Summary for Elementary Probability

SEUNGWOO HAN

Contents

Chapter 1

Basic Concepts

1.1 Events and Probability

Definition 1.1.1: Probability Space

A *probability space* contains of a triple $(\Omega, \mathcal{F}, P)$ where
- Ω is the sample space,
- $\mathcal{F} \subseteq 2^\Omega$ (each $A \in \mathcal{F}$ is called an *event*), and
- $P : \mathcal{F} \to [0,1]$ maps each event $A \in \mathcal{F}$ to the *probability* of A

which satisfies the following conditions:

Axioms Relative to the Events The family $\mathcal{F}$ of events must be a σ-field on Ω:
(1) $\Omega \in \mathcal{F}$;
(2) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (where A^c is the complement of A);
(3) If $\langle A_n \rangle_{n \in \mathbb{Z}_+}$ is a sequence on $\mathcal{F}$, then $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

Axioms Relative to the Probability The function P must satisfy the following conditions:
(1) $P(\Omega) = 1$;
(2) σ-additivity holds: if $\langle A_n \rangle_{n \in \mathbb{Z}_+}$ is a sequence of pairwise disjoint events, then

$$P\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty P(A_n).$$

Note

Here are immediate properties of probability:
- $P(A^c) = 1 - P(A)$;
- $\varnothing = \Omega^c \in \mathcal{F}$ and $P(\varnothing) = 0$;
- If $\langle A_n \rangle_{n \in \mathbb{Z}_+}$ is a sequence of events, then $\bigcap_{n=1}^\infty A_n$ is also an event;
- $A, B \in \mathcal{F}$ and $A \subseteq B$ implies $P(A) \le P(B)$.

Lemma 1.1.2 sub-σ-additivity

If $\langle A_n\rangle_{n\in\mathbb{Z}_+}$ is a sequence of events, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \le \sum_{n=1}^{\infty} P(A_n).$$

Proof. Let $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$ for each $n \ge 1$ and use σ-additivity. □

Lemma 1.1.3 Inclusion-Exclusion Principle

If $A_1, \cdots, A_n$ are events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\emptyset \neq I \subseteq [n]} (-1)^{|I|-1} P\left(\bigcap_{i\in I} A_i\right).$$

Proof. Classic. □

Theorem 1.1.4 Sequential Continuity of Probability

(1) Let $\langle B_n\rangle_{n\in\mathbb{Z}_+}$ be a sequence of events such that $B_n \subseteq B_{n+1}$ for all $n \ge 1$. Then,

$$P\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} P(B_n).$$

(2) Let $\langle C_n\rangle_{n\in\mathbb{Z}_+}$ be a sequence of events such that $C_n \supseteq C_{n+1}$ for all $n \ge 1$. Then,

$$P\left(\bigcap_{n=1}^{\infty} C_n\right) = \lim_{n\to\infty} P(C_n).$$

Proof.

(1) Let $B_n' := B_n \setminus B_{n-1}$ for each $n \ge 2$ and $B_1' := B_1$. so that $B_m = \bigcup_{n=1}^{m} B_n'$ and B_i''s are pairwise disjoint. Hence, by σ-additivity, we have

$$P\left(\bigcup_{n=1}^{\infty} B_n\right) = P\left(\bigcup_{n=1}^{\infty} B_n'\right) = \sum_{n=1}^{\infty} P(B_n') = P(B_1) + \sum_{n=1}^{\infty}\big(P(B_n) - P(B_{n-1})\big) = \lim_{n\to\infty} P(B_n).$$

(2) Let $C_n' := C_n^c$ for each $n \ge 1$ so that $C_n' \subseteq C_{n+1}'$ for all n. Hence, by (1), we have $P\left(\bigcup_{n=1}^{\infty} C_n'\right) = \lim_{n\to\infty} P(C_n')$. The result follows from the fact that $\bigcup_{n=1}^{\infty} C_n' = \Omega \setminus \bigcap_{n=1}^{\infty} C_n$. □

1.2 Random Variables and Their Distributions

Definition 1.2.1: Random Variable

A *random variable* on $(\Omega, \mathcal{F})$ is any mapping $X : \Omega \to \overline{\mathbb{R}}$ such that for all $a \in \mathbb{R}$, $\{X \le a\} \triangleq \{\omega \in \Omega \mid X(\omega) \le a\} \in \mathcal{F}$. Here, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$.

- If X only takes finite values, X is called a *real random variable*.
- If X only takes only a countable set of values $\{a_n\}_{n\in\mathbb{Z}_{\ge 0}}$, X is called a *discrete random variable*.

Definition 1.2.2: Cumulative Distribution Function

The *cumulative distribution function* (CDF) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P(X \leq x) \triangleq P(\{X \leq x\}).$$

Lemma 1.2.3

Let F be a cumulative distribution function of a random variable X.

(1) F is monotone increasing.

(2) F is right-continuous.

(3) If we define $F(\infty) := \lim_{x\to\infty} F(x)$ and $F(-\infty) = \lim_{x\to-\infty} F(x)$, then $1 - F(\infty) = P(X = \infty)$ and $F(-\infty) = P(X = -\infty)$.

Proof.

(1) Take any $x, y \in \mathbb{R}$ with $x \leq y$. Then, $\{X \leq x\} \subseteq \{X \leq y\}$. Hence, $F(x) = P(X \leq x) \leq P(X \leq y) \leq F(y)$.

(2) Take any decreasing nonnegative sequence $\langle \varepsilon_n \rangle_{n\in\mathbb{Z}_+}$ of real numbers converging to zero and a real number x. Let $C_n := \{X \leq x + \varepsilon_n\}$ so that $\langle C_n \rangle_{n\in\mathbb{Z}_+}$ is a decreasing sequence of events. Note also that $\{X \leq x\} = \bigcap_{n=1}^{\infty} C_n$ Then, by Theorem 1.1.4 (2),

$$F(x) = P(X \leq x) = \lim_{n\to\infty} P(X \leq x + \varepsilon_n) = \lim_{n\to\infty} F(x + \varepsilon_n).$$

(3) Let $B_n := \{X \leq n\}$ for each $n \in \mathbb{Z}_+$ so that $\bigcup_{n=1}^{\infty} B_n = \{X < \infty\}$ and $\langle B_n \rangle_{n\in\mathbb{Z}_+}$ is an increasing sequence of events. By Theorem 1.1.4 (1),

$$1 - P(X = \infty) = P(X < \infty) = P\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} F(n) = F(\infty).$$

The last equality is due to (1). □

Definition 1.2.4: Probability Density

If a real random variable X admits a cumulative distribution function F such that

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$$

for some nonnegative function f, then X is said to admit the *probability density* f.

Note

Note that the probability density f satisfies

$$\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y = 1.$$

1.3 Conditional Probability and Independence

Definition 1.3.1: Conditional Probability

Let B be an event with $P(B) > 0$. For any event A, we define

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

and it is called the *probability of A given B*.

Definition 1.3.2: Independent Events

(1) Two events A and B are said to be *indepenent* if $P(A \cap B) = P(A)P(B)$.

(2) Let $\mathcal{A}$ be a nonempty family of events. $\mathcal{A}$ is said to be a *family of independent events* if for any finite subfamily $\langle A_1, \cdots, A_n \rangle$ of $\mathcal{A}$,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Note

When $P(B) > 0$, A and B are indepenent if and only if $P(A \mid B) = P(A)$.

Definition 1.3.3: Independent Random Variables

Two random variables X and Y defined on $(\Omega, \mathcal{F}, P)$ are said to be *independent* if

$$\forall a, b \in \mathbb{R},\ P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq a).$$

A family $\mathcal{X}$ of random variables is said to be *independent* if, for any finite subfamily $\{X_1, \cdots, X_n\} \subseteq \mathcal{X}$, and for any $a_1, \cdots, a_n \in \mathbb{R}$, we have

$$P(X_1 \leq a_1, \cdots, X_n \leq a_n) = \prod_{i=1}^{n} P(X_i \leq a_i).$$

Note

If X and Y takes values $\langle a_n \rangle_{n \in \mathbb{Z}_+}$ and $\langle b_n \rangle_{n \in \mathbb{Z}_+}$, respectively, then X and Y are independent if and only if

$$P(X = a_i, Y = b_j) = P(X = a_i)P(Y = b_j)$$

for all $i, j \in \mathbb{Z}_+$. It is analogous to family of discrete random variables.

Lemma 1.3.4 Bayes' Retrodiction Formula

If A and B are events of positive probability, then

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}.$$

Lemma 1.3.5 Bayes' Sequential Formula

Let $A_1, \cdots, A_n$ be events such that $P(A_1 \cap \cdots \cap A_n) > 0$. Then,

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Proof. Mathematical induction. □

Lemma 1.3.6 Law of Total Probability

Let A be an event, and let $\langle B_n \rangle_{n \in \mathbb{Z}_{>0}}$ be an exaustive sequence of events. In other words, $\bigcup_{n=1}^{\infty} B_n = \Omega$ and $B_i \cap B_j = \varnothing$ for all $1 \le i < j$. Then, we have

$$P(A) = \sum_{n=1}^{\infty} P(A \mid B_n)P(B_n)$$

where we agree to have $P(A \mid B_n)P(B_n) = 0$ when $P(B_n) = 0$. Moreover, for all $m \in \mathbb{Z}_{>0}$, we have

$$P(B_m \mid A) = \frac{P(A \mid B_m)P(B_m)}{\sum_{n=1}^{\infty} P(A \mid B_n)P(B_n)}$$

if $P(A) > 0$.

Proof. $A = A \cap \Omega = A \cap \left(\bigcup_{n=1}^{\infty} B_n\right) = \bigcup_{n=1}^{\infty}(A \cap B_n)$. Apply σ-additivity to obtain the result. Note that $P(A \cap B_n) = P(A \mid B_n)P(B_n)$ always according to our convention. □

1.4 Counting and Probability

If Ω is finite and we let $p(\omega) := P(\{\omega\})$ with equal probabilities, then we must have $P(A) = (\operatorname{card} A)/(\operatorname{card} \Omega)$ for all $A \subseteq \Omega$. Hence, we should *count*.

