

# Advanced Probability

LMD3 Course Notes

## General Information

- **Course:** Advanced Probability
- **Unit:** Fundamental
- **Credits:** 6
- **Coefficient:** 4
- **Prerequisites:** Real Analysis (1,2,3), Probability 1

## Objectives

This course provides a detailed study of the main concepts and methods of probability, including:

- Probability of events
- Laws and moments of random variables
- Conditioning and regression
- Transforms of random variables
- Gaussian distributions

# Course Outline

## Motivation

The real world is full of uncertainty. Weather, measurement errors, financial markets, genetics, and even everyday decisions all involve randomness. **Probability theory** provides the mathematical framework we need to model and analyze this uncertainty. It allows us to make predictions, quantify risks, and build models that are used in fields as diverse as science, engineering, economics, and data science.

In this module, we will study the foundations of probability at an advanced level, with the goal of developing both intuition and rigorous mathematical tools.

## 1. Fundamental Review on Random Variables

Just like in algebra or calculus, we deal with variables that take specific values. In probability, however, we work with **random variables**, which do not take a fixed value, but instead have a probability of taking certain values.

Each random variable has important **numerical characteristics**, such as the mean and the standard deviation. These parameters already tell us a lot about the distribution.

We will also study several key distributions, each describing different kinds of events:

- **Bernoulli distribution:** models a system with two outcomes (e.g., heads/tails, success/failure).
- **Binomial distribution:** the sum of independent Bernoulli random variables. This distribution is very important and widely used.
- **Normal (Gaussian) distribution:** one of the most important distributions in probability and statistics, used in many fields.
- **Poisson distribution:** models the number of occurrences of rare events in a fixed time or space interval.
- **Exponential distribution:** describes waiting times between events in a Poisson process.

## 2. Functions of Random Variables

We will then study important functions of random variables:

- Expectation, variance, and standard deviation,
- Moment generating functions (MGFs),
- Characteristic functions.

## 3. Modes of Convergence

Different types of convergence for sequences of random variables will be introduced, and the relationships between them will be studied.

## 4. Limit Theorems

This part covers the cornerstone results of probability:

- Weak Law of Large Numbers (WLLN),
- Strong Law of Large Numbers (SLLN),
- Central Limit Theorem (CLT).

The CLT is particularly important: it tells us that if we take a large number of independent random variables and look at their average, then this average behaves approximately like a Gaussian random variable, regardless of the original distribution.

## 5. Random Vectors

Finally, we extend the theory to random vectors:

- Probability laws of random vectors,
- Numerical characteristics (expectation, covariance matrix),
- Moment and characteristic functions,
- Conditional expectation,
- Multivariate normal distribution,
- Convergence and the multivariate Central Limit Theorem.

## Review of Basic Concepts

**Sample Space:** The set of all possible outcomes of an experiment is called the *sample space*, denoted by  $\Omega$ .

**Probability of an event:**

$$P(A) = \frac{\text{number of ways } A \text{ can happen}}{\text{total number of possible outcomes}}$$

**Examples:**

- Coin flips: If I flip a coin twice,

$$\Omega = \{HH, HT, TH, TT\}$$

Let  $A =$  "at least one head". Then  $A = \{HH, HT, TH\}$  and  $P(A) = 3/4$ . Let  $B =$  "no heads". Then  $B = \{TT\}$  and  $P(B) = 1/4$ .

- Dice rolls: If I roll two dice, then

$$\Omega = \{11, 12, \dots, 66\}, \quad |\Omega| = 36$$

Let  $A =$  "at least one die is a 5". Then

$$A = \{15, 25, 35, 45, 55, 65, 51, 52, 53, 54, 56\}, \quad |A| = 11,$$

so  $P(A) = 11/36$ .

### Axioms of Probability:

1.  $0 \leq P(E) \leq 1$  for any event  $E$ ,
2.  $P(\Omega) = 1$ ,
3. For any sequence of mutually exclusive events  $E_1, E_2, \dots$ ,

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i), \quad n = 1, 2, \dots$$

## Conditional Probability and Independence

**Conditional Probability:** Given two events  $A$  and  $B$  with  $P(B) > 0$ , the conditional probability of  $A$  given  $B$  is defined as

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

**Example:** Suppose we roll a fair die. Let

$$A = \{\text{even number}\} = \{2, 4, 6\}, \quad B = \{\text{number greater than 3}\} = \{4, 5, 6\}.$$

Then

$$P(A \cap B) = P(\{4, 6\}) = \frac{2}{6}, \quad P(B) = \frac{3}{6},$$

so

$$P(A | B) = \frac{2/6}{3/6} = \frac{2}{3}.$$

—

## Independence of Events

**Definition:** Two events  $A$  and  $B$  are said to be *independent* if

$$P(A \cap B) = P(A) \cdot P(B).$$

**Example:** Flip two fair coins. Let  $A$  = “first coin is heads” and  $B$  = “second coin is heads.” Then

$$P(A) = P(B) = \frac{1}{2}, \quad P(A \cap B) = P(\{HH\}) = \frac{1}{4}.$$

Since  $P(A \cap B) = (1/2)(1/2)$ , the events  $A$  and  $B$  are independent.

**Generalization:** A collection of events  $E_1, E_2, \dots, E_n$  is called *mutually independent* if, for every subset  $\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$ ,

$$P(E_{i_1} \cap \dots \cap E_{i_k}) = P(E_{i_1}) \cdots P(E_{i_k}).$$

**Remark 0.1.** *Disjoint means two events cannot occur together, i.e.,  $A \cap B = \emptyset$ . Independent means the occurrence of one does not affect the probability of the other; the events may still happen together.*

—

## From Events to Random Variables

So far, we have described uncertainty in terms of events within a sample space and assigned probabilities to these events. However, in practice, we often want to associate *numbers* with outcomes in order to perform calculations, compute averages, and apply mathematical tools.

This brings us to the notion of a **random variable**: a measurable function that assigns a real number to each outcome of the experiment. In other words, random variables translate the abstract language of outcomes and events into the numerical framework of mathematics.

**Example:** If the experiment is flipping a coin once, the sample space is

$$\Omega = \{\text{Heads, Tails}\}.$$

Define a random variable  $X$  by

$$X(\text{Heads}) = 1, \quad X(\text{Tails}) = 0.$$

Here,  $X$  converts qualitative outcomes (Heads/Tails) into quantitative values (1/0) that we can analyze with probability tools.

—

## Definition of a $\sigma$ -algebra

A collection  $\mathcal{F}$  of subsets of a sample space  $\Omega$  is called a  **$\sigma$ -algebra** if:

1.  $\Omega \in \mathcal{F}$ ,
2. If  $A \in \mathcal{F}$ , then  $A^c \in \mathcal{F}$ ,
3. If  $A_1, A_2, \dots \in \mathcal{F}$ , then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ .

# Chapter 1

## Fundamental Review on Random Variables

### 1.1 Introduction

The world around us is full of uncertainty, and probability is the mathematical tool we use to describe and analyze it. We cannot predict events like weather, measurement errors, or financial markets with certainty, but we can model their likelihood. In this course, we will learn how to turn uncertainty into something we can measure and work with.

Probability theory provides the mathematical framework for modeling uncertainty. Random variables and their probability distributions constitute the central tools of this framework. In this chapter, we review the essential notions that will serve as the foundation for advanced probability: numerical characteristics of random variables, common probability laws, and operations involving random variables.

#### 1.1.1 Random Variables and Probability Laws

##### Simple Intuitive Definition

A random variable is just a rule that assigns a number to every possible outcome of an experiment.

**Example 1.1.** *If the experiment is flipping a coin:*

$$X = \begin{cases} 1 & \text{if the outcome is Heads,} \\ 0 & \text{if the outcome is Tails.} \end{cases}$$

Here,  $X$  is a random variable because it turns outcomes (Heads/Tails) into numbers (1/0).

**Definition 1.1** (Random Variable). *A random variable is a measurable function*

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R},$$

where

- $\Omega$  is the **sample space** (the set of all possible outcomes),
- $\mathcal{F}$  is a  **$\sigma$ -algebra** of events (the collection of subsets of  $\Omega$  on which probabilities are defined),

- $\mathbb{P}$  is the **probability measure**.

We denote by

$$(X = x) = \{\omega \in \Omega : X(\omega) = x\}$$

for simplicity.

- If  $X$  takes only finitely or countably many values, it is called **discrete**. **Examples:**
  - Flip a coin 4 times. The sample space is

$$\Omega = \{HHHH, HHHT, HHTH, \dots, TTTT\}.$$

Define  $X$  as the number of heads. Then

$$X \in \{0, 1, 2, 3, 4\},$$

which is a finite set.

- Number of arrivals in a queue:

$$X \in \{0, 1, 2, 3, \dots\},$$

which is countable.

- If  $X$  takes values in an interval of  $\mathbb{R}$  with a density, it is called **continuous**. **Example:** The height of students in a classroom can be modeled as a continuous random variable.

**Definition 1.2** (Probability Law). The probability law (or distribution) of a random variable  $X$  is the probability measure  $\mathbb{P}_X$  defined on  $\mathbb{R}$  by

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}),$$

where  $\mathcal{B}(\mathbb{R})$  denotes the Borel  $\sigma$ -algebra of  $\mathbb{R}$ .

