# Probability Theory

*Humbly scribed by* Ivan Kirev *and* Samuel Lam
*based on previous lectures by* Igor Krasovsky.

Version: 0.1.2, Date: August 24, 2022


Department of Mathematics
*Imperial College London*

We have written this set of unofficial notes to facilitate your understanding of important concepts in Probability Theory. Although most parts of the notes are based the lecture videos and in-person lectures given by Dr. Igor Krasovsky in Spring 2022, we have included additional materials drawn from various references. In particular:

- Any unexaminable materials are highlighted using gray colorboxes.
- Any unfinished/planned materials are written in red.

**Disclaimer.** These notes have not been checked by Dr. Krasovsky and should not be regarded as a replacement of the official notes and lectures for the course. In particular, all the errors are made by us, and we don't take any responsibility for their consequences; use at your own risk (and please do attend lectures – they are fun).

Please email us at (insert email) with any comments or corrections.

Ivan Kirev and Samuel Lam
September 2022

Version: 0.1.2, Date: August 24, 2022

# Contents


## I Measure Theory and Random Variables

1 Events, Probability and Random Variables . . . 4
1.1 Algebras and $\sigma$-algebras 5
1.2 Measurable Spaces 7
1.2.1 The measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ . . . 7
1.2.2 The measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ . . . 8
1.2.3 The measurable space $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ . . . 9
1.3 Probability Measures on Measurable Spaces 10
1.3.1 Probability Distributions . . . 10
1.3.2 Continuity of Measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ . . . 11
1.3.3 Probability Measures on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ . . . 14
1.3.4 Probability Measures on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ . . . 14
1.4 Random Variables 16
1.4.1 Operations of Random Variables . . . 17
1.4.2 Distributions of random variables . . . 19
1.4.3 Extension to Higher Dimensions . . . 19

2 Expectation and Integrals . . . 20
2.1 The Lebesgue Integral 20
2.2 Properties 20
2.2.1 Exchanging limits and expectations . . . 20
2.2.2 Change of variables . . . 21
2.3 Exchanging the Order of Integration 22
2.4 Jensen's Inequality and $L^p$ Spaces 23
2.4.1 Convex Functions and Jensen Inequality . . . 24
2.4.2 Inclusions of $\mathscr{L}^p$ spaces . . . 25
2.4.3 $\mathscr{L}^p(\Omega)$ and its seminorm . . . 26
2.5 Tail Bounds 29
2.5.1 Chernoff Bound and Moment Generating Function (MGF) . . . 31

3 More on Random Variables . . . 34
3.1 Transformation of Random Variables 34
3.2 Independent and Uncorrelated Random Variables 36
3.2.1 Independence . . . 36
3.2.2 Convolution of Independent Random Variables . . . 38
3.2.3 Correlation . . . 39

## II Concepts of Convergence

4 Coin Flips: Convergence in Probability . . . 41
4.1 Constructing the sample space 41
4.1.1 Radamecher Functions . . . 42
4.2 Weak Law of Large Numbers 43
4.2.1 A high probability statement for coin flips . . . 43
4.2.2 $L^2$ Weak Law of Large Numbers . . . 44
4.2.3 Weak Law of Large Number for uniformly integrable sequences* . . . 45

**4.3 Local and Central Limit Theorem** **46**
4.3.1 A crash course in asymptotic analysis . . . 46
4.3.2 Proving the Central Limit Theorem . . . 47
**4.4 Poisson Convergence** **50**
**4.5 Interlude: An overview to upcoming chapters** **51**

**5 Weak Convergence** . . . 52
**5.1 Definition of weak convergence** **52**
**5.2 Portmanteau Theorem and related facts** **54**
5.2.1 Slutsky's Theorem and Convergence of Probability . . . 57
5.2.2 Convergence of distribution function . . . 59
**5.3 Skorohod Representation Theorem** **61**
**5.4 Relative Compactness and Tightness** **62**
**5.5 Vague Convergence** **67**
5.5.1 A functional analysis view of vague convergence . . . 69

**6 Characteristic Functions** . . . 71
**6.1 Obtaining moments** **72**
**6.2 Inversion Formula** **75**
**6.3 Central Limit Theorem via Characteristic Functions** **78**
**6.4 More about constructing characteristic function** **82**
6.4.1 Bochner-Khinchin Theorem . . . 82
6.4.2 Polya's Criterion . . . 82
6.4.3 Marcinkiewicz Theorem . . . 84
6.4.4 Cumulants . . . 84
6.4.5 Degenerate distributions . . . 84

**7 Almost Sure Convergence of Random Series** . . . 86
**7.1 Important Zero-One Laws** **86**
7.1.1 Borel-Cantelli Lemma . . . 86
7.1.2 Kolmogorov's 0-1 Law . . . 88
**7.2 More on Almost Sure Convergence** **90**
7.2.1 Almost sure convergence implies convergence in probability . . . 90
7.2.2 Convergence in probability does not imply almost sure convergence . . . 91
7.2.3 Almost sure convergence and $L^p$ convergence . . . 94
**7.3 Strong Law of Large Numbers** **94**
**7.4 Random Walk** **100**

**III Foundations of Stochastic Processes**

**8 Conditioning and Disintegration** . . . 102
**8.1 Conditional Probability** **102**
8.1.1 The Discrete Case . . . 102
8.1.2 The General Case . . . 103
**8.2 Conditional Expectation** **104**
**8.3 Properties of conditional expectation** **106**
**8.4 Conditioning on a random variable** **108**
**8.5 Regular conditional distribution** **110**
8.5.1 Existence of regular conditional distribution . . . 112
8.5.2 Further Examples . . . 113

# Part I. Measure Theory and Random Variables

## 1 Events, Probability and Random Variables

The course aims to develop an abstract mathematical framework to describe the likelihood of certain events to happen in a random experiment. We would like to also formalise the large-sample results as developed in the probability classes you have taken in lower years, including the *Law of Large Numbers* and *Central Limit Theorem*.

To begin the story, we should understand how one can describe an experiment. You should have seen in previous probability courses that there are few key steps [1] to describe an experiment:

1. First describe the sets of possible outcomes $\omega$ of the experiment, which is known as *sample space* $\Omega$.
2. Then describe the *events* $A$ which we might observe as subsets of the sample space $\Omega$. We usually write the collection of such subset as $\mathscr{F}$.
3. Finally, assign a number $\mathbb{P}(A) \in [0,1]$ (included) to each of the subsets $A$ to quantify how likely this event might happen.

This leads to the following fundamental problem:

- What should we include in our collection of events $\mathscr{F}$, and
- How should we assign a number to those subsets in $\mathscr{F}$?

Let us first address the first question. For the description to make sense in reality, we should let $\mathscr{F}_* := \{\varnothing, \Omega\} \subseteq \mathscr{F}$, since we should definitely observe nothing or something in an experiment. In fact, we should probably including more subsets of $\Omega$ in $\mathscr{F}$ for our description to be useful.

A natural choice is to choose $\mathscr{F} = \mathscr{F}^* := 2^\Omega$, the *power set* of $\Omega$ (which contains all subsets of $\Omega$). This is fine if $\Omega$ is countable, but will raise some issue if $\Omega$ is uncountable, e.g. $\Omega = \mathbb{R}$. The main concern comes from our second question: there are many properties we wish $\mathbb{P}$ to satisfy, e.g. finite additivity:

$$\forall A, B \in \mathscr{F} \text{ disjoint}, \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

We can extend this naturally to countable additivity, and you should have seen that those properties would lead to contradictions! [2] Therefore, we should choose $\mathscr{F}$ that contains $\mathscr{F}_*$ but is strictly smaller than $2^\Omega$.

Mathematicians had therefore attempted to find suitable criteria on $\mathscr{F}$ and $\mathbb{P}$ so that they would not give rise to contradictions, but still allow $\mathscr{F}$ to be large enough for our framework to be useful. The most successful attempt was, perhaps, made by Andrey Kolmogorov in 1933 when he devised the axioms of probability in his *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of the Theory of Probability), which we will cover in this chapter. His work lead to the development of the *measure theory*, which forms the fundamentals of our probability theory.

As a final remark, the first two chapters of this course have many overlap with the courses you have done in year 1-2, so if you are familiar with the notions in these chapters, you can safely skip ahead.

[1] as suggested in Dr. Chris Hallsworth notes on MATH50010 Probability for Statistics.
[2] e.g. Vitali sets, Banach-Tarski paradox. See MATH50006 Lebesgue Measure and Integration.

## 1.1 Algebras and $\sigma$-algebras

Let $\Omega$ be a set of points $\omega$.

**Definition 1.1 — Algebra and $\sigma$-algebra.** A nonempty system of subsets of $\Omega$ is called an **algebra** $\mathscr{A}$ if

- $\Omega \in \mathscr{A}$
- $A, B \in \mathscr{A} \implies A \cup B \in A \quad (A \cap B \in \mathscr{A})$
- $A \in \mathscr{A} \implies A^c \in \mathscr{A}$

In addition, if all countable union $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$ whenever $A_1, A_2, \cdots \in \mathscr{A}$, then $\mathscr{A}$ is a $\sigma$-algebra. Note that then also $\bigcap_{n=1}^{\infty} A_n \in \mathscr{F}$ (consider $\Omega \backslash A_k = \hat{A}_k$).

**Definition 1.2 — Set function and measures.**

- A set function $\mu : \mathscr{A} \longrightarrow [0, \infty]$ is **finitely additive** if for any disjoint $A, B \in \mathscr{A}$

  $$\mu(A \cup B) = \mu(A) + \mu(B).$$

  Note that then $\forall A, B \in \mathscr{A}$

  $$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B).$$

- Let $\mathscr{F}$ be a $\sigma$-algebra. A set function $\mu : \mathscr{F} \to [0, \infty]$ is called $\boldsymbol{\sigma}$**-additive** if for any disjoint $A_1, A_2, ... \in \mathscr{F}$,

  $$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n).$$

  Such $\mu$ is called a **measure** on $\mathscr{F}$. A measure $\mu$ is called a **probability measure** if $\mu(\Omega) = 1$. Note that $\mu(\varnothing) = 0$ since $\mu(\varnothing) = \mu(\varnothing \cup \varnothing) = 2\mu(\varnothing)$.

- A measure is called $\boldsymbol{\sigma}$**-finite** if there exists a representation $\Omega = \bigcup_{k=1}^{\infty} \Omega_k$, $\Omega_k$ - pairwise disjoint, with

  $$\mu(\Omega_k) < \infty, \quad k = 1, 2, ...$$

**Definition 1.3** A **probability space** is a triple $(\Omega, \mathscr{F}, \mathbb{P})$, where $\Omega$ is a set called sample space, $\mathscr{F}$ is a $\sigma$-algebra of subsets of $\Omega$, $\mathbb{P}$ is a probability measure on $\mathscr{F}$. Any element of $\mathscr{F}$ is called an **event**.

**Property 1.4** Probability measures have the following properties:

- $\mathbb{P}(\varnothing) = 0$;
- If $A, B \in \mathscr{F}$ then
  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B);$$
- If $A, B \in \mathscr{F}$ and $B \subseteq A$ then
  $$\mathbb{P}(B) \leq \mathbb{P}(A);$$
- If $A_1, A_2, \cdots \in \mathscr{F}$, then
  $$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

**Proposition 1.5** Let $\mathbb{P}$ be a finitely additive set function defined over an algebra $\mathscr{A}$, with $\mathbb{P}(\Omega) = 1$. The following four conditions are equivalent:

(1) $\mathbb{P}$ is $\sigma$-additive (therefore it is a probability)

(2) $\mathbb{P}$ is continuous from below, i.e. for any sets $A_1, A_2, \cdots \in \mathscr{A}$ such that $A_1 \subset A_2 \subset \ldots$ and $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$,

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \lim_{n\to\infty} \mathbb{P}(A_n);$$

(3) $\mathbb{P}$ is continuous from above, i.e. for any sets $B_1, B_2, \ldots$ such that $B_1 \supset B_2 \supset \ldots$ and $\bigcap_{n=1}^{\infty} B_n \in \mathscr{A}$,

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big);$$

(4) $\mathbb{P}$ is continuous at $\varnothing$, i.e. for any sets $B_1, B_2, \cdots \in \mathscr{A}$ such that $B_1 \supset B_2 \supset \ldots$ and $\bigcap_{n=1}^{\infty} B_n = \varnothing$,

$$\lim_{n\to\infty} \mathbb{P}(B_n) = 0.$$

*Proof.* (1) $\implies$ (2). Since

$$\bigcup_{n=1}^{\infty} A_n = A_1 + (A_2 \backslash A_1) + (A_3 \backslash A_2) + \cdots,$$

we have

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) &= \mathbb{P}(A_1) + \mathbb{P}(A_1\backslash A_2) + \mathbb{P}(A_3\backslash A_2) + \cdots \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1) + \mathbb{P}(A_3) - \mathbb{P}(A_2) + \cdots \\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned}$$

(2) $\implies$ (3). Let $n \geq 1$; then

$$\mathbb{P}(B_n) = \mathbb{P}(B_1\backslash(B_1\backslash B_n)) = \mathbb{P}(B_1) - \mathbb{P}(B_1\backslash B_n).$$

The sequence $\{B_1\backslash B_n\}_{n\geq 1}$ of sets is non-decreasing and

$$\bigcup_{n=1}^{\infty}(B_1\backslash B_n) = B_1\backslash\bigcap_{n=1}^{\infty} B_n.$$

Then by (2)

$$\lim_{n\to\infty} \mathbb{P}(B_1\backslash B_n) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty}(B_1\backslash B_n)\Big)$$

and therefore

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(B_n) &= \mathbb{P}(B_1) - \lim_{n\to\infty} \mathbb{P}(B_1\backslash B_n) \\ &= \mathbb{P}(B_n) - \mathbb{P}\Big(\bigcup_{n=1}^{\infty}(B_1\backslash B_n)\Big) = \mathbb{P}(B_1) - \mathbb{P}\Big((B_1\backslash\bigcap_{n=1}^{\infty} B_n)\Big) \\ &= \mathbb{P}(B_1) - \mathbb{P}(B_1) + \mathbb{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big).\end{aligned}$$

(3) $\implies$ (4). Obvious.
(4) $\implies$ (1). Let $A_1, A_2, \cdots \in \mathscr{A}$ be pairwise disjoint and let $\sum_{n=1}^{\infty} A_n \in \mathscr{A}$. Then

$$\mathbb{P}\Big(\sum_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\sum_{i=1}^{n} A_i\Big) + \mathbb{P}\Big(\sum_{i=n+1}^{\infty} A_i\Big),$$

and since $\sum_{i=n+1}^{\infty} A_i \downarrow \varnothing, n \to \infty$, we have

$$\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i) = \lim_{n\to\infty} \mathbb{P}\Big(\sum_{i=1}^{n} A_i\Big)$$

$$= \lim_{n\to\infty} \left[ \mathbb{P}\Big( \sum_{i=1}^{\infty} A_i \Big) - \mathbb{P}\Big( \sum_{i=n+1}^{\infty} A_i \Big) \right]$$
$$= \mathbb{P}\Big( \sum_{i=1}^{\infty} A_i \Big) - \lim_{n\to\infty} \mathbb{P}\Big( \sum_{i=n+1}^{\infty} A_i \Big) = \mathbb{P}\Big( \sum_{i=1}^{\infty} A_i \Big).$$

■

**Proposition 1.6** Let $\mu$ be a finitely additive measure on an algebra $\mathscr{A}$ and let the sets $A_1, A_2, \cdots \in \mathscr{A}$ be pairwise disjoint and satisfy $A = \bigcup_{i=1}^{\infty} A_i \in \mathscr{A}$. Then

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

**Example 1.7 — $\sigma$-algebras on $\Omega$.** Let $\Omega$ be a sample space. The following are $\sigma$-algebras:

$$\mathscr{F}_* = \{\varnothing, \Omega\}; \quad \mathscr{F}^* = \{A : A \in \Omega\} = 2^{\Omega};$$

**Lemma 1.8** For any collection $\mathscr{E}$ of subsets of $\Omega$ there exists minimal algebra $a(\mathscr{E})$ and minimal $\sigma$-algebra $\sigma(\mathscr{E})$ that contains all elements of $\mathscr{E}$ (intersection of all algebras (resp. $\sigma$-algebras) containing $\mathscr{E}$).

*Proof.* Intersection, countable or uncountable, of algebras (resp. $\sigma$-algebras) containing $\mathscr{E}$ is an algebra (resp. $\sigma$-algebra) containing $\mathscr{E}$. ■

We say that $\sigma(\mathscr{E})$ is **generated** by $\mathscr{E}$.

**Example 1.9 — $\sigma$-algebra generated by partitions.** Let $D = \{D_1, D_2, \dots\}$ be a countable partition of $\Omega$ such that $\Omega = \bigsqcup_{j=1}^{\infty} D_j$. Then

$$\sigma(D) = \left\{ \bigcup_{j\in I} D_j : I \subset \mathbb{N} \right\},$$

To show this, we first note that the RHS is indeed a $\sigma$-algebra containing $D$. (Exercise!) Therefore we must have $\sigma(D) \subseteq$ the RHS. But since $\sigma(D)$ is closed under countable union we know that the RHS must be a subset of $\sigma(D)$. This proves the claim.

## 1.2 Measurable Spaces

**Definition 1.10** A **measurable space** is a pair $(E, \mathscr{E})$, where $E$ is a set and $\mathscr{E}$ is a $\sigma$-algebra on $E$.

### 1.2.1 The measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$

Let $\mathbb{R} = (-\infty, \infty)$ be the real line and

$$(a, b] = \{x \in \mathbb{R} : a < x \leq b\},$$

for all $a, b$ with $-\infty \leq a < b < \infty$. Let $\mathscr{A}$ be the algebra of subsets of $\mathbb{R}$ such that

$$A \in \mathscr{A} \quad \text{if} \quad A = \bigcup_{i=1}^{n} (a_i, b_i] \quad n < \infty.$$

Let $\mathscr{B}(\mathbb{R})$ be the smallest $\sigma$-algebra $\sigma(\mathscr{A})$ containing $\mathscr{A}$. We observe that

$$(a, b) = \bigcup_{n=1}^{\infty} \left( a, b - \frac{1}{n} \right], \quad a < b,$$

$$[a,b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right], \quad a < b,$$
$$\{a\} = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, a\right].$$

Thus the Borel algebra contains not only intervals $(a,b]$ but also the singletons $\{a\}$ and all sets of the forms

$$(a,b), \quad [a,b], \quad [a,b), \quad (-\infty,b), \quad (-\infty,b], \quad (a,\infty).$$

Let us also notice that the construction of $\mathscr{B}(\mathbb{R})$ could have been based on any of the intervals above instead of on $(a,b]$, since all the minimal $\sigma$-algebras generated by systems of intervals of any of the forms are the same as $\mathscr{B}(\mathbb{R})$.

**Exercise 1.11** Show that $\mathscr{B}(\mathbb{R})$ is generated by the collection of (1) open intervals of the form $(a,b)$, (2) closed intervals $[a,b]$, (3) half intervals, (4) intervals of the form $(-\infty,a]$ or $[a,\infty)$, (5) open sets and (6) closed sets with respect to the Euclidean metric.

### 1.2.2 The measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$

Let $\mathbb{R}^n = \mathbb{R} \times \cdots \mathbb{R}$ be the direct product of $n$ copies of the real line, i.e. the set of ordered $n$-tuples $x = (x_1, \ldots, x_n)$, where $x_k \in \mathbb{R}, k = 1, \ldots, n$. The set

$$I = I_1 \times \cdots I_n,$$

where $I_k = (a_k, b_k]$, i.e. the set $\{x \in \mathbb{R}^n : x_k \in I_k, k = 1, ..., n\}$ is called a rectangle and $I_k$ is a side of the rectangle. Let $\mathscr{I}$ be the set of all rectangles $I$. The smallest $\sigma$-algebra $\sigma(\mathscr{I})$ generated by the system $\mathscr{I}$ is the Borel algebra of subsets of $\mathbb{R}^n$. Let us show that we can arrive at this Borel algebra by starting in a different way.

Instead of the rectangles $I = I_1 \times \cdots I_n$ let us consider the rectangles $B = B_1 \times B_2 \times \cdots \times B_n$ with Borel sides ($B_k$ is the Borel subset of the real line that appears in the $k$th place in the direct product $\mathbb{R} \times \cdots \times \mathbb{R}$).The smallest $\sigma$-algebra containing all rectangles with Borel sides is denoted by

$$\mathscr{B}(\mathbb{R}) \otimes \cdots \otimes \mathscr{B}(\mathbb{R})$$

and called the direct product of the $\sigma$-algebras $\mathscr{B}(\mathbb{R})$. Let us show that in fact

$$\mathscr{B}(\mathbb{R}^n) = \mathscr{B}(\mathbb{R}) \otimes \cdots \otimes \mathscr{B}(\mathbb{R}).$$

In other words, the smallest $\sigma$-algebra generated by the rectangles $I = I_1 \times \cdots \times I_n$ and the (broader) class of rectangles $B = B_1 \times \cdots \times B_n$ with Borel sides are actually the same. We need the following lemma

**Lemma 1.12** Let $\mathscr{E}$ be a class of subsets of $\Omega$ and let $\mathscr{B} \subseteq \Omega$, and define

$$\mathscr{E} \cap B = \{A \cap B : A \in \mathscr{E}\}.$$

Then

$$\sigma(\mathscr{E} \cap B) = \sigma(\mathscr{E}) \cap B.$$

Now we show that $\mathscr{B}(\mathbb{R}^n)$ and $\mathscr{B}(\mathbb{R}) \otimes \cdots \otimes \mathscr{B}(\mathbb{R})$ are the same. This is obvious for $n = 1$. We now show that it is true for $n = 2$.

**Lemma 1.13**

$$\mathscr{B}(\mathbb{R}^2) = \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R})$$

*Proof.*

- $\mathscr{B}(\mathbb{R}^2) \subset \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R})$, since any open $A \subset \bigcup_{x \in A \cap \mathbb{Q}^2} R(x, \sigma(x)) \in \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R})$, $R(x,\tau)$ - open square centered at $x$ of side length $\tau(x)$.

- $\mathscr{B}(\mathbb{R}) \otimes \mathscr{B}(\mathbb{R}) \subset \mathscr{B}(\mathbb{R}^2)$. It is sufficient to check that $B_1 \times B_2 \in \mathscr{B}(\mathbb{R}^2)$ for any 2 Borel sets $B_1, B_2$. Note that $B_1 \times \mathbb{R} \in \mathscr{B}(\mathbb{R}^2)$ since

$$B_1 \times \mathbb{R} \in \sigma(\{\text{open subsets of } \mathbb{R}\}) \times \mathbb{R} = \sigma(\{\text{open subsets of } \mathbb{R} \times \mathbb{R}\}),$$

Similarly, $\mathbb{R} \times B_2 \in \mathscr{B}(\mathbb{R}^2)$, and so $B_1 \times B_2 = (B_1 \times \mathbb{R}) \cap (\mathbb{R} \times B_2) \in \mathscr{B}(\mathbb{R}^2)$.

■

The case for any $n, n > 2$ can be discussed in the same way.

**Remark 1.14** Let $\mathscr{B}_0(\mathbb{R}^n)$ be the smallest $\sigma-$ algebra generated by the open "balls"

$$S_\rho(x^0) = \{x \in \mathbb{R}^n : \rho_n(x, x^0) < \rho\}, \quad x^0 \in \mathbb{R}^n, \rho > 0,$$

in the metric

$$\rho_n(x, x^0) = \sum_{k=1}^{n} 2^{-k} \rho_1(x_k, x_k^0),$$

where $x = (x_1, \dots, x_n)$, $x^0 = (x_1^0, \dots, x_n^0)$. Then $\mathscr{B}_0(\mathbb{R}^n) = \mathscr{B}(\mathbb{R}^n)$ (exercise).

### 1.2.3 The measurable space $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$

This space plays a significant role in probability theory, since it is used as the basis for constructing probabilistic models of experiments with infinitely many steps. Let $\mathbb{R}^\infty = \{x = (x_1, x_2, \dots), x_k \in \mathbb{R}\}$.

**Definition 1.15** A set $C \subset \mathbb{R}^\infty$ is called **cylindrical** if it is of the form $C = \{x \in \mathbb{R}^\infty : (x_1, x_2, ..., x_n) \in \tilde{C}_n\}$ for some $n \geq 1$ and $\tilde{C}_n \in \mathscr{B}(\mathbb{R}^n)$.

Cylindrical sets form an algebra (check!) which generates a $\sigma$-algebra called cylindrical $\sigma$-algebra, denoted $\mathscr{B}(\mathbb{R}^\infty)$. One can verify that

$$\mathscr{B}(\mathbb{R}^\infty) = \sigma(\{A_1 \times A_2 \times \cdots \subset \mathbb{R}^\infty, A_k \in \mathscr{B}(\mathbb{R})\}).$$

**Example 1.16** For all $c \in \mathbb{R}$, let $A = \{x \in \mathbb{R}^\infty : \limsup_n x_n = \inf_n \sup_{k>n} x_k > c\}$. We have $A \in \mathscr{B}(\mathbb{R}^\infty)$: indeed

$$A = \bigcap_{n=1}^{\infty} \bigcup_{k=n+1}^{\infty} \{x \in \mathbb{R}^\infty : x_k > c\} = (x_k > c \quad i.o.).$$

For all $c$ let

$$B = \{x \in \mathbb{R}^\infty : \liminf_n x_n = \sup_n \inf_{k>n} x_k > c\}.$$

We have $B \in \mathscr{B}(\mathbb{R}^\infty)$: indeed

$$B = \bigcap_{n=1}^{\infty} \bigcup_{k=n+1}^{\infty} \{x \in \mathbb{R}^\infty : x_k > c\} = (x_k > c \quad ev.).$$

For all $c$, $D = \{x \in \mathbb{R}^\infty : \lim_{n\to\infty} x_n = c\} \in \mathscr{B}(\mathbb{R}^\infty)$ (exercise).

Recall: non-decreasing function $g(x)$ on $\mathbb{R}$ is continuous up to possibly countably many discontinuities of first kind: $f(x+0)$, $f(x-0)$ both exist, but $f(x+0) - f(x-0) = h_x > 0$. Moreover, the derivative $g'(x)$ exists Lebesgue a.e.

**Remark 1.17** We can generalise the notion of Borel $\sigma$-algebra to other sample spaces. Assume $\Omega$ is **Polish**, i.e. a metric space which is

- **complete**: all Cauchy sequences $(\omega_k)_{k\geq 1}$ has a limit $\omega \in \Omega$, and
- **separable**: there exists a countable subset $E := \{\omega_k\}_{k\geq 1}$ which is dense in $\Omega$, or $\bar{E} = \Omega$

then we can define the Borel $\sigma$-algebra of $\Omega$ $\mathscr{B}(\Omega)$ as the smallest $\sigma$-algebra containing all open subsets of $\Omega$. We can then talk about the Borel $\sigma$-algbera of e.g. the space of all continuous function of $[0,1]$, $\Omega = C^0([0,1])$, equipped with the supremum metric $d(f,g) = \sup_{[0,1]} |f-g|$.

