the decision to stop a gamble at time n is determined by the information $\mathscr{F}_n$ available at time n.

A fundamental but very important result is that martingales stopped at a stopping time is a martingale. Recall that we denote $a \wedge b = \min\{a, b\}$.

Theorem 5.2.17 (Stopped submartingale is a submatringale)**.** *If X_n is a submartingale w.r.t. $\mathscr{F}_n$ and if N is a stopping time w.r.t. $\mathscr{F}_n$, then $X_{n\wedge N}$ is a submartingale w.r.t. $\mathscr{F}_n$.*

PROOF. The key observation is the following altenative expression for $X_{n\wedge N}$:

$$X_{n\wedge N} = \sum_{m=1}^{n} \mathbf{1}_{\{N\geq m\}} \cdot (X_m - X_{m-1}). \tag{58}$$

The above identity can be checked by verifying it on each event $\{N = m\}$ for $m \geq 0$. Then triangle inequality and the finite expectation assumption shows that $\mathbb{E}[|X_{n\wedge N}|]$. For adaptedness, note that $X_{n\wedge N} \in \mathscr{F}_n$ since for each $m = 1, \ldots, n$,

$$\mathbf{1}_{\{N\geq m\}} = \mathbf{1}_{\{N\neq 0\}}\mathbf{1}_{\{N\neq 1\}}\cdots\mathbf{1}_{\{N\neq m-1\}} \in \mathscr{F}_{m-1} \subseteq \mathscr{F}_{n-1} \tag{59}$$

and $X_1, \ldots, X_n \in \mathscr{F}_n$. For the conditional increment condition, use linearity of conditional expectation and (59) to get

$$\begin{aligned}
\mathbb{E}\left[X_{(n+1)\wedge N} \,|\, \mathscr{F}_n\right] &= \left(\sum_{m=1}^{n} \mathbf{1}_{\{N\geq m\}} \cdot (X_m - X_{m-1})\right) + \mathbb{E}\left[\mathbf{1}_{\{N\geq n+1\}} \cdot (X_{n+1} - X_n) \,|\, \mathscr{F}_n\right] \\
&= X_{n\wedge N} + \mathbf{1}_{\{N\geq n+1\}} \underbrace{\mathbb{E}\left[X_{n+1} - X_n \,|\, \mathscr{F}_n\right]}_{\geq 0} \\
&\geq X_{n\wedge N}.
\end{aligned}$$

Hence $X_{n\wedge N}$ is a submartingale w.r.t. $\mathscr{F}_n$. □

The key property of stopping time we used in the proof of Theorem 5.2.17 was that $\mathbf{1}_{\{N\geq n+1\}} \in \mathscr{F}_n$. That is, at time n, one already knows that if $N \geq n+1$ or not. This property is generalized as the following notion of 'predictability'.

Definition 5.2.18 (Predictable sequence)**.** Let $(\mathscr{F}_n)_{n\geq 0}$ be a filteration. A stochastic process $(H_n)_{n\geq 1}$ is *predictible* w.r.t. $(\mathscr{F}_n)_{n\geq 0}$ if $H_n \in \mathscr{F}_{n-1}$ for $n \geq 1$.

Example 5.2.19. If N is a stopping time w.r.t. $\mathscr{F}_n$, then $H_n = \mathbf{1}_{\{N\geq n\}}$ is predictable w.r.t. $\mathscr{F}_n$.

A natural interpretation of a predictable sequence H_n is the amount of shares of a stock between time $n-1$ and n, determined by the information available at time $n-1$ by the investor. If we let X_n be the value of one share of that stock at time n, then the gain we have between time $n-1$ and n is $H_n(X_n - X_{n-1})$. Then the total gain from time 0 to time n is[5]

$$\int_0^n H\,dX := \sum_{m=1}^{n} H_m(X_m - X_{m-1}).$$

Example 5.2.20 (Doubling strategy)**.** Let $(\xi_n)_{n\geq 1}$ are i.i.d. RVs with $\mathbb{P}(\xi_n = 1) = p$ and $\mathbb{P}(\xi_n = -1) = 1 - p$. Let $X_n = X_0 + \xi_1 + \cdots + x_n$ so that $\xi_n = X_n - X_{n-1}$. A famous gampling strategy known as "martingale" is defined by a predictable sequence $(H_n)_{n\geq 1}$, where

$$H_n = \begin{cases} 2H_{n-1} & \text{if } \xi_{n-1} = -1 \\ 1 & \text{if } \xi_{n-1} = 1. \end{cases}$$

In words, we double our bet H_n when we lose ($\xi_{n-1} = -1$), so that if we lose k times and then win, our net winnings will be $1 = 2^k - 2^{k-1} - 2^{k-2} - \cdots - 1$. ▲

[5]This is known as the stochastic integral of H against X.

Exercise 5.2.21 (Casino always win)**.** Let $X = (X_n)_{n\geq 0}$ be a supermartingale w.r.t. a filtration $\mathscr{F}_n$ and let $H = (H_n)_{n\geq 1}$ be any predictable sequence w.r.t. $(\mathscr{F}_n)_{n\geq 1}$. Suppose that H_n is bounded and nonnegative for $n \geq 1$. Show that $\int_0^n H\,dX$ is a supermartingale w.r.t. $\mathscr{F}_n$. (*Hint:* Mimic the proof of Theorem 5.2.17.) Also show the similar results for submartingales and martingales. (For the martingale case, it holds without assuming $H_n \geq 0$.)

Next, we introduce the notion of 'upcrossings'.

Definition 5.2.22 (Upcrossing)**.** $(X_n)_{n\geq 0}$ be a supermartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Fix $a, b \in \mathbb{R}$ with $a < b$. Let $N_0 = -1$ and for all $k \geq 1$, define random times

$$\begin{aligned} N_{2k-1} &:= \inf\{m > N_{2k-2} : X_m \leq a\} \\ N_{2k} &:= \inf\{m > N_{2k-1} : X_m \geq b\}. \end{aligned} \tag{60}$$

Namely, between time N_{2k-1} and N_{2k}, X_n crosses from below a to above b, which is called the kth upcrossing of $(X_n)_{n\geq 0}$ between levels a and b (see Figure 5.2.1). The total number of upcrossings completed by time n is denoted

$$U_n := \inf\{k \geq 1 : N_{2k} \leq n\}. \tag{61}$$

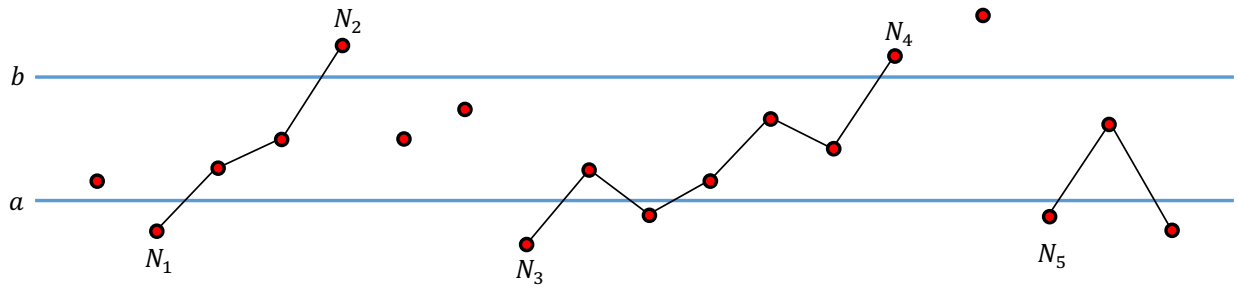


FIGURE 5.2.1. Illustration of stopping times N_{2n-1} and N_{2n} and upcrossings during $[N_{2n-1}, N_{2n}]$ for $n \geq 1$. Solid lines depict the increments counted by H_n in (62).

Lemma 5.2.23 (Upcrossing inequality)**.** *Let $X = (X_n)_{n\geq 0}$ be a submartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Fix $a < b$ and let U_n be as in (61). Then*

$$(b-a)\,\mathbb{E}[U_n] \leq \mathbb{E}[(X_n - a)^+] - \mathbb{E}[(X_0 - a)^+].$$

PROOF. By using a similar argument as in Exercise 5.2.16, the random times N_{2k-1} and N_{2k} in (60) are stopping times. Define a sequence $H = (H_n)_{n\geq 1}$ by

$$H_m = \begin{cases} 1 & \text{if } N_{2k-1} < m \leq N_{2k} \text{ for some } k \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{62}$$

Note that

$$\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m-1\} \cap \{M_{2k} \leq m-1\}^c \in \mathscr{F}_{m-1}.$$

Hence

$$\{H_m = 1\} = \bigcup_{k\geq 1} \{N_{2k-1} < m \leq N_{2k}\} \in \mathscr{F}_{m-1},$$

which shows that $H_m \in \mathscr{F}_{m-1}$. Thus $(H_n)_{n\geq 1}$ is predictable w.r.t. $\mathscr{F}_n$[6].

[6]In stock market terms, $(H_n)_{n\geq 1}$ is a trading strategy that buys one share when $X_n \leq a$ and sells when $X_n \geq b$. Thus one makes a profit of at least $b-a$ during each upcrossing.

Define $Y_n := a + (X_n - a)^+$ for $n \geq 0$. By Proposition 5.2.13, $(Y_n)_{n\geq 0}$ is a submartingale. Note that $Y_n = X_n$ if $X_n \geq a$ and $Y_n = a$ if $X_n < a$. Hence Y_n has an upcrossing whenever X_n has. From this we deduce

$$(b-a)U_n \leq \int_0^n H\,dY. \tag{63}$$

Indeed, for every upcrossing, we make a contribution to the RHS above by at least $b-a$. If there is an imcomplete upcrossing after the last complete upcrossing (see Fig. 5.2.1), then this gives a nonneagtive contribution to the RHS above[7].

Let $K_m := 1 - H_m$ for $m \geq 1$. Note that $K = (K_n)_{n\geq 1}$ is predictable w.r.t. $\mathscr{F}_n$ (since H is). So by Exc. 5.2.21, $\int_0^n K\,dY$ is a submartingale. By Prop. 5.2.11,

$$\mathbb{E}\left[\int_0^n K\,dY\right] \geq \int_0^0 K\,dY = 0.$$

Noting that

$$Y_n - Y_0 = \int_0^n 1\,dY = \int_0^n H\,dY + \int_0^n K\,dY$$

and using (63), we can now deduce

$$\mathbb{E}[Y_n - Y_0] \geq \mathbb{E}\left[\int_0^n H\,dY\right] \geq (b-a)\mathbb{E}[U_n].$$

This finishes the proof. □

We can now prove one of the main results in martingale theory.

Theorem 5.2.24 (Martingale Convergence Theorem)**.** *Let $X = (X_n)_{n\geq 0}$ be a submartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. If $\sup_{n\geq 0}\mathbb{E}[X_n^+] < \infty$, then as $n \to \infty$, X_n converges almost surely to a limit X with $\mathbb{E}[|X|] < \infty$.*

PROOF. Fix $a < b$. Let U_n denote the upcrossings of $[a,b]$ up to time n. Let $U := \sup_{n\geq 0} U_n$, the number of upcrossings on the entire horizon $[0,\infty)$. Then $U_n \nearrow U$. Since $(X_n - a)^+ \leq X_n^+ + |a|$, by Lemma 5.2.23 and the hypothesis,

$$(b-a)\,\mathbb{E}[U] = (b-a)\sup_{n\geq 0}\mathbb{E}[U_n] \leq \sup_{n\geq 0}\mathbb{E}[X_n^+] + |a| < \infty,$$

where the first equality follows from monotone convergence theroem (Thm. 1.3.19) and the fact that U_n is increasing in n. It follows that $\mathbb{E}[U] < \infty$. Since $U \geq 0$ almost surely, it implies that $U < \infty$ almost surely. This yields

$$\mathbb{P}\left(\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty}\right) = 0, \tag{64}$$

since the event in the probability above is the event of having infinitely many upcrossings of $[a,b]$. Now since (64) holds for all reals $a < b$, in particular it holds for all rationals $a < b$. By union bound, we deduce

$$\begin{aligned}\mathbb{P}\left(\liminf_{n\to\infty} X_n \neq \limsup_{n\to\infty} X_n \text{ a.s.}\right) &\leq \mathbb{P}\left(\bigcup_{a<b\in\mathbb{Q}}\left\{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\right\}\right)\\ &\leq \sum_{a<b\in\mathbb{Q}} \mathbb{P}\left(\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty}\right) = 0.\end{aligned}$$

This shows that X_n converges to some limit X almost surely. (We can write $X = \liminf_{n\to\infty} X_n$.) It also follows that both $X_n^+ := X_n \vee 0$ and $X_n^- := (-X_n) \vee 0$ converge almost surely.

[7] Note that the last incomplete upcrossing may result in negative contribution to $\int_0^n H\,dX$, so (63) may be false with X instead of Y.

Now it remains to show that $\mathbb{E}[|X|] < \infty$. On the one hand, by Fatou's lemma (Thm. 1.3.18),

$$\mathbb{E}[X^+] = \mathbb{E}[\liminf_{n\to\infty} X_n^+] \le \liminf_{n\to\infty} \mathbb{E}[X_n^+ \le \sup_{n\ge 0} \mathbb{E}[X_n^+] < \infty.$$

On the other hand, since X_n is a submartingale,

$$\mathbb{E}[X_n^-] = \mathbb{E}[X_n^+] - \mathbb{E}[X_n] \le \mathbb{E}[X_n^+] - \mathbb{E}[X_0].$$

Hence again by Fatou's lemma,

$$\mathbb{E}[X^-] = \mathbb{E}[\liminf_{n\to\infty} X_n^-] \le \liminf_{n\to\infty} \mathbb{E}[X_n^-] \le \sup_{n\ge 0} \mathbb{E}[X_n^+] - \mathbb{E}[X_0] < \infty.$$

Thus we conclude

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

□

An important consequence of Theorem 5.2.24 is the following.

Corollary 5.2.25. *If $X_n \ge 0$ is a supermartingale, then as $n \to \infty$, $X_n \to X$ almost surely and $\mathbb{E}[X] \le \mathbb{E}[X_0]$.*

PROOF. $Y_n := -X_n$ is a submartingale uniformly bounded above by 0. Thus by Theorem 5.2.24, $Y_n \to Y$ converges almost surely for some integrable Y. Then $X_n \to -Y =: X$ almost surely and $\mathbb{E}[|X|] = \mathbb{E}[|Y|] < \infty$. Lastly, by Fatou's lemma and since X_n is a supermartingale,

$$\mathbb{E}[X] = \mathbb{E}[\liminf_{n\to\infty} X_n] \le \liminf_{n\to\infty} \mathbb{E}[X_n] \le \mathbb{E}[X_0].$$

This finishes the proof. □

Recall that almost sure convergence does not necessarily imply convergence in expectation (i.e., convergnece in L^1). The following counterexample shows that this is indeed the case for martingales.

Example 5.2.26 (Martingale convergene do not hold in L^1)**.** Let $(S_n)_{n\ge 0}$ be a simple symmetric random walk on $\mathbb{Z}$ with $S_0 = 1$. Let $N = \inf\{m \ge 0 \mid S_m = 0\}$, the first hitting time of the origin (i.e., stop gambling when broke). Since S_n is a martingale w.r.t. the filtration $\mathscr{F}_n = \sigma(S_0, \dots, S_n)$ and since N is a stopping time w.r.t. the same filtration, by Theorem 5.2.17, $X_n := S_{n\wedge N}$ is also a martingale. Since $X_n \ge 0$, it is a nonnegative supermartingale, so by Corollary 5.2.25, $X_n \to X$ a.s. as $n \to \infty$. Note that $X \equiv 0$, since for any $k > 0$,

$$\mathbb{P}\{X = k\} \le \mathbb{P}\{X_n = k \text{ for all but finitely many } n\} \le \mathbb{P}\{X_n = X_{n+1} = k \text{ for some } n\} = 0.$$

Thus $\mathbb{P}(X = 0) = 1$. But note that since X_n is a martingale, $\mathbb{E}[X_n] = \mathbb{E}[X_0] = \mathbb{E}[S_0] = 1 \ne 0 = \mathbb{E}[X]$. ▲

Example 5.2.27. This example is brought from [Dur19, Ex. 4.2.14]. We will construct a martingale $(X_n)_{n\ge 0}$ on a large enough common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ recursively as follows. Let $X_0 = 0$ and for $k \ge 1$, define $\mathscr{F}_{k-1} := \sigma(X_1, \dots, X_{k-1})$, and on the event that $X_{k-1} = 0$, set

$$X_k = \begin{cases} 1 & \text{indepenently from } \mathscr{F}_{k-1} \text{ with prob. } 1/2k \\ -1 & \text{indepenently from } \mathscr{F}_{k-1} \text{ with prob. } 1/2k \\ 0 & \text{indepenently from } \mathscr{F}_{k-1} \text{ with prob. } 1 - k^{-1} \end{cases}$$

and on the event that $X_{k-1} \ne 0$, set

$$X_k = \begin{cases} kX_{k-1} & \text{indepenently from } \mathscr{F}_{k-1} \text{ with prob. } 1/k \\ 0 & \text{indepenently from } \mathscr{F}_{k-1} \text{ with prob. } 1 - k^{-1}. \end{cases}$$

Often we define a RV by only specifying its *distribution* as above without specifying where in the sample space Ω it takes a specific value. We claim that X_n is a martingale w.r.t. the filtration $\mathscr{F}_n$. This should be intuitive since conditional on knowing the value of X_{k-1}, the conditional expectation of X_k is 0 if $X_{k-1} = 0$ and X_{k-1} if $X_{k-1} \ne 0$. Below we will give a detailed verification.

We first each X_k, $k \geq 1$ more precisely. Choose arbitrary events $A_k, B_k, C_k \in \mathscr{F}$ independent from $\mathscr{F}_{k-1}$[8] with $\mathbb{P}(A_k) = \mathbb{P}(B_k) = 1/2k$ and $\mathbb{P}(C_k) = 1/k$. Then we define

$$X_k(\omega) = \begin{cases} 1 & \text{if } X_{k-1}(\omega) = 0 \text{ and } \omega \in A_k \\ -1 & \text{if } X_{k-1}(\omega) = 0 \text{ and } \omega \in B_k \\ kX_{k-1}(\omega) & \text{if } X_{k-1}(\omega) \neq 0 \text{ and } \omega \in C_k \\ 0 & \text{otherwise.} \end{cases}$$

It is not really important what these sets A_k, B_k, C_k actually are, in the sense that, as long as they are independent from $\mathscr{F}_{k-1}$ and have the probabilities as above, the distribution of X_k is completely determined so we can compute any probability involving X_k.

Now we show that X_n is a martingale w.r.t $\mathscr{F}_n$. Finite expectation and adaptedness are clear. For martingale increment, we need to show that X_{n-1} is a version of $\mathbb{E}[X_n \,|\, \mathscr{F}_{n-1}]$. To verify this, fix $B \in \mathscr{F}_{n-1}$ and we will show

$$\mathbb{E}[X_n \mathbf{1}(B)] = \mathbb{E}[X_{n-1}\mathbf{1}(B)]. \tag{65}$$

We do this by partitioning the sample space Ω suitably. Namely, note that

$$\begin{aligned} \mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} = 0)\mathbf{1}(A_k)] &= 1 \cdot \mathbb{P}(B \cap \{X_{n-1} = 0\} \cap A_k) \\ &= \mathbb{P}(B \cap \{X_{n-1} = 0\})\,\mathbb{P}(A_k) \\ &= \mathbb{P}(B \cap \{X_{n-1} = 0\})\,(1/2k), \end{aligned}$$

since $B \cap \{X_{n-1} = 0\} \in \mathscr{F}_{n-1}$ and $A_k \perp \mathscr{F}_{n-1}$. Similarly,

$$\begin{aligned} \mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} = 0)\mathbf{1}(B_k)] &= -\mathbb{P}(B \cap \{X_{n-1} = 0\})\,(1/2k), \\ \mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} = 0)\mathbf{1}(A_k^c \cap B_k^c)] &= 0, \end{aligned}$$

so by linearity of expectation,

$$\mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} = 0)] = 0 = \mathbb{E}[X_{n-1}\mathbf{1}(B)\mathbf{1}(X_{n-1} = 0)]. \tag{66}$$

Also observe that

$$\begin{aligned} \mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)\mathbf{1}(C_k)] &= n\,\mathbb{E}[X_{n-1}\mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)]\,\mathbb{P}(C_k) \\ &= n\,\mathbb{E}[X_{n-1}\mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)]\,(1/n) \\ &= \mathbb{E}[X_{n-1}\mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)], \\ \mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)\mathbf{1}(C_k^c)] &= 0. \end{aligned}$$

Again by linearity of expectation,

$$\mathbb{E}[X_n \mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)] = \mathbb{E}[X_{n-1}\mathbf{1}(B)\mathbf{1}(X_{n-1} \neq 0)]. \tag{67}$$

We can then combine (66) and (67) by linearity of expectation to obtain (65). Thus X_n is a martingale w.r.t. $\mathscr{F}_n$.

Now note that $X_n \to 0$ in probability since

$$\begin{aligned} \mathbb{P}(X_n = 0) &= \mathbb{P}(X_n = 0 \,|\, X_{n-1} = 0)\,\mathbb{P}(X_{n-1} = 0) + \mathbb{P}(X_n = 0 \,|\, X_{n-1} \neq 0)\,\mathbb{P}(X_{n-1} \neq 0) \\ &= (1 - k^{-1})\mathbb{P}(X_{n-1} \neq 0) + (1 - k^{-1})\mathbb{P}(X_{n-1} \neq 0) = (1 - k^{-1}) \to 1. \end{aligned}$$

However, X_n does not converge to 0 almost surely. To see this, note that $\mathbb{P}(X_n \neq 0 \,|\, \mathscr{F}_{n-1}) = 1/n$, so we have

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq 0 \,|\, \mathscr{F}_{n-1}) = \infty.$$

[8] This is where we need our probability space 'large enough'. If it weren't, we could have expanded it by taking the product space with another probability space so that it can accomodate events independent from $\mathscr{F}_{k-1}$.

So by Borel-Cantelli lemma (Lem. 5.3.3), $X_n \neq 0$ infintely often with probability 1. But notice that X_n is an integer-valued process, so if it converges to 0 with a positive probability, then $X_n \equiv 0$ for all sufficiently large n with a positive probabilty. Thus $\mathbb{P}(\lim_{n\to\infty} X_n = 0) = 0$. ▲

5.3. Applications of martingale convergence

In this section, we will apply the martingale convergence theorem (Thm. 5.2.24) to generalize the second Borel-Cantelli lemma (Lem. 3.4.15) and to study Polya's urn scheme, Radon-Nikodym derivatives, and branching processes. The four topics are independent of each other.

5.3.1. Bounded increments. We first show that martingales with bounded increments either converge or oscillate between $+\infty$ and $-\infty$.

Theorem 5.3.1 (Asymptotic behavior of martingales with bounded increments)**.** *Let $(X_n)_{n\geq 0}$ be a martingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Suppose that it has bounded increments:* $\sup_{n\geq 1}|X_{n+1}-X_n| \leq M$ *almost surely for some constant $M > 0$. Let*

$$\mathscr{C} := \{\lim_{n\to\infty} X_n \text{ exists and is finite a.s. }\}$$
$$\mathscr{D} := \{\liminf_{n\to\infty} X_n = -\infty \text{ and } \limsup_{n\to\infty} X_n = \infty.\}$$

Then $\mathbb{P}(\mathscr{C}\cup\mathscr{D}) = 1$.

PROOF. Since $X_n - X_0$ is a martingale, we can WLOG assume $X_0 = 0$. Fix $K \in (0,\infty)$ and define $N := \inf\{n \geq 0 \,|\, X_n \leq -K\}$, the first hitting time of $(-\infty, K]$. This is a stopping time, so by Theorem 5.2.17, $X_{n\wedge N}$ is a martingale. Note that $X_{(n\wedge N)-1} > -K$, so

$$X_{n\wedge N} = X_{(n\wedge N)-1} + \left(X_{n\wedge N} - X_{(n\wedge N)-1}\right) \geq -K - M.$$

It follows that $Y_n := X_{n\wedge N} + K + M$ is a nonnegative martingale, so by Theorem 5.2.24, Y_n converges a.s. to some limit Y. It follows that

$$\left\{\inf_{n\geq 0} X_n > -K\right\} = \{N = \infty\} \subseteq \{X_n \text{ converges a.s.}\}.$$

Since the above holds for all $K > 0$, letting $K \to \infty$[9],

$$\left\{\liminf_{n\to\infty} X_n > -\infty\right\} = \left\{\inf_{n\geq 0} X_n > -\infty\right\} \subseteq \{X_n \text{ converges a.s.}\}. \tag{68}$$

Applying this result for $-X_n$, we can also deduce

$$\left\{\limsup_{n\to\infty} X_n < \infty\right\} \subseteq \{X_n \text{ converges a.s.}\}.$$

This shows $\mathscr{D}^c$ implies X_n converges a.s.,which is enough to conclude. □

Theorem 5.3.2 (Doob's decomposition)**.** *Any submartingale $(X_n)n \geq 0$, can be written in a unique way as $X_n = M_n + A_n$, where M_n is a martingale and A_n is a predictable increasing sequence with A_0. More specifically,*

$$A_n = \sum_{m=1}^{n} \mathbb{E}[X_m - X_{m-1} \,|\, \mathscr{F}_{m-1}]. \tag{69}$$

[9] To see the first identity in (68), note that on the event $\liminf_{n\to\infty} X_n > -\infty$, we have $X_n > -\infty$ for all $n \geq N_0$ for some (random) N_0. Since X_n's are integrable, $X_n > -\infty$ for all $n < N_0$. Hence $\inf X_n > -\infty$ on this event.

PROOF. The statement says that the submartingale X_n has its 'stationary part' M_n and the 'increasing part' A_n. The formula (69) makes sense since $\mathbb{E}[X_m - X_{m-1} \,|\, \mathscr{F}_{m-1}] \geq 0$ is the one-step conditional increment of the submartingale X_n, which must be subtracted to make it a martingale. Furthermore, this conditional increment is $\in \mathscr{F}_{m-1} \subseteq \mathscr{F}_{n-1}$.

We first show that the formula (69) works. As we noted, A_n defined in (69) is predictable and increasing. To show that $M_n := X_n - A_n$ is a martingale (finite expectation and adaptededness are easy),

$$\begin{aligned}
\mathbb{E}[M_n \,|\, \mathscr{F}_{n-1}] &= \mathbb{E}[X_n - A_n \,|\, \mathscr{F}_{n-1}] \\
&= \mathbb{E}[X_{n-1} + X_n - X_{n-1} - A_n \,|\, \mathscr{F}_{n-1}] \\
&= X_{n-1} - A_n + \underbrace{\mathbb{E}[X_n - X_{n-1} \,|\, \mathscr{F}_{n-1}]}_{= A_n - A_{n-1}} \\
&= X_{n-1} - A_{n-1} \\
&= M_{n-1}.
\end{aligned}$$

Next, suppose there exists such a Doob's decomposition $X_n = M_n + A_n$. We show that it must satisfy (69). Indeed, M_n and A_n should satisfy

$$\mathbb{E}[M_n \,|\, \mathscr{F}_{n-1}] = M_{n-1} \quad \text{and} \quad A_n \in \mathscr{F}_{n-1} \quad \text{and} \quad A_n \nearrow .$$

It follows that

$$\begin{aligned}
\mathbb{E}[X_n \,|\, \mathscr{F}_{n-1}] &= \mathbb{E}[M_n \,|\, \mathscr{F}_{n-1}] + \mathbb{E}[A_n \,|\, \mathscr{F}_{n-1}] \\
&= M_{n-1} + A_n \\
&= X_{n-1} - A_{n-1} + A_n.
\end{aligned}$$

Since $X_{n-1} \in \mathscr{F}_{n-1}$, this shows that $A_n - A_{n-1} = \mathbb{E}[X_n - X_{n-1} \,|\, \mathscr{F}_{n-1}]$, which then yields (69). □

An important application of Theorems 5.3.1 and 5.3.2 is the following generalization of the second Borel-Cantelli Lemma 3.4.15.

Lemma 5.3.3 (Second Borel-Cantelli Lemma, conditional ver.)**.** *Let $(\mathscr{F}_n)_{n\geq 0}$ be a filtration with $\mathscr{F}_0 = \{\varnothing, \Omega\}$. Let $(B_n)_{n\geq 0}$ be a sequence of events such that $B_n \in \mathscr{F}_n$. Then possibly except a set of probability zero,*

$$\{B_n \text{ i.o.}\} = \left\{ \sum_{n=1}^{\infty} \mathbf{1}(B_n) = \infty \right\} = \left\{ \sum_{n=1}^{\infty} \mathbb{P}(B_n \,|\, \mathscr{F}_{n-1}) = \infty \right\}.$$

PROOF. Let $X_0 = 0$ and $X_n = \sum_{m=1}^{n} \mathbf{1}(B_m)$. Then X_n is a submartingale. Applying Doob's decomposition (Thm. 5.3.2), we can write $X_n = M_n + A_n$ where

$$A_n = \sum_{m=1}^{n} \mathbb{P}(B_m \,|\, \mathscr{F}_{m-1}) \quad \text{and} \quad M_n = \sum_{m=1}^{n} \mathbf{1}(B_m) - \mathbb{P}(B_m \,|\, \mathscr{F}_{m-1}).$$

Here M_n is a martingale with incremenet bounded by 1. Using the notation in Theorem 5.3.1, we have

$$\begin{aligned}
&\text{on } \mathscr{C}, \quad \sum_{n=1}^{\infty} \mathbf{1}(B_n) = \infty \quad \Longleftrightarrow \quad \sum_{n=1}^{\infty} \mathbb{P}(B_n \,|\, \mathscr{F}_{n-1}) = \infty, \\
&\text{on } \mathscr{D}, \quad \sum_{n=1}^{\infty} \mathbf{1}(B_n) = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \mathbb{P}(B_n \,|\, \mathscr{F}_{n-1}) = \infty.
\end{aligned}$$

Since $\mathbb{P}(\mathscr{C} \cup \mathscr{D}) = 1$ by Theorem 5.3.1, the result follows. □

5.3.2. Polya's Urn. An urn contains r red and g green balls. At each time we draw a ball out, then replace it, and add c more balls of the color drawn. Let X_n be the fraction of green balls after the nth draw. Will X_n converge to some random fraction? This may not be obvious, but we can see it by using the martingale convergence theorem.

Let $\mathscr{F}_n$ denote the σ-algebra generated by the first n draws. Then X_n is a martingale w.r.t. $\mathscr{F}_n$. To see this, assuming that there are i red balls and j green balls at time n, then

$$X_{n+1} = \begin{cases} \frac{j+c}{i+j+c} & \text{with prob. } \frac{j}{i+j} \\ \frac{j}{i+j+c} & \text{with prob. } \frac{i}{i+j}. \end{cases}$$

So we have

$$\mathbb{E}[X_{n+1} \,|\, \mathscr{F}_n] = \frac{j+c}{i+j+c} \cdot \frac{j}{i+j} + \frac{j}{i+j+c} \cdot \frac{i}{i+j} = \frac{j(i+j+c)}{(i+j+c)(i+j)} = \frac{j}{i+j} = X_n.$$

Since X_n is a nonnegative martingale, by Corollary 5.2.25, X_n converges to some limiting RV X_∞, which is the limiting fraction of green balls. We also know that $\mathbb{E}[X_\infty] \le \mathbb{E}[X_0]$. But the actual value of X_∞ is random. What is the distribution of X_∞?

We make the following observation. Let ξ_k denote the indicator that we get a green ball at the kth draw. Then

$$\mathbb{P}\big((\xi_1,\dots,\xi_n) = (\underbrace{1,\dots,1}_{m\text{ 1's}},\underbrace{0,\dots,0}_{\ell\text{ 0's}})\big) = \frac{g}{r+g} \cdot \frac{g+c}{r+g+c} \cdots \frac{g+(m-1)c}{r+g+(m-1)c} \cdot \frac{r}{r+g+mc} \cdots \frac{r+(\ell-1)c}{r+g+(n-1)c}.$$

In fact, the above probability is unchanged for any binary sequence of length n with the same number of 1's, since the denominators in the RHS are unchanged and the numerators are only permuted.

Now for concrete examples, suppose $r = g = c = 1$. By the above computation, for any $m = 0,\dots,n$,

$$\mathbb{P}\left(\sum_{k=1}^n \xi_k = m\right) = \binom{n}{m} \frac{m!\,\ell!}{(n+1)!} = \frac{1}{n+1}.$$

This shows that X_n is uniform among all possible values. Since $X_n \to X_\infty$ a.s. implies convergence in probability, for each $x \in [0,1]$,

$$\mathbb{P}(X_\infty \le x) = \lim_{n\to\infty} \mathbb{P}(X_n \le x) = x.$$

This shows that $X_\infty \sim \text{Uniform}(0,1)$.

Next, suppose $r = c = 1$ and $g = 2$. Then

$$\mathbb{P}\left(\sum_{k=1}^n \xi_k = m\right) = \binom{n}{m} \frac{(2\cdot 3\cdots m+1)\,\ell!}{3\cdot 4\cdots(n+2)} = \frac{2(m+1)}{n+2} \to 2x$$

provided $n \to \infty$ and $m/n \to x$.

In general, X_∞ is a continuous RV on $[0,1]$ with density

$$\frac{\Gamma((g+r)/c)}{\Gamma(g/c)\Gamma(r/c)} x^{(g/c)-1}(1-x)^{r/c} - 1,$$

which is the Beta distribution with parameters g/c and r/c (see Exc. 1.6.19).

Below in Figure 5.3.1 (generated by using Python), we provide some simulation results of Polya's urn for various parameter choices.

Exercise 5.3.4. Use your favorate programming language (e.g., python, R, matlab, C++) and reproduce plots similar to the ones in Figure 5.3.1.

5.3.3. Branching processes. The martingale theory makes an important application in analyzing branching processes, a stochastic model for population growth.

Imagine a specie where every individual gives birth to a random number of offsprings for the next generation and dies at the end of the current generation. Suppose for simplicity that the number of offsprings that the ith individual at generation n, denoted as ξ_i^n, are i.i.d. from some fixed 'offspring distribution' $\mathbf{p} = (p_0, p_1, \dots,)$, i.e., $\mathbb{P}(\xi_i^n = k) = p_k$ for $k \ge 0$. Suppose there are Z_n individuals at generation n. Then Z_{n+1}, the population at the next generation, should be given by the sum of all offsprings of every individuals at generation n. This is the idea behind the branching process in the following definition.

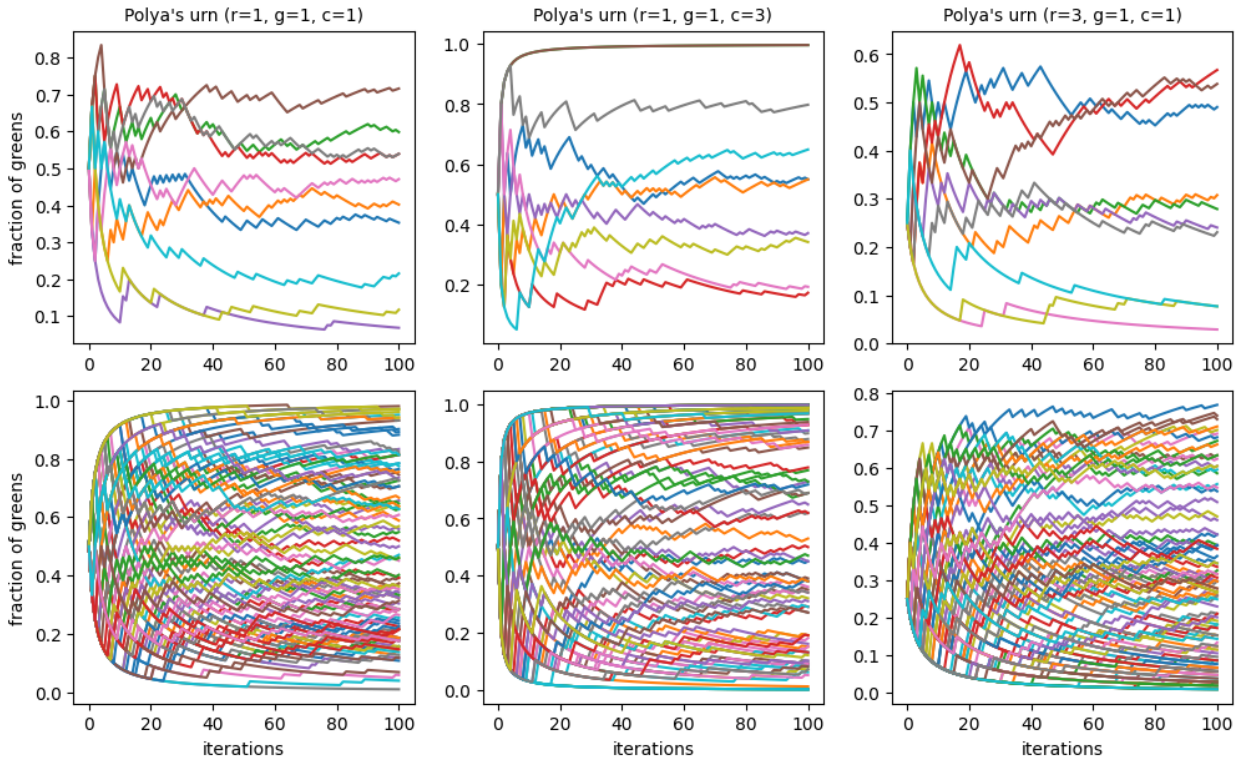


FIGURE 5.3.1. 10 (top row) and 100 (bottom row) sample trajectories of Polya's urn process.

Definition 5.3.5 (Branching processes)**.** Let $(\xi_i^n)_{i,n\geq 1}$ be a sequence of doubly indexed i.i.d. nonnegative integer-valued random variables with distribution $\mathbf{p} = (p_1, p_2, \dots)$. Define a sequence Z_n, $n \geq 0$, by $Z_0 = 1$ and

$$Z_{n+1} := \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & \text{if } Z_n \geq 1 \\ 0 & \text{if } Z_n = 0. \end{cases}$$

Then $(Z_n)_{n\geq 0}$ is called the *branching process* (or a Galton-Watson process) with *offspring distribution* $\mathbf{p}$. We call $\mu := \mathbb{E}[\xi_i^n] = \sum_{k=1}^{\infty} k p_k$ the *mean offspring number*.

Every individual on average gives μ number of offsprings. It would be reasonable to guess that the following behavior:

(*Subcritical phase*): $Z_n \to 0$ almost surely and exponentially fast if $\mu < 1$;

(*Critical phase*): Z_n is 'unbiased' but $Z_n \to 0$ almost surely if $\mu = 1$;[10]

(*Supercritical phase*): $Z_n \sim \mu^n$ almost surely if $\mu > 1$.

We will use martingale theory to justify the above speculations. The key connection between the branching processes and martingales is the following.

Lemma 5.3.6 (Normalized branching process is a martingale)**.** *Let* $\mathscr{F}_n := \sigma(\xi_i^m; i \geq 1, 1 \leq m \leq n)$ *and* $\mu := \mathbb{E}[\xi_i^n] \in (0, \infty)$. *Then* Z_n/μ^n *is a martingale w.r.t. the filtration* $\mathscr{F}_n$.

PROOF. Finite expectation and adaptedness are clear. For martingale increments, since $Z_n \in \mathscr{F}_n$, we have

$$\mathbb{E}[Z_{n+1} \,|\, \mathscr{F}_n] = \mathbb{E}[\xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} \,|\, \mathscr{F}_n] = \mu Z_n. \tag{70}$$

[10] Exclusing the trivial case when $Z_n \equiv 1$ a.s., which occurs when $\xi_i^n \equiv 1$ a.s. for all i, n.

The desired result follows from multiplying both sides above with $1/\mu^{n+1}$.[11] □

Proposition 5.3.7. *Z_n/μ^n converges almost surely to some limiting RV.*

PROOF. Since Z_n/μ^n is a nonnegative martingale, it follows from Corollary 5.2.25. □

Note that since Z_n/μ^n is a martingale by Prop. 5.3.7, $\mathbb{E}[Z_n/\mu^n] = \mathbb{E}[Z_0] = 1$. Therefore

$$\mathbb{E}[Z_n] = \mu^n. \tag{71}$$

This confirms the conjectured phase transition behavior of the branching process in expectation.

Next, we identify the limit in Prop. 5.3.7. We start with the subcritical phase $\mu < 1$, in which case the limit should be zero.

Proposition 5.3.8 (Extinction of subcritical branching process)**.** *If $\mu < 1$, then $Z_n \equiv 0$ for all n sufficiently large, so $Z_n/\mu^n \to 0$ almost surely.*

PROOF. By Markov's inequality and (71),

$$\mathbb{P}(Z_n > 0) = \mathbb{P}(Z_n \geq 1) \leq \mathbb{E}[Z_n] = \mu^n.$$

Hence if $\mu \in (0,1)$, then $\mathbb{P}(Z_n > 0) \to 0$ exponentially fast. By Borel-Canteli lemma (Lem. 3.4.6), $Z_n > 0$ only for finitely many ns almost surely. Thus $Z_n = 0$ for all sufficiently large n almost surely. □

Next we consider the critical case $\mu = 1$. Clearly if every individual gives birth to exactly one offspring, then $Z_n \equiv 1$ almost surely. If we exclude this trivial case, then any small amount of randomness is enough to get the entrie population extinct.[12]

Proposition 5.3.9 (Extinction of critical branching process)**.** *If $\mu = 1$ and $\mathbb{P}(\xi_i^m \neq 1) > 0$, then $Z_n \equiv 0$ for all n sufficiently large.*

PROOF. Since $\mu = 1$, Z_n itself is a nonnegative martingale by Prop. 5.3.7. Hence $Z_n \to Z_\infty$ almost surely for some limiting RV Z_∞ by Corollary 5.2.25. Since Z_n is integer-valued, so is Z_∞. Hence $Z_n = Z_\infty$ almost surely when n is sufficiently large. Fix an integer $k \geq 1$. Then

$$\begin{aligned}\mathbb{P}(Z_\infty = k) &= \mathbb{P}(Z_n = k \text{ for all } n \text{ sufficiently large}) \\ &\leq \sum_{N\geq 1} \mathbb{P}(Z_n = k \text{ for all } n \geq N).\end{aligned}$$

Now for any $N \geq 1$,

$$\begin{aligned}\mathbb{P}(Z_n = k \text{ for all } n \geq N) &= \mathbb{P}(\,\xi_1^n + \cdots + \xi_k^n = k \text{ for all } n \geq N) \\ &= \prod_{n\geq N} \mathbb{P}(\,\xi_1^n + \cdots + \xi_k^n = k).\end{aligned}$$

Note that the probability $\mathbb{P}(\,\xi_1^n + \cdots + \xi_k^n = k)$ does not depend on n. Furthermore, since the offspring distribution is not the point mass at 1, $\mathbb{P}(\xi_1^n + \cdots + \xi_k^n = k)$ is strictly less than one. Thus the infinite product in the last expression above is zero. This shows $\mathbb{P}(Z_\infty = k) = 0$ for all $k \geq 1$. Since Z_∞ is integer-valued, it follows that $Z_\infty = 0$ almost surely. □

---

[11] If you are not completely satisfied with the argument for (70) since Z_n is random (albeit being in $\mathscr{F}_n$), we can proceed by showing the claimed identity with partitioning the sample space. That is, write $\Omega = \bigsqcup_{k=0}^\infty \{Z_n = k\}$. Then for each k, $\{Z_n = k\} \in \mathscr{F}_n$ and $Z_{n+1} = \xi_1^{n+1} + \cdots + \xi_k^{n+1}$ on $\{Z_n = k\}$. Hence by Exc. 5.1.6,

$$\mathbb{E}[Z_{n+1} \,|\, \mathscr{F}_n] = \mathbb{E}[\xi_1^{n+1} + \cdots + \xi_k^{n+1} \,|\, \mathscr{F}_n] = \mu k = \mu Z_n.$$

The above holds for all $k \geq 0$, so $\mathbb{E}[Z_{n+1} \,|\, \mathscr{F}_n] = \mu Z_n$ almost surely.