Equivalently:

- For a **discrete** random variable  $X$ , the law is described by its **probability mass function (pmf)**:

$$p(x) = \mathbb{P}(X = x), \quad x \in \mathbb{R}.$$

- For a **continuous** random variable  $X$ , the law is described by its **probability density function (pdf)**  $f(x)$ , satisfying

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx, \quad a < b.$$

**Definition 1.3** (Cumulative Distribution Function). The cumulative distribution function (cdf) of  $X$ , denoted  $F$ , is defined for all  $x \in \mathbb{R}$  by

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

Thus,  $F(x)$  gives the probability that  $X$  takes a value less than or equal to  $x$ .



**Notation:** We will write  $X \sim F$  to mean that  $F$  is the distribution function of  $X$ . All probability questions about  $X$  can be answered in terms of its cdf  $F$ .

**Remark:**

- If  $X$  is discrete with pmf  $p(x)$ , then

$$F(a) = \sum_{x \leq a} p(x).$$

- If  $X$  is continuous with pdf  $f(x)$ , then

$$F(a) = \int_{-\infty}^a f(x) dx.$$

## 1.2 Numerical Characteristics

**Definition 1.4** (Expectation). *The expectation (or mean) of a random variable  $X$  is defined by*

$$\begin{aligned} \mathbb{E}[X] &= \sum_x x p(x), & (\text{discrete case}), \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx, & (\text{continuous case}), \end{aligned}$$

*provided the sum or integral is absolutely convergent.*

**Remark 1.1.** *The expectation represents the “average value” or the “center of gravity” of the distribution. It is also called the first moment. Intuitively, it tells us the long-run average outcome if the experiment is repeated many times.*

*However, expectation alone is not sufficient to fully describe a distribution. For example, two distributions can have the same expectation: one may be unimodal while another is bimodal. In such cases, we need an additional measure to capture the spread of the distribution—this is given by the variance or the standard deviation.*

**Definition 1.5** (Variance and Standard Deviation). *The variance of  $X$  is*

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

*An equivalent formula is*

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

*The variance  $\sigma^2$  is related to the second moment and measures the average squared deviation of  $X$  from its mean  $\mu = \mathbb{E}[X]$ . The standard deviation is defined as*

$$\text{SD}(X) = \sqrt{\text{Var}(X)},$$

*and represents the spread of the distribution in the same units as  $X$ .*

**Definition 1.6** (Higher Moments). *The  $k$ -th (raw) moment of  $X$  is*

$$\mu'_k = \mathbb{E}[X^k].$$

*The  $k$ -th central moment is*

$$\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^k].$$

**Remark 1.2.** Taken together, higher-order moments serve as a kind of “fingerprint” of the distribution. Two important shape measures are:

- **Skewness:**  $\gamma_1 = \frac{\mu_3}{\sigma^3}$ , which measures the asymmetry of the distribution.
- **Kurtosis:**  $\gamma_2 = \frac{\mu_4}{\sigma^4}$ , which measures the concentration of mass and the heaviness of the tails.

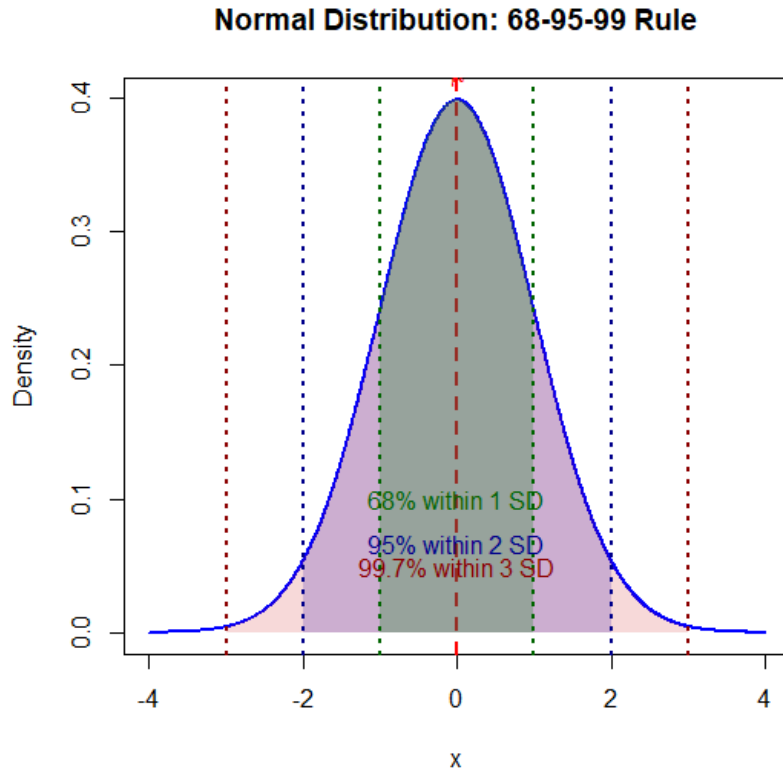


Figure 1.1: Normal distribution curve with mean  $\mu$  and variance  $\sigma^2$ . Shaded areas (e.g.,  $\mu \pm \sigma$ ) correspond to probabilities such as  $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$ ,  $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$ , and  $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$ .

**Remark 1.3** (68–95–99 Rule for the Normal Distribution). Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Then:

$$\begin{aligned}\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 0.68, \\ \mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 0.95, \\ \mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 0.997.\end{aligned}$$

**Example 1.2.** Suppose the heights of students are Normally distributed with mean  $\mu = 170$  cm and standard deviation  $\sigma = 10$  cm.

- About 68% of students are between 160 cm and 180 cm.
- About 95% of students are between 150 cm and 190 cm.
- Almost all (99.7%) students are between 140 cm and 200 cm.

## 1.3 Main Probability Distributions

### Discrete Distributions

- **Bernoulli distribution:**  $X \sim \text{Bernoulli}(p)$ , with

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p.$$

Expectation:  $\mathbb{E}[X] = p$ , Variance:  $\text{Var}(X) = p(1 - p)$ . A Bernoulli random variable  $X$  can take two values: 0 (event does not happen) or 1 (event happens). Examples: coin flip (Head = 1, Tail = 0), dice roll (six = 1, otherwise = 0).

- **Binomial distribution:**  $X \sim \text{Bin}(n, p)$ , with

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, \dots, n.$$

Expectation:  $\mathbb{E}[X] = np$ , Variance:  $\text{Var}(X) = np(1 - p)$ . If  $n$  independent Bernoulli trials are performed ( $B_1, \dots, B_n$ ), then  $X = B_1 + \dots + B_n \sim \text{Bin}(n, p)$ .

- **Poisson distribution:**  $X \sim \text{Poisson}(\lambda)$ , with

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Expectation:  $\mathbb{E}[X] = \lambda$ , Variance:  $\text{Var}(X) = \lambda$ . The Poisson distribution can be obtained as an approximation of the Binomial distribution when  $n$  is large and  $p$  is small. It is extremely useful for modeling rare events such as accidents, catastrophes, or defective items in a factory.

- **Geometric distribution:** The probability that the first success occurs on the  $n$ -th trial is

$$\mathbb{P}(X = n) = (1 - p)^{n-1} p, \quad n = 1, 2, \dots$$

Example: sequence of Bernoulli trials with success probability  $p$ .

### Continuous Distributions

- **Uniform distribution:**  $X \sim U(a, b)$ , with

$$f(x) = \frac{1}{b - a}, \quad a \leq x \leq b.$$

Expectation:  $\mathbb{E}[X] = \frac{a+b}{2}$ , Variance:  $\text{Var}(X) = \frac{(b-a)^2}{12}$ .

- **Exponential distribution:**  $X \sim \text{Exp}(\lambda)$ , with

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Expectation:  $\mathbb{E}[X] = \frac{1}{\lambda}$ , Variance:  $\text{Var}(X) = \frac{1}{\lambda^2}$ . Models waiting times, lifetimes, or times between rare events. Memoryless property:  $\mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t)$ .