In fact, we will look at probability measures on Polish space quite often in Chapter 5 when we talk about weak convergence. If you are not comfortable with dealing with general Polish spaces, just treat them as $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ for the first reading. However, we will still try to incorporate discussions about measures on a general Polish space as they laid the foundation of stochastic processes. A few applications include:

- Continuous-time random walk (e.g. Brownian motion) - this could obviously be viewed as an indexed family of real-valued random variable $B_t(\omega), t \geq 0$, but one can also treat it as **one single** random variable $B(\omega) \in C^0[0,1]$.
- Empirical processes (i.e. "sample" cumulative distribution function) - this is usually treated as a random variable with an output of a distribution function (as defined below).

Therefore, we would encourage you to understand important results related to measures on a general Polish space before delving deep into the theory of stochastic processes.

## 1.3 Probability Measures on Measurable Spaces

### 1.3.1 Probability Distributions

We begin by the following observation.

**Exercise 1.18** Let $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P})$ be a probability space and denote $F(x) := \mathbb{P}(-\infty, x] \quad x \in \mathbb{R}$. Show that:

- $F(x)$ is non-decreasing,
- $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$, and
- $F(x)$ is continuous on the right for all $x \in \mathbb{R}$.

**Definition 1.19** Every function $F : \mathbb{R} \longrightarrow [0,1]$ satisfying the above three conditions is called a **distribution function** (on the real line $\mathbb{R}$).

Thus to every probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, there corresponds a distribution function. In fact, the opposite is also true and there exists one to one correspondence between distribution functions and probability measures:

**Theorem 1.20** Let $F = F(x)$ be a distribution function on $\mathbb{R}$. Then there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that

$$\mathbb{P}(a,b] = F(b) - F(a),$$

for all $a, b, -\infty \leq a < b < \infty$.

This relies on the following fundamental result in measure theory.

**Theorem 1.21 — Caratheodory Theorem.** Let $\mu_0$ be a $\sigma$-additive (pre-)measure on $(\Omega, \mathscr{A})$, where $\mathscr{A}$

is an algebra of subsets of $\Omega$. Then there exists a unique measure $\mu$ on $(\Omega, \sigma(\mathscr{A}))$, such that

$$\mu(A) = \mu_0(A) \quad \forall A \in \mathscr{A}.$$

We make a few remarks on this theorem, the first one being the completeness of measure. Recall the following:

**Definition 1.22 — Complete measure.** A measure $\mu$ on a $\sigma$-algebra $\Sigma$ on $\Omega$ is called complete if any subset of a set of measure zero (null sets) is measurable, i.e. belongs to $\Sigma$.

We introduce this notion of completeness to avoid any caveats in proving results relating to measures. Note that $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P})$ with $\mathbb{P}$ constructed from the Caratheodory theory above is not complete ($\exists$ subsets of a Borel set which are not Borel). Fortunately, one can enlarge the $\sigma$-algebra so that it includes all null sets: if a measure $\mu$ on $\Sigma$ is not complete, it can be completed by extending $\Sigma$ to

$$\overline{\Sigma} = \sigma(\Sigma \cup \{B \in \Omega : B \subset A \in \Sigma, \mu(A) = 0\}),$$

and the definition of $\mu$ so that $\mu(B) = 0$ for any $B$ being a subset of a null set. This is a fundamental result that should be checked by yourself. The completion of the measure obtained from the Caratheodory theorem is called the *Lebesgue-Stiltjes measure*. In particular, the distribution function $F(x) = x$ corresponds to the Lebesgue measure on $\mathbb{R}$.

The second remark concerns measure determining sets. Given two measures $\mu, \nu$ defined on a common measurable space $(\Omega, \mathscr{F})$. How many sets do we need to check for us to show that $\mu = \nu$? The Caratheodory Theorem suggests that one may look at the algebra $\mathscr{A}$ such that $\mathscr{F} = \sigma(\mathscr{A})$, but the following theorem suggests that one might look at a far smaller collection of subsets of $\Omega$:

**Theorem 1.23 — Measure Determining Set.** Consider a measurable space $(\Omega, \mathscr{F})$ with measures $\mu, \nu$, and let $\mathscr{C}$ be a $\pi$-system containing $\Omega$ (in the sense that $A, B \in \mathscr{C}$ implies $A \cap B \in \mathscr{C}$). If $\mu(B) = \nu(B)$ for all $B \in \mathscr{C}$, then $\mu = \nu$.

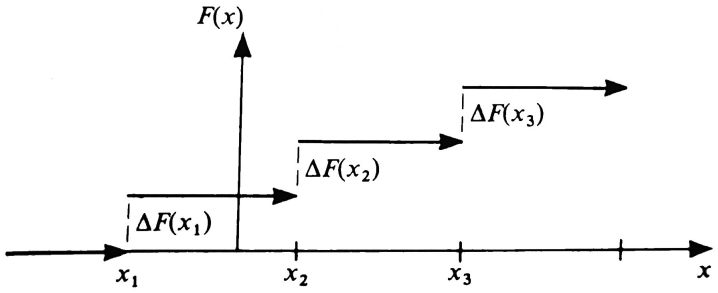
The key idea is to note that the collection $\{B \in \mathscr{F} \,|\, \mu(B) = \nu(B)\}$ is a Dynkin $\lambda$-system (which is non-empty, closed under complement and countable disjoint union). Since $\mathscr{C}$ above is a sub-collection of this collection, we know from $\pi$-$\lambda$ theorem that $\sigma(\mathscr{C})$ is contained in this collection. Therefore if one can show that $\sigma(\mathscr{C}) = \mathscr{F}$ then we have proven $\mu = \nu$. We will not go through the technical details here.

**Example 1.24** As an application, any probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is determined by open, closed or half-open intervals. In other words, e.g. if $\mu, \nu$ are probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\mu(O) = \nu(O)$ for all $O$ being open intervals, then $\mu = \nu$.

**Example 1.25** We can generalise this to the following: for probability measure defined on a Polish space $(X, \mathscr{B}(X))$, it is determined by open or closed sets in the above sense.

### 1.3.2 Continuity of Measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$

**Discrete/Atomic measures** are measures $\mathbb{P}$ for which the corresponding distributions $F = F(x)$ are piecewise constant, changing their values at the points $x_1, x_2, \ldots$ ($\Delta F(x_i) > 0$, where $\Delta F(x) = F(x) - F(x^-)$).




In this case the measure is concentrated at the points $x_1, x_2, \dots$, known as **atoms**:

$$\mathbb{P}(\{x_k\}) = \Delta F(x_k) > 0, \quad \sum_k \mathbb{P}(\{x_k\}) = 1$$

The set of numbers $(p_1, p_2, \dots)$, where $p_k = \mathbb{P}\{x_k\}$, is called a **discrete probability distribution** and the corresponding distribution function

$$F(x) = F_{\text{disc}}(x) = \sum_{x_k \le x} p_k, \qquad \dots, p_1, p_2, \cdots > 0$$

is called **discrete**.

**Example 1.26**

- Discrete Uniform distribution: $p_k = \frac{1}{N}, \quad k = 1, 2, \dots, N, \quad N$ - fixed.
- Bernoulli $\mathsf{B}(1, p)$: $p_1 = p, p_2 = 1 - p, \quad 0 \le p \le 1.$
- Binomial $\mathsf{B}(n, p)$: $p_k = \binom{n}{k} p^k (1-p)^k, \quad k = 0, 1, \dots, n \quad 0 \le p \le 1.$
- Poisson $\mathsf{Po}(\lambda)$: $p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \lambda > 0, \quad k = 0, 1, \dots$

**Absolutely continuous measures**. We begin by noting the following observation:

**Proposition 1.27** Let $f$ be an integrable function [a] $f(x) \ge 0$ such that

$$F(x) = F_{\mathsf{ac}}(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \quad \text{w.r.t Lebesgue measure}$$

Then the set function $\mathbb{P}_{ac}(A) = \int_A f(t)\mathrm{d}t \quad A \in \mathscr{F}$ is a measure. In particular, we say that $f$ is a density of $\mathbb{P}_{\mathsf{ac}}$.

[a] see Chapter 2

*Proof.* First assign $\mathbb{P}_{ac}((a, b]) = \int_a^b f(t)\mathrm{d}t$, then use Caratheodory theorem to extend to $\sigma$-algebra. ■

Measures of such kind is **absolutely continuous** (with respect to the Lebesgue measure $\mu$) in the following sense: if $\mu(A) = 0$ then $\mathbb{P}_{\mathsf{ac}}(A) = 0$. Indeed,

**Theorem 1.28 — Radon-Nikodym.** If $\mathbb{P}$ is a measure such that $\mu(A) = 0 \implies \mathbb{P}(A) = 0$, then $\mathbb{P}$ has a density.

Note that there is a connection between absolute continuity of measures and absolute continuity of measures: if $\mathbb{P}$ is an absolutely continuous measure then $F_{ac}(x)$ is an absolutely continuous function, and that $F'_{ac}(x) = f(x)$ almost everywhere.

Finally, the notion of absolute continuity can be generalised. This is useful in constructing conditional expectations: see Chapter 8.

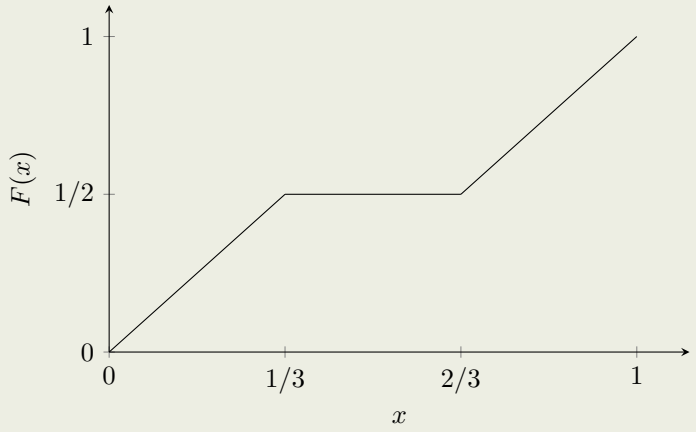
**Example 1.29**

- Uniform distribution on $[a, b]$: $f(x) = 1/(b-a), \quad a \le x \le b, \quad f(x) = 0$ otherwise.
- Normal or Gaussian: $f(x) = (2\pi\sigma^2)^{-1/2} \exp \frac{-(x-m)^2}{2\sigma^2}, \quad x \in \mathbb{R}, \quad \sigma > 0.$
- Gamma: $f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad x \ge 0 \quad \alpha, \beta > 0$

Recall that a measure $\nu$ is said to be **concentrated** on a measurable set $A$ if $\nu(E) = 0$ for any $E \subset \mathbb{R}\backslash A$.

**Singular continuous**. These are measures whose distribution functions are continuous but have all their points of increases on sets of zero Lebesgue measure. We have that $F(x) = F_{\mathsf{sc}}(x)$ is continuous at any $x$ and $\mathbb{P}_{\mathsf{sc}}$ is concentrated on a set of Lebesgue measure zero. In particular this distribution has no atoms. For $x$ in this set, $F'_{\mathsf{sc}}(x) \neq 0$ or does not exist. Thus $F'_{sc}(x) = 0$ a.e. Note that, by continuity, $\mathbb{P}_{\mathsf{sc}}\{x\} = 0$ for each point $x \in \mathbb{R}$.

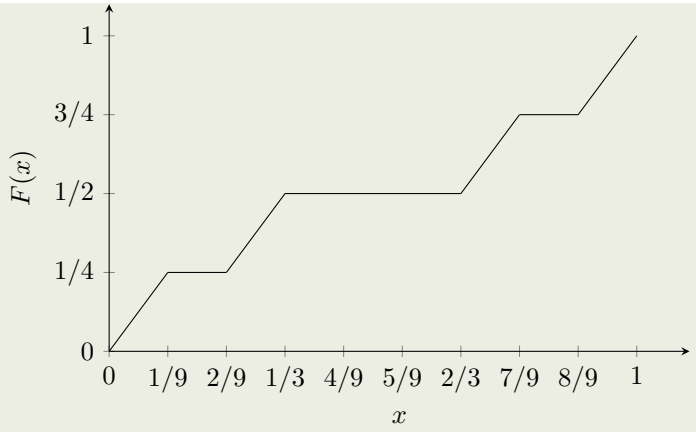
**Example 1.30 — Cantor's Devil staircase.** We consider the interval $[0, 1]$ and construct $F(x)$ by the following procedure originated by Cantor.




We divide $[0, 1]$ into thirds and put

$$F_2(x) = \begin{cases} 1/2, & x \in (\frac{1}{3}, \frac{2}{3}) \\ 1/4, & x \in (\frac{1}{9}, \frac{2}{9}) \\ 3/4, & x \in (\frac{7}{9}, \frac{8}{9}) \\ 0, & x = 0 \\ 1, & x = 1 \end{cases}$$

defining it in the intermediate intervals by linear interpolation. Then we divide each of the intervals $[0, 1/3]$ and $[2/3, 1]$ into three parts and define the function shown below with its values at other points determined by linear interpolation.




Continuing with this process, we construct a sequence of functions $F_n(x)$, $n = 1, 2, \ldots$ which converges to a non-decreasing continuous function $F(x)$ (the Cantor function), whose points of increase form a set of Lebesgue measure zero. In fact, it is clear from the construction of $F(x)$ that the total length of the intervals $(\frac{1}{3}, \frac{2}{3}), (\frac{1}{9}, \frac{2}{9}), (\frac{7}{9}, \frac{8}{9}), \ldots$ on which the function is constant is

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = 1.$$

Let $N$ be the set of points of increase of the Cantor function $F(x)$. It follows from the sum above that $\mathsf{Leb}(N) = 0$. At the same time, if $\mu$ is the measure corresponding to the Cantor function $F(x)$, we have $\mu(N) = 1$. (We then say that the measure is singular with respect to the Lebesgue measure Leb.)

**Theorem 1.31 — Hahn decomposition.** Any probability distribution has a representation:

$$F(x) = a_1 F_{\mathsf{disc}}(x) + a_2 F_{\mathsf{ac}}(x) + a_3 F_{\mathsf{sc}}(x), \quad a_1 + a_2 + a_3 = 1.$$

### 1.3.3 Probability Measures on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$

Distribution functions on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ are defined similarly, e.g.

$$F(x, y) = \mathbb{P}((-\infty, x] \times (-\infty, y]), \quad n = 2$$

The product measure on $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2)$ is defined as follows. First set

$$\mathbb{P}_0(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$$

for $A_1 \in \mathscr{F}_1, A_2 \in \mathscr{F}_2$. (Probability spaces $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$, $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$) Then extend $p_0$ to the algebra generated by $A_1 \times A_2$, show that $\mathbb{P}_0$ is a $\sigma$-additive measure on this algebra, and apply Caratheodory theorem to obtain the extension. This extension is called the product measure, denoted $\mathbb{P}_1 \otimes \mathbb{P}_2$.

### 1.3.4 Probability Measures on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$

For the spaces $\mathbb{R}^n$, $n \geq 1$, the probability measures were constructed in the following way: first for elementary sets (rectangles $(a, b]$), then, in a natural way, for sets $A = \sum(a_i, b_i]$, and finally, by using Caratheodory's theorem, for sets in $\mathscr{B}(\mathbb{R}^n)$.

A similar procedure of constructing probability measures also works for the space $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$. Let

$$\mathscr{I}_n(B) = \{x \in \mathbb{R}^\infty : (x_1, \ldots, x_n) \in B\}, \quad B \in \mathscr{B}(\mathbb{R}^n),$$

denote a cylinder set in $\mathbb{R}^\infty$ with base $B \in \mathscr{B}(\mathbb{R}^n)$. As we will see now, it is natural to take the cylinder sets for the elementary sets in $R^\infty$ whose probabilities enable us to determine the probability measure on the sets of $\mathscr{B}(\mathbb{R}^\infty)$.

We now define the notion of consistent sequence of measures:

**Definition 1.32 — Consistent Sequence.** The sequence $\mathbb{P}_n$ of probability measures on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ is said to be **consistent** if for all $n = 1, 2, \ldots$ and $B \in \mathscr{B}(\mathbb{R}^n)$,

$$\mathbb{P}_{n+1}(B \times \mathbb{R}) = \mathbb{P}_n(B).$$

The following theorem states that we can always construct a probability measure on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ based on a consistent sequence of measures:

**Theorem 1.33 — Kolmogorov Extension Theorem.** For any consistent sequence $\mathbb{P}_n$ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ such that

$$\mathbb{P}(\mathscr{I}_n(B)) = \mathbb{P}_n(B), \quad B \in \mathscr{B}(\mathbb{R}^n),$$

for $n = 1, 2, \ldots$.

*Proof.* Let $B^n \in \mathscr{B}(\mathbb{R}^n)$ and let $\mathscr{I}_n(B^n)$ be the cylinder with base $B^n$. We assign the measure $\mathbb{P}(\mathscr{I}_n(B^n))$ to this cylinder by taking

$$\mathbb{P}(\mathscr{I}_n(B^n)) = \mathbb{P}_n(B^n).$$

Let us show that, in virtue of the consistency condition, this definition is consistent, i.e., the value of $\mathbb{P}(\mathscr{I}_n(B^n))$ is independent of the representation of the set $\mathscr{I}_n(B^n)$. In fact, let the same cylinder be represented in two ways:

$$\mathscr{I}_n(B^n) = \mathscr{I}_{n+k}(B^{n+k}).$$

It follows that, if $(x_1, \ldots, x_{n+k}) \in \mathbb{R}^{n+k}$, we have

$$(x_1, \ldots, x_n) \in B^n \iff (x_1, \ldots, x_{n+k}) \in B^{n+k},$$

and therefore,

$$\begin{aligned}\mathbb{P}_n(B^n) &= \mathbb{P}_{n+1}((x_1, \ldots, x_{n+1}) : (x_1, \ldots, x_n) \in B^n) \\ &= \cdots = \mathbb{P}_{n+k}((x_1, \ldots, x_{n+k}) : (x_1, \ldots, x_n) \in B^n) \\ &= \mathbb{P}_{n+k}(B^{n+k}).\end{aligned}$$

Let $\mathscr{A}(\mathbb{R}^\infty)$ denote the collection of all cylinder sets $\hat{B}^n = \mathscr{I}_n(B^n), B^n \in \mathscr{B}(\mathbb{R}^n), n = 1, 2, \ldots$. It is easy to see that $\mathscr{A}(\mathbb{R}^\infty)$ is an algebra.

Now let $\hat{B}_1, \ldots, \hat{B}_k$ be disjoint sets in $\mathscr{A}(\mathbb{R}^\infty)$. We may suppose without loss of generality that $\hat{B}_i = \mathscr{I}_n(B_i^n), i = 1, \ldots, k$, for some $n$, where $B_1^n, \ldots, B_k^n$ are disjoint sets in $\mathscr{B}(\mathbb{R}^n)$. Then

$$\mathbb{P}\left(\sum_{i=1}^k \hat{B}_i\right) = \mathbb{P}\left(\sum_{i=1}^k \mathscr{I}_n(B_i^n)\right) = \mathbb{P}_n\left(\sum_{i=1}^k B_i^n\right) = \sum_{i=1}^n \mathbb{P}_n(B_i^n) = \sum_{i=1}^n \mathbb{P}(\hat{B}_i),$$

i.e. the set function $\mathbb{P}$ is finitely additive on the algebra $\mathscr{A}(\mathbb{R}^\infty)$.
Let us show that $\mathbb{P}$ is continuous at zero (and therefore $\sigma$-additive on $\mathscr{A}(\mathbb{R}^\infty)$), i.e., if a sequence of sets $\hat{B}_n \downarrow \emptyset$, $n \to \infty$, then $\mathbb{P}(\hat{B}_n) \to 0$, $n \to \infty$. Suppose the contrary, i.e., let $\lim \mathbb{P}(\hat{B}_n) = \delta > 0$ (the limit exists due to monotonicity). We may suppose without loss of generality that $\{\hat{B}_n\}$ has the form

$$\hat{B}_n = \{x : (x_1, \ldots, x_n) \in B_n\}, \quad B_n \in \mathscr{B}(\mathbb{R}^n).$$

We use the following property of probability measures $\mathbb{P}_n$ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$: if $B_n \in \mathscr{B}(\mathbb{R}^n)$, for a given $\delta > 0$ we can find a compact set $A_n \in \mathscr{B}(\mathbb{R}^n)$ such that $A_n \subset B_n$ and

$$\mathbb{P}_n(B_n \backslash A_n) \leq \delta / 2^{n+1}.$$

Therefore, if

$$\hat{A}_n = \{x : (x_1, \ldots, x_n) \in A_n\},$$

we have

$$\mathbb{P}(\hat{B}_n \backslash \hat{A}_n) = \mathbb{P}(B_n \backslash A_n) \leq \delta/2^{n+1}.$$

Form the set $\hat{C}_n = \bigcap_{k=1}^n \hat{A}_k$ and let $C_n$ be such that

$$\hat{C}_n = \{x : (x_1, \dots, x_n) \in C_n\}.$$

Then, since the sets $\hat{B}_n$ decrease, we obtain

$$\mathbb{P}(\hat{B}_n \backslash \hat{C}_n) \leq \sum_{k=1}^n \mathbb{P}(\hat{B}_k \backslash \hat{A}_k) \leq \delta/2.$$

But by assumption, $\lim_n \mathbb{P}(\hat{B}_n) = \delta > 0$, and therefore $\lim_n \mathbb{P}(\hat{C}_n) \geq \delta/2 > 0$. Let us show that this contradicts the condition $\hat{C}_n \downarrow \emptyset$.

Let us choose a point $\hat{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \dots)$ in $\hat{C}_n$. Then $(x_1^{(n)}, \dots, x_n^{(n)}) \in C_n$ for $n \geq 1$.
Let $(n_1)$ be a subsequence of $(n)$ such that $x_1^{(n_1)} \to x_1^0$, where $x_1^0$ is a point in $C_1$. (Such a sequence exists since $x_1^{(n_1)} \in C_1$ and $C_1$ is compact.) Then select a subsequence $(n_2)$ of $(n_1)$ such that $(x_1^{(n_2)}, x_2^{(n_2)}) \to (x_1^0, x_2^0) \in C_2$. Similarly let $(x_1^{(n_k)}, \dots, x_k^{(n_k)}) \to (x_1^0, \dots, x_k^0) \in C_k$. Finally, from the diagonal sequence $(m_k)$, where $m_k$ is the $k$th term of $(n_k)$. Then $x_i^{(m_k)} \to x_i^0$ as $m_k \to \infty$ for $i = 1, 2, \dots$, and $(x_1^0, x_2^0, \dots) \in \hat{C}_n$ for $n = 1, 2, \dots$, which evidently contradicts the assumption that $\hat{C}_n \downarrow \emptyset$, $n \to \infty$. Thus the set function $\mathbb{P}$ is $\sigma$-additive on the algebra $\mathscr{A}(\mathbb{R}^\infty)$ and hence, by the Caratheodory's theorem, it can be extended to a (probability) measure on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$. This completes the proof of the theorem. ■

## 1.4 Random Variables

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

**Definition 1.34 — Random variable.** A real function $\xi : \Omega \to \mathbb{R}$ is an **$\mathscr{F}$-measurable function**, or a **random variable** if

$$\{\omega : \xi(\omega) \in B\} \in \mathscr{F}$$

for every $B \in \mathscr{B}(\mathbb{R})$; or equivalently, if the inverse image

$$\xi^{-1}(B) \equiv \{\omega : \xi(\omega) \in B\}$$

is a measurable set in $\Omega$. When $(\Omega, \mathscr{F}) = (\mathbb{R}^n, \mathscr{B}(R^n))$, the $\mathscr{B}(R^n)$-measurable functions are called **Borel functions**.

Random variables are used to *summarise* certain abstract outcomes $\omega$ with a real number / vector etc. A good example is the following:

**Example 1.35 — Sum of two dices.** Consider the experiment of throwing two independent fair six-faced dices. This can be represented by the probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1) \otimes (\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$, where for $i = 1, 2$, $\Omega_i = \{1, 2, 3, 4, 5, 6\}$ is the outcome from dice $i$, $\mathscr{F}_i = 2^{\Omega_i}$ and $\mathbb{P}_i(\{j\}) \equiv 1/6$ for all $j \in \{1, 2, 3, 4, 5, 6\}$. We can then consider the function $X : \Omega_1 \times \Omega_2 \to \mathbb{R}$ by

$$X(\omega_1, \omega_2) = \omega_1 + \omega_2$$

which summarises the outcome of two dices by their sum. One can easily check that $X$ is a measurable function by looking at the possible pre-images of $X$, and we will leave it as an exercise.

**Exercise 1.36** In the example above,

- What are the possible outcomes of $X$?
- What are the possible pre-images of $X$? Notice they are subsets of $\mathscr{F}_1 \otimes \mathscr{F}_2 = 2^{\Omega_1 \times \Omega_2}$.

**Remark 1.37** Let $\xi$ be a random variable. If we consider sets from $\mathscr{F}$ of the form $\{\omega : \xi(\omega) \in B\}, B \in \mathscr{B}(\mathbb{R})$, it is easily verified that they form a $\sigma$-algebra, called the **$\sigma$-algebra generated by** $\xi$, denoted by $\mathscr{F}_\xi$ and $\mathscr{F}_\xi \subset \mathscr{F}$.

### 1.4.1 Operations of Random Variables

**Lemma 1.38** Let $\mathscr{D}$ be a collection of subsets on $\mathbb{R}$ such that $\sigma(\mathscr{D}) = \mathscr{B}(\mathbb{R})$. A necessary and sufficient condition that a function $\xi = \xi(\omega)$ is a random variable is that

$$\xi^{-1}(D) = \{\omega : \xi(\omega) \in D\} \in \mathscr{F}$$

for all $D \in \mathscr{D}$.

**Corollary 1.39** A necessary and sufficient condition for $\xi = \xi(\omega)$ to be a random variable is that

$$\{\omega : \xi(\omega) < x\} \in \mathscr{F}$$

for every $x \in \mathbb{R}$, or that

$$\{\omega : \xi(\omega) \leq x\} \in \mathscr{F}$$

for every $x \in \mathbb{R}$.

**Lemma 1.40** Let $\varphi = \varphi(x)$ be a Borel function and $\xi = \xi(\omega)$ a random variable. Then the composition $\eta = \varphi \circ \xi$, i.e. the function $\eta(\omega) = \varphi(\xi(\omega))$, is also a random variable (and, in fact, $\mathscr{F}_\xi$- measurable).