[12] This is just like the fact that symmetric random walk on $\mathbb{Z}$ will eventually visit zero.

Lastly, we consider the supercritical branching process ($\mu > 1$). In this case, we will show that the population survives forever with a positive probability, meaning that $Z_n > 0$ for all n with a positive probability. Furthermore, we can exactly compute this survival probability. using the generating function of the offspring distribution.

Define the *generating function* of the offspring distribution $\mathbf{p} = (p_0, p_1, \dots)$ as

$$\varphi(s) := \mathbb{E}[s^{Z_1}] = \sum_{k=0}^{\infty} p_k s^k.$$

Observe that

$$\varphi'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1} > 0, \qquad \varphi''(s) = \sum_{k=2}^{\infty} k(k-1) p_k s^{k-2} > 0.$$

Hence φ has the following properties:

1. φ is strictly increasing on $[0,1)$;

2. $\varphi'' > 0$ on $(0,1)$, and hence φ' is strictly increasing and φ is strictly convex on $(0,1)$;

3. $\varphi(1) = 1$.

4. $\varphi(0) = p_0$ and $\varphi'(0) = p_1$.

We make some observations on the generating functions.

Lemma 5.3.10. *Define $\varphi_n(t) := \mathbb{E}[t^{Z_n}]$. Then φ_n is the n-fold composition of φ by itself, that is,*

$$\varphi_0(t) = t,$$
$$\varphi_{n+1}(t) = \varphi(\varphi_n(t)) = \varphi_n(\varphi(t)).$$

PROOF. We present two argumetns depending on conditioning on $\mathscr{F}_1$ or $\mathscr{F}_n$. First, note that Z_{n+1} is the sum of Z_1 independent copies of Z_n, due to the recursive structure. Hence, denoting Z_n^i for $i \geq 1$ to be i.i.d. copies of Z_n,

$$\begin{aligned}
\varphi_{n+1}(t) &= \mathbb{E}\left[t^{Z_{n+1}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[t^{Z_{n+1}} \mid \mathscr{F}_1\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[t^{Z_n^1 + \cdots + Z_n^{Z_1}} \mid \mathscr{F}_1\right]\right] \\
&= \mathbb{E}\left[\varphi_n(t)^{Z_1}\right] \\
&= \varphi(\varphi_n(t)).
\end{aligned}$$

For the second argument, note that

$$\begin{aligned}
\varphi_{n+1}(t) &= \mathbb{E}\left[t^{Z_{n+1}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[t^{Z_{n+1}} \mid \mathscr{F}_n\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[t^{\xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1}} \mid \mathscr{F}_n\right]\right] \\
&= \mathbb{E}\left[\varphi(t)^{Z_n}\right] \\
&= \varphi_n(\varphi(t)).
\end{aligned}$$

□

A key notion for describing the behavior of a branching process is its *extinction time*, which is the first time that the population reaches zero:

$$\tau := \inf\{n \geq 0 \mid Z_n = 0\}. \tag{72}$$

With the convention that $\inf \emptyset = \infty$, we have $\tau = \infty$ if Z_n never goes extinct. What is the probability of extinction, $\mathbb{P}(\tau < \infty)$? The following lemma is a key result that relates the extinction probability and the fixed point of the generating function.

Lemma 5.3.11 (Extinction probability)**.** *The extinction probability* $\zeta := \mathbb{P}(\tau < \infty)$ *is the smallest nonnegative root of the fixed point equation* $\varphi(t) = t$.

PROOF. We present two approaches. First we appeal to the recursive nature of the branching process. Consider what has to happen for the event $\tau < \infty$. There are Z_1 first-generation nodes, and descending from them there are independent copies of the entire branching process. Therefore by partioning on the values of Z_1,

$$\zeta = p_0 + \sum_{k=1}^{\infty} p_k \zeta^k = \varphi(\zeta).$$

Second, observe that (since $0^0 = 1$)

$$\mathbb{P}(Z_n = 0) = \varphi_n(0).$$

Note that $Z_{n+1} = 0$ if $Z_n = 0$, so $\mathbb{P}(Z_n = 0)$ is non-decreasing in n. It follows that, by continuity of probability measure (Thm. 1.1.16), $\zeta = \lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \lim_{n\to\infty} \varphi_n(0)$. It follows that by Lem. 5.3.10,

$$\zeta = \lim_{n\to\infty} \varphi(\varphi_n(0)) = \varphi(\lim_{n\to\infty} \varphi_{n-1}(0)) = \varphi(\zeta).$$

Lastly, we conclude that ζ is the smallest nonnegative root of $\varphi(t) = t$. This follows from the monotonicity of φ. Indeed, suppose ζ' is another nonnegative root of the fixed point equation. Then since φ is strictly increasing and by Lem. 5.3.10, $\varphi_n = \varphi \circ \cdots \circ \varphi$ is also strictly increasing. Since $0 \le \zeta'$,

$$\varphi_n(0) \le \varphi_n(\zeta').$$

Letting $n \to \infty$, we get $\zeta \le \zeta'$. So we must have $\zeta = \zeta'$. □

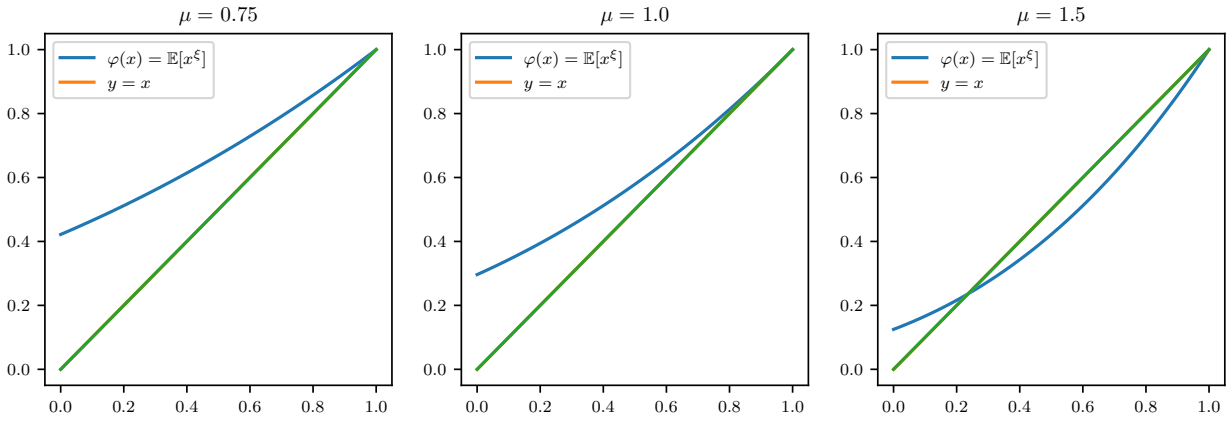


FIGURE 5.3.2. Phase transition in the fixed point of the generating function φ of the offspring distribution $\xi \sim \mathbf{p} = \text{Binomial}(3, p)$, where the mean offspring number $\mu = 3p$. When $\mu \le 1$, $\zeta = 1$ is the only fixed point of φ. When $p > 1/3$ so that $\mu > 1$, there appears a new fixed point $\zeta < 1$.

The following is a key observation for the generating function φ of the offpsring distribution $\mathbf{p}$.

Theorem 5.3.12 (Phase transition in extinction probability)**.** *Suppose* $p_1 < 1$ *and let* $\zeta = \mathbb{P}(\tau < \infty)$ *denote the exinction probability. Then the followings hold:*

(i) *(*subcritical and critical regime*) If* $\mu \le 1$, *then* $\zeta = 1$.

(ii) *(*supercritical regime*) If* $\mu > 1$, *then* $\zeta < 1$.

PROOF. By Lemma 5.3.11, we know that ζ is the smallest nonnegative fixed point of the generationg function φ of the offspring distribution $\mathbf{p}$.

Suppose $\mu \le 1$. Recall that φ is strictly convex with strictly increasing first derivative and positive second derivative. Note that $\lim_{t\nearrow 1} \varphi'(t) = \mu \le 1$, so $\varphi'(t) < 1$ for all $t \in (0,1)$. Then by mean value

theorem, there cannot be any fixed point of φ in $[0,1)$, for if $\zeta = \varphi(\zeta)$ for some $\zeta \in [0,1)$, then since $\varphi(1) = 1$, there must be some $\zeta' \in (\zeta, 1)$ such that $\varphi'(\zeta') = 1$, but this contradicts the fact that $\varphi' < 1$ on $(0,1)$. Since $\varphi(1) = 1$, the smallest nonnegative fixed point of φ is 1. Therefore $\zeta = 1$.

Next, suppose $\mu > 1$. Recall that $\varphi(0) = p_0$. Hence if p_0, then $\varphi(0) = 0$ and $\varphi(1) = 1$. Since φ is strictly convex, there is no other fixed point in $(0,1)$. Now suppose $p_0 > 0$ so that $\varphi(0) = p_0 > 0$. Since $\varphi'(1) = \mu > 1$, when $t \nearrow 1$ we should have $\varphi'(t) < t$ by Taylor expansion. Hence there exists $t^* \in (0,1)$ such that $\varphi(t^*) < t^*$. Since $\varphi(0) > 0$, by the intermediate value theorem, there exists a fixed point ζ in $(0, t^*)$. It follows that $\zeta < 1$. □

Theorem 5.3.12 only states that $\zeta < 1$ in the supercritical case but does not provide what that value is. Lemma 5.3.11 tells us that we only need to solve the fixed point equation $\varphi(t) = t$ to find it, but solving this equation exactly might be difficult. In Exercise 5.3.13 below, we provide a simple iterative algorithm that converges to the extinction probability ζ for the supercritcal case $\mu > 1$ at an exponential rate.

Exercise 5.3.13 (Approximating the extinction probability by fixed point iteration)**.** Suppose $\mu > 1$. In order to compute the extinction probability $\zeta = \mathbb{P}(\tau < \infty) \in [0,1)$, consider the following 'fixed point iterates': $\theta_0 = 0$ and

$$\theta_n := \varphi(\theta_{n-1}),$$

where $\varphi(s) = \mathbb{E}[s^{Z_1}]$ is the generating function of the offspring distribution.

(i) Show that $\theta_n = \varphi_n(0)$ for all $n \geq 1$.

(ii) Show that $\varphi'(\zeta) \in [0,1)$. (*Hint*: See Fig. 5.3.2.)

(iii) By induction, show that $\theta_n \leq \zeta$ for all $n \geq 0$.

(iv) Show that (*Hint*: Use convexity of φ)

$$\zeta - \theta_n = \varphi(\zeta) - \varphi(\theta_{n-1}) \leq \varphi'(\zeta)(\zeta - \theta_{n-1}).$$

By induction, deduce that

$$0 \leq \zeta - \theta_n \leq \varphi'(\zeta)^n.$$

Conclude that $\theta_n \nearrow \zeta$ at an exponential rate.

Our last point of investigation for the branching processes is how does the tail of the extinction time τ behaves. For instance, Theorem 5.3.12 shows that for $\mu \leq 1$, the extinction time τ is almost surely finite. But how likely is it to exceed a certain value? That is, how does the probability of the branching process surviving up to time n behave? We can answer this question again by using generating fuctions. The key relation is

$$\mathbb{P}(\tau > n) = \mathbb{P}(Z_n > 0) = 1 - \mathbb{P}(Z_n = 0) = 1 - \varphi_n(0).$$

Thus the speed at which $\mathbb{P}(\tau > n)$ decays is determined by the speed at which $\varphi_n(0)$ converges to one.

Exercise 5.3.14 (Tail of the extinction time in the subcritical regime)**.** Suppose $\mu < 1$ and $\mathbb{E}[Z_1^2] < \infty$.

(i) Use a similar approach illustrated in Exc. 5.3.13 to show that $1 - \varphi_n(0) \leq \mu^n$. Conclude that

$$\mathbb{P}(\tau > n) \leq \mu^n.$$

(ii)* By Taylor's theorem, show that

$$1 - \varphi_{n+1}(0) = \varphi(1) - \varphi(\varphi_n(0)) = \mu(1 - \varphi_n(0)) + O\big((1 - \varphi_n(0))^2\big).$$

Using convexity of φ, deduce that, for some constant $C > 0$,

$$\mu(1 - \varphi_n(0)) - C\big((1 - \varphi_n(0))^2\big) \leq 1 - \varphi_{n+1}(0) \leq \mu(1 - \varphi_n(0)).$$

Deduce that (using **(i)**)

$$1 - C\mu^{n-1} \le \frac{\mu^{-n-1}(1-\varphi_{n+1}(0))}{\mu^{-n}(1-\varphi_n(0))} \le 1.$$

Use Weierstrass' theorem on convergence of products to conclude that the limit

$$C_{\mathbf{p}} := \lim_{n\to\infty} \frac{\mu^{-n-1}(1-\varphi_{n+1}(0))}{(1-\varphi_0(0))} = \lim_{n\to\infty} \mu^{-n-1}(1-\varphi_{n+1}(0))$$

exists and is positive. Finally, conclude that

$$\mathbb{P}(\tau > n) \sim C_{\mathbf{p}}\,\mu^n.$$

Exercise 5.3.15 (Tail of the extinction time in the critical regime)**.** Suppose $\mu = 1$, $p_1 \neq 1$, and $\mathbb{E}[Z_1^2] < \infty$.

(i)* Show that there exists a constant $C > 0$ such that

$$\mathbb{P}(\tau > n) \sim C/n.$$

(*Hint:* Use the Taylor expansion of φ near $\zeta = 1$:

$$1 - \varphi_{n+1}(0) \approx 1 - \varphi_n(0) - \frac{\varphi''(1)}{2}(1-\varphi_n(0))^2.$$

Letting $x_n := 1-\varphi_n(0)$, they satisfy the recursion $x_{n+1} = x_n - bx_n^2$. Then making a further change of variable $y_n = x_n^{-1}$, deduce

$$y_{n+1} = \frac{y_n^2}{y_n - b} = y_n + \frac{b}{1-bx_n} = y_n + \frac{b}{1-o(1)}.$$

Hence y_n is asymptotically an arithmetic sequence.)

(ii) Use **(i)** to deduce that the extinction time has the following scaling property: For each $x > 1$,

$$\mathbb{P}(\tau > xn \,|\, \tau > n) = C/x + o(1).$$

5.4. Martingale concentration inequalities

Martingales can be used prove extremely useful (expeically in many machine learning and statistical applications) concentration inequalities without assuming much on independence structure of the increments. We will study a few of them in this section.

The first is a martingale-version of the classical Hoeffding's inequality called the Azuma-Hoeffding inequality, which is obtained by applying moment generating function, Markov's inequality, and optimization for best bound. This inequality is used a lot in ML, for inatence, in analyzing no regret algorithms for online sequential prediction [CBL06].

Theorem 5.4.1 (Azuma-Hoeffding inequality)**.** *Let $(X_k)_{0\le k\le n}$ be a martingale w.r.t. a filtration $(\mathscr{F}_k)_{0\le k\le n}$ with bounded increments: $|X_k - X_{k-1}| \le \sigma_k$ for some constant $\sigma_k \in (0,\infty)$ for $k = 1,\dots,n$. Then*

$$\mathbb{P}\left(X_n - X_0 \ge t\right) \le \exp\left(-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}\right). \tag{73}$$

If $\sigma_k = O(1)$, then the sum $\sum_{k=1}^n \sigma_k^2$ grows linearly in n. This growing denominator in the exponential function above is killed only when $t \gg \sqrt{n}$; if $t = o(\sqrt{n})$, then the concentration bound (73) is not useful. Thus (73) implies that the martingale X_n cannot exceed $\sqrt{n}$ by much with high probability. For instance, $\mathbb{P}(X_n \ge n) = \exp(-O(n))$ is exponentially small. If X_n had independent increments, this $O(\sqrt{n})$ scaling is expected by the central limit theorem (Thm. 4.4.5). We will use a 'variational Jensen's inequality' in Exc. 5.4.2 in the proof of Theorem 5.4.1.

PROOF OF THEOREM 5.4.1. The proof follows the usual recipe for proving Hoeffding's inequality but we use conditional expectation everywhere. Without loss of generality, assume $X_0 = 0$.

Write $\xi_n = X_n - X_{n-1}$ for the increments. Since $|\xi_n| \le \sigma_n$ a.s., ξ_n is a bounded RV so it has finite exponential moments everywhere: $\mathbb{E}[\exp(\theta\sigma_n)] < \infty$ for all $\theta \in \mathbb{R}$. Note that, for any parameter $\theta > 0$, by exponentiating and taking Markov's inequality,

$$\begin{aligned}\mathbb{P}(X_n \ge t) &\le \mathbb{P}(\exp(\theta X_n) \ge \exp(\theta t)) \\ &\le \exp(-\theta t)\mathbb{E}[\exp(\theta X_n)].\end{aligned} \tag{74}$$

Next we use iterated expectation to write

$$\begin{aligned}\mathbb{E}[\exp(\theta X_n)] &= \mathbb{E}[\mathbb{E}[\exp(\theta X_n)\,|\,\mathscr{F}_{n-1}]] \\ &= \mathbb{E}[\mathbb{E}[\exp(\theta X_{n-1} + \theta\xi_n)\,|\,\mathscr{F}_{n-1}]] \\ &= \mathbb{E}[\exp(\theta X_{n-1})\,\mathbb{E}[\exp(\theta\xi_n)\,|\,\mathscr{F}_{n-1}]]\end{aligned}$$

since $\exp(\theta X_{n-1}) \in \mathscr{F}_{n-1}$. Now, ξ_n conditional on $\mathscr{F}_{n-1}$ is a bounded random variable taking values from $[-\sigma_k, \sigma_k]$. By Exercise 5.4.2 (this is where we use the martingale condition $\mathbb{E}[\xi_n\,|\,\mathscr{F}_{n-1}] = 0$), we have

$$\mathbb{E}[\exp(\theta\xi_n)\,|\,\mathscr{F}_{n-1}]] \le \exp(\theta^2\sigma_n^2/2).$$

Since $\exp(\theta X_{n-1}) \ge 0$, it follows that

$$\mathbb{E}[\exp(\theta X_n)] \le \mathbb{E}[\exp(\theta X_{n-1})]\exp(\theta^2\sigma_n^2/2).$$

Proceeding by an induction, we deduce

$$\mathbb{E}[\exp(\theta X_n)] \le \exp\left(\frac{\theta^2}{2}\sum_{k=1}^n \sigma_k^2\right).$$

Combining with (74), we get

$$\mathbb{P}(X_n \ge t) \le \exp\left(-\theta t + \frac{\theta^2}{2}\sum_{k=1}^n \sigma_k^2\right).$$

The above bound holds for all $\theta > 0$ so we can optimize the above bound in θ. Note that the exponent in the RHS above is a convex quadratic function in θ, which is mimized at $\theta = \frac{t}{\sum_{k=1}^n \sigma_k^2}$ with minimum value $-t^2/2\sum_{k=1}^n \sigma_k^2$. This shows the assertion. □

Exercise 5.4.2 (A variational Jensen's inequality)**.** Let X be a mean zero RV taking values from an interval $[-A, B]$. Fix a convex function $\varphi : \mathbb{R} \to \mathbb{R}$. We will show that

$$\mathbb{E}[\varphi(X)] \le \varphi(-A)\frac{B}{A+B} + \varphi(B)\frac{A}{A+B}. \tag{75}$$

In words, over all possible distributions of X over $[-A, B]$, the most extreme distribution that maximizes $\mathbb{E}[\varphi(X)]$ is the one that puts point mass on $-A$ and B as in the right-hand side.

(i) Let Y be a RV taking values from $[0,1]$ and mean $p \in [0,1]$. Suppose that for any convex function $\psi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\psi(Y)] \le (1-p)\psi(0) + p\psi(1). \tag{76}$$

Then deduce (75) from this. (*Hint:* Rescale X and make appropriate change to φ.)

(ii) Here we will deduce (76). Let Y be as before. Let $U \sim \text{Uniform}(0,1)$ independent from Y. Argue that

$$\mathbf{1}(U \le Y)\,|\,Y \sim \text{Bernoulli}(Y) \quad \text{and} \quad \mathbf{1}(U \le Y) \sim \text{Bernoulli}(p).$$

(You may use Ex. 5.1.14 for the first part.) Then use Jensen's inequality to deduce

$$(1-p)\varphi(0) + p\varphi(1) = \mathbb{E}[\varphi(\mathbf{1}(U \le Y))] \ge \mathbb{E}[\varphi(Y)].$$

(iii) (Hoeffding's lemma) Let $\varphi(x) = e^{\theta x}$ for a fixed $\theta > 0$ and assume $A = B > 0$. Deduce that

$$\mathbb{E}[\exp(\theta X)] \leq \frac{\mathbb{E}[\exp(-\theta A)] + \mathbb{E}[\exp(\theta A)]}{2} \leq \exp(\theta^2 A^2/2).$$

A useful consequence of Azuma-Hoeffding's inequality is the following McDiarmid's inequality, which is also known as the 'bounded difference inequality'.

Theorem 5.4.3 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be indepdendent RVs. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a "Lipschitz" function in the following sense: If $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$ differ only in the kth coordinate, then*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq \sigma_k \tag{77}$$

for some constant $\sigma_k \in (0, \infty)$. Then for any $t > 0$,

$$\mathbb{P}\left(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}\right).$$

PROOF. Define a filtration $(\mathscr{F}_k)_{0 \leq k \leq n}$ by $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and $\mathscr{F}_k := \sigma(X_1, \ldots, X_k)$ for $k = 1, \ldots, n$. For $k = 1, \ldots, n$, denote

$$Y_k := \mathbb{E}[f(X_1, \ldots, X_n) \,|\, \mathscr{F}_k].$$

Then $|Y_k| \leq \sum_{k=1}^n \sigma_k$ by using the Lipschitz property of f, so $\mathbb{E}[|Y_k|] < \infty$. Clearly $Y_k \in \mathscr{F}_k$. Also by iterated expectation, $\mathbb{E}[Y_{k+1} \,|\, \mathscr{F}_k] = Y_k$, so $(Y_k)_{1 \leq k \leq n}$ is a martingale w.r.t. $(\mathscr{F}_k)_{1 \leq k \leq n}$[13]. Furthermore, by (77), $|Y_k - Y_{k-1}| \leq \sigma_k$. To see this, let X_k' be an independent copy of X_k, so $X_k' \perp \mathscr{F}_k$. Then (using Ex. 5.1.15)

$$\mathbb{E}[f(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n) \,|\, \mathscr{F}_k] = \mathbb{E}[f(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n) \,|\, \mathscr{F}_{k-1}] = Y_{k-1},$$

so by (77) and a triangle inequality,

$$\begin{aligned} |Y_k - Y_{k-1}| &= \left|\mathbb{E}[f(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) - f(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n) \,|\, \mathscr{F}_k]\right| \\ &\leq \sigma_k. \end{aligned}$$

Now the statement follows from Azuma-Hoeffding (Thm. 5.4.1) and noting that $Y_0 = \mathbb{E}[f(X_1, \ldots, X_n)]$. □

Below we will see some interesting applications of the martingale concentration inequalities in the context of Erdős-Rényi random graphs.

Definition 5.4.4 (Erdős-Renyí random graphs)**.** Construct a random graph with n nodes in the following manner. For each pair of nodes (i, j), include an edge ij independently with probability p and leave it as a non-adjacent pair with probability $1 - p$. The resulting random graph is denoted as $G(n, p)$ and is called a *Erdős-Renyí random graph.*

Exercise 5.4.5 (Number of triangles in $G(n, p)$)**.** Let $T = T(n, p)$ denote the total number of triangles in $G(n, p)$.

(i) For each three distinct nodes i, j, k in G, let $Y_{ijk} := \mathbf{1}(ij, jk, ki \in E)$, which is the indicator variable for the event that there is a triangle with node set $\{i, j, k\}$. Show that

$$Y_{ijk} \sim \text{Bernoulli}(p^3).$$

(ii) Show that we can write

$$T = \sum_{1 \leq i < j < k \leq n} \mathbf{1}(ij, jk, ki \in E). \tag{78}$$

Deduce that the expected number of triangles is

$$\mathbb{E}[T] = \binom{n}{3} p^3.$$

[13] This is called the 'exposure martingale' because we are revealing the random coordinates $X_1, \ldots, X_n$ one by one.

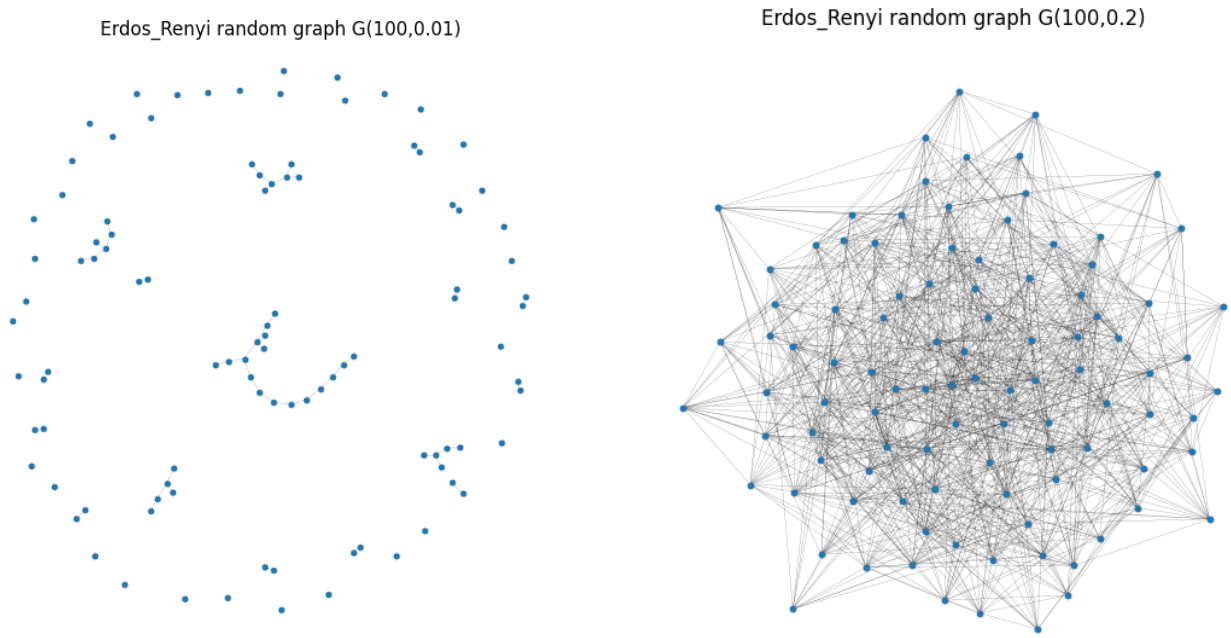


FIGURE 5.4.1. Samples of Erdős-Renyí random graphs.

(iii) Show that

$$\mathrm{Var}(T(n,p)) = \binom{n}{3}(p^3-p^6) + 12\binom{n}{4}(p^5-p^6) \sim \frac{n^4}{2}(p^5-p^6).$$

(*Hint:* First compute $\mathbb{E}[T^2]$ and use the fact that $\mathrm{Var}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2$. For computing $\mathbb{E}[T^2]$, use (78) and consider possible cases according to the number of overlapping edges.) Thus $\mathrm{Std}(T(n,p)) = \Theta(n^2)$. If CLT holds for $T(n,p)$, then $T(n,p)$ should fluctuate around its mean by $\Theta(n^2)$. Can we conclude this by CLT?

(iv) Show that for each $t \geq 0$,

$$\mathbb{P}\left(\left|T(n,p) - \binom{n}{3}p^3\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{n(n-1)(n-2)^2}\right).$$

Deduce that the above probability is $o(1)$ if $t \gg n^2$. Specifically, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|T(n,p) - \binom{n}{3}p^3\right| \geq n^{2+\varepsilon}\right) \leq 2\exp\left(-n^{2\varepsilon}\right).$$

Thus, McDiarmid's inequality almost confirms the upper tail of fluctuation of $T(n,p)$ predicted by CLT. (*Hint:* Let $X_1, \ldots, X_{\binom{n}{2}}$ denote the indicator of there being an edge for the kth pair of distinct nodes. Let $f(X_1, \ldots, X_{\binom{n}{2}})$ denote the number of triangles using the edges indicated by X_ks. Consider the "edge exposure filtration" $(\mathscr{F}_n)_{0 \leq n \leq \binom{n}{2}}$, where we reveal the connectedness of every pair of distinct nodes (i,j) sequentially. Argue that there at most $n-2$ triangles that contains a given edge. Then use Theorem 5.4.3.)

Exercise 5.4.6 (A moment bound for sub-exponential RVs)**.** Suppose X is a sub-exponential RV, i.e., there exists a constant $c > 0$ such that for all $x \geq 0$,

$$\mathbb{P}(|X| \geq x) \leq \exp(-cx).$$

Then show that for all integers $p \geq 1$,

$$\mathbb{E}[|X|^p] \leq c^{-p} p! \quad \text{for all } p \geq 1.$$

(*Hint:* Use the tail-sum formula 1.5.10 and integration by parts to show $\int_0^\infty e^{-cx} x^{p-1}\, dx = c^{-p}(p-1)!$.)

Generalize the above result to conditionally sub-exponential RVs. Namely, suppose

$$\mathbb{P}(|X| \geq x \,|\, \mathscr{F}) \leq \exp(-cx) \quad \text{for all } x \geq 0 \text{ almost surely}$$

where $\mathscr{F}$ is a sub σ-algebra. Show that for all integers $p \geq 1$,

$$\mathbb{E}[|X|^p \,|\, \mathscr{F}] \leq c^{-p} p! \quad \text{a.s. for all } p \geq 1.$$

Exercise 5.4.7 (A useful bound on MGF of mean-zero sub-exponential RVs)**.** Suppose X is a mean-zero sub-exponential RV, i.e., there exists a constant $c > 0$ such that for all $x \geq 0$,

$$\mathbb{P}(|X| \geq x) \leq \exp(-cx).$$

Then we have

$$\mathbb{E}[\exp(\lambda |X|)] \leq \exp(2(\lambda/c)^2) \quad \text{for all } 0 \leq \lambda \leq c/2.$$

(*Hint:* Use Taylor expansion, $\mathbb{E}[X] = 0$, the moment bounds in Exc. 5.4.6, and $1 + x \leq e^x$ for $x \geq 0$.)

Generalize the above result to conditionally sub-exponential RVs. Namely, suppose

$$\mathbb{P}(|X| \geq x \,|\, \mathscr{F}) \leq \exp(-cx) \quad \text{for all } x \geq 0 \text{ almost surely}$$

where $\mathscr{F}$ is a sub σ-algebra. Then show that

$$\mathbb{E}[\exp(\lambda |X|) \,|\, \mathscr{F}] \leq \exp(2(\lambda/c)^2) \quad \text{a.s. for all } 0 \leq \lambda \leq c/2.$$

Exercise 5.4.8 (Azuma-Hoeffding inequality for increments with exponential tail)**.** Let $(X_k)_{0 \leq k \leq n}$ be a martingale w.r.t. a filtration $(\mathscr{F}_k)_{0 \leq k \leq n}$. Suppose that the increment has an exponential tail: For all $1 \leq k \leq n$ and $t \geq 0$, almost surely,

$$\mathbb{P}(|X_k - X_{k-1}| \geq t \,|\, \mathscr{F}_{k-1}) \leq \exp(-ct),$$

where $c > 0$ is a constant. In this exercise, we will show that the following variant of Azuma-Hoeffding inequality holds:

$$\mathbb{P}\left(|X_n - X_0| \geq t\right) \leq 2\exp\left(-\frac{c^2 t}{8n}(t \wedge (2n/c))\right). \tag{79}$$

(i) Show that

$$\mathbb{P}(X_n \geq t) \leq \exp\left(-\theta t + \frac{2n\theta^2}{c^2}\right) \quad \text{for all } \theta \in [0, c/2].$$

(*Hint*: Use iterated expectation to write $\mathbb{E}[\exp(\theta X_n)] = \mathbb{E}[\exp(\theta X_{n-1})\mathbb{E}[\exp(\theta \xi_n) \,|\, \mathscr{F}_{n-1}]]$. By Exc. 5.4.7, the conditional MGF $\mathbb{E}[\exp(\theta \xi_n) \,|\, \mathscr{F}_{n-1}]$ is almost surely bounded above by $\exp(2(\theta/c)^2)$ for all $\theta \in [0, c/2]$. Then use Markov's inequality.)

(ii) Let $\bar{t} := t \wedge (2n/c)$. From **(ii)**, show that for all $t \geq 0$,

$$\mathbb{P}(X_n \geq t) \leq \exp\left(-\frac{c^2 t \bar{t}}{8n}\right).$$

Lastly, deduce (79) by a union bound.

(*c.f.* Due to the constaint $\theta \leq c/2$, we cannot simply choose θ to be the global minimizer $4nt/c^2$ for the quadratic function in **(ii)**, in case when $t > 2n/c$. This issue does not occur when ξ_k has bounded support.)

(*Hint:* For minimizing $-\theta t + \frac{L\theta^2}{2}$ over $\theta \in [0, C]$, let $\bar{t} := t \wedge CL$, write the objective as $-\theta(t - \bar{t}) + (-\theta \bar{t} + \frac{L}{2}\theta^2)$ and use the fact that the second quadratic term is minimized at $\theta = \bar{t}/L \leq C$. The objective value at such θ is at most $-\frac{\bar{t}t}{2L}$. Apply this with $L = 4n/c^2$ and $C = c/2$.)

(iii) Using (79), deduce that

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}|X_n - X_0| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{c^2t^2}{8}\right) & \text{for } t \leq \frac{2\sqrt{n}}{c} \\ 2\exp\left(-\frac{ct\sqrt{n}}{4}\right) & \text{for } t > \frac{2\sqrt{n}}{c}. \end{cases}$$

Thus, for 'small deviations' $t \leq \frac{2\sqrt{n}}{c}$, we have a Gaussian tail bound, as if the normalized sum had a normal distribution with constant variance (i.e., the central limit behavior). However, for 'large deviations' $t > \frac{2\sqrt{n}}{c}$, we have much heavier, sub-exponential tail bound. This is because a single increment ξ_k could be as large as $\sqrt{n}$ with probability $O(\exp(-c\sqrt{n}))$, which already matches the sub-exponential tail bound.

5.5. Doob's inequality and convergence in L^p for $p>1$

Recall that a sequence of RVs $X_n \to X$ for some X in L^p for some $p \in (0,\infty)$ if

$$\|X_n - X\|_p := \mathbb{E}[|X_n - X|^p]^{1/p} \to 0 \quad \text{as } n \to \infty.$$

In this section, we will investigate when does a martingale converge to the limit in the L^p sense for $p>1$, which is a stronger notion of convergence than the a.s. convergence in the martingale convergence theorem (Thm. 5.2.24).

In order to derive the key results in this section, we need to analyze submartingales evaluated at a bounded stopping time. The following lemma provides a natural result on this situation.

Lemma 5.5.1 (Optional stopping for bounded stopping time)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Let N be a stopping time such that $\mathbb{P}(N \leq k) = 1$. Then almost surely,*

$$\mathbb{E}[X_0] \leq \mathbb{E}[X_N] \leq \mathbb{E}[X_k].$$

In particular, if $(X_n)_{n\geq 1}$ is a martingale,

$$\mathbb{E}[X_0] = \mathbb{E}[X_N].$$

PROOF. Since $N \leq k$ almost surely, X_N is submartignale evaluated at some random time between 0 and k, so the result should be resonable.

By Theorem 5.2.17, $X_{n\wedge N}$ is a submartingale so

$$\mathbb{E}[X_0] \leq \mathbb{E}[X_{N\wedge 0}] \leq \mathbb{E}[X_{N\wedge k}] = \mathbb{E}[X_N],$$

where the equality follows from the fact that $N \leq k$ almost surely. This shows the first inequality in the assertion.

To show the second inequality, let $K_n := \mathbf{1}(N < n) = \mathbf{1}(N \leq n-1)$. Then $K = (K_n)_{n\geq 1}$ is bounded predictible and note that (recall (58))

$$\int_0^n K\,dX = \sum_{m=1}^n \left(1 - \mathbf{1}_{\{N\geq m\}}\right)(X_m - X_{m-1}) = X_n - X_{N\wedge n}.$$

By Exercise 5.2.21, $X_n - X_{N\wedge n}$ is a submartingale. So

$$\mathbb{E}[X_k] - \mathbb{E}[X_N] = \mathbb{E}[X_k] - \mathbb{E}[X_{N\wedge k}] = \mathbb{E}[X_k - X_{N\wedge k}] \geq \mathbb{E}[X_0 - X_{N\wedge 0}] = 0.$$

This finishes the proof. □

Example 5.5.2. Let $(S_n)_{n\geq 0}$ be a simple random walk on $\mathbb{Z}$ with $S_0 = 1$. Suppose $\mathbb{E}[S_1 - S_0] \geq 0$ so that S_n is a submartingale. Let $N = \inf\{n \geq 1 \mid S_n = 0\}$. Then

$$\mathbb{E}[S_0] = 1 > 0 = \mathbb{E}[0] = \mathbb{E}[S_N].$$

Therefore Lemma 5.5.1 does not hold. This is because the first hitting time of zero, N, is not bounded. ▲

Martingales are the most useful when stopped or evaluated at appropriately designed stopping times. The following result known as Doob's inequality is a classical example of this.

Theorem 5.5.3 (Doob's inequality)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Fix $\lambda>0$ and let $A:=\{\max_{0\leq m\leq n} X_m^+ > \lambda\}$. Then*

$$\lambda\mathbb{P}(A)\leq \mathbb{E}[X_n\mathbf{1}_A]\leq \mathbb{E}[X_n^+].$$

In particular,

$$\mathbb{P}\left(\max_{0\leq m\leq n} X_m^+ > \lambda\right)\leq \lambda^{-1}\mathbb{E}[X_n^+].$$

PROOF. Let $M:=\inf\{m\geq 1\mid X_m\geq\lambda\}$ and let $N:=M\wedge n$. Then N is a stopping time bounded by n. Since $X_N\geq\lambda$ on A, we have $X_N\mathbf{1}_A\geq\lambda\mathbf{1}_A$ almost surely, so

$$\lambda\mathbb{P}(A)\leq\mathbb{E}[X_N\mathbf{1}_A].$$

Also note that $X_N=X_n$ on A^c, so by Lemma 5.5.1 we have

$$\begin{aligned}\mathbb{E}[X_N\mathbf{1}_A]&=\mathbb{E}[X_N]-\mathbb{E}[X_N\mathbf{1}_{A^c}]\\&\leq\mathbb{E}[X_n]-\mathbb{E}[X_n\mathbf{1}_{A^c}]=\mathbb{E}[X_n\mathbf{1}_A].\end{aligned}$$

Combining the above inequalities, we obtain the first inequality in the assertion. The second inequality is trivial since $X_n\mathbf{1}_A\leq X_n^+$ almost surely. □

Example 5.5.4 (Kolmogorov's maximal inequality)**.** Let $(S_n)_{n\geq 0}$ be a random walk with independent increments ξ_i with $\mathbb{E}[\xi_i]=0$ and $\mathbb{E}[\xi_i^2]=\sigma_m^2\in(0,\infty)$. Then S_n is a martingale w.r.t. $\mathscr{F}_n:=\sigma(\xi_1,\dots,\xi_n)$, so by Prop. 5.2.12, S_n^2 is a submartingale w.r.t. $\mathscr{F}_n$. Now applying Doob's inequality (Thm. 5.5.3) with $\lambda=x^2$ for $x>0$ gives

$$\mathbb{P}\left(\max_{1\leq k\leq n}|S_k|\geq x\right)\leq x^{-2}\mathbb{E}[S_n^2]=x^{-2}\mathrm{Var}(S_n).$$

Thus we recover Kolmogorov's maximal inequality (see Exc. 3.7.1) from Doob's inequality. ▲

Theorem 5.5.5 (L^p maximum inequality)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale w.r.t. a filtration $(\mathscr{F}_n)_{n\geq 0}$. Denote $\overline{X}_n:=\max_{0\leq m\leq n}X_m^+$. Then for $1<p<\infty$,*

$$\mathbb{E}[\overline{X}_n^p]\leq\left(\frac{p}{p-1}\right)^p\mathbb{E}[(X_n^+)^p].$$

PROOF. Fix a constant $M\in(0,\infty)$. We will consider the truncated variable $\overline{X}_n\wedge M$ instead of $\overline{X}_n$ to ensure integrability. By Thm. 5.5.3, we have

$$\mathbb{P}(\overline{X}_n>x)\leq x^{-1}\mathbb{E}[X_n^+\mathbf{1}(\overline{X}_n>x)].$$

Note that $\{\overline{X}_n\wedge M>x\}=\{\overline{X}_n>x\}$ if $x<M$ and is $\emptyset$ if $x\geq M$. Thus the above yields

$$\mathbb{P}(\overline{X}_n\wedge M>x)\leq x^{-1}\mathbb{E}[X_n^+\mathbf{1}(\overline{X}_n\wedge M>x)].$$

Then by Prop. 1.5.10 and Fubini's theorem, we get

$$\begin{aligned}\mathbb{E}[(\overline{X}_n\wedge M)^p]&=\int_0^\infty px^{p-1}\mathbb{P}((\overline{X}_n\wedge M)>x)\,dx\\&\leq\int_0^\infty px^{p-2}\mathbb{E}[X_n^+\mathbf{1}(\overline{X}_n\wedge M>x)]\,dx\\&=\mathbb{E}\left[X_n^+\int_0^\infty px^{p-2}\mathbf{1}(\overline{X}_n\wedge M>x)\,dx\right]\\&=\frac{p}{p-1}\mathbb{E}\left[X_n^+\int_0^{\overline{X}_n\wedge M}(p-1)x^{p-2}\,dx\right]\\&=\frac{p}{p-1}\mathbb{E}\left[X_n^+(\overline{X}_n\wedge M)^{p-1}\right].\end{aligned}$$

Now let $q := p/(p-1)$ denote the conjugate of p. Applying Hölder's inequality (Prop. 1.3.13), we get

$$\mathbb{E}\left[X_n^+(\overline{X}_n \wedge M)^{p-1}\right] \le \mathbb{E}[|X_n^+|^p]^{1/p}\mathbb{E}[(\overline{X}_n \wedge M)^p]^{1/q}.$$

Since $1-(1/q) = 1-\frac{p-1}{p} = 1/p$, dividing both sides by $\mathbb{E}[(\overline{X}_n \wedge M)^p]^{1/q}$[14], it follows that

$$\mathbb{E}[(\overline{X}_n \wedge M)^p] = \mathbb{E}[(\overline{X}_n \wedge M)^{p(1-(1/q))}] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_n^+|^p].$$

Lastly, we let $M \nearrow \infty$ and use monotone convergence theorem (Thm. 1.3.19) to conclude. □

Theorem 5.5.6 (L^p convergence theorem)**.** *If X_n is a martingale with $\sup_{n\ge1}\mathbb{E}[|X_n|^p] < \infty$ for some $p>1$, then $X_n \to X$ a.s. and in L^p as $n\to\infty$.*

PROOF. Almost sure convergence follows directly from the martingale convergence theorem (Thm. 5.2.24) since $(\mathbb{E}[X_n^+])^p \le \mathbb{E}[|X_n|^p] \le \mathbb{E}[|X_n|]^p$. For the L^p convergence, we will use the fact that

$$|X_n - X|^p \le \left(2\sup_{n\ge0}|X_n|\right)^p,$$

which follows from $|X_n| \le \sup_{n\ge0}|X_n|$ and $|X| = \lim_{n\to\infty}|X_n| \le \sup_{n\ge0}|X_n|$. Thus if $\sup_{n\ge0}|X_n|$ is in L^p, then by dominated convergence theorem (Thm. 1.3.20), $\mathbb{E}[|X_n - X|^p] \to 0$ as desired. Indeed, $|X_n|$ is a submartingale (Prop. 5.2.12) so by Theorem 5.5.5,

$$\mathbb{E}\left[\left(\sup_{0\le n\le m}|X_n|\right)^p\right] \le \left(\frac{p}{p-1}\right)\mathbb{E}[|X_n|]^p < \infty.$$

Taking $m\to\infty$ and using MCT (Thm. 1.3.19), we deduce that $\sup_{n\ge0}|X_n|$ is in L^p, as desired. □

Exercise 5.5.7 (Orthogonality of martingale increments)**.** Let X_n be a martingale with $\mathbb{E}[X_n^2] < \infty$ for all n. Show that

(i) If $m \le n$ and $Y \in F_m$, then $\mathbb{E}[(X_n - X_m)Y] = 0$.