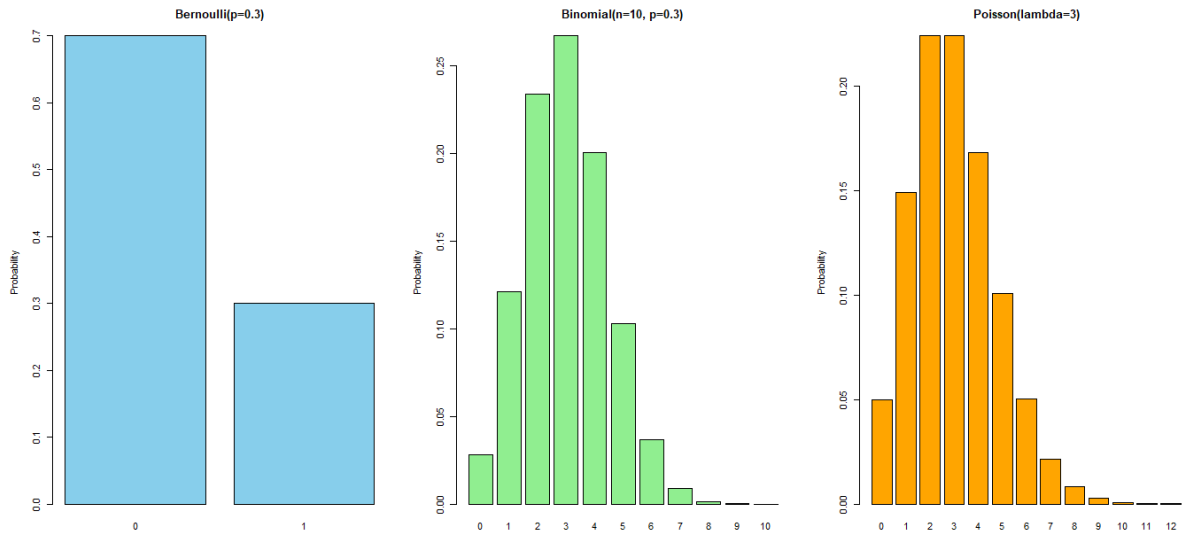


Figure 1.2: Discrete distributions

- **Gamma distribution:**  $X \sim \text{Gamma}(r, \lambda)$ , with

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0,$$

where  $r > 0$  (shape) and  $\lambda > 0$  (rate). Expectation:  $\mathbb{E}[X] = \frac{r}{\lambda}$ , Variance:  $\text{Var}(X) = \frac{r}{\lambda^2}$ .

- **Normal distribution:**  $X \sim \mathcal{N}(\mu, \sigma^2)$ , with

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

Expectation:  $\mathbb{E}[X] = \mu$ , Variance:  $\text{Var}(X) = \sigma^2$ .

- **Chi-square distribution:**  $X \sim \chi_k^2$ , with

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0,$$

where  $k$  is the degrees of freedom. Expectation:  $\mathbb{E}[X] = k$ , Variance:  $\text{Var}(X) = 2k$ .

- **Beta distribution:**  $X \sim \text{Beta}(\alpha, \beta)$ , with

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1,$$

where  $\alpha, \beta > 0$ . Expectation:  $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ , Variance:  $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ . Useful for modeling proportions and probabilities.

- **Cauchy distribution:**  $X \sim \text{Cauchy}(x_0, \gamma)$ , with

$$f(x) = \frac{1}{\pi\gamma} \frac{1}{1 + \left(\frac{x-x_0}{\gamma}\right)^2}, \quad -\infty < x < \infty,$$

where  $x_0$  is the location and  $\gamma > 0$  the scale. Heavy-tailed distribution: expectation and variance are *undefined*. Appears in physics (resonance) and as the ratio of two standard normals.

- **Lognormal distribution:**  $X \sim \text{Lognormal}(\mu, \sigma^2)$ , with

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Expectation:  $\mathbb{E}[X] = e^{\mu + \frac{\sigma^2}{2}}$ , Variance:  $\text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$ . Models skewed positive data such as incomes, stock prices, and survival times.

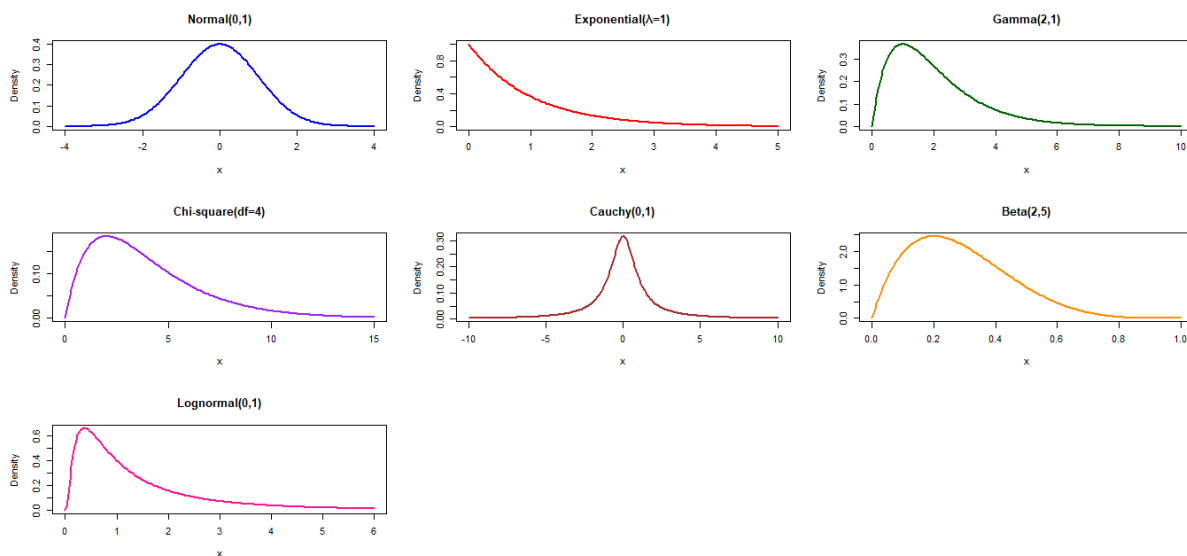


Figure 1.3: Discrete distributions

Distribution	PMF $p(x)$	Support	$\mathbb{E}[X]$	$\text{Var}(X)$	Applications
Bernoulli ( $p$ )	$\begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$	$\{0, 1\}$	$p$	$p(1 - p)$	Single trial: coin toss/failure events
Binomial ( $n, p$ )	$\binom{n}{k} p^k (1 - p)^{n-k}$	$k = 0, 1, \dots, n$	$np$	$np(1 - p)$	Number of successes in independent trials
Geometric ( $p$ )	$(1 - p)^{k-1} p$	$k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	Number of trials until success
Poisson ( $\lambda$ )	$e^{-\lambda} \frac{\lambda^k}{k!}$	$k = 0, 1, 2, \dots$	$\lambda$	$\lambda$	Counts of rare events, defects, arrivals
Negative Binomial ( $r, p$ )	$\binom{k+r-1}{k} (1 - p)^k p^r$	$k = 0, 1, 2, \dots$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	Number of failures before successes
Discrete Uniform ( $a, b$ )	$\frac{1}{b-a+1}$	$a, a + 1, \dots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2 - 1}{12}$	Random integers, etc.

Table 1.1: Summary of common discrete distributions.

Distribution	PDF $f(x)$	Support	$\mathbb{E}[X]$	$\text{Var}(X)$	Applications
Uniform $(a, b)$	$\frac{1}{b-a}$	$a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	Equal probabilities, number generation
Exponential $(\lambda)$	$\lambda e^{-\lambda x}$	$x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Waiting times, Poisson processes
Gamma $(r, \lambda)$	$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$	$x > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	Waiting time for $r$ -t Bayesian stats
Normal $(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x \in \mathbb{R}$	$\mu$	$\sigma^2$	Heights, measurement errors, CLT limit law
Chi-square $(k)$	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	$x > 0$	$k$	$2k$	Hypothesis testing, inference
Beta $(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	Modeling proportions, Bayesian priors
Cauchy $(x_0, \gamma)$	$\frac{1}{\pi\gamma} \frac{1}{1+(\frac{x-x_0}{\gamma})^2}$	$x \in \mathbb{R}$	–	–	Physics (resonance), tailed phenomena
Lognormal $(\mu, \sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$	$x > 0$	$e^{\mu + \frac{\sigma^2}{2}}$	$(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$	Incomes, stock price, survival times

Table 1.2: Summary of common continuous distributions.

### 1.3.1 Operations on Random Variables

**Proposition 1.1** (Linear Transformations). *If  $Y = aX + b$ , then*

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b, \quad \text{Var}(Y) = a^2 \text{Var}(X).$$

**Proposition 1.2** (Independent Random Variables). *If  $X$  and  $Y$  are independent, then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y), \quad \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y].$$

**Example 1.3.** *If  $X \sim \text{Bin}(n_1, p)$  and  $Y \sim \text{Bin}(n_2, p)$  are independent, then*

$$X + Y \sim \text{Bin}(n_1 + n_2, p).$$

**Example 1.4.** *If  $X \sim \text{Poisson}(\lambda_1)$  and  $Y \sim \text{Poisson}(\lambda_2)$  are independent, then*

$$X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

### Functions of a Random Variable

Let  $X$  be a random variable with pdf  $f_X$  and cdf  $F_X(x) = P(X \leq x)$ , and let  $Y = g(X)$ .

If  $g$  is a strictly monotone differentiable function, then the pdf of  $Y$  is given by

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|.$$

**Example 1.5** (Celsius to Fahrenheit). Suppose the temperature in Celsius follows  $X \sim \mathcal{N}(\mu, \sigma^2)$ , and we define

$$Y = g(X) = aX + b, \quad (a = 1.8, b = 32).$$

Then

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

**Remark 1.4** (Standardization). If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

This allows us to compute probabilities using the standard normal cdf  $\Phi$ :

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

# Chapter 2

## Moment Generating Function (MGF)

The **moment generating function (MGF)** is a powerful tool that allows us to compute moments such as the expectation, variance, and higher-order moments (e.g., skewness, kurtosis) through differentiation. Moreover, MGFs play a central role in probability theory and statistics, particularly in proving limit theorems such as the Central Limit Theorem (CLT).