*Proof.* The proof follows from the equations:

$$\{w : \eta(\omega) \in B\} = \{w : \eta(\xi(\omega)) \in B\} = \{w : \xi(\omega) \in \varphi^{-1}(B)\} \in \mathscr{F}$$

for $B \in \mathscr{B}(\mathbb{R})$, since $\varphi^{-1}(B) \in \mathscr{B}(\mathbb{R})$. ■

**Example 1.41**

- If $\xi$ is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a continuous function then $f(\xi)$ is a random variable.
- If $\xi$ is a random variable, then so are $\xi^n, \xi^+ = \max(\xi, 0), \xi^- = -\min(\xi, 0)$, and $|\xi| = \xi^+ + \xi^-$

**Lemma 1.42** If $\xi, \eta$ are random variables, then

$$\xi + \eta, \xi - \eta, \xi\eta, \xi/\eta, \max(\xi, \eta), \min(\xi, \eta)$$

are random variables (assuming that they are defined, i.e. no forms like $\infty - \infty, \infty/\infty, a/0$ occur).

**Lemma 1.43** If $f_n$, $n = 1, 2, \ldots$ are random variables and if $\forall\omega$

$$s(\omega) = \sup_n f_n(\omega)$$

exits, then $s(\omega)$ is a random variable. Similarly, we can replace $\sup_n$ with $\inf_n$ or $\lim_n$.

**Definition 1.44** A random variable $\xi$ is called **simple** if

$$\xi(\omega) = \sum_{j=1}^{n} x_j \chi_{D_j}(\omega)$$

for some $n \geq 1$ and a partition $D_1, D_2, \ldots, D_n$ of $\Omega$, where

$$\chi_D(\omega) = \begin{cases} 1, & \omega \in D \\ 0, & \text{otherwise} \end{cases}$$

**Lemma 1.45**

- For every random variable $\xi = \xi(\omega)$ there is a sequence of simple random variables $\xi_1, \xi_2, \ldots$, such that $|\xi_n| \leq |\xi|$ and $\xi_n(\omega) \to \xi(\omega), n \to \infty$, for all $\omega \in \Omega$.
- For any random variable $\xi(\omega) \geq 0$ there exists a pointwise non-decreasing sequence of simple random variables $\xi_1(\omega) \leq \xi_2(\omega) \leq \cdots \leq \xi(\omega)$ such that

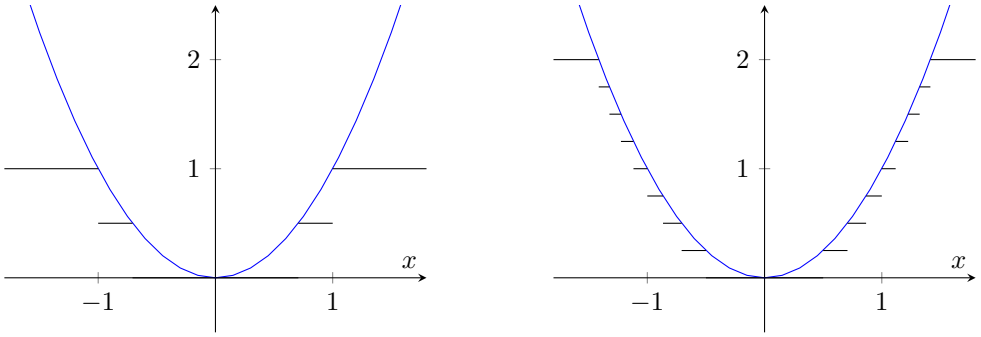
$$\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega) \quad \forall \omega \in \Omega \quad (\text{in short } \xi_n \nearrow \xi)$$

*Proof.* We begin by proving the second statement. For $n = 1, 2, \ldots$, put

$$\xi_n(\omega) = \sum_{j=0}^{n2^n - 1} \frac{j}{2^n} \chi_{\left\{\omega : \frac{j}{2^n} \leq \xi(\omega) < \frac{j+1}{2^n}\right\}} + n\chi_{\{\omega : \xi(\omega) \geq n\}}$$

It is easy to verify that the sequence $\xi_n(\omega)$ so constructed is such that $\xi_n \nearrow \xi$ for all $\omega \in \Omega$. The first statement follows from this if we merely observe that $\xi$ can be represented in the form $\xi = \xi^+ - \xi^-$, where $\xi^+ = \max(\xi, 0)$ and $\xi^- = \max(-\xi, 0)$. ■

The below figures represent how one can build simple function approximation to the function $f(x) = x^2$ for all $x \in \mathbb{R}$. For simplicity, only the first two steps are shown.




**(a)** $n = 1$ **(b)** $n = 2$

This construction gives a routine for proving statements related to random variables:

1. We first prove the statement for indicator functions.
2. We extend the statement for simple random variables by considering linearity.
3. We extend the statement to non-negative random variables by taking limits.
4. We extend the statement for any random variables by considering its positive and negative parts.

This is known as the *four-step* proof.

**Lemma 1.46** Consider a measurable space $(\Omega, \mathscr{F})$ and a finite or countable decomposition $\mathscr{D} = \{D_1, D_2, \dots\}$ of the space $\Omega$. Let $\xi = \xi(\omega)$ be a $\sigma(\mathscr{D})$-measurable random variable. Then $\xi$ is representable in the form

$$\xi(\omega) = \sum_{k=1}^{\infty} \alpha_k \chi_{D_k}(\omega),$$

where $\alpha_k \in \mathbb{R}$, i.e. $\xi(\omega)$ is constant on the elements $\mathscr{D}_k$ of the decomposition.

### 1.4.2 Distributions of random variables

**Definition 1.47** A probability measure $\mathbb{P}_\xi$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with

$$\mathbb{P}_\xi(B) = \mathbb{P}\{\omega : \xi(\omega) \in B\}, \quad B \in \mathscr{B}(\mathbb{R}),$$

is called the **probability distribution of** $\xi$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

It is quite clear that the above definition makes sense since $\mathbb{P}_\xi(B)$ is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

**Definition 1.48** The function

$$F_\xi(x) \equiv \mathbb{P}_\xi(-\infty, x] = \mathbb{P}\{\omega : \xi(\omega) \le x\}, \quad x \in \mathbb{R},$$

is called the **distribution function** of $\xi$.

**Example 1.49 — Sum of two dices - continued.** One can easily verify that the probability distribution of $X$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ satisfies the following:

$$\mathbb{P}_X(\{j\}) = \frac{6 - |7 - j|}{36}, \quad j \in \{2, 3, ..., 12\}$$

and extend the definition of $\mathbb{P}_X$ to other sets in $\mathscr{B}(\mathbb{R})$.

Notice that there are multiple random variables (on potentially different probability spaces) which gives the same distribution function! Indeed, we have established in previous section that we can always construct a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ given a distribution function $F_\xi$. Therefore for any random variables $\xi$ on probability $(\Omega, \mathscr{F}, \mathbb{P})$, the identity random variable $I(\omega) = \omega$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P}_\xi)$ has same distribution as $\xi$! For example, consider the space $(\Omega, 2^\Omega, Q)$, with $\Omega = \{2, 3, ..., 12\}$ and $Q(\{j\}) = \mathbb{P}_X(\{j\})$ as defined above. Consider the random variable $\xi(j) = j$ for all $j \in \Omega \subseteq \mathbb{R}$. Then $Q_\xi(A) \equiv \mathbb{P}_X(A)$ for all $A \in \mathscr{B}(\mathbb{R})$.

### 1.4.3 Extension to Higher Dimensions

**Definition 1.50 — Extension to** $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$.

$$\xi : \Omega \to \mathbb{R}^n \quad \xi = (\xi_1, \xi_2, \dots, \xi_n)$$

is called a **random vector** if for any $B \subset \mathscr{B}(\mathbb{R}^n)$, $\xi^{-1}(B) \in F$. $\mathbb{P}_\xi$ is also defined as before and we say that $\mathbb{P}_\xi = \mathbb{P}_{(\xi_1, \dots, \xi_n)}$ is a **joint distribution** of $\xi_1, \dots, \xi_n$, given by

$$F_\xi(x_1, \dots, x_n) = \mathbb{P}(\omega : \xi_1 \le x_1, \dots, \xi_n \le x_n).$$

Note that $\xi$ is a random vector if and only if $\xi_1, \dots, \xi_n$ are random variables. For $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ we can define random sequences $\xi = (\xi_1, \xi_2, \dots)$.

# 2 Expectation and Integrals

## 2.1 The Lebesgue Integral

First we recall the construction of the Lebesgue integral. Define the expectation of a simple random variable $\xi = \sum_{j=1}^{n} x_j \chi_{D_j}$ by

$$\mathbb{E}[\xi] = \sum_{j=1}^{n} x_j \mathbb{P}(D_j), \tag{2.1}$$

where the sets $D_j$ are a decomposition of $\Omega$.

For an arbitrary non-negative random variable $\xi = \xi(\omega)$ we construct a sequence of simple nonnegative random variables $\{\xi_n\}_{n\geq 1}$ such that $\xi_n(\omega) \uparrow \xi(\omega)$, as $n \to \infty$ for each $\omega \in \Omega$. We then set $\mathbb{E}[\xi] = \lim_{n\to\infty} \mathbb{E}[\xi_n]$, which exists since $\mathbb{E}[\xi_n] \leq \mathbb{E}[\xi_{n+1}]$ (possibly taking the value $+\infty$).

**Definition 2.1 — Expectation.** The **expectation** $\mathbb{E}[\xi]$ of a random variable $\xi$ is the Lebesgue integral w.r.t. $\mathbb{P}$

$$\mathbb{E}[\xi] := \lim_{n\to\infty} \mathbb{E}[\xi_n] = \int_\Omega \xi \mathrm{d}\mathbb{P} = \int_\Omega \xi(\omega)\mathbb{P}(\mathrm{d}\omega) \tag{2.2}$$

if it exists. We say that $\xi$ is **integrable** if $\mathbb{E}[|\xi|]$ exists and is finite ($\mathbb{E}[|\xi|]$ exists and is finite $\iff$ $\mathbb{E}[\xi]$ exists and is finite).

To see that this definition is consistent, one has to show it is independent of the choice of $\xi_n \uparrow \xi$. For general random variables (not necessary non-negative), we use the fact that $\xi = \xi^+ - \xi^-$.

## 2.2 Properties

We recall the following basic properties:

**Property 2.2** Let $\xi, \eta$ be integrable random variables and $c$ a constant. Then

- $\mathbb{E}[c] = c$
- $\mathbb{E}[c\xi] = c\mathbb{E}[\xi]$
- $\xi + \eta$ is integrable and $\mathbb{E}(\xi + \eta) = \mathbb{E}[\xi] + \mathbb{E}[\eta]$
- $\xi \leq \eta \implies \mathbb{E}[\xi] \leq \mathbb{E}[\eta]$
- if $\xi = \eta$ a.e. w.r.t $\mathbb{P}$ (a.e. $\equiv$ almost surely (a.s) i.e. up to sets of $\mathbb{P}$-measure zero), then $\mathbb{E}[\xi] = \mathbb{E}[\eta]$
- if $\xi \geq 0$ and $\mathbb{E}[\xi] = 0$, then $\xi = 0$ a.s.

### 2.2.1 Exchanging limits and expectations

We begin by recalling the following monotone convergence thoerem.

**Theorem 2.3 — Monotone convergence theorem (MCT).** Let $0 \leq \xi_1 \leq \xi_2 \leq \cdots$ be random variables. Then there exists (finite or infinite)

$$\lim_{n\to\infty} \mathbb{E}[\xi_n] = \mathbb{E}\left[\lim_{n\to\infty} \xi_n\right]. \tag{2.3}$$

**Remark 2.4**

- $0 \leq \xi_1 \leq \cdots$ can be replaced by $\eta \leq \xi_1 \leq \cdots$ with $\mathbb{E}[\eta] > -\infty$ (just consider $\xi_n - \eta$ instead of $\xi_n$).
- $0 \leq \xi_1 \leq \cdots$ can be replaced by $\cdots \leq \xi_2 \leq \xi_1 \leq \eta$, with $\mathbb{E}[\eta] < \infty$.

**Corollary 2.5** Let $\{\eta_k\}_{k\geq 1}$ be a sequence of non-negative random variables. Then

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\eta_k\right] = \sum_{k=1}^{\infty}\mathbb{E}[\eta_k]. \tag{2.4}$$

**Theorem 2.6 — Fatou's Lemma.** Let $\{\xi_n\}_{n\geq 1}$ be non-negative random variables. Then

$$\mathbb{E}[\liminf_n \xi_n] \leq \liminf_n \mathbb{E}[\xi_n]. \tag{2.5}$$

**Remark 2.7**

- $\xi_n \geq 0$ can be replaced by $\xi_n \geq \eta$, if $\mathbb{E}[\eta] > -\infty$
- If $\xi_n < \eta$, $\mathbb{E}[\eta] < \infty$, the statement holds for $\limsup$ instead.

**Hint.** *Apply monotone convergence theorem to* $\lambda = \inf_{k>n}\xi_k$.

**Theorem 2.8 — Lebesgue's Theorem on Dominated Convergence.** Let $\{\xi_n\}_{n\geq 1}$ be random variables such that $\xi_n \to \xi$ (a.s.). If there exists an integrable random variable $\eta$ ($\mathbb{E}[\eta] < \infty$) such that $|\xi_n| \leq \eta, \forall n$, then $\xi$ is integrable ($\mathbb{E}[\xi] < \infty$),

$$\mathbb{E}[\xi_n] \to \mathbb{E}[\xi], \tag{2.6}$$

and

$$\mathbb{E}[|\xi_n - \xi|] \to 0 \tag{2.7}$$

as $n \to \infty$.

**Corollary 2.9** Let $\eta, \xi, \xi_1, \ldots$ be random variables such that $|\xi_n| \leq \eta, \xi_n \to \xi$ (a.s.) and $\mathbb{E}[\eta^p] < \infty$ for some $p > 0$. Then $\mathbb{E}[|\xi|^p] < \infty$ and $\mathbb{E}[|\xi - \xi_n|^p] \to 0$ as $n \to \infty$.

**Remark 2.10** In all the above theorems, integral over $\Omega$ can be replaced by integral over any measurable $\hat{A} \subset \Omega$.

### 2.2.2 Change of variables

**Theorem 2.11 — Change of variables / Law of Unconscious Statistician (LOTUS).** Let $\xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a random variable with probability distribution $\mathbb{P}_\xi$. Then if $g = g(x)$ is a Borel function, then for all $A \in \mathscr{B}(\mathbb{R})$,

$$\int_A g(x)\mathrm{d}\mathbb{P}_\xi = \int_{\xi^{-1}(A)} g(\xi(\omega))\mathrm{d}\mathbb{P}, \tag{2.8}$$

where both integrals exist or not simultaneously. In particular, for $A = \mathbb{R}$ we obtain

$$\mathbb{E}[g(\xi(\omega))] = \int_\Omega g(\xi(\omega))\mathrm{d}\mathbb{P} = \int_{-\infty}^{\infty} g(x)\mathrm{d}\mathbb{P}_\xi \equiv \int_{-\infty}^{\infty} g(x)\mathrm{d}F_\xi. \tag{2.9}$$

*Proof.* We use the four-step proof. The result clearly holds for $g = \chi_B(x), B \in \mathscr{B}(\mathbb{R})$. Therefore, also holds for simple $g(x)$ by linearity of the integral. For any measurable $g(x) \geq 0$ consider a sequence of simple $g_n \nearrow g$. The result for $g$ then follows from monotone convergence theorem. For arbitrary measurable $g(x)$ we use $g(x) = g_+ - g_-$. ■

**Remark 2.12** The theorem guarantees that the expectation only depends on the probability distribution, but NOT the underlying probability space! In particular,

1. If $\xi$ is discrete ($\mathscr{F}_\xi$ is discrete) taking values $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots$ then the theorem gives
$$\mathbb{E}[g(\xi)] = \sum_j g(x_j)p_j \tag{2.10}$$
2. If $\xi$ is absolutely continuous (i.e. $F_\xi$ is absolutely continuous) with density $f(x)$, then
$$\mathbb{E}[g(\xi)] = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x. \tag{2.11}$$

This provides a way to calculate expectations of $g(\xi)$ without first being "conscious" with the actual distribution of $g(\xi)$. In addition, it makes sense to talk about expectations of probability distributions **without** specifying its underlying probability space.

So are there any points to specify the underlying probability space of a random variable instead of assuming it to be $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P}_\xi)$? Unfortunately the answer *can be* no, since there is a way to develop probability theory (and notions of random variables) without going through the Kolmogorov constructions. Indeed, the underlying probability space is not important at all in most of the applications in statistics. Nevertheless, this formulation is still helpful in understanding more complicated random variables like stochastic processes.

## 2.3 Exchanging the Order of Integration

We now consider product measures, and more specifically expressing integration on product measures as a repeated integral.

Suppose $(X_1, \mathscr{M}_1, \mu_1)$ and $(X_2, \mathscr{M}_2, \mu_2)$ are a pair of measure spaces and consider the product measure $(\mu_1 \times \mu_2)$ on the space
$$X = X_1 \times X_2 = \{(x_1, x_2) : x_1 \in X_1, x_2 \in X_2\}. \tag{2.12}$$
We assume that the two measure spaces are complete and $\sigma$-finite .

Given a set $E$ in $\mathscr{M}$ we consider the *slices*
$$E_{x_1} = \{x_2 \in X_2 : (x_1, x_2) \in E\} \quad \text{and} \quad E^{x_2} = \{x_1 \in X_1 : (x_1, x_2) \in E\}. \tag{2.13}$$

**Theorem 2.13 — Fubini's Theorem.** In the setting above, suppose that $f(x_1, x_2)$ is an integrable function on $(X_1 \times X_2, \mu_1 \times \mu_2)$. Then

- For almost every $x_2 \in X_2$, the slice $f^{x_2}(x_1) = f(x_1, x_2)$ is integrable on $(X_1, \mu_1)$.
- $\int_{X_1} f(x_1, x_2)\mathrm{d}\mu_1$ is an integrable function on $X_2$.
- We can exchange integrals as followed
$$\int_{X_2}\left(\int_{X_1} f(x_1, x_2)\mathrm{d}\mu_1\right)\mathrm{d}\mu_2 = \int_{X_1}\left(\int_{X_2} f(x_1, x_2)\mathrm{d}\mu_2\right)\mathrm{d}\mu_1 = \int_{X_1\times X_2} f\mathrm{d}\mu_1 \times \mu_2. \tag{2.14}$$

**Remark 2.14** In general, the product space $(X, \mathscr{M}, \mu)$ is not complete. One can define the completion of this space as follows. Let $\overline{\mathscr{M}}$ be the collection of sets of the form $E \cup Z$, where $E \in \mathscr{M}$ and $Z \subset F$ with $F \in \mathscr{M}$ and $\mu(F) = 0$. Also, define $\overline{\mu}(E \cup Z) = \mu(E)$. Then

- $\overline{\mathscr{M}}$ is the smallest $\sigma-$algebra containing $\mathscr{M}$ and all subsets of elements of $\mathscr{M}$ of measure zero.
- The function $\overline{\mu}$ is a measure on $\overline{\mathscr{M}}$, and this measure is complete.

The theorem continues to hold in this completed space.

In the above theorem we assume that the function $f$ is integrable over the product space. We can relax this condition, but instead we need to assume that $f$ is a non-negative measurable function. This gives us the following theorem

**Theorem 2.15 — Tonelli's Theorem.** Suppose that $f(x_1, x_2) : X_1 \times X_2 \to [0, \infty]$ is a non-negative measurable function on $(X_1 \times X_2, \mu_1 \times \mu_2)$. Then

$$\int_{X_2} \left( \int_{X_1} f(x_1, x_2) \mathrm{d}\mu_1 \right) \mathrm{d}\mu_2 = \int_{X_1} \left( \int_{X_2} f(x_1, x_2) \mathrm{d}\mu_2 \right) \mathrm{d}\mu_1 = \int_{X_1 \times X_2} f \mathrm{d}\mu_1 \times \mu_2. \tag{2.15}$$

Combining Fubini's theorem with Tonelli's theorem gives

**Theorem 2.16 — Fubini-Tonelli Theorem.** If $f$ is a measurable function, then

$$\int_{X_1} \left( \int_{X_2} |f(x_1, x_2)| \mathrm{d}\mu_2 \right) \mathrm{d}\mu_1 = \int_{X_2} \left( \int_{X_1} |f(x_1, x_2)| \mathrm{d}\mu_1 \right) \mathrm{d}\mu_2 = \int_{X_1 \times X_2} |f| \mathrm{d}\mu_1 \times \mu_2. \tag{2.16}$$

Besides if any one of these integrals is finite, then

$$\int_{X_1} \left( \int_{X_2} f(x_1, x_2) \mathrm{d}\mu_2 \right) \mathrm{d}\mu_1 = \int_{X_2} \left( \int_{X_1} f(x_1, x_2) \mathrm{d}\mu_1 \right) \mathrm{d}\mu_2 = \int_{X_1 \times X_2} f \mathrm{d}\mu_1 \times \mu_2.$$

The absolute value of $f$ in the conditions above can be replaced by either the positive or the negative part of $f$. These forms include Tonelli's theorem as a particular case as the negative part of a non-negative function is zero and has a finite integral. Informally all these conditions say that the double integral of $f$ is well defined, though possibly infinite.

The advantage of the Fubini–Tonelli over Fubini's theorem is that the repeated integrals of the absolute value of $|f|$ may be easier to study than the double integral. As in Fubini's theorem, the single integrals may fail to be defined on a measure $0$ set.

Here is an application of the Fubini-Tonelli theorem:

**Proposition 2.17 — $\mathbb{E}$ and tail probabilities.** Let $\xi$ be a non-negative integrable random variable. Then

$$\mathbb{E}[\xi] = \int_{[0,\infty)} \mathbb{P}(\xi \geq x) \, \mathrm{d}x. \tag{2.17}$$

*Proof.* We have

$$\mathbb{E}[\xi] = \int_{[0,\infty)} x \, \mathrm{d}\mathbb{P}_\xi = \int_{[0,\infty)} \left( \int_0^x \mathrm{d}t \right) \mathrm{d}\mathbb{P}_\xi = \int_{[0,\infty)} \mathbb{P}(\xi \geq t) \, \mathrm{d}t, \tag{2.18}$$

where we applied Fubini's theorem to

$$g(t, x) = \begin{cases} 1, & 0 \leq t \leq x, \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

■

**Exercise 2.18** Generalise the above proof to prove that if $\xi \geq 0$ and $p \geq 1$ then

$$\mathbb{E}[\xi^p] = \int_0^\infty p y^{p-1} \mathbb{P}(\xi \geq y) \, dy \tag{2.20}$$

Verify the formula 2.17 and 2.20 for some simple distributions, e.g. the exponential $\mathrm{Exp}(\lambda)$ distribution with density $f(x) = \lambda e^{-\lambda x}$.

## 2.4 Jensen's Inequality and $L^p$ Spaces

For a random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ we want to study how integrable it is. We define the following notions of integrability:

**Definition 2.19 — The space $\mathscr{L}^p(\Omega)$.** Let $p \in [1, \infty)$.

- We say that $\xi$ is in $\mathscr{L}^p(\Omega)$ if $\mathbb{E}(|\xi|^p)$ exists and is finite. For such $\xi$ we define the function

$$\|\,.\,\|_{\mathscr{L}^p(\Omega)} : \mathscr{L}^p(\Omega) \to \mathbb{R}_{\geq 0} \tag{2.21}$$

$$\xi \mapsto (\mathbb{E}(|\xi|^p))^{1/p} \tag{2.22}$$

- We also say that $\xi \in \mathscr{L}^\infty(\Omega)$ if there is an $M < \infty$ such that $|\xi(\omega)| \leq M$ almost everywhere. For such $\xi$ we define the function

$$\|\,.\,\|_{\mathscr{L}^\infty(\Omega)} : \mathscr{L}^\infty(\Omega) \to \mathbb{R}_{\geq 0} \tag{2.23}$$

$$\xi \mapsto \inf\{M \geq 0 \,|\, |\xi| \leq M \text{ almost everywhere }\} \tag{2.24}$$

If there is no ambiguity, we drop the $\Omega$ when writing $\mathscr{L}^p$ spaces.

We would like to address two central questions:

1. Do we know anything about the inclusions of $\mathscr{L}^p$ spaces?
2. Do we know any information about the $\mathscr{L}^p$ spaces itself?

The ultimate goal is to establish that $\mathscr{L}^p(\Omega)$ is almost a normed vector space, so that we can carry further analysis using tools from functional analysis. Before we address the above questions, let us recall the following important inequality in analysis:

### 2.4.1 Convex Functions and Jensen Inequality

We recall some notions of convexity.

**Definition 2.20 — Convexity.**

- A set $E \subseteq \mathbb{R}^n$ [a] is convex if, for all $x, y \in \Omega$ and $\lambda \in [0, 1]$, the point $(1-\lambda)x + \lambda y \in \Omega$.
- Let $E \subseteq \mathbb{R}^n$ be a convex set. A function $g : E \to \mathbb{R}$ is convex (downward) if, for all $x, y \in \Omega$ and $\lambda \in [0, 1]$,

$$g((1-\lambda)x + \lambda y) \leq (1-\lambda)g(x) + \lambda g(y) \tag{2.25}$$

- Let $E \subseteq \mathbb{R}^n$ be a convex set. A function $g : E \to \mathbb{R}$ is concave (upward) if $-g$ is convex.

[a] You can safely assume $n = 1$ for this section, and treat any gradient/Hessian as derivatives.

For the following analysis we assume $E = \mathbb{R}^n$ for simplicity, although removing this assumption won't affect much. We recall the following characterisations of a convex function:

**Proposition 2.21 — Characterisations of a convex function.** Let $g : \mathbb{R}^n \to \mathbb{R}$

- If $g \in C^1(\mathbb{R}^n)$ (i.e. first-time continuously differentiable), then $g$ is convex iff for all $x, x_0 \in \mathbb{R}^n$, we have

$$g(x) \geq g(x_0) + \nabla g(x_0)^\top (x - x_0) \tag{2.26}$$

  where $\nabla g(x_0)$ is the gradient (total derivative) of $g$ at $x_0$.
- If $g \in C^2(\mathbb{R}^n)$ (i.e. second-time continuously differentiable), then $g$ is convex iff for all $x \in \mathbb{R}^n$, we have the Hessian $\nabla\nabla^\top g(x)$ being positive semi-definite.

**Exercise 2.22** Prove the above characterisations.

We also recall the following theorem

**Proposition 2.23 — Existence of subgradient.** Let $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for all $x_0 \in \mathbb{R}^n$, there is a vector $v \in \mathbb{R}^n$ (depending on $x_0$) such that for all $x \in \mathbb{R}^n$, we have

$$g(x) \geq g(x_0) + v^\top (x - x_0) \tag{2.27}$$

Any vectors $v$ satisfying (2.27) is called a subgradient of $g$ at $x_0$.

The proof is beyond our scope.