Example 1.4.1

- The number of injections from E to F with $p = \operatorname{card}(E)$ and $n = \operatorname{card}(F)$ when $p \le n$ is $A_p^n = \frac{n!}{(n-p)!}$.
- In particular, if $p = n$, we have A_n^n, the number of permutations of n elements, which is $n!$.
- The number of subsets of F with p elements is $\binom{n}{p} = \frac{n!}{p!(n-p)!}$.
- (Binomial formula) $(x+y)^n = \sum_{p=0}^{n} x^p y^{n-p}$. $2^n = \sum_{p=0}^{n} \binom{n}{p}$.
- $\binom{n}{p} = \binom{n}{n-p}$.
- (Pascal's formula) $\binom{n}{p} = \binom{n-1}{p-1} + \binom{n-1}{p}$.

Chapter 2

Discrete Probability

2.1 Discrete Random Elements

Definition 2.1.1: Discrete Random Element

Let E be a denumerable set and let $(\Omega, \mathcal{F}, P)$ be a probability space. Any function $X : \Omega \to E$ such that

$$\forall x \in E, \{\, \omega \mid X(\omega) = x \,\} \in \mathcal{F}$$

is called a *discrete random element* of E. When $E \subseteq \mathbb{R}$, we refer to X as a *discrete random variable*. This allows us to define

$$p(x) := P(X = x)$$

for $x \in E$. The collection $\{p(x)\}_{x \in E}$ is the *distribution* of X. It satisfies

$$0 \le p(x) \le 1 \quad \text{and} \quad \sum_{x \in E} p(x) = 1.$$

Note

E being denumerable enables us to define in such way. Note the difference from Definition 1.2.1.

Example 2.1.2 Bernoulli Distribution

The coin tossing experiment of a single coin with bias p $(0 \le p \le 1)$ is described by a discrete random variable X taking its values in $E = \{0, 1\}$ with the distribution

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

This is called the *Bernoulli distribution* of parameter p.

Example 2.1.3 Binomial Distribution

Let $X_1, \cdots, X_n$ be n indepenent random variables with the Bernoulli distribution of parameter p. The distribution of a discrete random variable $S_n = \sum_{i=1}^{n} X_i$ satisfies

$$P(S_n = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

for $0 \le k \le n$. This is called the *binomial distribution* of size n and parameter p.

Example 2.1.4 Geometric Distribution

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of independent random variables with the Bernoulli distribution of parameter p. Let T be a random element such that

$$T = \begin{cases} \min\{\, n \mid X_n = 1 \,\} & \text{if } \{\, n \mid X_n = 1 \,\} \neq \varnothing \\ +\infty & \text{otherwise.} \end{cases}$$

Then, we have

$$P(T = k) = p(1-p)^{k-1}$$

for $k \geq 1$ and $P(T = \infty) = 0$ or 1 according to whether $p > 0$ or $p = 0$. We call T a *geometry random variable* of paramter p. This is symbolized by $T \sim \mathcal{G}(p)$.

Example 2.1.5 Multinomial Distribution

Suppose you have k boxes in which you place n balls at random in the following manner. The balls are thrown into the boxes independently of one another, and the probability that a given ball falls in a box i is p_i. Of course, $0 \leq p_i \leq k$ and $\sum_{i=1}^{k} p_i = 1$. Let N_i $(1 \leq i \leq k)$ denote the number of balls that fall into box i. The random vector $N = (N_1, \cdots, N_k)$ takes its values in the k-tuples of integers $(n_1, \cdots, n_k)$ satisfying

$$n_1 + \cdots + n_k = n.$$

The probability that $N_i = n_i$ for all i is given by

$$P(N_1 = n_1, \cdots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},$$

where $n_1 + \cdots + n_k = n$. This type of distribution is called the *multinomial distribution* of size (n, k) and of parameters $(p_1, \cdots, p_k)$. Notation $(N_1, \cdots, N_k) \sim \mathcal{M}(n, k, p_i)$ expresses that $(N_1, \cdots, N_k)$ is a multinomial random variable.

Example 2.1.6 Poisson Distribution

A random variable X that takes its values in $E = \mathbb{Z}_{\geq 0}$ and admits the distribution

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

for $k \geq 0$, where λ is a nonnegative real number, is called a *Poisson random variable* with parameter λ. This is denoted by $X \sim \text{Poisson}(\lambda)$.

2.2 Expectation

Definition 2.2.1: Expectation of Discrete Random Variable

Let X be a random element taking its values in E, and let $f : E \to \mathbb{R}$ be a function such that

$$\sum_{x \in E} |f(x)| p(x) < \infty. \tag{2.1}$$

One then defines the *expectation* of $f(X)$, denoted $\mathbb{E}[f(X)]$, by

$$\mathbb{E}[f(X)] := \sum_{x \in E} f(x) p(x).$$

Note

If ⟨2.1⟩ is satisfied, $\mathbb{E}[f(X)]$ is well-defined and finite. If ⟨2.1⟩ is not satisfied and f is nonnegative, then $\mathbb{E}[f(X)]$ is well-defined but can be infinite. Otherwise, $\mathbb{E}[f(X)]$ may not be well-defined.

Exercise 2.2.1

Let X be a Poisson random variable with parameter λ. We have

$$\mathbb{E}[X] = \lambda \qquad \text{and} \qquad \mathbb{E}[X^2] = \lambda^2 + \lambda.$$

Solution:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \\
\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda \sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} + \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
&= \lambda \mathbb{E}[X] + \lambda = \lambda^2 + \lambda.
\end{aligned}$$

□

Note

Definition 2.2.1 easily extends to $f : E \to \mathbb{C}$ with the same condition. Writing $f = g + ih$, ⟨2.1⟩ is equivalent to

$$\sum_{x \in E} |g(x)| p(x) < \infty \quad \text{and} \quad \sum_{x \in E} |h(x)| p(x) < \infty.$$

Note

Some properties of expectation:

- *Linearity*. $\mathbb{E}[\lambda_1 f_1(X) + \lambda_2 f_2(X)] = \lambda_1 \mathbb{E}[f_1(X)] + \lambda_2 \mathbb{E}[f_2(X)]$.
- *Monotonicity*. If $\forall x \in E$, $f_1(x) \leq f_2(x)$, then $\mathbb{E}[f_1(X)] \leq \mathbb{E}[f_2(X)]$.
- $|\mathbb{E}[f(X)]| \leq \mathbb{E}[|f(X)|]$.
- Let $C \subseteq E$ and let I_C be the *indicator function* of C defined by

$$I_C(x) := \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\mathbb{E}[I_C(X)] = \sum_{x\in E} I_C(x)p(x) = \sum_{x\in C} p(x) = \sum_{x\in C} P(X=x) = P\left(\bigcup_{x\in C}\{X=x\}\right)$.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A \in \mathcal{F}$. Defining the indicator function $I_A\colon \Omega \to \{0,1\}$ for A, I_A is clearly a discrete random variable taking values on $\{0,1\}$. We have $\mathbb{E}[I_A] = P(A)$.

Theorem 2.2.2 Markov's Inequality

Let $f: E \to \mathbb{R}$ satisfy $\langle 2.1 \rangle$. Then, for $a > 0$, we have

$$P(|f(X)| \geq a) \leq \frac{\mathbb{E}[|f(X)|]}{a}.$$

Proof. Let $C := \{x \in E \mid |f(x)| \geq a\} \subseteq E$. Then, $|f(x)| \geq |f(x)|I_C(x)$ and thus

$$\begin{aligned}\mathbb{E}[|f(X)|] &\geq \mathbb{E}[|f(X)|I_C(X)] \\ &\geq \mathbb{E}[aI_C(X)] \\ &= a\mathbb{E}[I_C(X)] = aP(|f(X)| \geq a).\end{aligned}$$

□

2.3 Independence

Definition 2.3.1: Independence of Discrete Random Elements

Let X and Y be two discrete random elements with values in the denumerable spaces E and F, respectively. Now, one can define another random element Z on $G := E \times F$ by $Z(\omega) = (X(\omega), Y(\omega))$. We say X and Y are *independent* if

$$P(X=x, Y=y) := P(Z=(x,y)) = P(X=x)P(Y=y)$$

for all $x \in E$ and $y \in F$. This can be ge

Lemma 2.3.2 Product Formula

Let X and Y be two discrete random elements with values in the denumerable spaces E and F, respectively. If $f: E \to \mathbb{R}$ and $g: F \to \mathbb{R}$ satisfy $\langle 2.1 \rangle$, and if X and Y are independent, then $\mathbb{E}[f(X)g(Y)]$ is well-defined and

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)].$$

Proof. We have

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \sum_{(x,y)\in E\times F} f(x)g(y)P(X=x, Y=y) \\ &= \sum_{x\in E} f(x)P(X=x) \sum_{y\in F} g(y)P(Y=y) \\ &= \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)].\end{aligned}$$

□

Lemma 2.3.3 Convolution Formula

Let X and Y be two discrete random elements with values in the denumerable spaces E

and F, respectively. If X and Y, the random variable $S = X + Y$ admits the distribution

$$P(S = k) = \sum_{j=0}^{k} P(X = j) \cdot P(Y = k - j)$$

for $k \geq 0$.

Proof. Note that $\{S = k\} = \biguplus_{j=0}^{k} (\{X = j\} \cap \{Y = k - j\})$. Hence,

$$P(S = k) = \sum_{j=0}^{k} P(X = j, Y = k - j) = \sum_{j=0}^{k} P(X = j) \cdot P(Y = k - j).$$ □

Note

Definition 2.3.1 and Lemma 2.3.2 can readily be generalized to finite number of discrete random elements.