Intuitively, the collection of moments uniquely characterizes the distribution in many cases, in the same way that the derivatives of a function characterize it in a Taylor series expansion. Thus, the MGF can be viewed as a “Taylor series for probability distributions”, making it a fundamental and very useful concept.

**Definition 2.1** (Moment Generating Function). *Let  $X$  be a random variable. The **moment generating function (MGF)** of  $X$  is defined as*

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R} \text{ such that } \mathbb{E}[e^{tX}] < \infty.$$

For different types of random variables:

$$M_X(t) = \begin{cases} \sum_x e^{tx} p(x), & \text{if } X \text{ is discrete with pmf } p(x), \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx, & \text{if } X \text{ is continuous with pdf } f(x). \end{cases}$$

Here, the parameter  $t$  serves as a transform variable. In fact, the MGF is closely related to the *Laplace transform* of the probability density function.

**Theorem 2.1** (Moments from the MGF). *If the MGF  $M_X(t)$  exists in an open interval containing  $t = 0$ , then the  $n$ -th derivative of  $M_X(t)$  at  $t = 0$  yields the  $n$ -th moment of  $X$ :*

$$M_X^{(n)}(0) = \mathbb{E}[X^n], \quad n = 1, 2, \dots$$

*Proof.* Starting from the definition,

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right].$$

By exchanging the expectation and the series (justified if  $M_X(t)$  exists in a neighborhood of 0),

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}[X^n].$$



This is a power series expansion where the coefficients are precisely the moments of  $X$ . Differentiating  $n$  times and evaluating at  $t = 0$  isolates  $\mathbb{E}[X^n]$ .  $\square$

## 2.1 Monotone Convergence Theorem (MCT)

The Monotone Convergence Theorem is one of the fundamental results in probability and integration theory. It provides the conditions under which we can interchange a limit and an expectation (or integral).

**Theorem 2.2** (Monotone Convergence Theorem). *Let  $(X_n)_{n \geq 1}$  be a sequence of non-negative random variables such that*

$$X_1(\omega) \leq X_2(\omega) \leq \dots \leq X(\omega) \quad \text{for all } \omega,$$

*and suppose that  $X_n(\omega) \uparrow X(\omega)$  as  $n \rightarrow \infty$ . Then,*

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}\left[\lim_{n \rightarrow \infty} X_n\right] = \mathbb{E}[X].$$

**Remark 2.1.** *This theorem guarantees that when we have an **increasing sequence of non-negative random variables**, we may safely move the limit inside the expectation. This property is essential in probability theory — it allows us to justify steps like interchanging expectations and infinite sums or limits, as we did in the proof of the MGF theorem.*

**Remark 2.2** (Application of MCT in the MGF proof). *In the proof of the previous theorem, we wrote:*

$$M_X(t) = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}[X^n].$$

*Formally, this means that we interchanged the expectation and the infinite sum. To justify this rigorously, we use the **Monotone Convergence Theorem (MCT)**.*

*Indeed, for  $t > 0$  and  $X \geq 0$ , the partial sums*

$$S_N = \sum_{n=0}^N \frac{(tX)^n}{n!}$$

*form an increasing sequence of nonnegative random variables:*

$$S_1 \leq S_2 \leq \dots \leq e^{tX}.$$

*Therefore, by the Monotone Convergence Theorem,*

$$\mathbb{E}\left[\lim_{N \rightarrow \infty} S_N\right] = \lim_{N \rightarrow \infty} \mathbb{E}[S_N].$$

*This justifies the interchange between the limit (or the infinite sum) and the expectation.*

*For a general random variable  $X$  (not necessarily nonnegative), we can apply the same reasoning separately to  $X^+ = \max(X, 0)$  and  $X^- = \max(-X, 0)$ .*

**Example 2.1.** *Let  $X_n = \min(X, n)$  for a non-negative random variable  $X$ . Then  $X_n \uparrow X$  as  $n \rightarrow \infty$ . By the MCT,*

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

*Thus, the theorem allows us to compute  $\mathbb{E}[X]$  as the limit of expectations of the truncated variables.*

## 2.2 Dominated Convergence Theorem (DCT)

The Monotone Convergence Theorem applies only to increasing sequences of non-negative random variables. The Dominated Convergence Theorem is a more general and powerful result that works for bounded (not necessarily monotone) sequences.

**Theorem 2.3** (Dominated Convergence Theorem). *Let  $(X_n)_{n \geq 1}$  be a sequence of random variables that converges almost surely to a random variable  $X$ . Assume there exists an integrable random variable  $Y$  such that*

$$|X_n(\omega)| \leq Y(\omega) \quad \text{for all } n \text{ and all } \omega.$$

Then:

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

**Remark 2.3.** *The function  $Y$  is called a **dominating function**. It provides a uniform bound ensuring that all  $X_n$  are “controlled” in magnitude. This theorem justifies interchanging limits and expectations when the sequence is dominated by an integrable bound.*

**Example 2.2.** *Let  $X_n = X \mathbf{1}_{\{|X| \leq n\}}$ . Then  $X_n \rightarrow X$  almost surely and  $|X_n| \leq |X|$ . If  $\mathbb{E}[|X|] < \infty$ , then by the DCT:*

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

*This means we can approximate an integrable random variable by truncated versions without changing its expectation in the limit.*

## 2.3 Connection to Convergence in $L^1$

The DCT naturally leads to the concept of **convergence in  $L^1$**  (mean convergence).

**Definition 2.2** (Convergence in  $L^1$ ). *A sequence of random variables  $(X_n)$  converges to  $X$  in  $L^1$  if*

$$\mathbb{E}[|X_n - X|] \rightarrow 0.$$

**Proposition 2.1.** *If  $X_n \rightarrow X$  almost surely and there exists an integrable random variable  $Y$  such that  $|X_n| \leq Y$  for all  $n$ , then  $X_n \rightarrow X$  in  $L^1$ . In particular,*

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

**Remark 2.4.** *The Dominated Convergence Theorem is the key tool that bridges almost sure convergence and convergence in  $L^1$ . It ensures not only pointwise convergence but also convergence of expectations.*

### 2.3.1 Examples of MGFs

**Example 2.3** (Bernoulli( $p$ )). *If  $X \sim \text{Bernoulli}(p)$ , then*

$$M_X(t) = \mathbb{E}[e^{tX}] = (1 - p) + pe^t.$$

**Example 2.4** (Poisson( $\lambda$ )). *If  $X \sim \text{Poisson}(\lambda)$ , then*

$$M_X(t) = \exp(\lambda(e^t - 1)).$$

**Example 2.5** (Standard Normal  $N(0, 1)$ ). Let  $X \sim \mathcal{N}(0, 1)$ . Its moment generating function is

$$M_X(t) = \mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx.$$

Completing the square in the exponent:

$$tx - \frac{1}{2}x^2 = -\frac{1}{2}(x^2 - 2tx) = -\frac{1}{2}((x-t)^2 - t^2),$$

so

$$M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} dx e^{t^2/2}.$$

Since the integral is 1 (it is the pdf of a  $\mathcal{N}(t, 1)$ ), we obtain

$$M_X(t) = e^{t^2/2}, \quad t \in \mathbb{R}.$$

### 2.3.2 Properties of MGFs

**Proposition 2.2.** Let  $X$  and  $Y$  be random variables. The moment generating function (MGF) satisfies the following properties:

- 1) **Normalization:**  $M_X(0) = 1$ .
- 2) **Uniqueness:** If  $M_X(t) = M_Y(t)$  for all  $t$  in a neighborhood of 0, then  $X$  and  $Y$  have the same distribution.
- 3) **Linear transformation:** If  $Y = aX + b$ , then

$$M_Y(t) = e^{bt} M_X(at).$$

- 4) **Sum of independent variables:** If  $X$  and  $Y$  are independent and their MGFs exist (finite) for  $t$  in some neighborhood of 0, then

$$M_{X+Y}(t) = M_X(t) M_Y(t), \quad \text{for such } t.$$

*Proof.* By definition,

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} e^{tY}].$$

Since  $X$  and  $Y$  are independent, the random variables  $e^{tX}$  and  $e^{tY}$  are also independent. Hence, the expectation of their product equals the product of their expectations:

$$\mathbb{E}[e^{tX} e^{tY}] = \mathbb{E}[e^{tX}] \mathbb{E}[e^{tY}] = M_X(t) M_Y(t).$$

This equality holds for all  $t$  in the neighborhood where both MGFs are finite. □

## 2.4 Characteristic Function (CF)

The basic idea is to transform a random variable into a new space. Each random variable has its own unique CF, and this mapping is one-to-one: if two random variables  $X_1$  and  $X_2$  have the same CF, then they have the same distribution.

**Definition 2.3.** The **characteristic function (CF)** of a random variable  $X$  is defined by

$$\varphi_X(t) = \mathbb{E}[e^{itX}], \quad t \in \mathbb{R}.$$

Explicitly,

$$\varphi_X(t) = \begin{cases} \sum_x e^{itx} p(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{itx} f(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

**Remark 2.5.** Unlike the MGF, the characteristic function always exists, since

$$|e^{itX}| = 1 \quad \text{for all } X, t.$$

**Remark 2.6.** The CF is the Fourier transform of the probability distribution (or density). Conversely, the distribution can be recovered from the CF via the inverse Fourier transform. This duality makes CFs a central tool in probability theory.