With the above lemma we can prove the Jensen's inequality for expectation:

**Theorem 2.24 — Jensen's Inequality.** Let $\xi$ be an integrable random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable, convex downward function. Then

$$g(\mathbb{E}[\xi]) \leq \mathbb{E}[g(\xi)] \tag{2.28}$$

*Proof.* If $g(x)$ is convex downward, for each $x_0 \in \mathbb{R}$ there is a number $v$ such that

$$g(x) \geq g(x_0) + v(x - x_0) \tag{2.29}$$

for all $x \in \mathbb{R}$. Putting $x = \xi$ and $x_0 = \mathbb{E}[\xi]$, we find that

$$g(\xi) \geq g(\mathbb{E}[\xi]) + v(\xi - \mathbb{E}[\xi]) \tag{2.30}$$

and by taking expectation for both sides we get $g(\mathbb{E}[\xi]) \leq \mathbb{E}[g(\xi)]$. ■

### 2.4.2 Inclusions of $\mathscr{L}^p$ spaces

We can immediately observe that if $\xi \in \mathscr{L}^\infty(\Omega)$ then $\xi \in \mathscr{L}^p(\Omega)$ for any $p \in [1, \infty)$. To show this we let $E$ be the set $\{\omega \in \Omega \,|\, |\xi(\omega)| > M\}$. Then $\mathbb{P}(E) = 0$. Therefore

$$\mathbb{E}(|\xi|^p) = \mathbb{E}(|\xi|^p \chi_{\Omega \setminus E}) \leq M^p \mathbb{P}(E) = M^p < \infty \tag{2.31}$$

We therefore have $\mathscr{L}^\infty(\Omega) \subseteq \mathscr{L}^p(\Omega)$ for all $p \in [1, \infty)$.

Let us also compare $L^p(\Omega)$ and $L^q(\Omega)$ for $p < q$. From the Jensen's inequality we can prove the following:

**Corollary 2.25 — Lyapunov's Inequality.** If $0 < p < q < \infty$,

$$(\mathbb{E}[|\xi|^p])^{1/p} \leq (\mathbb{E}[|\xi|^q])^{1/q} \tag{2.32}$$

and therefore $\mathscr{L}^q(\Omega) \subseteq \mathscr{L}^p(\Omega)$

**Hint.** *Consider* $f(x) = x^{q/p}$

*Proof.* To prove this, let $r = q/p$. Then putting $\eta = |\xi|^p$ and applying Jensen's inequality to $g(x) = |x|^r$ (check that it is convex), we obtain $|\mathbb{E}[\eta]|^r \leq \mathbb{E}[|\eta|^r]$, i.e.

$$(\mathbb{E}[|\xi|^p])^{q/p} \leq \mathbb{E}[|\xi|^q]$$

from which the result follows. Therefore if $\mathbb{E}[|\xi|^q] < \infty$ then $\mathbb{E}[|\xi|^p] = (\mathbb{E}[|\xi|^q])^{p/q} < \infty$, and then $\xi \in \mathscr{L}^p(\Omega)$. ■

**Exercise 2.26** Prove that composition of convex function is convex. Check that $g_1(x) = x^r$ is convex whenever $x \geq 0$, and that $g_2(x) = |x|$ is convex whenever $x \in \mathbb{R}$. Conclude that $g(x) = |x|^r$ is convex.

The following chain of inequalities among absolute moments is a consequence of Lyapunov's inequality:

$$\mathbb{E}[|\xi|] \leq (\mathbb{E}[|\xi|^2])^{1/2} \leq \cdots \leq (\mathbb{E}[|\xi|^n])^{1/n} \leq \ldots$$

or equivalently

$$\|\xi\|_{\mathscr{L}^1(\Omega)} \leq \|\xi\|_{\mathscr{L}^2(\Omega)} \leq \cdots \leq \|\xi\|_{\mathscr{L}^n(\Omega)} \leq \ldots \leq \|\xi\|_{\mathscr{L}^\infty(\Omega)}.$$

which yields

$$\mathscr{L}^\infty(\Omega) \subseteq \ldots \subseteq \mathscr{L}^n(\Omega) \subseteq \mathscr{L}^2(\Omega) \subseteq \mathscr{L}^1(\Omega)$$

**Remark 2.27** As a warning, Lyapunov inequality is only true when $\mathbb{P}$ is a **finite** (e.g. probability measure). It is NOT true if we try to generalise the definition of $\mathscr{L}^p$ spaces to other measure spaces.

We will later show that, indeed when $p \nearrow \infty$ we have $\|\xi\|_{\mathscr{L}^p(\Omega)} \nearrow \|\xi\|_{L^\infty(\Omega)}$. We will leave it as an exercise for the next part.

We also define the notion of moments

**Definition 2.28 — Moments.** Let $\xi \in \mathscr{L}^p(\Omega)$. For integers $0 \leq k \leq p$, we define the its $k$-th moments being equal to $\mathbb{E}(\xi^k)$.

### 2.4.3 $\mathscr{L}^p(\Omega)$ and its seminorm

We note that for all $p \in [1, \infty]$, the space $\mathscr{L}^p(\Omega)$ is a *vector space*:

- Let $p \in [1, \infty)$. Notice by convexity, for all $a, b \in \mathbb{R}$,

  $$\left|\frac{a+b}{2}\right|^p \leq \frac{|a|^p + |b|^p}{2} \iff |a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$$

  We therefore have, for all $\xi, \eta \in \mathscr{L}^p(\Omega)$ and $\lambda \in \mathbb{R}$,

  $$\mathbb{E}[|\xi + \lambda\eta|^p] \leq 2^{p-1}(\mathbb{E}[|\xi|^p] + |\lambda|\mathbb{E}[|\eta|^p]) < \infty$$

  and hence $\xi + \lambda\eta \in \mathscr{L}^p(\Omega)$.

- Let $\xi, \eta \in \mathscr{L}^\infty(\Omega)$, such that $|\xi|, |\eta| \leq M$ almost everywhere. Define $E_1 := \{\omega \,|\, |\xi| > M\}$ and $E_2 := \{\omega \,|\, |\eta| > M\}$. For $\omega \in \Omega \setminus (E_1 \cup E_2)$, we have

  $$|\xi(\omega) + \lambda\eta(\omega)| \leq (1+\lambda)M < \infty \tag{2.33}$$

  Finally note that $E_1 \cup E_2$ has measure zero, so $\xi + \lambda\eta \in \mathscr{L}^\infty(\Omega)$.

Recall, in the definition of $\mathscr{L}^p(\Omega)$, we have define the function $\|\,.\,\|_{\mathscr{L}^p(\Omega)}$. One would hope that this function is a norm of $\mathscr{L}^p(\Omega)$ in the following sense:

**Definition 2.29 — Seminorm and Norm.** Consider a vector space $V$, and let $\|\,.\,\| : V \to \mathbb{R}_{\geq 0}$ be a function. $\|\,.\,\|$ is a **seminorm** if it is

- Absolutely homogeneous: for all $\lambda \in \mathbb{R}, v \in V$

  $$\|\lambda\xi\| = |\lambda|\,\|\xi\| \tag{2.34}$$

- Triangle inequality: for all $u, v \in V$,

  $$\|u+v\| \leq \|u\| + \|v\| \tag{2.35}$$

In particular if it holds that $\|v\| = 0 \iff v = 0$, then $\|\,.\,\|$ is a norm.

Unfortunately this is not true: in fact $\|\xi\|_{\mathscr{L}^p(\Omega)} = 0 \iff \xi = 0$ **almost** everywhere. Nevertheless, we can prove that $\|\,.\,\|_{\mathscr{L}^p(\Omega)}$ is a **semi**-norm. To begin, it is trivial to see from linearity that $\|\,.\,\|_{L^p(\Omega)}$ is absolutely homogeneous. The main task is hence to check if triangle inequality is satisfied for $\|\,.\,\|_{L^p(\Omega)}$. Showing this for the case $p = \infty$ is not hard: we can run through the steps in proving (2.33) to show that

$$|\xi(\omega) + \eta(\omega)| \leq \|\xi\|_{\mathscr{L}^\infty(\Omega)} + \|\eta\|_{\mathscr{L}^\infty(\Omega)}$$

almost everywhere, and therefore $\|\xi+\eta\|_{L^\infty}$ as an infinmum must not be greater than $\|\xi\|_{\mathscr{L}^\infty(\Omega)} + \|\eta\|_{\mathscr{L}^\infty(\Omega)}$.

We now try to prove the triangle inequality for the case when $p < \infty$. We need the following important stepping stones:

**Proposition 2.30 — Young's Inequality.** Let $p \in (1,\infty)$ and $q$ such that $(1/p)+(1/q) = 1$. Then $\forall a, b > 0$

$$ab \leq \frac{a^p}{p} + \frac{a^q}{q} \tag{2.36}$$

**Hint.** *Consider* $f(x) = -\ln x$.

*Proof.* Note that $f(x) = -\ln x$ is a convex function for all $x > 0$, we have

$$-\ln\left(\frac{a^p}{p} + \frac{a^q}{q}\right) \leq -\frac{1}{p}\ln a^p - \frac{1}{q}\ln b^q$$

which simplifies to our Young's inequality (2.36). ■

**Proposition 2.31 — Hölder's Inequality.** Let $p \in [1,\infty]$. Define $q \in [1,\infty]$ such that $(1/p)+(1/q) = 1$ (in particular $q = \infty$ when $p = 1$). If $\xi \in \mathscr{L}^p(\Omega)$ and $\eta \in \mathscr{L}^q(\Omega)$, then

$$\|\xi\eta\|_{\mathscr{L}^1(\Omega)} := \mathbb{E}[|\xi\eta|] \leq \|\xi\|_{\mathscr{L}^p(\Omega)}\,\|\eta\|_{\mathscr{L}^q(\Omega)} \tag{2.37}$$

*Proof.* We first prove a more trivial case when $p = 1$ and $q = \infty$: define the set $E := \left\{\omega \,|\, |\xi(\omega)| > \|\xi\|_{L^\infty(\omega)}\right\}$, which has measure zero. Then, similar to the inequality (2.31), we have

$$\mathbb{E}[|\xi\eta|] = \mathbb{E}[|\xi\eta|\chi_{\Omega\setminus E}] \leq \|\xi\|_{L^\infty(\Omega)}\,\mathbb{E}[|\eta|] = \|\xi\|_{L^\infty(\Omega)}\,\|\eta\|_{L^1(\Omega)} \tag{2.38}$$

We can similarly prove for the case when $p = \infty$ and $q = 1$, so wlog assume $p, q < \infty$. We can apply Young's inequality (2.36) to show that for all $\lambda > 0$, we have

$$\mathbb{E}\left[|\lambda\xi\eta|\right] \leq \mathbb{E}\left[\frac{|\lambda|^p|\xi|^p}{p} + \frac{|\eta|^q}{q}\right] = \frac{\lambda^p\mathbb{E}[|\xi|^p]}{p} + \frac{\mathbb{E}[|\eta|^q]}{q}$$

Divide both sides by $\lambda$ yields

$$\mathbb{E}\left[|\xi\eta|\right] = \frac{\lambda^{p-1}\mathbb{E}[|\xi|^p]}{p} + \frac{\mathbb{E}[|\eta|^q]}{\lambda q} =: f_{\xi,\eta}(\lambda) \tag{2.39}$$

Here comes the important part - since (2.39) is true for all $\lambda > 0$, *we can make the inequality as tight as possible by minimising the RHS*. Specifically, we can rewrite (2.39) as

$$\mathbb{E}\left[|\xi\eta|\right] \leq \inf_{\lambda>0} f_{\xi,\eta}(\lambda) \tag{2.40}$$

Let us carefully find the minimum of $f_{\xi,\eta}(\lambda)$. We note that

$$f'_{\xi,\eta}(\lambda) = \underbrace{\frac{p-1}{p}}_{=1/q}\mathbb{E}[|\xi|^p]\lambda^{p-2} - \frac{1}{q}\mathbb{E}[|\eta|^q]\lambda^{-2} = \frac{\mathbb{E}[|\xi|^p]}{q\lambda^2}\left(\lambda^p - \frac{\mathbb{E}[|\eta|^q]}{\mathbb{E}[|\xi|^p]}\right) \tag{2.41}$$

So we have $f'_{\xi,\eta}(\lambda) = 0 \iff \lambda = \lambda^* := \left(\frac{\mathbb{E}[|\eta|^q]}{\mathbb{E}[|\xi|^p]}\right)^{1/p}$. In particular, $f'_{\xi,\eta}(\lambda) < 0$ whenever $\lambda < \lambda^*$ and $f'_{\xi,\eta}(\lambda) > 0$ whenever $\lambda > \lambda^*$, so indeed $f_{\xi,\eta}(\lambda^*)$ is indeed the global minimum of $f_{\xi,\eta}(\lambda)$. Plugging into (2.41), we have

$$\mathbb{E}[|\xi\eta|] \leq \left(\frac{\mathbb{E}[|\eta|^q]}{\mathbb{E}[|\xi|^p]}\right)^{\frac{p-1}{p}}\frac{\mathbb{E}[|\xi|^p]}{p} + \left(\frac{\mathbb{E}[|\eta|^q]}{\mathbb{E}[|\xi|^p]}\right)^{-1/p}\frac{\mathbb{E}[|\eta|^q]}{q} = \|\xi\|_{\mathscr{L}^p(\Omega)}\,\|\eta\|_{\mathscr{L}^q(\Omega)} \tag{2.42}$$

completing the proof. ■

We can finally utilise the Hölder's inequality to prove the triangle inequality.

**Proposition 2.32 — Triangle inequality for $\|.\|_{\mathscr{L}^p}$/Minkowski's Inequality.** If $\mathbb{E}[|\xi|^p] < \infty$, $\mathbb{E}[|\eta|^p] < \infty, 1 \leq p \leq \infty$, then we have $\mathbb{E}[|\xi+\eta|^p] < \infty$ and

$$(\mathbb{E}[|\xi+\eta|^p])^{1/p} \leq (\mathbb{E}[|\xi|^p])^{1/p} + (\mathbb{E}[|\eta|^p])^{1/p}.$$

**Hint.** *Note $|\xi+\eta|^p \leq |\xi+\eta|^{p-1}(|\xi|+|\eta|)$ and use Holder's inequality.*

*Proof.* We only need to take care the case when $p < \infty$. The Hölder inequality says

$$\mathbb{E}[|\xi+\eta|^p] \leq \mathbb{E}[|\xi+\eta|^{p-1}(|\xi|+|\eta|)] \leq \left(\mathbb{E}\left[|\xi+\eta|^{p-1})^{\frac{p}{p-1}}\right]\right)^{1-1/p} (\|\xi\|_{\mathscr{L}^p(\Omega)} + \|\eta\|_{\mathscr{L}^p(\Omega)}) \tag{2.43}$$

which simplifies to the Minkowski inequality. ■

We have therefore shown that, for all $p \in [1,\infty]$, the function $\|.\|_{\mathscr{L}^p(\Omega)}$ is a seminorm of the vector space $\mathscr{L}^p(\Omega)$. In fact this seminorm induces a norm in the following quotient space $L^p(\Omega) = \mathscr{L}^p(\Omega)/\sim$, where $\sim$ is the following equivalent relation

$$f \sim g \iff f = g \ \mathbb{P}\text{-almost everywhere.} \tag{2.44}$$

**Exercise 2.33** First check that $\sim$ is an equivalent relation. Then check that $L^p(\Omega)$ is a vector space by defining a suitable addition and scalar multiplication. (Remember to check that these operations are well defined!)

The norm is given by the following: writing $[f]_\sim$ being the equivalent class containing $f$, then

$$\|[f]_\sim\|_{L^p(\Omega)} = \|f\|_{\mathscr{L}^p(\Omega)} \tag{2.45}$$

**Exercise 2.34** First check that the above norm is well-defined in the sense that the value is independent of the choice of elements in $[f]_\sim$. Then check that it is a norm.

For the following chapters, we abuse notations and replace $\|[f]_\sim\|_{L^p(\Omega)}$ with $\|f\|_{L^p}$. Just note that if we have $\|f\|_{L^p} = 0$ only imply that $f = 0$ almost surely. Since $L^p$ is a normed vector space, the norm induces a metric $d_{L^p}(\xi,\eta) = (\mathbb{E}[|\xi-\eta|^p])^{1/p}$, and we can define a notion of convergence from this metric:

**Definition 2.35 — $L^p$ convergence.** The sequence $\xi_1, \xi_2, \ldots$ of (equivalence classes of) random variables converges in $L^p$ (moments of order $p$) to the random variable $\xi$ if as $n \to \infty$

$$(\mathbb{E}[|\xi_n - \xi|^p])^{1/p} \to 0 \iff \mathbb{E}[|\xi_n - \xi|^p] \to 0 \tag{2.46}$$

**Remark 2.36 — Reverse Triangle Inequality.** Note that if the sequence $(\xi_i)_{i\geq 1}$ converges in $L^p$ to $\xi$, then as $i \to \infty$,

$$0 \leq |\,\|\xi_i\|_{L^p} - \|\xi\|_{L^p}\,| \leq \|\xi_i - \xi\|_{L^p} \to 0 \tag{2.47}$$

i.e. $\|\xi_i\|_{L^p} \to \|\xi\|_{L^p}$.

Let us also define the following space for convenience

**Definition 2.37 — The $L^{\infty-}(\Omega)$ space.** Define

$$L^{\infty-}(\Omega) = \bigcap_{p\geq 1} L^p(\Omega) \tag{2.48}$$

Note that $L^\infty(\Omega) \neq L^{\infty-}(\Omega)$. For instance, any random variable $X$ which are normally distributed are in $L^{\infty-}(\Omega)$ but not in $L^\infty(\Omega)$.

## 2.5 Tail Bounds

Most large samples results concern about extreme events, e.g. whether the value of a random variable deviates from its mean. This section builds necessary tools to develop upper bounds of the probability of such events. These bounds are usually called the "tail bounds" since they corresponds to the "tail" of the densities of random variables (if they exist). In particular, we will see how the tail bounds are related to integrability of the random variables.

**Example 2.38 — Tail bounds for specific random variables.** To motivate, consider the following random variables living on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which has zero mean (expectation). We would like to bound the probability of tail event $\mathbb{P}(X > c)$ for $c \gg 1$ [a].

1. $\xi_1$ following a normal distribution $\mathbf{N}(0,1)$ with density

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}$$

For such case, we have

$$\mathbb{P}(\xi_1 > c) = \int_c^\infty \frac{1}{\sqrt{2\pi}} \frac{x}{c} \exp\left(-\frac{x^2}{2}\right) dx \le \int_c^\infty \frac{1}{\sqrt{2\pi}} \frac{x}{c} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{c\sqrt{2\pi}} \exp\left(-\frac{c^2}{2}\right) =: r_1(c)$$

2. $\xi_2$ following a double exponential (Laplace) distribution Laplace$(1)$ with density

$$f_2(x) = \frac{1}{2} e^{-|x|}$$

For such case, we have

$$\mathbb{P}(\xi_2 > c) = \int_c^\infty \frac{1}{2} \exp(-x)\, dx = \frac{1}{2} \exp(-c) =: r_2(c)$$

3. $\xi_3$ following a Cauchy distribution Cauchy$(1)$ with density

$$f_3(x) = \frac{1}{\pi(1+x^2)}$$

For such case, we have

$$\mathbb{P}(\xi_3 > c) = \frac{1}{2} - \frac{1}{\pi} \arctan c = \frac{1}{\pi} \arctan \frac{1}{c} \le \frac{1}{\pi c} =: r_3(c)$$

From this it is clear that $\mathbb{P}(X > c)$ decays much faster for normal distribution than double exponential distribution, and Cauchy distribution admits the slowest decay. In particular, we have $r_3(c) \gg r_2(c) \gg r_1(c)$ in the sense that $r_1(c)/r_2(c) \to 0$ and $r_2(c)/r_3(c) \to 0$ when $c \to \infty$.

[a] This means much greater than. We will also use this when performing formal asymptotic analysis.

**Exercise 2.39 — Moments for some distributions.** Verify the following observations for the above random variables concerning their even moments ($\xi_1, \xi_2$ has zero odd moments, why?)

1. $\xi_1$ has $2k$-th moments $m_{1,k} = (2k-1)!! := (2k-1) \times \ldots \times 3 \times 1 = (2k)!/(2^k k!)$ for all $k \in \mathbb{Z}_{\ge 1}$.
2. $\xi_2$ has $2k$-th moments $m_{2,k} = (2k)!$, and
3. $\xi_3$ has $2k$-th moments $m_{3,k} = \infty$.

Therefore, we see that $\infty = m_{3,k} \gg m_{2,k} \gg m_{1,k}$ as $k \to \infty$ (in the sense that as $k \to \infty$ we have

$m_{1,k}/m_{2,k} \to 0$.) We therefore suspect that there is a connection between the tail bounds and the growth of moments.

**Exercise 2.40 — Sharpness of tail bounds.** Note that the tail bounds we have derived for $\mathbb{P}(\xi_1 > c)$ and $\mathbb{P}(\xi_3 > c)$ are not sharp. Try to prove the lower bounds for them and show that both $\mathbb{P}(\xi_1 > c)/\mathbb{P}(\xi_2 > c)$ and $\mathbb{P}(\xi_2 > c)/\mathbb{P}(\xi_3 > c) \to 0$ as $c \to \infty$.

**Hint.** *To control $\mathbb{P}(\xi_3 > c)$ consider the Taylor series (with remainder) for $\arctan x$. To control $\mathbb{P}(\xi_1 > c)$, use a suitable integration by part to obtain an asymptotic expansion of $\mathbb{P}(\xi_1 > c)$. Here is another slicky way to derive the desired asymptotic expansion: consider the function $M(c) := \lambda\mathbb{P}(\xi_1 > c)/\phi(c)$ with $\phi(c) = \exp(-c^2/2)/\sqrt{2\pi}$ (the Mill's ratio). Verify that*

$$M(c) = \int_0^\infty \exp\left(-x - \frac{x^2}{2c^2}\right) dx \tag{2.49}$$

*Then expand $\exp(-x^2/2c^2)$ and exchange integral and infinite sum **with care** to conclude that*

$$\mathbb{P}(\xi_1 > c) = c^{-1}\phi(c) \sum_{k=0}^{\infty} \frac{(-1)^k (2k-1)!!}{c^{2k}} \tag{2.50}$$

*The inequalities derived from truncating the above expansion is called the Mill-Ratio inequalities. One of the example is:*

$$\left(\frac{1}{c} - \frac{1}{c^3}\right)\phi(c) \le \mathbb{P}(\xi_1 > c) \le \frac{1}{c}\phi(c) \tag{2.51}$$

Before we delve into further discussions, let us define the following notions of moments for our convenience

**Definition 2.41 — Central Moments.** The $k$-th central moment (for $k \ge 1$) of a random variable $\xi$ is the expectation $\mathbb{E}((\xi - \mathbb{E}(\xi))^k)$ whenever $\mathbb{E}(|\xi|^k) < \infty$. In particular the first central moment is zero. Some of the central moments have special names:

- The 2nd central moment $\mathbb{V}(\xi) := \mathbb{E}((\xi - \mathbb{E}(\xi))^2)$ is called the **variance**.
- The 3rd central moment $\mathbb{E}((\xi - \mathbb{E}(\xi))^3)$ is called the **skewness**.
- The 4th central moment $\mathbb{E}((\xi - \mathbb{E}(\xi))^4)$ is called the **kurtosis**.

With the above notions, we can state the central inequality which we use for deriving useful tail bounds.

**Theorem 2.42 — Markov Inequality.** Let $\xi$ be a non-negative integrable random variable and $c > 0$ being a constant. Then

$$\mathbb{P}(\xi \ge c) \le \frac{\mathbb{E}[\xi]}{c}. \tag{2.52}$$

*Proof.*

$$\mathbb{E}[\xi] \ge \mathbb{E}[\xi \cdot \chi_{\xi \ge c}] \ge c\mathbb{E}[\chi_{\xi \ge c}] = c\mathbb{P}(\xi \ge c).$$

■

**Remark 2.43** A natural generalisation of the Markov inequality is the following: let $\xi$ be a random variable, $g$ be a Borel function such that $g(\xi)$ is non-negative. Assume $c > 0$ is a constant and that $\mathbb{E}[g(\xi)]$ exists, then

$$\mathbb{P}(g(\xi) \ge c) \le \frac{\mathbb{E}[g(\xi)]}{c}. \tag{2.53}$$

Let us interrupt our discussion of tail bound by proving an interesting result regarding $L^p$ norms:

**Example 2.44 — Limits of $L^p$ norms.** Let the random variable $\xi$ defined on probability space $(\Omega, \mathscr{F}, \mathbb{P})$ be in $L^\infty(\Omega)$. We want to establish the fact that $\lim_{p\to\infty} \|\xi\|_{L^p} = \|\xi\|_{L^\infty}$. Consider an arbitrary increasing sequence $(p_i)_{i\geq 1}$ such that $p_i \to \infty$ as $i \to \infty$. The Lyapunov inequality shows that the sequence $(\|\xi\|_{L^{p_i}})_{i\geq 1}$ is a monotonic increasing sequence upper bounded by $\|\xi\|_{L^\infty}$. Therefore the sequence has a limit satisfying

$$\limsup_{i\to\infty} \|\xi\|_{L^{p_i}} \leq \|\xi\|_{L^\infty} \tag{2.54}$$

We now show that the limit of sequence $(\|\xi\|_{L^{p_i}})_{i\geq 1}$ is lower bounded by $\|\xi\|_{L^\infty}$. Let $\alpha \in (0, \|\xi\|_{L^\infty})$, then by Markov inequality, for all $i$

$$\alpha^{p_i}\mathbb{P}(|\xi| > \alpha) \leq \|\xi\|_{L^{p_i}}^{p_i} \tag{2.55}$$

equivalently

$$\|\xi\| \geq \alpha(\mathbb{P}(|\xi| > \alpha))^{1/p_i} \tag{2.56}$$

Sending $p_i$ **with** $\alpha$ **being fixed** we have

$$\liminf_{i\to\infty} \|\xi\|^{p_i} \geq \alpha \underbrace{\liminf_{i\to\infty} (\mathbb{P}(|\xi| > \alpha))^{1/p_i}}_{=1} = \alpha \tag{2.57}$$

We finally finish by sending $\alpha \to \|\xi\|_{L^\infty}$ to conclude that

$$\liminf_{i\to\infty} \|\xi\|^{p_i} \geq \|\xi\|_{L^\infty} \tag{2.58}$$

as desired.