Exercise 2.3.1

Let X and Y be two independent Poisson random variables with parameters λ and μ, respectively. Show that $S = X + Y \sim \text{Poisson}(\lambda + \mu)$.

Solution:

$$\begin{aligned} P(S = k) &= \sum_{j=0}^{k} P(X = j) \cdot P(Y = k - j) && \triangleright \text{ Convolution Formula} \\ &= \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} \cdot \frac{\mu^{k-j}}{(k-j)!} e^{-\mu} \\ &= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda^j \mu^{k-j} \\ &= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}. && \triangleright \text{ Binomial Formula} \end{aligned}$$

Hence, $S \sim \text{Poisson}(\lambda + \mu)$. □

2.4 Mean and Variance

Definition 2.4.1: Mean and Variance of Discrete Random Variable

If X is a discrete random variable, the quantities

$$m \triangleq \mathbb{E}[X] \quad \text{and} \quad \sigma^2 \triangleq \text{Var}[X] \triangleq \mathbb{E}[(X - m)^2]$$

are called the *mean* and *variance* of X, respectively. The quantity $\sigma \triangleq \sqrt{\sigma^2}$ is called the *standard deviation* of X.

Note

Some properties of mean and variance:

- $\text{Var}(aX) = a^2 \text{Var}(X)$.

- $\sigma^2 = 0$ implies that $p(x) = 0$ for all $x \neq m$.
- If $X_1, \cdots, X_n$ are independent discrete random variables, then $\mathrm{Var}\big(\sum_{i=1}^n X_i\big)$ equals $\sum_{i=1}^n \mathrm{Var}(X_i)$.
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Exercise 2.4.1

Show that the variance of a Poisson random variable of parameter λ is λ. Show that the mean and variance of a geometric random variable of parameter $p > 0$ is $1/p$ and $(1-p)/p^2$.

Solution: Let $X \sim \mathrm{Poisson}(\lambda)$. By Exercise 2.2.1, we have $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$.

Let $Y \sim \mathcal{G}(p)$. Then,

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} \\
&= p + \sum_{k=2}^{\infty} kp(1-p)^{k-1} \\
&= p + (1-p)\sum_{k=1}^{\infty} (k+1)p(1-p)^{k-1} \\
&= p + (1-p)\sum_{k=1}^{\infty} kp(1-p)^{k-1} + (1-p)\sum_{k=1}^{\infty} p(1-p)^{k-1} \\
&= (1-p)\mathbb{E}[Y] + 1.
\end{aligned}$$

Hence, $\mathbb{E}[Y] = 1/p$. Moreover,

$$\begin{aligned}
\mathbb{E}[Y^2] &= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} \\
&= \sum_{k=1}^{\infty} \big((k-1)^2 + 2k - 1\big)p(1-p)^{k-1} \\
&= (1-p)\sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} + 2\mathbb{E}[Y] - 1 \\
&= (1-p)\mathbb{E}[Y^2] + \frac{2}{p} - 1.
\end{aligned}$$

Hence, $\mathbb{E}[Y^2] = (2-p)/p^2$. Therefore, $\mathrm{Var}(Y) = (2-p)/p^2 - 1/p^2 = (1-p)/p^2$. $\square$

Exercise 2.4.2

Let X be a discrete random variable with values in $\mathbb{N}_0 = \{0, 1, 2, \cdots\}$. Show that

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

Solution: Note that $\{X \geq n\} = \biguplus_{k=n}^{\infty} \{X = k\}$ for $n \in \mathbb{N}_0$. Hence, by σ-additivity,

$$\begin{aligned}
\sum_{n=1}^{\infty} P(X \geq n) &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(X = k) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{k} P(X = k) && \triangleright \text{ Fubini's theorem} \\
&= \sum_{k=1}^{\infty} kP(X = k) \\
&= \sum_{k=0}^{\infty} kP(X = k) = \mathbb{E}[X].
\end{aligned}$$

□

Exercise 2.4.3

Show that the mean and variance corresponding to the binomial distribution of size n and parameter p are np and $np(1-p)$, respectively.

Solution: Let $X \sim \text{Binomial}(n, p)$. Then, $X \sim \sum_{i=1}^{n} X_i$ where X_i are independent Bernoulli random variables with parameter p. We have $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$. Hence, $\mathbb{E}[X] = np$ and $\text{Var}(X) = np(1-p)$. □

Theorem 2.4.2 Chebyshev's Inequality

Let X be a discrete random variable. Then, for any $\varepsilon > 0$, we have

$$P(|X - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Apply Markov's Inequality to X with $f(x) = (x-m)^2$ and $a = \varepsilon^2$ to get

$$\begin{aligned}
P(|X - m| \geq \varepsilon) &= P((X-m)^2 \geq \varepsilon^2) \\
&\leq \frac{\mathbb{E}[|X-m|^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.
\end{aligned}$$

□

Theorem 2.4.3 Weak Law of Large Numbers

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of discrete random variables, identically distributed with common mean m and common variance σ^2. Consider the empirical mean $S_n/n = (X_1 + \cdots + X_n)/n$. Then,

$$\lim_{n \to \infty} P\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) = 0$$

for every $\varepsilon > 0$.

Proof. We have $\text{Var}[S_n/n] = \frac{\sigma^2}{n}$. By Chebyshev's Inequality, $P\left(\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$. □

Definition 2.4.4: Convergence in Probability

A sequence of random variables $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ is said to *converge in probability* to a random variable X if if, for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0.$$

This is denoted by $X_n \xrightarrow{P} X$.

Note

There are various notions of convergence: convergence in quadratic mean, convergence in law, convergence in probability, and almost-sure convergence. The strong law of large numbers states that S_n/n converges to m almost surely.

2.5 Generating Functions

Definition 2.5.1: Generating Function

Let X be a discrete random variable taking its values in $\mathbb{Z}_{\geq 0}$. The *generating function* of X is the function g from the unit disc of $\mathbb{C}$ into $\mathbb{C}$ defined by

$$g(s) \triangleq \mathbb{E}[s^X] = \sum_{k=0}^{\infty} s^k P(X = k).$$

Note

Inside the unit disk, the power series $\sum_{k=0}^{\infty} s^k P(X = k)$ uniformly and absolutely convergent since

$$\sum_{k=1}^{\infty} P(X = k)|s|^k \leq \sum_{k=1}^{\infty} P(X = k) = 1.$$

Hence, we can add, differentiate, and integrate term-by-term.

Moreover, the generating function uniquely determines the determines the distribution. If $\sum_{k=0}^{\infty} P(X_1 = k)s^k = \sum_{k=0}^{\infty} P(X_2 = k)s^k$ in the unit disk, then the corresponding coefficients must be equal.

Exercise 2.5.1

Let $X \sim \text{Binomial}(n, p)$. Show that the generating function of X is $g(s) = (ps + 1 - p)^n$.

Solution:

$$\begin{aligned} g(s) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k \\ &= \sum_{k=0}^{n} \binom{n}{k} (ps)^k (1-p)^{n-k} \\ &= (ps + 1 - p)^n \end{aligned}$$

□

Definition 2.5.2: Multivariate Generating Function

Let $X_1, \cdots, X_k$ be k discrete random variables taking their values in $\mathbb{Z}_{\geq 0}$. The *generating function* of $(X_1, \cdots, X_k)$ is the function g from D^k into $\mathbb{C}$ defined by

$$g(s_1, \cdots, s_k) \triangleq \mathbb{E}[s_1^{X_1} \cdots s_k^{X_k}] = \sum_{i_1=0}^{\infty} \cdots \sum_{i_k=0}^{\infty} s_1^{i_1} \cdots s_k^{i_k} P(X_1 = i_1, \cdots, X_k = i_k)$$

where D is the unit disc of $\mathbb{C}$.

Note

- If g is a multivariate generating function, then $g(s_1, 1, \cdots, 1)$ is the generating function of X_1.
- If X_i's are independent, then by Product Formula, we have $\mathbb{E}[s_1^{X_1} \cdots s_k^{X_k}] = \prod_{i=1}^{k} \mathbb{E}[s_i^{X_i}]$, i.e.,

$$g(s_1, \cdots, s_k) = \prod_{i=1}^{k} g(s_i).$$

  Moreover, $\mathbb{E}[s^{X_1} \cdots s^{X_k}] = \mathbb{E}[s^{X_1 + \cdots + X_k}]$, i.e., $g(s, \cdots, s)$ is the generating function of $X_1 + \cdots + X_k$.

Note

Differentiation of Generating Functions and Moments As $g(s)$ is absolutely convergent in the unit disc, we can differentiate term-by-term to get

$$g'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1}$$

for $|s| < 1$. If $\mathbb{E}[X] = \sum_{k=0}^{\infty} k p_k$ exists, then by Abel's lemma, we get $\mathbb{E}[X] = g'(1) := \lim_{\substack{s \to 1 \\ |s|<1}} g'(s)$. Doing this once more, we have $g''(1) = \sum_{k=2}^{\infty} k(k-1) p_k s^{k-2} = \mathbb{E}[X^2] - m$. Moreover, we have $\sigma^2 = g''(1) + g'(1) - g'(1)^2$.

Exercise 2.5.2

Using generating functions, show that if X_1 and X_2 are independent Poisson random variables $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Solution: The generating function of a Poisson random variable of parameter λ is

$$g(s) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} s^k = \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{\lambda s}.$$

Letting $X \sim \text{Poisson}(\lambda_1 + \lambda_2)$, we thus have $g_{X_1 + X_2}(s) = g_X(s)$ in some neighborhood of the origin. Hence, $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$. $\square$

Theorem 2.5.3 Wald's Equality

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be an i.i.d. sequence of discrete random variables with values in $\mathbb{Z}_{\geq 0}$ and the common generating function g_X. Let T be a discrete random variable taking its values in $\mathbb{Z}_{>0}$ and the generating function g_T. Suppose moreover that T is independent from the X_n's. Let

$$Y \triangleq X_1 + \cdots + X_T$$

be a random variable. Then, $\mathbb{E}[Y] = \mathbb{E}[T] \cdot \mathbb{E}[X_1]$.