## 2.4.1 Examples of CFs

**Example 2.6** (Normal Distribution). If  $X \sim \mathcal{N}(0, 1)$ , then

$$\varphi_X(t) = \exp\left(-\frac{1}{2}t^2\right).$$

**Example 2.7** (Characteristic function of an Exponential( $\lambda$ )). Let  $X \sim \text{Exp}(\lambda)$  with density  $f_X(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$  and  $\lambda > 0$ . The characteristic function is

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_0^{\infty} e^{itx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda-it)x} dx.$$

Since  $\Re(\lambda - it) = \lambda > 0$ , the integral converges and we obtain

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}, \quad t \in \mathbb{R}.$$

Equivalently, by multiplying numerator and denominator by  $\lambda + it$ ,

$$\varphi_X(t) = \frac{\lambda(\lambda + it)}{\lambda^2 + t^2} = \frac{\lambda^2}{\lambda^2 + t^2} + i \frac{\lambda t}{\lambda^2 + t^2}.$$

Thus the real part is  $\Re\varphi_X(t) = \frac{\lambda^2}{\lambda^2 + t^2}$  and the imaginary part is  $\Im\varphi_X(t) = \frac{\lambda t}{\lambda^2 + t^2}$ .

**Moments from derivatives.** Using the derivative relation  $\mathbb{E}[X^k] = \frac{1}{i^k} \varphi_X^{(k)}(0)$ , we get

$$\varphi_X(t) = \lambda(\lambda - it)^{-1}, \quad \varphi_X'(t) = \lambda i(\lambda - it)^{-2},$$

hence

$$\varphi_X'(0) = \frac{i}{\lambda} \quad \Rightarrow \quad \mathbb{E}[X] = \frac{1}{i} \varphi_X'(0) = \frac{1}{\lambda}.$$

Also

$$\varphi_X''(t) = -2\lambda(\lambda - it)^{-3}, \quad \varphi_X''(0) = -\frac{2}{\lambda^2},$$

so

$$\mathbb{E}[X^2] = \frac{1}{i^2} \varphi_X''(0) = -\varphi_X''(0) = \frac{2}{\lambda^2},$$

and therefore

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2},$$

as expected.

**Remarks.**

- The CF  $\varphi_X(t) = \frac{\lambda}{\lambda - it}$  exists for all real  $t$  (this is a general advantage of CFs over MGFs).
- The MGF  $M_X(s) = \mathbb{E}[e^{sX}]$  exists only for  $s < \lambda$  and satisfies  $M_X(s) = \varphi_X(-is) = \frac{\lambda}{\lambda - s}$  for  $s < \lambda$ .

### 2.4.2 Properties of CFs

The CF encodes all the moments of a distribution (when they exist). Expanding  $e^{itx}$  into a Taylor series and interchanging sum/integral with expectation gives

$$\varphi_X(t) = 1 + it \mathbb{E}[X] - \frac{t^2}{2} \mathbb{E}[X^2] + \dots + \frac{(it)^k}{k!} \mathbb{E}[X^k] + \dots$$

Hence, the derivatives of  $\varphi_X$  at  $t = 0$  recover the moments:

$$\mathbb{E}[X^k] = \frac{1}{i^k} \varphi_X^{(k)}(0).$$

**Proposition 2.3.** *Let  $X$  be a random variable with CF  $\varphi_X(t)$ . Then:*

- (i)  $\varphi_X(0) = 1$ .
- (ii)  $|\varphi_X(t)| \leq 1$  for all  $t$ .
- (iii) If  $X$  and  $Y$  are independent, then  $\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t)$ .
- (iv) The distribution of  $X$  is uniquely determined by  $\varphi_X(t)$ .
- (v)  $\varphi_X(-t) = \overline{\varphi_X(t)}$  (complex conjugate symmetry).

# Chapter 3

## Modes of Convergence

### 3.1 Introduction

In probability theory, various notions of convergence are used to describe how a sequence of random variables  $(X_n)_{n \geq 1}$  behaves as it approaches a limiting random variable  $X$ . When we say that  $X_n$  converges to  $X$ , we mean that the sequence  $(X_n)$  becomes “closer and closer” to  $X$  in some sense. However, the notion of “closeness” can be defined in several different ways, leading to different **types of convergence**. Understanding these distinctions is crucial for analyzing the asymptotic behavior of random variables.

The most common modes of convergence are:

- Almost sure convergence,
- Convergence in probability,
- Convergence in  $L^p$ ,
- Convergence in distribution.

Before introducing these types of convergence, we will first present several **fundamental inequalities and lemmas**, which serve as essential tools for proving limit theorems in probability.

### 3.2 Fundamental Inequalities

The main idea behind these inequalities is to use partial information about a distribution—such as its mean or variance—to control the probabilities of extreme events, that is, the likelihood that a random variable takes unusually large or small values.

In many situations, we may not have enough information to compute a desired quantity exactly, such as the probability of an event or the expected value of a random variable. In other cases, the problem may be too complex for an exact calculation. These inequalities therefore provide general and powerful tools that allow us to obtain useful bounds and approximations, even when an explicit solution is not possible.

### 3.2.1 Markov's Inequality

**Proposition 3.1** (Markov's Inequality). *Let  $X$  be a non-negative random variable and  $a > 0$ . Then*

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

#### Intuition

If  $\mathbb{E}[X]$  is small, then the probability that  $X$  exceeds a large value  $a$  must also be small. Moreover, as  $a$  increases, the ratio  $\mathbb{E}[X]/a$  decreases, meaning that very large deviations are increasingly unlikely.

*Idea of the proof.* For a discrete random variable:

$$\mathbb{E}[X] = \sum_x x \mathbb{P}(X = x) \geq \sum_{x \geq a} x \mathbb{P}(X = x) \geq a \sum_{x \geq a} \mathbb{P}(X = x) = a \mathbb{P}(X \geq a).$$

Dividing both sides by  $a$  gives the result. □

**Example 3.1.** *If  $\mathbb{E}[X] = 10$ , then*

$$\mathbb{P}(X > 20) \leq \frac{10}{20} = \frac{1}{2}.$$

**Example 3.2.** *Let  $X \sim \text{Binomial}(n, p)$ . Using Markov's inequality, find an upper bound for*

$$\mathbb{P}(X \geq \alpha n),$$

*where  $0 < p < \alpha < 1$ . Evaluate the bound for  $p = \frac{1}{2}$  and  $\alpha = \frac{3}{4}$ .*

*Proof.* Since  $X \geq 0$ , we may apply Markov's inequality to the nonnegative random variable  $X$  (or to  $X/n$ ). Note that

$$\mathbb{P}(X \geq \alpha n) = \mathbb{P}\left(\frac{X}{n} \geq \alpha\right).$$

Markov's inequality gives, for any nonnegative r.v.  $Y$  and  $a > 0$ ,  $\mathbb{P}(Y \geq a) \leq \mathbb{E}[Y]/a$ . Taking  $Y = X/n$  and  $a = \alpha$ ,

$$\mathbb{P}(X \geq \alpha n) \leq \frac{\mathbb{E}[X/n]}{\alpha} = \frac{\mathbb{E}[X]}{n\alpha}.$$

For a binomial  $X \sim \text{Binomial}(n, p)$  we have  $\mathbb{E}[X] = np$ , so

$$\boxed{\mathbb{P}(X \geq \alpha n) \leq \frac{p}{\alpha}}.$$

**Numerical example.** For  $p = \frac{1}{2}$  and  $\alpha = \frac{3}{4}$ ,

$$\mathbb{P}(X \geq \frac{3}{4}n) \leq \frac{1/2}{3/4} = \frac{2}{3} \approx 0.6667.$$

□

### 3.2.2 Chebyshev's Inequality

Chebyshev's inequality is a direct application of Markov's inequality to the squared deviation from the mean.

**Proposition 3.2** (Chebyshev's Inequality). *Let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $a > 0$ ,*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

*Idea of the proof.* Let  $Y = (X - \mu)^2$  and  $b = a^2$ . Then

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}(Y \geq b) \leq \frac{\mathbb{E}[Y]}{b} = \frac{\sigma^2}{a^2},$$

where we have applied Markov's inequality to  $Y$ . □

#### Interpretation

If the variance  $\sigma^2$  is small (meaning that  $X$  exhibits little randomness), then the random variable  $X$  is unlikely to deviate much from its mean  $\mu$ .

Chebyshev's inequality formally quantifies this idea: it shows that the probability of  $X$  being far from its mean is limited by its variance. In other words, the smaller the variance, the more concentrated  $X$  is around  $\mu$ . This matches our intuition that the variance measures, on average, how far observations tend to lie from the mean.