How powerful Markov inequality (2.53) is for proving tail bounds for specific distribution depend on the integrability of $\xi$. For instance, if we know that $\xi \in L^p(\Omega)$ when $p \geq 1$, then

**Corollary 2.45** For all $\varepsilon > 0$, we have

$$\mathbb{P}(|\xi - \mathbb{E}[\xi]| \geq \varepsilon) = \mathbb{P}(|\xi - \mathbb{E}[\xi]|^p \geq \varepsilon^p) \leq \frac{\mathbb{E}[|\xi - \mathbb{E}[\xi]|^p]}{\varepsilon^p}. \tag{2.59}$$

We emphasise the special case when $p = 2$: in such case we have the **Chebyshev Inequality**:

$$\mathbb{P}(|\xi - \mathbb{E}[\xi]| \geq \varepsilon) \leq \frac{\mathbb{V}[\xi]}{\varepsilon^2} \tag{2.60}$$

### 2.5.1 Chernoff Bound and Moment Generating Function (MGF)

For the case when $\xi \in L^{\infty-}(\Omega)$ and that the $k$-th moment does not grow "too quick", one may consider choosing an optimal $p$ such that the RHS of (2.59) is minimised. This is rarely done in practice. Instead, we consider the moment generating function:

**Definition 2.46 — Moment Generating Function (MGF).** The moment generating function of a random variable $\xi$ is defined as

$$M_\xi(t) = \mathbb{E}[\exp(tX)] = \int_{\mathbb{R}} e^{tx} dF_\xi(x) \tag{2.61}$$

Moment generating function does not necessary exists for all values of $t \in \mathbb{R}$ (e.g. for random variables $\xi$ with Cauchy distribution, $M_\xi(t) = \infty$ for all $t \neq 0$, and is equal to 1 for $t = 0$). However, if we show that $M_\xi(t) < \infty$ for a small neighborhood of zero (say $t \in (-h, h)$), then we can show the following

**Corollary 2.47 — Chernoff (Exponential Chebyshev) Inequality.** Let $\xi$ be a non-negative random variable, then for all $\varepsilon > 0, t \in (0, h)$

$$\mathbb{P}(\xi \geq \varepsilon) = \mathbb{P}(e^{t\xi} \geq e^{t\varepsilon}) \leq \frac{\mathbb{E}[e^{t\xi}]}{e^{t\varepsilon}} = \frac{M_X(t)}{e^{t\varepsilon}} \tag{2.62}$$

The existence of moment generating function at a neighborhood of zero captures how the $k$-th moments grow. In fact, if $\xi \in L^{\infty-}(\Omega)$, then one can show that:

**Proposition 2.48 — Generating moments from MGF.** $M_\xi(t)$ is smooth (infinitely diffentiable) at $t = 0$ with, for all $k \in \mathbb{Z}_{\geq 0}$

$$\mathbb{E}[X^\xi] = \left.\frac{\mathrm{d}M_\xi}{\mathrm{d}t}\right|_{t=0} \tag{2.63}$$

**Hint.** *Use differentiation under integral.*

*Proof.* We show for the case when $k = 1$ - the other cases can be generalised by induction etc. We note that for all $x$, $e^{tx}$ is differentiable, and for all $t \in (-h/2, h/2)$, $e^{tx} \in L^1(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P}_\xi)$. We finally check that the function $|x|e^{tx}$ is integrable under $\mathbb{P}_\xi$ (i.e. $\mathbb{E}[|\xi|e^{t\xi}] < \infty$). Note that for all $t \in (-h/2, h/2)$

$$\mathbb{E}[|\xi|e^{t\xi}] \overset{\text{(Hölder)}}{\leq} \|\xi\|_{L^2} M_\xi(2t) < \infty \tag{2.64}$$

So we can use differentiation under integral to conclude that $\mathbb{E}[\xi] = M'_\xi(0)$. ■

Moreover, we can show

**Proposition 2.49 — Bounds for moments.** If in addition $\xi \geq 0$, then for all $t \in (-h, h)$

$$\mathbb{E}[\xi^k] \leq \left(\frac{k}{te}\right)^k M_\xi(t) \tag{2.65}$$

*Proof.* We utilise the inequality $1 + x \leq e^x$ for $x \in \mathbb{R}$ (check by e.g. calculus!), then we have $x \leq e^{x-1}$. Replacing $x$ with $t\xi/k$ yields

$$\frac{t\xi}{k} \leq \exp\left(\frac{t\xi}{k} - 1\right) \tag{2.66}$$

Raising power by $k$ and take expectation completes the proof. ■

This inequality shows that the if we can control the moment generating function $M_\xi(t)$, then we can control the growth of moments as $k \to \infty$.

Let us use the following example to illustrate how Markov inequality can be used in proving several tail bounds

**Example 2.50 — Tail bounds for Gaussian.** Consider $\xi$ following a standard normal distribution $\mathbf{N}(0, 1)$. The $(2k)$-th moments of $\xi$ is given by $(2k-1)!!$, and therefore we have, for all $k \geq 1$ and $c > 0$,

$$\mathbb{P}(X > c) \leq \frac{(2k-1)!!}{c^{2k}} \tag{2.67}$$

As a warning, even though using larger $k$ will lead to a faster rate of decay as $c \to \infty$, the numerator is also larger, so it is harder to use the bound for practical applications. We can obtain a much sharper bound than any of the (2.67) by consider the Chernoff bounds: notice that

$$\mathbb{E}[e^{t\xi}] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx - x^2/2}\, dx = \exp\frac{t^2}{2} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-(x-t)/2}\, dx = \exp\frac{t^2}{2} \tag{2.68}$$

and therefore by Chernoff bound (2.62) we have

$$\mathbb{P}(\xi > c) \leq \exp\left(\frac{t^2}{2} - ct\right) = \exp\left(-\frac{c^2}{2} + \frac{1}{2}(t-c)^2\right) \tag{2.69}$$

Since (2.69) holds for all $t > 0$, we can choose the optimal $t$ such that the RHS is minimised. In our

case we choose $c > 0$ to obtain

$$\mathbb{P}(\xi > c) \leq \exp\left(\frac{t^2}{2} - ct\right) = \exp\left(-\frac{c^2}{2}\right) \tag{2.70}$$

This is almost optimal comparing with our previous Mill-ratio inequalities, in the sense that the RHS is off by a factor of $C/\lambda$, with $C$ being a constant independent of $c$. It is also surprisingly useful in practice.

We make a final remark regarding MGF

**Remark 2.51 — MGF and large sample results.** As seen in previous classes in probability, we can use MGF to establish large-sample results like Central Limit Theorem, since MGF determines all moments of a random variable. However, we will not use this approach here, as MGF really doesn't exist for random variables $\xi \notin L^{\infty-}(\Omega)$, and even if it exists, it will not be defined for all $t \in \mathbb{R}$. Instead, we will consider the characteristic functions (CF) for proving large-sample results, since it exists for random variables (and is uniformly continuous on $\mathbb{R}$). We will discuss this further in Chapter 6.

# 3 More on Random Variables

## 3.1 Transformation of Random Variables

Let us consider the problem of determining the distribution functions of random variables that are functions of other random variables. Let $\xi$ be a random variable with distribution function $F_\xi(x)$ (and density $f_\xi(x)$, if it exists), let $\varphi = \varphi(x)$ be a Borel function and $\eta = \varphi(\xi)$. We have

$$F_\eta(y) = \mathbb{P}(\eta \le y) = \mathbb{P}(\xi \in \varphi^{-1}(-\infty, y]) = \int_{\varphi^{-1}(-\infty,y]} \mathrm{d}F_\xi, \tag{3.1}$$

which expresses the distribution function $F_\eta(y)$ in terms of $F_\xi(x)$ and $\varphi$.

**Example 3.1**

1. (Location-scale family) Let $\eta = a\xi + b, \quad a > 0$. Then

$$F_\eta(y) = \mathbb{P}(\eta \le y) = \mathbb{P}\left(\xi \le \frac{y-b}{a}\right) = F_\xi\left(\frac{y-b}{a}\right). \tag{3.2}$$

2. ($\chi^2_1$ distribution) Let $\eta = \xi^2$. Then it is evident that $F_\eta(y) = 0$ if $y < 0$. while for $y \ge 0$,

$$\begin{aligned} F_\eta(y) = \mathbb{P}(\xi^2 \le y) &= \mathbb{P}(-\sqrt{y} \le \xi \le \sqrt{y}) \\ &= \mathbb{P}_\xi(-\infty, \sqrt{y}] - \mathbb{P}_\xi(-\infty, -\sqrt{y}) \\ &= F_\xi(\sqrt{y}) - F_\xi(-\sqrt{y}) + \mathbb{P}(\xi = -\sqrt{y}). \end{aligned}$$

As a further example, let us prove a result that connects the distribution function and probability distribution of a random variable.

**Proposition 3.2 — Probability Integral Transform.** Let $\xi$ be a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ with distribution function $F_\xi(x)$. Let $U$ be a random variable on $([0,1], \mathscr{B}([0,1]), \mathsf{Leb})$ ($\mathsf{Leb}$ being Lebesgue measure) such that it is uniformly distributed on $[0,1]$ (i.e. $\mathbb{P}_U = \mathsf{Leb}$). Define the right inverse of $F_\xi$ on $[0,1]$:

$$F_\xi^{-1}(y) = \sup\{x \mid F_\xi(x) < y\}. \tag{3.3}$$

and extend so that $F_\xi(0) = -\infty$ and $F_\xi(1) = \infty$. Then $\xi$ has the same distribution function as $F_\xi^{-1}(U)$. In such case we say $F_\xi^{-1}(U)$ is *equally distributed* as $\xi$, or $\xi \overset{d}{=} F_\xi^{-1}(U)$.

**Hint.** *Let us try to gain a better understanding to the definition of the right inverse:*

- *To begin, try to verify if $F_\xi$ is strictly increasing (so that $F_\xi(x)$ is a continuous bijection from $(-\infty, \infty)$ to $(0,1)$), then this right inverse of $F_\xi$ agrees with the inverse $F_\xi^{-1}$.*
- *For this special case, we really have*

$$F_{F_\xi^{-1}}(y) = \mathsf{Leb}\left(\left\{F_\xi^{-1}(U) \le y\right\}\right) = \mathsf{Leb}\left(\{U \le F_\xi(y)\}\right) = F_\xi(y) \tag{3.4}$$

  *The actual proof won't differ too much.*

- *One can prove a simpler version of this theorem, that the random variable $F_\xi(\xi)$ is equally distributed as $U$.*

*Proof.* The only caveat of proof is to justify the second equality, that is to prove the equivalence

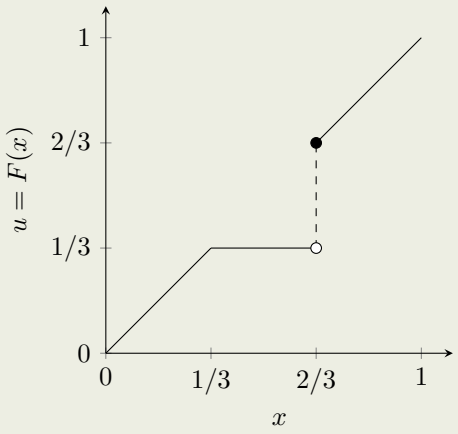
$$F_\xi^{-1}(u) \le y \iff u \le F_\xi(y)$$

($\Leftarrow$) **Assume $u \le F_\xi(y)$.** Then clearly whenever $F_\xi(x) \le u$ then $F_\xi(x) \le F_\xi(y) \implies x \in y$, thus clearly $y$ is an upper bound of the set $\{x \mid F_\xi(x) < u\}$.

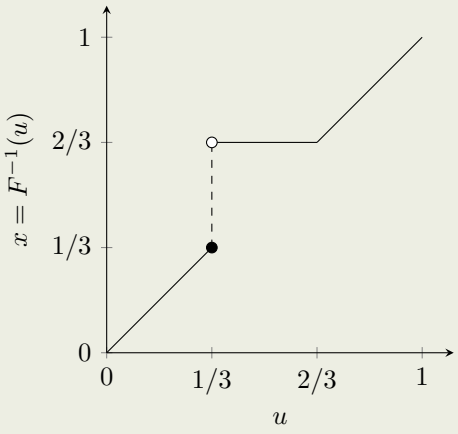
($\Rightarrow$) **Assume $F_\xi^{-1}(u) \le y$ but $u > F_\xi(y)$ by contradiction.** Note in such case we have $y \in \{x \mid F_\xi(x) < u\}$, so $y$ is indeed the *maximum* of $\{x \mid F_\xi(x) < u\}$. In other words, for any $x > y$ we have $F_\xi(x) \ge u$. Consider an arbitrary monotonic decreasing sequence $(x_n)$ with $x_n > y$ that converges to $y$. We then know

that $\lim_{n\to\infty} F_\xi(x_n) = F(y)$ by right continuity, and that $F_\xi(y) \geq u$. This contradicts with our assumption, so we must have $u \leq F_\xi(y)$. ■

**Example 3.3** Suppose that the distribution function $F$ for $\xi$ uniform on $[0, 1/3] \cup [2/3, 1]$ with an atom at $2/3$.




The inverse $F^{-1}(u)$ is shown below




Now we turn to the problem of determining $f_\eta(y)$. Let us suppose that the range of $\xi$ is a (finite or infinite) open interval $I = (a, b)$, and that the function $\varphi = \varphi(x)$, with domain $(a, b)$, is continuously differentiable and either strictly increasing or strictly decreasing. We also suppose that $\varphi'(x) \neq 0, x \in I$. Let us write $h(y) = \varphi^{-1}(y)$ and suppose for definiteness that $\varphi$ is strictly increasing. Then when $y \in \varphi(I)$,

$$\begin{aligned} F_\eta(y) = \mathbb{P}(\eta \leq y) = \mathbb{P}(\varphi(\xi) \leq y) = \mathbb{P}(\xi \leq \varphi^{-1}(y)) = \mathbb{P}(\xi \leq h(y)) \\ = \int_{-\infty}^{h(y)} f_\xi(x)\mathrm{d}x \\ = \int_{-\infty}^{y} f_\xi(h(z))h'(z)\mathrm{d}z. \end{aligned}$$

Therefore,

$$f_\eta(y) = f_\xi(h(y))h'(y). \tag{3.5}$$

Similarly, if $\varphi(x)$ is strictly decreasing,

$$f_\eta(y) = f_\xi(h(y))(-h'(y)). \tag{3.6}$$

Hence in either case

$$f_\eta(y) = f_\xi(h(y))|h'(y)|. \tag{3.7}$$

**Example 3.4** If $\eta = a\xi + b, a \neq 0$, we have

$$h(y) = \frac{y-b}{a} \quad \text{and} \quad f_\eta(y) = \frac{1}{|a|} f_\xi\left(\frac{y-b}{a}\right). \tag{3.8}$$

If $\varphi = \varphi(x)$ is neither strictly increasing nor strictly decreasing, the above formula is inapplicable. However, the following generalisation suffices for many applications.

**Lemma 3.5** Let $\varphi = \varphi(x)$ be defined on the set $\sum_{k=1}^{n}[a_k, b_k]$, continuously differentiable and either strictly increasing or strictly decreasing on each open interval $I_k = (a_k, b_k)$, and with $\varphi'(x) \neq 0$ for $x \in I_k$. Let $h_k = h_k(y)$ be the inverse of $\varphi(x)$ for $x \in I_k$, Then

$$f_\eta(y) = \sum_{k=1}^{n} f_\xi(h_k(y))|h_k'(y)| \cdot I_{D_k}(y), \tag{3.9}$$

where $D_k$ is the domain of $h_k(y)$.

**Example 3.6** Let $\eta = \xi^2$ and take $I_1 = (-\infty, 0), I_2 = (0, \infty)$, and find that $h_1(y) = -\sqrt{y}, h_2(y) = \sqrt{y}$, and therefore

$$f_\eta(y) = \begin{cases} \frac{1}{2\sqrt{y}}[f_\xi(\sqrt{y}) + f_\xi(-\sqrt{y})], & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{3.10}$$

In particular, if $\xi \sim N(0,1)$,

$$f_{\xi^2}(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-y/2}, & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{3.11}$$

A straightforward calculation shows that

$$f_{|\xi|}(y) = \begin{cases} f_\xi(y) + f_\xi(-y), & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{3.12}$$

$$f_{\sqrt{|\xi|}}(y) = \begin{cases} 2y(f_\xi(y^2) + f_\xi(-y^2)), & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{3.13}$$

Let $\xi$ and $\eta$ be random variables with joint distribution $F_{\xi,\eta}(x,y)$, and $\varphi = \varphi(x,y)$ be a Borel function, then

$$F_{\varphi(\xi,\eta)}(z) = \int_{\{x,y:\varphi(x,y)\leq z\}} \mathrm{d}F_{\xi,\eta}(x,y). \tag{3.14}$$

Plan: Should we also include some discussion about multivariable transformations?

## 3.2 Independent and Uncorrelated Random Variables

### 3.2.1 Independence

**Definition 3.7 — (Mutual) Independence.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a measure space.

- Assume there are events $A_1, A_2, \ldots$. The finite collection of events $\{A_1, \ldots, A_n\}$ is independent if $\mathbb{P}(\cap_{i=1}^n A_n) = \prod_{i=1}^n \mathbb{P}(A_i)$. The infinite collection $\{A_1, \ldots\}$ is (mutually) independent if any finite sub-collections is independent.

- Assume $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are sub-$\sigma$-algebras of $\mathscr{F}$. Then the collection of sub-$\sigma$-algebras $\{\mathscr{F}_1, \mathscr{F}_2, \ldots\}$ is mutually independent if for any $A_1 \in \mathscr{F}_1, A_2 \in \mathscr{F}_2, \ldots$, we have $\mathbb{P}(\cap_{i=1}^n A_n) = \prod_{i=1}^n \mathbb{P}(A_i)$. The infinite collection $\{\mathscr{F}_1, \ldots\}$ is (mutually) independent if any finite sub-collections is independent.

- Let $\xi_1, \xi_2, ...$ be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$. Then the collection $\{\xi_1, \ldots, \xi_n\}$ is (mutually) independent if the sub-$\sigma$-algebras $\sigma(\xi_1), ..., \sigma(\xi_n)$ are independent. In particular if $B_1, \ldots, B_n \in \mathscr{B}(\mathbb{R})$, then

$$\mathbb{P}(\xi_1 \in B_1, \ldots, \xi_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(\xi_i \in B_i) = \mathbb{P}_{\xi_i}(B_i).$$

  The definition can be extended to the infinite case as above.

**Remark 3.8** As seen in elementary probability classes, there is another notion of independence. Let's say there are events $A_1, A_2, ...$, then the collection of events $\{A_1, ..., A_n\}$ is *pairwise* independent if for all $i, j$ with $i \neq j$ we have $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. It is clear that mutual independence implies pairwise independence but not vice-versa. There are similar notions when considering collections of sub-$\sigma$-algebra or random variables. Note that the notion of mutual independence is far more applicable than the notion of pairwise independence in probability theory.

Let us note the following shortcut in establishing mutual independence of sub-$\sigma$-algebra (and hence random variables).

**Lemma 3.9** Let's say $\mathscr{F}_i = \sigma(\mathscr{C}_i)$ for some $\pi$-system containing $\Omega$ (that are closed under finite intersection, see theorem 1.23). If we know that for any $C_i \in \mathscr{C}_i$ (with $\mathscr{F}_i$ belongs to an arbitrary finite sub-collection), $\mathbb{P}(\cap_{i=1}^{n} C_i) = \prod_{i=1}^{n} \mathbb{P}(C_i)$, then $\mathscr{F}_1, \mathscr{F}_2, ...$ are independent.

*Proof.* It suffices to consider the case when there are only two $\sigma$-algebra in our collection, denoted as $\{\mathscr{F}_i := \sigma(\mathscr{C}_i)\}_{i=1,2}$. We break the proof into two steps:

- Fix $A \in \mathscr{C}_1$ and consider the measures $Q_{1,A}(B) := \mathbb{P}(A \cap B)$ and $Q_{2,A}(B) := \mathbb{P}(A)\mathbb{P}(B)$ on $\mathscr{C}_2$. Since these two measures agree on $\mathscr{C}_2$, we know from theorem 1.23 that for all $A \in \mathscr{C}_1$ we have $Q_{1,A} = Q_{2,A}$.
- Now let $B \in \mathscr{F}_2$ and consider the measures $\tilde{Q}_{1,B}(A) := \mathbb{P}(B \cap A)$ and $\tilde{Q}_{2,B}(A) := \mathbb{P}(B)\mathbb{P}(A)$. As seen in the above step, these two measures agree on $\mathscr{C}_2$, so by theorem 1.23 we know that for all $B \in \mathscr{C}_2$ we have $\tilde{Q}_{1,B} = \tilde{Q}_{2,B}$. This completes the proof.

■

As an immediate corollary, we know that

**Corollary 3.10** A necessary and sufficient condition for the random variables $\xi_1, \xi_2, \ldots, \xi_n$ to be independent is that

$$F_\xi(x_1, x_2, \ldots, x_n) = F_{\xi_1}(x_1) \ldots F_{\xi_n}(x_n) \tag{3.15}$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

*Proof.* This comes from the fact that the Borel sets are generated by the collection of left-half-intervals $(-\infty, a]$ and $\mathbb{R}$, with this collection being a $\pi$-system. ■

Combining with Fubini-Tonelli theorem, we have

**Corollary 3.11** If $\xi = (\xi_1, \ldots, \xi_n)$ has a density $f_\xi$, then each $\xi_i$ has a density $f_{\xi_i}$. Furthermore, $\xi_1, \ldots, \xi_n$ are independent if and only if

$$f_\xi(x_1, x_2, \ldots, x_n) = f_{\xi_1}(x_1) \cdots f_{\xi_n}(x_n) \tag{3.16}$$

for all $(x_1, \ldots, x_n)$ except possibly for a Borel subset of $\mathbb{R}^n$ with Lebesgue measure zero.

**Corollary 3.12** If $\xi_1, \ldots, \xi_n$ are independent and $\xi_i$ has density $f_{\xi_i}$, $i = 1, \ldots, n$, then $\xi$ has a density $f_\xi$ given by

$$f_\xi(x_1, x_2, \ldots, x_n) = f_{\xi_1}(x_1) \cdots f_{\xi_n}(x_n).$$

**Remark 3.13** Note that if $\xi_1, \dots, \xi_n$ each have a density, it does not follow that $(\xi_1, \dots, \xi_n)$ has a density.

### 3.2.2 Convolution of Independent Random Variables

Let $\xi, \eta$ be two independent random variables, so $F_{(\xi,\eta)}(x,y) = F_\xi(x)F_\eta(y)$. Consider the random variable $\xi + \eta$. We get

$$\begin{aligned} F_{(\xi,\eta)}(z) &= \int_{\{x,y:x+y\le z\}} \mathrm{d}F_\xi(x)\cdot \mathrm{d}F_\eta(y) \\ &= \int_{\mathbb{R}^2} \chi_{x+y\le z}\,\mathrm{d}F_\xi(x)\cdot \mathrm{d}F_\eta(y) \\ &= \int_{-\infty}^{\infty} \mathrm{d}F_\xi(x)\Big\{\int_{-\infty}^{\infty} \chi_{x+y\le z}\,\mathrm{d}F_\eta(y)\Big\} \\ &= \int_{-\infty}^{\infty} F_\eta(z-x)\,\mathrm{d}F_\xi(x) \end{aligned}$$

and similarly

$$F_{(\xi,\eta)}(z) = \int_{-\infty}^{\infty} F_\xi(z-y)\mathrm{d}F_\eta(y).$$

Thus we obtained the following result

**Proposition 3.14** The distribution function $F_{\xi+\eta}$ of the sum of two independent random variables is the convolution of their distribution functions.

$$F_{(\xi,\eta)}(z) = F_\xi * F_\eta = \int_{-\infty}^{\infty} F_\xi(z-y)\mathrm{d}F_\eta(y) = \int_{-\infty}^{\infty} F_\eta(z-x)\mathrm{d}F_\xi(x). \tag{3.17}$$

Similarly, we can easily obtain

**Corollary 3.15** If $\xi, \eta$ are independent a.c. random variables, then the density $f_{\xi+\eta}$ is the convolution of the densities.

$$f_{\xi+\eta} = f_\xi * f_\eta = \int_{-\infty}^{\infty} f_\xi(z-y)f_\eta(y)\mathrm{d}y = \int_{-\infty}^{\infty} f_\eta(z-x)f_\xi(x)\mathrm{d}x. \tag{3.18}$$

**Example 3.16 — $\xi, \eta$ - independent a.c. random variables.**

- Let $\xi \sim N(m_1, \sigma_1^2)$ and $\eta \sim N(m_2, \sigma_2^2)$, i.e.

  $$f_\xi(x) = \frac{1}{\sigma_1}\varphi\Big(\frac{x-m_1}{\sigma_1}\Big), \quad f_\eta(x) = \frac{1}{\sigma_2}\varphi\Big(\frac{x-m_2}{\sigma_2}\Big),$$

  where

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

  Then

  $$f_{\xi+\eta}(z) = \int_{-\infty}^{\infty} f_\eta(z-x)f_\xi(x)\mathrm{d}x = \frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}}\varphi\Big(\frac{z-(m_1+m_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\Big).$$

  Thus the sum of two independent normal random variables is the normal random variable $N(m_1+m_2, \sigma_1^2+\sigma_2^2)$.

- Let $\xi_1, \xi_2, \dots \xi_n$ be independent $N(0,1)$. Then

  $$f_{\xi_1^2+\dots+\xi_n^2}(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)}x^{(n/2)-1}e^{-x/2}, & x>0, \\ 0, & x\le 0. \end{cases}$$

The random variable $\xi_1^2 + \cdots + \xi_n^2$ is usually denoted by $\chi_n^2$ and its distribution is the $\chi^2$-distribution with $n$ degrees of freedom.

**Proposition 3.17** Let $\xi$ and $\eta$ be independent random variables with $\mathbb{E}[\xi] < \infty$, $\mathbb{E}[\eta] < \infty$. Then $\mathbb{E}[\xi\eta] < \infty$ and $\mathbb{E}[\xi\eta] = \mathbb{E}[\xi]\mathbb{E}[\eta]$.

*Proof.* We utilise the four-step proofs. First assume $\xi, \eta \geq 0$. Consider

$$\xi_n = \sum_{k=0}^{\infty} \frac{k}{n} \chi_{\{\frac{k}{n} \leq \xi(\omega) < \frac{k+1}{n}\}},$$

$$\eta_n = \sum_{k=0}^{\infty} \frac{k}{n} \chi_{\{\frac{k}{n} \leq \eta(\omega) < \frac{k+1}{n}\}}.$$

Then $\xi_n \leq \xi, \eta_n \leq \eta$, $|\xi - \xi_n| \leq \frac{1}{n}, |\eta - \eta_n| \leq \frac{1}{n}$ for all $n$. Since $\xi, \eta$ are integrable, by dominated convergence theorem

$$\lim_{n\to\infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi], \quad \lim_{n\to\infty} \mathbb{E}[\eta_n] = \mathbb{E}[\eta].$$

Now write

$$\begin{aligned}
\mathbb{E}[\xi_n\eta_n] &= \sum_{i,j\geq 0} \frac{jk}{n^2} \mathbb{E}[\chi_{\{\frac{j}{n} \leq \xi < \frac{j+1}{n}\}} \chi_{\{\frac{k}{n} \leq \eta < \frac{k+1}{n}\}}] \quad \text{(Monotone Convergence)} \\
&= \sum_{i,j\geq 0} \frac{jk}{n^2} \mathbb{E}[\chi_{\{\frac{j}{n} \leq \xi < \frac{j+1}{n}\}}] \mathbb{E}[\chi_{\{\frac{k}{n} \leq \eta < \frac{k+1}{n}\}}] \quad \text{(Independence)} \\
&= \mathbb{E}[\xi_n]\mathbb{E}[\eta_n]
\end{aligned}$$

Since

$$|\mathbb{E}[\xi\eta] - \mathbb{E}[\xi_n\eta_n]| \leq \mathbb{E}[|\xi\eta - \xi_n\eta_n|] = \mathbb{E}[|\xi(\eta - \eta_n) + \eta_n(\xi - \xi_n)|] \leq \frac{1}{n}\mathbb{E}[|\xi|] + \frac{1}{n}\mathbb{E}\left[|\eta| + \frac{1}{n}\right] \to 0$$

as $n \to \infty$, we have that

$$\mathbb{E}[\xi\eta] = \lim_{n\to\infty} \mathbb{E}[\xi_n\eta_n] = \lim_{n\to\infty} \mathbb{E}[\xi_n] \lim_{n\to\infty} \mathbb{E}[\eta_n] = \mathbb{E}[\xi]\mathbb{E}[\eta],$$

and $\mathbb{E}[\xi\eta] < \infty$. The result in the general case follows by using the representation

$$\xi = \xi^+ - \xi^-, \eta = \eta^+ - \eta^-.$$

■

In fact, we can prove a stronger converse

**Proposition 3.18** Under the above setting, $\xi, \eta$ are independent iff for all Borel-measurable functions $f, g$ we have $\mathbb{E}[f(\xi)g(\eta)] = \mathbb{E}[f(\xi)]\mathbb{E}[g(\eta)]$.