(ii) If $\ell \le m \le n$, then $\mathbb{E}[(X_n - X_m)(X_m - X_\ell)] = 0$.

(iii) (*Conditional variance formula*) If $m \le n$, then

$$\mathbb{E}[(X_n - X_m)^2 \,|\, \mathscr{F}_m] = \mathbb{E}[X_n^2 \,|\, \mathscr{F}_m] - X_m^2.$$

A nice application of the L^p martingale convergence theorem is the exponential growth of supercritical BP conditioned to survive.

Exercise 5.5.8 (Supercritical branching process on survival)**.** Consider supercritical branching process $(Z_n)_{n\ge0}$ (recall the notations in Sec. 5.3.3) with mean offspring number $\mu = \mathbb{E}[\xi_n^i] > 1$ and suppose $\mathrm{Var}(\xi_n^i) = \sigma^2 < \infty$. By Theorem 5.3.12, we know that the extinction probability $\zeta = \mathbb{P}(\tau < \infty)$ is nonzero, where τ denotes the extinction time (see (72)). By Prop. 5.3.7, we also know that $X_n := Z_n/\mu^n$ converges a.s. to some limiting RV X. It is reasonable that on the survival event $\tau = \infty$, X should be positive so the population grows asymptotically exponentially as $X\mu^n$. The goal of this exercise is to justify this:

$$\left\{X := \lim_{n\to\infty} Z_n/\mu_n > 0\right\} = \{\tau = \infty\}. \tag{80}$$

(i) Show that

$$\mathbb{E}[Z_n^2] = \mu^2\mathbb{E}[Z_{n-1}^2] + \sigma^2\mathbb{E}[Z_{n-1}].$$

(*Hint*: Use $\mathbb{E}[Z_n^2 \,|\, \mathscr{F}_{n-1}] = \mathbb{E}[(\xi_n^1 + \cdots + \xi_n^{Z_{n-1}}) \,|\, \mathscr{F}_{n-1}] = Z_{n-1}\mathbb{E}[(\xi_n^1)^2] + Z_{n-1}(Z_{n-1}-1)\mathbb{E}[\xi_n^1]^2$.)

[14] We can do so since $\mathbb{E}[(\overline{X}_n \wedge M)^p] < \infty$. Without the truncation, this is not necessarily true.

(ii) Deduce that for all $n \geq 1$,

$$\mathbb{E}[X_n^2] = \mathbb{E}[X_{n-1}^2] + \sigma^2/\mu^{n+1}.$$

By induction, show that

$$\mathbb{E}[X_n^2] = 1 + \sigma^2 \sum_{k=2}^{n+1} \mu^{-k}.$$

(iii) Show that $X_n \to X$ in L^2 and $\mathbb{E}[X_n] \to \mathbb{E}[X]$ for some RV X. (*Hint:* Use L^p martingale convergence and Jensen's inequality.)

(iv) Deduce that $\mathbb{E}[X] = 1$, so $\theta := \mathbb{P}(X = 0) < 1$. Show that θ satisfies the fixed point equation

$$\theta = \sum_{k=0}^{\infty} p_k \theta^k = \varphi(\theta),$$

where $\varphi(s) = \mathbb{E}[s^{Z_1}]$ is the generating function of the offspring distribution. Deduce that $\theta = \zeta = \mathbb{P}(\tau < \infty)$ and conclude (80). (*Hint:* "$\subseteq$" in (80) is trivial, so it is enough to show $\theta = \zeta$. For this, use a first-step analysis using the recursive property of BP. Then use Lemma 5.3.11.)

5.6. Uniform integrability and convergence in L^1

Given a sequence of RVs X_n converging almost surely to another RV X, when can we interchange the pointwise limit and integral and say

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

We can certainly do whenever we can apply any of the convergence theorems (i.e., MCT, BCT, DCT). But there are some counterexamples as well. One such example is given in Ex. 1.3.17. Below we give another one that will motivate a key definition in this section.

Example 5.6.1 (Mass escaping to infinity)**.** Consider $(\mathbb{R}, \mathscr{B}, \mu)$, where μ =Lebesgue measure. Let $f_n := \mathbf{1}_{[n,n+1]}$. Then $f_n \leq 1$ for $n \geq 1$ and $\int f_n \, d\mu = 1$ for all $n \geq 1$. Also, $f_n \to f = 0$ almost everywhere, since for each $x \in \mathbb{R}$, $f_n(x) = \mathbf{1}(n \leq x \leq n+1) \equiv 0$ for all $n > x$. Thus

$$0 = \int f \, d\mu \neq \lim_{n\to\infty} \int f_n \, d\mu = 1.$$

A similar but more probabilistic example is the following. Let $X_1, X_2, \cdots$ be independent RVs with

$$X_n = \begin{cases} (n+1)^2 & \text{with prob. } 1/(n+1)^2 \\ 0 & \text{with prob. } 1 - 1/(n+1)^2. \end{cases}$$

Then $\sum_{n\geq 1} \mathbb{P}(X_n \neq 0) = \sum_{n\geq 1} (n+1)^{-2} < \infty$, so X_n is nonzero only for finitely many n's almost surely by Borel-Cantelli. Hence $X_n \to X = 0$ almost surely. Now

$$0 = \mathbb{E}[X] \neq \lim_{n\to\infty} \mathbb{E}[X_n] = 1.$$

In both examples, some amount of mass is escaping to the infinity and the pointwise limit does not hold it in the limit. ▲

In this section, we will give necessary and sufficient conditions for a martingale to converge in L^1. The key to this is the following definition.

Definition 5.6.2 (Uniform integrability)**.** A collection of random variables $(X_i)_{i\in I}$ is said to be *uniformly integrable* (UI) if

$$\lim_{M\to\infty} \left(\sup_{i\in I} \mathbb{E}[|X_i| \mathbf{1}(|X_i| \geq M)] \right) = 0.$$

Note that uniform integrability implies a uniform bound on the expectations. Indeed, if we choose M large enough so that the above supremum is at most 1, then writing $\mathbb{E}[|X_i|] \le \mathbb{E}[|X_i|\mathbf{1}(|X_i| \ge M)] + M$, we have

$$\sup_{i\in I} \mathbb{E}[|X_i|] \le M + 1.$$

We will use this observation several times in this section.

Example 5.6.3 (Dominated sequence)**.** Let $(X_i)_{i\in I}$ be a collection of RVs domiated by an integrable RV Y, i.e., $|X_i| \le Y$ for all $i \in I$ and $\mathbb{E}[Y] < \infty$. Then $(X_i)_{i\in I}$ is UI. Indeed, note that

$$\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}(|X_i| \ge M)] \le \mathbb{E}[Y\mathbf{1}(Y \ge M)]. \tag{81}$$

Since $Y\mathbf{1}(Y < M) \nearrow Y$ a.s. $M \to \infty$, MCT (Thm. 1.3.19) yields $\mathbb{E}[Y\mathbf{1}(Y < M)] \to \mathbb{E}[Y] < \infty$. Consequently, $\mathbb{E}[Y\mathbf{1}(Y \ge M)] = \mathbb{E}[Y] - \mathbb{E}[Y\mathbf{1}(Y < M)] \to 0$ as $M \to \infty$. This shows the UI of $(X_i)_{i\in I}$. As a consequence, any uniformly bounded collection of RVs is UI. ▲

Another important collection of UI RVs is provided by the following result.

Proposition 5.6.4 (Collection of conditional expectations)**.** *Given a probability space $(\Omega, \mathscr{F}_0, \mathbb{P})$ and a RV $X \in L^1$, let I denote the set of all sub-σ algebras of $\mathscr{F}_0$. Then we will show that $(\mathbb{E}[X\,|\,\mathscr{F}])_{\mathscr{F}\in I}$ is uniformly integrable.*

PROOF. Fix $M > 0$, $\mathscr{F} \in I$, and denote $Y := \mathbb{E}[X\,|\,\mathscr{F}]$. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}[|Y|\mathbf{1}(|Y| \ge M)] &= \mathbb{E}[|\mathbb{E}[X\,|\,\mathscr{F}]|\mathbf{1}(|Y| \ge M)] \\ &\le \mathbb{E}[\mathbb{E}[|X|\,|\,\mathscr{F}]\mathbf{1}(|Y| \ge M)] \\ &= \mathbb{E}[|X|\mathbf{1}(|Y| \ge M)]. \end{aligned} \tag{82}$$

So while the bound looks similar to that in the dominated sequence case (81), it is not quite uniform over the choice in I.

To handle this issue, we claim that for each $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\mathbb{E}[|X|\mathbf{1}(A)] \le \varepsilon \quad \text{for all events } A \in \mathscr{F}_0 \text{ with } \mathbb{P}(A) \le \delta.$$

Suppose not. Then there exists some $\varepsilon > 0$ and a sequence of events $(A_n)_{n\ge1}$ in $\mathscr{F}_0$ such that for all $n \ge 1$,

$$\mathbb{P}(A_n) \le 1/n^2 \quad \text{and} \quad \mathbb{E}[|X|\mathbf{1}(A_n)] > \varepsilon.$$

But this is impossible since then $|X|\mathbf{1}(A_n) \to 0$ a.s. by Borel-Cantelli and $|X|\mathbf{1}(A_n) \le |X|$ with $X \in L^1$, so $\mathbb{E}[|X|\mathbf{1}(A_n)] \to 0$ as $n \to \infty$ by DCT.

We now go back to (82). Fix $\varepsilon > 0$. Then by the claim above, there exists $\delta > 0$ such that whenever $\mathbb{P}(A) \le \delta$, $\mathbb{E}[|X|\mathbf{1}(A)] \le \varepsilon$. Choose M large enough so that $\mathbb{E}[|X|]/M \le \delta$. Then by Markov's and Jensen's inequalities,

$$\mathbb{P}(|Y| \ge M) \le M^{-1}\mathbb{E}[|\mathbb{E}[X\,|\,\mathscr{F}]|] \le M^{-1}\mathbb{E}[\mathbb{E}[|X|\,|\,\mathscr{F}]] = M^{-1}\mathbb{E}[|X|] \le \delta,$$

regardless of the choice of $\mathscr{F} \in I$. Thus by the choice of δ, it follows that

$$\sup_{\mathscr{F}\in I} \mathbb{E}[|X|\mathbf{1}(|Y| \ge M)] \le \varepsilon.$$

This shows the desired UI. □

A common way to verify uniform integrability is by using the following result.

Proposition 5.6.5. *Let $\varphi : \mathbb{R} \to [0,\infty)$ be any function such that $\varphi(x) \gg x$. (e.g., $\varphi(x) = x^p$ for $p > 1$ or $\varphi(x) = x(0 \vee \log x)$). Then a collection $(X_i)_{i\in I}$ of RVs is UI if $\sup_{i\in I} \mathbb{E}[\varphi(|X_i|)] \le C$ for some constant $C > 0$.*

PROOF. For each $M > 0$, denote $\varepsilon_M := \sup\{x/\varphi(x) : x \geq M\}$. Note that for each $x \geq 0$,

$$x\mathbf{1}(x \geq M) = \frac{x}{\varphi(x)}\varphi(x)\mathbf{1}(x \geq M) \leq \varepsilon_M \varphi(x)\mathbf{1}(x \geq M).$$

It follows that

$$\mathbb{E}[|X_i|\mathbf{1}(|X_i| \geq M)] \leq \varepsilon_M \mathbb{E}[\varphi(|X_i|)\mathbf{1}(|X_i| \geq M)] \leq C\varepsilon_M.$$

By the hypothesis, $\varepsilon_M \to 0$ as $M \to \infty$. This verifies the UI. □

Uniform integrability is in fact necessary and sufficient for convergence in L^1, as the following result states.

Theorem 5.6.6 (UI $\iff$ L^1 convergence)**.** *Suppose that* $\mathbb{E}[|X_n|] < \infty$ *for all* $n \geq 1$. *If* $X_n \to X$ *in probability then the following are equivalent:*

(i) $\{X_n : n \geq 0\}$ *is UI;*

(ii) $X_n \to X$ *in* L^1*;*

(iii) $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|] < \infty$.

PROOF. "**(i)** $\Rightarrow$**(ii)**:" We wish to show that $\mathbb{E}[|X_n - X|] = o(1)$. To this effect, define the 'clipping function'

$$\varphi_M(x) := \begin{cases} M & \text{if } x > M \\ x & \text{if } |x| \leq M \\ -M & \text{if } x < -M. \end{cases}$$

By triangle inequality, write

$$|X_n - X| \leq |X_n - \varphi_M(X_n)| + |\varphi_M(X_n) - \varphi_M(X)| + |\varphi_M(X) - X|.$$

Note that for each $x \in \mathbb{R}$, $|\varphi_M(x) - x| = (|x| - M)^+ \leq |x|\mathbf{1}(|x| \geq M)$. Hence taking expecation, we get

$$\begin{aligned} \mathbb{E}[|X_n - X|] &\leq \mathbb{E}[|X_n|\mathbf{1}(|X_n| \geq M)] + \mathbb{E}[|X|\mathbf{1}(|X| \geq M)] + \mathbb{E}\left[|\varphi_M(X_n) - \varphi_M(X)|\right] \\ &\leq \underbrace{\sup_{n\geq 0} \mathbb{E}[|X_n|\mathbf{1}(|X_n| \geq M)]}_{=(a)} + \underbrace{\mathbb{E}[|X|\mathbf{1}(|X| \geq M)]}_{=(b)} + \underbrace{\mathbb{E}\left[|\varphi_M(X_n) - \varphi_M(X)|\right]}_{=(c)}. \end{aligned}$$

Note that (a) tends to zero as $M \to \infty$ by UI in **(i)**. For (b), first note that UI implies $\sup_{n\geq 1} \mathbb{E}[|X_n|] < \infty$. Then by the in-probability Fatou's lemma (Ex. 3.4.11)

$$\mathbb{E}[|X|] \leq \liminf_{n\to\infty} \mathbb{E}[|X_n|] < \infty.$$

Thus (b) tends to zero as $M \to \infty$ by MCT. Lastly, to show that (c) also tends to zero, we note that $X_n \to X$ in probability φ_M being continuous imply that $\varphi_M(X_n) \to \varphi_M(X)$ in probability as $n \to \infty$ (see Prop. 3.4.12). Since φ_M is also bounded, then BCT (Thm. 1.3.16) yields that (c) tends to zero as $n \to \infty$.

To finish, we fix $\varepsilon > 0$ and choose M large enough so that (a) and (b) combined are $\leq \varepsilon$. For this M, (c) tends to zero as $n \to \infty$, so this shows

$$\limsup_{n\to\infty} \mathbb{E}[|X_n - X|] \leq \varepsilon.$$

Now since $\varepsilon > 0$ was arbitrary, the limsup above must be zero.

"**(ii)** $\Rightarrow$**(iii)**:" By Jensen's and triangle inequalities,

$$|\mathbb{E}[|X_n| - |X|]| \leq \mathbb{E}\left[\left||X_n| - |X|\right|\right] \leq \mathbb{E}[|X_n - X|].$$

Since L^1 convergence means the last expression tends to zero, this yields **(iii)**.

"**(iii)** $\Rightarrow$**(i)**:" Fix $\varepsilon > 0$. For each $M \geq 0$, let ψ_M denote a continuous, piece-wise linear approximation of $x\mathbf{1}(x \geq M)$:

$$\psi_M(x) := \begin{cases} x & \text{if } x \in [0, M-1] \\ 0 & \text{if } x \geq M \\ \text{linear} & \text{if } |x| \leq M. \end{cases}$$

Since $X_n \to X$ in probability, $\psi_M(|X_n|) \to \psi_M(|X|)$ in probability (Prop. 3.4.12) so by BCT (Thm. 1.3.16), $\mathbb{E}[\psi_M(|X_n|)] \to \mathbb{E}[\psi_M(|X|)]$ as $n \to \infty$. Hence combining with **(iii)**, we can choose $N \geq 1$ large such that

$$|\mathbb{E}[|X_n| - |X|]| \leq \varepsilon/3 \quad \text{and} \quad |\mathbb{E}[\psi_M(|X_n|)] \to \mathbb{E}[\psi_M(|X|)]| \leq \varepsilon/3 \qquad \text{for all } n \geq N.$$

Next, by MCT, we will choose M large enough so that

$$\max_{1 \leq n < N} \mathbb{E}[|X_n| \mathbf{1}(|X_n| > M)] \leq \varepsilon.$$

Since $\psi_M(|X|) \to |X|$ a.s. as $M \to \infty$, $|\psi_M(X)| \leq |X|$, and $\mathbb{E}[|X|] < \infty$, DCT (Thm. 1.3.20) implies that $\mathbb{E}[\psi_M(|X|)] \to \mathbb{E}[|X|]$ as $M \to \infty$. Thus, we may choose larger M, if necessary, so that

$$\mathbb{E}[|X|] - \mathbb{E}[\psi_M(|X|)] \leq \varepsilon/3.$$

It follows that for all $n \geq N$,

$$\begin{aligned} \mathbb{E}[|X_n| \mathbf{1}(|X_n| > M)] &\leq \mathbb{E}[|X_n|] - \mathbb{E}[\psi_M(|X_n|)] \\ &= (\mathbb{E}[|X_n|] - \mathbb{E}[|X|]) + \big(\mathbb{E}[|X|] - \mathbb{E}[\psi_M(|X|)]\big) + \big(\mathbb{E}[\psi_M(|X_n|)] - \mathbb{E}[\psi_M(|X|)]\big) \leq \varepsilon. \end{aligned}$$

Then we conclude

$$\sup_{n \geq 1} \mathbb{E}[|X_n| \mathbf{1}(|X_n| > M)] \leq \max\left\{ \max_{1 \leq n < N} \mathbb{E}[|X_n| \mathbf{1}(|X_n| > M)], \sup_{n \geq N} \mathbb{E}[|X_n| \mathbf{1}(|X_n| > M)] \right\} \leq \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows the UI in **(i)**. □

We are now ready to state and derive the first main result in this section.

Theorem 5.6.7 (L^1 convergence of supermartingales)**.** *For a supermartingale, the following are equivalent:*

(i) *It is uniformly integrable;*

(ii) *It converges a.s. and in L^1;*

(iii) *It converges in L^1.*

PROOF. "**(i)** $\Rightarrow$**(ii)**:" UI implies $\sup \mathbb{E}[|X_n|] < \infty$ so the martingale convergence theorem (Thm. 5.2.24) implies $X_n \to X$ a.s. (hence in probability), and Theorem 5.6.6 implies $X_n \to X$ in L^1.

"**(ii)** $\Rightarrow$**(iii)**:" Trivial.

"**(iii)** $\Rightarrow$**(i)**:" $X_n \to X$ in L^1 implies $X_n \to X$ in probability by Markov's inequality. Hence **(i)** holds by Theorem 5.6.6. □

Next, we aim to deduce an analogue of Theorem 5.6.7 for martingales. We will use the following two simple observations in the proof.

Proposition 5.6.8. *Let $(X_n)_{n \geq 1}$ be a sequence of RVs on the common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that converges to some RV X in L^1. Then for any $A \in \mathscr{F}$, $\mathbb{E}[X_n \mathbf{1}(A)] \to \mathbb{E}[X\mathbf{1}(A)]$.*

PROOF. By linearity of expectation and Jensen's inequality,

$$|\mathbb{E}[X_n \mathbf{1}(A)] - \mathbb{E}[X\mathbf{1}(A)]| = |\mathbb{E}[(X_n - X)\mathbf{1}(A)]| \leq \mathbb{E}[|X_n - X|] = o(1).$$

□

Proposition 5.6.9. *If $(X_n)_{n \geq 0}$ is a martingale w.r.t. a filtration $(\mathscr{F}_n)_{n \geq 0}$ converging to some RV X in L^1, then $X_n = \mathbb{E}[X \,|\, \mathscr{F}_n]$.*

PROOF. Since $(X_n)_{n\geq 0}$ is a martingale, $\mathbb{E}[X_m \,|\, \mathscr{F}_n] = X_n$ for all $m \geq n$. By definition of conditional expectation, for each $A \in \mathscr{F}_n$ and $m \geq n$, $\mathbb{E}[X_m \mathbf{1}(A)] = \mathbb{E}[X_n \mathbf{1}(A)]$. By Prop. 5.6.8, we also have $\mathbb{E}[X_m \mathbf{1}(A)] \to \mathbb{E}[X\mathbf{1}(A)]$ as $m \to \infty$. It follows that

$$\mathbb{E}[X_n \mathbf{1}(A)] = \lim_{m\to\infty} \mathbb{E}[X_m \mathbf{1}(A)] = \mathbb{E}[X\mathbf{1}(A)].$$

Since the above holds for all $A \in \mathscr{F}_n$ and since $X_n \in \mathscr{F}_n$, it follows that $\mathbb{E}[X \,|\, \mathscr{F}_n] = X_n$. □

Theorem 5.6.10 (L^1 convergence of martingales)**.** *For a martingale, the following are equivalent:*

(i) *It is uniformly integrable;*

(ii) *It converges a.s. and in L^1;*

(iii) *It converges in L^1;*

(iv) *There is an integrable random variable X so that $X_n = \mathbb{E}[X \,|\, \mathscr{F}_n]$.*

PROOF. "**(i)** $\Rightarrow$**(ii)**:" Since martingales are also submartingales, this follows from Theorem 5.6.7.
"**(ii)** $\Rightarrow$**(iii)**:" Trivial.
"**(iii)** $\Rightarrow$**(iv)**:" Follows from Prop. 5.6.9.
"**(iv)** $\Rightarrow$**(i)**:" Follows from Prop. 5.6.4. □

Theorem 5.6.11 (Lévy's upward convergence)**.** *Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and let $(\mathscr{F}_n)_{n\geq 0}$ be filtration on Ω such that $\mathscr{F}_n \subseteq \mathscr{F}$ for $n \geq 0$. Denote $\mathscr{F}_\infty := \sigma(\bigcup_{n\geq 0} \mathscr{F}_n)$ (i.e., $\mathscr{F}_n \nearrow \mathscr{F}_\infty \subseteq \mathscr{F}$). Then for any $\mathscr{F}$-measurable and integrable RV X,*

$$\mathbb{E}[X \,|\, \mathscr{F}_n] \to \mathbb{E}[X \,|\, \mathscr{F}_\infty] \quad \text{as } n \to \infty \text{ a.s. and in } L^1.$$

In particular, if $X \in \mathscr{F}_\infty$, then

$$\mathbb{E}[X \,|\, \mathscr{F}_n] \to X \quad \text{as } n \to \infty \text{ a.s. and in } L^1.$$

PROOF. Denote $Y_n := \mathbb{E}[X \,|\, \mathscr{F}_n]$. The tower property of conditional expectation shows that $(Y_n)_{n\geq 0}$ is a martingale w.r.t. $(\mathscr{F}_n)_{n\geq 0}$. By Prop. 5.6.4, $(Y_n)_{n\geq 0}$ is uniformly integrable. Then by Theorem 5.6.10, we have that $Y_n \to Y_\infty$ a.s. and in L^1 for some RV Y_∞. Then by Prop. 5.6.9, for all $n \geq 0$,

$$\mathbb{E}[X \,|\, \mathscr{F}_n] = Y_n = \mathbb{E}[Y_\infty \,|\, \mathscr{F}_n].$$

It follows that, denoting $\mathscr{R} := \bigcup_{n\geq 0} \mathscr{F}_n$,

$$\mathbb{E}[X\mathbf{1}(A)] = \mathbb{E}[Y_\infty \mathbf{1}(A)] \quad \text{for all } A \in \mathscr{R}.$$

Since X, Y_∞ are $\mathscr{F}_\infty$-measurable and integrable, $A \mapsto \mathbb{E}[X\mathbf{1}(A)]$ and $A \mapsto \mathbb{E}[Y_\infty \mathbf{1}(A)]$ define measures on $\mathscr{F}_\infty = \sigma(\mathscr{R})$. Since $\mathscr{R}$ is a π-system (i.e., closed under intersection), the two measures must agree on the entire $\mathscr{F}_\infty$ (see Lem. 1.1.38). In other words,

$$\mathbb{E}[X\mathbf{1}(A)] = \mathbb{E}[Y_\infty \mathbf{1}(A)] \quad \text{for all } A \in \mathscr{F}_\infty = \sigma(\mathscr{R}).$$

Since $Y_\infty \in \mathscr{F}_\infty$, the definition of conditional expectation yields $\mathbb{E}[X \,|\, \mathscr{F}_\infty] = Y_\infty$. This shows the assertion. □

Exercise 5.6.12. In Theorem 5.6.11, further assume that $X \in L^p$ for some $p \geq 1$. Conclude that in Theorem 5.6.11, $X_n = \mathbb{E}[X \,|\, \mathscr{F}_n] \to \mathbb{E}[X \,|\, \mathscr{F}_\infty]$ in L^p. (*Hint*: Use Jensen's inequality to show that $\mathbb{E}[|X_n|^p] \leq \mathbb{E}[|X|^p] < \infty$. Then use $\sup_{n\geq 1} \mathbb{E}[|X_n|^p] < \infty$ and L^p maximum ineq. (Thm. 5.5.5) to deduce $\sup_{n\geq 1} |X_n| \in L^p$. Then conclude by DCT.)

The second part of Theorem 5.6.11 should be intuitive. Given the information $\mathscr{F}_\infty$, we can determine the value of X competely, so the best guess $\mathbb{E}[X \,|\, \mathscr{F}_n]$ of X given $\mathscr{F}_n$ should approach X. However, that the convergence occurs both almost surely and in L^1 is nontrivial.

An immediate consequence is the following Lévy's 0-1 law.

Corollary 5.6.13 (Lévy's 0-1 law)**.** *If $\mathscr{F}_n \nearrow \mathscr{F}_\infty$ and if $A \in \mathscr{F}_\infty$, then $\mathbb{E}[\mathbf{1}(A) \,|\, \mathscr{F}_n] \to \mathbf{1}(A)$ as $n \to \infty$ almost surely.*

PROOF. By Theorem 5.6.11, $\mathbb{E}[\mathbf{1}(A)\,|\,\mathscr{F}_n] \to \mathbb{E}[\mathbf{1}(A)\,|\,\mathscr{F}_\infty] = \mathbf{1}(A)$ a.s. and in L^1. □

For the same reason as before, the above result might not be very surprising. However, we can immediately deduce Kolmogorov's 0-1 law from it, so it should be far from trivial.

Example 5.6.14 (Kolmogorov's 0-1 law)**.** Let $X_1, X_2, \dots$ be independent RVs on the same probability space. For each $n \geq 1$, let $\mathscr{F}_n := \sigma(X_1, \dots, X_n)$ and let $\mathscr{T}$ denote the tail σ-algebra for $(X_n)_{n\geq 1}$. Let $A \in \mathscr{T}$. We would like to show that $\mathbb{P}(A) \in \{0,1\}$ (Kolmogorov's 0-1 law, see Exc. 3.7.8).

To this effect, first observe that $(\mathscr{F}_n)_{n\geq 1}$ is a filtration. Also since A is a tail event and X_n's are independent, $\mathbf{1}(A) \perp \mathscr{F}_n$ for each $n \geq 1$[15]. Hence by Lévy's 0-1 law, as almost surely as $n \to \infty$,

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}(A)\,|\,\mathscr{F}_n] \to \mathbf{1}(A).$$

Thus, the constant function $\omega \mapsto \mathbb{P}(A)$ is almost surely the same as the indicator function $\mathbf{1}(A)$. It follows that $\mathbb{P}(A) \in \{0,1\}$, as desired. ▲

Another nice consequence of Levy's upward convergence thoerem is the following DCT for conditional expectation.

Theorem 5.6.15 (Dominated convergence theorem for conditional expectations)**.** *Suppose $Y_n \to Y$ almost surely and $|Y_n| \leq Z$ for all n where $\mathbb{E}[Z] < \infty$. If $\mathscr{F}_n \nearrow \mathscr{F}_\infty$ then*

$$\mathbb{E}[Y_n|\mathscr{F}_n] \to \mathbb{E}[Y|\mathscr{F}_\infty] \quad \textit{almost surely.}$$

PROOF. Let $W_N = \sup\{|Y_n - Y_m| : n, m \geq N\}$. By triangle inequality $W_N \leq 2Z$, so $\mathbb{E}[W_N] < \infty$ by the hypothesis. Note that for $n \geq N$,

$$|Y_n - Y| = \lim_{m\to\infty} |Y_n - Y_m| \leq \lim_{m\to\infty} W_N = W_N.$$

Using monotonicity and applying Theorem 5.6.11 to W_N gives

$$\limsup_{n\to\infty} \mathbb{E}[|Y_n - Y||\mathscr{F}_n] \leq \lim_{n\to\infty} \mathbb{E}[W_N|\mathscr{F}_n] = \mathbb{E}[W_N|\mathscr{F}_\infty].$$

The above holds for all N and $W_N \searrow 0$ as $N \to \infty$, so continuity of conditional expectation from below (Prop. 5.1.21) implies $\mathbb{E}[W_N|\mathscr{F}_\infty] \searrow 0$. Then Jensen's inequality and the above give us

$$|\mathbb{E}[Y_n|\mathscr{F}_n] - \mathbb{E}[Y|\mathscr{F}_n]| \leq \mathbb{E}[|Y_n - Y||\mathscr{F}_n] \to 0 \quad \text{a.s.}$$

as $n \to \infty$. Theorem 5.6.11 then implies $\mathbb{E}[Y|\mathscr{F}_n] \to \mathbb{E}[Y|\mathscr{F}_\infty]$ almost surely. The desired result follows from the last two conclusions and the triangle inequality. □

Exercise 5.6.16 (Approximation of measurable function by stepfunction in L^1)**.** Let $f : [0,1) \to \mathbb{R}$ be a Borel measurable function. In this exercise, we will show that for each $k \geq 1$, there exists a stepfunction g_k with stepsize 2^{-k} such that $\|f - g_k\|_1 \to 0$ as $k \to \infty$, a well-known fact in real analysis[16]. We will use a filteration given by diadic partition and Levy's upward converence theorem (Thm. 5.6.11).

(i) Fix an integer $L \geq 1$ and denote the intervals $I_{L;i} := [\frac{i-1}{L}, \frac{i}{L})$ for $i = 1, \dots, L$ that partition $[0,1)$. Let U be an independent Uniform$([0,1])$ RV and let $\mathscr{F}_L$ denote the σ-algebra generataed by the events $\{U \in I_{L;i}\}$ for $i = 1, \dots, L$. Define

$$f_L := \mathbb{E}\left[f(U)\,|\,\mathscr{F}_L\right].$$

Show that f_L is the block average of f over the interval partition $[0,1) = I_{L;1} \sqcup \cdots \sqcup I_{L;L}$, that is, for each $\omega \in I_{L;i}$ for $i = 1, \dots, L$,

$$f_L(\omega) = \frac{1}{|I_i|} \int_{I_i} f(x)\,dx = L \int_{I_i} f(x)\,dx.$$

[15] If $X_n = X$ a.s. for all $n \geq 1$ for some RV X, clearly this is not the case.

[16] We can also use the fact that continuous functions are desne in $L^1([0,1])$ and then use uniform continuity for a direct construction.

(ii) Now take $L = 2^k$ for $k = 1, 2, \ldots$. Show that $(\mathscr{F}_{2^k})_{k\geq 0}$ defines a filtration and that $(f_{2^k})_{k\geq 0}$ is a martingale w.r.t. this filtration. Conclude that

$$\|f_{2^k} \to f\|_1 \to 0 \quad \text{as } k \to \infty.$$

(*Hint:* Use Lévy's upward convergence theorem).

Exercise 5.6.17 (Stepfunction approximation of graphons)**.** A symmetric integrable function $W : [0,1]^2 \to [0,1]$ is called a *graphon*, a continuum generalization of graphs which also arise as the limit object for sequences of dense graphs. A 'block graphon' is a special graphon that takes constant values over rectangles that partition $[0,1]^2$. Use the approach in Exc. 5.6.16 to show that, for each $k \geq 1$, there exists a block graphon W_k with square blocks of side lengths 2^{-k} such that

$$\|U - U_k\|_1 \to 0 \quad \text{as } k \to \infty.$$

5.7. Optional Stopping Theorems

In this section, we will prove several results that allow us to conclude that if X_n is a submartingale and $M \leq N$ are stopping times, then $\mathbb{E}[X_M] \leq \mathbb{E}[X_N]$. Such results are called 'optional stopping theorems' and have numerous applications, for instance in analyzing random walks. While this type of results are not always true (e.g., Ex 5.5.2), we have already seen such a result in Lemma 5.5.1 when N is bounded, which use used to prove Doob's inequality (Thm. 5.5.3). Our attention in this section will be focused on proving optional stopping theormes for the case of unbounded N.

We begin with a simple but useful observation.

Proposition 5.7.1 (Optional stopping for bounded submartingales)**.** *Let X_n be a submartingale and let N be a stopping time with $\mathbb{P}(N < \infty) = 1$. Suppose that $\sup_{n\geq 1} |X_{N\wedge n}| \leq Y$ a.s. for some integrable RV Y. Then $\mathbb{E}[X_0] \leq \mathbb{E}[X_N]$.*

PROOF. By Theorem 5.2.17, $X_{N\wedge n}$ is a submartingale. Also note that $X_{N\wedge n} \to X_N$ (which is well-defined since $N < \infty$ a.s.) a.s. as $n \to \infty$. Hence by DCT,

$$\mathbb{E}[X_0] \leq \mathbb{E}[X_{T_1\wedge n}] \to \mathbb{E}[X_N]$$

as $n \to \infty$. □

We can generalize the domination condition in the previous result to uniform integrability.

Theorem 5.7.2 (Optional stopping for UI submartingale I)**.** *Let $(X_n)_{n\geq 1}$ be a stochastic process adapted to a filtration $(\mathscr{F}_n)_{n\geq 1}$ and N be an almost surely finite stopping time. Suppose that $\mathbb{E}[|X_N|] < \infty$ and $X_n\mathbf{1}(N > n)$ is uniformly integrable. Then $X_{N\wedge n}$ is uniformly integrable. Furthermore, if X_n is a submartingale, then $\mathbb{E}[X_0] \leq \mathbb{E}[X_N]$.*

PROOF. Note that for $M > 0$,

$$\mathbf{1}(|X_n\mathbf{1}(N > n)| \geq M) = \mathbf{1}(|X_n| \geq M)\mathbf{1}(N > n) = \mathbf{1}(|X_n| \geq M)\mathbf{1}(N > n)^2.$$

Write $Y_n = X_n\mathbf{1}(N > n)$. Then

$$\begin{aligned}
\mathbb{E}[|X_{N\wedge n}|\mathbf{1}(|X_{N\wedge n}| \geq M)] &= \mathbb{E}[|X_{N\wedge n}|\mathbf{1}(|X_{N\wedge n}| \geq M)\mathbf{1}(N \leq n)] + \mathbb{E}[|X_{N\wedge n}|\mathbf{1}(|X_{N\wedge n}| \geq M)\mathbf{1}(N > n)] \\
&= \mathbb{E}[|X_N|\mathbf{1}(|X_N| \geq M)\mathbf{1}(N \leq n)] + \mathbb{E}[|X_n|\mathbf{1}(|X_n| \geq M)\mathbf{1}(N > n)] \\
&\leq \mathbb{E}[|X_N|\mathbf{1}(|X_N| \geq M)] + \mathbb{E}[|Y_n|\mathbf{1}(|Y_n| \geq M)].
\end{aligned}$$

Since $\mathbb{E}[|X_N|] < \infty$, the first term in the last expression tends to zero as $M \to \infty$ by DCT. The second term in the last expression also tends to zero as $M \to \infty$ since Y_n is UI.

Lastly, further assume that X_n is a submartingale. Since $X_{N\wedge n}$ is UI, $X_{N\wedge n} \to X_N$ as $n \to \infty$ in L^1 by Theorem 5.6.7. Then since $\mathbb{E}[X_0] \leq \mathbb{E}[X_{N\wedge n}]$ by Lemma 5.5.1, taking $n \to \infty$ shows $\mathbb{E}[X_0] \leq \mathbb{E}[X_N]$. □

Theorem 5.7.3 (Optional stopping for UI submartingale II)**.** *Let X_n be a uniformly integrable submartingale and let N be any stopping time.*

(i) *$X_{N\wedge n}$ is uniformly integrable.*

(ii) $\mathbb{E}[X_0] \le \mathbb{E}[X_N] \le \mathbb{E}[X_\infty]$, *where* $X_\infty = \lim_{n\to\infty} X_n$.

PROOF. Since X_n^+ is a supermartingale, Lemma 5.5.1 implies that $\mathbb{E}[X_{N\wedge n}^+] \le \mathbb{E}[X_n^+]$. Since X_n^+ is UI submartingale (since X_n is), we have

$$\sup_{n\ge 1} \mathbb{E}[X_{N\wedge n}^+] \le \sup_{n\ge 1} \mathbb{E}[X_n^+] < \infty.$$

Hence by martingale convergence theorem (Thm. 5.2.24), $X_{N\wedge n}$ converges almost surely as $n\to\infty$ to an integrable limit, which must be X_N. So $\mathbb{E}[|X_N|] < \infty$. Now $X_n \mathbf{1}(N > n)$ is UI since X_n is so. Hence we can conclude **(i)** from Lemma 5.7.2.

To show **(ii)**, first note that $\mathbb{E}[X_0] \le \mathbb{E}[X_{N\wedge n}] \le \mathbb{E}[X_n]$ by Lemma 5.5.1. Since both X_n and $X_{N\wedge n}$ are UI, Theorem 5.6.7 implies that they converge to X_∞ and X_N in L^1 as $n\to\infty$, respectively. Hence we can conclude. □

The next result on optional stopping does not require uniform integrability:

Proposition 5.7.4 (Optional stopping for nonnegative supermartingale)**.** *If X_n is a nonnegative supermartingale and N is a stopping time, then* $\mathbb{E}[X_0] \ge \mathbb{E}[X_N]$.

PROOF. Note that $\mathbb{E}[X_0] \ge \mathbb{E}[X_{N\wedge n}] \ge \mathbb{E}[X_n]$ by Lemma 5.5.1. Using the first inequality and Fatou's lemma,

$$\mathbb{E}[X_0] \ge \liminf_{n\to\infty} \mathbb{E}[X_{N\wedge n}] \ge \mathbb{E}\left[\liminf_{n\to\infty} X_{N\wedge n}\right] = \mathbb{E}[X_N].$$

□

The last optimal stopping theorem is for submartingales with increments whose conditional expectations are uniformly bounded. This result is quite useful since its hypothesis is easy to check for many applications (e.g., i.i.d. integerable increments).

Theorem 5.7.5 (Optional stopping for submartingales with increments of uniformly bounded conditional expectation)**.** *Suppose $(X_n)_{n\ge 1}$ is a submartingale and suppose that there exists a constant $B > 0$ such that $\mathbb{E}[|X_n - X_{n-1}| \,|\, \mathscr{F}_{n-1}] \le B$ a.s. for all $n \ge 1$. If N is a stopping time with $\mathbb{E}[N] < \infty$, then $X_{N\wedge n}$ is uniformly integrable and hence $\mathbb{E}[X_0] \le \mathbb{E}[X_N]$.*

PROOF. If we know that $X_{N\wedge n}$ is UI, then $X_{N\wedge n} \to X_N$ as $n\to\infty$ in L^1 by Theorem 5.6.7. Then since $\mathbb{E}[X_0] \le \mathbb{E}[X_{N\wedge n}]$ by Lemma 5.5.1, taking $n\to\infty$ shows $\mathbb{E}[X_0] \le \mathbb{E}[X_N]$.

It remains to show that $X_{N\wedge n}$ is UI. We begin by writing (recall (58))

$$|X_{N\wedge n}| \le \sum_{m=1}^{n} |X_m - X_{m-1}| \mathbf{1}(N \ge m) =: Y.$$

We will show that Y is integrable. Then by the result in Example 5.6.3, we can conclude that $X_{N\wedge n}$ is UI. To this end, note that since $\{N \ge m\} = \{N < m\}^c \in \mathscr{F}_{m-1}$,

$$\begin{aligned}
\mathbb{E}[|X_m - X_{m-1}| \mathbf{1}(N \ge m)] &= \mathbb{E}\left[\mathbb{E}\left[|X_m - X_{m-1}| \,|\, \mathscr{F}_{m-1}\right] \mathbf{1}(N \ge m)\right] \\
&\le \mathbb{E}\left[B \mathbf{1}(N \ge m)\right] \\
&= B\,\mathbb{P}(N \ge m).
\end{aligned}$$

Then

$$\mathbb{E}[|Y|] \le B \sum_{m=1}^{n} \mathbb{P}(N \ge m) \le B\,\mathbb{E}[N] < \infty.$$

This finishes the proof. □

5.7.1. Application of martingales to random walks. In this section, we harvest some interesting results from the optional stopping theorems for random walks. We first recall some notations. Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. increments with $\mathbb{E}[\xi_i] = \mu < \infty$. Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$, where S_0 is a constant. Denote $\mathscr{F}_n := \sigma(\xi_1, \ldots, \xi_n)$ and let $\mathscr{F}_0 = \{\emptyset, \Omega\}$. Then $(S_n)_{n\geq 0}$ defines a random walk with increments ξ_k. Recall the three martingales for random walks — the linear martingale (Ex. 5.2.4), the quadratic martingale (Ex. 5.2.5), and the exponential martingale (Ex. 5.2.7):

1. (*Linear martingale*) $X_n = S_n - \mu n$;

2. (*Quadratic martingale*) $X_n = S_n^2 - \sigma^2 n$, where $\mathbb{E}[\xi_k = 0]$ and $\mathbb{E}[\xi_k^2] = \sigma^2 < \infty$;

3. (*Exponential martingale*) $X_n = \exp(\theta S_n)/\varphi(\theta)$, where $\varphi(\theta) := \mathbb{E}[\exp(\theta \xi_k)] < \infty$.

We first deduce Wald's equaiton from the linear martingale.