**Example 3.3.** *By Chebyshev's inequality,*

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

*For  $k = 3$ , this means  $\mathbb{P}(|X - \mu| \geq 3\sigma) \leq \frac{1}{9}$  — regardless of the distribution of  $X$ .*

**Example 3.4.** *Let us continue example 3.2. Chebyshev's inequality gives a bound that improves with  $n$ :*

$$\mathbb{P}(X \geq \alpha n) = \mathbb{P}(X - np \geq n(\alpha - p)) \leq \frac{\text{Var}(X)}{n^2(\alpha - p)^2} = \frac{np(1-p)}{n^2(\alpha - p)^2} = \frac{p(1-p)}{n(\alpha - p)^2},$$

*which tends to 0 as  $n \rightarrow \infty$  (provided  $\alpha > p$ ). Thus for large  $n$  Chebyshev (or Chernoff bounds) give much sharper control than Markov's inequality.*

### 3.2.3 Cauchy–Schwarz Inequality

**Theorem 3.1** (Cauchy–Schwarz Inequality). *For any two random variables  $X$  and  $Y$  with finite second moments, we have*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \mathbb{E}[Y^2]}.$$

*Proof.* Consider the random variable  $Z = X - \lambda Y$ , where  $\lambda$  is a real number. Since  $\mathbb{E}[Z^2] \geq 0$ , we have

$$\mathbb{E}[(X - \lambda Y)^2] = \mathbb{E}[X^2] - 2\lambda\mathbb{E}[XY] + \lambda^2\mathbb{E}[Y^2] \geq 0.$$



This is a quadratic inequality in  $\lambda$ . For it to hold for all  $\lambda$ , its discriminant must be non-positive:

$$(-2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0.$$

Simplifying gives

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Taking the square root on both sides yields the desired result.  $\square$

#### Interpretation

The Cauchy–Schwarz inequality expresses a fundamental relationship between two random variables. It states that the absolute value of their covariance (or correlation) cannot exceed the product of their standard deviations.

Geometrically, it is analogous to the angle between two vectors not being smaller than zero — it ensures that the “inner product” between  $X$  and  $Y$  is bounded by their lengths in the  $L^2$  space. Equality holds if and only if  $X$  and  $Y$  are linearly dependent (i.e.,  $X = aY$  for some constant  $a$ ).

### 3.2.4 Jensen’s Inequality

Jensen’s inequality compares  $\mathbb{E}[g(X)]$  and  $g(\mathbb{E}[X])$  when  $g$  is a convex function.

**Proposition 3.3** (Jensen’s Inequality). *Let  $X$  be an integrable random variable and  $g$  a convex function. Then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

#### Intuition

If  $g$  is linear, then  $\mathbb{E}[g(X)] = g(\mathbb{E}[X])$ . For convex  $g$ , the average of the transformed values is at least as large as the transformation of the average.

*Sketch of the proof.* If  $g$  is convex, then for any  $c$  and  $x$ ,

$$g(x) \geq g(c) + g'(c)(x - c).$$

Taking  $c = \mathbb{E}[X]$  and applying the expectation gives

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) + g'(\mathbb{E}[X])\mathbb{E}[X - \mathbb{E}[X]] = g(\mathbb{E}[X]).$$

$\square$

**Example 3.5.** • With  $g(x) = x^2$ , we get  $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ , confirming that variance is non-negative.

- With  $g(x) = x^4$ , we obtain  $\mathbb{E}[X^4] \geq (\mathbb{E}[X])^4$ .
- With  $g(x) = -\log x$  (a convex function), we have  $-\log(\mathbb{E}[X]) \leq \mathbb{E}[-\log X]$ , or equivalently:

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X].$$

To apply Jensen's inequality, we first need to determine whether the function  $g$  is convex. A practical way to check convexity is by using the second derivative.

**Criterion for Convexity:** A twice-differentiable function  $g : I \rightarrow \mathbb{R}$  is convex on the interval  $I$  if and only if

$$g''(x) \geq 0 \quad \text{for all } x \in I.$$

**Example 3.6.** Let  $X$  be a positive random variable. Compare  $\mathbb{E}[X^a]$  with  $(\mathbb{E}[X])^a$  for all real values of  $a \in \mathbb{R}$ .

**Solution 3.1.** We apply Jensen's inequality with  $g(x) = x^a$ , defined for  $x > 0$ .

- If  $a \geq 1$  or  $a \leq 0$ , then  $g''(x) = a(a-1)x^{a-2} \geq 0$ , so  $g$  is convex. By Jensen's inequality:

$$\mathbb{E}[X^a] \geq (\mathbb{E}[X])^a.$$

- If  $0 < a < 1$ , then  $g''(x) = a(a-1)x^{a-2} \leq 0$ , so  $g$  is concave. Therefore, the inequality is reversed:

$$\mathbb{E}[X^a] \leq (\mathbb{E}[X])^a.$$

### Interpretation

This example illustrates how the sign of the second derivative determines the direction of Jensen's inequality. When the function  $x^a$  is convex ( $a \geq 1$  or  $a \leq 0$ ), the expectation of the power is larger. When it is concave ( $0 < a < 1$ ), the inequality is reversed.

## 3.3 Modes of Convergence

In this section, we discuss what it means for a sequence of random variables to converge. Recall that, in any probability model, we have a sample space  $\Omega$  and a probability measure  $\mathbb{P}$ . For simplicity, suppose that  $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$  is finite. A random variable  $X$  is a mapping that assigns a real number to each possible outcome  $\omega_i$ , i.e.,  $X(\omega_i) = x_i$ .

When we have a sequence of random variables  $X_1, X_2, X_3, \dots$ , it is important to remember that each  $X_n$  is a function

$$X_n : \Omega \rightarrow \mathbb{R}.$$

Hence, for each  $\omega_i \in \Omega$ , we have a numerical sequence  $X_n(\omega_i) = x_{ni}$ . After the random experiment is performed and one outcome  $\omega_j$  occurs, we observe the real sequence

$$X_1(\omega_j), X_2(\omega_j), X_3(\omega_j), \dots$$

and can ask whether it converges, and if so, to what limit.

**Definition 3.1** (Almost Sure Convergence). A sequence of random variables  $(X_n)$  is said to converge **almost surely** (a.s.) to a random variable  $X$  if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1,$$

that is, if

$$\mathbb{P}\left(\{\omega \in \Omega : X_n(\omega) \rightarrow X(\omega)\}\right) = 1.$$

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

—

**Example 3.7.** Let  $X$  be a real random variable and  $(a_n)$  a sequence of real numbers such that  $a_n \rightarrow 0$  as  $n \rightarrow \infty$ . Define

$$X_n = a_n X.$$

**Claim:**  $X_n \xrightarrow{a.s.} 0$ .

**Proof.** For each fixed  $\omega \in \Omega$ , we have

$$X_n(\omega) = a_n X(\omega).$$

Since  $a_n \rightarrow 0$  (a deterministic limit) and  $X(\omega)$  is finite for each  $\omega$ , the product  $a_n X(\omega) \rightarrow 0$ . Thus, for every  $\omega \in \Omega$ ,  $X_n(\omega) \rightarrow 0$ . Hence,

$$\mathbb{P}(\{\omega : X_n(\omega) \rightarrow 0\}) = 1,$$

which means  $X_n \xrightarrow{a.s.} 0$ . □

—

**Theorem 3.2** (Sufficient Condition for Almost Sure Convergence). Let  $(X_n)$  be a sequence of random variables and  $X$  a random variable. If for all  $\varepsilon > 0$ ,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty,$$

then  $X_n \xrightarrow{a.s.} X$ .

**Remark.** This result is a direct application of the **\*\*Borel–Cantelli Lemma\*\***, which states that if  $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ , then  $\mathbb{P}(A_n \text{ infinitely often}) = 0$ .

—

**Example 3.8** (Illustration of the Theorem). Let  $X_n = \frac{Z}{n}$ , where  $Z$  is any integrable random variable.

We show that  $X_n \xrightarrow{a.s.} 0$  using the theorem.

For any  $\varepsilon > 0$ , we have

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(|Z| > n\varepsilon).$$

Since  $\mathbb{P}(|Z| > n\varepsilon)$  decreases to 0 as  $n \rightarrow \infty$  and

$$\sum_{n=1}^{\infty} \mathbb{P}(|Z| > n\varepsilon) < \infty$$

(because  $\mathbb{E}|Z| < \infty$  implies this tail sum is finite by Markov's inequality), the condition of the theorem holds. Hence, by the theorem, we conclude that

$$X_n = \frac{Z}{n} \xrightarrow{a.s.} 0.$$

□

—

**Summary.** Almost sure convergence is the strongest mode of convergence for random variables. It requires that the sequence of values  $X_n(\omega)$  converges pointwise for all  $\omega$  in a set of probability one.

### 3.3.1 Convergence in Probability

**Definition 3.2** (Convergence in Probability). A sequence of random variables  $(X_n)$  is said to **converge in probability** to a random variable  $X$ , written

$$X_n \xrightarrow{\mathbb{P}} X,$$

if for every  $\varepsilon > 0$ ,

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

#### Intuition

Convergence in probability means that as  $n$  increases, the random variable  $X_n$  becomes *closer to  $X$  with high probability*. Small deviations may still occur, but they become increasingly rare as  $n$  grows.