Proof. Using $1 = \sum_{n=1}^{\infty} I_{\{T=n\}}$, we have

$$g_Y(s) = \mathbb{E}[s^Y] = \mathbb{E}[s^{X_1+\cdots+X_T}] = \mathbb{E}\left[\sum_{n=1}^{\infty} I_{\{T=n\}} s^{X_1+\cdots+X_n}\right].$$

By Lebesgue's dominated convergence theorem, we can interchange the sum and the expectation to get

$$\begin{aligned} g_Y(s) &= \sum_{n=1}^{\infty} \mathbb{E}[I_{\{T=n\}} s^{X_1+\cdots+X_n}] \\ &= \sum_{n=1}^{\infty} P(T=n)\mathbb{E}[s^{X_1}]^n && \triangleright \text{ Product Formula} \\ &= \sum_{n=1}^{\infty} P(T=n) g_X(s)^n \\ &= g_T(g_X(s)). \end{aligned}$$

Then, we have

$$\mathbb{E}[Y] = g_Y'(1) = g_T'(g_X(s)) g_X'(s)\Big|_{s=1} = \mathbb{E}[T] \cdot \mathbb{E}[X_1]. \qquad \square$$

Chapter 3

Probability Densities

3.1 Univariate Probability Densities

Recall Definition 1.2.4.

Example 3.1.1 Uniform Density
A random variable X with the probability density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

is said to be *uniformly distributed* on $[a, b]$. This is denoted by $X \sim U([a, b])$.

Example 3.1.2 Exponential Density
For $\lambda \in \mathbb{R}_{>0}$, the random variable X with the probability density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

is called an *exponential random variable*. This is denoted by $X \sim \mathcal{E}(\lambda)$.

Example 3.1.3 Gaussian Density
For $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$, the random variable X with the probability density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

is called a *Gaussian random variable*. This is denoted by $X \sim \mathcal{N}(m, \sigma^2)$. When $X \sim \mathcal{N}(0, 1)$, we say that X is a *standard Gaussian random variable*.

Example 3.1.4 Gamma Density

Let $\alpha, \beta \in \mathbb{R}_{>0}$. The random variable X with the probability density

$$f(x) = \begin{cases} \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

is called a *gamma distributed random variable* where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x.$$

This is denoted by $X \sim \Gamma(\alpha, \beta)$.

Note

When $\alpha = 1$, the gamma distribution is simply the exponential distribution:

$$\Gamma(1, \beta) = \mathcal{E}(\beta).$$

When $\alpha = n/2$ and $\beta = 1/2$, the corresponding distribution is called the *chi-squared distribution* with n degrees of freedom. When X admits this density, we denote this by

$$X \sim \chi_n^2.$$

3.2 Mean and Variance

Definition 3.2.1: Mean and Variance

Let X be a real random variable with the probability density function f. The *mean* of X is defined as

$$m_X \triangleq \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx,$$

provided that the integral exists. The *variance* of X is defined as

$$\sigma_X^2 \triangleq \mathrm{Var}(X) = \mathbb{E}[X - \mathbb{E}[X]]^2 = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f(x)\, dx,$$

provided that the integral exists.

Exercise 3.2.1

Show that if $X \sim \Gamma(\alpha, \beta)$, then $\mathbb{E}[X] = \alpha/\beta$ and $\mathrm{Var}(X) = \alpha/\beta^2$.

Solution:

$$
\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}\,\mathrm{d}x \\
&= \frac{1}{\beta\Gamma(\alpha)} \int_0^\infty u^\alpha e^{-u}\,\mathrm{d}u && \triangleright\ u = \beta x \\
&= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty x^2 \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}\,\mathrm{d}x \\
&= \frac{1}{\beta^2\Gamma(\alpha)} \int_0^\infty u^{\alpha+1} e^{-u}\,\mathrm{d}u && \triangleright\ u = \beta x \\
&= \frac{\Gamma(\alpha+2)}{\beta^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\beta^2}.
\end{aligned}
$$

Hence, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \alpha/\beta^2$. □

Exercise 3.2.2

Compute the mean and variance of X when $X \sim U([a,b])$, $X \sim \mathcal{E}(\lambda)$, and $X \sim \mathcal{N}(0,1)$.

Solution: Let $X \sim U([a,b])$. Then,

$$
\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}
$$

and

$$
\mathbb{E}[X^2] = \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right) = \frac{a^2+ab+b^2}{3}.
$$

Hence, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (a-b)^2/12$.

Let $X \sim \mathcal{E}(\lambda)$. Then,

$$
\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}\int_0^\infty u e^{-u}\,\mathrm{d}u = \frac{\Gamma(2)}{\lambda} = \frac{1}{\lambda}
$$

and

$$
\mathbb{E}[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda^2}\int_0^\infty u^2 e^{-u}\,\mathrm{d}u = \frac{\Gamma(3)}{\lambda^2} = \frac{2}{\lambda^2}.
$$

Hence, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1/\lambda^2$.

Let $X \sim \mathcal{N}(0,1)$. Then, it is evident that $\mathbb{E}[X] = 0$. We first have

$$
\begin{aligned}
\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty \frac{e^{-x}}{\sqrt{x}}\,\mathrm{d}x \\
&= \int_0^\infty 2e^{-u^2}\,\mathrm{d}u && \triangleright\ u = \sqrt{x} \\
&= \int_{-\infty}^\infty e^{-u^2}\,\mathrm{d}u = \sqrt{\pi}
\end{aligned}
$$

Moreover,

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\infty} u^2 e^{-u^2}\,\mathrm{d}u && \triangleright\ x = \sqrt{2}u \\
&= \frac{4}{\sqrt{\pi}} \int_{0}^{\infty} u^2 e^{-u^2}\,\mathrm{d}u \\
&= \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} x^{1/2} e^{-x}\,\mathrm{d}x && \triangleright\ u = \sqrt{x} \\
&= \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) = 1.
\end{aligned}$$

□

Note

Let X be a random variable admitting the following probability density:

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Then, although f is even, $\mathbb{E}[X]$ is not defined.

3.3 Chebyshev's Inequality

Theorem 3.3.1 Markov's Inequality

Let X be a random variable and let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a function. Then, for each $a \in \mathbb{R}_{>0}$,

$$P(f(X) \geq a) \leq \frac{\mathbb{E}[f(X)]}{a}$$

given that $\mathbb{E}[f(X)]$ exists.

Proof. Let $C := \{ x \in \mathbb{R} \mid f(x) \geq a \}$ so that $|f(x)| \leq f(x) \cdot I_C(x)$. Then,

$$\begin{aligned}
\mathbb{E}[f(X)] &\geq \mathbb{E}[f(x) \cdot I_C(X)] \\
&\geq \mathbb{E}[af(x)] = a\mathbb{E}[f(X)].
\end{aligned}$$

□

Theorem 3.3.2 Chebyshev's Inequality

Let X be a random variable for with the mean m and the variance σ^2 are defined. Then, for each $\varepsilon \in \mathbb{R}_{>0}$,

$$P(|X-m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Same as the proof of Theorem 2.4.2. □

Definition 3.3.3: P-Almost Surely Null/Constant

- A random variable X is said to be *P-almost surely null* if $P(X=0) = 1$.
- A random variable X is said to be *P-almost surely constant* if $P(X=c) = 1$ for some constant c.

Lemma 3.3.4

Let X be a random variable with the mean m and the variance 0. Then, X is P-almost surely m.

Proof. Note that $\{\omega \in \Omega : |X(\omega) - m| > 0\} = \bigcup_{n=1}^{\infty} \{\omega \in \Omega : |X(\omega) - m| \geq 1/n\}$ so that

$$P(|X - m| > 0) \leq \sum_{n=1}^{\infty} P\left(|X - m| \geq \frac{1}{n}\right).$$

By Chebyshev's Inequality, we have

$$P\left(|X - m| \geq \frac{1}{n}\right) \leq \operatorname{Var}(X) \cdot n^2 = 0.$$

Therefore, $P(X = m) = 1 - P(|X - m| > 0) = 1$. □

3.4 Characteristic Function of a Random Variable

Definition 3.4.1: Characteristic Function

Let X be a real random variable with the probability density function f_X. The *characteristic function* $\phi_X : \mathbb{R} \to \mathbb{C}$ of X is defined as

$$\phi_X(u) \triangleq \mathbb{E}[e^{iuX}] = \int_{-\infty}^{\infty} e^{iux} f(x)\, dx.$$

Note

- Definition 3.4.1 is well-defined as cos and sin are bounded.
- $\phi_{aX+b}(u) = \mathbb{E}[e^{iuaX} e^{iub}] = e^{iub} \phi_X(au)$ for any real numbers a and b.
- If two real random variables X and Y satisfy $\mathbb{E}[e^{iuX}] = \mathbb{E}[e^{iuY}]$ for all $u \in \mathbb{R}$, then $P(X \leq x) = P(Y \leq x)$ for all $x \in \mathbb{R}$. Hence, the characteristic function uniquely determines the distribution of a random variable.
- It should be emphasized that two random variables with the same distribution function are not necessarily identical random variables. For instance, take $X \sim \mathcal{N}(0, 1)$ and $Y = -X$.

3.5 Multivariate Probability Densities

Definition 3.5.1: Random Vector

Let $X_1, X_2, \cdots, X_n$ be real random variables. The vector $X = (X_1, \cdots, X_n)$ is then called a *real random vector* of dimension n. The function $F_X : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ defined by

$$F_X(x_1, \cdots, x_n) \triangleq P(X_1 \leq x_1, \cdots, X_n \leq x_n)$$

is the cumulative distribution function of X. If

$$F_X(x_1, \cdots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(y_1, \cdots, y_n)\, \mathrm{d}y_n \cdots \mathrm{d}y_1,$$

for some nonnegative function $f_X : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, then f_X is called a *(joint) probability density function* of X.