Theorem 5.7.6 (Wald's equation)**.** *Let $(\xi_n)_{n\geq 1}$ be a sequence of i.i.d. increments with $\mathbb{E}[\xi_i] = \mu < \infty$. Let $S_n = \xi_1 + \cdots + \xi_n$, where S_0 is a constant. Let N be a stopping time with $\mathbb{E}[N] < \infty$. Then $\mathbb{E}[S_N] = \mu \mathbb{E}[N]$.*

PROOF. Let $X_n = S_n - \mu n$ be the linear martingale. Note that $\mathbb{E}[|X_n - X_{n-1}| \,|\, \mathscr{F}_{n-1}] = \mathbb{E}[|\xi_n|] = \mu$ for all $n \geq 1$. Hence by Theorem 5.7.5 applied to X_n and $-X_n$, we get

$$0 = \mathbb{E}[X_0] = \mathbb{E}[X_N] = \mathbb{E}[S_N] - \mu \mathbb{E}[N],$$

as desired. □

Theorem 5.7.7 (Symmetric Gambler's ruin)**.** *Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$ for $n \geq 0$ be a random walk with symmetric increments ξ_k with $\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = 1/2$. Fix integers $a \leq x \leq b$ and let $N := \inf\{n \geq 0 \,|\, S_n \in \{a, b\}$, the first time that S_n hits either a or b. Write $\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot \,|\, S_0 = x)$ and $\mathbb{E}_x(\cdot) = \mathbb{E}[\cdot \,|\, S_0 = x]$.*

(i) $\mathbb{P}_x(S_N = a) = \dfrac{b-x}{b-a}$ *and* $\mathbb{P}_x(S_N = b) = \dfrac{x-a}{b-a}$.

(ii) $\mathbb{E}_x[N] = (b-x)(x-a)$.

(iii) *Let $T_y := \inf\{n \geq 0 \,|\, S_n = y\}$. Then for $m \geq 1$,*

$$\mathbb{P}_1(T_M < T_0) = \frac{1}{M} \quad \text{and} \quad \mathbb{P}_1(T_M > T_0) = \frac{M-1}{M}.$$

(iv) $\mathbb{P}_1(T_0 < \infty) = 1$.

PROOF. First we show **(i)**. The key is to show optional stopping equation: $\mathbb{E}[S_N] = \mathbb{E}[S_0] = x$. Indeed, noting that $S_N \in \{a, b\}$ with probability one, we would deduce

$$\begin{aligned} x = \mathbb{E}[S_N] &= a\mathbb{P}_x(S_N = a) + b\mathbb{P}_x(S_N = b) \\ &= a\mathbb{P}_x(S_N = a) + b(1 - \mathbb{P}_x(S_N = a)), \end{aligned}$$

and solving this equaiton for $\mathbb{P}_x(S_N = a)$ gives the desired formula for it. Subtracting this from one also gives the desired formula for the other probability.

We will use Theorem 5.7.5 to verify optional stopping equation. For this, we only need to check $\mathbb{E}[N] < \infty$. For this, observe that a consecutive run of $+1$ increments of length $b-a$ will get the random walk out of (a, b). This occurs with probability $2^{-(b-a)}$. Hence in order to be confined in the interval (a, b) for $m(b-a)$ steps, we need to keep avoiding such runs m times. This gives

$$\mathbb{P}_x(N > m(b-a)) \leq \left(1 - 2^{-(b-a)}\right)^m$$

for all $m \geq 1$. Since the right-hand side decays exponentially fast, using the tail-sum formula the fact that $\mathbb{P}_x(N \geq k)$ decreases in k,

$$\mathbb{E}_x[N] = \sum_{k=0}^{\infty} \mathbb{P}(N \geq k) \leq \sum_{m=0}^{\infty} (b-a)\, \mathbb{P}(N \geq m(b-a)) < \infty.$$

Next, we show **(ii)**. For this, we will leverage the quardaric martingale $S_n^2 - n$ (note that $\sigma^2 = \mathbb{E}[\xi_1^2] = 1$). Applying Theorem 5.2.17, we get

$$x^2 = \mathbb{E}_x[S_0^2 - 0] = \mathbb{E}_x[S_{N\wedge n}^2 - (N \wedge n)].$$

By MCT, we have $\mathbb{E}[N \wedge n] \nearrow \mathbb{E}[N]$ as $n \to \infty$. Also, recall that $S_{N\wedge n} \in [a, b]$ for all $n \geq 0$, so by BCT,

$$\begin{aligned}\mathbb{E}_x[S_{N\wedge n}^2] \to \mathbb{E}_x[S_N^2] &= a^2 \frac{b-x}{b-a} + b^2 \frac{x-a}{b-a} \\ &= \frac{1}{b-a}(-ab(b-a) + x(b-a)(b+a)) \\ &= -ab + x(b+a).\end{aligned}$$

It follows that

$$\mathbb{E}[N] = -ab + x(b+a) - x^2 = (b-x)(x-a).$$

(iii) follows easily from the previous parts by taking $a = 0$, $x = 1$, and $b = M \geq 1$.

Lastly, **(iv)** follows from **(iii)** since $\{T_0 < T_M\} \nearrow \{T_0 < \infty\}$ as $M \to \infty$ (see continuity of measures from below, Thm. 1.1.16),

$$\mathbb{P}_1(T_0 < \infty) = \lim_{M\to\infty} \mathbb{P}_1(T_0 < T_M) = \lim_{M\to\infty} \frac{M-1}{M} = 1.$$

□

The technique we used in the proof of Theorem 5.7.7 **(ii)** is noteworthy. Namely, we first apply the optional stopping theorem for bounded stopping time $N \wedge n$ (Lem. 5.5.1) and the take $n \to \infty$ with appropriate convergence theorem.

Next, we 'apply' optional stopping to the exponential martingale and obtain the exponential moment generating function for the first hitting time. This will in tern give us the exact distribution of hitting time.

Theorem 5.7.8. *Let S_n be the simple symmetric random walk on $\mathbb{Z}$ with $S_0 = 0$. Let $T_1 = \inf\{n \geq 1 \mid S_1 = 1\}$. Then*

$$\mathbb{E}[s^{T_1}] = \frac{1 - \sqrt{1-s^2}}{s}.$$

Inverting the generating function, we get

$$\mathbb{P}(T_1 = 2n-1) = \frac{1}{2n-1} \cdot \frac{(2n)!}{n!\,n!} 2^{-2n}.$$

PROOF. Let $X_n = \exp(\theta S_n)/\phi(\theta)^n$ denote the exponential martingale, where $\phi(\theta) = \mathbb{E}[\exp(\theta\xi_1)]$ is the generating function of the increment. By Theorem 5.7.7 and the symmetry of SRW, we know $\mathbb{P}_0(T_1 < \infty) = 1$, so X_{T_1} is well-defined. Note that

$$\phi(\theta) = \frac{1}{2}(e^\theta + e^{-\theta}).$$

So $\phi(0) = 1$, $\phi'(0) = \mathbb{E}[\xi_1] = 0$, and $\phi'' = \phi > 0$ (convesity of a log MGF is a general fact, see Exc. 5.7.9). So $\phi(\theta) > 1$ for $\theta > 0$ in the domain. Choose any $\theta > 0$. Consider the stopped martingale $X_{T_1\wedge n}$ (Thm 5.2.17). Noting that $S_{T_1\wedge n} \leq 1$, for all $n \geq 1$,

$$0 \leq X_{T_1\wedge n} \leq \exp(\theta S_{T_1\wedge n}) \leq \exp(\theta).$$

Thus by BCT (Thm. 1.3.16)[17],

$$1 = \mathbb{E}[X_0] = \mathbb{E}[X_{T_1\wedge n}] \to \mathbb{E}[X_{T_1}] = \mathbb{E}\left[\exp(\theta)/\phi(\theta)^{T_1}\right].$$

[17] Here, we in fact have proved a version of optional stopping theorem for any stopping time N that holds when the stopped martingale $X_{N\wedge n}$ is uniformly bounded, by using the fact that $X_{N\wedge n}$ itself is a martingale and BCT. We can also apply Prop. 5.7.1 directly.

This yields

$$\mathbb{E}\left[\phi(\theta)^{-T_1}\right] = \exp(-\theta). \tag{83}$$

To convert the above as the probability generating function of T_1, choose s so that

$$\phi(\theta) = \frac{1}{2}(e^{\theta} + e^{-\theta}) = 1/s.$$

Substituting $e^{\theta} = x$, this becomes $x + x^{-1} = 2/s$, or $sx^2 - 2x + s = 0$. Solving for $x = e^{\theta} \geq 0$, we get

$$x = \frac{1 + \sqrt{1 - s^2}}{s}.$$

Now (83) gives

$$\sum_{m \geq 0} s^m \, \mathbb{P}(T_1 = m) = \mathbb{E}[s^{T_1}] = x^{-1} = \frac{s}{1 + \sqrt{1 - s^2}} = \frac{1 - \sqrt{1 - s^2}}{s}.$$

Expanding the function in the RHS as a power-series and matching the coefficients will then give us the formula for the moments of T_1. Some standard computation shows (denoting $\binom{x}{r} = x(x-1)\cdots(x-r+1)/r!$ for real $x > 0$)

$$\begin{aligned} \frac{1 - \sqrt{1 - s^2}}{s} &= \binom{1/2}{1} s - \binom{1/2}{3} s^3 + \binom{1/2}{5} s^5 - \cdots \\ &= \sum_{n \geq 1} \left(\frac{1}{2n-1} \cdot \frac{(2n)!}{n!\, n!} 2^{-2n} \right) s^{2n-1}. \end{aligned}$$

This shows the assertion. □

Exercise 5.7.9 (Convexity of log MGF)**.** Let Z be a random variable and let $\varphi(s) := \log \mathbb{E}[\exp(sZ)]$ denote its logarithmic moment generating function. Show that φ is convex on its domain. (*Hint:* Fix $\theta \in [0,1]$ and x, y in the domain of φ. Let $U := \exp((1-\theta)xZ)$ and $V := \exp(\theta x Z)$. Apply Hölder's inequality (Prop. 1.3.13) with $p = \frac{1}{1-\theta}$ and $q = \frac{1}{\theta}$ and take log.)

Lastly, we deduce some results on asymmetric Gamber's ruin.

Theorem 5.7.10 (Asymmetric Gambler's ruin)**.** *Let $S_n = S_0 + \xi_1 + \cdots + \xi_n$ for $n \geq 0$ be a random walk with symmetric increments ξ_k with $\mathbb{P}(\xi_k = 1) = p$ and $\mathbb{P}(\xi_k = -1) = 1 - p$ for some $p \neq 1/2$.*

(i) *Let $T_y := \inf\{n \geq 0 \mid S_n = y\}$ and $\varphi(y) := \left(\frac{1-p}{p}\right)^y$. Then for $a < x < b$,*

$$\mathbb{P}_x(T_a < T_b) = \frac{\varphi(b) - \varphi(x)}{\varphi(b) - \varphi(a)} \quad \text{and} \quad \mathbb{P}_x(T_b > T_a) = \frac{\varphi(x) - \varphi(a)}{\varphi(b) - \varphi(a)}$$

(ii) *Suppose $p \in (1/2, 1)$. If $a < 0$, then*

$$\mathbb{P}_0\left(\inf_{n \geq 0} S_n \leq a\right) = \mathbb{P}_0(T_a < \infty) = \left(\frac{1-p}{p}\right)^{-a}.$$

(iii) *Suppose $p \in (1/2, 1)$. If $b > 0$, then*

$$\mathbb{P}_0\left(\sup_{n \geq 0} S_n \geq b\right) = \mathbb{P}_0(T_b < \infty) = 1 \quad \text{and} \quad \mathbb{E}_0[T_b] = \frac{b}{2p-1}.$$

PROOF. Recall that $\varphi(S_n)$ is a martingale (Ex. 5.2.9). Let $N = T_a \wedge T_b$. As in the proof of Theorem 5.7.7, we can argue that $\mathbb{P}_x(N < \infty) = 1$. Since $S_{N \wedge n} \leq [a, b]$ for all $n \geq 1$, by Prop. 5.7.1, we have

$$\varphi(x) = \mathbb{E}[\varphi(S_0)] = \mathbb{E}[\varphi(S_N)] = \varphi(b)\mathbb{P}_x(T_b < T_a) + \varphi(a)\mathbb{P}_x(T_a < T_b).$$

Noting that $\mathbb{P}_x(T_a < T_b) + \mathbb{P}_x(T_b < T_a) = 1$, this gives **(i)**.

For **(ii)**, choose $x = 0$, $a < 0$, and $b > 0$. Observe that $\mathbb{P}_0(T_b \geq b) = 1$ since the increments are bounded by one. Hence $\{T_a < T_b\} \nearrow \{T_a < \infty\}$, so by **(i)** and letting $b \to \infty$,

$$\mathbb{P}_0(T_a < \infty) = \lim_{b\to\infty} \mathbb{P}_0(T_a < T_b) = \lim_{b\to\infty} \frac{\varphi(b) - \varphi(0)}{\varphi(b) - \varphi(a)} = \frac{1}{\varphi(a)},$$

where we have also used the fact that $\varphi(b) \to 0$ as $b \to \infty$ since $p > 1/2$.

For **(iii)**, proceeding as before,

$$\mathbb{P}_0(T_b < \infty) = \lim_{a\to-\infty} \mathbb{P}_0(T_b < T_a) = \lim_{b\to\infty} \frac{\varphi(0) - \varphi(a)}{\varphi(b) - \varphi(a)} = 1$$

where we have also used the fact that $\varphi(a) \to 0$ as $a \to -\infty$ since $p > 1/2$.

For the last conclusion (which makes sense since the average speed is $2p - 1$), we use the linear martingale $S_n - (2p-1)n$. By Theorem 5.2.17,

$$0 = \mathbb{E}\left[S_{T_b \wedge n}\right] - (2p-1)\mathbb{E}[T_b \wedge n].$$

By MCT, $\mathbb{E}[T_b \wedge n] \nearrow \mathbb{E}[T_b]$ as $n \to \infty$. Also, note that

$$b \geq S_{T_b \wedge n} \geq \inf_{n\geq 0} S_n$$

and by tail-sum formula and **(ii)**,

$$\mathbb{E}\left[\left|\inf_{n\geq 0} S_n\right|\right] = \mathbb{E}\left[-\inf_{n\geq 0} S_n\right] = \sum_{a\geq 0} \mathbb{P}\left(\inf_{n\geq 0} S_n \leq -a\right) < \infty.$$

Thus by DCT, $\mathbb{E}\left[S_{T_b \wedge n}\right] \to \mathbb{E}[S_{T_b}] = b$ as $n \to \infty$. It follows that

$$b = (2p-1)\,\mathbb{E}[T_b],$$

as desired. □

CHAPTER 6

Markov chains

In this chapter, we delve into the study of a pivotal class of stochastic processes known as *Markov chains*. Roughly speaking, Markov chains are used to model temporally changing systems where the future state depends only on the current state. For instance, if the price of bitcoin tomorrow depends only on its price today, then the bitcoin price can be modeled as a Markov process. (Of course, the entire history of prices often influences the decisions of buyers/sellers, so this assumption may not be realistic.)

The significance of Markov chains is twofold: Firstly, they find application across diverse realms including the physical, biological, social, and economic domains. Secondly, their theory is robustly established, enabling computation and accurate prediction through model utilization.

6.1. Definition and examples

In this chapter, we will mostly discuss Markov chains on a *countable* state space.[1]

Definition 6.1.1 (Markov chains on countable state spaces)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $\mathscr{S}$ be a countable set. Let $(X_n)_{n\geq 0}$ be a sequence of $\mathscr{S}$-valued random variables (i.e., $X_n : \Omega \to \mathscr{S}$ is $(\mathscr{F} - 2^{\mathscr{S}})$-measurable). We say $(X_n)_{n\geq 0}$ is a (time-homogeneous) *Markov chain* on state space $\mathscr{S}$ with transition matrix $P : \mathscr{S}^2 \to [0,1]$ and initial state $X_0 \in \mathscr{S}$ if for each $x' \in \mathscr{S}$, [2] [3]

$$\begin{aligned} \mathbb{P}(X_{n+1} = x' \,|\, \mathscr{F}_n) &\overset{a.s.}{=} \mathbb{P}(X_{n+1} = x' \,|\, X_n) \qquad \text{(Markov property)} \\ &= P(X_n, x'). \end{aligned} \tag{84}$$

When the state space $\mathscr{S}$ is finite, we will often identify it with the set of integers $\{1, 2, \ldots, m\}$ for $m = |\mathscr{S}|$. If $\mathscr{S}$ is countably infinite, then we will identify with the set $\mathbb{N}$ of natural numbers. Then P is a matrix of size $|\mathscr{S}| \times |\mathscr{S}|$. We denote

$$P = (p_{ij})_{i,j\in\Omega},$$

where the (i, j) entry p_{ij} has the meaning of the transition probability from state i to state j. When $|\mathscr{S}| = m$, the transition matrix P becomes the $m \times m$ matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix},$$

Observe that each row of P is a probability mass function (PMF) on $\mathscr{S}$. Since the state X_t of the chain is a (countable) $\mathscr{S}$-valued RV, we represent its PMF via a row vector

$$\mathbf{r}_t = [\mathbb{P}(X_t = 1), \mathbb{P}(X_t = 2), \cdots, \mathbb{P}(X_t = m), \ldots].$$

[1] Defining and constructing Markov chains on general (possibly uncountable) state spaces need more work. See [Dur19, Sec.5.2].

[2] The conditioning on the LHS of (84) is with respect to the restricted sub σ-algebra $\mathscr{F}'_n := \{\{X_n = x\} \cap A \,|\, A \in \mathscr{F}_n\}$.

[3] According to Exercise 6.1.6, an equivalent formulation of the Markov property (84) is, for all $x_0, \ldots, x_{n-1} \in \mathscr{S}$,

$$\mathbb{P}(X_{n+1} = x' \,|\, X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x' \,|\, X_n = x) = P(x, x').$$

Example 6.1.2. Let $\Omega = \{1,2\}$ and let $(X_t)_{t\geq 0}$ be a Markov chain on Ω with the following transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}.$$

We can also represent this Markov chain pictorially as in Figure 6.3.2, which is called the 'state space diagram' of the chain $(X_t)_{t\geq 0}$.

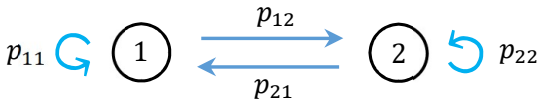


FIGURE 6.1.1. State space diagram of a 2-state Markov chain

For some concrete example, suppose

$$p_{11} = 0.2, \quad p_{12} = 0.8, \quad p_{21} = 0.6, \quad p_{22} = 0.4.$$

If the initial state of the chain X_0 is 1, then

$$\begin{aligned} \mathbb{P}(X_1 = 1) &= \mathbb{P}(X_1 = 1 \,|\, X_0 = 1)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_1 = 1 \,|\, X_0 = 2)\mathbb{P}(X_0 = 2) \\ &= \mathbb{P}(X_1 = 1 \,|\, X_0 = 1) = p_{11} = 0.2 \end{aligned}$$

and similarly,

$$\begin{aligned} \mathbb{P}(X_1 = 2) &= \mathbb{P}(X_1 = 2 \,|\, X_0 = 1)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_1 = 2 \,|\, X_0 = 2)\mathbb{P}(X_0 = 2) \\ &= \mathbb{P}(X_1 = 2 \,|\, X_0 = 1) = p_{12} = 0.8. \end{aligned}$$

Also we can compute the distribution of X_2. For example,

$$\begin{aligned} \mathbb{P}(X_2 = 1) &= \mathbb{P}(X_2 = 1 \,|\, X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1 \,|\, X_1 = 2)\mathbb{P}(X_1 = 2) \\ &= p_{11}\mathbb{P}(X_1 = 1) + p_{21}\mathbb{P}(X_1 = 2) \\ &= 0.2 \cdot 0.2 + 0.6 \cdot 0.8 = 0.04 + 0.48 = 0.52. \end{aligned}$$

In general, the distribution of X_{t+1} can be computed from that of X_t via a simple linear algebra. Note that for $i = 1,2$,

$$\begin{aligned} \mathbb{P}(X_{t+1} = i) &= \mathbb{P}(X_{t+1} = i \,|\, X_t = 1)\mathbb{P}(X_t = 1) + \mathbb{P}(X_{t+1} = i \,|\, X_t = 2)\mathbb{P}(X_t = 2) \\ &= p_{1i}\mathbb{P}(X_t = 1) + p_{2i}\mathbb{P}(X_t = 2). \end{aligned}$$

This can be written as

$$[\mathbb{P}(X_{t+1} = 2), \mathbb{P}(X_{t+1} = 2)] = [\mathbb{P}(X_{t+1} = 2), \mathbb{P}(X_{t+1} = 2)] \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}.$$

That is, if we represent the distribution of X_t as a row vector, then the distribution of X_{t+1} is given by multiplying the transition matrix P to the left. ▲

Example 6.1.3 (Gambler's ruin)**.** Suppose a gambler has fortune of k dolors initially and starts gambling. At each time he wins or loses 1 dolor independently with probability p and $1-p$, respectively. The game ends when his fortune reaches either 0 or N dolors. What is the probability that he wins N dolors and goes home happy?

We use Markov chains to model his fortune after betting t times. Namely, let $\Omega = \{0,1,2,\cdots,N\}$ be the state space. Let $(X_t)_{t\geq 0}$ be a sequence of RVs where X_t is the gambler's fortune after betting t times. We first draw the state space diagram for $N = 4$ below: Next, we can write down its transition probabilities as

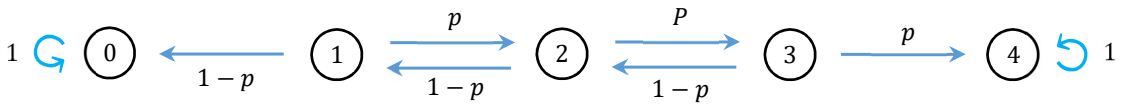


FIGURE 6.1.2. State space diagram of a 5-state gambler's chain

$$\begin{cases} \mathbb{P}(X_{t+1} = k+1 \,|\, X_t = k) = p & \forall 1 \le k < N \\ \mathbb{P}(X_{t+1} = k \,|\, X_t = k+1) = 1-p & \forall 1 \le k < N \\ \mathbb{P}(X_{t+1} = 0 \,|\, X_t = 0) = 1 \\ \mathbb{P}(X_{t+1} = N \,|\, X_t = N) = 1. \end{cases}$$

For example, the transition matrix P for $N = 5$ is given by

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 & 0 \\ 0 & 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We call the resulting Markov chain $(X_t)_{t\ge 0}$ the *gambler*'s chain. ▲

Example 6.1.4 (Ehrenfest Chain)**.** This chain is originated from the physics literature as a model for two cubical volumes of air connected by a thin tunnel. Suppose there are total N indistinguishable balls split into two "urns" A and B. At each step, we pick up one of the N balls uniformly at random, and move it to the other urn. Let X_t denote the number of balls in urn A after t steps. This is a Markov chain called the *Ehrenfest chain*. (See the state space diagram in Figure 6.1.3.)

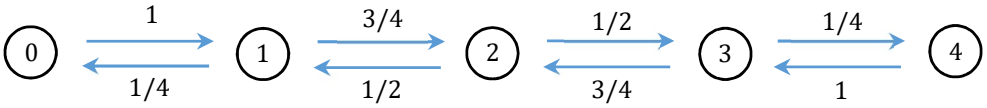


FIGURE 6.1.3. State space diagram of the Ehrenfest chain with 4 balls

It is easy to figure out the transition probabilities by considering different cases. If $X_t = k$, then urn B has $N-k$ balls at time t. If $0 < k < N$, then with probability k/N we move one ball from A to B and with probability $(N-k)/N$ we move one from B to A. If $k = 0$, then we must pick up a ball from urn B so $X_{t+1} = 1$ with probability 1. If $k = N$, then we must move one from A to B and $X_{t+1} = N-1$ with probability 1. Hence, the transition kernel is given by

$$\begin{cases} \mathbb{P}(X_{t+1} = k+1 \,|\, X_t = k) = (N-k)/N & \forall 0 \le k < N \\ \mathbb{P}(X_{t+1} = k-1 \,|\, X_t = k) = k/N & \forall 0 < k \le N \\ \mathbb{P}(X_{t+1} = 1 \,|\, X_t = 0) = 1 \\ \mathbb{P}(X_{t+1} = N-1 \,|\, X_t = N) = 1. \end{cases}$$

For example, the transition matrix P for $N = 5$ is given by

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/5 & 0 & 4/5 & 0 & 0 & 0 \\ 0 & 2/5 & 0 & 3/5 & 0 & 0 \\ 0 & 0 & 3/5 & 0 & 2/5 & 0 \\ 0 & 0 & 0 & 4/5 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

▲

Next, we take a look at an example of an important class of Markov chains, which is called the random walk on graphs. This is the basis of many algorithms involving machine learning on networks (e.g., Google's PageRank).

Example 6.1.5 (Random walk on graphs)**.** A *graph* G consists of a pair (V,E) of sets of nodes V and edges $E \subseteq V^2$. A graph G can be concisely represented as a $|V| \times |V|$ matrix A, which is called the *adjacency matrix* of G. Namely, the (i,j) entry of A is defined by

$$A(i,j) := \mathbf{1}(\text{nodes } i \text{ and } j \text{ are adjacent in } G) = \mathbf{1}(i \sim j).$$

We say G is *simple* if $(i,j) \in E$ implies $(j,i) \in E$ and $(i,i) \notin E$ for all $i \in V$. For a simple graph $G = (V,E)$, we say a node j is *adjacent* to i if $(i,j) \in E$. We denote $\deg_G(i)$ the number of neighbors of i in G, which we call the *degree* of i.

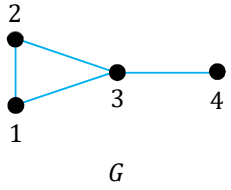
Consider we hop around the nodes of a given simple graph $G = (V,E)$: at each time, we jump from one node to one of the neighbors with equal probability. For instance, if we are currently at node 2 and if 2 is adjacent to 3, 5, and 6, then we jump to one of the three neighbors with probability $1/3$. The location of this jump process at time t can be described as a Markov chain. Namely, a Markov chain $(X_t)_{t\geq 0}$ on the node set V is called a *random walk on* G if

$$\mathbb{P}(X_{t+1} = j \,|\, X_t = i) = \frac{A(i,j)}{\deg_G(i)}.$$

Note that its transition matrix P is obtained by normalizing each row of the adjacency matrix A by the corresponding degree. That is,

$$P = D^{-1}A,$$

where D is the diagonal matrix of the degrees of nodes in G (i.e., the degree matrix of G).



$$A_G = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

FIGURE 6.1.4. A 4-node simple graph G, its adjacency matrix A, and associated random walk transition matrix P

Random walk on G can be defined (and more importantly, simulated) by the following recursive sampling rule:

$$X_{t+1} \,|\, X_t \quad \sim \quad \text{Uniform}(N(X_t)),$$

where $N(v)$ denote the set of all neighbors of node v in G. So X_{t+1} given X_t is chosen uniformly at random among the neighbors of X_t. ▲

Exercise 6.1.6 (Natural filtration for stochastic processes with countable state spaces)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(\mathscr{S}, 2^{\mathscr{S}})$ be a countable measurable space. Let $(X_n)_{n\geq 0}$ be a sequence of $\mathscr{S}$-valued random variables. Let $\mathscr{F}_n := \sigma(X_0, \ldots, X_n)$. Show that

$$\mathscr{F}_n = \sigma\left(\{\{X_0 = x_0, \cdots, X_n = x_n\} : x_0, \ldots, x_n \in \mathscr{S}\}\right).$$

Exercise 6.1.7. Repeat rolling two four sided dices with numbers 1,2,3, and 4 on them. Let Y_k be the some of the two dice at the kth roll. Let $S_n = Y_1 + Y_2 + \cdots + Y_n$ be the total of the first n rolls, and define $X_t = S_t \,(\text{mod} 6)$. Show that $(X_t)_{t\geq 0}$ is a Markov chain on the state space $\Omega = \{0,1,2,3,4,5\}$. Furthermore, identify its transition matrix.

We generalize our observation in Example 6.1.2 in the following exercise.

Exercise 6.1.8 (Time evolution and left-multiplication with transition matrix)**.** Let $(X_t)_{t\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. Let $\mathbf{r}_t$ denote the row vector of the distribution of X_t.

(i) Show that for each $t \geq 0$,

$$\mathbf{r}_{t+1} = \mathbf{r}_t P.$$

(ii) Show by induction that for each $t \geq 0$,

$$\mathbf{r}_t = \mathbf{r}_0 P^t.$$

Exercise 6.1.9. Let $(X_t)_{t\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P.

(i) (Multi-step transition prob.) Show that for each $x, y \in \mathscr{S}$ and $a, b \geq 1$,

$$\mathbb{P}\left(X_{a+b} = y \,|\, X_a = x\right) = P^b(x, y).$$

Hint: Use Exercise 6.1.8.

(ii) (The Chapman-Kolmogorov eq.) Show that for each $x, y \in \mathscr{S}$ and $n, m \geq 1$,

$$P^{n+m}(x, y) = \sum_{z\in\Omega} P^n(x, z)P^m(z, y).$$

While right-multiplication of P advances a given row vector of distribution one step forward in time, left-multiplication of P on a column vector computes the expectation of a given function with respect to the one-step future distribution. This point is clarified in the following exercise.

Exercise 6.1.10 (Right-multiplication by transition matrix)**.** Let $(X_t)_{t\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. Let $f : \mathscr{S} \to \mathbb{R}$ be a function. Let $\mathbf{f} = [f(1), f(2), \cdots]^T$ be the column vector representing the reward function f. Show that

$$[P\mathbf{f}](i) = \sum_{j\in\mathscr{S}} P(i, j)f(j) = \mathbb{E}[f(X_1) \,|\, X_0 = i].$$

That is, $P\mathbf{f}$ is a function on the state space whose value on each state i is the one-step conditional expectation of $f(X_1)$ given $X_0 = i$.

Consider quadratic forms of the form $\mathbf{r}P\mathbf{v}$, where P is the Markov transition matrix, $\mathbf{r}$ is a PMF on the state space, and $\mathbf{v}$ is a column vector reprenting a function f on the state space. The value of this quaratic form is in fact the expectation $\mathbb{E}_{X_0\sim\mathbf{r}}[f(X_1)]$, where the initial state X_0 is drawn from $\mathbf{r}$. A generalization of this observation is given in the following exercise.

Exercise 6.1.11 (Expected reward in the future)**.** Let $(X_t)_{t\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. Let $f : \mathscr{S} \to \mathbb{R}$ be a function. Suppose that if the chain X_t has state x at time t, then we get a 'reward' of $f(x)$. Let $\mathbf{r}_t = [\mathbb{P}(X_t = 1), \mathbb{P}(X_t = 2), \ldots]$ be the distribution of X_t. Let $\mathbf{v} = [f(1), f(2), \cdots]^T$ be the column vector representing the reward function f.

(i) Show that the expected reward at time t is given by

$$\mathbb{E}[f(X_t)] = \sum_{i\in\mathscr{S}} f(i)\mathbb{P}(X_t = i) = \mathbf{r}_t\mathbf{v}.$$

(ii) Use part (i) and Exercise 6.1.8 to show that

$$\mathbb{E}[f(X_t)] = \mathbf{r}_0 P^t\mathbf{v}.$$

(iii) The total reward up to time t is a RV given by $R_t = \sum_{k=0}^t f(X_t)$. Show that

$$\mathbb{E}[R_t] = \mathbf{r}_0(I + P + P^2 + \cdots + P^t)\mathbf{v}.$$

If a stochastic process depends on the past two consecutive states, then one can simply extend the state space into the set of paired states and define a Markov chain on that extended state space. The following example illustrates this.

Exercise 6.1.12 (Multi-step dependence)**.** Suppose that the probability it rains today is 0.4 if neither of the last two days was rainy, but 0.5 if at least one of the last two days was rainy. Let $\Omega = \{S, R\}$, where S=sunny and R=rainy. Let W_t be the weather of day t.

(i) Show that $(W_t)_{t\geq 0}$ is *not* a Markov chain.

(ii) Expand the state space into the set of pairs $\Sigma := \Omega^2$. For each $t \geq 0$, define $X_t = (W_{t-1}, W_t) \in \Sigma$. Show that $(X_t)_{t\geq 0}$ is a Markov chain on Σ. Identify its transition matrix.

(iii) What is the two-step transition matrix?

(iv) What is the probability that it will rain on Wednesday if it didn't rain on Sunday and Monday?

6.2. Strong Markov property

6.2.1. Strong Markov property. A fundamental property of Markov chains is that one can restart a Markov chain upon any stopping time and the restarted process is itself a Markov chain with the same transition matrix and is independent from the past. This is called the strong Markov property, since specializing it for deterministic stopping times yield the usual Markov property. Moreover, the strong Markov property allows one to use 'probabilistic induction' for Markov chains.

Theorem 6.2.1 (Strong Markov property)**.** *Let $(X_n)_{n\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. Let τ denote a stopping time w.r.t. the natural filtration $\mathscr{F}_n := \sigma(X_0, \dots, X_n)$. Define the sub-$\sigma$-algebra*

$$\mathscr{F}_\tau := \{A \in \mathscr{F} : \{\tau \leq n\} \cap A \in \mathscr{F}_n \ \forall n \geq 0\}.$$

For each integer $k \geq 0$, denote $X_{t\geq k} := (X_k, X_{k+1}, \dots)$. Then on the event that $\tau < \infty$, the future process $X_{t\geq \tau}$ given X_τ is a Markov chain with transition matrix P that is is independent of $\mathscr{F}_\tau$. That is, if S is an event in $\mathscr{S}^{\mathbb{N}}$, then on the event $\tau < \infty$, for any initial distribution μ,

$$\mathbb{E}_\mu\left[\mathbf{1}(X_{t\geq\tau} \in S) \,|\, \mathscr{F}_\tau\right] = \mathbb{E}_{X_\tau}\left[\mathbf{1}(X_{t\geq 0} \in S)\right].$$

(see Ex. 5.1.8.)

PROOF. Fix arbitrary $A \in \mathscr{F}_\tau$ and $n \geq 0$. Note that

$$\begin{aligned}
\mathbb{E}_\mu\left[\mathbf{1}(X_{t\geq n} \in S)\mathbf{1}(\tau = n)\mathbf{1}(A)\right] &\overset{(a)}{=} \mathbb{E}_\mu\left[\mathbb{E}[\mathbf{1}(X_{t\geq n} \in S) \,|\, \mathscr{F}_n]\mathbf{1}(\tau = n)\mathbf{1}(A)\right] \\
&\overset{(b)}{=} \mathbb{E}_\mu\left[\mathbb{E}[\mathbf{1}(X_{t\geq n} \in S) \,|\, X_n]\mathbf{1}(\tau = n)\mathbf{1}(A)\right] \\
&\overset{(c)}{=} \mathbb{E}_\mu\left[\mathbb{E}[\mathbf{1}(X_{t\geq \tau} \in S) \,|\, X_\tau]\mathbf{1}(\tau = n)\mathbf{1}(A)\right].
\end{aligned}$$

Here, (a) follows from the definition of conditional expectation given $\mathscr{F}_n$ with $\mathbf{1}(\tau = n)\mathbf{1}(A) \in \mathscr{F}_n$ and (b) follows from the Markov property. For (c), we have used that

$$\begin{aligned}
\mathbb{E}[\mathbf{1}(X_{t\geq n} \in S) \,|\, X_n]\mathbf{1}(\tau = n) &= \mathbb{E}_{X_n}[\mathbf{1}(X_{t\geq 0} \in S)]\mathbf{1}(\tau = n) \\
&= \mathbb{E}_{X_\tau}[\mathbf{1}(X_{t\geq 0} \in S)]\mathbf{1}(\tau = n) \\
&= \mathbb{E}[\mathbf{1}(X_{t\geq \tau} \in S) \,|\, X_\tau]\mathbf{1}(\tau = n).
\end{aligned}$$

Then since $A \in \mathscr{F}_\tau$ was arbitrary and since $\mathbb{E}[\mathbf{1}((X_{\tau+k})_{k\geq 1} \in S) \,|\, X_\tau]\mathbf{1}(\tau = n) \in \mathscr{F}_\tau$, this shows that

$$\mathbb{E}_\mu\left[\mathbf{1}((X_{n+k})_{k\geq 1} \in S)\mathbf{1}(\tau = n) \,|\, \mathscr{F}_\tau\right] = \mathbb{E}_\mu[\mathbf{1}((X_{\tau+k})_{k\geq 1} \in S) \,|\, X_\tau]\mathbf{1}(\tau = n).$$

Now summing over all $n \geq 0$ and using Fubini's theorem with $\mathbf{1}(\tau < \infty) \in \mathscr{F}_\tau$,

$$\mathbb{E}_\mu\left[\mathbf{1}((X_{\tau+k})_{k\geq 1} \in S) \,|\, \mathscr{F}_\tau\right]\mathbf{1}(\tau < \infty) = \mathbb{E}_\mu[\mathbf{1}((X_{\tau+k})_{k\geq 1} \in S) \,|\, X_\tau]\mathbf{1}(\tau < \infty).$$

This shows the desired conclusion. □

Exercise 6.2.2 (Restarting Markov chain at stopping times)**.** Let $(X_n)_{n\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. Let τ denote an a.s. finite stopping time w.r.t. the natural filtration $\mathscr{F}_n := \sigma(X_0, \dots, X_n)$. Show that $(X_{\tau+k})_{k\geq 0}$ is a Markov chain with transition matrix P. Furthermore, show that it is independent from $(X_n)_{1\leq n<\tau}$ given X_τ. (*Hint*: The independence from the past

follows directly from the strong Markov property. For the first part, partition the sample space according to the value of τ.)

6.2.2. Irreducibility, transience, and recurrence. Being able to restart a Markov chain at stopping times using the strong Markov property allows us to decompose the trajectory of the chain into i.i.d. 'excursions' from a 'recurrent' state. This gives a powerful means to analyze the behavior of Markov chains.

A Markov chain is 'irreducible' if every state is 'accessible' from any other state. More formal definition is given below.

Definition 6.2.3 (Irreducibility)**.** Let P be the transition matrix of a Markov chain $(X_n)_{n\geq 0}$ on a countable state space $\mathscr{S}$. We say the chain (and P) is *irreducible* if for any $i, j \in \mathscr{S}$, there exists an integer $k = k(i,j) \geq 0$ such that

$$\mathbb{P}(X_k = j \mid X_0 = i) = P^k(i,j) > 0.$$

Exercise 6.2.4 (RW on connected graphs is irreducible)**.** A graph $G = (V, E)$ is *connected* if for every pair of distinct nodes $x, y \in V$, there exists a sequence of nodes $z_0, z_1, \ldots, z_m$ for some $m \geq 1$ such that $z_0 = x$, $z_m = y$, and $z_i \sim z_{i+1}$ for all $i = 1, \ldots, m-1$. Show that the RW on a graph G is irreducible if and only if G is connected.

Definition 6.2.5 (Hitting and return times)**.** Let $(X_t)_{t\geq 0}$ be a Markov chain on state space $\mathscr{S}$. For $x \in \mathscr{S}$, define the *hitting time* for x to be

$$\tau_x := \inf\{t \geq 0 \mid X_t = x\},$$

the first time at which the chain is at state x. We also define the *first return time* to x as

$$\tau_x^+ := \inf\{t \geq 1 \mid X_t = x\}.$$

In general, for each $k \geq 1$, let $\tau_x^{(k)}$ denote the kth return time to x, which satisfies the recursion

$$\tau_x^{(k+1)} := \inf\{t > \tau_x^{(k)} \mid X_t = x\}, \qquad \tau_x^{(1)} = \tau_x^+.$$

For states $x, y \in \mathscr{S}$, let

$$\rho_{xy} := \mathbb{P}_x(\tau_y^+ < \infty),$$

which is the probability that the chain X_t eventually visits y starting from x.

Using the hitting times, we can classify the states into a few classes.

Definition 6.2.6 (Classification of states)**.** Let $(X_n)_{n\geq 0}$ be a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P. We classify the states into the following classes:

(*Transient*) We say a state $x \in \mathscr{S}$ is *transient* if $\rho_{xx} < 1$.

(*Recurrent*) We say a state $x \in \mathscr{S}$ is *recurrent* if $\rho_{xx} = 1$. A recurrent state is said to be

(Positive recurrent) if $\mathbb{E}_x[\tau_x^+] < \infty$; and

(Null recurrent) if $\mathbb{E}_x[\tau_x^+] = \infty$.

We say the chain and the transition matrix P is *recurrent* (resp., positive/null recurrent) if all states are recurrent (resp., positive/null recurrent).

Example 6.2.7 (SRW on $\mathbb{Z}$ is null recurrent)**.** Coinsier SRW on $\mathbb{Z}$. We remark the following asymptotic of 'persistence probability' of SRW on $\mathbb{Z}$:

$$\mathbb{P}(X_1 \geq 1,\, X_2 \geq 1, \ldots, X_n \geq 1 \mid X_0 = 0) \sim \sqrt{\frac{2}{\pi n}}.$$

(For reference, see, e.g., [LS19].) It follows that, by the tail-sum formula for the expectation of nonnegative RVs,

$$\mathbb{E}_0[\tau_0^+] = \sum_{n=1}^{\infty} \mathbb{P}_0(\tau_0^+ \geq n) \sim \sum_{n=1}^{\infty} \sqrt{\frac{2}{\pi n}} = \infty.$$

Thus the origin is null recurrent. By symmetry, all other states are also null recurrent. ▲

Exercise 6.2.8. Show the following implications:

$$\rho_{xx} > 0 \qquad \Longleftrightarrow \qquad P^m(x,x) > 0 \text{ for some } m \geq 1.$$

Hint: Use

$$\sup_{n\geq 1} \mathbb{P}\left(X_n = x \,\middle|\, X_0 = x\right) \leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} \{X_n = x\} \,\middle|\, X_0 = x\right) \leq \sum_{n=1}^{\infty} \mathbb{P}\left(X_n = x \,\middle|\, X_0 = x\right).$$

In the following lemma implies that every state for a finite-state irreudiclbe MC is positive recurrent.

Lemma 6.2.9 (Expected return time)**.** *Let $(X_t)_{t\geq 0}$ be an irreducible MC with a transition matrix P on a finite state space $\mathscr{S}$. Then for each $x, y \in \mathscr{S}$, we have (denoting $\mathbb{E}_x[\cdot] = \mathbb{E}[\cdot \mid X_0 = x]$)*

$$\mathbb{E}_x[\tau_y^+] < \infty.$$

In particular, P is positive recurrent.

PROOF. Since P is irreducible, for each $x, y \in \mathscr{S}$, there exists an integer $k(x,y) \geq 0$ such that $P^{k(x,y)}(x,y) > 0$. Let

$$r := \max_{x,y} k(x,y) \quad \text{and} \quad \varepsilon = \min_{x,y} P^{k(x,y)}(x,y).$$

Since $\mathscr{S}$ is finite, we have $r < \infty$ and $\varepsilon > 0$. Then note that for each $x, y \in \mathscr{S}$, there exists $j \in \{1, \dots, r\}$[4] such that

$$P^j(x,y) \geq \varepsilon.$$

Thus, for arbitrary $y \in \mathscr{S}$ and $t \geq 0$, the chain hits y between some time t and $t+r$ with probability at least ε regardless of X_t. It follows that, for $k \geq 1$,

$$\mathbb{P}_x(\tau_y^+ > kr) \leq (1-\varepsilon)\mathbb{P}_x(\tau_y^+ > (k-1)r).$$

By induction and strong Markov property, this shows

$$\mathbb{P}_x(\tau_y^+ > kr) \leq (1-\varepsilon)^k.$$

Thus, it is exponentially unlikely to not visit y during k consecutive time intervals of length r. Then by the tail-sum formula (Prop. 1.5.10),

$$\begin{aligned}
\mathbb{E}_x[\tau_y^+] &= \sum_{m=0}^{\infty} \mathbb{P}_x(\tau_y^+ \geq m) \\
&\leq \sum_{k=0}^{\infty} \sum_{m=kr}^{(k+1)r-1} \mathbb{P}_x(\tau_y^+ \geq m) \\
&\leq \sum_{k=0}^{\infty} \sum_{m=kr}^{(k+1)r-1} \mathbb{P}_x(\tau_y^+ \geq kr) \\
&\leq \sum_{k=0}^{\infty} r\,(1-\varepsilon)^k < \infty,
\end{aligned}$$

where we have used the fact that the tail probability $\mathbb{P}_x(\tau_y^+ \geq m)$ is non-increasing in m.