To better understand this, recall the ordinary convergence of a sequence of real numbers  $(a_n)$ :

$$a_n \rightarrow a \iff \forall \varepsilon > 0, \exists n_0 \text{ such that } n \geq n_0 \Rightarrow |a_n - a| \leq \varepsilon.$$

In the probabilistic setting, we replace numbers with random variables. For convergence in probability, we require that for any fixed  $\varepsilon > 0$ , the probability of large deviations,  $\mathbb{P}(|Y_n - a| \geq \varepsilon)$ , becomes smaller and smaller as  $n$  increases. Geometrically, the mass of the distribution of  $Y_n$  concentrates more and more tightly around  $a$ .

**Remark 3.1** (Properties). *Convergence in probability shares similar properties with ordinary convergence. If  $X_n \xrightarrow{\mathbb{P}} a$  and  $Y_n \xrightarrow{\mathbb{P}} b$ , then:*

- If  $g$  is continuous, then  $g(X_n) \xrightarrow{\mathbb{P}} g(a)$  (e.g., if  $X_n \xrightarrow{\mathbb{P}} a$ , then  $X_n^2 \xrightarrow{\mathbb{P}} a^2$ ).
- $X_n + Y_n \xrightarrow{\mathbb{P}} a + b$ .
- $cX_n \xrightarrow{\mathbb{P}} ca$  for any constant  $c$ .

**Example 3.9** (Simple Discrete Case). *To show convergence in probability, we usually:*

1. Guess the candidate limit.
2. Compute or bound  $\Pr(|Y_n - Y| \geq \varepsilon)$ .
3. Show that it tends to 0 as  $n \rightarrow \infty$ .

Let

$$Y_n = \begin{cases} 0, & \text{with probability } 1 - \frac{1}{n}, \\ n^2, & \text{with probability } \frac{1}{n}. \end{cases}$$

Fix  $\varepsilon > 0$ . Then

$$\Pr(|Y_n - 0| \geq \varepsilon) = \Pr(Y_n = n^2) = \frac{1}{n} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Hence,  $Y_n \xrightarrow{\mathbb{P}} 0$ .

**Example 3.10.** Let  $X_n = X + \frac{1}{n}$ , where  $X$  is a random variable. Then

$$\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}\left(\frac{1}{n} \geq \varepsilon\right) = 0, \quad \text{for all } n > \frac{1}{\varepsilon}.$$

Thus,  $X_n \xrightarrow{\mathbb{P}} X$ .

**Example 3.11.** Let  $X_n \sim \text{Exponential}(n)$ . We show that  $X_n \xrightarrow{p} 0$ . Recall that for an exponential random variable with rate  $n$ ,

$$\Pr(X_n > \varepsilon) = e^{-n\varepsilon}.$$

Then,

$$\Pr(|X_n - 0| \geq \varepsilon) = \Pr(X_n > \varepsilon) = e^{-n\varepsilon} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Therefore,  $X_n \xrightarrow{p} 0$ .

**Example 3.12.** Let  $X$  be a random variable, and let  $X_n = X + Y_n$ , where

$$\mathbb{E}[Y_n] = \frac{1}{n}, \quad \text{Var}(Y_n) = \frac{\sigma^2}{n},$$

for some constant  $\sigma > 0$ . By Chebyshev's inequality,

$$\Pr(|Y_n - \frac{1}{n}| \geq \varepsilon) \leq \frac{\text{Var}(Y_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0.$$

Hence,  $Y_n \xrightarrow{p} 0$ , and therefore  $X_n = X + Y_n \xrightarrow{p} X$ .

### 3.3.2 Convergence in $L^p$

**Definition 3.3** (Convergence in  $L^p$ ). Let  $p \geq 1$ . A sequence of random variables  $(X_n)$  is said to **converge in  $L^p$**  to  $X$  if

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

We write  $X_n \xrightarrow{L^p} X$ .

#### Intuition

Convergence in  $L^p$  means that the *expected distance* between  $X_n$  and  $X$ , measured in the  $p$ -norm, becomes arbitrarily small. In other words, on average,  $X_n$  gets very close to  $X$ , and large deviations contribute less and less to the expectation.

One way to interpret convergence is in terms of the “distance” between random variables. For instance:

- If we define the distance as  $\mathbb{P}(|X_n - X| \geq \varepsilon)$ , we obtain **convergence in probability**.
- If we define the distance as  $\mathbb{E}(|X_n - X|^p)$ , we obtain **convergence in  $L^p$**  (also called *convergence in mean of order  $p$* ).

The most common special cases are:

- $p = 1$ : convergence in mean (or *mean absolute convergence*);
- $p = 2$ : convergence in mean square (or *mean-square convergence*).

**Example 3.13.** Let  $X_n = \frac{1}{n}X$ , where  $\mathbb{E}[|X|^p] < \infty$ . Then

$$\mathbb{E}[|X_n - 0|^p] = \frac{1}{n^p} \mathbb{E}[|X|^p] \rightarrow 0,$$

so  $X_n \xrightarrow{L^p} 0$ .

**Example 3.14.** Let  $X_n \sim \text{Uniform}(0, \frac{1}{n})$ . We show that  $X_n \xrightarrow{L^p} 0$  for any  $p \geq 1$ .

Since  $\mathbb{E}[|X_n|^p] = \frac{1}{(p+1)n^p}$ , it follows that

$$\mathbb{E}[|X_n - 0|^p] = \frac{1}{(p+1)n^p} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and hence  $X_n \xrightarrow{L^p} 0$ .

**Theorem 3.3.** Let  $1 \leq r \leq s$ . If  $X_n \xrightarrow{L^s} X$ , then  $X_n \xrightarrow{L^r} X$ .

*Idea of proof.* Since  $|x|^r \leq 1 + |x|^s$  for all  $x$ , the sequence  $\{|X_n - X|^r\}$  is dominated by an integrable random variable when  $\mathbb{E}[|X_n - X|^s] < \infty$ . Applying Hölder's inequality and taking limits shows that

$$\mathbb{E}[|X_n - X|^r] \rightarrow 0.$$

□

### 3.3.3 Convergence in Distribution

**Definition 3.4** (Convergence in distribution). A sequence of random variables  $(X_n)$  is said to **converge in distribution** (or **in law**) to a random variable  $X$  if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

for all points  $x$  at which the cumulative distribution function  $F_X$  is continuous. We write this as

$$X_n \xrightarrow{d} X \quad \text{or} \quad X_n \xrightarrow{\mathcal{L}} X.$$

#### Intuition

Convergence in distribution means that the *distributions* of  $X_n$  approach the distribution of  $X$ . It focuses only on the behavior of the CDFs, not on the random variables themselves. This is the weakest form of convergence — it does not require that  $X_n(\omega)$  converges to  $X(\omega)$  for each outcome  $\omega$ .

Convergence in distribution is therefore the weakest mode of convergence. It simply states that the CDF of  $X_n$  converges to the CDF of  $X$  as  $n \rightarrow \infty$ . Unlike convergence in probability or almost sure convergence, it does not require any relationship between  $X_n$  and  $X$  on the same probability space.

**Example 3.15.** Let  $X_n \sim \text{Binomial}(n, \lambda/n)$  for  $n \in \mathbb{N}$ , with  $\lambda > 0$  fixed. Then, as  $n \rightarrow \infty$ ,

$$X_n \xrightarrow{d} \text{Poisson}(\lambda).$$

Indeed, the Binomial distribution with parameters  $(n, \lambda/n)$  approximates the Poisson distribution when  $n$  is large and the success probability  $\lambda/n$  is small.

**Example 3.16.** If  $X_n \sim \mathcal{N}(0, 1/n)$ , then as  $n \rightarrow \infty$ ,

$$X_n \xrightarrow{d} 0,$$

because the variance tends to zero, and the distribution of  $X_n$  becomes increasingly concentrated at 0.

**Example 3.17.** Let  $X_2, X_3, X_4, \dots$  be a sequence of random variables whose cumulative distribution functions are, for each  $n \geq 2$ ,

$$F_{X_n}(x) = \begin{cases} 1 - \left(1 - \frac{1}{n}\right)^{nx}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

Show that  $X_n$  converges in distribution to an Exponential(1) random variable.

*Proof.* Let  $F(x) = 1 - e^{-x}$  ( $x \in \mathbb{R}$ ) be the CDF of  $\text{Exp}(1)$ . We will show that for every fixed  $x \in \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F(x).$$

If  $x \leq 0$ , then  $F_{X_n}(x) = 0$  for every  $n$ , and  $F(x) = 1 - e^{-x} = 1 - e^{-(\leq 0)}$ . In particular  $F(x) = 0$  when  $x \leq 0$ . Hence the convergence holds for  $x \leq 0$ .