Note

Let $X = (X_1, \cdots, X_n)$ be a real random vector admitting a probability density function $f(x_1, \cdots, x_n)$. Let $Y = (X_1, \cdots, X_\ell)$ for $1 \leq \ell \leq n$. Then,

$$\begin{aligned} F_Y(y_1, \cdots, y_\ell) &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_\ell} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \cdots, z_n)\, \mathrm{d}z_n \cdots \mathrm{d}z_1 \\ &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_\ell} \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(z_1, \cdots, z_n)\, \mathrm{d}z_n \cdots \mathrm{d}z_{\ell+1} \right] \mathrm{d}z_\ell \cdots \mathrm{d}z_1; \end{aligned}$$

hence

$$f_Y(y_1, \cdots, y_\ell) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \cdots, y_\ell, z_{\ell+1}, \cdots, z_n)\, \mathrm{d}z_n \cdots \mathrm{d}z_{\ell+1}$$

is a probability density function of Y.

3.6 Covariance, Cross-Covariance, and Correlation

Definition 3.6.1: Mean and Covariance Matrix of Random Vector

Let $X = (X_1, \cdots, X_n)$ be a real random vector of dimension n. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function. Then,

$$\mathbb{E}[g(X_1, \cdots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_n) f_X(x_1, \cdots, x_n)\, \mathrm{d}x_n \cdots \mathrm{d}x_1$$

is called the *expectation* of $g(X_1, \cdots, X_n)$. The *mean* of X is defined as

$$m = \mathbb{E}[X] \triangleq \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_2] \end{bmatrix}.$$

The *covariance matrix* of X is defined as

$$\Gamma = \mathbb{E}[(X-m)(X-m)^\mathsf{T}] \triangleq \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

where $\sigma_{ij} = \mathbb{E}[(X_i - m_i)(X_j - m_j)]$.

Note

The covariance matrix Γ is symmetric and positive semi-definite. For any $(u_1, \cdots, u_n) \in \mathbb{R}^n$, we have

$$u^\mathsf{T}\Gamma u = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j \sigma_{ij} = \mathbb{E}\left[\left(\sum_{i=1}^{n} u_i(X_i - m_i)\right)^2\right] \geq 0.$$

Definition 3.6.2: Cross-Covariance Matrix

Let $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_p)$ be two real random vectors. The *cross-covariance matrix* of X and Y is defined by

$$\Sigma_{XY} \triangleq \mathbb{E}[(X - m_X)(Y - m_Y)^\mathsf{T}].$$

X and Y are said to be *uncorrelated* if $\Sigma_{XY} = 0$.

Note

- In particular, $\Sigma_{XX} = \Gamma_X$.
- Obviously, $\Sigma_{XY} = \Sigma_{YX}^\mathsf{T}$.
- Let A be a $k \times n$ matrix, C be a $\ell \times p$ matrix, and b and d be vectors of dimension k and ℓ, respectively. Then,

$$m_{AX+b} = Am_X + b$$

and

$$\Sigma_{AX+b, CY+d} = A\Sigma_{XY}C^\mathsf{T}.$$

In particular, $\Gamma_{AX+b} = A\Gamma_X A^\mathsf{T}$.

Definition 3.6.3: Characteristic Function of Random Vector

Let $X = (X_1, \cdots, X_n)$ be a random vector that admits a probability density function. is the fuction $\phi_X : \mathbb{R}^n \to \mathbb{C}$ defined by

$$\phi_X(u_1, \cdots, u_n) = \mathbb{E}\left[e^{iu(X_1+\cdots+X_n)}\right].$$

Note

We have

$$\frac{\partial^k}{\partial^{k_1} u_1 \cdots \partial^{k_n} u_n} \phi_X(u_1, \cdots, u_n)$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} i^k x_1^{k_1} \cdots x_n^{k_n} e^{i(u_1 x_1 + \cdots + u_n x_n)} f_X(x_1, \cdots, x_n) \, \mathrm{d}x_n \cdots \mathrm{d}x_1$$

where $k = k_1 + \cdots + k_n$. Hence,

$$\frac{\partial^k}{\partial^{k_1} u_1 \cdots \partial^{k_n} u_n} \phi_X(0, \cdots, 0) = i^k \mathbb{E}\left[X_1^{k_1} \cdots X_n^{k_n}\right].$$

This will be justified in the advanced cources and is valid whenever

$$\mathbb{E}\left[|X_1|^{k_1} \cdots |X_n|^{k_n}\right] < \infty.$$

Exercise 3.6.1

Compute $\mathbb{E}[X^n]$ when $X \sim \mathcal{E}(\lambda)$.

Solution: We have

$$\phi_X(u) = \mathbb{E}[e^{iuX}] = \int_0^\infty e^{iux} \lambda e^{-\lambda x} \, \mathrm{d}x = \lambda \int_0^\infty e^{(iu-\lambda)x} \, \mathrm{d}x = \frac{\lambda}{\lambda - iu}.$$

Then, we have

$$\frac{\mathrm{d}^n}{\mathrm{d}^n u} \phi_X(u) = \frac{i^n \lambda n!}{(\lambda - iu)^{n+1}};$$

hence $\mathbb{E}[X^n] = i^{-n} \dfrac{i^n \lambda n!}{\lambda^{n+1}} = \dfrac{n!}{\lambda^{n+1}}$. □

3.7 Independence of Random Variables

Theorem 3.7.1

Let $X = (X_1, \cdots, X_n)$ be a real random vector. X_i's are independent random variables admitting probability density functions f_i if and only if f_i's are probability densities such that

$$f_X(x_1, \cdots, x_n) = \prod_{i=1}^n f_i(x_i)$$

is a probability density function of X.

Proof.

(⇒) We have, by independence and Fubini's theorem,

$$F_X(x_1,\cdots,x_n)=\prod_{i=1}^n F_{X_i}(x_i)=\prod_{i=1}^n \int_{-\infty}^{x_i} f_i(y_i)\,\mathrm{d}y_i=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\prod_{i=1}^n f_i(y_i)\,\mathrm{d}y_n\cdots\mathrm{d}y_1.$$

Hence, $\prod_{i=1}^n f_i(x_i)$ is a probability density function of X.

(⇐)

$$\begin{aligned}
P(X_1\le x_1)&=\int_{-\infty}^{x_1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^n f_i(y_i)\,\mathrm{d}y_n\cdots\mathrm{d}y_2\,\mathrm{d}y_1\\
&=\left(\int_{-\infty}^{x_1} f_1(y_1)\,\mathrm{d}y_1\right)\left(\int_{-\infty}^{\infty} f_2(y_2)\,\mathrm{d}y_2\right)\cdots\left(\int_{-\infty}^{\infty} f_2(y_n)\,\mathrm{d}y_n\right)\\
&=\int_{-\infty}^{x_1} f_1(y_1)\,\mathrm{d}y_1.
\end{aligned}$$

Hence, f_1 is a probability density function of X_1. Similarly, f_i is a probability density function of X_i for all i.

Moreover, by Fubini's theorem,

$$\begin{aligned}
F(x_1,\cdots,x_n)&=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\prod_{i=1}^n f_i(y_i)\,\mathrm{d}y_n\cdots\mathrm{d}y_1\\
&=\prod_{i=1}^n\int_{-\infty}^{x_i} f_i(y_i)\,\mathrm{d}y_i\\
&=\prod_{i=1}^n F_i(x_i).
\end{aligned}$$

Hence, X_i's are independent random variables. □

Lemma 3.7.2 Product Formula

Let $X_1,\cdots,X_n$ be real random variables admitting probability density functions $f_1,\cdots,f_n$, respectively. Then, for any functions $g_i:\mathbb{R}\to\mathbb{C}$ for $i\in[n]$, we have

$$\mathbb{E}\left[\prod_{i=1}^n g_i(X_i)\right]=\prod_{i=1}^n\mathbb{E}[g_i(X_i)].$$

Proof. Fubini's theorem and Theorem 3.7.1. □

Note

In particular, we get

$$\phi_X(u_1,\cdots,u_n)=\prod_{i=1}^n\phi_{X_i}(u_i)$$

for all $u_i\in\mathbb{R}$ where ϕ's are characteristic functions of corresponding random vector or random variable by applying Product Formula.

Although we cannot prove in this stage, the converse is also true.

Lemma 3.7.3 Convolution Formula

Let X and Y be independent real random variables admitting probability density functions f_X and f_Y, respectively. Then, a probability density function f_Z of the random variable $Z = X + Y$ is given by:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,\mathrm{d}x.$$

Proof. Fix $z_0 \in \mathbb{R}$ and let $C = \{(x, y) \mid x + y \le z_0\}$. We have

$$\begin{aligned}
\int_{-\infty}^{z_0} f_Z(z)\,\mathrm{d}z &= \int_{-\infty}^{z_0} \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z_0} f_X(x) f_Y(z-x)\,\mathrm{d}z\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z_0 - x} f_X(x) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \iint_{\mathbb{R}^2} I_C(x, y) f_X(x) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \mathbb{E}[I_C(X, Y)] = P(X + Y \le z_0).
\end{aligned}$$

□

Definition 3.7.4: Independence of Random Vector

Let X and Y are real random vectors of dimension n and p, respectively. We say X and Y are *independent* if

$$P(X \le x, Y \le y) = P(X \le x) P(Y \le y)$$

for all $x \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$.