**Hint.** *For $(\Leftarrow)$ we directly apply the assumption for $f = \chi_{A_1}, g = \chi_{A_2}$, such that $A_1 = \xi_1^{-1}(B_1)$ for an arbitrary $B_1 \in \mathscr{B}(\mathbb{R})$ and similarly for $A_2$. For $(\Rightarrow)$ we utilise the four-step proof similar to above.*

### 3.2.3 Correlation

We also define another notion of "unrelatedness" of two random variables. First we define the notion of **covariance**:

**Definition 3.19 — Covariance.** Let $\xi, \eta$ be a pair of random variables on the same probability space. Their **covariance** is

$$\mathrm{Cov}[\xi, \eta] := \mathbb{E}[(\xi - \mathbb{E}[\xi])(\eta - \mathbb{E}[\eta])]. \tag{3.19}$$

if the expectation above exists.

**Remark 3.20** Note that

$$\mathbb{V}[\xi+\eta] = \mathbb{V}[\xi] + \mathbb{V}[\eta] + 2\mathsf{Cov}[\xi,\eta],$$

so

$$\mathsf{Cov}[\xi,\eta] = 0 \implies \mathbb{V}[\xi+\eta] = \mathbb{V}[\xi] + \mathbb{V}[\eta].$$

**Definition 3.21 — Uncorrelated variables.** The random variables $\xi$ and $\eta$ are called **uncorrelated** if

$$\mathsf{Cov}[\xi,\eta] = 0.$$

**Corollary 3.22** Independent random variables are uncorrelated.

*Proof.* Indeed, using the theorem from above,

$$\mathsf{Cov}[\xi,\eta] = \mathbb{E}[\xi\eta] - \mathbb{E}[\xi]\mathbb{E}[\eta] = 0.$$

■

The converse is not true.

**Example 3.23** Consider for example the random variable $\alpha$ which takes the values $0, \frac{\pi}{2}, \pi$ with probability $\frac{1}{3}$. Then $\xi = \sin\alpha, \eta = \cos\alpha$ are uncorrelated ($\mathbb{E}[\xi] = \frac{1}{3}, \mathbb{E}[\eta] = 0$), but they are not independent since

$$\mathbb{P}(\xi = 1, \eta = 1) = 0 \neq \frac{1}{9} = \mathbb{P}(\xi = 1)\mathbb{P}(\eta = 1).$$

Indeed we can also see that the random variables $\xi, \eta^2$ are correlated, since $\mathbb{E}[\eta^2] = 2/3$ and $\mathbb{E}[\xi]\mathbb{E}[\eta + 1] = 2/9$, but $\mathsf{Cov}[\xi,\eta^2] = -2/9$.

# Part II. Concepts of Convergence

## 4 Coin Flips: Convergence in Probability

We now have sufficient tools from measure theory to get into the first serious topic of probability - limiting theorems. Given a random sequence $(\xi_1, \xi_2, \dots)$ with $\xi_i$ *independently and identically distributed* (i.i.d.), we would like to study the deviation between the empirical mean (or time average) $S_n/n$ (with $S_n = \xi_1 + \cdots + \xi_n$) and the actual mean $\mathbb{E}[\xi_1]$ (space average). In particular, do we know anything when $n \to \infty$?

In this chapter, we consider our simplest example of random events: flipping $n$ independent unfair coins.

### 4.1 Constructing the sample space

So how do we represent the flipping of $n$ independent unfair coins in the mathematical framework we have built in the previous sections? Let us consider the case of flipping just one coin. Assume the outcomes are 0 and 1 with probability of getting 1 being $p \in (0,1)$. We hope to express this as a random variable $\xi$ with value in $\{0,1\}$ on suitable probability space $(\Omega, \mathscr{A}, \mathbb{P})$, such that $\mathbb{P}_\xi(\{0\}) = 1-p$ and $\mathbb{P}_\xi(\{1\}) = p$. A more elaborate way to write the above condition is:

$$\mathbb{P}_\xi(\{x\}) = p^x (1-p)^{1-x}, \quad x = 0, 1 \tag{4.1}$$

There are several choices of sample spaces we can choose:

- The natural choice:

$$\Omega = \{0,1\}, \ \mathscr{A} = 2^\Omega, \ \mathbb{P}(\{\omega\}) = p^\omega (1-p)^{1-\omega}, \ \xi(\omega) = \omega$$

- A more complicated choice:

$$\Omega = [0,1], \ \mathscr{A} = \mathscr{B}([0,1]), \ \mathbb{P}(E) = \mathsf{Leb}(E), \ \xi(\omega) = \chi_{(p,1]}(\omega)$$

  with $\lambda$ being the Lebesgue measure. This represents how a computer simulates a flipping of biased coin: first generate a random number $r \in [0,1]$ from uniform distribution (using e.g. the `numpy.random.rand` function in Python), then return 0 if $r < 1-p$ and 1 otherwise.

In both cases we see that $\mathbb{P}_\xi(\{0\}) = 1-p$ and $\mathbb{P}_\xi(\{1\}) = p$. In fact, we have shown from LOTUS (theorem 2.11) that the distribution functions (hence expectation etc.) will not depend on our choice of probability spaces and random variables, as long as $\mathbb{P}_\xi$ satisfies (4.1).

How can we extend to the experiment of flipping $n$ coins? The wrong way is to assume that $\xi_1, ..., \xi_n$ are on the same sample space $(\Omega, \mathscr{A}, \mathbb{P})$ such that $\xi_1 = ... = \xi_n$, since these random variables really mean flipping a single coin once and recording the result $n$ times. (In particular the random variables are not independent for sure). In fact, it will be hard to write down a large number of independent random variables defined on any of the sample spaces $(\Omega, \mathscr{A}, \mathbb{P})$ in the above example.

A standard way of describing $n$ independent coin flips (or $n$ independent trials in general) is to assume that the random variables $\xi_1, \xi_2, ..., \xi_n$ in different sample spaces (i.e. $\xi_1$ defines on $(\Omega_1, \mathscr{A}_1, \mathbb{P}_1)$, $\xi_2$ defines on $(\Omega_2, \mathscr{A}_2, \mathbb{P}_2)$ and so on...) However, we need a way to understand any operations involving more than one $\xi_i$'s. The way to mitigate is to consider the product space $(\Omega^{(n)}, \mathscr{A}^{(n)}, \mathbb{P}^{(n)}) = \otimes_{i=1}^n (\Omega_i, \mathscr{A}_n, \mathbb{P}_n)$. As a reminder, the sample space of this new probability space is

$$\Omega^{(n)} = \Omega_1 \times ... \times \Omega_n = \{\omega = (\omega_1, ..., \omega_n) \ : \ \forall i, \ \omega_i \in \Omega_i\},$$

the new collection of events are

$$\mathscr{A}^{(n)} = \sigma(\{A_1 \times ... \times A_n \ : \ \forall i, \ A_i \in \mathscr{A}_i\}) = 2^{\Omega^{(n)}}$$

and the new probability measures are $\mathbb{P}_n$ satisfying

$$\mathbb{P}^{(n)}(A_1 \times ... \times A_n) = \prod_{i=1}^{n} \mathbb{P}_i(A_i)$$

Then we can define the family of projection function onto the $i$-th component, $\text{proj}_i^{(n)} : (\Omega^{(n)}, \mathscr{A}^{(n)}) \to (\Omega_i, \mathscr{A}_i)$ such that $\text{proj}_i^{(n)}(\omega_1, ..., \omega_n) = \omega_i$. For convenience, we drop the superscript $(n)$ if there is no ambiguity. Notice that the projection functions are measurable, since the preimage of any sets in $\mathscr{A}_i$ is

$$\text{proj}_i^{-1}(A_i) = \Omega_1 \times ... \times \Omega_{i-1} \times A_i \times \Omega_{i+1}, ..., \Omega_n \in \mathscr{A}^{(n)}$$

Now we can define new random variables $\tilde{\xi} : (\Omega^{(n)}, \mathscr{A}^{(n)}, \mathbb{P}^{(n)}) \to \{0, 1\}$ such that $\tilde{\xi}_i(\omega) = \xi_i(\text{proj}_i(\omega))$. These random variables are an accurate description of flipping $n$ coins since:

1. The marginal distribution of $\xi_i$, defined as the measure $A \mapsto \mathbb{P}_{\tilde{\xi}_i}(\Omega_1 \times ... \times A \times ... \times \Omega_n)$ satisfies (4.1).
2. The family $(\tilde{\xi}_i)$ of random variables are independent.

**Exercise 4.1** Verify the above assertions.

Finally, we want to extend the above construction to $n \to \infty$, i.e. consider the space $(\Omega, \mathscr{A}) = \otimes_{i=1}^{\infty}(\Omega_i, \mathscr{A}_i)$ on a suitable probability measure $\mathbb{P}$, so that we can discuss large-sample theorems. We want the probability measure to satisfies:

$$\mathbb{P}\left(A_1 \times ... \times A_n \times \Omega_{n+1} \times ...\right) = \prod_{i=1}^{n} \mathbb{P}_i(A_i) \tag{4.2}$$

The good news is, such probability measure exists if we use $\Omega_i \equiv \Omega$ and $\mathscr{A}_i \equiv \mathscr{A}$ in our above examples. If we assume the natural choice, then we can safely set

$$\mathbb{P}(\{(\omega_1, \omega_2, ...)\}) = \prod_{i=1}^{\infty} p^{\omega_i}(1-p)^{1-\omega_i} = p^{\sum \omega_i}(1-p)^{\sum(1-\omega_i)}$$

since the probability measure is well-defined for all singletons $\{(\omega_1, \omega_2, ...)\}$. If we use the example when $\Omega_i = [0, 1]$, we can check that our sequence of measures $(\mathbb{P}^{(n)})$ is consistent (see definition 1.32) and apply the Kolmogorov Extension Theorem (theorem 1.33) to define $\mathbb{P}$.

Note the above construction can be generalised to describe a sequence of independent experiments. The above construction means that we need not worried about specifying a single probability space to describe a sequence of independent experiments. As a summary, **if we want to describe infinite sequence of experiment with underlying distribution $\mathbb{P}_\xi$, we consider the infinite product space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))^{\otimes\infty}$ equipped with the probability measure $\mathbb{P}$ as determined by the Kolmogorov extension theorem, then the projections onto the $i$-th component $\text{proj}_i$ are random variables with distribution $\mathbb{P}_\xi$.** From now on, we abuse notation by not mentioning the underlying probability space, dropping the tilde sign above $\xi$ and interpreting any operations (especially addition) in the above sense.

### 4.1.1 Radamecher Functions

There is a third way to construct a **fair** coin flip without using the Kolmogorov extension theorem. Consider the probability space $([0, 1], \mathscr{B}([0, 1]), \mathbb{P})$ with $\mathbb{P} = \lambda$ being the Lebesgue measure. Consider the binary expansions $\omega = 0.\overline{\omega_1\omega_2\ldots}_2$ of numbers $\omega \in \Omega$, and define random variables $\xi_1(\omega), \xi_2(\omega), \ldots$ by putting $\xi_k(\omega) = \omega_k$. We extend these functions so that for all $k$, $\xi_i(1) = 1$. A neater way to express these random variables is:

$$\xi_k(\omega) = H(\sin 2^k \pi t) \tag{4.3}$$

where $H(t) = \chi_{[0,\infty)}$ is the heavyside function. Since, for all $n \geq 1$ and all $x_1, \ldots, x_n$ taking a value 0 or 1,

$$\{\omega : \xi_1(\omega) = x_1, \ldots, \xi_n(\omega) = x_n\} = \left\{\omega : \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n} \leq \omega < \frac{x_1}{2} + \cdots + \frac{x_n}{2^n} + \frac{1}{2^n}\right\},$$

the $\mathbb{P}-$measure of this set is $\frac{1}{2^n} = \prod_{i=1}^n \mathbb{P}(\xi_n(\omega) = x_n)$. It follows that $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables with

$$\mathbb{P}(\xi_1 = 0) = \mathbb{P}(\xi_1 = 1) = \frac{1}{2}.$$

As we will see in later chapters, such construction give rises to some important number theoretic results regarding binary expansion of numbers in $[0, 1]$.

## 4.2 Weak Law of Large Numbers

### 4.2.1 A high probability statement for coin flips

Let $S_n = \xi_1 + \ldots \xi_n$. Then

$$\mathbb{E}[S_n] = \sum_{j=1}^n \mathbb{E}[\xi_n] = \sum_{j=1}^n (1 \cdot \mathbb{P}_{\xi_j}(\xi_j = 1) + 0 \cdot \mathbb{P}_{\xi_j}(\xi_j = 0)) = np.$$

Thus the mean value of $S_n/n$ is equal to $p$. Our central question is: what does $|\frac{1}{n}S_n(\omega) - p|$ converges to for large $n$? Moreover, in what sense we can consider the convergence? It cannot be that

$$\left| \frac{S_n(\omega)}{n} - p \right| \to 0$$

uniformly/pointwise in $\omega$, because there is always an $\omega$ such that $\xi_i(\omega) = 1$ for all $i$, so $S_n(\omega)/n \equiv 1 \nrightarrow p$, so we must consider a weaker notion of convergence. Before delving into our discussion, let us recall the following observation

**Exercise 4.2** Verify that $S_n \sim \mathsf{B}(n, p)$. Hence show that $\|S_n/n - p\|_{L^2}^2 = \mathbb{V}[S_n/n] = p(1-p)/n \overset{n\to\infty}{\to} 0$.

This shows that, in fact, we have $L^2$ convergence. With more careful analysis we show that we actually have $L^p$ convergence for any $p \in [1, \infty)$. By Chebyshev inequality, we also shows that for all $\epsilon > 0$ (fixed)

$$\mathbb{P}\left( \left| \frac{S_n}{n} - p \right| > \epsilon \right) \leq \frac{\mathbb{V}[S_n/n]}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \tag{4.4}$$

So for all $\epsilon > 0$ fixed, we can always make the probability $\mathbb{P}\left(|S_n/n - p| > \epsilon\right)$ arbitrary small by choosing sufficiently large $n$. This is known as *convergence of probability*.

**Remark 4.3** *We can make (4.4) sharper by considering Chernoff bound. Indeed, we know that the random variable $S_n$ has moment generating function

$$M_{S_n}(t) = (1 - p + pe^t)^n, \quad \forall t \in \mathbb{R} \tag{4.5}$$

so from Chernoff bound, we know for all $t > 0$ we have

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{n} - p > \epsilon\right) &= \mathbb{P}\left(\exp(t(S_n - np)) > \exp(tn\epsilon)\right) \\ &\leq \frac{\exp(-tnp)\mathbb{E}[\exp(tS_n)]}{\exp(tn\epsilon)} \\ &= \exp\left(n\left(\ln(1 - p + pe^t) - t(p + \epsilon)\right)\right) =: \exp(n(\phi(t) - t\epsilon))\end{aligned}$$

where $\phi(t) := \ln(1 - p + pe^t) - pt$. We claim is that $\phi(t)$ is bounded by a quadratic function. We can prove this by bounding the second derivative of $\phi(t)$. Note that

$$\phi'(t) = -p + \frac{pe^t}{1 - p + pe^t} = 1 - p - \frac{1 - p}{1 - p + pe^t} \tag{4.6}$$

$$\phi''(t) = \frac{(1 - p)pe^t}{(1 - p + pe^t)^2} = \frac{1 - p}{1 - p + pe^t}\left(1 - \frac{1 - p}{1 - p + pe^t}\right) \tag{4.7}$$

Let $u(t) = \dfrac{1 - p}{1 - p + pe^t}$. Then clearly $\forall t$, $0 < u(t) < 1$. Moreover, we see that $\phi''(t) = u(1 - u) = -(u - 1/2)^2 + 1/4$. Therefore we know for all $t$, $\phi''(t) \in (0, 1/4]$. Hence by Taylor Theorem, there exists a $c$ between $0$ and $t$ such that

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(c) \leq \frac{t^2}{8} \tag{4.8}$$

Substitute into our Chernoff bound yields

$$\mathbb{P}\left(\frac{S_n}{n} - p > \epsilon\right) \leq \exp(n(t^2/8 - \epsilon t)) = \exp(n((t - 4\epsilon)^2/8 - 2\epsilon^2)) \tag{4.9}$$

So by choosing $t = 4\epsilon$ (such that the RHS) is minimised we yield a much sharper bound:

$$\mathbb{P}\left(\frac{S_n}{n} - p > \epsilon\right) \leq \exp(-2n\epsilon^2) \tag{4.10}$$

By symmetry we also obtain

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2) \tag{4.11}$$

We obtain a much sharper bound than (4.4) since we now have exponential decay rather than a quadratic decay. The trick of isolate the term $-nt\epsilon$ and bounding $\phi(t)$ by a quadratic function is common, since this can be generalised with any random variables $S_n \in L^\infty(\Omega)$. Such a bound is called a *Hoeffding bound*, and we will discuss that further in latter chapter.

### 4.2.2 $L^2$ Weak Law of Large Numbers

Let us study how the weak law of large numbers can be generalised. We first formally define the definition of convergence in probability. The definition holds for not only real-valued random variables, but also random variables taking values in a Polish space $(X, \mathscr{B}(X))$ (equipped with the Borel $\sigma$-algebra) associated with a metric $d$. Recall that a metric space is Polish if it is complete and separable (see remark 1.17).

**Definition 4.4** A sequence $\xi_1, \xi_2, \ldots$ of random variables from probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to Polish space $(X, \mathscr{B}(X))$ with metric $d$ converges *in probability*, or *in measure* $\mathbb{P}$, to the random variable $\xi$

(denoted by $\xi_n \xrightarrow{p} \xi$) if for every $\varepsilon > 0$

$$\mathbb{P}(d(\xi_n, \xi) > \varepsilon) \to 0, \quad n \to \infty.$$

**Exercise 4.5 — Properties of convergence in probability.** Let $(\xi_i)_{i\geq 1}$, $(\eta_i)_{i\geq 1}$ are sequences of random variables on $(\Omega, \mathscr{F}, \mathbb{P})$, and also let $\xi$, $\eta$ be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$

1. Check that the limit of convergence in probability is almost surely unique: if $\xi_i$ converges in probability to $\xi$ and $\xi'$ then $\xi = \xi'$ almost surely.
2. Prove that if $\xi_i \xrightarrow{p} \xi$ and $\eta_i \xrightarrow{p} \eta$ then for all real numbers $a, b$ we have $a\xi_i + b\eta_i \xrightarrow{p} a\xi + b\eta$.
3. Prove that if $\xi_i \xrightarrow{p} \xi$ and $\eta_i \xrightarrow{p} \eta$ then for all real numbers $a, b$ we have $\xi_i\eta_i \xrightarrow{p} \xi\eta$. Notice this is not necessary true if we have $L^p$ convergence. What's wrong with the argument, and can we refine the statements?
4. (Slutsky's lemma) Show that if $\xi_i \xrightarrow{p} \xi$ and $\eta_i \xrightarrow{p} \eta$ and if $\varphi(x, y)$ is a continuous function, then

$$\varphi(\xi_i, \eta_i) \xrightarrow{p} \varphi(\xi, \eta).$$

Note that by Markov inequality that if $\xi_n \to \xi$ in $L^p$ ($p \geq 1$, see definition 2.35), then we must have $\xi_n \xrightarrow{p} \xi$, because

$$\mathbb{P}(|\xi_n - \xi| > \varepsilon) \leq \frac{\|\xi_n - \xi\|_{L^p}^p}{\varepsilon^p} \to 0 \tag{4.12}$$

Let $\xi_1, \ldots, \xi_n$ be a random variables. Denote

$$S_n^{(c)} = \sum_{j=1}^{n} (\xi_j - \mathbb{E}[\xi_j]).$$

Note $\mathbb{E}[S_n^{(c)}] = 0$. How can we fully use Chebyshev inequality to make the assumptions on $\xi_1, ..., \xi_n$ as weak as possible? For simplicity, we first assume $\xi_i \in L^2$ such that $\mathbb{V}[\xi_i] \leq C$ for some constant $C$ independent of $i$. Then, if we assume that $\xi_1, ..., \xi_n$ are pairwise uncorrelated (which is a much weaker assumption than independence), then we have

$$\mathbb{V}[S_n] = \sum_{i=1}^{n} \mathbb{V}[\xi_i] \leq Cn \tag{4.13}$$

and hence we have the following

**Theorem 4.6 — $L^2$ Weak Law of Large Numbers.** Let $\xi_1, \ldots, \xi_n$ be uncorrelated $L^2$ random variables such that $\mathbb{V}[\xi_j] \leq C$ for some $C > 0$ and all $n \geq 1$. Then $S_n^{(c)}/n \xrightarrow{p} 0$

*Proof.* By Chebyshev's inequality and since $\xi_j$ are uncorrelated, for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n^{(c)}}{n}\right| > \varepsilon\right) \leq \frac{\mathbb{V}[S_n^{(c)}/n]}{\varepsilon^2} \leq \frac{C}{n\varepsilon^2} \xrightarrow{n\to\infty} 0 \tag{4.14}$$

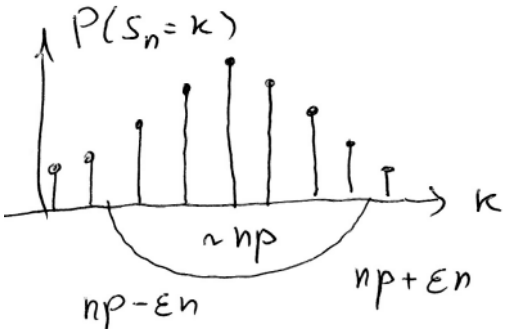
■

### 4.2.3 Weak Law of Large Number for uniformly integrable sequences*

Can we try to weaken the assumption that $\xi \in L^2$, to say, $\xi \in L^1$?

## 4.3 Local and Central Limit Theorem

We now return to our coin flipping scenario. Recall we have $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$, where $\xi_j$ are iid as $\mathsf{B}(1,p)$ as constructed in section 4.1. As discussed in previous section, $S_n$ tends to be close to $np$ for large $n$.




Specifically, let us take some interval $\mathscr{I}_n = (n(p-\varepsilon), n(p+\varepsilon))$. If we pick some sufficiently large $\varepsilon$, say $\varepsilon = n^\alpha$ where $\alpha > 1/2$, then by Chebyshev inequality (4.4) (or Hoeffding bound (4.11)) we know that

$$\mathbb{P}(S_n \in \mathscr{I}_n^c) = \mathbb{P}(|S_n/n - p| > n^{\alpha-1}) \leq \frac{p(1-p)}{n^{1+2\alpha-2}} = \frac{p(1-p)}{n^{2\alpha-1}} \overset{n\to\infty}{\to} 0 \tag{4.15}$$

and the decaying bounds of $\mathbb{P}(S_n \in \mathscr{I}_n^c)$ no longer exist when $\alpha \leq 1/2$. How much do we know about $\mathbb{P}(S_n \in \mathscr{I}_n^c)$ for this case? Will it tends to some non-trivial constant in $(0,1)$, or will it increase and tend to 1? We will see that the central limit theorem shows that at the boundary case $\alpha = 1/2$, $\mathbb{P}(S_n \in \mathscr{I}_n^c)$ tends to some non-trivial constant in $(0,1)$ depending on the constant $C$ when defining $\varepsilon = C/n^{1/2}$. This also suggests that the rescaled mean $S_n/n^\alpha$ at the threshold $\alpha = 1/2$ will "converge" in some way to a non-trivial distribution.

### 4.3.1 A crash course in asymptotic analysis

Before we delve into the main discussions, we define two important order notations for sequences $f_n$:

**Definition 4.7 — Order notations.** Consider two sequences $f_n, g_n$. As $n \to \infty$, we say that

- (Big $O$) $g_n = O(f_n)$ if $|g_n/f_n|$ is bounded for sufficiently large $n$, i.e. there exists constant $C > 0$ and $N$ such that for all $n \geq N$, $|g_n| \leq C|f_n|$
- (Small $o$) $g_n = o(f_n)$ if $|g_n/f_n| \to 0$ as $n \to \infty$. In other words, for all constants $\epsilon > 0$, there exists $N := N(\epsilon)$ such that for all $n \geq N$, $|g_n| \leq \epsilon|f_n|$. We also sometimes write $g_n \ll f_n$ or $f_n \gg g_n$.
- (Asymptotic equivalence) $g_n \sim f_n$ if $|g_n/f_n| \to 1$ as $n \to \infty$. Equivalently, we have $g_n = (1 + o(1))f_n$.
- (Order) $g_n = \Theta(f_n) = \mathsf{ord}(f_n)$ if $g_n = O(f_n)$ but $g_n$ is not $o(f_n)$.

**Remark 4.8**

- Notice the use of equal sign is an abuse of notation.
- The definitions can be extended to any functions $f(x)$ defined on real or complex numbers, in such case we can assume $x$ tends to some points $x_0$ including $\infty$.
- We can also consider order notations for sequences of functions. Let $g_n = g_n(\alpha)$, $f_n = f_n(\alpha)$. We say $g_n(\alpha) = O(|f_n(\alpha)|)$ **uniformly** if above definition holds with constants $C, N$ independent of $\alpha$. We also have analogous definition for $g_n(\alpha) = o(|f_n(\alpha)|)$.