Positive recurrence of P follows immediately by noting that $\mathbb{E}_x[\tau_x^+] < \infty$ for all x. □

[4] Take $j = k(x,y)$.

The following is a typical application of strong Markov property.

Proposition 6.2.10 (Irreducible chain with a recurrent state)**.** *Suppose P is an irreducible Markov transition matrix on a countable state space $\mathscr{S}$. Fix two states $x, y \in \mathscr{S}$ and suppose x is recurrent. Then*

$$\rho_{xy} = \mathbb{P}_x(\tau_y^+ < \infty) = 1.$$

PROOF. The idea is to look at excursions from x. Namely, starting from x, the chain returns to x at least once in some a.s. finite time τ_x^+ since x is assumed to be recurrent. Restarting the chain from the stopping time τ_x^+ and again using the recurrence of x with the strong Markov property, one can deduce that the chain returns to x for the second time in some finite time. Each trajectory of the chain from x to x is called an excursion from x. By the strong Markov property, we see that (any measurable functions of) the excursions are i.i.d. and the duration of each excursion is distributed as τ_x^+.

Now during each excursion, there is a positive probability (say $\delta > 0$) to visit y. To see this, suppose not. Then since the excursions are i.i.d., the probability of visiting y during any excursion is zero. This means that the chain will never visit y, which violates the irreducibility. Then by strong Markov property,

$$\begin{aligned}\mathbb{P}_x(\tau_y^+ = \infty) &\le \mathbb{P}_x(\text{the chain does not visit } y \text{ during the first } k\text{th excursions})\\ &\le (1-\delta)^k.\end{aligned}$$

We can conclude by taking $k \to \infty$. □

Exercise 6.2.11 (Probability of mth visit)**.** Fix states $x, y \in \mathscr{S}$ and let $\tau_y^{(m)}$ denote the time of mth visit to y. Then for each $m \ge 1$, show that

$$\mathbb{P}_x(\tau_y^{(m)} < \infty) = \rho_{xy}\rho_{yy}^{m-1}.$$

(*Hint:* Use strong Markov property.)

Next, we will give a more quantitative characterization of recurrent and transient states by using the expected total number of visits. That is, for each state y, define the following counting variable

$$N_y := \sum_{n=1}^{\infty} \mathbf{1}(X_n = y) = \sum_{m=1}^{\infty} \mathbf{1}(\tau_y^{(m)} < \infty),$$

which equals the total number of visits to state y by the Markov chain.

Proposition 6.2.12. *The following hold:*

$$\mathbb{E}_x[N_y] = \begin{cases} 0 & \text{if } \rho_{xy} = 0 \\ \infty & \text{if } \rho_{xy} > 0 \text{ and } y \text{ is recurrent} \\ \frac{\rho_{xy}}{1-\rho_{yy}} & \text{if } \rho_{xy} > 0 \text{ and } y \text{ is transient.} \end{cases}$$

PROOF. By using Fubini's theorem, one can write

$$\begin{aligned}\mathbb{E}_x[N_y] = \mathbb{E}_x\left[\sum_{m=1}^{\infty} \mathbf{1}(\tau_y^{(m)} < \infty)\right] &= \sum_{m=1}^{\infty} \mathbb{E}_x\left[\mathbf{1}(\tau_y^{(m)} < \infty)\right] \\ &= \sum_{m=1}^{\infty} \mathbb{P}_x\left(\tau_y^{(m)} < \infty\right) \\ &= \sum_{m=1}^{\infty} \rho_{xy}\rho_{yy}^{m-1},\end{aligned}$$

where the last equality uses Exercise (6.2.11). Note that the last expression equals zero if $\rho_{xy} = 0$, infinity if $\rho_{xy} > 0$ and $\rho_{yy} = 1$, and a converging geometric series if $\rho_{xy} > 0$ and $\rho_{yy} < 1$. Hence we can conclude. □

Now we can give an alternative characterization of the recurrence of a state in terms of the sum of multi-step transition probabilities.

Theorem 6.2.13. *For each $x, y \in \mathscr{S}$, it holds that*

$$\mathbb{E}[N_y \,|\, X_0 = x] = \sum_{n=1}^{\infty} P^n(x,y).$$

Furthermore, a state $x \in \mathscr{S}$ is recurrent if and only if $\mathbb{E}[N_x \,|\, X_0 = x] = \sum_{m=1}^{\infty} P^m(x,x) = \infty$.

PROOF. Recall that $N_y = \sum_{n=1}^{\infty} \mathbf{1}(X_n = y)$. Hence

$$\mathbb{E}_x[N_y] = \mathbb{E}_x\left[\sum_{n=1}^{\infty} \mathbf{1}(X_n = y)\right] = \sum_{n=1}^{\infty} \mathbb{E}_x\left[\mathbf{1}(X_n = y)\right] = \sum_{n=1}^{\infty} \mathbb{P}_x\left(X_n = y\right) = \sum_{n=1}^{\infty} P^n(x,y),$$

verifying the first assertion. But Proposition 6.2.12 implies that x is recurrent if and only if $\mathbb{E}[N_x \,|\, X_0 = x] = \infty$. Hence the assertion follows. □

6.3. Stationary distribution

6.3.1. Definition and examples. In Exercise 6.1.8, we observed that we can simply multiply the transition matrix P to a given row vector $\mathbf{r}_t$ of distribution on the state space $\mathscr{S}$ in order to get the next distribution $\mathbf{r}_{t+1}$. Hence if the initial distribution of the chain is π that satisfies $\pi = \pi P$, then its distribution is invariant in time. This motivates the following definition of stationary distributions.

Definition 6.3.1. A probability distribution π (viewed as a row vector $\mathscr{S} \to [0,1]$) on a countable state space $\mathscr{S}$ is a *stationary distribution* for a Markov transition matrix P on $\mathscr{S}$ if

$$\pi = \pi P$$

that is,

$$\pi(x) = \sum_{y \in \mathscr{S}} \pi(y) P(y,x).$$

A Markov chain may have multiple stationary distributions, as the following example illustrates.

Example 6.3.2. Let $(X_t)_{t \geq 0}$ be a 2-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then any distribution $\pi = [p, 1-p]$ is a stationary distribution for the chain $(X_t)_{t \geq 0}$.

Example 6.3.3 (Stationary distribution of RW on G)**.** What is the stationary distribution of random walk on G? There is a typical one that always works. Define a probability distribution π on V by

$$\pi(i) = \frac{\deg_G(i)}{\sum_{j \in V} \deg_G(i)} = \frac{\deg_G(i)}{2|E|}.$$

Namely, the probability given to node i is proportional to the degree of node i. Then observe that π is a stationary distribution for P the transition matrix for RW on G. Indeed,

$$\sum_{i \in V} \pi(j) P(j,i) = \sum_{i \in V} \frac{\deg_G(i)}{2|E|} \frac{A(i,j)}{\deg_G(i)} = \frac{1}{2|E|} \sum_{i \in V} A(i,j) = \frac{\deg_G(j)}{2|E|} = \pi(j).$$

As we will see later, stationary distribution is crucially related to the long-term behavior of the Markov chain. Roughly speaking, for irreducible Markov chains, the stationary distribution equals the limiting frequency of the Markov chain visiting each state in the long run (see Thm. **??**). See Figure 6.3.1 for an empirical validation of this claim for RW on the Facebook network among students in Caltech in 2007. ▲

In the following exercise, we compute the stationary distribution of the so-called birth-death chain.

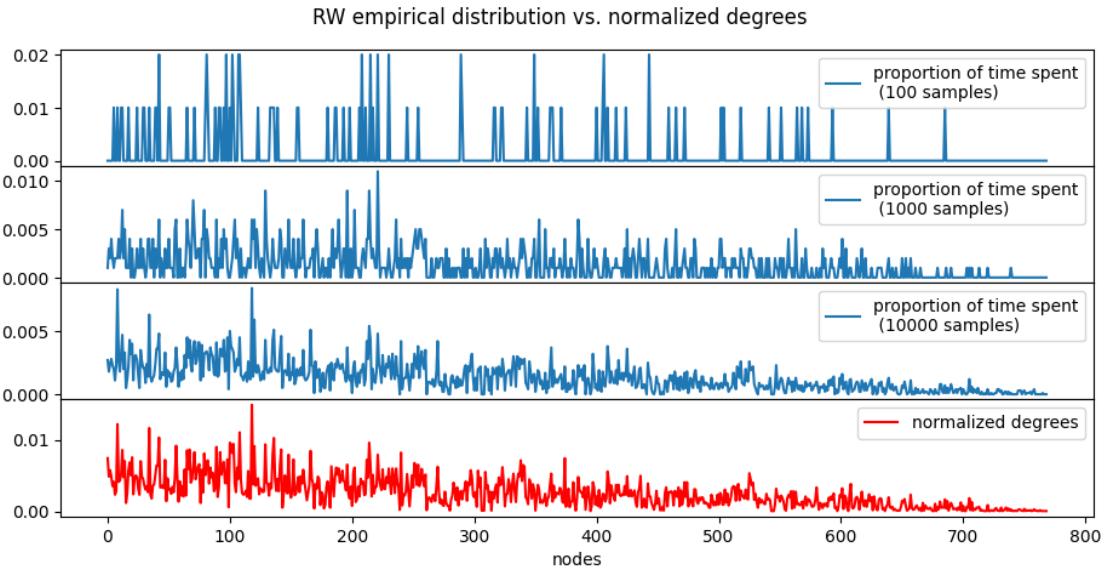


FIGURE 6.3.1. Normalized proportion of times that RW on CALTECH spends at each node for $N = 100, 1000, 10000$ steps (top three rows) and the normalized degrees $\deg(v)/2|E|$ (bottom).

Exercise 6.3.4 (Birth-Death chain)**.** Let $\mathscr{S} = \{0, 1, 2, \cdots, N\}$ be the state space. Let $(X_t)_{t\geq 0}$ be a Markov chain on $\mathscr{S}$ with transition probabilities

$$\begin{cases} \mathbb{P}(X_{t+1} = k+1 \,|\, X_t = k) = p & \forall 0 \leq k < N \\ \mathbb{P}(X_{t+1} = k-1 \,|\, X_t = k) = 1-p & \forall 1 \leq k \leq N \\ \mathbb{P}(X_{t+1} = 0 \,|\, X_t = 0) = 1-p \\ \mathbb{P}(X_{t+1} = N \,|\, X_t = N) = p. \end{cases}$$

This is called a Birth-Death chain. Its state space diagram is as below.

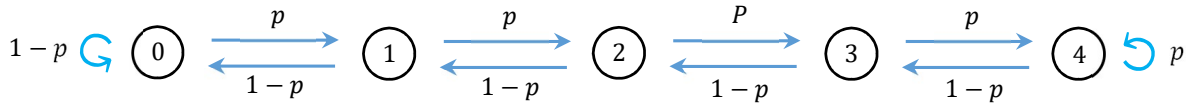


FIGURE 6.3.2. State space diagram of a 5-state Birth-Death chain

(i) Let $\pi = [\pi_0, \pi_1, \cdots, \pi_N]$ be a distribution on $\mathscr{S}$. Show that π is a stationary distribution of the Birth-Death chain if and only if it satisfy the following 'balance equation'

$$p\pi_k = (1-p)\pi_{k+1} \qquad 0 \leq k < N.$$

(ii) Let $\rho = p/(1-p)$. From (ii), deduce that $\pi_k = \rho^k \pi_0$ for all $0 \leq k < N$.

(iii) Using the normalization condition $\pi_0 + \pi_1 + \cdots + \pi_N = 1$, show that $\pi_0 = 1/(1 + \rho + \rho^2 + \cdots + \rho^N)$. Conclude that

$$\pi_k = \frac{\rho^k}{1 + \rho + \rho^2 + \cdots + \rho^N} \qquad 0 \leq k \leq N. \tag{85}$$

Conclude that the Birth-Death chain has a unique stationary distribution given by (85), which becomes the uniform distribution on $\mathscr{S}$ when $p = 1/2$.

Exercise 6.3.5 (Success run chain)**.** Consider a Markov chain $(X_n)_{n\geq 0}$ on the state space $\mathscr{S} = \mathbb{Z}_{\geq 0}$ of nonnegative integers with the following transition matrix P:

$$P(x, x+1) = p, \quad P(x,0) = 1-p \qquad \text{for all } x \in \mathbb{Z}_{\geq 0},$$

where $p \in [0,1]$ is a fixed parameter. In words, one flips independent coins with success probability p; every time it comes up heads with probability p X_n increases by 1, and when it comes up tails, X_n resets back to 0.

(i) Let $\pi : \mathbb{Z}_{\geq 0} \to [0,\infty)$ be a function that satisfies $\pi P = \pi$. Show that the following recursion holds:

$$\begin{aligned}\pi(0) &= (1-p)\sum_{k=0}^{\infty} \pi(k)\\ \pi(1) &= p\pi(0)\\ \pi(2) &= p\pi(1)\\ &\vdots\end{aligned}$$

Deduce that $\pi(k) = p^k\pi(0)$ for all $k \geq 1$.

(ii) Show that π is an stationary distribution on $\mathbb{Z}_{\geq 0}$ if and only if $\pi(k) = (1-p)p^k$ for $k \geq 0$. (Such π is called the Geometric(p) distribution that puts probability $p^k(1-p)$ on each $k \in \mathbb{Z}_{\geq 0}$. The number of success run of probability-p coin flips has this distribution.)

Next, we introduce reversibilitiy and time-reversal of a transition matrix. While the equation $\pi = \pi P$ that defines a stationary distribution requires solving a linear equation, a much simpler 'detailed balance equation', once its satisfied, implies stationarity.

Definition 6.3.6 (Reversibility and time-reversal)**.** Let P be a transition matrix on a countable state space $\mathscr{S}$ and let π be a probability distribution on $\mathscr{S}$. A matrix $\tilde{P}$ on $\mathscr{S}$ is a *time-reversal* of P w.r.t. π if it satisfies the following 'detailed balance equation':

$$\pi(x)\tilde{P}(x,y) = \pi(y)P(y,x) \quad \forall x, y \in \mathscr{S}. \tag{86}$$

We say P is *reversible* w.r.t. π if it satisfies the following detailed balance equation:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall x, y \in \mathscr{S}.$$

The following proposition will be used crucially in the proof of uniqueness of stationary distribution (Thm. 6.3.12).

Proposition 6.3.7 (Properties of time-reversal matrix)**.** *Let $\tilde{P}$ be a time-reversal of P w.r.t. π. Assume that $\pi = \pi P$ and $\pi > 0$ (entrywise). Then the following hold:*

(i) $\pi\tilde{P} = \pi$.

(ii) *$\tilde{P}$ is itself a Markov transition matrix on $\mathscr{S}$. Furthermore, π is a stationary distribution for $\tilde{P}$.*

(iii) *Let μ be another stationary distribution for P. Define a column vector $h = \mu^T/\pi^T$[5]. Then $\tilde{P}h = h$, that is, h is $\tilde{P}$-harmonic (see Def. 6.3.14).*

(iv) *If P is irreducible and recurrent, then so is $\tilde{P}$.*

PROOF. To show **(i)**, note that for each state x, since the rows of P sum to one,

$$[\pi\tilde{P}](x) = \sum_y \pi(y)\tilde{P}(y,x) = \sum_y \pi(x)P(x,y) = \pi(x)\left(\sum_y P(x,y)\right) = \pi(x).$$

[5]Entrywise division

So **(i)** holds.[6] To show **(ii)**, it is enough to show that the rows of $\tilde{P}$ sum to one. Indeed, for each state x, using that $\pi P = \pi$ and $\pi > 0$,

$$\sum_y \tilde{P}(x,y) = \sum_y \pi(x)^{-1} P(y,x)\pi(y) = \pi(x)^{-1} \sum_y P(y,x)\pi(y) = \pi(x)^{-1}[\pi P](x) = \pi(x)^{-1}\pi(x) = 1.$$

Thus $\tilde{P}$ is a Markov transition matrix on $\mathscr{S}$. By **(i)**, we also have that π is a stationary distribution for $\tilde{P}$.

To show **(iii)**, fix a state x and note that

$$\begin{aligned}
\sum_y \tilde{P}(x,y)h(y) &= \sum_y \pi(x)^{-1} P^T(x,y)\pi(y)h(y) \\
&= \sum_y \pi(x)^{-1} P^T(x,y)\mu(y) \\
&= \pi(x)^{-1} \sum_y P^T(x,y)\mu(y) \\
&= \pi(x)^{-1}\mu(x) \\
&= h(x).
\end{aligned}$$

This shows **(iii)**.

Lastly, suppose that P is irreducible and recurrent. For each states x, y, there is a path from x to y of positive probability under P. Since $\pi > 0$, traversing the same path backwards gives a path from x to y of positive probability under $\tilde{P}$. Hence $\tilde{P}$ is irreducible. For recurrence, first note that we can write

$$\tilde{P} = \mathrm{diag}(\pi)^{-1} P^T \mathrm{diag}(\pi).$$

Taking the tth power, we get

$$\tilde{P}^t = \mathrm{diag}(\pi)^{-1} (P^t)^T \mathrm{diag}(\pi).$$

Evaluating the matrices on both sides at (x,x), we get $\tilde{P}^t(x,x) = P^t(x,x)$. This holds for all states x and times $t \geq 0$. Since P is recurrent, $\sum_{t\geq 0} P^t(x,x) = \infty$ for all x (see Thm. 6.2.13). Thus $\sum_{t\geq 0} \tilde{P}^t(x,x) = \infty$ for all x, so $\tilde{P}$ is recurrent. □

Exercise 6.3.8 (Reversibility of random walk)**.** Let $P = D^{-1}A$ denote the transition matrix of a random walk on a finite graph $G = (V,E)$, where D denotes the diagonal matrix of degrees and A denotes the adjacency matrix of G. Define a probability distribution π on the node set $V = \{1,\ldots,m\}$ as

$$\pi := \frac{1}{2|E|}\big[\deg(1),\ldots,\deg(m)\big].$$

Show that P is reversible w.r.t. π. Deduce that π is a stationary distribution for P.

6.3.2. Uniqueness and existence of stationary distribution: Finite state space. In this subsection, we will show the fundamental result that for irreducible transition matrix P on a finite state space, there is a unique stationary distribution. The uniqueness uses uniqueness of harmonic functions; the existence uses a linear algebraic argument with a lazy chain.

Theorem 6.3.9 (Uniqueness and existence of stationary distribution for finite irreducible chain)**.** *Let P be an irreducible transition matrix on a finite state space $\mathscr{S} = \{1,\ldots,m\}$. Then there exists a unique stationary distribution π on $\mathscr{S}$ for P (i.e., $\pi = \pi P$) and $\pi > 0$.*

PROOF. *Existence.* Stationary distributions are closely related with eigenvectors and eigenvalues of the transition matrix P. Namely, suppose $\pi = \pi P$ and by taking transpose,

$$\pi^T = P^T \pi^T.$$

Hence, the column vector π^T is an eigenvector of P^T associated with eigenvalue 1. By Exercise 6.3.10, the dimension of the eigenspace of P^T with eigenvalue 1 equals to the dimension of eigenspace of P

[6] In fact, the argument above shows that **(i)** holds for general probability distribution π on $\mathscr{S}$.

with eigenvalue 1. But note that 1 is an eigenvalue of P with an eigenvector $\mathbf{1}$ (all-ones vector) since P is transition matrix. Thus,

$$\dim \ker(P^T - I) = \dim \ker(P - I) \geq 1. \tag{87}$$

Now we can argue for the uniqueness and the existence of stationary distribution.[7]

From (87), it follows that there exists an eigenvector v of P^T with eigenvalue 1. If the entries of v had the same sign, then $\text{sign}(v)\, v^T/\|v\|_1$ is a stationary distribution of P. In order to check this, we will use a 'lazy' version of P. Namely, define the 'lazy' transition matrix $Q = (P+I)/2$[8]. Clearly Q is irreducible since P is so (use the probabilistic interpretation of Q). Then for each states x, y, $P^{k(x,y)} > 0$ for some integer $k(x,y) \geq 0$. Since there are finitely many states, $r := \max_{x,y} k(x,y) \geq 0$ is a finite integer. By lazyness, Q^r is a positive matrix. Then $W := Q^r$ is a positive markov transition matrix.

Now let v be an eigenvector of P^T with eigenvalue 1. We claim that the entries of v must have the same sign. Suppose not. Note that v is an eigenvector of Q^T with eigenvalue 1, so it is an eigenvector of W^T with eigenvalue 1. So

$$|v_x| = \left| \sum_y W^T(x,y)\, v_y \right| < \sum_y W^T(x,y)|v_y|,$$

where the strict inequality follows since the sum in the middle has some cancellation due to mixed signs of v_y's and $W^T > 0$. Then summing over x,

$$\sum_x |v_x| < \sum_x \sum_y W^T(x,y)|v_y| = \sum_y \left(\sum_x W(y,x) \right) |v_y| = \sum_y |v_y|,$$

which is a contradiction. In turn, this shows the existence of stationary distribution for P.

Positivity. Now we know that there exists a nonnegative eigenvector v for P^T (hence for W^T) with eigenvalue 1. Then since $W^T > 0$, $v = W^T v$ gives that each entry of v is positive. This shows the existence of positive stationary distribution for P.

Uniqueness. We have shown that every eigenvector of P^T with eigenvalue 1 must have all coordinates in the same sign. In particular, it follows that the eigenspace of P^T with eigenvalue 1 must have dimension one (otherwise we can choose two orthogonal eigenvectors of the same sign, a contradiction). Since a positive eigenvalue exists, this shows the uniqueness of stationary distribution for P. □

Exercise 6.3.10 (Transpose and eigenspaces)**.** Let A be a real square matrix. Show that A and A^T have the same eigenvalues and the corresponding eigenspaces have the same dimension. (*Hint:* To show that they have the same set of eigenvalues, note $\det(A - \lambda I) = \det(A - \lambda I)^T = \det(A^T - \lambda I)$. To show the corresponding eigenspaces have the same dimension, use the fact that $\text{rank}(A - \lambda I) = \text{rank}(A^T - \lambda I)$.)

Exercise 6.3.11 (Spectral radius of a transition matrix)**.** Let P be a Markov transition matrix on a finite state space. Show that the spectral radius of P (i.e., the maximum modulus of the eigenvalues of P) is one. (*Hint:* Gershgorin circle theorem and the fact that P is a transition matrix. Namely, the eigenvalues of P reside in the union of circles in the complex plane with center P_{ii} and radius $\sum_{j \neq i} |P_{ij}| = \sum_{j \neq i} P_{ij} = 1 - P_{ii}$. Then show that 1 is an eigenvalue of P.)

6.3.3. Uniqueness of stationary distribution: Countable state space. In this section, we will use harmonicity and a marginale argument to show uniqueness of stationary distribution for an irreducible and recurrent transition matrix P on a (possibly infinite) countable state space $\mathscr{S}$. Note that irreducibility implies (positive) recurrence by Lemma 6.2.9 for finite state spaces, but it is not necessarily true for countably infinite state spaces (e.g., Asymmetyric SRW on $\mathbb{Z}$).

[7] Here both parts use the irreducibility of P. Later, we will see that irreducilibity is not needed for the exitence of stationary distribution. See Lemma 6.3.20.

[8] For each transition with Q, flip an independent fair coin, and stay put upon heads and move according to P upon tails.

Our goal in this section is to show the uniqueness of stationary distribution in the general countable state space case:

Theorem 6.3.12 (Uniqueness of positive stationary distribution)**.** *Suppose P is an irreducible and recurrent transition matrix on a countable state space $\mathcal{S}$. If there is a stationary distribution for P, then P must have a unique stationary distribution.*

A simple but important observation is that stationary distribution for an irreducible transition matrix must be positive everywhere.

Exercise 6.3.13 (Stationary distribution for irreducible chain is positive)**.** Let a transition matrix P on $\mathcal{S}$ has a stationary distribution π. Then $\pi(x) > 0$ for all $x \in \mathcal{S}$. *Hint*: Show that 0's in π propagate to outgoing neighbors: If $\pi(x) = 0$ for some x, then $\pi(y) = 0$ for all y s.t. $P(x, y) = 0$.)

Our argument for this result is based on harmonic functions and martingales.

In Exercise 6.1.8, we have seen that left multiplication by the transition matrix advances the PMF of the current state of the Markov chain in one time step. In Exercise 6.1.10, we also have seen that right-multiplying a function (represented as a column vector) by the transition matrix gives the expectation of the function w.r.t. the next-time PMF. We call distributions invariant under right multiplication by P stationary. What about functions that are invariant under left multiplication by P?

Definition 6.3.14 (Harmonic functions w.r.t. P)**.** Let P be a transition matrix on a countable state space $\mathcal{S}$. A function $h : \mathcal{S} \to \mathbb{R}$ is said to be *P-harmonic* at x if [9],

$$h(x) = \sum_{y \in \mathcal{S}} P(x, y) h(y).$$

For each subset $D \subseteq \mathcal{S}$, we say h is *P-harmonic* on D if h is harmonic on every state in D. If h is P-harmonic everywhere on $\mathcal{S}$ and if h is viewed as a column vector, then it satisfies the matrix eqution $h = Ph$.

Exercise 6.3.15 (P-harmonic on finite state space is constant)**.** Let P be an irreducible transition matrix on a finite state space $\mathcal{S}$ and let h be a P-harnomic function on $\mathcal{S}$. Show that h is a constant function. (*Hint:* Use the maximum principle. Namely, h attains a maximizer, and using harmonicity, show that all the 'neighboring' states are also global maximizer. Iteratively use irreducibility to conclude that h attains global maximum everywhere.)

Lemma 6.3.16 (Nonnegative P-harmonic functions are constant)**.** *Suppose P is an irreducible and recurrent transition matrix on a countable state space $\mathcal{S}$. Let h be a nonnegative P-harmonic function on $\mathcal{S}$. Then h is a constant.*

PROOF. Let $(X_n)_{n \geq 0}$ be a Markov chain on $\mathcal{S}$ with transition matrix P. Recall that $h(X_n)$ is a martingale (see Lem. 5.2.8). Thus if τ_y is the hitting time of a state y, then by using the stopped martingale (Thm. 5.2.17)

$$h(x) = \mathbb{E}_x[h(X_0)] = \mathbb{E}_x[h(X_{n \wedge \tau_y})] \geq h(y)\, \mathbb{P}_x(\tau_y \leq n),$$

where we have used that $h \geq 0$ for the inequality. The above inequality holds for all states x, y and time $t \geq 0$. Since P is irreducible and recurrent, by Prop. 6.2.10, we have $\mathbb{P}_x(\tau_y < \infty) = 1$. Thus by letting $n \to \infty$ and using continuity of measure (Thm. 1.1.16), we get $\mathbb{P}_x(\tau_y < n) \to 1$ giving us $h(x) \geq h(y)$. Since x, y are arbitrary states, this shows that h is a constant function. □

Now we are ready to show Theorem 6.3.12.

[9] Recall Lem. 5.2.8 on martingales obtained by plugging in a Markov chain into a harmonic function.

PROOF OF THEOREM 6.3.12. Suppose there is a stationary distribution π for P. Since P is irreducible, π is positive by Exercise 6.3.13. Let μ be any other stationary distribution for P. We wish to show that $\pi = \mu$. Since $\pi > 0$, the function (column vector) $h = \mu^T/\pi^T \geq 0$ (entrywise division) is well-defined. Let $\overleftarrow{P}$ be the time-reversal of P w.r.t. π. by Prop. 6.3.7, $\overleftarrow{P}$ is irreducible and recurrent with $\overleftarrow{P}h = h$, that is, h is $\overleftarrow{P}$-harmonic. Then h is constant by Lemma 6.3.16, so there exists a constant $c \geq 0$ such that $h(x) = \mu(x)/\pi(x) \equiv c$. Since both μ and π are probability distributions, we must have $c = 1$, so $\pi = \mu$ as desired. □

6.3.4. Characterization of stationary distribution. In this section, we will show that if an irreducible Markov chain has a stationary distribution, then the stationary distribution should be given by the reciprocal of the expected return time and all states must be positive recurrent. This is provided the following well-known lemma by Kac.

Lemma 6.3.17 (Kac)**.** *Let $(X_t)_{t\geq 0}$ be an irreducible Markov chain with transition matrix P on a countable state space $\mathcal{S}$. Suppose that there is a stationary distribution π solving $\pi = \pi P$.*

(i) *For any set $S \subseteq \mathcal{S}$, the expected return time to S when starting at the stationary distribution conditioned π on S is $\pi(S)^{-1}$. That is,*

$$\sum_{x\in S} \pi(x)\,\mathbb{E}_x[\tau_S^+] = 1. \tag{88}$$

In particular, for all states x,

$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

In particular, π is the unique stationary distribution of P.

(ii) *All states are positive recurrent.*

PROOF. Let $(Y_t)_{t\geq 0}$ be the time-reversed chain with transition matrix $\overleftarrow{P}$ defined w.r.t. π. We will first show that both P and $\overleftarrow{P}$ are recurrent. Fix a state x and define

$$\alpha(t) := \mathbb{P}_\pi(X_t = x,\, X_s \neq x \text{ for all } s > t).$$

By using stationarity and strong Markov property,

$$\alpha(t) = \mathbb{P}_\pi(X_t = x)\,\mathbb{P}_x(\tau_x^+ = \infty) = \pi(x)\,\mathbb{P}_x(\tau_x^+ = \infty).$$

Since the events $\{X_t = x,\, X_s \neq x \text{ for all } s > t\}$ are disjoint for distinct t, we have

$$1 \geq \sum_{t\geq 0} \alpha(t) = \sum_{t\geq 0} \pi(x)\,\mathbb{P}_x(\tau_x^+ = \infty).$$

Since the summand in the last sum does not depend on t and since $\pi > 0$ by Exc. 6.3.13, it follows that $\mathbb{P}_x(\tau_x^+ = \infty) = 0$ for all x. This shows that P is recurrent. Since π is a stationary distribution for $\overleftarrow{P}$ (see Prop. 6.3.7), rhe same argument shows that $\overleftarrow{P}$ is also recurrent.

By using the detailed balance equation (86) recursively, for each sequence of states $(z_0, \ldots, z_t)$,

$$\pi(z_0)P(z_0,z_1)P(z_1,z_2)\cdots P(z_{t-1},z_t) = \pi(z_t)\hat{P}(z_t,z_{t-1})\cdots\hat{P}(z_1,z_0).$$

Summing the aboe over all sequences where $z_0 = x$, $z_1, \ldots, z_{t-1} \notin S$, and $z_t = y$, we obtain

$$\pi(x)\,\mathbb{P}_x(\tau_{S^+} \geq t, X_t = y) = \pi(y)\,\overleftarrow{\mathbb{P}}_y(\tau_{S^+} = t, Y_t = x).$$

(We write $\overleftarrow{\mathbb{P}}$ for the probability measure corresponding to the reversed chain.) Then summing over all $x \in S$, $y \in \mathcal{S}$, and $t \geq 0$ shows that

$$\sum_{x\in S}\sum_{t=1}^{\infty} \pi(x)\,\mathbb{P}_x(\tau_{S^+} \geq t) = \overleftarrow{P}_\pi(\tau_{S^+} < \infty) = 1,$$

where the last equality follows from recurrence of (Y_t). Since τ_{S^+} takes only positive integer values, by the tail-sum formula, we get (88). Furthermore, deviding both sides of this by $\pi(S) = \sum_{x\in S}\pi(x)$, we get

$$\frac{1}{\pi(S)} = \sum_{x\in S}\frac{\pi(x)}{\pi(S)}\mathbb{E}_x[\tau_S^+] = \mathbb{E}_\pi[\tau_s^+ \mid X_0 \in S].$$

The second conclusion follows immediately by taking $S = \{x\}$. This shows **(i)**.

For **(ii)**, recall that $\pi > 0$ due to the irreducibility and Exc. 6.3.13. Thus $\mathbb{E}_x[\tau_x^+] < \infty$ for all x by part **(i)**. This shows positive recurrence of P. □

6.3.5. Construction of stationary distribution. Does a Markov chain always have a stationary distribution? In Theorem 6.3.9, we answered this question positively for irreucible transition matrices on a finite state space, using an indirect linear algebra argument. But such algebraic approach do not provide us any useful intuition for the behavior of the Markov chain itself. In this section, we give a probabilistic and constructive argument to show that every Markov chain has a stationary distribution that admit a certain canonical form. Importantly, we will show that irreducibility of the transition matrix is not required for there be a stationary distribution[10].

For each $y \in \mathscr{S}$, let $V_n(y)$ denote the *number of visits* to y in the first n steps:

$$V_n(y) := \sum_{k=1}^{n} \mathbf{1}(X_k = y).$$

We will see that with reasonable assumptions, a Markov chain will admit a unique stationary distribution given by

$$\pi(y) := \lim_{t\to\infty}\frac{1}{n}\mathbb{E}_x[V_n(y)], \tag{89}$$

which is the expected proportion of times spending at y starting at x and it turns out that it does not depend on x. The main result in this section, Theorem 6.3.18, states that when the Markov chain satisfies certain reasonable conditions, (89) is indeed a stationary distribution of the Markov chain. However, it is not a priori clear that the limit in (89) exists. Instead of taking the limiting frequency of visits to y, we will instead construct a measure by only counting visits to y during a single 'excursion' from x to x, which occurs during the time interval $[0, \tau_x^+]$. As we will see later, such a construction will give a valid probability distribution if and only if $\mathbb{E}_x[\tau_x^+]$ is finite.

The main result we will prove in this section is the following:

Theorem 6.3.18 (Existence, uniqueness, and characterization of stationary distribution)**.** *Assume P is irreducible. Then P has a stationary distribution if and only if all states are positive recurrent. In this case the stationary distribution π is unique, positive, and is characterized by*

$$\pi(x) = \lim_{n\to\infty}\frac{V_n(x)}{n} = \frac{1}{\mathbb{E}_x[T_x]} \qquad \forall\, x \in \mathscr{S}. \tag{90}$$

In particular, the above result shows that the empirical frequency of visiting each state (or the proportion of time spent at that state) converges almost surely to the stationary probabability at that state. We have shown

Since finite irreducible Markov chains are positive recurrent (i.e., all states are positive recurrent) (see Lem. 6.2.9), we obtain the following corollary immediately.

Corollary 6.3.19. *Assume $\mathscr{S}$ is finite and P is irreducible. Then P has a unique stationary distribution π given by* (90).

PROOF. Follow from Lemma 6.2.9 and Theorem 6.3.18. □

[10] This is clear for random walks on disconnected graphs.

The gist of the to construct of a stationary distribution using the visit counts $V_n(x,y)$ is given in the following proposition, where we construct a stationary distribution for general Markov chains with a positive recurrent state. Namely, we consider the expected number of visits to each state starting from a positive recurrent state during a single excursion. These expected counts turn out to give an invariant measure. Normalizing it to be a probability measure then gives a stationary distribution.

Lemma 6.3.20 (Excursion and stationary measure)**.** *Consider a Markov chain $(X_n)_{n\geq 0}$ on $\mathscr{S}$ with transition matrix P with at least one recurrent state, say x. For each $y \in \mathscr{S}$, define*

$$\lambda_x(y) := \mathbb{E}\left[V_{\tau_x^+}(y)\right],$$

which is the expected number of visits to y during $[0, \tau_y^+]$ (a single excursion from x to x).

(i) *$\lambda_x(x) = 1$ and $\sum_{y\in\mathscr{S}} \lambda_x(y) = \mathbb{E}_x[\tau_x^+]$.*

(ii) *$\lambda_x = \lambda_x P$. (That is, λ_x is a 'stationary measure' for P.)*

(iii) *If x is positive recurrent, then $\pi_x := \frac{1}{\mathbb{E}_x[\tau_x^+]}\lambda_x$ is a stationary distribution for P.*

PROOF. To show **(i)**, note that $V_{\tau_x^+}(x) = 1$ so $\lambda_x(x) = 1$. Also,

$$\begin{aligned}\sum_{y\in\mathscr{S}} \lambda_x(y) = \sum_{y\in\mathscr{S}} \mathbb{E}_x\left[V_{\tau_x^+}(y)\right] &= \sum_{y\in\mathscr{S}} \mathbb{E}_x\left[\sum_{n=1}^{\tau_x^+} \mathbf{1}(X_n = y)\right] \\ &= \mathbb{E}_x\left[\sum_{y\in\mathscr{S}}\sum_{n=1}^{\tau_x^+} \mathbf{1}(X_n = y)\right] = \mathbb{E}_x\Bigg[\sum_{n=1}^{\tau_x^+} \underbrace{\sum_{y\in\mathscr{S}} \mathbf{1}(X_n = y)}_{=1}\Bigg] = \mathbb{E}\left[\sum_{n=1}^{\tau_x^+} 1\right] = \mathbb{E}_x[\tau_x^+],\end{aligned}$$

where the third equality uses Fubini's theorem for nonnegative summands.

To show **(ii)**, our goal is to verify for each $y \in \mathscr{S}$,

$$\sum_{z\in\mathscr{S}} \lambda_x(z)P(z,y) = \lambda_x(y).$$

Indeed, first observe that

$$\begin{aligned}\mathbb{E}_x\left[\sum_{n=1}^{\tau_x^+} \mathbf{1}(X_n = z)\right] &= \mathbb{E}_x\left[\sum_{n=1}^{\infty} \mathbf{1}(X_n = z, \tau_x^+ \geq n)\right] \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x\left(X_n = z, \tau_x^+ \geq n\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}_x\left(X_1 \neq x, \ldots, X_{n-1} \neq x, X_n = z\right),\end{aligned}$$

where for the first equality, we have converted a random number of summation into an infinite sum with additioanl indicator. Next, by time-homogeneity and Markov property, we observe that

$$\begin{aligned}\mathbb{P}_x\left(\tau_x^+ \geq n, X_n = z\right) P(z,y) &= \mathbb{P}_x\left(\tau_x^+ \geq n, X_n = z\right)\mathbb{P}(X_{n+1} = y \,|\, X_n = z) \\ &= \mathbb{P}_x\left(\tau_x^+ \geq n, X_n = z\right)\mathbb{P}(X_{n+1} = y \,|\, \tau_x^+ \geq n, X_n = z) \\ &= \mathbb{P}_x\left(\tau_x^+ \geq n, X_n = z, X_{n+1} = y\right).\end{aligned}$$

Hence, noting that $\lambda_x(x) = 1$ from **(i)**,

$$\begin{aligned}
\sum_{z\in\mathscr{S}} \lambda_x(z)P(z,y) &= P(x,y) + \sum_{z\in\mathscr{S}\setminus\{x\}} \mathbb{E}_x\left[\sum_{n=1}^{\tau_x^+} \mathbf{1}(X_n = z)\right] P(z,y) \\
&= P(x,y) + \sum_{z\in\mathscr{S}\setminus\{x\}} \sum_{n=1}^{\infty} \mathbb{P}_x\left(\tau_x^+ \geq n,\, X_n = z,\, X_{n+1} = y\right) \\
&= P(x,y) + \sum_{n=1}^{\infty} \sum_{z\in\mathscr{S}\setminus\{x\}} \mathbb{P}_x\left(\tau_x^+ \geq n,\, X_n = z,\, X_{n+1} = y\right) \\
&= P(x,y) + \sum_{n=1}^{\infty} \mathbb{P}_x\left(\tau_x^+ \geq n,\, X_n \neq x,\, X_{n+1} = y\right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x\left(\tau_x^+ \geq n,\, X_n = y\right) \\
&= \mathbb{E}_x\left[\sum_{n=1}^{\tau_x^+} \mathbf{1}(X_n = y)\right] = \lambda_x(y).
\end{aligned}$$

This shows **(ii)**.

Lastly, if x is positive recurrent, then one can normalize λ_x by the total sum $\mathbb{E}_x[\tau_x^+] < \infty$ (see **(i)**) to obtain a probability distribution π_x that is invariant under right-multiplication of P. This shows **(iii)**. □

Now we can deduce Theorem 6.3.18.

PROOF OF THEOREM 6.3.18. If P has a stationary distribution, then all states are positive recurrent by Lemma 6.3.17. Conversely, suppose all states are positive recurrent. Then for each $x \in \mathscr{S}$, we have a stationary distribution π_x defined in Lemma 6.3.20 **(iii)**. By irreducibility, $\pi_x > 0$ (see Exc. 6.3.13). Hence by Theorem 6.3.12, there is a unique stationary distribution for P, which we will denote as π. Then for each $x \in \mathscr{S}$, it holds that

$$\pi(x) = \pi_x(x) = \frac{\lambda_x(x)}{\mathbb{E}_x[\tau_x^+]} = \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

Lastly, the return times to x form a renewal process (Def. 3.6.3) due to the strong Markov property. Thus by the renewal theorem (Thm. 3.6.4), almost surely,

$$\lim_{n\to\infty} \frac{V_n(x)}{n} = \frac{1}{\mathbb{E}_x[\tau_x^+]}.$$

This finishes the proof. □

Exercise 6.3.21 (Poisitive recurrence is a class property)**.** Let x, y be two communicating states, meaning that there exists a positive probability to reach y from x and vice versa. Show that if one is positive recurrent, then so is the other.

6.4. Convergence rate and Markov chain mixing

In this section, we are interested in obtaining bounds on the rate of convergence of the time-t PMF of a Markov chain toward the stationary distribution, for the irreducible and recurrent case. This notion is captured by the 'mixing' of the Markov chain, which is a crucial concept in Markov chain theory and also has many applications in statatistics (e.g., for obtaining bounds on the estimation error using Markov chain Monte Carlo estimators) and in machine learning (e.g., analyzing policy gradient descent algorithms in reinforcement learning).

6.4.1. Total variance distance and mixing time. We first need to introduce a metric that measures the 'distance' between two probability distributions on the same measurable space. The one we use a lot in Markov chain theory is called the total variation distance, which is the worst-case difference between the probabilities assigned to a single event by the two distributions.

Definition 6.4.1 (Total variation distance)**.** Let μ and λ be two probability distribution on a measurable space $(\mathscr{S},\mathscr{F})$. Then the *total variation distance* between μ and λ, $\|\mu-\lambda\|_{TV}$, is defined by

$$\|\mu-\lambda\|_{TV} := \sup_{A\in\mathscr{F}} |\mu(A)-\lambda(A)|. \tag{91}$$

The definition in (91) involves supremum over all events so it is not easy to work with. Below we provide three usful characterizations of it.

Proposition 6.4.2 (TV distance; pointwise form)**.** *Let μ and ν be two probability distributions on a countable state space $\mathscr{S}$. Then*

$$\|\mu-\lambda\|_{TV} = \sum_{x:\mu(x)\geq\nu(x)} |\mu(x)-\nu(x)| = \frac{1}{2}\sum_{x\in\mathscr{S}} |\mu(x)-\nu(x)|. \tag{92}$$

PROOF. Let $B=\{x:\mu(x)\geq\nu(x)\}$ and let $A\subseteq\mathscr{S}$ be any event. Then

$$\mu(A)-\nu(A)\leq\mu(A\cap B)-\nu(A\cap B)\leq\mu(B)-\nu(B). \tag{93}$$

The first inequality is true because any $x\in A\cap B^c$ satisfies $\mu(x)-\nu(x)<0$, so the difference in probability cannot decrease when such elements are eliminated. For the second inequality, note that including more elements of B cannot decrease the difference in probability. Since (93) holds for all events A, the first equality in (92) follows.