Theorem 3.7.5

Let $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_p)$ be real random vectors. Then, X and Y are independent random vectors admitting probability density functions f_X and f_Y, respectively, if and only if f_X and f_Y are probability density functions such that $f_Z(x, y) = f_X(x) f_Y(y)$ is a probability density function of $Z = (X_1, \cdots, X_n, Y_1, \cdots, Y_p)$.

Proof. Same as Theorem 3.7.1. □

Lemma 3.7.6

Let $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_p)$ be independent real random vectors. Let $g : \mathbb{R}^n \to \mathbb{R}$ and $h \to \mathbb{R}^p \to \mathbb{R}$. Then,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]$$

provided that the quantities are well-defined.

Proof. Same as Lemma 3.7.2. □

Chapter 4

Convergences

4.1 Almost-Sure Convergence

Definition 4.1.1: Almost-Sure Convergence

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of random variables. One says that $X_n \xrightarrow{\text{a.s.}} X$ (read X_n converges to X almost surely when $n \to \infty$) if there exists an event N of null probability such that for all $\omega \in N^{\mathsf{c}}$, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$. In other words, $P(\lim_{n\to\infty} X_n = X) = 1$. (See Lemma 4.1.2.)

Lemma 4.1.2

If the almost-sure limit of a sequence $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ exists, it is *essentially unique*. If $X_n \xrightarrow{\text{a.s.}} X$ and $X_n \xrightarrow{\text{a.s.}} X'$, then $X = X'$ P-a.s., i.e., $P(X = X') = 1$.

Proof. There are events of null probability $N, N' \subseteq \Omega$ such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in N \cup N'$. Now, note that $P(N \cup N') = 0$; hence $X(\omega) = X'(\omega)$ for all $\omega \in (N \cup N')^{\mathsf{c}}$. □

Note

Notation 4.1.3

Let $\langle A_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of evenets. We write

$$\{A_n \text{ i.o.}\} \triangleq \{\omega : \omega \in A_n \text{ infinitely often}\}.$$

In other words,

$$\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Theorem 4.1.4 First Borel–Cantelli Lemma

For any sequence of events $\langle A_n \rangle_{n \in \mathbb{Z}_{>0}}$,

$$\sum_{n=1}^{\infty} P(A_n) < \infty \implies P(A_n \text{ i.o.}) = 0.$$

Proof. Let $B_n \triangleq \bigcup_{k=n}^{\infty} A_k$. Then, we have

$$
\begin{aligned}
P(A_n \text{ i.o.}) &= P\left(\bigcap_{n=1}^{\infty} B_n\right) \\
&= \lim_{n\to\infty} P(B_n) && \triangleright \text{ Sequential Continuity of Probability} \\
&= \lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) \\
&\leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) = 0.
\end{aligned}
$$

□

Theorem 4.1.5 Second Borel–Cantelli Lemma

For any sequence of independent events $\langle A_n \rangle_{n\in\mathbb{Z}_{>0}}$,

$$
\sum_{n=1}^{\infty} P(A_n) = \infty \implies P(A_n \text{ i.o.}) = 1.
$$

Proof. Let $B_n \triangleq \bigcap_{k=n}^{\infty} A_k$. Note that $P((A_n \text{ i.o.})^c) = P\left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c\right)$. Then,

$$
\begin{aligned}
P\left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c\right) &\leq \sum_{n=1}^{\infty} P\left(B_n\right) \\
&= \sum_{n=1}^{\infty} \lim_{m\to\infty} P\left(\bigcap_{k=n}^{m} A_k^c\right) && \triangleright \text{ Sequential Continuity of Probability} \\
&= \sum_{n=1}^{\infty} \lim_{m\to\infty} \prod_{k=n}^{m} (1 - P(A_k)) \\
&\leq \sum_{n=1}^{\infty} \lim_{m\to\infty} \exp\left(-\sum_{k=n}^{m} P(A_k)\right) && \triangleright\ e^{-x} \leq 1 - x \\
&= \sum_{n=1}^{\infty} \exp\left(-\lim_{m\to\infty} \sum_{k=n}^{m} P(A_k)\right) \\
&= \sum_{n=1}^{\infty} 0 = 0.
\end{aligned}
$$

□

Exercise 4.1.1 Borel's Law of Large Numbers

Consider a sequence of independent random variables $\langle X_n \rangle_{n\in\mathbb{Z}_{>0}}$ with values in $\{0, 1\}$ such that $P(X_n = 1) = p$ for all $n \in \mathbb{Z}_{>0}$. Define the empirical frequency of "1" as

$$
\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.
$$

Show that $\overline{X}_n \xrightarrow{\text{a.s.}} p$ as $n \to \infty$.

Solution: Apply Strong Law of Large Numbers.

4.2 A Criterion for Almost-Sure Convergence

Theorem 4.2.1
Let $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ be a sequence of random variables. It converges almost surely to the random variable X if and only if

$$\forall\varepsilon\in\mathbb{R}_{>0},\ P(|X_n-X|\geq\varepsilon\text{ i.o.})=0.$$

Proof.

$$\begin{aligned}
&P\left(\lim_{n\to\infty}X_n=X\right)=1\\
\iff&\exists N\in\mathcal{F},\ \left(P(N)=0\land\forall\omega\in N^c,\ \lim_{n\to\infty}X_n(\omega)=X(\omega)\right)\\
\iff&\forall\varepsilon\in\mathbb{R}_{>0},P(|X_n-X|<\varepsilon\text{ for all but finitely many }n)=1\\
\iff&\forall\varepsilon\in\mathbb{R}_{>0},P(|X_n-X|\geq\varepsilon\text{ for infinitely many }n)=0
\end{aligned}$$ □

Corollary 4.2.2
Let $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ be a sequence of random variables. If

$$\forall\varepsilon\in\mathbb{R}_{>0},\ \sum_{n=1}^{\infty}P(|X_n-X|\geq\varepsilon)<\infty$$

for a random variable X, then $X_n\xrightarrow{\text{a.s.}}X$.

Proof. Combine First Borel–Cantelli Lemma and Theorem 4.2.1. □

4.3 The Strong Law of Large Numbers

Theorem 4.3.1 Strong Law of Large Numbers
Let $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ be identically distributed random variables. Assume that their mean $\mu=\mathbb{E}[X_1]$ is defined with finite variance σ^2. Moreover, assume that they are uncorrelated, i.e.,

$$\operatorname{Cov}(X_i,X_j)=\mathbb{E}[(X_i-\mu)(X_j-\mu)]=0$$

for all $i\neq j$. Then, letting $S_n=\sum_{i=1}^n X_i$, we have

$$\frac{S_n}{n}\xrightarrow{\text{a.s.}}\mu\quad\text{as}\quad n\to\infty.$$

Proof. WLOG, $\mu=0$. For each $m\in\mathbb{Z}_{>0}$, let $Z_m:=\max_{k=1}^{2m+1}\left|\sum_{i=1}^k X_{m^2+i}\right|$. Moreover, for each $n\in\mathbb{Z}_{>1}$, let $m(n)$ be the unique integer such that

$$m(n)^2<n\leq[m(n)+1]^2.$$

Then, we have

$$\left|\frac{S_n}{n}\right|\leq\left|\frac{S_{m(n)^2}}{m(n)^2}\right|+\frac{Z_{m(n)}}{m(n)^2}$$

for all $n>1$. Hence, we only need to prove $\frac{S_{m^2}}{m^2}\xrightarrow{\text{a.s.}}0$ and $\frac{Z_m}{m^2}\xrightarrow{\text{a.s.}}0$ as $m\to\infty$.

- Fix any $\varepsilon \in \mathbb{R}_{>0}$. By Chebyshev's Inequality, we have
$$P\left(\left|\frac{S_{m^2}}{m^2}\right| \geq \varepsilon\right) \leq \frac{\operatorname{Var}(S_{m^2})}{m^4\varepsilon^2} = \frac{\sigma^2}{m^2\varepsilon^2}.$$
Hence, we have $\sum_{m=1}^{\infty} P\left(\left|\frac{S_{m^2}}{m^2}\right| \geq \varepsilon\right) < \infty$. Therefore, by Corollary 4.2.2, $\frac{S_{m^2}}{m^2} \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$.
- Fix any $\varepsilon \in \mathbb{R}_{>0}$. Let $\xi_{m,k} := \sum_{i=1}^{k} X_{m^2+i}$ so that
$$\left\{\frac{Z_m}{m^2} \geq \varepsilon\right\} \subseteq \bigcup_{k=1}^{2m+1} \{|\xi_{m,k}| \geq m^2 k\}$$
for each $m \in \mathbb{Z}_{>0}$. Note that $\mathbb{E}[\xi_{m,k}] = 0$ and $\operatorname{Var}(\xi_{m,k}) = \sum_{i=1}^{k} \operatorname{Var}(X_{m^2+i}) = k\sigma^2$ as X_i's are uncorrelated. Therefore, by σ-subadditivity, we have
$$\begin{aligned}
P\left(\frac{Z_m}{m^2} \geq \varepsilon\right) &\leq \sum_{k=1}^{2m+1} P(|\xi_{m,k}| \geq m^2 k) && \triangleright\ \sigma\text{-subadditivity} \\
&\leq \sum_{k=1}^{2m+1} \frac{\operatorname{Var}(\xi_{m,k})}{m^4 k^2} && \triangleright\ \text{Chebyshev's Inequality} \\
&\leq \frac{\sigma^2(2m+1)}{m^4}.
\end{aligned}$$
Hence, $\sum_{m=1}^{\infty} P\left(\frac{Z_m}{m^2} \geq \varepsilon\right) < \infty$. Therefore, by Corollary 4.2.2, $\frac{Z_m}{m^2} \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$. □

Theorem 4.3.2 Kolmogorov's Strong Law of Large Numbers

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of independent and identically distributed random variables with mean μ. Then,
$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu \quad \text{as} \quad n \to \infty.$$

Note

Theorem 4.3.1 requires the random varaibles to have finite variance and to be uncorrelated, while Theorem 4.3.2 requires the random varaibles to be independent.