With this, we can prove one of the most important result in asymptotic analysis

**Lemma 4.9 — Stirling's Approximation.** As $n \to \infty$, we have

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + O(1/n)) \tag{4.16}$$

*Proof.* (Sketch)* There are many ways to prove this famous results, perhaps the quickest way is to notice that $n! = \Gamma(n+1)$, where $\Gamma(z)$ is the Gamma function satisfying

$$\Gamma(z+1) = \int_0^\infty t^z e^{-t}\, dt = \int_0^\infty \exp(z \ln t - t)\, dt \tag{4.17}$$

Apply a change of variable $s = t/z$, we have

$$\Gamma(z+1) = \int_0^\infty t^z e^{-t}\, dt = z^{z+1} \int_0^\infty \exp(z(\ln s - s))\, ds \tag{4.18}$$

Notice the function $\phi(s)$ has the following Taylor series at $s = 1$:

$$\ln s = \ln(1 + (s-1)) = (s-1) - \frac{(s-1)^2}{2} + O((s-1)^3) \tag{4.19}$$

and therefore

$$\Gamma(z+1) = z\left(\frac{z}{e}\right)^z \int_0^\infty \exp\left(-z\left(\frac{(s-1)^2}{2} + O(s-1)^3\right)\right) dz \tag{4.20}$$

With careful analysis we see that the integral is equivalent to

$$\int_{-\infty}^\infty \exp\left(-z\frac{(s-1)^2}{2}\right) ds = \sqrt{\frac{2\pi}{z}} \tag{4.21}$$

The main difference is that we now consider range of integration to be $(-\infty, \infty)$ instead of $[0, \infty)$ and ignore the higher order term, in both cases would introduce error which is exponentially small (so can be ignored). Hence we prove that

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{4.22}$$

Refining the analysis to incorporate the $O(1/n)$ correction is hard as it involves more advanced techniques in asymptotic analysis, so will not be included here. ■

With this we can briefly sketch the asymptotic analysis of binomial coefficient. Let's say $n, k, n-k$ all tends to infinity (e.g. $k = np$ for $k \in (0,1)$), then we can use Stirling's formula to obtain

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \frac{(n/e)^n}{(k/e)^k((n-k)/e)^{n-k}} \frac{1 + O(1/n)}{(1+O(1/n))1 + O(1/n)} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \exp\left(n\ln n - k\ln k - (n-k)\ln(n-k)\right) \frac{1 + O(1/n)}{(1+O(1/n))(1+O(1/n))}
\end{aligned}$$

We notice that the purple term only gives a correction of $1 + O(1/n)$. Instead of going through careful analysis, we can build intuition by treating the $O(1/n)$ correction terms as being exactly equal to $1/n$. Then the denominator satisfies $(1+1/n)^{-2} = 1 - 2/n + ... = 1 + O(1/n)$. Then conclude that $(1 + 1/n)(1 - 2/n) = 1 - 1/n + ... = 1 + O(1/n)$. To sum up, we have

$$\binom{n}{k} = \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \exp\left(n\ln n - k\ln k - (n-k)\ln(n-k)\right)\left(1 + O\left(\frac{1}{n}\right)\right) \tag{4.23}$$

### 4.3.2 Proving the Central Limit Theorem

We recall the following result regarding the probability mass function of a binomial distribution $\mathsf{B}(n,p)$:

**Exercise 4.10 — Monotonicity of Binomial probability.** Show that $\mathbb{P}(S_n = k)$ is monotone in $k$ below and above its point of maximum.

With this we can prove the local limit theorem, which specifies the local asymptotics of probability mass distribution at the point $S_n = k$.

**Theorem 4.11 — Local Limit Theorem.** For any $0 < p < 1$,

$$\max_{0\le k\le n}\left|\mathbb{P}(S_n=k)-\frac{1}{\sqrt{2\pi p(1-p)}\sqrt{n}}e^{-\frac{x^2}{2p(1-p)}}\right| = o\left(\frac{1}{\sqrt{n}}\right) \quad n\to\infty, \tag{4.24}$$

where $x = x_{k,n} := \dfrac{k-np}{\sqrt{n}}$

*Proof.* The main subtlety is that we cannot always apply Stirling's formula. We have to first consider $k$ that are "sufficiently close" to $np$. Specifically, we consider $k$ such that

$$|x_{k,n}| \le \frac{A_n}{\sqrt{n}}, \quad A_n = n^\epsilon \text{ with } \epsilon \in (0,1) \tag{4.25}$$

Then we have $k = np + x\sqrt{n}$, $k$

$$\begin{aligned} k &= np + x\sqrt{n} = np(1+O(A_n/n)) \\ n-k &= n(1-p) - x\sqrt{n} = n(1-p)(1+O(A_n/n)) \end{aligned}$$

These inequalities ensure that both $k$ and $n-k$ tends to infinity as $n\to\infty$, and we can safely use Stirling's approximation to show that

$$\mathbb{P}(S_n=k) = \underbrace{\frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}}}_{\text{(A)}}\underbrace{\exp\left(n\ln n - k(\ln k - \ln p) - (n-k)(\ln(n-k)-\ln(1-p))\right)}_{\text{(B)}}\left(1+O\left(\frac{1}{n}\right)\right)$$

We first analyse (A): notice that

$$\begin{aligned} \text{(A)} &= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \\ &= \frac{1}{\sqrt{2\pi np(1-p)(1+O(A_n/n))(1+O(A_n/n))}} \\ &= \frac{1}{\sqrt{2\pi np(1-p)}}(1+O(A_n/n)) \end{aligned}$$

The the correction factor can be obtained using similar arguments above. We can then analyse (B) by noticing that

$$\begin{aligned} \text{(B)} &= \exp\left(n\ln n - k\left(\ln n + \ln\left(1+\frac{x}{p\sqrt{n}}\right)\right) - (n-k)\left(\ln n + \ln\left(1-\frac{x}{(1-p)\sqrt{n}}\right)\right)\right) \\ &= \exp\left(-\left[(np+x\sqrt{n})\ln\left(1+\frac{x}{p\sqrt{n}}\right) + (n(1-p)-x\sqrt{n})\ln\left(1-\frac{x}{(1-p)\sqrt{n}}\right)\right]\right) \\ &= \exp\Bigg(-\Bigg[np\left(\frac{x}{p\sqrt{n}} - \frac{x^2}{2p^2n} + O\left(\frac{x^3}{n^{3/2}}\right)\right) + \frac{x^2}{p} + O\left(\frac{x^3}{n^{1/2}}\right)\Big) \\ &\quad + n(1-p)\left(-\frac{x}{(1-p)\sqrt{n}} - \frac{x^2}{2(1-p)^2n} + O\left(\frac{x^3}{n^{3/2}}\right)\right) + \frac{x^2}{(1-p)} + O\left(\frac{x^3}{n^{1/2}}\right)\Bigg]\Bigg) \\ &= \exp\left(-\frac{x^2}{2p(1-p)} + O\left(\frac{x^3}{\sqrt{n}}\right)\right) \\ &= \exp\left(-\frac{x^2}{2p(1-p)} + O\left(\frac{A_n^3}{n^2}\right)\right) = \exp\left(-\frac{x^2}{2p(1-p)}\right)\left(1+O\left(\frac{A_n^3}{n^2}\right)\right) \end{aligned}$$

Combining, we have

$$\mathbb{P}(S_n = k) = \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{x^2}{2p(1-p)}\right)\left(1 + O\left(\frac{A_n}{n}\right) + O\left(\frac{A_n^3}{n^2}\right)\right) \tag{4.26}$$

We want to select $\varepsilon < 2/3$ for $A_n^3/n^2 \ll 1$. We select $\varepsilon = 7/12$, then $A_n^3/n^2 = n^{-1/4}$ and $A_n/n = n^{-5/12}$, combining yield

$$\max_{|x|\le A_n/\sqrt{n}} \mathbb{P}(S_n = k) = \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{x^2}{2p(1-p)}\right)\left(1 + \underbrace{O\left(\frac{1}{n^{5/12}}\right)}_{=o(1/\sqrt{n})}\right) \tag{4.27}$$

We are not done yet, since we still have to consider the case when $x_{n,k}$ satisfies $|x| > \frac{A_n}{\sqrt{n}}$. Fortunately both $\mathbb{P}(S_n = k)$ and the Gaussian tails are very small. Specifically,

$$\begin{aligned}
&\max_{|x|>A_n/\sqrt{n}} \left|\mathbb{P}(S_n = k) - \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{x^2}{2p(1-p)}\right)\right| \\
&\le \max_{|x|>A_n/\sqrt{n}} |\mathbb{P}(S_n = k)| + \max_{|x|>A_n/\sqrt{n}} \left|\frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{x^2}{2p(1-p)}\right)\right| \\
&\le \max\left(\mathbb{P}(S_n = \lfloor np + A_n \rfloor), \mathbb{P}(S_n = \lceil np - A_n \rceil)\right) + \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{A_n^2}{2np(1-p)}\right)
\end{aligned}$$

The bound of first term is a direct application of exercise 4.10. Keeping the choice $A_n = n^{7/12}$, we see immediately that the second term is of $o(1/\sqrt{n})$. Now note that

$$n^{1/12} - n^{-1/2} = \frac{np + A_n - np - 1}{\sqrt{n}} \le \frac{\lfloor np + A_n \rfloor - np}{\sqrt{n}} \le \frac{np + A_n - np}{\sqrt{n}} = n^{1/12} \tag{4.28}$$

so $x_{\lfloor np+A_n \rfloor, k} \sim n^{1/12}$, and this holds similarly for $x_{\lceil np+A_n \rceil, k}$, so the first term is also $o(1/\sqrt{n})$. Combining these results with (4.27) completes the proof. ∎

The local limit theorem really tells us that

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{n}} = x\right) = \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{x^2}{2p(1-p)}} + o(1)\right] \quad n \to \infty \tag{4.29}$$

At first glance you may find this result not useful, as it only tells us that the probability decays to zero at a rate of $O(1/\sqrt{n})$. However, since $\dfrac{S_n - np}{np(1-p)}$ seems to converge to a *continuous* distribution, we really should look at the **density** by ignoring the $1/\sqrt{n}$. The things inside the square bracket suggests that the density function of $n^{-1/2}(S_n - np)$ "converges" to a normal distribution $\mathbf{N}(0, p(1-p))$, which is equivalent to the distribution of $\dfrac{S_n - np}{np(1-p)}$ converging to standard normal $\mathbf{N}(0,1)$. The above heuristics can be formalised by adding the local probabilities and consider the cumulative distribution function. We therefore arrive the central limit theorem (CLT).

**Theorem 4.12 — de Moivre-Laplace CLT.** For any $0 < p < 1$, $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) = \Phi(x),$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \mathrm{d}y,$$

is the distribution of $\mathbf{N}(0,1)$.

*Proof.* (Sketch) We note that

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) = \mathbb{P}\left(S_n \leq np + x\sqrt{np(1-p)}\right)$$
$$= \sum_{k=0}^{\lfloor np - n^{7/12} \rfloor - 1} \mathbb{P}(S_n = k) + \sum_{k=\lfloor np - n^{7/12} \rfloor}^{\left\lfloor np + x\sqrt{np(1-p)} \right\rfloor} \mathbb{P}(S_n = k)$$

Now note that the first term is a sum of polynomial number of terms in exponentially small order $O(\exp(-n^{1/2}))$, so will vanish as $n \to \infty$. Notice the second term is a Riemann sum: writing

$$T_n = \left\{k \mid \left\lfloor np - n^{7/12} \right\rfloor \leq k \leq \left\lfloor np + x\sqrt{np(1-p)} \right\rfloor\right\}$$

we have

$$\sum_{k \in T_n} \mathbb{P}(S_n = k) = \sum_{k \in T_n} \frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi p(1-p)}} \exp\left(-\frac{1}{2}\left(\frac{k - np}{\sqrt{np(1-p)}}\right)^2\right)$$
$$= \sum_{k \in T_n - np} \frac{1}{\sqrt{n}} \frac{1}{\sqrt{2\pi p(1-p)}} \exp\left(-\frac{1}{2}\left(\frac{k/\sqrt{n}}{\sqrt{p(1-p)}}\right)^2\right)$$

Notice this is almost a Riemann sum on a partition of $(-\infty, x]$ with mesh of partition $1/\sqrt{n}$, missing some boundary terms. One can then show that the boundary terms leads to an $o(1)$ contribution, and conclude that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \mathrm{d}y \tag{4.30}$$

We will omit the details here. ■

**Remark 4.13** It also holds, for all $a < b$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y \tag{4.31}$$

With careful analysis, we can let $a = -\epsilon\sqrt{\frac{n}{p(1-p)}}$ and $b = \epsilon\sqrt{\frac{n}{p(1-p)}}$ to conclude that

$$0 \leq \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) = \left(\int_{-\infty}^{-\epsilon\sqrt{\frac{n}{p(1-p)}}} + \int_{\epsilon\sqrt{\frac{n}{p(1-p)}}}^{\infty}\right) \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y + o(1) \tag{4.32}$$

By Mill's ratio inequality (2.51), we can further control the integral by

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) \leq 2\sqrt{\frac{p(1-p)}{2\pi n\epsilon^2}} \exp\left(-\frac{1}{2}\left(\frac{n\epsilon^2}{p(1-p)}\right)\right) + o(1) \tag{4.33}$$

So the tail probability tends to zero, leading to the WLLN for coin flipping.

The de Moivre-Laplace CLT demonstrates that the sequence of random variables $\sqrt{n}((S_n/n) - p)$ converges *in distribution* (converges weakly) to a random variable with normal distribution $\mathsf{N}(0, p(1-p))$. We will formally define the notion of weak convergence in chapter 5-6, and prove a generalised version of central limit theorem.

## 4.4 Poisson Convergence

For completion, let us prove another result concerning convergence in distribution.

**Theorem 4.14 — Poisson distribution.** Fix $k$ and let $p := p(n) \to 0$ as $n \to \infty$ s.t. $p(n) \cdot n \to \lambda > 0$. Then

$$\mathbb{P}(S_n = k) = \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right)^k \left(1 - \frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right)^{n-k} \longrightarrow \frac{1}{k!}\lambda^k e^{-\lambda}. \quad (4.34)$$

## 4.5 Interlude: An overview to upcoming chapters

Let us give an overview to the upcoming chapters. In previous sections we have discussed $L^p$ convergence and convergence in probability in detail using the coin flip example. In chapter 5-6 we will discuss weak convergence, and in chapter 7 we will discuss almost sure convergence. The concept of almost sure convergence should have been covered in elementary courses in measure theory. If you have not seen it before, here is the formal definition:

**Definition 4.15 — Almost sure convergence.** A sequence $(\xi_n)_{n\geq 1}$ of random variables on probability space $(\Omega, \mathscr{F}, \mathbb{P})$ converges $\mathbb{P}$-almost surely to the random variable $\xi$ (denoted by $\xi_n \overset{a.e.}{\to} \xi$) if

$$\mathbb{P}\left(\left\{\omega \,|\, \xi_n(\omega) \overset{n\to\infty}{\not\to} \xi(\omega)\right\}\right) \quad (4.35)$$

Further discussions will follow, but having this definition is probably enough for now. It will be useful to note the following implications of convergence, which is summarised by the following figure. The following figure illustrates the implications.

$$\begin{array}{ccccc} & & \overset{L^p}{\to} & & \\ & & \Downarrow & & \\ \overset{a.e.}{\to} & \Longrightarrow & \overset{p}{\to} & \Longrightarrow & \overset{d}{\to} \end{array}$$

The double arrows refer to implications. This means, in particular, that almost sure convergence and convergence in $L^p$ both imply convergence in probability, and convergence in probability implies convergence in distribution. Notice that all these implications cannot be reversed.

One should also note that most of the above notions of convergence are related to some metrics. For instance, it is apparent that $L^p$ convergence involves the metric induced by the $L^p$ norm, and we will see in the next chapter that convergence in distribution involves the so-called Levy-Prokhorov metric. The following exercise shows that convergence in probability induces a metric in the space of all random variables.

**Exercise 4.16 — Metric for convergence in probability.** Let $\mathscr{M}$ be the set of the real-valued random variables on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Recall the equivalence relation as defined on (2.44), that $\xi \sim \eta$ iff $\xi = \eta$ $\mathbb{P}$-almost everywhere. Define function $d : \mathscr{M}/\sim \times \mathscr{M}/\sim \to \mathbb{R}$ such that

$$d([\xi]_\sim, [\eta]_\sim) = \mathbb{E}(|\xi - \eta| \wedge 1) \quad (4.36)$$

1. Show that $d$ is a *well-defined* metric on $\mathscr{M}/\sim$.
2. Let $(\xi_n)_{n\geq 1}$ be a sequence in $\mathscr{M}$ and let $\xi$ be an element of $\mathscr{M}$. Show that $([\xi_n]_\sim)_{n\geq 1}$ converges to $[\xi]_\sim$ with respect to this metric iff $(\xi_n)_{n\geq 1}$ converges to $\xi$ in probability.

Note that the notion of metric can be generalised when $\mathscr{M}$ is the set of random variables taking value on a generic Polish space with metric $d$.

# 5 Weak Convergence

In this section, we focus on the notion of weak convergence of measures. Discussions for measures of $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are enough for our discussions of the Central Limit Theorem for real-valued random variables, but we will go further and discuss random variables with values taken from a Polish space $(X, \mathscr{B}(X))$. A good reference would be chapter 13 of [1] or chapter 2 of [2].

## 5.1 Definition of weak convergence

We now define the notion of weak convergence of measures and convergence in distribution.

**Definition 5.1 — Weak convergence of measures and random variables.**

- Let $\mu_n, \mu$ be measures on the above Polish space $(X, \mathscr{B}(X))$. We say that $\mu_n$ converges to $\mu$ weakly as $n \to \infty$ if for all $f \in C_b(X)$, we have
$$\int_X f(x)\, \mu_n(dx) \overset{n\to\infty}{\to} \int_X f(x)\, \mu(dx),$$
where $C_b(X)$ represents the set of all bounded continuous functions on $X$.
- Let $\xi_n : (\Omega_n, \mathscr{F}_n, \mathbb{P}_n) \to (X, \mathscr{B}(X))$ be random variables for $n = 1, 2, ..., \infty$. Then $\xi_n \to \xi_\infty$ in distribution as $n \to \infty$ if the push forward measures $\xi_n^* \mathbb{P}_n$ converges to weakly to $\xi_\infty^* \mathbb{P}_\infty$. If $\mathbb{E}_\mathbb{P}$ denotes the expectation with respect to $\mathbb{P}$, then by the change of variable formula (theorem 2.11), the above definition is equivalent to saying that for all $f \in C_b(X)$,
$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}[f(\xi_n)] = \mathbb{E}_\mathbb{P}[f(\xi)]$$

It is important to note that we do not need to specify the probability space of $\xi_n$ when establishing convergence in distribution - what matters is the distribution of $\xi_n$. The definition comes with the following fact of measure theory, that a finite measure $\mu$ on a Polish space $(X, \mathscr{B}(X))$ with metric $d$ is *regular* in the following sense: for all $A \in \mathscr{B}$ we have

$$\mu(A) = \sup_{E \subseteq A, E \text{ compact}} \mu(E) = \inf_{A \subseteq O, O \text{ open}} \mu(O). \tag{5.1}$$

For the special case on $([0,1], \mathscr{B}([0,1]), \text{Leb})$ this could be seen as we approximate the Lebesgue measure of set $A \in \mathscr{B}([0,1])$ by considering coverings of open (or half-open) intervals. For general cases please see theorem 13.6 of [1]. As a result, a measure could be determined by looking at integrals of a certain type of function. Formally,

**Lemma 5.2** If $\mu, \nu$ are probability measures defined on a Polish space $(X, \mathscr{B}(X))$ with metric $d$ such that for all $f \in C_b(X)$, we have $\int_X f\, \mu(dx) = \int_X f\, \nu(dx)$, then $\mu = \nu$.

**Remark 5.3** The condition can be further weakened. Recall that a function $f : X \to \mathbb{R}$ is $K$-Lipschitz (denoted as $f \in \text{Lip}_K(X)$) if for all $x, y \in X$,
$$|f(x) - f(y)| \leq K d(x, y) \tag{5.2}$$
We denote the set of all Lipschitz functions to be $\text{Lip}(X) = \cup_{K>0} \text{Lip}_K(X)$. Then if $\int_X f\, \mu(dx) = \int_X f\, \nu(dx)$ for all **bounded** $f \in \text{Lip}(X)$, then $\mu = \nu$. This is enough for our later purpose.

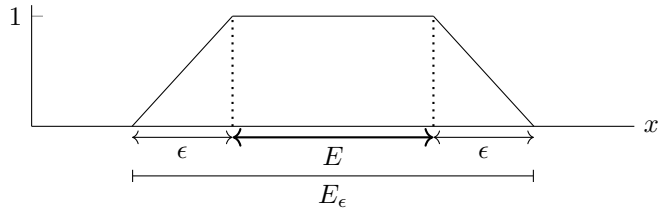
The above statement says that the family of functions $C_b(X)$ or the set of all bounded functions in $\text{Lip}(X)$ is **separating**.

*Proof.* As pointed out in example 1.24 it is enough to show that $\mu(E) = \nu(E)$ for all $E$ closed. For each closed set $E$, we may define the distance between a point $x \in X$ and $E$ by taking infinmum:
$$\rho(x, A) = \inf \{|x - y| : y \in E\} \tag{5.3}$$

As a result, for any $\epsilon > 0$, any indicator function of closed sets $\chi_E(x)$ can be approximated (from above) by the following bounded $1/\epsilon$-Lipschitz function:

$$g_\epsilon(x) = \max\left(0, 1 - \frac{\rho(x,E)}{\epsilon}\right) = \left(1 - \frac{\rho(x,E)}{\epsilon}\right) \vee 0 \tag{5.4}$$

If $E \subseteq \mathbb{R}$ is a closed interval, then the function $g(x)$ can be visualised as below:




Let $E_\epsilon := \{x \,|\, \rho(x,E) < \epsilon\}$. Then clearly

$$\chi_E \le g_\epsilon \le \chi_{E_\epsilon} \implies \mu(E) \le \int_X g_\epsilon(x)\,\mu(dx) = \int_X g_\epsilon(x)\,\nu(dx) \le \nu(E_\epsilon).$$

But as $\epsilon \searrow 0$ we have $E_\epsilon \searrow E$ and hence by continuity from above we have $\nu(E_\epsilon) \searrow \nu(E)$, so $\mu(E) \le \nu(E)$. Interchanging $\mu$ and $\nu$ completes the proof. ■

As a corollary,

**Corollary 5.4 — Uniqueness of weak limit.** If $\mu_n \to \mu$ and $\mu_n \to \nu$ weakly then $\mu = \nu$.

We may relate the weak convergence of a sequence of measures to functional analysis. This discussion is optional if you have not previously studied functional analysis before, and a good reference would be chapter 3 of [3] or section 4.9 of [4]. Recall that $C_b(X)$ is itself a **Banach** space (i.e. complete normed vector space) equipped with the supremum norm:

$$\|f\|_\infty := \|f\|_{C_b(X)} := \sup_{x\in X} |f(x)| \tag{5.5}$$

We may consider **functionals** $T : f \mapsto Tf$ that are linear maps from $C_b(X)$ to $\mathbb{R}$. A functional $T$ is bounded (equivalent to being continuous) if there is a constant $C > 0$ such that for all $x$ one have $|Tf(x)| \le \|f\|_{C_b(X)}$. The supremum norm now induces a norm on the **dual space** of $C_b(X)$, i.e. vector space of all bounded linear functional $C_b(X)'$, which is defined as followed:

$$\|T\|_{C_b(X)'} := \sup_{f\in C_b(X)} \frac{|Tf|}{\|f\|_{C_b(X)}} = \inf\left\{C > 0 \,|\, \forall f \in C_b(X), |Tf| \le C\,\|f\|_{C_b(X)}\right\}. \tag{5.6}$$

**Exercise 5.5** Let $\mu$ be a probability measure on $X$. Check that the map

$$T_\mu : f \mapsto \int_X f(x)\,\mu(dx)$$

is a bounded linear functional on $C_b(X)$ with $\|T_\mu\|_{C_b(X)'} = 1$.

We note that there are at least two notions of convergence of functionals in $C_b(X)$: consider functionals $T_n, T \in C_b(X)'$ with $n \in \mathbb{Z}_{\ge 1}$

- $T_n \to T$ **uniformly** if $\|T_n - T\|_{C_b(X)'} \to 0$ as $n \to \infty$.
- $T_n \to T$ **weakly*** if for all $x \in X$ we have $T_n x \to Tx$. (This corresponds to pointwise convergence).

As an important example (which is the point of going through this discussion on this functional analysis topic):

**Example 5.6** $\mu_n \to \mu$ weakly $\iff T_{\mu_n} \to T_\mu$ weakly*.

**Remark 5.7** Do not be confused with the notion of weak convergence! Weak convergence usually describes the convergence of elements inside a Banach space, not its dual! Of course, one could say that $T_n \to T$ **weakly** as elements of $C_b(X)'$ as the "intended" Banach space, which refers to $gT_n \to gT$ for all $g$ in the **double dual** (dual of dual) space $C_b(X)''$. Note that this is strictly stronger than weak* convergence. This is because all evaluation maps $\mathsf{ev}_x : T \mapsto Tx$ are elements in $C_b(X)''$, which means

$$T_n \to T \text{ weakly } \implies \mathsf{ev}_x T_n = T_n x \to \mathsf{ev}_x T = Tx;$$

but not all elements in $C_b(X)''$ can be represented by the evaluation of an element in $C_b(X)$. In general, the notions of weak and weak* convergence coincide only when we are looking at Banach space for which its double dual coincides with itself.

We may therefore use the tools from functional analysis to prove theorems relating to weak convergence of measures. For example, we may draw parallels from proposition 3.12 to prove that weakly* convergence induces a topology on the set of probability measures on $(X, \mathscr{B}(X))$, which is generated by neighbourhoods of the following form:

$$U_{f_1,\dots,f_k,\epsilon}(\mu) := \left\{ \nu \,\middle|\, \forall i = 1, \dots, k, \left| \int_X f_i \, \mu(dx) - \int_X f_i \, \nu(dx) \right| < \epsilon \right\} \tag{5.7}$$

where $\mu, \nu$ is a probability measure on $X$. This also motivates the following definition of the Levy-Prokhorov metric:

$$\pi(\mu, \nu) := \inf \{ \epsilon > 0 \,|\, \mu(A) \le \nu(A_\epsilon) + \epsilon \text{ and } \nu(A) \le \mu(A_\epsilon) + \epsilon \}, \tag{5.8}$$

where $A_\epsilon := \{x \,|\, \rho(x, A) < \epsilon\}$ is as defined in the proof of lemma 5.2. The induced topology is also Polish, which is particularly useful when we want to talk about random probability measures.

**Exercise 5.8** Show that for the case when $(X, \mathscr{B}(X)) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ then the probability measures $\mu_n \to \mu$ weakly iff $\pi(\mu_n, \mu) \to 0$.