By exactly parallel reasoning,

$$\nu(A)-\mu(A)\leq\nu(B^c)-\mu(B^c). \tag{94}$$

Fortunately, the upper bounds on the right-hand sides of (93) and (94) are actually the same (as can be seen by subtracting them; see Figure 4.1). Furthermore, when we take $A=B$ (or B^c), then $|\mu(A)-\nu(A)|$ is equal to the upper bound. Thus

$$\|\mu-\nu\|_{\mathrm{TV}} = \frac{1}{2}[\mu(B)-\nu(B)+\nu(B^c)-\mu(B^c)] = \frac{1}{2}\sum_{x\in\mathscr{S}} |\mu(x)-\nu(x)|.$$

□

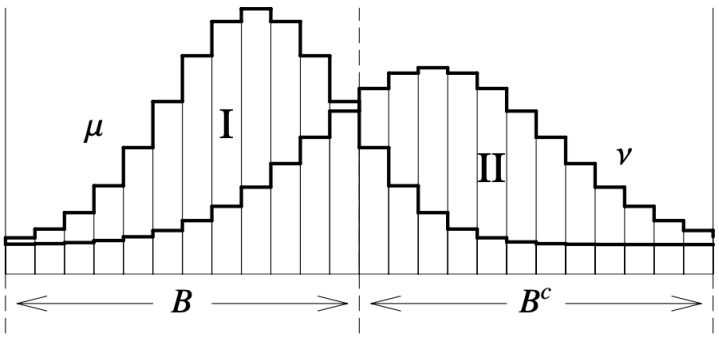


FIGURE 6.4.1. Recall that $B=\{x:\mu(x)\geq\nu(x)\}$. Region I has area $\mu(B)-\nu(B)$. Region II has area $\nu(B^c)-\mu(B^c)$. Since the total area under each of μ and ν is 1, regions I and II must have the same area and that area is $\|\mu-\nu\|_{\mathrm{TV}}$. Figure exerpted from [LP17].

Remark 6.4.3. From Proposition 6.4.2 and the triangle inequality for real numbers, it is easy to see that total variation distance satisfies the triangle inequality: for probability distributions μ, ν, and η,

$$\|\mu-\nu\|_{\mathrm{TV}}\leq\|\mu-\eta\|_{\mathrm{TV}}+\|\eta-\nu\|_{\mathrm{TV}}.$$

Proposition 6.4.4 (TV distance; variational form)**.** *Let μ and ν be two probability distributions on a countable state space $\mathcal{S}$. Then*

$$\|\mu-\lambda\|_{TV}=\frac{1}{2}\sup\left\{\sum_{x\in\mathcal{S}}f(x)\mu(x)-\sum_{x\in\mathcal{S}}f(x)\nu(x):\|f\|_\infty\le 1\right\}. \tag{95}$$

PROOF. On the one hand, since f has supremum norm ≤ 1, using triangle inequality we get

$$\sum_{x\in\mathcal{S}}f(x)\mu(x)-\sum_{x\in\mathcal{S}}f(x)\nu(x)\le\sum_{x\in\mathcal{S}}|f(x)||\mu(x)-\nu(x)|\le\sum_{x\in\mathcal{S}}|\mu(x)-\nu(x)|.$$

Thus "$\ge$" in (95) follows from Prop. 6.4.2.

On the other hand, consider the function f on $\mathcal{S}$ defined as

$$f^*(x)=\mathbf{1}(\mu(x)\ge\nu(x))-\mathbf{1}(\mu(x)<\nu(x)).$$

Then

$$\begin{aligned}\sum_{x\in\mathcal{S}}f^*(x)\mu(x)-\sum_{x\in\mathcal{S}}f^*(x)\nu(x)&=\sum_{x:\mu(x)\ge\nu(x)}\mu(x)-\nu(x)-\sum_{x:\mu(x)<\nu(x)}\mu(x)-\nu(x)\\&=\sum_x|\mu(x)-\nu(x)|\\&=2\|\mu-\nu\|_{TV},\end{aligned}$$

where the last equality follows from Prop. 6.4.2. This shows "$\le$" in (95). □

It is useful to introduce a parameter that measures the time required by a Markov chain for the distance to stationarity to be small.

Definition 6.4.5 (Mixing time)**.** Let P be a Markov transition matrix on a countable state space $\mathcal{S}$ with a stationary distribution π. The *mixing time* of P is defined by

$$t_{\text{mix}}(\varepsilon):=\min\{t:d(t)\le\varepsilon\},\quad t_{\text{mix}}:=t_{\text{mix}}(1/4),$$

where $d(t):=\sup_{x\in\mathcal{S}}\|P^t(x,\cdot)-\pi\|_{TV}$.

Exercise 6.4.6. Let P be a Markov transition matrix on a countable state space $\mathcal{S}$ with a stationary distribution π.

(i) (*Standardizing Distance from Stationarity*) Define

$$d(t):=\sup_x\|P^t(x,\cdot)-\pi\|_{TV}\quad\text{and}\quad\bar{d}(t):=\sup_{x,y}\|P^t(x,\cdot)-P^t(y,\cdot)\|_{TV}.$$

Show that $d(t)\le\bar{d}(t)\le 2d(t)$. (*Hint:* For the first inequality, use that since $\pi=\pi P$, we have $\pi(A)=\sum_y\pi(y)P(y,A)$. Then $|P^t(x,A)-\pi(A)|\le\sum_y\pi(y)|P^t(x,A)-P^t(y,A)|\le\bar{d}(t)$.)

(ii) (submultiplicativity) Show that $\bar{d}(s+t)\le\bar{d}(s)\bar{d}(t)$. (*Hint*: Use an optimal coupling, see Prop. 6.4.23.)

(iii) (multiples of mixing time) Show that $d(\ell t_{mix})\le 2^{-\ell}$.

6.4.2. Examples and aperiodicity. Let $(X_n)_{n\ge0}$ be an irreducible positive recurrent Markov chain on a countable state space Ω. By Theorem 6.3.18, we know that the chain has unique stationary distribution π. Denote by π_t the distribution of X_t. When can we expect that $\pi_t\to\pi$ in TV distance?

We first look at some examples. We start with an example, where we show the convergence of a 2-state chain using a diagonalization of its transition matrix.

Example 6.4.7. Let $(X_n)_{n\ge0}$ be a Markov chain on $\mathcal{S}=\{1,2\}$ with the following transition matrix

$$P=\begin{bmatrix}0.2&0.8\\0.6&0.4\end{bmatrix},$$

which has a unique stationary distribution $\pi=[3/7,4/7]$. By diagonalizing P, we can write

$$P^t=\begin{bmatrix}1&-4/3\\1&1\end{bmatrix}\begin{bmatrix}1&0\\0&(-2/5)^t\end{bmatrix}\begin{bmatrix}1&-4/3\\1&1\end{bmatrix}^{-1}.$$

Hence, letting π_t denote the row vector of distribution of X_t, we deduce

$$\pi_t = \pi_0 \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}.$$

Writing

$$\pi = \pi_0 \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1},$$

we get

$$\pi_t - \pi = \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}.$$

It follows that $\|\pi_t - \pi\|_{TV} \le C(2/5)^t$ for an explicit constant $C > 0$. ▲

Next, we observe that an irreducible MC may not always converge to the stationary distribution. The key issue there is the notion of 'periodicity'.

Example 6.4.8. Let $(X_n)_{n\ge 0}$ be a 2-state MC on state space $\Omega = \{0,1\}$ with transition matrix

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Namely, the chain deterministically alternates between the two states. Note that it is irreducible and has a unique stationary distribution

$$\pi = [1/2, 1/2].$$

Let π_t be the distribution of X_n, where the initial distribution is given by $\pi_0 = [1,0]$. Then we have

$$\pi_1 = [0,1]$$
$$\pi_2 = [1,0]$$
$$\pi_3 = [0,1]$$
$$\pi_4 = [1,0],$$

and so on. Hence the distributions π_n do not converge to the stationary distribution π. ▲

What goes wrong with RWs on bipartite graphs?

Example 6.4.9 (RW on torus)**.** Let $\mathbb{Z}_n$ be the set of integers modulo n. Let $G = (V,E)$ be a graph where $V = \mathbb{Z}_n \times \mathbb{Z}_n$ and two nodes $u = (u_1, u_2)$ and $v = (v_1, v_2)$ are adjacent if and only if

$$|u_1 - u_2| + |v_1 - v_2| = 1.$$

Such a graph G is called the $n \times n$ *torus* and we write $G = \mathbb{Z}_n \times \mathbb{Z}_n$. Intuitively, it is obtained from the $n \times n$ square grid by adding boundary edges to wrap around (see Figure 6.4.2 left).

Now let $(X_t)_{t\ge 0}$ be a random walk on G. Since G is connected, X_t is irreducible. Since all nodes in G have degree 4, the uniform distribution on $\mathbb{Z}_n \times \mathbb{Z}_n$, which we denote by π, is the unique stationary distribution of X_t. Let π_t denote the distribution of X_t.

For instance, consider $G = \mathbb{Z}_6 \times \mathbb{Z}_6$. As illustrated in Figure 6.4.2 below, observe that if X_0 is one of the red nodes (where sum of coordinates is even), then X_t is at a red node for any t = even and at a black node (where sum of coordinates is odd) at t = odd. Hence, π_t is supported only on the 'even' nodes for even times and on the 'odd' nodes for the odd times. Hence π_t does not converge in any sense to the uniform distribution π.

The following example of random walks on cycles in Figure 6.4.3 illustrates the same point on the 'one-dimensional torus'. Note that even after 200 steps, RW on C_{14} has zero probability of being at every other nodes, while for RW on C_{15}, the probabilities evens out.

▲

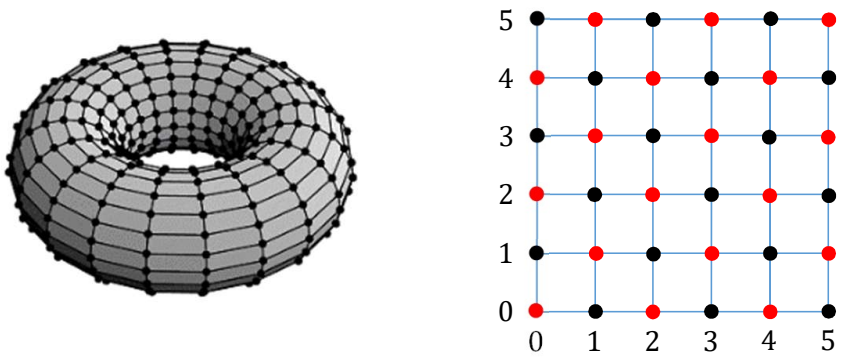


FIGURE 6.4.2. (Left) Torus graph (Figure excerpted from [LP17]). (Right) RW on torus $G = \mathbb{Z}_6 \times \mathbb{Z}_6$ has period 2.

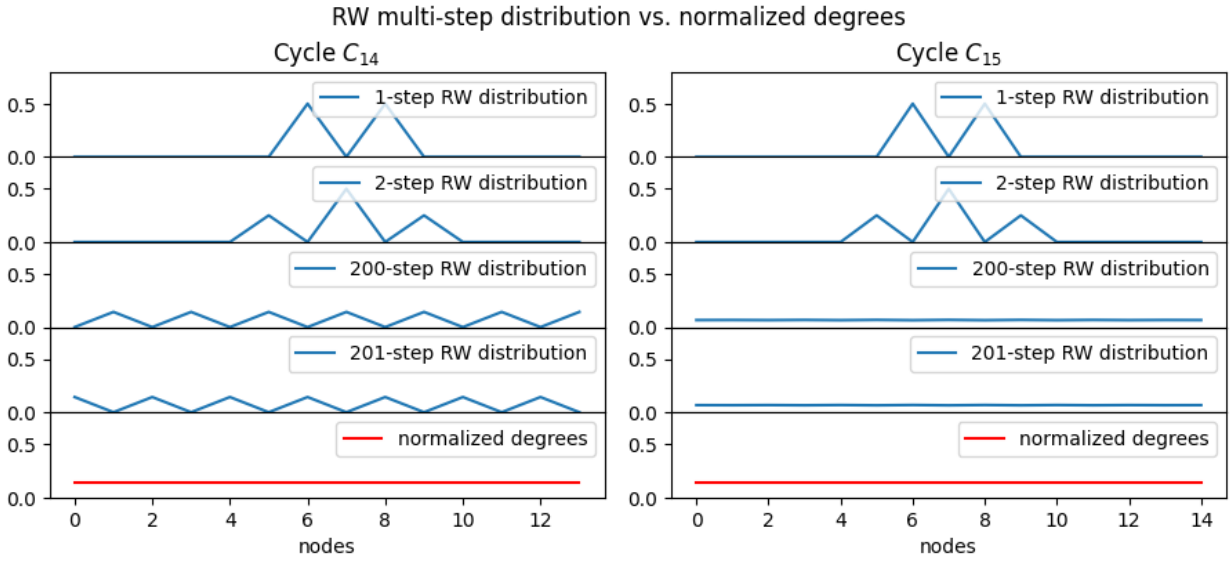


FIGURE 6.4.3. Comparison of multi-step distribution of RW on C_n

The key issue in the 2-periodicity of RW on $G = \mathbb{Z}_6 \times \mathbb{Z}_6$ is that it takes even number of steps to return to any given node. Generalizing this observation, we introduce the following notion of periodicity.

Definition 6.4.10. Let P be the transition matrix of a Markov chain $(X_t)_{t\geq 0}$ on a countable state space $\mathcal{S}$. For each state $x \in \mathcal{S}$, let $\mathcal{T}(x) := \{t \geq 1 \,|\, P^t(x,x) > 0\}$ be the set of times when it is possible for the chain to return to starting state x. We define the *period* of x by the greatest common divisor of $\mathcal{T}(x)$. We say the chain X_t is *aperiodic* if all states have period 1.

Example 6.4.11. Let $\mathcal{T}(x) = \{4,6,8,10,\cdots\}$. Then the period of x is 2, even through it is not possible to go from x to x in 2 steps. If $\mathcal{T}(x) = \{4,6,8,10,\cdots\} \cup \{3,6,9,12,\cdots\}$, then the period of x is 1. This means the return times to x is irregular. For the RW on $G = \mathbb{Z}_6 \times \mathbb{Z}_6$ in Example 6.4.9, all nodes have period 2. ▲

Exercise 6.4.12 (Aperiodicity of RW on graphs)**.** Let $(X_t)_{t\geq 0}$ be a random walk on a connected graph $G = (V, E)$.

(i) Show that all nodes have the same period.
(ii) If G contains an odd cycle C (e.g., triangle), show that all nodes in C have period 1.
(iii) Show that X_t is aperiodic if and only if G contains an odd cycle.

(iv) Show that X_t is periodic if and only if G is *bipartite*. (A graph G is bipartite if there exists a partition $V = A \cup B$ of nodes such that all edges are between A and B.)

Remark 6.4.13. If $(X_t)_{t\geq 0}$ is an irreducible Markov chain on a finite state space Ω, then all states $x \in \Omega$ must have the same period. The argument is similar to that for Exercise 6.4.12 (i).

Exercise 6.4.14 (Aperiodicity implies irreducibility for the product chain)**.** Let P be an irreducible transition matrix for a Markov chain on $\mathcal{S}$. Let Q be a matrix on $\mathcal{S} \times \mathcal{S}$ defined by

$$Q((x,y),(z,w)) = P(x,z)\,P(y,w) \qquad \text{for } x,y,z,w \in \mathcal{S}.$$

(i) Consider a Markov chain $Z_t := (X_t, Y_t)$ on $\mathcal{S} \times \mathcal{S}$ that evolves by independently evolving its two coordinates according to P. Show that Q above is the transition matrix for Z_t.
(ii) Show that Q is irreducible if P is aperiodic.

The following exercise shows that if a RW on G is irreducible and aperiodic, then it is possible to reach any node from any other node in a fixed number of steps.

Exercise 6.4.15. Let $(X_t)_{t\geq 0}$ be a RW on a connected graph $G = (V,E)$. Let P denote its transition matrix. Suppose G contains an odd cycle, so that the walk is irreducible and aperiodic. For each $x \in V$, let $\mathcal{T}(x)$ denote the set of all possible return times to the state x.

(i) For any $x \in V$, show that $a, b \in \mathcal{T}(x)$ implies $a + b \in \mathcal{T}(x)$.
(ii) For any $x \in V$, show that $\mathcal{T}(x)$ contains 2 and some odd integer b.
(iii) For each $x \in V$, let b_x be the smallest odd integer contained in $\mathcal{T}(x)$. Show that $m \in \mathcal{T}(x)$ whenever $m \geq b_x$.
(iv) Let $b^* = \max_{x\in V} b_x$. Show that $m \in \mathcal{T}(x)$ for all $x \in V$ whenever $m \geq b^*$.
(v) Let $r = |V| + b^*$. Show that $P^r(x,y) > 0$ for all $x, y \in V$.

With a little more work, one can also show a similar statement for general irreducible and aperiodic Markov chains.

Exercise 6.4.16. Let P be the transition matrix of a Markov chain $(X_t)_{t\geq 0}$ on a finite state space $\mathcal{S}$. Show that (i) implies (ii):

(i) P is irreducible and aperiodic.
(ii) There exists an integer $r \geq 0$ such that every entry of P^r is positive.

Furthermore, show that (ii) implies (i) if X_t is a RW on some graph $G = (V,E)$. (Hint for (i)⇒(ii): You may use the fact that if a subset $\mathcal{T}$ of positive integers $\mathbb{N}$ is closed under addition with $\gcd(\mathcal{T}) = 1$, then it must contain all but finitely many integers. Proof is similar to Exercise 6.4.15, but a bit more number-theoretic. See [LP17, Lem. 1.27].)

Exercise 6.4.17. Let (X_t) and (Y_t) be independent random walks on a connected graph $G = (V,E)$. Suppose that G contains an odd cycle. Let P be the transition matrix of the random walk on G.

(i) Let $t \geq 0$ be arbitrary. Use Exercise 6.4.15 to deduce that for some integer $r \geq 1$, we have

$$\begin{aligned}\mathbb{P}(X_{t+r} = Y_{t+r} \mid X_t = x, Y_t = y) &= \sum_{z\in V} \mathbb{P}(X_{t+r} = z \mid X_t = x)\mathbb{P}(Y_{t+r} = z \mid Y_t = y)\\ &= \sum_{z\in V} P^r(x,z)P^r(y,z) > 0.\end{aligned}$$

(ii) Let $r \geq 1$ be as in (i) and let

$$\delta = \min_{x,y\in V} \sum_{z\in V} P^r(x,z)P^r(y,z) > 0.$$

Use (i) and Markov property that in every r steps, X_t and Y_t meet with probability $\geq \delta > 0$.
(iii) Let τ be the first time that X_t and Y_t meet. From (ii) and Markov property, deduce that

$$\mathbb{P}(\tau \geq kr) \leq (1-\delta)^k \to 0 \quad \text{as } k \to \infty.$$

6.4.3. Convergence theorem: Finite state space. In this section, we prove the convergence theorem for irreducible aperiodic chains on finite state spaces.

Theorem 6.4.18 (Exponential mixing for finite irreducible aperiodic chains)**.** *Let $(X_t)_{t\geq 0}$ be an irreducible aperiodic Markov chain on a finite state space $\mathscr{S}$. Let π denote the unique stationary distribution of X_t. Then there exists constants $C>0$ and $\alpha\in[0,1)$ such that for all $t\geq 0$,*

$$\max_{x\in\mathscr{S}} \|P^t(x,\cdot)-\pi\|_{TV} \leq C\alpha^t. \tag{96}$$

PROOF. Since P is irreducible and aperiodic, by Exercise 6.4.16, there exists an $r\geq 0$ such that P^r has strictly positive entries. Let Π be the $\mathscr{S}\times\mathscr{S}$ matrix obtained by stacking the row vector π. For sufficiently small $\delta>0$, we have

$$P^r(x,y)\geq \delta\pi(y) \quad \text{for all } x,y\in\mathscr{S}. \tag{97}$$

Let $\theta := 1-\delta\in[0,1)$. If $\delta\geq 1$, then $P^r(x,\cdot)=\pi$ for all x, so (96) holds with $\theta=0$ (i.e., the chain mixes completely in r steps). Hence we may assume $\theta\in(0,1)$.

Let Q be a matrix satisfting the following equation

$$P^r = (1-\theta)\Pi+\theta Q. \tag{98}$$

Multyplying by the colum vector $\mathbf{1}$ and using $P^r\mathbf{1}=\mathbf{1}=\Pi\mathbf{1}$ with $\theta>0$, it follows that $Q\mathbf{1}=\mathbf{1}$. Moreover, (97) implies that $Q\geq 0$, so Q is itself a Markov transition matrix on $\mathscr{S}$[11].

Next, we use induction to show that, for any $k\geq 1$,

$$P^{rk} = (1-\theta^k)\Pi+\theta^k Q^k. \tag{99}$$

The above holds for $k=1$ due to (98). For the induction step, note that $M\Pi=\Pi$ for any Makov transition matrix M on $\mathscr{S}$ and that $\Pi M=\Pi$ for any matrix M such that $\pi M=\pi$. In particular, $\Pi P^n=\Pi$ and $Q^n\Pi=\Pi$. Hence, assuming that (98) holds for $k=n$, using (98),

$$\begin{aligned} P^{r(n+1)} = P^{rn}P^r &= \left[(1-\theta^n)\Pi+\theta^nQ^n\right]P^r \\ &= (1-\theta^n)\Pi+\theta^nQ^n((1-\theta)\Pi+\theta Q) \\ &= (1-\theta^{n+1})\Pi+\theta^nQ^n. \end{aligned}$$

This verifies (99).

Multiplying by P^j on both sides of (99) and rearranging terms now yields

$$P^{rk+j}-\Pi = \theta^k\left(Q^{rk}P^j-\Pi\right). \tag{100}$$

To finish, sum the absolute values of the elements in row x on both sides of (100) and divide by 2. By Proposition 6.4.2, this gives

$$\|P^{rk+j}(x,\cdot)-\pi\|_{TV}\leq\theta^k\|Q^{rk}(x,\cdot)P^j-\pi\|_{TV}\leq\theta^k.$$

In the middle term, think of initializing a Markov chain with distribution $Q^{rk}(x,\cdot)$ and evolve according to P j steps and then compare the resulting distribution with π. By definition, the TV distance between any two probability distributions is at most one, which gives the last inequality. Then taking $\alpha=\theta^{1/r}$ and $C=1/\theta$ finishes the proof. □

[11] (98) can be interpreted probabilistically as follows: In each step, flip an independent probability θ coin. If heads, evolve the state according to Q; if tails, sample the next state independently from π.

6.4.4. Coupling and Total Variation Distance. One of the favorite arguments of probabilists involves the use of 'coupling'. The coupling technique in probability is a powerful tool used to compare and relate different probability distributions. It involves constructing a joint probability distribution for two random variables on a single probability space such that each variable has the desired marginal distribution. By aligning the marginals in this way, coupling allows for direct comparison between distributions, enabling to establish relationships, bounds, and convergence properties. Coupling is particularly valuable in stochastic processes, where it facilitates the analysis of complex systems and the study of convergence behavior.

Definition 6.4.19 (Coupling)**.** A coupling of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν. That is, a coupling (X, Y) satisfies $\mathbb{P}(X = x) = \mu(x)$ and $\mathbb{P}(Y = y) = \nu(y)$.

Example 6.4.20. Let μ and ν both be the "fair coin" measure giving weight $1/2$ to the elements of $\{0, 1\}$.

(i) One way to couple μ and ν is to define (X, Y) to be a pair of independent coins, so that $\mathbb{P}\{X = x, Y = y\} = 1/4$ for all $x, y \in \{0, 1\}$.

(ii) Another way to couple μ and ν is to let X be a fair coin toss and define $Y = X$. In this case, $\mathbb{P}\{X = Y = 0\} = 1/2$, $\mathbb{P}\{X = Y = 1\} = 1/2$, and $\mathbb{P}\{X \neq Y\} = 0$.

Example 6.4.21. Let $\mu = \text{Bernoulli}(1/3)$ and $\nu = \text{Bernoulli}(1/4)$. Let $U \sim \text{Uniform}(0, 1)$ and define $X = \mathbf{1}(U \leq 1/3)$ and $Y = \mathbf{1}(U \leq 1/2)$. Then (X, Y) is a coupling of μ and ν and satisfies $X \leq Y$ almost surely.

Exercise 6.4.22 (Coupling between ER random graphs)**.** Let $\mu = G(n, p)$ and $\nu = G(n, q)$ for $0 \leq p \leq q \leq 1$ (recall Def. 5.4.4). Construct a coupling (X, Y) between μ and ν such that $X \sim G(n, p)$, $Y \sim G(n, q)$, and X is a subgraph of Y almost surely.

Suppose μ, ν are probability distributions on a countable state space $\mathcal{S}$. Given a coupling (X, Y) of μ and ν, if q is the joint distribution of (X, Y) on $\mathcal{S} \times \mathcal{S}$, meaning that $q(x, y) = \mathbb{P}(X = x, Y = y)$, then the row and column margins of q equals μ and ν, respectively. That is,

$$\sum_{y \in \mathcal{S}} q(x, y) = \sum_{y \in X} \mathbb{P}\{X = x, Y = y\} = \mathbb{P}\{X = x\} = \mu(x) \qquad \text{for } x \in \mathcal{S},$$
$$\sum_{x \in \mathcal{S}} q(x, y) = \sum_{x \in \mathcal{S}} \mathbb{P}\{X = x, Y = y\} = \mathbb{P}\{Y = y\} = \nu(y) \qquad \text{for } y \in \mathcal{S}.$$

Conversely, given a probability distribution q on the product space $\mathcal{S} \times \mathcal{S}$ with margin (μ, ν), then there is a pair of random variables (X, Y) having q as their joint distribution – and consequently this pair (X, Y) is a coupling of μ and ν (why?). In summary, a coupling can be specified either by a pair of random variables (X, Y) defined on a common probability space or by a distribution q on $\mathcal{S} \times \mathcal{S}$.

Every pair of distributions μ and ν possesses an independent coupling. Nonetheless, in cases where μ and ν differ, it becomes unattainable for X and Y to consistently assume identical values. What degree of closeness can a coupling achieve to ensure X and Y are nearly identical? The total variation distance provides the answer.

Proposition 6.4.23 (Coupling and TV distance)**.** *Let μ and ν be two probability distributions on $\mathcal{S}$. Then*

$$\|\mu - \nu\|_{TV} = \inf\{\mathbb{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$$

Furthermore, there exists a coupling that achieves the infemum above, which is called an 'optimal coupling'.

PROOF. First, we note that for any coupling (X, Y) of μ and ν and any event $A \subseteq \mathcal{S}$,

$$\mu(A) - \nu(A) = \mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\} \leq \mathbb{P}\{X \in A, Y \notin A\} \leq \mathbb{P}\{X \neq Y\}.$$

It immediately follows that

$$\|\mu - \nu\|_{\text{TV}} \leq \inf\{\mathbb{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

For the converse direction, we will construct a coupling that attains the infemum. Let $p = \sum_{x\in\mathscr{S}} \mu(x)\wedge \nu(x) \in [0,1]$. Flip a coin with probability of heads equal to p. If it lands heads, draw Z from the distribution $p^{-1}(\mu\wedge\nu)$ and let $X = Y = Z$ (this case never occurs if $p = 0$). Otherwise, draw independently X and Y from the distributions $(1-p)^{-1}(\mu-\nu)\mathbf{1}(\mu > \nu)$ and $(1-p)^{-1}(\nu-\mu)\mathbf{1}(\nu > \mu)$, respectively. It is easy to verify that X and Y have distributions μ and ν, respectively, and that $X = Y$ if and only if the coin lands heads. Furthermore, note that

$$\sum_{x\in\mathscr{S}} \mu(x)\wedge\nu(x) = \sum_{x\in\mathscr{S};\mu(x)\le\nu(x)} \mu(x) + \sum_{x\in\mathscr{S};\mu(x)>\nu(x)} \nu(x).$$

Adding and subtracting $\sum_{x\in\mathscr{S};\mu(x)>\nu(x)} \mu(x)$ from both sides, we deduce

$$\mathbb{P}(X\neq Y) = 1-p = 1-\sum_{x\in\mathscr{S}} \mu(x)\wedge\nu(x) = \sum_{x\in\mathscr{S};\mu(x)>\nu(x)} \mu(x)-\nu(x) = \|\mu-\nu\|_{\mathrm{TV}}.$$

This is enouth to conclude. □

We define a coupling of Markov chains with transition matrix P to be a process $(X_t, Y_t)_{t\ge0}$ with the property that both (X_t) and (Y_t) are Markov chains with transition matrix P, although the two chains may possibly have different starting distributions. More precisely,

Definition 6.4.24 (Markovian coupling)**.** Given a Markov chain on a countable state space $\mathscr{S}$ with transition matrix P, a *Markovian coupling* of two P-chains on $\mathscr{S}$ is a Markov chain $\{(X_t, Y_t)\}_{t\ge0}$ with state space $\mathscr{S}\times\mathscr{S}$ which satisfies the following:

(i) *(coupling)* For each $x, y, x', y' \in \mathscr{S}$,

$$\mathbb{P}\{X_{t+1} = x' | X_t = x, Y_t = y\} = P(x,x') \quad \text{and} \quad \mathbb{P}\{Y_{t+1} = y' | X_t = x, Y_t = y\} = P(y,y').$$

(ii) *(coalescence)* Let $\tau_{\text{couple}} := \inf\{t \ge 0 \,|\, X_t = Y_t\}$ denote the first time that the two chains meet. Then $X_s = Y_s$ a.s. for all $s \ge t$.

We will use $\mathbb{P}_{x,y}$ to denote the probability measure for the coupled Markov chain $\{(X_t, Y_t)\}_{t\ge0}$ with initial state (x, y).

Any coupling of Markov chains with transition matrix P can be modified so that the two chains stay together at all times after their first simultaneous visit to a single state—more precisely, so that if $X_s = Y_s$, then $X_t = Y_t$ for $t \ge s$. To construct a coupling satisfying (5.2), simply run the chains according to the original coupling until they meet, then run them together.

Example 6.4.25 (Coupled random walks on $\mathbb{Z}_{\ge0}$)**.** Let P be the Markov transition matrix for the simple symmetric random walk on $\mathbb{Z}_{\ge0}$, where we place a self-loop at the origin so that the negative jump attempted at the origin results in not moving. Let X_t, Y_t be two such random walks, and suppose $X_0 = x \le y = Y_0$. It is intuitive that

$$\mathbb{P}_x(X_t = 0) \ge \mathbb{P}_y(Y_t = 0)$$

for all t, since they play the same games and Y_t starts with larger or equal amount of fortune as X_t does.

Here is a simple proof of the above fact by using (Markovian) coupling. Let $\xi_1, \xi_2, \dots$ denote i.i.d. Bernoulli(1/2) variables. We use these common RVs to evolve the two RWs X_t and Y_t simultaneously: If $\xi_t = 1$, then both X_t and Y_t increase by one; If $\xi_t = 0$, then both decrease by one, except when any one of them is at zero, they stay put. Once the two chains meet (necessarily either at 0), they stay together thereafter.

Clearly, the distribution of X_t is $P^t(x,\cdot)$, and the distribution of Y_t is $P^t(y,\cdot)$. Importantly, X_t and Y_t are defined on the same underlying probability space, as both chains use the common randomness $(\xi_t)_{t\ge1}$ to determine their moves.

It is clear that if $x \le y$, then $X_t \le Y_t$ almost surely for all t. In particular, if $X_t = 0$, the top state, then it must be that $Y_t = 0$ also. From this, we can conclude that

$$P^t(x,n) = \mathbb{P}\{X_t = n\} \le \mathbb{P}\{Y_t = n\} = P^t(y,n).$$

This argument showcases the power of coupling. We were able to couple the two chains together in such a way that $X_t \le Y_t$ always, and from this fact about the random variables, we could easily derive information about the distributions. ▲

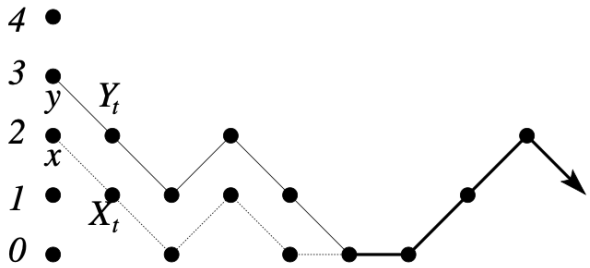


FIGURE 6.4.4. Coupling between two random walks on $\mathbb{Z}_{\ge 0}$. Time goes from left to right. Once they meet, they stay synchronized. Figure exerpted from [LP17].

Lemma 6.4.26 (Bounding TV distance by coalescence time)**.** *Let $\{(X_t, Y_t)\}$ be a Markovian coupling for transition matrix P with for which $X_0 = x$ and $Y_0 = y$. Then*

$$\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \le \mathbb{P}_{x,y}\{\tau_{\text{couple}} > t\}.$$

PROOF. Notice that $P^t(x,z) = \mathbb{P}_{x,y}\{X_t = z\}$ and $P^t(y,z) = \mathbb{P}_{x,y}\{Y_t = z\}$. Consequently, (X_t, Y_t) is a coupling of $P^t(x,\cdot)$ with $P^t(y,\cdot)$. So by Prop. 6.4.23,

$$\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \le \mathbb{P}_{x,y}\{X_t \ne Y_t\} = \mathbb{P}_{x,y}\{\tau_{\text{couple}} > t\}.$$

This shows the assertion. □

We now prove the convergence theorem for irreducible and aperiodic Markov chains on countable state spaces. The proof uses coupling.

Theorem 6.4.27 (Mixing for irreducible aperiodic chains)**.** *Let $(X_t)_{t\ge 0}$ be an irreducible aperiodic and positive recurrent Markov chain on a countable state space $\mathcal{S}$ with transition matrix P. Then P has a unique stationary distribution π and for each state x,*

$$\|P^t(x,\cdot) - \pi\|_{TV} \to 0 \quad \text{as } t \to \infty.$$

PROOF. The existence of the unique stationary distribution π for P follows from Theorem 6.3.18.

For the convergence, we will couple two P-chains (X_t) and (Y_t), where $X_0 = x$ and $Y_0 \sim \pi$, moving independently according to P in each step until the first time τ_{couple} they meet, after which they stay synchronized and move together according to P. Let δ_x denote the Dirac delta mass at x and $\mathbb{P}_{\delta_x \otimes \pi}$ the probability measure for the joint chain (X_t, Y_t). By Prop. 6.4.23 and since $X_t \sim P^t(x,\cdot)$ and $Y_t \sim \pi$,

$$\|P^t(x,\cdot) - \pi\|_{TV} \le \mathbb{P}_{\delta_x \otimes \pi}(X_t \ne Y_t) \le \mathbb{P}_{\delta_x \otimes \pi}(\tau_{\text{couple}} > t).$$

Therefore, it suffices to show that $\tau_{\text{couple}} < \infty$ almost surely.

We will use aperiodicity to show that $\tau_{\text{couple}} < \infty$ almost surely. (Otherwise it is not true. Why?) Consider first the 'product chain', a Markov chain on $\mathcal{S} \times \mathcal{S}$ with transition matrix Q given by

$$Q((x,y),(z,w)) = P(x,z)\,P(y,w) \quad \text{for all } (x,y),(z,w) \in \mathcal{S} \times \mathcal{S}.$$

This chain makes independent moves in the two coordinates, each according to the matrix P. Aperiodicity of P implies that Q is irreducible (Exc. 6.4.14). Hence, if (X_t, Y_t) is a chain started with product distribution $\mu \otimes \nu$ and run with transition matrix Q, then (X_t) is a Markov chain with transition matrix P and initial distribution μ, and (Y_t) is a Markov chain with transition matrix P and initial distribution ν. Note that

$$(\pi \otimes \pi)Q(z,w) = \sum_{(x,y)\in \mathcal{S}\times\mathcal{S}} \pi(x)\pi(y)P(x,z)P(y,w) = \sum_{x\in\mathcal{S}} \pi(x)P(x,z) \sum_{y\in\mathcal{S}} \pi(y)P(y,w).$$

Since $\pi = \pi P$, the right-hand side equals $\pi(z)\pi(w) = (\pi \otimes \pi)(z, w)$. Thus $\pi \otimes \pi$ is a stationary distribution for Q. By Theorem 6.3.18, the chain (X_t, Y_t) is positive recurrent. In particular, for any fixed x_0, if

$$\tau := \min\{t > 0 : (X_t, Y_t) = (x_0, x_0)\},$$

then by Prop. 6.2.10,

$$\mathbb{P}_{x,y}\{\tau < \infty\} = 1 \quad \text{for all } x, y \in X.$$

It follows that

$$\mathbb{P}_{x,\pi}(\tau < \infty) = \sum_{y \in \mathscr{S}} \pi(y)\, \mathbb{P}_{x,y}(\tau < \infty) = 1.$$

Thus if we initialize the product chain at $\delta_x \times \pi$ (δ_x =point mass at x), then it eventually visits a fixed 'diagonal state' (x_0, x_0). Since $\tau_{\text{couple}} \le \tau$ (why?), this is enough to conclude. □

6.5. Markov chain Monte Carlo

So far, we were given a Markov chain $(X_t)_{t \ge 0}$ on a finite state space Ω and studied existence and uniqueness of its stationary distribution and convergence to it. In this section, we will consider the reverse problem. Namely, given a distribution π on a sample space Ω, can we construct a Markov chain $(X_t)_{t \ge 0}$ such that π is a stationary distribution? If in addition the chain is irreducible an aperiodic, then by the convergence theorem (Theorem 6.4.18), we know that the distribution π_t of X_t converges to π. Hence if we run the chain for long enough, the state of the chain is asymptotically distributed as π. In other words, we can sample a random element of Ω according to the prescribed distribution π by emulating it through a suitable Markov chain. This method of sampling is called *Markov chain Monte Carlo* (MCMC).

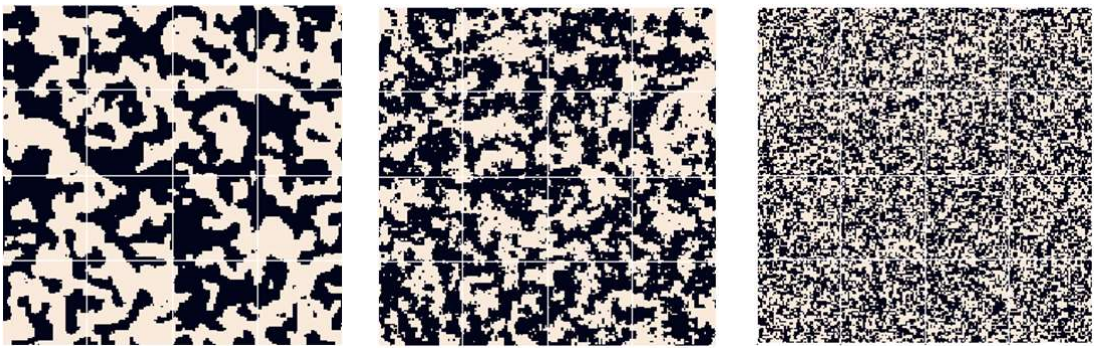


FIGURE 6.5.1. MCMC simulation of Ising model on 200 by 200 torus at temperature $T = 1$ (left), 2 (middle), and 5 (right).

Example 6.5.1 (Uniform distribution on regular graphs)**.** Let $G = (V, E)$ be a connected *regular* graph, meaning that all nodes have the same degree. Let μ be the uniform distribution on the node set V. How can we sample a random node according to μ? If we have a list of all nodes, then we can label them from 1 to $N = |V|$, choose a random number between 1 and N, and find corresponding node. But often times, we do not have the full list of nodes, especially when we want to sample a random node from a social network. Given only local information (set of neighbors for each given node), can we still sample a uniform random node from G?

One answer is given by random walk. Indeed, random walks on graphs are defined by only using local information of the underlying graph: Choose a random neighbor and move there. Moreover, since

G is connected, there is a unique stationary distribution π for the walk, which is given by

$$\pi(x) = \frac{\deg_G(x)}{2|E|}.$$

Since G is regular, any two nodes have the same degree, so $\pi(x) = \pi(y)$ for all $x, y \in V$. This means π equals the uniform distribution μ on V. Hence the sampling algorithm we propose is as follows:

$(*)$ *Run a random walk on G for $t \gg 1$ steps, and take the random node that the walk sits on.*

However, there is a possible issue of convergence. Namely, if the graph G does not contain any odd cycle, then random walk on G is periodic (see Exercise 6.4.12), so we are not guaranteed to have convergence. We can overcome this by using a lazy random walk instead, which is introduced in Exercise 6.5.2. We know that the lazy RW on G is irreducible, aperiodic, and has the same set of stationary distribution as the ordinary RW on G. Hence we can apply the sampling algorithm $(*)$ above for lazy random walk on G to sample a uniform random node in G. ▲

Exercise 6.5.2 (Lazy RW on graphs)**.** Let $G = (V, E)$ be a graph. Let $(X_t)_{t\geq 0}$ be a Markov chain on the node set V with transition probabilities

$$\mathbb{P}(X_{t+1} = j \,|\, X_t = i) = \begin{cases} 1/2 & \text{if } j = i \\ 1/(2\deg_G(i)) & \text{if } j \text{ is adjacent to } i \\ 0 & \text{otherwise.} \end{cases}$$

This chain is called the *lazy random walk* on G. In words, the usual random walker on G now flips a fair coin to decide whether it stays at the same node or make a move to one of its neighbors.

(i) Show that for any connected graph G, the lazy random walk on G is irreducible and aperiodic.
(ii) Let P be the transition matrix for the usual random walk on G. Show that the following matrix

$$Q = \frac{1}{2}(P + I)$$

is the transition matrix for the lazy random walk on G.
(iii) For any distribution π on V, show that $\pi Q = \pi$ if and only if $\pi P = \pi$. Conclude that the usual and lazy random walks on G have the same set of stationary distribution.

Example 6.5.3 (Finding local minima)**.** Let $G = (V, E)$ be a connected graph and let $f : V \to [0, \infty)$ be a 'cost' function. The objective is to find a node $x^* \in V$ such that f takes global minimum at x^*. This problem has a lot of application in machine learning, for example. Note that if the domain V is very large, then an exhaustive search is too expensive to use.

Here is simple form of the popular algorithm of *stochastic gradient descent*, which lies at the heart of most of the important machine learning algorithms.

(i) Initialize the first guess $X_0 = x_0 \in V$.
(ii) Suppose $X_t = x \in V$ is chosen. Let

$$D_t = \{y \text{ a neighbor of } x \text{ or } x \text{ itself} \,|\, f(y) \leq f(x)\}.$$

Define X_{t+1} to be a uniform random node from D_t.
(iii) The algorithm terminates if it finds a local minima.

In words, at each step we move to a random neighbor which could possibly decrease the current value of f. It is easy to see that one always converges to a local minima, which may not be a global minimum. In a very complex machine learning task (e.g., training a deep neural network), this is often good enough. Is this a Markov chain? Irreducible? Aperiodic? Stationary distribution? ▲

There is a neat solution to finding global minimum. The idea is to allow that we go uphill with a small probability.

Example 6.5.4 (Finding global minimum)**.** Let $G = (V, E)$ be a connected regular graph and let $f : V \to [0, \infty)$ be a cost function. Let

$$V^* = \left\{ x \in V \mid f(x) = \min_{y \in V} f(y) \right\}$$

be the set of all nodes where f attains global minimum.

Fix a parameter $\lambda \in (0, 1]$, and define a probability distribution π_λ on V by

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z_\lambda},$$

where $Z_\lambda = \sum_{x \in V} \lambda^{f(x)}$ is the normalizing constant. Since $\pi_\lambda(x)$ is decreasing in $f(x)$, it favors nodes x for which $f(x)$ is small.

Let $(X_t)_{t \geq 0}$ be Markov chain on V, whose transition is defined as follows. If $X_t = x$, then let y be a uniform random neighbor of x. If $f(y) \leq f(x)$, then move to y; If $f(y) > f(x)$, then move to y with probability $\lambda^{f(y)-f(x)}$ and stay at x with probability $1 - \lambda^{f(y)-f(x)}$. We analyze this MC below:

(i) (Irreducibility) Since G is connected and we are allowing any move (either downhill or uphill) we can go from one node to any other in some number of steps. Hence the chain $(X_t)_{t \geq 0}$ is irreducible.