4.4 Convergence in Law

Definition 4.4.1: Convergence in Law

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ and X be real random variables with respective cumulative distribution functions $\langle F_{X_n} \rangle_{n \in \mathbb{Z}_{>0}}$ and F_X. One says that $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ *converges in law to* X if
$$\forall x \in \mathbb{R}, \left(\lim_{a \to x^-} F(a) = F(x) \implies \lim_{n \to \infty} F_{X_n}(x) = F_X(x)\right). \tag{4.1}$$
This is denoted by $X_n \xrightarrow{\mathcal{L}} X$.

Note

In the definition of convergence in law, the discontinuity points of the cumulative distribution function do not play a spacial part. If ⟨4.1⟩ were required to hold without the premise $\lim_{a \to x^-} F(a) = F(x)$, then defining $X_n \equiv a + \frac{1}{n}$ and $X \equiv a$, we could not say that $X_n \xrightarrow{\mathcal{L}} X$ because $P(X_n \leq a) = 0$ does not converge toward $P(X \leq a) = 1$.

Exercise 4.4.1

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of independent random variables such that $Z_n \sim U([0,1])$. Define

$$Z_n := \min\{X_1, \ldots, X_n\}.$$

Show that $nZ_n \xrightarrow{\mathcal{L}} X$ where $X \sim \mathcal{E}(1)$.

Solution: For $x \in \mathbb{R}$, we have

$$P(nZ_n \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \left(1 - \frac{x}{n}\right)^n & \text{if } 0 \leq x \leq n \\ 1 & \text{otherwise.} \end{cases}$$

Therefore, for every $x \in \mathbb{R}_{\geq 0}$,

$$\lim_{n \to \infty} P(nZ_n \leq x) = \lim_{n \to \infty} \left(1 - \left(1 - \frac{x}{n}\right)^n\right) = 1 - e^{-x},$$

which is the cumulative distribution function of $\mathcal{E}(1)$. □

Theorem 4.4.2 Characteristic Function Criterion

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be real random variables with respective characteristic distribution functions $\langle \phi_{X_n} \rangle_{n \in \mathbb{Z}_{>0}}$. If the sequence $\langle \phi_{X_n} \rangle_{n \in \mathbb{Z}_{>0}}$ converges pointwise to some function $\phi : \mathbb{R} \to \mathbb{C}$ that is continuous at 0, then ϕ is a characteristic function of some real random variable X, and moreover, $X_n \xrightarrow{\mathcal{L}} X$.

4.5 The Central Limit Theorem

Theorem 4.5.1 Central Limit Theorem

Let $\langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ be a sequence of independent and identically distributed random variables with common (finite) mean μ and (finite) variance σ^2, respectively. Then,

$$\frac{\left(\sum_{i=1}^n X_i\right) - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} Z \quad \text{as} \quad n \to \infty$$

where $Z \sim \mathcal{N}(0,1)$.

Proof Sketch. WLOG, $\mu = 0$. Let $\phi(u)$ denote the characteristic function of X_1. Then, the characteristic function of $\left(\sum_{i=1}^n X_i\right)/\sigma\sqrt{n}$ is $\phi(u/\sigma\sqrt{n})^n$. Since $\phi(0) = 1$, $\phi'(0) = 0$, and $\phi''(0) = -\sigma^2$, we have

$$\phi\left(\frac{u}{\sigma\sqrt{n}}\right) = 1 - \frac{1}{2n}u^2 + o\left(\frac{1}{n}\right).$$

Therefore,

$$\lim_{n\to\infty} \phi\left(\frac{u}{\sigma\sqrt{n}}\right)^n = \lim_{n\to\infty}\left(1-\frac{u^2}{2n}\right)^n = e^{-u^2/n},$$

which is the characteristic function of Z. The result follows from Characteristic Function Criterion. □

4.6 Convergence in L^p and Hierarchy of Convergences

Definition 4.6.1: Convergence in Probability

(Restatement of Definition 2.4.4) A sequence of random variables $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ is said to *converge in probability* to a random variable X if if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0.$$

This is denoted by $X_n \xrightarrow{P} X$.

Definition 4.6.2: Convergence in L^p

For any $p \geq 1$, a sequence of random variables $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ such that $\mathbb{E}[|X_n|^p] < \infty$ for $n \in \mathbb{Z}_{>0}$ is said to *converge in L^p* to a random variable X such that $\mathbb{E}[|X|^p] < \infty$ if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

This is denoted by $X_n \xrightarrow{L^p} X$.

Theorem 4.6.3

Let $\langle X_n\rangle_{n\in\mathbb{Z}_{>0}}$ be a sequence of random variables and X be a random variable.

(1) If $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{P} X$.

(2) If $X_n \xrightarrow{L^p} X$ for some $p \geq 1$, then $X_n \xrightarrow{P} X$.

(3) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{\mathcal{L}} X$.

Proof.

(a) Fix any $\varepsilon \in \mathbb{R}_{>0}$. By Theorem 4.2.1, we have $P\left(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty \{|X_k - X| \geq \varepsilon\}\right) = 0$. By Theorem 1.1.4 (2), we get

$$0 = \lim_{n\to\infty} P\left(\bigcup_{k=n}^\infty \{|X_k - X| \geq \varepsilon\}\right) \geq \lim_{n\to\infty} P(\{|X_n - X| \geq \varepsilon\}).$$

Hence, $X_n \xrightarrow{P} X$ as $n \to \infty$.

(b) We have

$$P(|X_n - X| \geq \varepsilon) = P(|X_n - X|^p \geq \varepsilon^p) \leq \frac{\mathbb{E}[|X_n - X]^p}{\varepsilon^p} \to 0$$

as $n \to \infty$.

(c) We need the following lemma:

Claim 1. Let X and Y be random variables. Let $a \in \mathbb{R}$ and $\varepsilon \in \mathbb{R}_{>0}$. Then,

$$P(Y \le a) \le P(X \le a+\varepsilon) + P(|Y-X| \ge \varepsilon).$$

Proof. We have:

$$\begin{aligned} P(Y \le a) &\le P(Y \le a, X \le a+\varepsilon) + P(Y \le a, X \ge a+\varepsilon) \\ &\le P(X \le a+\varepsilon) + P(Y-X \le a-X, a-X \le -\varepsilon) \\ &\le P(X \le a+\varepsilon) + P(Y-X \le -\varepsilon) \\ &\le P(X \le a+\varepsilon) + P(|Y-X| \le \varepsilon). \end{aligned}$$ □

Applying Claim 1 twice, we get

$$P(X \le x-\varepsilon) - P(|X_n - X| \ge \varepsilon) \le P(X_n \le x) \tag{4.2}$$
$$P(X_n \le x) \le P(X \le x+\varepsilon) + P(|X_n - X| \ge \varepsilon) \tag{4.3}$$

for every $\varepsilon \in \mathbb{R}_{>0}$. Then, we have

$$\begin{aligned} P(X \le x-\varepsilon) &\le P(X_n \le x) && \triangleright \langle 4.2 \rangle \text{ and } X_n \xrightarrow{P} X \\ &\le P(X \le x+\varepsilon) && \triangleright \langle 4.3 \rangle \text{ and } X_n \xrightarrow{P} X \end{aligned}$$

for every $n \in \mathbb{Z}_{>0}$. Therefore, if F_X is continuous at x, limiting $n \to \infty$, we have

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x).$$ □

Chapter 5

Markov Chain

5.1 Markov Chain

Definition 5.1.1: Stochastic Process

A *stochastic process with state space* $\mathbb{S}$ is a sequence $X = \langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$ of random variables taking values in $\mathbb{S}$.

Definition 5.1.2: Markov Chain

A stochastic process $X = \langle X_n \rangle_{n \in \mathbb{Z}_{>0}}$, with a discrete state space $\mathbb{S}$, is a *(homogeneous) Markov chain* with transition probabilities $p = \langle p(i,j) \rangle_{i,j \in \mathbb{S}}$ if for any $i_0, j_0, \cdots, i_{n-1}, j_{n-1}, i, j \in \mathbb{S}$ such that

$$P(X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) > 0,$$

we have

$$\begin{aligned} &P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) \\ = &P(X_{n+1} = j \mid X_n = i, X_{n-1} = j_{n-1}, \cdots, X_0 = j_0) = p(i,j). \end{aligned}$$

When $\mathbb{S}$ is *finite*, we refer to p as a *transition matrix*.

Definition 5.1.3: Stochastic Matrix

Any square matrix p satisfying $p(i,j) \geq 0$ and $\sum_{j \in \mathbb{S}} p(i,j) = 1$ is called a *stochastic matrix*.

5.2 Multistep Transition Probabilities

Theorem 5.2.1

The m-step transition probabilities of a Markov chain are independent of the past.

$$P(X_{n+m} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) = P(X_{n+m} = j \mid X_n = i) = p^m(i,j)$$

where p^m is the m-th power of the transition matrix.

Proof. Just feel it. □

5.3 Classification of States

Notation 5.3.1

For any event A and state x, we introduce the following notation:

$$P_x(A) \triangleq P(A \mid X_0 = x)$$

Expectations under this probability measure are denoted by $\mathbb{E}_x$. These simply mean that, when computing P_x or $\mathbb{E}_x$, we assume that the associated Markov chain starts from x.

Definition 5.3.2: Time of the First Jump

Let T_y be the random variable of the *time of the first jump* to state y:

$$T_y \triangleq \min\{\, n \geq 1 \mid X_n = y \,\}$$

Note that, if the chain starts from y, the time zero does *not* count as a visit.