Now let  $x > 0$ . Using the identity  $(1 - \frac{1}{n})^{nx} = \exp\left(nx \log\left(1 - \frac{1}{n}\right)\right)$  and the expansion  $\log(1 - t) = -t + o(t)$  as  $t \downarrow 0$ , we get

$$\lim_{n \rightarrow \infty} n \log\left(1 - \frac{1}{n}\right) = \lim_{n \rightarrow \infty} n\left(-\frac{1}{n} + o\left(\frac{1}{n}\right)\right) = -1.$$

Multiplying by  $x$  and exponentiating yields

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^{nx} = \exp\left(\lim_{n \rightarrow \infty} nx \log\left(1 - \frac{1}{n}\right)\right) = \exp(-x) = e^{-x}.$$

Therefore

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = \lim_{n \rightarrow \infty} \left(1 - \left(1 - \frac{1}{n}\right)^{nx}\right) = 1 - e^{-x} = F(x).$$

Since the limiting CDF  $F$  is continuous for every  $x \in \mathbb{R}$ , pointwise convergence of  $F_{X_n}(x)$  to  $F(x)$  for all  $x$  implies convergence in distribution:

$$X_n \xrightarrow{d} \text{Exp}(1).$$

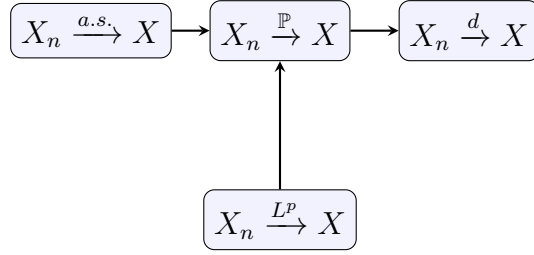
□

We end this section by presenting a version of the **Continuous Mapping Theorem**, which is often useful when proving the convergence of random variables through transformations.

**Theorem 3.4** (Continuous Mapping Theorem). *Let  $(X_n)$  be a sequence of random variables, and let  $h : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function. Then, the following statements hold:*

1. If  $X_n \xrightarrow{d} X$ , then  $h(X_n) \xrightarrow{d} h(X)$ ,  $n = 1, 2, \dots$ .
2. If  $X_n \xrightarrow{\mathbb{P}} X$ , then  $h(X_n) \xrightarrow{\mathbb{P}} h(X)$ .
3. If  $X_n \xrightarrow{a.s.} X$ , then  $h(X_n) \xrightarrow{a.s.} h(X)$ .

### 3.4 Relationships Between Modes of Convergence



In general:  $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{d} X$   
 and  $X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X$ .

#### Summary

- Almost sure convergence  $\Rightarrow$  convergence in probability  $\Rightarrow$  convergence in distribution.
- Convergence in  $L^p$  also  $\Rightarrow$  convergence in probability.
- The converses generally do not hold.

**Proposition 3.4.** *Almost sure convergence implies convergence in probability:*

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X.$$

*Proof.* Fix  $\varepsilon > 0$  and define the indicator random variables

$$Y_n := \mathbf{1}_{\{|X_n - X| > \varepsilon\}}.$$

If  $X_n \rightarrow X$  almost surely, then  $Y_n(\omega) \rightarrow 0$  for almost every  $\omega$  (because for almost every  $\omega$  there exists  $N(\omega)$  such that  $|X_n(\omega) - X(\omega)| \leq \varepsilon$  for all  $n \geq N(\omega)$ ). Moreover  $0 \leq Y_n \leq 1$  for all  $n$ , so the dominated convergence theorem applies. Hence

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}[Y_n] \xrightarrow{n \rightarrow \infty} 0.$$

Since this holds for every  $\varepsilon > 0$ , we conclude  $X_n \xrightarrow{\mathbb{P}} X$ . □

**Remark 3.2.** *This shows the chain of (standard) implications*

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X,$$

*and that almost sure convergence is stronger than convergence in probability.*

**Proposition 3.5.** *Convergence in  $L^p$  implies convergence in probability:*

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X.$$

*Proof.* Recall  $X_n \xrightarrow{L^p} X$  means  $\mathbb{E}[|X_n - X|^p] \rightarrow 0$  for some  $p > 0$ . For any  $\varepsilon > 0$ , Markov's inequality (applied to the nonnegative random variable  $|X_n - X|^p$ ) gives

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

Since  $\mathbb{E}[|X_n - X|^p] \rightarrow 0$ , the right-hand side tends to 0; hence  $\mathbb{P}(|X_n - X| \geq \varepsilon) \rightarrow 0$  for every  $\varepsilon > 0$ . This is exactly  $X_n \xrightarrow{\mathbb{P}} X$ . □



**Proposition 3.6.** *Convergence in probability implies convergence in distribution:*

$$X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X.$$

*Sketch of proof.* Let  $F_n$  and  $F$  denote the distribution functions of  $X_n$  and  $X$ , respectively. Fix  $x$  at which  $F$  is continuous. For any  $\delta > 0$ ,

$$\begin{aligned} \mathbb{P}(X_n \leq x) &\leq \mathbb{P}(X \leq x + \delta) + \mathbb{P}(|X_n - X| > \delta), \\ \mathbb{P}(X_n \leq x) &\geq \mathbb{P}(X \leq x - \delta) - \mathbb{P}(|X_n - X| > \delta). \end{aligned}$$

Indeed, on the event  $\{|X_n - X| \leq \delta\}$  we have  $\{X_n \leq x\} \subseteq \{X \leq x + \delta\}$  and  $\{X \leq x - \delta\} \subseteq \{X_n \leq x\}$ . Since  $X_n \xrightarrow{\mathbb{P}} X$ ,  $\mathbb{P}(|X_n - X| > \delta) \rightarrow 0$  for each fixed  $\delta > 0$ . Taking lim sup and lim inf as  $n \rightarrow \infty$  and then letting  $\delta \downarrow 0$ , using the continuity of  $F$  at  $x$ , yields

$$\lim_{n \rightarrow \infty} F_n(x) = F(x).$$

Since this holds for every continuity point  $x$  of  $F$ , we conclude  $X_n \xrightarrow{d} X$ . □

Thus convergence in probability is stronger than convergence in distribution (i.e., it implies it). The converse need not hold in general, as the following example shows.

**Example 3.18** (Counterexample: convergence in distribution does not imply convergence in probability). *Let  $\{X_n\}_{n \geq 1}$  be an i.i.d. sequence with  $X_n \sim \text{Bernoulli}(\frac{1}{2})$  for every  $n$ . Let  $X$  be another  $\text{Bernoulli}(\frac{1}{2})$  random variable independent of the whole sequence  $(X_n)_{n \geq 1}$ . Then for each fixed  $n$  the law of  $X_n$  equals the law of  $X$ , hence*

$$X_n \xrightarrow{d} X \quad (\text{trivially, since the finite-dimensional distributions coincide}).$$

However  $X_n$  does not converge to  $X$  in probability. Indeed, for  $0 < \varepsilon < 1$ ,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}(X_n \neq X) = \frac{1}{2}$$

for every  $n$ , so this probability does not tend to 0. Therefore  $X_n \not\xrightarrow{\mathbb{P}} X$ .

A useful special case where convergence in distribution does imply convergence in probability is when the limiting random variable is a constant.

**Theorem 3.5.** *If  $X_n \xrightarrow{d} c$  for some constant  $c \in \mathbb{R}$ , then  $X_n \xrightarrow{\mathbb{P}} c$ .*

*Proof.* Let  $\varepsilon > 0$  be arbitrary and consider the open interval  $G = (c - \varepsilon, c + \varepsilon)$ . The degenerate random variable  $c$  satisfies  $\mathbb{P}(c \in G) = 1$ . By the Portmanteau theorem (or using the elementary CDF argument at the continuity points  $c \pm \varepsilon$ ), convergence in distribution implies

$$\liminf_{n \rightarrow \infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(c \in G) = 1.$$

Hence  $\mathbb{P}(|X_n - c| < \varepsilon) = \mathbb{P}(X_n \in G) \rightarrow 1$ , which is exactly  $X_n \xrightarrow{\mathbb{P}} c$ . □

**Proposition 3.7** (Relationship between almost sure and  $L^p$  convergence).

- In general, almost sure convergence does not imply convergence in  $L^p$ .

- Convergence in  $L^p$  does not necessarily imply almost sure convergence.

Nevertheless, the two notions can coincide under certain conditions.

**Theorem 3.6** (Sufficient conditions for the link).

1. If  $X_n \xrightarrow{a.s.} X$  and  $\{|X_n|^p\}$  is uniformly integrable, then

$$X_n \xrightarrow{L^p} X.$$

2. Conversely, if  $X_n \xrightarrow{L^p} X$ , then there exists a subsequence  $(X_{n_k})$  such that

$$X_{n_k} \xrightarrow{a.s.} X.$$

*Sketch of proof.* **(1)** By the dominated convergence theorem: If  $X_n \rightarrow X$  almost surely and  $|X_n|^p \leq Y$  for some integrable random variable  $Y$  (i.e.  $\mathbb{E}[Y] < \infty$ ), then

$$\mathbb{E}[|X_n - X|^p] \rightarrow 0,$$

which is convergence in  $L^p$ .

**(2)** Conversely, suppose  $X_n \xrightarrow{L^p} X$ . Then  $\mathbb{E}[|X_n - X|^p] \rightarrow 0$ . By the Markov inequality and Borel–Cantelli lemma, one can extract a subsequence  $(X_{n_k})$  such that

$$\sum_k \mathbb{P}(|X_{n_k} - X| > \varepsilon) < \infty$$

for all  $\varepsilon > 0$ , implying  $X_{n_k} \xrightarrow{a.s.} X$ . □

**Remark 3.3.** We summarize the hierarchy of convergence modes:

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X.$$

Almost sure convergence is stronger than convergence in probability, but weaker than  $L^p$  convergence only under integrability conditions.