(ii) (Aperiodicity) By (i) and Remark 6.4.13, all nodes have the same period. Moreover, let $x \in V^*$ be an arbitrary node where f takes global minimum. Then all return times are possible, so x has period 1. Hence all nodes have period 1, so the chain is aperiodic.

(iii) (Stationarity) Here we show that π_λ is a stationary distribution of the chain. To do this, we first need to write down the transition matrix P. Namely, if we let $A_G(y, z)$ denote the indicator that y and z are adjacent, then

$$P(x, y) = \begin{cases} \frac{A_G(x,y)}{\deg_G(x)} \min(1, \lambda^{f(y)-f(x)}) & \text{if } x \neq y \\ 1 - \sum_{y \neq x} P(x, y) & \text{if } y = x. \end{cases}$$

To show $\pi_\lambda P = \pi_\lambda$, it suffices to show for any $y \in V$ that

$$\sum_{z \in V} \pi_\lambda(z) P(z, y) = \pi_\lambda(y).$$

Note that

$$\begin{aligned} \sum_{z \in V} \pi_\lambda(z) P(z, y) &= \pi_\lambda(y) P(y, y) + \sum_{z \neq y} \pi_\lambda(z) P(z, y) \\ &= \pi_\lambda(y) - \sum_{z \neq y} \pi_\lambda(y) P(y, z) + \sum_{z \neq y} \pi_\lambda(z) P(z, y). \end{aligned}$$

Hence it suffices to show that

$$\pi_\lambda(y) P(y, z) = \pi_\lambda(z) P(z, y) \tag{101}$$

for any $z \neq y$. Indeed, considering the two cases $f(z) \leq f(y)$ and $f(z) > f(y)$, we have

$$\begin{aligned} \pi_\lambda(y) P(y, z) &= \frac{\lambda^{f(y)}}{Z_\lambda} \frac{A_G(y, z)}{\deg_G(y)} \min(1, \lambda^{f(z)-f(y)}) = \frac{1}{Z_\lambda} \frac{A_G(z, y)}{\deg_G(z)} \lambda^{\max(f(y), f(z))}, \\ \pi_\lambda(z) P(z, y) &= \frac{\lambda^{f(z)}}{Z_\lambda} \frac{A_G(z, y)}{\deg_G(z)} \min(1, \lambda^{f(y)-f(z)}) = \frac{1}{Z_\lambda} \frac{A_G(y, z)}{\deg_G(y)} \lambda^{\max(f(y), f(z))}. \end{aligned}$$

Now since $A_G(z, y) = A_G(y, z)$ and we are assuming G is a regular graph, this yields (101), as desired. Hence π_λ is a stationary distribution for the chain $(X_t)_{t \geq 0}$.

(iv) (Convergence) By (i), (iii), Theorem 6.3.18, we see that π_λ is the unique stationary distribution for the chain X_t. Furthermore, by (i)-(iii) and Theorem 6.4.18, we conclude that the distribution of X_t converges to π_λ.

(v) (Global minimum) Let $f_* = \min_{x\in V} f(x)$ be the global minimum of f. Note that by definition of V^*, we have $f(x) = f_*$ for any $x \in V^*$. Then observe that

$$\lim_{\lambda\to 0} \pi_\lambda(x) = \lim_{\lambda\to 0} \frac{\lambda^{f(x)}}{\sum_{y\in V}\lambda^{f(y)}} = \lim_{\lambda\to 0} \frac{\lambda^{f(x)}/\lambda^{f_*}}{\sum_{y\in V}\lambda^{f(y)}/\lambda^{f_*}}$$
$$= \lim_{\lambda\to 0} \frac{\lambda^{f(x)-f_*}}{|V^*| + \sum_{y\in V\setminus V^*}\lambda^{f(y)-f_*}} = \frac{\mathbf{1}(x\in V^*)}{|V^*|}.$$

Thus for λ very small, π_λ is approximately the uniform distribution on the set of all nodes V^* where f attains global minimum. ▲

Exercise 6.5.5 (Detailed Balance equation)**.** Let P be a transition matrix of a Markov chain on state space Ω. Suppose π is a probability distribution on Ω that satisfies the following *detailed balance equation*

$$\pi(x)P(x,y) = \pi(y)P(y,x) \qquad \text{for all } x, y \in \Omega.$$

(i) Show that for all $x \in \Omega$,

$$\sum_{y\in\Omega} \pi(y)P(y,x) = \sum_{y\in\Omega} \pi(x)P(x,y) = \pi(x)\sum_{y\in\Omega} P(x,y) = \pi(x).$$

(ii) Conclude that π is a stationary distribution for P, that is, $\pi P = \pi$.

Exercise 6.5.6 (Metropolis-Hastings algorithm)**.** Let P be a transition matrix of a Markov chain on state space $\Omega = \{1, 2, \cdots, m\}$. Let π be a probability distribution on Ω, which is not necessarily a stationary distribution for P. Our goal is to design a Markov chain on Ω that has π as a stationary distribution. Below we will derive the famous *Metropolis-Hastings algorithm* for MCMC sampling.

Fix a $m \times m$ matrix A of entries from $[0, 1]$. Consider a Markov chain $(X_t)_{t\geq 0}$ on Ω that uses additional rejection step on top of the transition matrix P as follows:

(`Generation`) Suppose the current location $X_t = a$. Generate a candidate $b \in \Omega$ from the conditional distribution $P(a, \cdot)$.

(`Rejection`) Flip an independent coin with success probability $A(a, b)$. Upon success, accept the proposed move and set $X_{t+1} = b$; Otherwise, reject the move and set $X_{t+1} = a$.

Here, the (a, b) entry $A(a, b)$ is called the *acceptance probability* of the move $a \to b$.

(i) Let Q denote the transition matrix of the chain $(X_t)_{t\geq 0}$ defined above. Show that

$$Q(a,b) = \begin{cases} P(a,b)A(a,b) & \text{if } b \neq a \\ 1 - \sum_{c:c\neq a} P(a,c)A(a,c) & \text{if } b = a. \end{cases}$$

(ii) Show that $\pi(x)Q(x, y) = \pi(y)Q(y, x)\ \forall x, y \in \Omega$, $x \neq y$ implies $\pi Q = \pi$. Deduce that if

$$\pi(x)P(x,y)A(x,y) = \pi(y)P(y,x)A(y,x) \qquad \forall x, y \in \Omega,\ x \neq y,$$

then π is a stationary distribution for $(X_t)_{t\geq 0}$.

(iii) Since we also want the Markov chain to converge fast, we want to choose the acceptance probability $A(a, b) \in [0, 1]$ as large as possible for each $a, b \in \Omega$. Show that the following choice (denoting $\alpha \wedge \beta := \min(\alpha, \beta)$)

$$A(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)P(x,y)} \wedge 1 \qquad \forall x, y \in \Omega,\ x \neq y$$

satisfies the condition in (ii) and each $A(x, y)$ is maximized for all $x \neq y$.

(iv) Let $(Y_t)_{t\geq 0}$ be a random walk on the 5-wheel graph $G = (V, E)$ as shown in Figure 6.5.2. Show that $\pi = \left[\frac{5}{20}, \frac{3}{20}, \frac{3}{20}, \frac{3}{20}, \frac{3}{20}, \frac{3}{20}\right]$ is the unique stationary distribution of Y_t. Apply the Metropolis-Hastings algorithm derived in (i)-(iii) above to modify Y_t to obtain a new Markov chain X_t on V that converges to Uniform(V) in distribution.

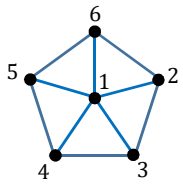


FIGURE 6.5.2. 5-wheel graph G

CHAPTER 7

Brownian Motion

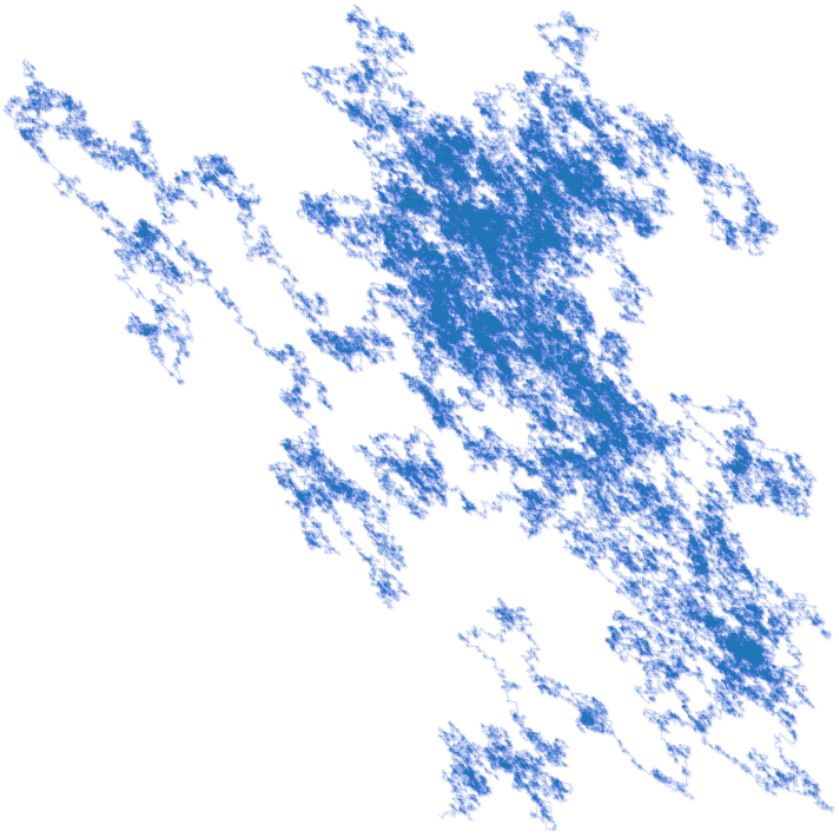


FIGURE 7.0.1. A sample path of a 2-dimensional Brownian motion

7.1. Definition and basic properties of Brownian motion

7.1.1. Definition of Brownian motion. A significant portion of probability theory focuses on elucidating the overarching patterns that emerge in random systems governed by myriad microscopic random influences. Brownian motion, for instance, illustrates the larger-scale behavior resulting from a particle's random movement in d-dimensional space. At the microscopic level, with each time step, the particle undergoes a random displacement, perhaps due to collisions with other particles or external forces. Thus, if its initial position at time zero is denoted as S_0, its position at time n can be expressed as $S_n = S_0 + \sum_{i=1}^{n} X_i$, where $X_1, X_2, X_3, \ldots$ represent independent, identically distributed random variables with values in $\mathbb{R}^d$. This sequence $\{S_n : n \geq 0\}$ constitutes a random walk, with the displacements portraying the microscopic inputs.

When contemplating the macroscopic perspective, we ponder questions such as:

- Does S_n trend towards infinity?

- Does S_n repeatedly return to the vicinity of the origin?
- How swiftly does $\max\{|S_1|,\dots,|S_n|\}$ increase as $n\to\infty$?
- What is the anticipated number of circuits made by $\{S_n : n\geq 0\}$ around the origin?

Notably, not all attributes of the microscopic inputs influence the macroscopic outlook. Specifically, only the mean and covariance of the displacements shape this perspective. Essentially, random walks sharing identical mean and covariance matrices yield the same macroscopic process. Moreover, the requirement for displacements to be independent and identically distributed can be considerably relaxed. This phenomenon, termed universality, underscores that the macroscopic process often referred to as a universal object. In probability studies, it is a common practice to explore diverse phenomena through their associated universal objects. Brownian motion is such a universal object for a large class of stochastic process including random walks.

Brownian motion is closely linked to the normal distribution. Recall that a random variable X is normally distributed with mean μ and variance σ^2 if

$$P\{X > x\} = \frac{1}{\sqrt{2\pi\sigma^2}}\int_x^\infty e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du, \quad \text{for all } x\in\mathbb{R}.$$

If $\mu = 0$ and $\sigma^2 = 1$, then the above becomes the complementary distribution function for the standard normal distribution, denoted $N(0,1)$.

Definition 7.1.1 (Multivariate normal distribution)**.** We say a random vector $X = (X_1, X_2, \dots, X_d)^T \in \mathbb{R}^d$ has a *multivariate normal distribution* $N(\mu,\Sigma)$ with mean $\mu\in\mathbb{R}^d$ and covariance matrix $\Sigma\in\mathbb{R}^{d\times d}$ if $\mathbb{E}[X] = \mu$ and

$$\Sigma[i,j] = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathrm{Cov}(X_i,X_j) \quad \text{for all } 1\leq i\leq j\leq d.$$

The joint probability density function f_X of X is given by: For $x = (x_1,\dots,x_d)^T\in\mathbb{R}^d$,

$$f_X(x_1,x_2,\dots,x_n) = \frac{1}{\sqrt{(2\pi)^n\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Exercise 7.1.2 (Gaussian random vector)**.**

(i) If $Z = (Z_1,\dots,Z_d)^T$ consists of i.i.d. standard normal coordinates, then show that $Z\sim N(\mathbf{0}, I_d)$. Such Z is called a 'standard normal vector'.

(ii) Let $\mu\in\mathbb{R}^p$ and $A\in\mathbb{R}^{p\times d}$. Let $X := AZ+\mu$ where $Z\sim N(\mathbf{0},I_d)$. Show that $X\sim N(\mu, AA^T)$.

(iii) Let $\mu\in\mathbb{R}^p$ and $\Sigma\in\mathbb{R}^{d\times d}$ be a symmetric and positive semi-definite matrix. Show that there exists a matrix $A\in\mathbb{R}^{p\times p}$ such that $X = AZ+\mu$ with $Z\in N(\mathbf{0},I_p)$ follows the multivariate normal distribution $N(\mu,\Sigma)$. (*Hint:* Use spectral decomposition of real symmetric matrices to write $\Sigma = UDU^T$ with U orthogonal and D diagonal. Since Σ is PSD, D is a diagonal matrix of nonnegative entries. Let $A = U\sqrt{D}$ and use the previous parts.)

Below we give an axiomatic definition of Brownian motion, listing all properties we desire to endow on it.

Definition 7.1.3 (Brownian motion)**.** A stochastic process $(B(t) : t\geq 0)$ in $\mathbb{R}^d$ is called a d-dimensional *Brownian motion* (BM) with initial location $x\in\mathbb{R}^d$, drift $\mu\in\mathbb{R}^d$, and diffusion matrix $\Sigma\in\mathbb{R}^{d\times d}$ if the following holds: (Denote $B\sim BM(x,\mu,\Sigma)$)

(i) $B(0) = x$;

(ii) the process has independent increments, i.e., for all times $0\leq t_1\leq t_2\leq\ldots\leq t_n$ the increments $B(t_n)-B(t_{n-1}), B(t_{n-1})-B(t_{n-2}),\dots,B(t_2)-B(t_1)$ are independent random vectors;

(iii) for all $t\geq 0$ and $h>0$, the increments $B(t+h)-B(t)\sim N(h\mu, h\Sigma)$;

(iv) almost surely, the function $t\mapsto B(t)$ is continuous.

We say that $(B(t) : t\geq 0)$ is a standard Brownian motion if $x = 0$, $\mu = \mathbf{0}$, and $\Sigma = I_d$.

In conditions **(i)**-**(iii)** in the definition of Brownian motion, we specified the joint law of $B(0)$ and all the increments $(B(t_1)-B(0), B(t_2)-B(t_1), \ldots, B(t_n)-B(t_{n-1}))$, for all $0 \le t_1 \le t_2 \le \ldots \le t_n$. This specifies the distribution of the tuple of 'snapshots' $(B(t_1), B(t_2), \ldots, B(t_n))$ of the Brownian motion for any given sequence of times $0 \le t_1 < t_2 < \cdots < t_n$. In other words, this specifies the 'marginal distributions' of Brownian motion.

Definition 7.1.4 (Marginal distribution)**.** Given a Brownian motion $(B(t) : t \ge 0)$, By the marginal distributions of a stochastic process $(B(t) : t \ge 0)$, we mean the laws of all the n-tuples of finite-dimensional random vectors $(B(t_1), B(t_2), \ldots, B(t_n))$, for all $0 \le t_1 \le t_2 \le \ldots \le t_n$.

Condition **(iv)** in Def. 7.1.3 concerns almost sure continuity of the 'sample paths' of the Brownian motion. We have defined Brownian motion as a stochastic process $(B(t) : t \ge 0)$, which is just a family of (uncountably many) random variables $\omega \mapsto B(t,\omega)$ defined on a single probability space $(\Omega, \mathscr{F}, \mathbb{P})$. At the same time, a stochastic process can also be interpreted as a 'random function' with the sample functions defined by $t \mapsto B(t,\omega)$, which we refer to the 'sample path' with randomness specified by ω (see Fig. 7.0.1). The sample path properties of a stochastic process are the properties of these random functions, and it is these properties we will be most interested in the study of Brownian motion.

Sample path properties (e.g., continuity and differentiability) goes beyond the marginal distributions of the process in the sense above. For instance, the set $\{\omega \in \Omega : t \mapsto B(t,\omega) \text{ continuous}\}$ is, in general, not in the σ-algebra generated by the random vectors $(B(t_1), , \ldots, B(t_n))$, for $n \in \mathbb{N}$. Hence, if we are interested in the sample path properties of a stochastic process, we may need to specify more than just its marginal distributions.

Definition 7.1.5 (Sample path properties)**.** Suppose $\mathfrak{X}$ is a property a function might or might not have, like continuity, differentiability, etc. We say that a process $(X(t) : t \ge 0)$ *has property* $\mathfrak{X}$ *almost surely* if there exists $A \in \mathscr{F}$ such that $P(A) = 1$ and

$$A \subseteq \{\omega \in \Omega : t \mapsto X(t,\omega) \text{ has property } \mathfrak{X}\}.$$

(Note that these sets on the right need not lie in $\mathscr{F}$.)

7.1.2. Existence of Brownian motion: Lévy's construction. In this section, we show the existence of Brownian motion by constructing it. The first step is to reduce it to the construction of 1D standard Brownian motion (SBM).

Exercise 7.1.6 (Constructing BM from 1D SBM)**.**

(i) Let $B^{(1)}, \ldots, B^{(d)}$ be i.i.d. standard one-dimensional Brownian motions. Show that $B = (B^{(1)}, \ldots, B^{(d)})$ is a standard Brownian motion in $\mathbb{R}^d$.

(ii) Suppose $(B_t)_{t \ge 0}$ is a stadard Brownian motion in $\mathbb{R}^d$. Fix $\mu \in \mathbb{R}^d$ and a symmetric PSD matrix $\Sigma \in \mathbb{R}^{d \times d}$. By spectral decomposition, write $\Sigma = AA^T$ for some $A \in \mathbb{R}^{p \times p}$. Show that $\tilde{B}_t := AB_t + \mu$, $t \ge 0$ defines a Brownian motion in $\mathbb{R}^d$ with drift μ and diffusion matrix Σ.

In the remainder of this section, we will prove the following celebrated result on the existence of SBM. The argument uses Lévy's construction of 1D SBM.

Theorem 7.1.7 (Wiener 1923)**.** *Standard Brownian motion in* $\mathbb{R}$ *exists.*

PROOF. The following is a famous construction of SBM due to Paul Lévy. We will construct Browian motion on the unit interval $[0,1]$ as a random element in the space $\mathscr{C}[0,1]$ of continuous functions $[0,1] \to \mathbb{R}$. The key idea for ensuring contiuity is to construct it as the uniform limit of continuous functions over discrete time points that become finer in the limit.

We will use 'diadic partition' of the unit interval $[0,1]$. Define for $n \ge 1$,

$$\mathscr{D}_n := \{k2^{-n} : 1 \le k \le 2^n\}.$$

Namely, if we partition $[0,1]$ into 2^n intervals of equal lenths, then the endpoints form $\mathscr{D}_n$. We will define values of the BM on the poitns on $\mathscr{D}_n$ and linearly interpolate them. We will then check that the uniform limit of these continuous functions exists as $n \to \infty$ and is a Brownian motion.

Step 1. Constructing a Gaussian process over diadic points

Let $\mathscr{D} = \bigcup_{n=0}^{\infty} \mathscr{D}_n$, and let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space on which a collection $\{Z_t : t \in \mathscr{D}\}$ of independent, standard normally distributed random variables can be defined. Inductively on $n \geq 0$, we will define $B(x)$ for all points $x \in \mathscr{D}$. For $n = 0$, define $B(0) := 0$ and $B(1) := Z_1$. Having defined B on $\mathscr{D}_{n-1}$, define $B(x)$ for $x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}$ by

$$\begin{cases} B(0) = 0 \quad B(1) = Z_1 \\ B(x) = \frac{B(x-2^{-n}) + B(x+2^{-n})}{2} + \frac{Z_x}{2^{(n+1)/2}} \quad \text{for all } n \geq 1 \text{ and } x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}. \end{cases}$$

This defines a process B on all points in $\mathscr{D}$.

Next, we show that the process B on $\mathscr{D}$ is a 'Gaussian process' in the following sense. For each $n \geq 1$,

(1) the vectors $(B(x) : x \in \mathscr{D}_n)$ and $(Z_t : t \in \mathscr{D} \setminus \mathscr{D}_n)$ are independent; and
(2) for all $r < s < t$ in $\mathscr{D}_n$, the random variable $B(t) - B(s) \sim N(0, t-s)$ and is independent of $B(s) - B(r)$.

From the above construction, it is clear that $B(x)$ for $x \in \mathscr{D}_n$ is determined by the normal variables Z_s for $s \in \mathscr{D}_n$. Since all these normal variables are independent, the first property follows.

Next, in order to show the second property above, it suffices to show that all increments $B(x) - B(x - 2^{-n})$ over an interal of length 2^{-n}, for $x \in D_n \setminus \{0\}$, are independent. To show the latter, this, it suffices to show that they are pairwise independent, as the vector of these increments is Gaussian (why?).

First, consider two increments of the form $B(x) - B(x-2^{-n})$, $B(x+2^{-n}) - B(x)$ for $x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}$ are independent. Indeed, as $\frac{1}{2}[B(x+2^{-n}) - B(x-2^{-n})]$ depends only on $(Z_t : t \in \mathscr{D}_{n-1})$, it is independent of $Z_x/2^{(n+1)/2}$. Both terms are normally distributed with mean zero and variance $2^{-(n+1)}$. Hence, their sum $B(x) - B(x-2^{-n})$ and their difference $B(x+2^{-n}) - B(x)$ are independent and normally distributed with mean zero and variance 2^{-n} by Exercise 7.1.8.

Second, the other possibility is that the increments are over intervals separated by some $x \in \mathscr{D}_{n-1}$. Choose $x \in \mathscr{D}_j$ with this property and minimal j, so that the two intervals are contained in $[x - 2^{-j}, x]$, respectively $[x, x + 2^{-j}]$. By induction, the increments over these two intervals of length 2^{-j} are independent, and the increments over the intervals of length 2^{-n} are constructed from the independent increments $B(x) - B(x-2^{-j})$, respectively $B(x+2^{-j}) - B(x)$, using a disjoint set of variables $(Z_t : t \in \mathscr{D}_n)$. Hence, they are independent, and this implies the first property, and completes the induction step.

Step 2. Linear interpolation, partial sums representation, and pointwise limit

Next, having thus chosen the values of the process B on all dyadic points in $\mathscr{D}$, we interpolate between them, thereby extending B to all points in $[0,1]$.

Formally, define

$$F_0(t) = \begin{cases} Z_1 & \text{for } t = 1 \\ 0 & \text{for } t = 0 \\ \text{linearly in between} & \end{cases}$$

and for each $n \geq 1$, define

$$F_n(t) = \begin{cases} 2^{-(n+1)/2} Z_t & \text{for } t \in \mathscr{D}_n \setminus \mathscr{D}_{n-1} \\ 0 & \text{for } t \in \mathscr{D}_{n-1} \\ \text{linear between consecutive points in } \mathscr{D}_n. & \end{cases}$$

These functions are piecewise linear, and hence continuous. Define a function $\overline{B} : [0,1] \to \mathbb{R}$ by

$$\overline{B}(x) = \sum_{i=0}^{\infty} F_i(x) \tag{102}$$

At this point, we have not shown that the infinite series in the RHS above is convergent at every point in $[0,1]$. But it is so trivially for diadic points $x \in \mathscr{D}$, since $F_i(x) \equiv 0$ for all $i > n$ and $x \in \mathscr{D}_n$ by construction.

We claim that for all $n \geq 0$,

$$B(x) = \sum_{i=0}^{n} F_i(x) = \sum_{i=0}^{\infty} F_i(x) = \overline{B}(x) \qquad \text{for all } x \in \mathscr{D}_n. \tag{103}$$

We have just shown that the second equality holds. The first equality can be seen by an inudction in n. It holds for $n = 0$ trivially. Suppose it holds for $n-1$. Let $x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}$. For $0 \leq i \leq n-1$ the function F_i is linear on $[x-2^{-n}, x+2^{-n}]$ and $x \pm 2^{-n} \in \mathscr{D}_{n-1}$. So by the induction hypothesis,

$$\begin{aligned}
\sum_{i=0}^{n-1} F_i(x) &= \sum_{i=0}^{n-1} \frac{F_i(x-2^{-n}) + F_i(x+2^{-n})}{2} \\
&= \frac{1}{2}\left(\sum_{i=0}^{n-1} F_i(x-2^{-n}) + \sum_{i=0}^{n-1} F_i(x+2^{-n}) \right) \\
&= \frac{1}{2}\left(B(x-2^{-n}) + B(x+2^{-n}) \right).
\end{aligned}$$

Since $F_n(x) = 2^{-(n+1)/2} Z_x$, this shows

$$\sum_{i=0}^{n} F_i(x) = \left(\sum_{i=0}^{n-1} F_i(x) \right) + F_n(x) = \frac{B\left(x-2^{-n}\right) + B\left(x+2^{-n}\right)}{2} + \frac{Z_x}{2^{(n+1)/2}} = B(x).$$

This completes the induction so we have shown (103) holds for all $n \geq 0$.

Step 4. Obtaining a uniform limit of continuous linear interpolations

Next, we show that the infinite series function $\overline{B}$ in (102) is uniformly convergent over $[0,1]$. Since each partial sum is continuous, this will imply that the limit $\overline{B}$ is continuous on $[0,1]$. For this, we will show that the supremum norm $\|F_n\|_\infty$ is so small that it is summable.

First note that for any $s > 0$,

$$\mathbb{P}(\|F_n\|_\infty \geq s) \leq \mathbb{P}\left(2^{-(n+1)/2}|Z_x| \geq s \text{ for some } x \in \mathscr{D}_n\right).$$

Hence choosing $s = c\sqrt{n}2^{-n/2}$ for a constant $c > 0$, we have

$$\begin{aligned}
\sum_{n=0}^{\infty} \mathbb{P}(\|F_n\|_\infty \geq c\sqrt{n}2^{-n/2}) &\leq \sum_{n=0}^{\infty} \mathbb{P}\left(|Z_x| \geq c\sqrt{n} \text{ for some } x \in \mathscr{D}_n\right) \\
&\leq \sum_{n=0}^{\infty} (2^n+1)\mathbb{P}\left(|Z_0| \geq c\sqrt{n}\right) \\
&\leq \sum_{n=0}^{\infty} (2^n+1)\exp\left(-cn^2/2\right),
\end{aligned}$$

where the second inequality follows from a union bound and $|\mathscr{D}_n| \leq 2^n+1$ and the third inequality follows from a tail bound on Gaussian distribution (Exc. 7.1.9). The last sum converges as soon as $c > \sqrt{2}\log 2$. Fix such a c. By the Borel-Cantelli lemma, there exists a random (but almost surely finite) N such that

$$\|F_n\|_\infty \leq c\sqrt{n}2^{-n/2} \quad \text{for all } n > N \text{ almost surely.} \tag{104}$$

This implies that, almost surely, the series

$$\overline{B}(t) = \sum_{n=0}^{\infty} F_n(t)$$

is uniformly convergent on $[0,1]$ almost surely, as desired. Since $\overline{B}$ agrees with B on $\mathscr{D}$, we can identify them and extend the values of B to the entire unit interval $[0,1]$ by $\overline{B}$. Henceforth, we will write B for $\overline{B}$.

Step 5. Verifying marginal distribution of SBM

It remains to check that the increments of the limiting continuous process have the right marginal distributions. This follows directly from the properties of B on the dense set $\mathscr{D} \subset [0,1]$ and the continuity of the paths. Indeed, suppose that $t_1 < t_2 < \ldots < t_n$ are in $[0,1]$. We find $t_{1,k} \le t_{2,k} \le \ldots \le t_{n,k}$ in $\mathscr{D}$ with $\lim_{k\to\infty} t_{i,k} = t_i$ and infer from the continuity of B that, for $1 \le i \le n-1$,

$$B(t_{i+1}) - B(t_i) = \lim_{k\to\infty} B(t_{i+1,k}) - B(t_{i,k}).$$

As $\lim_{k\to\infty} \mathbb{E}[B(t_{i+1,k}) - B(t_{i,k})] = 0$ and

$$\lim_{k\to\infty} \mathrm{Cov}\big(B(t_{i+1,k}) - B(t_{i,k}), B(t_{j+1,k}) - B(t_{j,k})\big) = \lim_{k\to\infty} \mathbf{1}_{(i=j)}(t_{i+1,k} - t_{i,k}) = \mathbf{1}_{(i=j)}(t_{i+1} - t_i),$$

the increments $B(t_{i+1}) - B(t_i)$ are, by Exc. 7.1.10, independent Gaussian random variables with mean 0 and variance $t_{i+1} - t_i$, as required.

Step 6. Extending to the whole real line

We have thus constructed a continuous process $B : [0,1] \to \mathbb{R}$ with the same marginal distributions as Brownian motion. Take a sequence $B_1, B_2, \ldots$ of independent $\mathscr{C}[0,1]$-valued random variables with the distribution of this process, and define $(B(t) : t \ge 0)$ by gluing together the parts. That is,

$$B(t) = B_{\lfloor t \rfloor}(t - \lfloor t \rfloor) + \sum_{i=0}^{\lfloor t \rfloor - 1} B_i(1) \quad \text{for } t \ge 0.$$

This defines a continuous random function $B : [0,\infty) \to \mathbb{R}$, and one can see easily from what we have shown so far that the requirements of a standard Brownian motion are fulfilled. □

Exercise 7.1.8. Let $X, Y \sim N(0,\sigma^2)$ be independent mean zero variance σ^2 normal random variables. Show that $X+Y, X-Y \sim N(0,2\sigma^2)$ and $X+Y$ and $X-Y$ are independent.

Exercise 7.1.9 (Tail bounds on standard normal distribution)**.** Suppose $Z \sim N(0,1)$. Then for all $x > 0$,

$$\frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \le \mathbb{P}(Z > x) \le \frac{1}{x}\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Exercise 7.1.10 (Gaussianity preserved under limits)**.** Suppose $\{X_n : n \in \mathbb{N}\}$ is a sequence of Gaussian random vectors and $\lim_{n\to\infty} X_n = X$ almost surely. If $b := \lim_{n\to\infty} \mathbb{E}[X_n]$ and $C := \lim_{n\to\infty} \mathrm{Cov}[X_n]$ exist, then X is Gaussian with mean b and covariance matrix C.

Exercise 7.1.11 (BM as a Gaussian process)**.** A stochastic process $(Y(t) : t \ge 0)$ is called a Gaussian process if for all $t_1 < t_2 < \ldots < t_n$, the vector $(Y(t_1), \ldots, Y(t_n))$ is a Gaussian random vector (see Exc. 7.1.2). Show that Brownian motion starting at $x \in \mathbb{R}$ is a Gaussian process.

Exercise 7.1.12 (Uniqueness of Brownian motion)**.** Let B be a standard Brownian motion. Let $\mathscr{D}_n := \{k2^{-n} : 1 \le k \le 2^n\}$ and $\mathscr{D} := \bigcup_{n=0}^{\infty} \mathscr{D}_n$.

(i) We will construct a collection $(Z_x : x \in \mathscr{D})$ of random variables using B recursively as follows:

$$\begin{cases} B(0) = 0 \quad B(1) = Z_1 \\ Z_x = -2^{(n-1)/2}\left[(B(x+2^{-n}) - B(x)) - (B(x) - B(x-2^{-n}))\right] \quad \text{for all } n \ge 1 \text{ and } x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}. \end{cases}$$

Show that Z_x's are i.i.d. $N(0,1)$ RVs. (*Hint*: By induction, show that Z_x's for all $x \in \mathscr{D}_n$ are i.i.d. $N(0,1)$. For the induction step, let $x \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}$. Note that $x \pm 2^{-n} \in \mathscr{D}_{n-1}$. By independent increments of B, $X = B(x+2^{-n}) - B(x)$ and $Y = B(x) - B(x-2^{-n})$ are i.i.d. $N(0,2^{-n})$. Then $Y+X = B(x+2^{-n}) - B(x-2^{-n})$ and $Y - X = 2^{(n+1)/2} Z_x$ are independent. By independent increments of B, this shows that Z_x is independent from Z_y for all $y \in \mathscr{D}_n \setminus \mathscr{D}_{n-1}$ and $y \ne x$. In order to show that Z_x is independent from Z_y for $y \in \mathscr{D}_{n-1}$, note that such Z_y's are functions of the increments of B over $\mathscr{D}_{n-1}$ by induction. We have already seen that $Z_x \perp B(x+2^{-n}) - B(x-2^{-x})$. Other increments of B over consecutive points in $\mathscr{D}_{n-1}$ are over disjoint intervals from $[x \pm 2^{-n}]$, so again by independent increments, they are independent from Z_x. This completes the induction step.)

(ii) Let $(Z_x : x \in \mathscr{D})$ be as constructed in **(i)**. Define piecewise linear functions F_n for $n \geq 0$ as follows: For F_0, let $F_0(0) = 0$ and $F_0(1) = Z_1$ and linearly interpolate in between. For $n \geq 1$, let

$$F_n(t) = \begin{cases} 2^{-(n+1)/2} Z_t & \text{for } t \in \mathscr{D}_n \setminus \mathscr{D}_{n-1} \\ 0 & \text{for } t \in \mathscr{D}_{n-1} \\ \text{linear between consecutive points in } \mathscr{D}_n. \end{cases}$$

In the proof of Lévy's construction of SBM, we have shown that

$$B(x) = \sum_{n=0}^{\infty} F_n(x) \quad \text{for } x \in \mathscr{D},$$

and for $c > \sqrt{2}\log 2$,

$$\sum_{n=0}^{\infty} \mathbb{P}(\|F_n\|_\infty \geq c\sqrt{n}\, 2^{-n/2}) \leq \sum_{n=0}^{\infty} (2^n + 1) \exp\left(-cn^2/2\right) < \infty.$$

By Borel-Cantelli lemma, this implies that $\sum_{n=0}^{\infty} F_n$ is uniformly convergent on $[0,1]$ almost surely. Show that, almost surely,

$$B(x) = \sum_{n=0}^{\infty} F_n(x) \quad \text{for } x \in [0,1].$$

(*Hint:* Show that $\mathscr{D}$ is a dense subset of $[0,1]$.)

7.1.3. Scaling invariance of Brownian motion. The behavior is significantly influenced by a fundamental property of Brownian motion known as scaling invariance. This property describes a transformation on the functions' space that alters individual Brownian random functions while preserving their distribution.

Lemma 7.1.13 (Scaling invariance)**.** *Suppose* $(B(t) : t \geq 0)$ *is a standard Brownian motion and let* $a > 0$. *Then the process* $(X(t) : t \geq 0)$ *defined by* $X(t) = \frac{1}{a} B(a^2 t)$ *is also a standard Brownian motion.*

PROOF. Continuity of the paths, independence, and stationarity of the increments remain unchanged under the rescaling. It remains to observe that $X(t) - X(s) = \frac{1}{a}\left(B(a^2 t) - B(a^2 s)\right)$ is normally distributed with expectation 0 and variance $\left(\frac{1}{a^2}\right)(a^2 t - a^2 s) = t - s$. □

Remark 7.1.14. Scaling invariance has many useful consequences. For instance, let $a < 0 < b$, and look at $T(a,b) = \inf\{t \geq 0 : B(t) = a \text{ or } B(t) = b\}$, the first exit time of a one-dimensional standard Brownian motion from the interval $[a,b]$. Then, with $X(t) = a^1 B(a^2 t)$, we have

$$\mathbb{E}[T(a,b)] = a^2 \mathbb{E}\left[\inf_{t \geq 0} : X(t) = 1 \text{ or } X(t) = \frac{b}{a}\right] = a^2 \mathbb{E}[T(b/a, 1)],$$

which implies that $\mathbb{E}[T(-b,b)]$ is a constant multiple of b^2. Also,

$$\mathbb{P}\{B(t) : t \geq 0 \text{ exits } [a,b] \text{ at } a\} = \mathbb{P}\{X(t) : t \geq 0 \text{ exits } [1, b/a] \text{ at } 1\}$$

is only a function of the ratio b/a.

The scaling invariance property will be used extensively in all the following sections, and we shall often use the phrase that a fact holds 'by Brownian scaling' to indicate this.

A further useful invariance property of Brownian motion, invariance under time inversion, can be easily identified. Similar to scaling invariance, this transformation on the space of functions alters individual Brownian random functions without changing their distribution.

Theorem 7.1.15 (Time inversion of BM)**.** *Suppose* $(B(t) : t \geq 0)$ *is a standard Brownian motion. Then the process* $(X(t) : t \geq 0)$ *defined by*

$$X(t) = \begin{cases} 0 & \text{if } t = 0 \\ tB(1/t) & \text{if } t > 0 \end{cases}$$

is also a standard Brownian motion.

PROOF. Recall that the finite-dimensional marginals $(B(t_1),\dots,B(t_n))$ of Brownian motion are Gaussian random vectors (Exc. 7.1.11) and are therefore characterized by $\mathbb{E}[B(t_i)] = 0$ and $\mathrm{Cov}(B(t_i),B(t_j)) = t_i$ for $0 \le t_i \le t_j$.

Obviously, $(X(t) : t \ge 0)$ is also a Gaussian process, and the Gaussian random vectors $(X(t_1),\dots,X(t_n))$ have expectation 0. The covariances, for $t > 0$ and $h \ge 0$, are given by

$$\begin{aligned}\mathrm{Cov}(X(t+h),X(t)) &= \mathrm{Cov}\left(tB\left(\frac{1}{t+h}\right), tB\left(\frac{1}{t}\right)\right)\\ &= t(t+h)\mathrm{Cov}\left(B\left(\frac{1}{t+h}\right), B\left(\frac{1}{t}\right)\right)\\ &= t(t+h)\cdot 1\\ &= t.\end{aligned}$$

Hence, the law of all the finite-dimensional marginals $(X(t_1),X(t_2),\dots,X(t_n))$, for $0 \le t_1 \le \dots \le t_n$, are the same as for Brownian motion. The paths of $t \mapsto X(t)$ are clearly continuous for all $t > 0$. At $t = 0$, we use the following two facts: First, the distribution of X on the rationals $\mathbb{Q}$ is the same as for a Brownian motion, hence

$$\lim_{t\searrow 0} X(t) = 0 \quad \text{almost surely.}$$

And second, X is almost surely continuous on $(0,\infty)$, so that

$$0 = \lim_{t\searrow 0:\, t\in\mathbb{Q}} X(t) = \lim_{t\searrow 0} X(t) \quad \text{almost surely.}$$

Hence, $(X(t) : t \ge 0)$ has almost surely continuous paths and is a Brownian motion. □

Remark 7.1.16. The symmetry inherent in the time inversion property becomes more apparent when considering the Ornstein-Uhlenbeck diffusion $(X(t) : t \in \mathbb{R})$, which is given by

$$X(t) = e^{-t}B(e^{2t}) \quad \text{for all } t \in \mathbb{R}.$$

This is a Markov process (which will be properly explained later) such that $X(t)$ is standard normally distributed for all t. It is a diffusion with a drift towards the origin proportional to the distance from the origin. Unlike Brownian motion, the Ornstein-Uhlenbeck diffusion is time reversible: The time inversion formula shows that $(X(t) : t \ge 0)$ and $(X(-t) : t \ge 0)$ have the same distribution. For t near $-\infty$, $X(t)$ relates to the Brownian motion near time 0, and for t near ∞, $X(t)$ relates to the Brownian motion near ∞.

Time inversion is a useful tool to relate the properties of Brownian motion in a neighborhood of time $t = 0$ to properties at infinity. To illustrate the use of time inversion, we exploit Theorem 1.9 to derive an interesting statement about the long-term behavior from a trivial statement at the origin.

Corollary 7.1.17 (Law of large numbers)**.** *Almost surely,* $\lim_{t\to\infty} B(t)/t = 0$.

PROOF. Let $(X(t) : t \ge 0)$ be as defined in Theorem 7.1.14. Using this theorem, we see that

$$\lim_{t\to\infty} B(t)/t = \lim_{t\to\infty} X(1/t) = X(0) = 0.$$

□

7.1.4. Modulus of continuity and nowhere differentiability of Brownian motion. The definition of Brownian motion already requires that the sample functions are continuous almost surely. This implies that on the interval $[0,1]$ (or any other compact interval), the sample functions are uniformly continuous. That is, there exists some (random) function φ with $\lim_{h\searrow 0}\varphi(h) = 0$, called a *modulus of continuity* of the function $B : [0,1] \to \mathbb{R}$, such that

$$\limsup_{h\searrow 0} \sup_{0\le t\le 1-h} \frac{|B(t+h)-B(t)|}{\varphi(h)} \le 1.$$

Can we achieve such a bound with a deterministic function φ, i.e., is there a nonrandom modulus of continuity for the Brownian motion? The answer is yes, as the following theorem shows.

Theorem 7.1.18. *Let $B \sim BM(0,0,1)$ be SBM. There exists a constant $C > 0$ such that, almost surely, for every sufficiently small $h > 0$ and all $0 \le t \le 1-h$,*

$$|B(t+h) - B(t)| \le C\sqrt{h \log(1/h)}.$$

PROOF. This follows quite elegantly from Lévy's construction of Brownian motion. Recall the notation introduced there and that we have represented Brownian motion as a series

$$B(t) = \sum_{n=0}^{\infty} F_n(t),$$

where each F_n is a piecewise linear function (see (103)). Hence for each t, $t+h \in [0,1]$, using the mean-value theorem, for any $l > N$, we can write

$$|B(t+h) - B(t)| \le \sum_{n=0}^{\infty} |F_n(t+h) - F_n(t)| \le \sum_{n=0}^{l} h\|F_n'\|_\infty + \sum_{n=l+1}^{\infty} 2\|F_n\|_\infty. \tag{105}$$

We will find that the RHS above is bounded above by $C\sqrt{h\log(1/h)}$ almost surely whenever h is sufficiently small.

The derivative of F_n exists almost everywhere, and by definition,

$$\|F_n'\|_\infty \le \sup_{x \in [0,1] \setminus \mathscr{D}_n} \frac{F_n(x+2^{-n}) - F_n(x)}{2^{-n}} \le \frac{2\|F_n\|_\infty}{2^{-n}} = 2^{n+1}\|F_n\|_\infty.$$

By (104), for any $c > \sqrt{2\log 2}$, there exists an almost surely finite $N \in \mathbb{N}$ such that, for all $n > N$,

$$\|F_n'\|_\infty \le 2^{n+1}\|F_n\|_\infty \le 2c\sqrt{n}\,2^{n/2}.$$

Hence, using (104) again, from (105) we deduce

$$|B(t+h) - B(t)| \le h\sum_{n=0}^{N} \|F_n'\|_\infty + 2ch\sum_{n=N}^{l} \sqrt{n}2^{n/2} + 2c\sum_{n=N}^{l} \sqrt{n}2^{-n/2}.$$

We now suppose that h is (again random and) small enough that the first summand is smaller than $\sqrt{h\log(1/h)}$, and for such h, choose l so that $2^{-l} < h \le 2^{-l+1}$. By choosing h sufficiently small, we can ensure such l exceeds N. For this choice of l, the second and the third summands are also bounded by a constant multiple of $\sqrt{h\log(1/h)}$ as both sums are dominated by their largest element. Hence we get the desired inequality with a deterministic function $\varphi(h) = C\sqrt{h\log(1/h)}$. □

This upper bound is pretty close to the optimal result. The following lower bound confirms that the only missing bit is the precise value of the constant.