Definition 5.3.3: Stopping Time

Assume $\langle X_n \rangle_{n \in \mathbb{Z}_{\geq 0}}$ is a Markov chain. Given a (extended) random variable T, with values in the set of time indices $\{0, 1, \cdots, \infty\}$, T is called a *stopping time* (with respect to $\langle X_n \rangle_{n \in \mathbb{Z}_{\geq 0}}$) if the occurrence or non-occurrence of the event $\{T = n\}$ can be determined only by looking at the values $X_0, \cdots, X_n$.

Definition 5.3.4: Ever Jumping

Denote the *probability of ever jumping to y, starting from x*, by ρ_{xy}:

$$\rho_{xy} \triangleq P_x(T_y < \infty).$$

Theorem 5.3.5 Strong Markov Property of a Markov Chain

Assume $\langle X_n \rangle_{n \in \mathbb{Z}_{\geq 0}}$ is a Markov chain and T is a stopping time. Conditional on $T < \infty$ and $X_T = y$, any other information about $X_0, \cdots, X_{T-1}$ is irrelevant for the future distribution of the Markov chain. Namely, the new process $\langle \tilde{X}_n \triangleq X_{T+n} \rangle_{n \in \mathbb{Z}_{\geq 0}}$ is also a Markov chain, with the same transition matrix and with initial state y, and it is independent of T and the past values $(X_0, \cdots, X_{T-1})$.

Proof. Skipped. □

Note

If we restrict Theorem 5.3.5 by forcing T to be *deterministic*, then the new property becomes a *regular (or, standard) Markov property*.

Definition 5.3.6: k-th Jump

For $k \geq 1$, we introduce the *time of the k-th jump* to state y:

$$\begin{aligned} T_y^1 &\triangleq T_y, \\ T_y^k &\triangleq \min\left\{ n > T_y^{k-1} \,\middle|\, X_n = y \right\}. \end{aligned}$$

Note that they are stopping times.

Lemma 5.3.7

$P_x(T_y^k < \infty) = \rho_{xy}\rho_{yy}^{k-1}$.

Proof. For each $k \geq 2$, we have

$$\begin{aligned} P_y(T_y^k < \infty) &= P_y(T_y^k < \infty \mid T_y^{k-1} < \infty) P_y(T_y^{k-1} < \infty) \\ &= \rho_{yy} \cdot P_y(T_y^{k-1} < \infty). && \triangleright \text{ Theorem 5.3.5} \end{aligned}$$

Hence, the result follows from mathematical induction. □

Definition 5.3.8: Transient and Recurrent States

- If $\rho_{yy} < 1$, then the state y is called *transient*.
- If $\rho_{yy} = 1$, then the state y is called *recurrent*.

Note

By Lemma 5.3.7, we have the following:
- If a state y is transient, then $\lim_{k\to\infty} P_y(T_y^k < \infty) = 0$.
- If a state y is recurrent, then $\lim_{k\to\infty} P_y(T_y^k < \infty) = 1$.

Definition 5.3.9: The Number of Visits

Let N_y be the random variable of the number of visits to state y, i.e.,

$$N_y \triangleq \sum_{k=1}^{\infty} \mathbf{1}_{\{T_y^k < \infty\}},$$

where $\mathbf{1}_A$ is an indicator of event A.

Lemma 5.3.10

If y is a transient state,

$$\mathbb{E}_x[N_y] = \frac{\rho_{xy}}{1 - \rho_{yy}}.$$

Proof.

$$\begin{aligned}
\mathbb{E}_x[N_y] &= \mathbb{E}_x\left[\sum_{k=1}^{\infty} \mathbf{1}_{\{T_y^k<\infty\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{E}_x\left[\mathbf{1}_{\{T_y^k<\infty\}}\right] && \triangleright \text{ Tonelli's Theorem} \\
&= \sum_{k=1}^{\infty} P(T_y^k < \infty) \\
&= \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1} && \triangleright \text{ Lemma 5.3.7} \\
&= \frac{\rho_{xy}}{1-\rho_{yy}}
\end{aligned}$$

□

Lemma 5.3.11

If y is a recurrent state, then

$$P_y(N_y = \infty) = 1.$$

Proof. Note that

$$\{N_y = \infty\} = \bigcap_{k=1}^{\infty} \{T_y^k < \infty\}.$$

Hence,

$$\begin{aligned}
P_y(N_y = \infty) &= \lim_{N\to\infty} P_y\left(\bigcap_{k=1}^{N} \{T_y^k < \infty\}\right) && \triangleright \text{ Sequential Continuity of Probability} \\
&= \lim_{N\to\infty} P_y\big(T_y^N < \infty\big) \\
&= \lim_{N\to\infty} \rho_{yy}^N && \triangleright \text{ Lemma 5.3.7} \\
&= 1.
\end{aligned}$$

□

Lemma 5.3.12

A state y is recurrent if and only if $\mathbb{E}_y[N_y] = \infty$.

Proof. Combine Lemmas 5.3.10 and 5.3.11. □

Definition 5.3.13: Communicating States

We say that x *communicates with* y, and denote it by $x \to y$, if

$$\rho_{xy} = P_x(T_y < \infty) > 0.$$

Lemma 5.3.14

x communicates with y if and only if there is some $m \in \mathbb{Z}_{>0}$ such that $p^m(x,y) > 0$.

Proof.

($\Rightarrow$) As $P_x(T_x < \infty) = \sum_{k=1}^{\infty} P_x(T_y = k) > 0$, there is some $m \in \mathbb{Z}_{>0}$ such that $P_x(T_y = k) > 0$. Such m satisfies $p^m(x,y) \geq P_x(T_y = k) > 0$.

($\Leftarrow$) Trivial. $\square$

Lemma 5.3.15

If $x \to y$ and $y \to z$, then $x \to z$.

Proof. By Lemma 5.3.14, there are $m_1, m_2 \in \mathbb{Z}_{>0}$ such that $p^{m_1}(x,y) > 0$ and $p^{m_2}(y,z) > 0$. Hence, we have $p^{m_1+m_2}(x,z) \geq p^{m_1}(x,y) \cdot p^{m_2}(y,z) > 0$. The result follows from Lemma 5.3.14. $\square$

Lemma 5.3.16

If $x \to y$ and $\rho_{yx} < 1$, then x is a transient state.

Proof. Let $K \triangleq \{k \in \mathbb{Z}_{>0} \mid p^k(x,y) > 0\}$. There is a sequence $y_1, \cdots, y_{K-1}$ of states so that

$$p(x,y_1)p(y_1,y_2)\cdots p(y_{K-2},y_{K-1})p(y_{K-1},y) > 0$$

Then, we have

$$P_x(T_x = \infty) \geq p(x,y_1)p(y_1,y_2)\cdots p(y_{K-2},y_{K-1})p(y_{K-1},y) \cdot (1-\rho_{yx}) > 0$$

so that x is transient. $\square$

Lemma 5.3.17

If x is recurrent and $x \to y$, then $\rho_{yx} = 1$.

Proof. Direct consequence of Lemma 5.3.16. $\square$

Lemma 5.3.18

$\mathbb{E}_x[N_y] = \sum_{n=1}^{\infty} p^n(x,y)$. Moreover, y is recurrent if and only if $\sum_{n=1}^{\infty} p^n(y,y) = \infty$.

Proof. Note that $N_y = \sum_{n=1}^{\infty} \mathbf{1}_{\{X_n = y\}}$. Hence, by the same argument as in the proof of Lemma 5.3.11, $\mathbb{E}_x[N_y] = \sum_{n=1}^{\infty} p^n(x,y)$. The result follows from Lemma 5.3.12. $\square$

Lemma 5.3.19

If x is recurrent and $x \to y$, then y is recurrent.

Proof. By Lemma 5.3.17, $\rho_{yx} = 1$. By Lemma 5.3.14, there are some $m_1, m_2 \in \mathbb{Z}_{>0}$ such that $p^{m_1}(x,y) > 0$ and $p^{m_2}(y,x) > 0$. Then, for $n \geq m_1 + m_2$, we have

$$p^n(y,y) \geq p^{m_2}(y,x) \cdot p^{n-m_1-m_2}(x,x) \cdot p^{m_1}(x,y).$$

Hence, by Lemma 5.3.18, y is recurrent. $\square$

Definition 5.3.20: Closed Set

A nonempty set A of states is *closed* if

$$\forall i \in A,\ \forall j \in \mathbb{S} \setminus A,\ p(i,j) = 0.$$

Lemma 5.3.21

If A is a finite closed set, then A has at least one recurrent state.

Proof. Suppose all states in A are transient for the sake of contradiction. Fix any state $x \in A$. Then,

$$\begin{aligned}
\infty &> \sum_{y \in A} \mathbb{E}_x[N_y] && \triangleright \text{ Lemma 5.3.10} \\
&= \sum_{y \in A} \sum_{n=1}^{\infty} p^n(x,y) && \triangleright \text{ Lemma 5.3.18} \\
&= \sum_{n=1}^{\infty} \sum_{y \in A} p^n(x,y) \\
&= \sum_{n=1}^{\infty} 1 && \triangleright\ A \text{ is closed} \\
&= \infty,
\end{aligned}$$

which is a contradiction. □

Definition 5.3.22: Irreducible Set

A nonempty set A of states is *irreducible* if

$$\forall x \in A,\ \forall y \in A,\ x \to y.$$

Theorem 5.3.23

All states in a finite closed irreducible set is recurrent.

Proof. Combine Lemmas 5.3.19 and 5.3.21. □

Theorem 5.3.24 Decomposition Theorem

If the state space $\mathbb{S}$ of a Markov chain is finite, then

$$\mathbb{S} = T \uplus R_1 \uplus \cdots \uplus R_k$$

where T is the set of transient states, and each R_i's are closed irreducible sets of recurrent states.

Proof. The relation $x \sim y$ defined by $x \to y$ and $y \to x$ is an equivalence relation on $\mathbb{S} \setminus T$ by Lemma 5.3.15. Let R be an equivalence class of $\mathbb{S} \setminus T$ under $\sim$. Then, it is closed and irreducible. □

End.