Theorem 7.1.19. *Let $B \sim BM(0,0,1)$ be a SBM. For every constant $c < \sqrt{2}$, for each $\varepsilon > 0$,*

$$\mathbb{P}\left(\text{there exists } h \in (0,\varepsilon) \text{ and } t \in [0,1-h] \text{ s.t. } |B(t+h) - B(t)| > c\sqrt{h\log(1/h)}\right) = 1.$$

PROOF. Let $c < \sqrt{2}$ and define, for integers $k, n \ge 0$, the events

$$A_{k,n} = \left\{B\left((k+1)e^{-n}\right) - B\left(ke^{-n}\right) > c\sqrt{n}e^{-n/2}\right\}.$$

Denoting $h = e^{-n}$ and $x = ke^n$, we can rewrite the above event as

$$A_{k,n} = \left\{B(x+h) - B(x) > c\sqrt{h\log(1/h)}\right\}.$$

Thus, for each $\varepsilon > 0$ and $n \geq 0$,

$$\mathbb{P}\left(\text{for all } h \in (0,\varepsilon) \text{ and } t \in [0,1-h],\, |B(t+h)-B(t)| \leq c\sqrt{h\log(1/h)}\right)$$
$$\leq \mathbb{P}\left(\bigcap_{k=0}^{\lfloor e^n-1\rfloor} A_{k,n}^c\right) \leq \left(1-\mathbb{P}(A_{0,n})\right)^{e^n} \leq \exp(-e^n\mathbb{P}(A_{0,n})),$$

where the second inequality follows since the increments defining $A_{k,n}$ are i.i.d. for $k \geq 0$ and the last inequality uses $1-x \leq e^{-x}$ for all x. Thus it remains to show $\mathbb{P}(A_{0,n}) \to \infty$ as $n \to \infty$. This follows easily from scaling invaraince (Lem. 7.1.12) and Gaussian tail bound (Exc. 7.1.9):

$$\mathbb{P}(A_{k,n}) = \mathbb{P}\{B(e^{-n}) > c\sqrt{n}e^{-n/2}\} = \mathbb{P}\{B(1) > c\sqrt{n}\} \geq \frac{c\sqrt{n}}{\sqrt{c^2n+1}} e^{-\frac{c^2n}{2}}.$$

By our assumption on c, we have $e^n\mathbb{P}(A_{k,n}) \to \infty$ as $n \to \infty$, as desired. □

One can determine the constant c in the best possible modulus of continuity $\phi(h) = cph\log(1/h)$ precisely. Indeed, our proof of the lower bound yields a value of $c = \sqrt{2}$, which turns out to be optimal. This striking result is due to Paul Lévy.

Theorem 7.1.20 (Lévy's Modulus of Continuity (1937))**.** *Let $B \sim BM(0,0,1)$ be a SBM. Almost surely,*

$$\limsup_{h\to 0} \sup_{0\leq t\leq 1-h} \frac{|B(t+h)-B(t)|}{\sqrt{2h\log(1/h)}} = 1.$$

PROOF. Omitted. □

We now turn our attention to nowhere differentiability of BM.

Proposition 7.1.21 (Non-monotonicity of BM)**.** *Almost surely, for all $0 < a < b < \infty$, Brownian motion is not monotone on the interval $[a,b]$.*

PROOF. First fix an interval $[a,b]$. If $[a,b]$ is an interval of monotonicity, i.e., if $B(s) \leq B(t)$ for all $a \leq s \leq t \leq b$, then we pick numbers $a = a_1 \leq \ldots \leq a_{n+1} = b$ and divide $[a,b]$ into n sub-intervals $[a_i, a_{i+1}]$. Each increment $B(a_i) - B(a_{i+1})$ has to have the same sign. As the increments are independent, this has probability $2\cdot 2^{-n}$, and taking $n \to \infty$ shows that the probability that $[a,b]$ is an interval of monotonicity must be 0. Taking a countable union gives that there is no interval of monotonicity with rational endpoints, but each monotone interval would have a nontrivial monotone rational sub-interval. □

We utilize the time-inversion technique to explore the differentiability of Brownian motion, connecting differentiability at $t = 0$ to a long-term property. This idea complements the law of large numbers: while Corollary 7.1.16 shows that Brownian motion grows slower than linearly, the subsequent proposition demonstrates that the maximum growth of $B(t)$ surpasses $\sqrt{t}$.

Proposition 7.1.22 ($\sqrt{n}$-fluctuation of BM)**.** *Almost surely,*

$$\limsup_{n\to\infty} \frac{B(n)}{\sqrt{n}} = +\infty, \quad \text{and} \quad \liminf_{n\to\infty} \frac{B(n)}{\sqrt{n}} = -\infty$$

.

PROOF. By Fatou's lemma and scaling invariance,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{B(n)}{\sqrt{n}} > c\right) \geq \limsup_{n\to\infty} \mathbb{P}\left(\frac{B(n)}{\sqrt{n}} > c\right) = P\left(B(1) > c\right) > 0.$$

Let $X_n := B(n) - B(n-1)$, and note that

$$\left\{\limsup_{n\to\infty} \frac{B(n)}{\sqrt{n}} > c\right\} = \{B(n) > c\sqrt{n} \text{ infinitely often}\} = \left\{\sum_{i=1}^{n} X_i > c\sqrt{n} \text{ infinitely often}\right\}$$

is an exchangeable[1] event for the sequence of i.i.d. RVs $X_1, X_2, \ldots$. Hence by Hewitt-Savage 0-1 law (Exc. 3.7.10), the above event has probability 0 or 1. Since it must be positive by the first display, it must occur with probability one. Taking the intersection over all positive integers c gives the first part of the statement and the second part is proved analogously. □

For a function f, we define the *upper right derivatives* and *lower right derivatives* as

$$D^* f(t) := \limsup_{h \searrow 0} \frac{f(t+h) - f(t)}{h} \quad \text{and} \quad D_* f(t) := \liminf_{h \searrow 0} \frac{f(t+h) - f(t)}{h},$$

respectively.

We now show that for any fixed time t, almost surely, Brownian motion is not differentiable at t. For this, we use Proposition 1.23 and the invariance under time inversion.

Theorem 7.1.23 (Non-differentiability of BM at a fixed point)**.** *Fix $t \geq 0$. Then, almost surely, Brownian motion is not differentiable at t. Moreover, $D^* B(t) = +\infty$ and $D^* B(t) = -\infty$.*

PROOF. By translation invariance of BM, it suffices to show that BM is not differentiable at 0[2]. Given a standard Brownian motion B, we construct a further Brownian motion X by time inversion as in Theorem 7.1.14. Then by Prop. 7.1.21,

$$D^* X(0) = \limsup_{n\to\infty} \frac{X(1/n) - X(0)}{1/n} \geq \limsup_{n\to\infty} \frac{X(1/n)}{\sqrt{n}} = \limsup_{n\to\infty} \frac{B(n)}{\sqrt{n}} = \infty.$$

Similarly, $D^* X(0) = -\infty$, showing that X is not differentiable at 0. Since X is an SBM, it is enough to conclude. □

The previous proof demonstrates that for every t, it's almost certain that it's a point of nondifferentiability for Brownian motion. However, this doesn't necessarily imply that almost every t is a point of nondifferentiability for Brownian motion! The arrangement of quantifiers, "for all t" and "almost surely," in results like Theorem 7.1.22 is crucial. In this case, the assertion holds for all Brownian paths except for those within a set of probability zero, which might vary depending on t. The accumulation of all such sets of probability zero may not, in itself, be a set of probability zero. To illustrate this point, consider the following example.

Example 7.1.24. The argument in the proof of Theorem 7.1.22 also shows that the Brownian motion X crosses 0 for arbitrarily small values $s > 0$. Defining the level sets $Z(t) = \{s > 0 : X(s) = X(t)\}$, this shows that every t is almost surely an accumulation point from the right for $Z(t)$. But not every point $t \in [0,1]$ is an accumulation point from the right for $Z(t)$. For example, the last zero of $\{X(t) : t \geq 0\}$ before time 1 is, by definition, never an accumulation point from the right for $Z(t) = Z(0)$. This example illustrates that there can be random exceptional times at which Brownian motion exhibits atypical behavior. These times are so rare that any fixed (i.e., nonrandom) time is almost surely not of this kind. ▲

Brownian motion is by definition (and construction) almost surely continuous. But the following classical result due to Paley, Winer, and Zygmund shows that it is nowhere differentiable!

Theorem 7.1.25 (Paley, Wiener and Zygmund 1933)**.** *Almost surely, Brownian motion is nowhere differentiable. Furthermore, almost surely, for all t, either $D^* B(t) = +\infty$ or $D^* B(t) = -\infty$ or both.*

PROOF. Omitted. □

[1] Does not depend on permutting the first finite number of RVs, see Exc. 3.7.10.

[2] Fix $t \geq 0$ and let $Y(s) := B(t+s) - B(t)$, which defines a standard Brownian motion. Then non-differentiability of Y at zero is equivalent to non-differentiability of B at t.

7.2. Brownian motion as a Markov process

7.2.1. The Markov property and Blumenthal's 0-1 Law. Suppose that $\{X(t) : t \geq 0\}$ is a stochastic process. Intuitively, the Markov property says that if we know the process $\{X(t) : t \geq 0\}$ on the interval $[0, s]$, for the prediction of the future $\{X(t) : t \geq s\}$, this is as useful as just knowing the endpoint $X(s)$. Moreover, a process is called a (time-homogeneous) Markov process if it starts afresh at any fixed time s. Slightly more precisely, this means that, supposing the process can be started in any point $X(0) = x \in \mathbb{R}^d$, the time-shifted process $\{X(s+t) : t \geq 0\}$ has the same distribution as the process started in $X(s) \in \mathbb{R}^d$. We shall formalize the notion of a Markov process later in this chapter, but start by giving a straightforward formulation of the facts for a Brownian motion.

Definition 7.2.1 (Independence between stochastic processes)**.** Two stochastic processes $\{X(t) : t \geq 0\}$ and $\{Y(t) : t \geq 0\}$ are called *independent* if for any sets $t_1, \ldots, t_n \geq 0$ and $s_1, \ldots, s_m \geq 0$ of times, the vectors $(X(t_1), \ldots, X(t_n))$ and $(Y(s_1), \ldots, Y(s_m))$ are independent.

Theorem 7.2.2 (Markov property of BM)**.** *Suppose that $\{B(t) : t \geq 0\}$ is a Brownian motion in $\mathbb{R}^d$ started at $x \in \mathbb{R}^d$. Let $s > 0$, then the process $\{B(t+s) - B(s) : t \geq 0\}$ is again a Brownian motion started at the origin and it is independent of the process $\{B(t) : 0 \leq t \leq s\}$.*

Proof. It is trivial to check that $\{B(t+s) - B(s) : t \geq 0\}$ satisfies the definition of a d-dimensional Brownian motion. The independence statement follows directly from the independence of the increments of a Brownian motion.

Definition 7.2.3 (Filteration)**.** A *filtration* on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a family $(\mathcal{F}(t) : t \geq 0)$ of σ-algebras such that $\mathcal{F}(s) \subseteq \mathcal{F}(t) \subseteq \mathcal{F}$ for all $s < t$. A probability space together with a filtration is sometimes called a *filtered probability space*. A stochastic process $\{X(t) : t \geq 0\}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is called *adapted* if $X(t)$ is $\mathcal{F}(t)$-measurable for any $t \geq 0$.

Suppose we have a Brownian motion $\{B(t) : t \geq 0\}$ defined on some probability space, then we can define a filtration $(\mathcal{F}^0(t) : t \geq 0)$ by letting $\mathcal{F}^0(t)$ be the σ-algebra generated by the random variables $\{B(s) : 0 \leq s \leq t\}$. With this definition, the Brownian motion is obviously adapted to the filtration. Intuitively, this σ-algebra contains all the information available from observing the process up to time t.

By Theorem 2.3, the process $\{B(t+s) - B(s) : t \geq 0\}$ is independent of $\mathcal{F}_0(s)$. In a first step, we improve this and allow a slightly larger (augmented) σ-algebra $\mathcal{F}^+(s)$ defined by

$$\mathcal{F}^+(s) := \bigcap_{t>s} \mathcal{F}^0(t).$$

Clearly, the family $(\mathcal{F}^+(t) : t \geq 0)$ is again a filtration and $\mathcal{F}^+(s) \supseteq \mathcal{F}_0(s)$, but intuitively $\mathcal{F}^+(s)$ is a bit larger than $\mathcal{F}_0(s)$ since it allows an additional 'infinitesimal glance into the future'. However, it turns out that $\mathcal{F}_0(s)$ still does not contain any information about the future.

Theorem 7.2.4 (Strict Markov property)**.** *For every $s \geq 0$, the process $\{B(t+s) - B(s) : t \geq 0\}$ is independent of $\mathcal{F}^+(s)$.*

PROOF. By continuity, $B(t+s) - B(s) = \lim_{n\to\infty} B(s_n + t) - B(s_n)$ for a strictly decreasing sequence $\{s_n : n \in \mathbb{N}\}$ converging to s. By Theorem 7.2.2, since $s_n > s$,

$$(B(t_1 + s_n) - B(s_n), \ldots, B(t_m + s_n) - B(s_n)) \perp \mathcal{F}^+(s) \quad \text{for all } n \geq 1.$$

By continuouty of BM, the limit of the vector above as $n \to \infty$ exists a.s. and equals $(B(t_1+s)-B(s), \ldots, B(t_m+s)-B(s))$. It follows that this limiting vector is independent of $\mathcal{F}^+(s)$ (consider multivariate CDF of the limit). Hence the process $\{B(t+s) - B(s) : t \geq 0\}$ is also independent of $\mathcal{F}^+(s)$. □

We now look at the *germ σ-algebra* $\mathcal{F}^+(0)$, which heuristically comprises all events defined in terms of Brownian motion on an infinitesimally small interval to the right of the origin. The next result, known as Blumenthal's 0-1 law, states that any event about the behavior of BM in an infintesimally small window must either occur for sure or not.

Theorem 7.2.5 (Blumenthal's 0-1 law)**.** *Let $x \in \mathbb{R}^d$ and $A \in \mathscr{F}^+(0)$. Then $\mathbb{P}_x(A) \in \{0,1\}$.*

PROOF. Using Theorem 7.2.4 for $s = 0$, we see that any $A \in \sigma(B(t) : t \geq 0)$ is independent of $\mathscr{F}^+(0)$. This applies in particular to $A \in \mathscr{F}^+(0)$, which therefore is independent of itself. Therefore $\mathbb{P}_x(A) = \mathbb{P}_x(A \cap A) = \mathbb{P}_x(A)\mathbb{P}_x(A)$, which yields that $\mathbb{P}_x(A)$ is zero or one. □

As a first application, we show that a standard linear Brownian motion has positive and negative values and zeros in every small interval to the right of 0. We have studied this remarkable property of Brownian motion already by different means, in the discussion following Theorem 7.1.22.

Theorem 7.2.6. *Suppose $\{B(t) : t \geq 0\}$ is a linear Brownian motion. Define $\tau = \inf\{t > 0 : B(t) > 0\}$ and $\sigma = \inf\{t > 0 : B(t) = 0\}$. Then*

$$\mathbb{P}_0\{\tau = 0\} = \mathbb{P}_0\{\sigma = 0\} = 1.$$

PROOF. The event

$$\{\tau = 0\} = \bigcap_{n=1}^{\infty} \left\{ \text{there is } 0 < \varepsilon < \frac{1}{n} \text{ such that } B(\varepsilon) > 0. \right\}$$

is clearly in $\mathscr{F}^+(0)$. Hence by Blumenthal's 0-1 law, we just have to show that this event has positive probability. This follows, as for $t > 0$, $\mathbb{P}_0\{\tau \leq t\} \geq \mathbb{P}_0\{B(t) > 0\} = 1/2$. Hence $\mathbb{P}_0\{\tau = 0\} \geq 1/2$, and we have shown the first part. The same argument works replacing $B(t) > 0$ by $B(t) < 0$, and from these two facts, $\mathbb{P}_0\{\sigma = 0\} = 1$ follows, using the intermediate value property of continuous functions. □

A further application is a 0-1 law for the *tail σ-algebra* of a Brownian motion. Define $\mathscr{G}(t) = \bigcap_{n=1}^{\infty} \sigma\{B(s) : s \geq t\}$. Let $\mathscr{T} = \lim_{t\to\infty} \mathscr{G}(t)$ be the σ-algebra of all *tail events*.

Theorem 7.2.7 (Kolmogorov's 0-1 Law)**.** *Let $x \in \mathbb{R}^d$ and $A \in \mathscr{T}$. Then $\mathbb{P}_x(A) \in \{0,1\}$.*

PROOF. It suffices to look at the case $x = 0$. Under the time inversion of Brownian motion, the tail σ-algebra is mapped onto the germ σ-algebra, which is trivial by Blumenthal's 0-1 law. □

As a final example of this section, we now exploit the Markov property to show that the set of local extrema of a linear Brownian motion is a countable, dense subset of $[0,\infty)$. We shall use the easy fact, proved in Exercise 7.2.9, that every local maximum of Brownian motion is a strict local maximum.

Proposition 7.2.8. *The set M of times where a linear Brownian motion assumes its local maxima is almost surely countable and dense.*

PROOF. Consider the function from the set of non-degenerate closed intervals with rational endpoints to $\mathbb{R}$ given by

$$[a,b] \mapsto \inf\left\{ t \geq a : B(t) = \max_{a\leq s\leq b} B(s) \right\}.$$

If $x \in M$ is a local maximum, then by Exercise 7.2.9 it is a strict local maximum almost surely, so there exists rationals $a < b$ such that $a < x < b$ and $B(t) = \max_{a\leq s\leq b} B(s)$[3]. It follows that M is contained almost surely in the image of the map defined above. Since the domain is countable, the image is also countable, so M is countable almost surely.

To show that M is dense, recall that Brownian motion has no interval of increase or decrease almost surely (Prop. 7.1.20). It follows that it almost surely has a local maximum in every non-degenerate interval. □

Exercise 7.2.9. Show that every local maximum of a one-dimensional Brownian motion is a strict local maximum. (*Hint*: In order to have a 'flat' local maximum in a small interval, one needs to have a Gaussian increment to take zero, which occurs with probability zero.)

[3]The max is achieved a.s. over the compact set $[a,b]$ since B is a.s. continuous.

7.2.2. The strong Markov property and the reflection principle. As we discussed for discrete-time Markov chains, Markov property of BM can be extended to the strong Markov property so that a BM can be restarted afresh at stopping times. Below we recall the definition of stopping time, this time for continuous processes.

Definition 7.2.10 (Stopping time)**.** A random variable T with values in $[0,\infty]$, defined on a probability space with filtration $(\mathscr{F}(t) : t \geq 0)$, is called a *stopping time* if $\{T < t\} \in \mathscr{F}(t)$, for every $t \geq 0$. It is called a *strict stopping time* if $\{T \leq t\} \in \mathscr{F}(t)$, for every $t \geq 0$.

It is easy to see that every strict stopping time is also a stopping time. This follows from

$$\{T < t\} = \bigcup_{n=1}^{\infty} \{T \leq t - n^{-1}\} \in \mathscr{F}(t).$$

For certain nice filtrations, strict stopping times and stopping times agree. In order to come into this situation, we are going to work with the 'germ filtration' $(\mathscr{F}^+(t) : t \geq 0)$ in the case of Brownian motion and refer to the notions of stopping time, etc., always with respect to this filtration. As this filtration is larger than $(\mathscr{F}_0(t) : t \geq 0)$, our choice produces more stopping times.

The crucial property which distinguishes $\{\mathscr{F}^+(t) : t \geq 0\}$ from $\{\mathscr{F}^0(t) : t \geq 0\}$ is *right-continuity*, which means that

$$\bigcap_{\varepsilon>0} \mathscr{F}^+(t+\varepsilon) = \mathscr{F}^+(t).$$

To see this, note that

$$\bigcap_{\varepsilon>0} \mathscr{F}^+(t+\varepsilon) = \bigcap_{n=1}^{\infty} \bigcap_{k=1}^{\infty} \mathscr{F}^+(t+n^{-1}+k^{-1}) = \mathscr{F}^+(t).$$

which implies that any stopping time with respect to any right-continuous filtration is automatically a strict stopping time.

Theorem 7.2.11 (Every stopping time is strict w.r.t. a right-continuous filtration)**.** *Every stopping time T with respect to the filtration $(\mathscr{F}^+(t) : t \geq 0)$, or indeed with respect to any right-continuous filtration, is automatically a strict stopping time.*

PROOF. Suppose that T is a stopping time. Then

$$\{T \leq t\} = \bigcap_{k=1}^{\infty} \{T < t + k^{-1}\} \in \bigcap_{k=1}^{\infty} \mathscr{F}(t+k^{-1}) \subseteq \bigcap_{k=1}^{\infty} \mathscr{F}^+(t+k^{-1}) = \mathscr{F}^+(t).$$

Note that we have only used right-continuity of $\mathscr{F}^+(t)$ in the proof above (for the last identity). □

Example 7.2.12 (Discrete approximation of continuous stopping time)**.** Let T be a stopping time. For each $n \geq 1$, define a stopping time T_n by

$$T_n := \min\{m2^{-n} \mid T < m2^{-n}\}.$$

In other words, we stop at the first time of the form $k2^{-n}$ after T. It is easy to see that T_n is a stopping time. We will use it later as a discrete approximation to T. ▲

We define, for every stopping time T, the corresponding σ-algebra[4]

$$\mathscr{F}^+(T) = \{A \in \mathscr{F} : A \cap \{T < t\} \in \mathscr{F}^+(t) \text{ for all } t \geq 0\}.$$

This means that the part of A that lies in $\{T < t\}$ should be measurable with respect to the information available at time t. Heuristically, this is the collection of events that happened before the stopping time T. As in the proof of the last theorem, we can infer that for right-continuous filtrations like our $(\mathscr{F}^+(t) : t \geq 0)$, the event $\{T \leq t\}$ may replace $\{T < t\}$ without changing the definition.

[4]Compare with the one in Lem. 6.2.1.

Theorem 7.2.13 (Strong Markov property of BM)**.** *Let B be a standard Brownian motion in $\mathbb{R}$. For every almost surely finite stopping time T, the process $\{B(T+t)-B(T) : t \geq 0\}$ is a standard Brownian motion independent of $\mathscr{F}^+(T)$.*

PROOF. We first show our statement for the stopping times $T_n = \min\{m2^{-n} \mid T < m2^{-n}\}$, which discretely approximate T from above (see Ex. 7.2.12). Write $B_k = \{B_k(t) : t \geq 0\}$ for the Brownian motion defined by $B_k(t) = B(t+k2^{-n}) - B(k2^{-n})$, and $B^* = \{B^*(t) : t \geq 0\}$ for the process defined by $B^*(t) = B(t+T_n) - B(T_n)$. By Theorem 7.2.2, we know that B_k is an SBM, restarted at time $k2^{-n}$. However, we do not yet know if B^* is also an SBM, restarted at the stopping time T_n.

We wish to show that B^* is an SBM independent of $\mathscr{F}^+(T_n)$. Suppose that $E \in \mathscr{F}^+(T_n)$. We wish to show that B^* is an SBM independent of E. Almost sure continuity of the sample paths of B^* is clear, so we only need to verify independent and Gaussian increments. To this end, we will show that, for every event $\{B^* \in A\}$[5], we have

$$\mathbb{P}\left(\{B^* \in A\} \cap E\right) = \mathbb{P}(B \in A)\,\mathbb{P}(E). \tag{106}$$

This would be enough to conclude. To show (106), first note that

$$\begin{aligned}\mathbb{P}\left(\{B^* \in A\} \cap E\right) &= \sum_{k=0}^{\infty} \mathbb{P}\left(\{B_k \in A\} \cap E \cap \{T_n = k2^{-n}\}\right) \\ &= \sum_{k=0}^{\infty} \mathbb{P}\left(B_k \in A\right) \cdot \mathbb{P}\left(E \cap \{T_n = k2^{-n}\}\right),\end{aligned}$$

using that $\{B_k \in A\}$ is independent of $E \cap \{T_n = k2^{-n}\} \in \mathscr{F}^+(k2^{-n})$ by the (strict) Markov property (Thm. 7.2.4). Now, by the Makov property (Thm. 7.2.2), B_k and B have the same distribution as an SBM, so $\mathbb{P}(B_k \in A) = \mathbb{P}(B \in A)$. It follows that

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbb{P}\left(B_k \in A\right) \cdot \mathbb{P}\left(E \cap \{T_n = k2^{-n}\}\right) &= \mathbb{P}\left(B \in A\right) \sum_{k=0}^{\infty} \mathbb{P}\left(E \cap \{T_n = k2^{-n}\}\right) \\ &= \mathbb{P}(B \in A)\,\mathbb{P}(E).\end{aligned}$$

This shows (106), as desired.

It remains to generalize this to general stopping times T. As $T_n \searrow T$, we have that $\{B(s+T_n) - B(T_n) : s \geq 0\}$ is a Brownian motion independent of $\mathscr{F}^+(T_n) \supset \mathscr{F}^+(T)$. Hence the increments $B(s+t+T) - B(t+T) = \lim_{n\to\infty} B(s+t+T_n) - B(t+T_n)$ of the process $\{B(r+T) - B(T) : r \geq 0\}$ are independent and normally distributed with mean zero and variance s. As the process is obviously almost surely continuous, it is a Brownian motion. Moreover, all increments, $B(s+t+T) - B(t+T) = \lim_{n\to\infty} B(s+t+T_n) - B(t+T_n)$, and hence the process itself, are independent of $\mathscr{F}^+(T)$. □

There are numerous applications of strong Markov property of Brownian motion. The next result, the reflection principle, is particularly interesting. The reflection principle states that Brownian motion reflected at some stopping time T is still a Brownian motion. More formally:

Theorem 7.2.14 (Reflection principle)**.** *If T is a stopping time and $\{B(t) : t \geq 0\}$ is a standard Brownian motion in $\mathbb{R}^d$, then the process $\{B^*(t) : t \geq 0\}$, called Brownian motion reflected at T and defined by*

$$B^*(t) = B(t)\mathbf{1}_{\{t \leq T\}} + (2B(T) - B(t))\mathbf{1}_{\{t > T\}},$$

is also a standard Brownian motion in $\mathbb{R}^d$.

PROOF. If T is finite, by the strong Markov property, both

$$\{B(t+T) - B(T) : t \geq 0\} \quad \text{and} \quad \{-(B(t+T) - B(T)) : t \geq 0\}$$

are Brownian motions and independent of the beginning $\{B(t) : t \in [0,T]\}$. Hence, the concatenation (gluing together) of the beginning with the first part and the concatenation with the second part have

[5] A is a Borel subset in $\mathbb{R}^{[0,\infty)}$, describing an event for the trajectory of a continous-time process

the same distribution. The first is just $\{B(t) : t \geq 0\}$, the second is the object $\{B^*(t) : t \geq 0\}$ introduced in the statement.

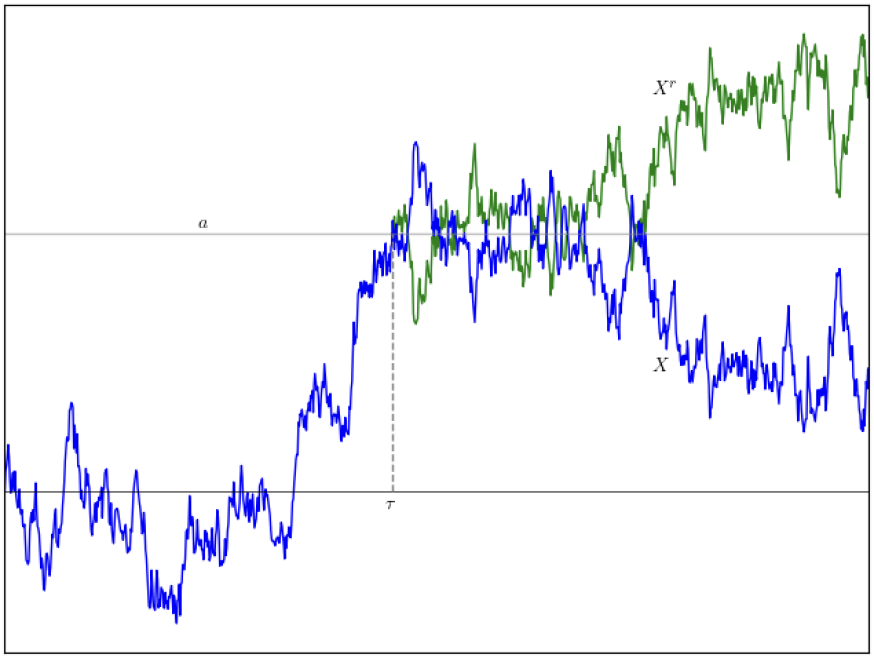


FIGURE 7.2.1. Reflection principle for 1D BM

□

Now we specialize to the case of linear Brownian motion. Let $M(t) = \max_{0\leq s\leq t} B(s)$. A priori it is not at all clear what the distribution of this random variable is, but we can determine it as a consequence of the reflection principle.

Theorem 7.2.15 (Distribution of maximum of BM)**.** *Let B be a BM in $\mathbb{R}$ and let $M(t) = \max_{0\leq s\leq t} B(s)$ denote the running maximum of B. If $a > 0$, then $\mathbb{P}\{M(t) \geq a\} = 2\mathbb{P}\{B(t) \geq a\} = \mathbb{P}\{|B(t)| \geq a\}$. That is, $M(t)$ and $|B(t)|$ have the same distribution. In particular,*

$$\mathbb{P}(M(t) \geq a) \leq \frac{\sqrt{2t}}{a\sqrt{\pi}} \exp\left(-\frac{a^2}{2t}\right).$$

PROOF. Let $T = \inf\{t \geq 0 : B(t) = a\}$ and let $\{B^*(t) : t \geq 0\}$ be Brownian motion reflected at T. Note that

$$\begin{aligned}\{M(t) \geq a\} = \{T \leq t\} &= \{T \leq t, B(t) > a\} \sqcup \{T \leq t, B(t) \leq a\} \\ &= \{T \leq t, B(t) > a\} \sqcup \{T \leq t, B^*(t) \geq a\}.\end{aligned}$$

By strong Markov property and reflection principle,

$$\begin{aligned}\mathbb{P}(M(t) \geq a) &= \mathbb{P}(T \leq t, B(t) > a) + \mathbb{P}(T \leq t, B(t) \geq a) \\ &= 2\mathbb{P}(T \leq t, B(t) > a) \\ &= 2\mathbb{P}(B(t) > a) \\ &= \mathbb{P}(|B(t)| > a),\end{aligned}$$

where the last equality uses that $B(t)$ and $B(-t)$ has the same distribution. □

7.2.3. The martingale property of Brownian motion. In the previous section we have taken a particular feature of Brownian motion, the Markov property, and introduced an abstract class of processes, the Markov processes, which share this feature. In this section we consider a different feature of Brownian motion, the martingale property.

Definition 7.2.16. A real-valued stochastic process $\{X(t) : t \geq 0\}$ is a *martingale* with respect to a filtration $\{F(t) : t \geq 0\}$ if it is adapted to the filtration, $\mathbb{E}|X(t)| < \infty$ for all $t \geq 0$, and for any pair of times $0 \leq s \leq t$,

$$\mathbb{E}[X(t)|\mathscr{F}(s)] = X(s) \quad \text{almost surely.}$$

The process is called a *submartingale* if $\geq$ holds, and a *supermartingale* if $\leq$ holds in the display above.

Example 7.2.17 (Drift-zero Brownian motion is a marginale)**.** For a drift-zero Brownian motion $\{B(t) : t \geq 0\}$, by the Markov property (Thm. 7.2.2),

$$\begin{aligned}\mathbb{E}[B(t)|\mathscr{F}^+(s)] &= \mathbb{E}[B(t) - B(s)|\mathscr{F}^+(s)] + B(s)\\ &= \mathbb{E}[B(t) - B(s)] + B(s)\\ &= B(s),\end{aligned}$$

where the last equality follows since the drift is zero. ▲

We now state two useful facts about martingales, which we will exploit extensively: The optional stopping theorem (specificlaly, Prop. 5.7.1) and L^p maximal inequality (Thm. 5.5.5). The natural extension of these results to the continuous time setting is the content of our propositions.

The optional stopping theorem provides a condition under which the defining equation for martingales can be extended from fixed times $0 \leq s \leq t$ to stopping times $0 \leq S \leq T$.

Proposition 7.2.18 (Optional stopping in continuous-time)**.** *Suppose $\{X(t) : t \geq 0\}$ is a continuous martingale, and $0 \leq S \leq T$ are stopping times. If the process $\{X(t \wedge T) : t \geq 0\}$ is dominated by an integrable random variable X, i.e., $|X(t \wedge T)| \leq X$ almost surely, for all $t \geq 0$, then*

$$\mathbb{E}[X(T)\,|\,\mathscr{F}(S)] = X(S) \quad \text{a.s.}$$

PROOF. We have proved the same result in Prop. 5.7.1 for discrete-time marginales. We will extend it to the continous-time by using discrete approximation. We will use similar strategies later to extend discrete-time martingale results to continuous-time marginales.

Fix $N \in \mathbb{N}$ and define a discrete time martingale by $X_n = X(n2^{-N} \wedge T)$ and stopping times $S' = \lceil 2^N S \rceil - 1$ and $T' = \lceil 2^N T \rceil - 1$ with respect to the filtration $(\mathscr{G}(n) : n \in \mathbb{N})$ given by $\mathscr{G}(n) = \mathscr{F}(n2^{-N})$[6]. Denote $S_N := 2^{-N}(\lceil 2^N S \rceil - 1)$. Then note that

$$\begin{aligned}X_{T'} &= X(T'2^{-N} \wedge T) = X(2^{-N}(\lceil 2^N T \rceil - 1) \wedge T) = X(T),\\ X_{S'} &= X(S'2^{-N} \wedge T) = X(2^{-N}(\lceil 2^N S \rceil - 1) \wedge T) = X(S_N \wedge T),\\ \mathscr{G}(S') &= \mathscr{F}(S'2^{-N}) = \mathscr{F}(S_N).\end{aligned}$$

Now clearly $|X_n| \leq X$ almost surely for all $n \geq 1$ by the hypothesis, so the discrete-time result (Prop. 5.7.1) gives $\mathbb{E}[X_{T'}|\mathscr{G}(S')] = X_{S'}$, which is equivalent to $\mathbb{E}[X(T)|\mathscr{F}(S_N)] = X(T \wedge S_N)$ from the above relations. Note that $S_N \nearrow S$ as $N \to \infty$. Hence letting $N \to \infty$ and using dominated convergence for conditional expectations (Thm. 5.6.15) gives the result. □

Theorem 7.2.19 (L^p maximal inequality in continuous time)**.** *Suppose $\{X(t) : t \geq 0\}$ is a continuous submartingale and $p > 1$. Then, for any $t \geq 0$,*

$$\mathbb{E}\left[\left(\sup_{0\leq s\leq t} X(s)^+\right)^p\right] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}\left[(X(t)^+)^p\right].$$

[6] Note that $\{T' \leq m\} = \{\lceil 2^N T \rceil \leq m+1\} = \{2^N T < m\} = \{T < m2^{-N}\} \in \mathscr{G}(m)$.

PROOF. Again, this is proved for martingales in discrete time in Theorem 5.5.5, and can be extended by approximation. Fix $N \in \mathbb{N}$ and define a discrete time martingale by $X_n = X(tn2^{-N})$ with respect to the filtration $(\mathscr{G}(n) : n \in \mathbb{N})$ given by $\mathscr{G}(n) = \mathscr{F}(tn2^{-N})$. By the L^p maximal inequality,

$$\mathbb{E}\left[\left(\sup_{1\le k\le 2N} X_k^+\right)^p\right] \le \left(\frac{p}{p-1}\right)\mathbb{E}\left[(X_{2N}^+)^p\right] = \left(\frac{p}{p-1}\right)^p \mathbb{E}\left[(X(t)^+)^p\right].$$

Note that the LHS is non-decreasing in N and $X_k \to X_t$ as $N \to \infty$. Hence letting $N \to \infty$ and using monotone convergence gives the claim. □

We now use the martingale property and the optional stopping theorem to prove Wald's lemmas for Brownian motion. These results identify the first and second moments of the value of Brownian motion at well-behaved stopping times.

Lemma 7.2.20 (Wald's lemma for Brownian motion)**.** *Let $\{B(t) : t \ge 0\}$ be a standard linear Brownian motion, and T be a stopping time such that either*

(i) $\mathbb{E}[T] < \infty$, *or*

(ii) $\{B(t \wedge T) : t \ge 0\}$ *is L^1-bounded.*

Then we have $\mathbb{E}[B(T)] = 0$.

PROOF. The assertion holds under condition **(ii)** by the optional stopping theorem (Prop. 7.2.18). Hence it is enough to show that condition **(i)** implies condition **(iii)**. Suppose $\mathbb{E}[T] < \infty$, and define

$$M_k := \max_{0\le t\le 1} |B(t+k) - B(k)|, \quad \text{and} \quad M := \sum_{k=1}^{\lceil T\rceil} M_k.$$

Note that $|B(t \wedge T)| \le M$, so that condition **(ii)** holds if $\mathbb{E}[M] < \infty$.

Note that M_k is independent from $\mathscr{F}(k)$ and they are i.i.d.. Since $\mathbf{1}(T \ge k) \in \mathscr{F}(k)$ and $\mathbb{E}[T] < \infty$,

$$\begin{aligned}\mathbb{E}[M] = \mathbb{E}\left[\sum_{k=1}^{\lceil T\rceil} M_k\right] = \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{E}[\mathbf{1}\{T \ge k\}M_k]\right] = \sum_{k=1}^{\infty} \mathbb{E}[\mathbf{1}\{T \ge k\}\mathbb{E}[M_k]] = \sum_{k=1}^{\infty} \mathbb{E}[\mathbf{1}\{T \ge k\}]\mathbb{E}[M_k] \\ = \mathbb{E}[M_0]\mathbb{E}[T] < \infty.\end{aligned}$$

Then by tail sum formula (Prop. 1.5.10) and reflection principle (Thm. 7.2.15),

$$\begin{aligned}\mathbb{E}[M_0] &= \int_0^\infty \mathbb{P}\left(\max_{0\le t\le 1} |B(t)| > x\right) dx \\ &\le 1 + \int_1^\infty \mathbb{P}\left(\max_{0\le t\le 1} |B(t)| > x\right) dx \\ &\le 1 + \le \int_1^\infty \sqrt{2/\pi}\, x^{-1} \exp(-x^2/2)\, dx < \infty.\end{aligned}$$

It follows that $\mathbb{E}[M] < \infty$, as desired. □

Corollary 7.2.21 (Orthogonal decomposition of variances)**.** *Let $S \le T$ be stopping times and $\mathbb{E}[T] < \infty$. Then*

$$\mathbb{E}[(B(T))^2] = \mathbb{E}[(B(S))^2] + \mathbb{E}[(B(T) - B(S))^2].$$

PROOF. The tower property of conditional expectation gives

$$\mathbb{E}[(B(T))^2] = \mathbb{E}[(B(S))^2] + 2\mathbb{E}[B(S)\mathbb{E}[B(T) - B(S) \,|\, \mathscr{F}(S)]] + \mathbb{E}[(B(T) - B(S))^2].$$

Note that $\mathbb{E}[T] < \infty$ implies $\mathbb{E}[T - S \,|\, \mathscr{F}(S)] < \infty$ almost surely. Hence the strong Markov property at time S together with Wald's lemma (Lem. 7.2.20) imply $\mathbb{E}[B(T) - B(S) \,|\, \mathscr{F}(S)] = 0$ almost surely, so that the middle term vanishes. □

Lemma 7.2.22. *Suppose $\{B(t) : t \ge 0\}$ is a linear Brownian motion. Then the process $\{B(t)^2 - t : t \ge 0\}$ is a martingale.*

PROOF. The process is adapted to the natural filtration of Brownian motion, and

$$\begin{aligned}\mathbb{E}[B(t)^2 - t \,|\, \mathscr{F}^+(s)] &= \mathbb{E}\left[(B(t) - B(s))^2 \,|\, \mathscr{F}^+(s)\right] + 2\mathbb{E}\left[B(t)B(s) \,|\, \mathscr{F}^+(s)\right] - B(s)^2 - t \\ &= (t-s) + 2B(s)^2 - B(s)^2 - t \\ &= B(s)^2 - s,\end{aligned}$$

which completes the proof. □

Theorem 7.2.23 (Wald's second lemma)**.** *Let T be a stopping time for standard Brownian motion such that $\mathbb{E}[T] < \infty$. Then*

$$\mathbb{E}[B(T)^2] = \mathbb{E}[T].$$

PROOF. By Lemma 7.2.22, $\{B(t)^2 - t : t \geq 0\}$ is a martingale. Define stopping times

$$T_n = \inf\{t \geq 0 : |B(t)| = n\}.$$

Then $\{B(t \wedge T \wedge T_n)^2 - t \wedge T \wedge T_n : t \geq 0\}$ is a martingale dominated by the integrable random variable $n^2 + T$. By the optional stopping theorem (Prop. 7.2.18), we get $\mathbb{E}[B(T \wedge T_n)^2] = \mathbb{E}[T \wedge T_n]$. By Corollary 7.2.21, we have $\mathbb{E}[B(T)^2] \geq \mathbb{E}[B(T \wedge T_n)^2]$. Hence, by monotone convergence,

$$\mathbb{E}[B(T)^2] \geq \lim_{n\to\infty} \mathbb{E}[B(T \wedge T_n)^2] = \lim_{n\to\infty} \mathbb{E}[T \wedge T_n] = \mathbb{E}[T].$$

Conversely, using Fatou's lemma in the first step,

$$\mathbb{E}[B(T)^2] \leq \liminf_{n\to\infty} \mathbb{E}[B(T \wedge T_n)^2] = \liminf_{n\to\infty} \mathbb{E}[T \wedge T_n] \leq \mathbb{E}[T].$$

This shows the assertion. □

Bibliography

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi, *Prediction, learning, and games*, Cambridge university press, 2006.

[DD99] Richard Durrett and R Durrett, *Essentials of stochastic processes*, vol. 1, Springer, 1999.

[DMRV06] Oleksiy Dovgoshey, Olli Martio, Vladimir Ryazanov, and Matt Vuorinen, *The cantor function*, Expositiones mathematicae **24** (2006), no. 1, 1–37.

[Dur19] Rick Durrett, *Probability: theory and examples*, vol. 49, Cambridge university press, 2019.

[Ete81] Nasrollah Etemadi, *An elementary proof of the strong law of large numbers*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **55** (1981), no. 1, 119–122.

[Kac87] Mark Kac, *Enigmas of chance: an autobiography*, Univ of California Press, 1987.

[LP17] David A Levin and Yuval Peres, *Markov chains and mixing times*, vol. 107, American Mathematical Soc., 2017.

[LS19] Hanbaek Lyu and David Sivakoff, *Persistence of sums of correlated increments and clustering in cellular automata*, Stochastic Processes and their Applications **129** (2019), no. 4, 1132–1152.