**Hint.** *Each probability measure of $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ can be represented by a distribution function. Rewrite the definition of the Levy-Prohorov metric to gain more insights.*

## 5.2 Portmanteau Theorem and related facts

But wait! We have defined the notion of convergence in distribution in another way in elementary probability classes. Specifically if $(X, \mathscr{B}(X)) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ then we say $\xi_n \to \xi$ in distribution if the distribution function $F_{\xi_n}(x) \to F_\xi(x)$ for all $x$ when $F_\xi$ is continuous. Turns out the Portmanteau theorem provides us with the required equivalence. This theorem was given its name since it involves a lot of equivalent statements, which look like lots of coats hanging on a coat hanger (*Portmanteau* in French). [3]

**Theorem 5.9 — Portmanteau Theorem.** Let $(X, \mathscr{B}(X))$ be a Polish space induced by a metric $d$ and $\mu_n, \mu$ are measures with $\mu_n(X), \mu(X) \le 1$. Writing $\mathbb{E}_\mu$ as expectation (or integral) with respect to the probability measure $\mu$, then the following are equivalent.

1. $\mu_n \to \mu$ weakly, that is, $\forall f \in C_b(X)$, $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_\mu[f]$.
2. For all **uniformly** continuous $f \in C_b(X)$, $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_\mu[f]$.
3. For all **bounded** $f \in \mathsf{Lip}(X)$, $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_\mu[f]$.
4. For all bounded measurable $f$ with $\mu(U_f) = 0$, $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_\mu[f]$, where $U_f$ is the set when $f$ is discontinuous.

[3] Figure provided by Hobvias Sudoneighm https://www.flickr.com/people/34427466731@N01 under the Creative Common 2.0 License: https://creativecommons.org/licenses/by-sa/2.0/

5. $\mu_n(X) \overset{n\to\infty}{\to} \mu(X)$, and for any closed set $E \subseteq X$, $\limsup \mu_n(E) \le \mu(E)$.
6. $\mu_n(X) \overset{n\to\infty}{\to} \mu(X)$, and for any open set $O \subseteq X$, $\liminf \mu_n(\xi_n \in O) \ge \mu(\xi \in O)$.
7. (Convergence in general) $\lim \mu_n(\xi_n \in C) = \mu(\xi \in C)$ for any $C$ such that $\mu(\xi \in \partial C) = 0$.

*Proof.* We will follow the order of proving (4) $\implies$ (1) $\implies$ (2) $\implies$ (3) $\implies$ (5) $\implies$ (6) $\implies$ (7) $\implies$ (4).

(4) $\implies$ (1) $\implies$ (2) $\implies$ (3). This is trivial.

(3) $\implies$ (5). The first condition can be proven by applying (3) with $f \equiv 1$. For the second condition, recall in the proof of lemma 5.2 we have defined a Lipschitz approximation for any indicator functions of a closed set $E$, which is denoted as $g_\varepsilon(x)$ in equation (5.4). Define also $E_\varepsilon = \{x : \rho(x, E) < \varepsilon\}$. Note that $E_\varepsilon \searrow E$ as $\varepsilon \searrow 0$, so we have

$$\mu_n(E) = \int f \,\mathrm{d}\mathbb{P}_n \le \int g_\epsilon \,\mathrm{d}\mu_n.$$

Therefore

$$\limsup_{n\to\infty} \mu_n(E) \le \limsup_{n\to\infty} \int g_\varepsilon \,\mathrm{d}\mu_n \overset{(3)}{=} \int g_\varepsilon \,\mathrm{d}\mu \le \mu(E_\varepsilon).$$

Since the above inequality holds for all $\epsilon > 0$, we can send $\epsilon \to 0$ to conclude that

$$\limsup_{n\to\infty} \mu_n(E) \le \mu(E)$$

as required.

(5 $\iff$ 6). consider $O = X \backslash E$, $B$ being closed.

(5, 6 $\implies$ 7). recall that $\overline{C} = C \cup \partial C$; $C^\circ := \operatorname{int} C = C \backslash \partial C$. Since $\mathbb{P}(\xi \in \partial C) = 0$.

$$\limsup \mu_n(C) \le \limsup \mu_n(\overline{C}) \overset{(5)}{\le} \mu(\overline{C}) = \mu(C),$$
$$\liminf \mu_n(C) \ge \liminf \mu_n(C^\circ) \overset{(6)}{\ge} \mu(C^\circ) = \mu(C),$$

so $\lim \mu_n(C) = \mu(C)$.

($7 \implies 4$). Let $f : X \to \mathbb{R}$ be a bounded and measurable function with $\mu(U_f) = 0$. We make the following claims:

**Claim 1:** For all $D \subseteq \mathbb{R}$, $\partial f^{-1}(D) \subseteq f^{-1}(\partial D) \cup U_f$.

**Hint.** *What happen if $x \in \partial f^{-1}(D)$ but $x \notin U_f$, i.e. $f$ is continuous at $x \in E$? Recall the definition of continuity at $x \in E$: $\forall \epsilon > 0$, $\exists \delta := \delta(\epsilon) > 0$ such that $f(B_\delta(x)) \subseteq B_\epsilon(f(x))$. Also recall that $f(x) \in \partial D$ iff $\forall \epsilon > 0$, $B_\epsilon(D) \cap A \neq \varnothing$ and $B_\epsilon(D) \cap A^c \neq \varnothing$.*

*Proof of claim 1.* Fix $\epsilon > 0$ and extract $\delta := \delta(\epsilon)$ as above. Provided that $x \in \partial f^{-1}(D)$, we know that there is a $y \in B_\delta(x) \cap f^{-1}(D)$ and $z \in B_\delta(x) \cap X \setminus f^{-1}(D)$. We therefore have

$$f(y) \in f(B_\delta(x) \cap f^{-1}(D)) \subseteq f(B_\delta(x)) \cap f(f^{-1}(D)) \subseteq B_\epsilon(f(x)) \cap D$$

and similarly $f(z) \in B_\epsilon(f(x)) \cap D^c$. Since $\epsilon > 0$ is arbitrary we have $f(x) \in \partial D$ as desired.

**Claim 2:** Consider the set $A := \{y \in \mathbb{R} \,|\, \mu(f(x) = y) > 0\}$, i.e. set of atoms of the finite push-forward measure $f^*\mu$. Then $A$ must be countable.

We leave that as an exercise.

*Finishing the proof.* Fix $\epsilon > 0$ and let $\|f\|_\infty = \sup_{x \in X} |f(x)|$ (for the case when $f$ is continuous then $\|f\|_\infty$ coincides with $\|f\|_{C_b(X)}$). Then by claim 2, we may choose $N + 1$ "grid points" such that

$$y_0 \leq -\|f\|_\infty < y_1 < ... < y_{N-1} < \|f\|_\infty < y_N,$$

such that for all $i$,

$$y_i \in \mathbb{R} \setminus A \quad \text{and} \quad y_{i+1} - y_i < \epsilon.$$

Now let $X_i = f^{-1}([y_{i-1}, y_i))$ for $i = 1, ..., N$. Then $X = \sqcup_{i=1}^N X_i$. By claim 1, we have

$$\mu(\partial X_i) \leq \mu(f^{-1}(\{y_{i-1}\})) + \mu(f^{-1}(\{y_i\})) + \mu(U_f) = 0$$

Therefore,

$$\limsup_{n\to\infty} \int f \, \mu_n(dx) \leq \limsup_{n\to\infty} \sum_{i=1}^N y_i \mu_n(E_i) \overset{(7)}{=} \sum_{i=1}^N y_i \mu(E_i) \leq \epsilon \mu(X) + \sum_{i=1}^N y_{i-1} \mu(E_i) \leq \varepsilon + \int f \, \mu(dx)$$

Since the above arguments hold for arbitrary $\epsilon > 0$, we see that

$$\limsup_{n\to\infty} \int f \, \mu_n(dx) \leq \int f \, \mu(dx)$$

We complete the proof by applying the above inequality with $-f$ to obtain the reverse inequality:

$$\limsup_{n\to\infty} \int -f \, \mu_n(dx) \leq \int -f \, \mu(dx) \iff \int f \, \mu(dx) \leq \liminf_{n\to\infty} \int f \, \mu_n(dx)$$

■

**Exercise 5.10**

1. Rewrite the statements of Portmanteau theorem for a sequence of random variables that converge in distribution.
2. Fill in the gap: prove that there are at most countable atoms for the finite measure $f^*\mu$.

**Hint.** *Argue by contradiction. The key step is to convince yourself that if $A$ is uncountable then there is an $\epsilon > 0$ such that $f^*\mu(\{x\})) \geq \epsilon$ for uncountably many $x \in A$ (otherwise you will have $A$ being countable!)*

We now look at some applications of the Portmanteau theorem:

### 5.2.1 Slutsky's Theorem and Convergence of Probability

Slutsky's theorem is the core of establishing the fact that convergence in probability implies convergence in distribution, as well as a couple of other facts relating to the arithmetic of weak limits of converging sequence of random variables. Let us now go through the statement of Slutsky's theorem:

**Theorem 5.11 — Slutsky.** Consider random variables $\xi, \xi_1, \ldots$ and $\eta, \eta_1, \ldots$ from $(\Omega, \mathscr{F}, \mathbb{P})$ to the Polish space $(X, \mathscr{B}(X))$ with metric $d$. Assume $\xi_n \overset{n\to\infty}{\to} \xi$ in distribution and $d(\xi_n, \eta_n) \overset{n\to\infty}{\to} 0$ in probability. Then $\eta_n \overset{n\to\infty}{\to} \xi$ in distribution.

*Proof.* Let $f : X \to \mathbb{R}$ be a bounded and Lipschitz function $f \in \mathsf{Lip}_K(X)$. Then for all $x \in E$,

$$|f(x) - f(y)| \leq Kd(x, y) \wedge 2 \|f\|_{C_b(X)}. \tag{5.9}$$

So by dominated convergence theorem one have $\limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}}[|f(\xi_n) - f(\eta_n)|] = 0$. But we also have $\limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}}[|f(\xi_n) - f(\xi)|] = 0$ by convergence in probability, so combining yields $\limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}}[|f(\eta_n) - f(\xi)|] = 0$. By statement (3) of the Portmanteau theorem we have $\eta_n \overset{n\to\infty}{\to} \xi$ in distribution. ■

As an immediate corollary:

**Corollary 5.12** Let $\xi_n, \xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (X, \mathscr{B}(X))$, then $\xi_n \overset{n\to\infty}{\to} \xi$ in probability implies $\xi_n \overset{n\to\infty}{\to} \xi$ in distribution. Moreover if $\xi_n \overset{d}{\to} \xi$ and $\xi \equiv c$ for some constant $c$, then $\xi_n \overset{p}{\to} c$.

$$\begin{array}{ccccc} & & \overset{L^p}{\to} & & \\ & & \Downarrow & & \\ \overset{a.e.}{\to} & \Longrightarrow & \overset{p}{\to} & \Longrightarrow & \overset{d}{\to} \end{array}$$

*Proof.* First statement is proven by applying $\xi_n \equiv \xi$. As for the second statement, let's say $\xi_n \overset{d}{\to} c$ where $c$ is a constant. Then by utilising statement (5) of the Portmanteau theorem, we know that

$$\lim_{n\to\infty} \mathbb{P}(d(\xi_n, c) \geq \epsilon) = \lim_{n\to\infty} \mathbb{P}(\xi_n \in X \setminus B_\epsilon(c)) \leq \mathbb{P}(c \in X \setminus B_\epsilon(c)) = 0.$$

■

**Remark 5.13**

- An alternative proof of the partial converse is as followed: if we assume the metric form of convergence in probability, then we can directly write

$$\mathbb{E}[d(X_n, c) \wedge 1] \overset{n\to\infty}{\to} 0, \tag{5.10}$$

  noticing that the function $f(.) = d(., c) \wedge 1$ is a continuous and bounded function on $X$.

- A counter-example showing convergence in distribution does not imply convergence in probability is as followed: consider a real-valued random variable $X$ that is symmetric about zero, e.g. $\xi \sim \mathsf{N}(0, 1)$. Then the sequence $\xi_n := (-1)^{n+1}\xi$ converges in distribution (since they are identically distributed) but not in probability.

Here is an example of using the above implication.

**Example 5.14 — Motivation of Weierstrass Approximation Theorem.** Recall the Bernoulli's law of large numbers, that for all $p \in [0, 1]$,

$$\frac{S_n}{n} \overset{p}{\to} p, \quad n \to \infty,$$

where $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$ and $\xi_i \sim \mathsf{B}(1,p)$. Putting

$$F_n(x) = \mathbb{P}\left(\frac{S_n}{n} \le p\right), \quad F(x) = \begin{cases} 1, & x \ge p, \\ 0, & x < p, \end{cases}$$

where $F(x)$ is the distribution function of the constant random variable $\xi \equiv p$. Also let $\mathbb{P}_n$ and $\mathbb{P}$ be the probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ corresponding to the distributions $F_n, F$. The above implication implies that $S_n/n \xrightarrow{d} p$, which means from statement (1) that

$$\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] \to \mathbb{E}[f(p)] = f(p), \quad n \to \infty,$$

for every bounded function $f(x) \in C_b(\mathbb{R})$. Since $S_n \sim \mathsf{B}(n,p)$, we know that

$$\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] = \int_{\mathbb{R}} f(x)\,\mathbb{P}_n(\mathrm{d}x) = \sum_{k=0}^{n}\left[\binom{n}{k}(1-p)^{n-k}p^k\right] f\left(\frac{k}{n}\right) =: f_n(p)$$

which is a polynomial in $p$. We therefore have for all $f \in C_b(\mathbb{R})$, $f_n(p) \to f(p)$ pointwise for all $p \in [0,1]$. Restricting $f$ to the compact interval $[0,1]$, we have recovered a weakened version of the Weierstrass approximation theorem, that there is a sequence of polynomials $f_n$ such that $f_n \to f$ pointwise on $[0,1]$. The actual approximation theorem goes further and establishes that the above convergence is actually uniform on $[0,1]$.

Here is another application of Slutsky's theorem regarding the addition of weak limits. This result is also named after Slutsky due to its relation with the original Slutsky's theorem.

**Proposition 5.15 — Slutsky, Addition of Limits.** Consider random variables $\xi, \xi_1, \ldots$ and $\eta_1, \ldots$ from $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ (or any Polish space with addition defined). Assume $\xi_n \overset{n\to\infty}{\to} \xi$ in distribution and $\eta_n \overset{n\to\infty}{\to} c$ in probability, with $c$ being a constant. Then $\xi_n + \eta_n \overset{n\to\infty}{\to} \xi + c$ in distribution.

*Proof.* First it is easily to check by that $\xi_n + c \overset{n\to\infty}{\to} \xi_n + c$ in distribution. (This can be done by e.g. shifting the closed sets and applying statement (5) of the Portmanteau theorem. Now note that $|(\xi_n + \eta_n) - (\xi_n + c)| \overset{n\to\infty}{\to} 0$ in probability, so by Slutsky's theorem we have $\xi_n + \eta_n \to \xi_n + c$ in distribution. ■

Note that it is not true in general that both $\xi_n \overset{n\to\infty}{\to} \xi$ and $\eta_n \overset{n\to\infty}{\to} \eta$ in distribution implies $\xi_n + \eta_n \overset{n\to\infty}{\to} \xi + \eta$ in distribution. As an example, consider again the real-valued random variable $\xi$ that is symmetric about zero. If $\xi_n \equiv \xi$ and $\eta_n = (-1)^{n+1}\xi$, then $\xi_n + \eta_n$ takes the form $(2\xi, 0, 2\xi, 0, \ldots)$ which does not converge in distribution.

Traditionally, Slutsky's theorem also contains the following statements

**Proposition 5.16 — Slutsky, Multiplication/Division of Limits.** Under the same setting as proposition 5.15, we have $\xi_n\eta_n \overset{n\to\infty}{\to} c\xi$ in distribution. In addition if $c \neq 0$ then $\xi_n/\eta_n \overset{n\to\infty}{\to} \xi/c$.

To prove this, we note the following result:

**Proposition 5.17 — Continuous Mapping Theorem.** Let $(X_1, \mathscr{B}(X_1))$ and $(X_2, \mathscr{B}(X_2))$ be metric spaces generated by metric $d_1$ and $d_2$ respectively, and let $\varphi : X_1 \to X_2$ be measurable. If $U_\varphi$ is the set of points of discontinuity of $\varphi$ (which is Borel measurable), then

1. If $\mu, \mu_1, \mu_2, \ldots$ are measures on $(X_1, \mathscr{B}(X_1))$ with $\mu(X_1), \mu_n(X_1) \le 1$ and $\mu(U_\varphi) = 0$ and $\mu_n \overset{n\to\infty}{\to} \mu$ weakly, then $\varphi^*\mu_n \overset{n\to\infty}{\to} \varphi^*\mu$ weakly.
2. If $\xi, \xi_1, \xi_2, \ldots : (\Omega, \mathscr{F}, \mathbb{P}) \to (X_1, \mathscr{B}(X_1))$ are random variables with $\mathbb{P}(\xi \in U_\varphi) = 0$ and $\xi_n \overset{n\to\infty}{\to} \xi$ in distribution, then $\varphi(\xi_n) \overset{n\to\infty}{\to} \varphi(\xi)$

*Proof.* Note that the second statement is an immediate result of the first one. To prove the first statement,

we are required to show that for all $f \in C_b(X_2)$,

$$\begin{aligned}\lim_{n\to\infty}\int_{X_2} f(x_2)\,\varphi^*\mu_n(dx_2) &\overset{\text{(LOTUS)}}{=} \lim_{n\to\infty}\int_{X_1} f(\varphi(x_1))\,\mu_n(dx_1)\\ &\overset{(*)}{=} \int_{X_1} f(\varphi(x_1))\,\mu(dx_1)\\ &\overset{\text{(LOTUS)}}{=} \int_{X_2} f(x_2)\,\varphi^*\mu(dx_1).\end{aligned}$$

The first and third equality is justified by the use of the change of variable formula (theorem 2.11). The second equality is justified by noting that $U_{f\circ\varphi} \subseteq U_\varphi$, so $\mu(U_{f\circ\varphi}) = 0$ and we may use statement (4) of the Portmanteau theorem to conclude. ■

We also note the following:

**Exercise 5.18** Under the setting of proposition 5.15, consider the random variables $T_n : \omega \in \Omega \mapsto (\xi_n(\omega), \eta_n(\omega))$ and $T : \omega \in \Omega \mapsto (\xi_n(\omega), c)$. Then $T, T_n$ are $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$-valued random variables. Show that $T_n \overset{n\to\infty}{\to} T$ in distribution.

**Hint.** *Try mimick the proof of proposition 5.15 by proving that* $(X_n, c) \overset{n\to\infty}{\to} (X_n, c)$ *in distribution and* $|(X_n, Y_n) - (X_n, c)| \overset{n\to\infty}{\to} 0$ *in probability, then utilise Slutsky's theorem.*

*Proof. of proposition 5.15/5.16.* Apply the result in the above exercise and continuous mapping theorem with map $\varphi(x, y) = x + y$, $xy$ or $x/y$. ■

### 5.2.2 Convergence of distribution function

Let us finally connect our definition of weak convergence with the definition of convergence of distribution in elementary probability classes, that is, the convergence of the distribution function. We restrict our discussion to real-valued random variables, although it can be extended to random variables taking value on $\mathbb{R}^n$.

**Theorem 5.19 — Convergence in distribution.**

1. Consider measures $\mu, \mu_n$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with $\mu(\mathbb{R}), \mu_n(\mathbb{R}) \le 1$ and define their distribution functions $F(x) := \mu((-\infty, x])$ and $F_n(x) := \mu_n((-\infty, x])$. Then $\mu_n \overset{n\to\infty}{\to} \mu$ weakly $\iff$
   - $F_n(x) \overset{n\to\infty}{\to} F(x)$ for all $x \in \mathbb{R} \setminus U_F$, $U_F$ being the set of discontinuities of $F$, and
   - $F_n(+\infty) \to F(+\infty) =: \lim_{x\to+\infty} F(x)$.
2. Consider random variables $\xi_n, \xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then $\xi_n \overset{n\to\infty}{\to} \xi$ in distribution $\iff$
   - $F_{\xi_n}(x) \overset{n\to\infty}{\to} F_\xi(x)$ when $x \in \mathbb{R} \setminus U_{F_\xi}$.
   - $F_{\xi_n}(+\infty) \overset{n\to\infty}{\to} F_\xi(+\infty)$

**Remark 5.20** We note that convergence at all points of continuity implies

$$F(+\infty) \le \liminf_{n\to\infty} F_n(+\infty) \tag{5.11}$$

To see this, we note that for all $x$, we have $F(x) \le \liminf_{n\to\infty} F_n(+\infty)$. Fix an $x \in \mathbb{R}$, then there is an $r > x$ such that $F$ is continuous at $r$. At this point we have $\liminf_{n\to\infty} F_n(r) = \lim_{n\to\infty} F_n(r) = F(r)$. Now note that, for all $n$, $F_n(r) \le F_n(+\infty)$, so we have $F(x) \le \liminf_{n\to\infty} F_n(+\infty)$. Since $x$ is arbitrary, we know that $F(+\infty) \le \liminf_{n\to\infty} F_n(+\infty)$.

From this, we know that if $F(+\infty) \ge \limsup_{n\to\infty} F_n(+\infty)$, then $\mu_n \overset{n\to\infty}{\to} \mu$ weakly.

*Proof.* We note that statement (2) immediately follows from (1). For the $(\Rightarrow)$ direction, it follows from the fact that $\mu(\partial(-\infty, x]) = \mu(\{x\}) = 0$ whenever $F$ is continuous at $x$, so by statement (7) of Portmanteau theorem we have $F_n(x) \overset{n\to\infty}{\to} F(x)$. For the $(\Leftarrow)$ direction, we attempt to use statement (3) of the

Portmanteau theorem; that is, if $f$ is a bounded Lipschitz function (wlog $f \in \mathsf{Lip}_1(\mathbb{R})$ and bounded by 1), then

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\,\mu_n(dx) = \int_{\mathbb{R}} f(x)\,\mu(dx).$$

Using arguments in proving (7) $\implies$ (4) in the Portmanteau theorem, we know that $U_F$ is countable. As a result, we may choose $N$ "knot points", but in a way different from the one in the proof of the Portmanteau theorem. Specifically, we want

$$y_1 < y_2 < \ldots < y_{N-1}$$

such that for all $i$,

$$y_i \notin U_F, \quad y_i - y_{i-1} < \epsilon, \quad F(y_1) < \epsilon, \quad F(y_{N-1}) > F(\infty) - \epsilon.$$

With the notation of $y_0 = -\infty$ and $y_N = \infty$ (so that $F(y_0) = F_n(y_0) = 0$), we obtain the bound:

$$\int f(x)\,\mu_n(dx) = \sum_{i=0}^{n} f(x)\,\mu_n(dx) \le F_n(y_1) + F_n(\infty) - F_n(y_{N-1}) + \sum_{i=1}^{N-1} \int_{y_i}^{y_{i+1}} f(x)\,\mu_n(dx).$$

But we know that for all $x \in [y_i, y_{i+1})$

$$|f(x) - f(y_i)| \le |x - y_i| \le y_i - y_{i-1} < \epsilon \implies f(y_i) - \epsilon < f(x) < f(y_i) + \epsilon,$$

so

$$\begin{aligned}\int f(x)\,\mu_n(dx) &< (F_n(y_1) + F_n(\infty) - F_n(y_{N-1})) + \sum_{i=1}^{N-1} (f(y_i) + \epsilon)(F_n(y_{i+1}) - F_n(y_i)) \\ &= (F_n(y_1) + F_n(\infty) - F_n(y_{N-1})) + \sum_{i=1}^{N-1} f(y_i)(F_n(y_{i+1}) - F_n(y_i)) + \epsilon F_n(\infty)\end{aligned}$$

Note that for all $i$, $F_n(y_i) \overset{n\to\infty}{\to} F(y_i)$. Recall that $F(\infty) \le 1$, we have

$$\begin{aligned}\limsup_{n\to\infty} \int f(x)\,\mu_n(dx) &< 3\epsilon + \sum_{i=1}^{N-1} f(y_i)(F(y_{i+1}) - F(y_i)) \\ &\le \int_{i=1}^{N-1} \int_{y_i}^{y_{i+1}} (f(x) + \epsilon)\,\mu(dx) \\ &\le 4\epsilon + \int f(x)\,\mu(dx)\end{aligned}$$

owing to the fact that $f(x) + \epsilon \ge f(y_i)$ for all $x \in [y_i, y_{i+1})$. Sending $\epsilon \searrow 0$ yields

$$\limsup_{n\to\infty} \int f(x)\,\mu_n(dx) \le \int f(x)\,\mu(dx)$$

Replacing $f$ by $(1 - f)$ yields the desired reverse inequality. ■

Here we raise some applications of the above result. First of all, (theorem 4.12)

**Example 5.21 — de Moivre's Central Limit Theorem.** For $\xi_i \overset{\text{iid}}{\sim} \mathsf{B}(1,p)$ and $S_n = \sum_{i=1}^n \xi_i$, recall the random variable $\eta_n = (S_n - np)/\sqrt{np(1-p)}$ satisfies

$$\lim_{n\to\infty} \mathbb{P}(\eta_n \le x) = \Phi(x),$$

where $\Phi(x)$ is the distribution function of the standard normal $\mathsf{N}(0,1)$ distribution. Therefore, **the probability distribution of $\eta_n$ converges to a $\mathsf{N}(0,1)$ distribution**. By definition of weak conver-

gence, for any $f \in C_b(\mathbb{R})$ we know that

$$\mathbb{E}\left[f\left(\frac{S_n - np}{\sqrt{np(1-p)}}\right)\right] \overset{n\to\infty}{\to} \int_{\mathbb{R}} f(t)\left(\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\right) dt \tag{5.12}$$

Here is another shortcut for proving convergence in distribution via the convergence of their densities.

**Corollary 5.22** Suppose that the random variables $\xi_n, \xi$ have densities $f_n(x), f(x)$, respectively. Also let $f_n(x) \to f(x)$ for any $x$. Then $\xi_n \xrightarrow{d} \xi$.

*Proof.* It is sufficient to show that

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

for any $x$. Using

$$|F_n(x) - F(x)| \le \int_{-\infty}^{x} |f_n(y) - f(y)|\,\mathrm{d}y,$$

we need to check that

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} |f_n(y) - f(y)|\,\mathrm{d}y = 0.$$

Note that $a = a_+ - a_-$ and $|a| = a_+ + a_-$ for any real $a$. Since $f_n$ and $f$ are densities, they integrate to 1, so for each $n$

$$0 = \int_{-\infty}^{\infty} (f(y) - f_n(y))\,\mathrm{d}y = \int_{-\infty}^{\infty} [(f(y) - f_n(y))_+ - (f(y) - f_n(y))_-]\,\mathrm{d}y.$$

Then

$$\int_{-\infty}^{\infty} |f(y) - f_n(y)|\,\mathrm{d}y = 2\int_{\infty}^{\infty} (f(y) - f_n(y))_+\,\mathrm{d}y,$$

which goes to zero by the dominated convergence theorem. Indeed

$$0 \le (f(y) - f_n(y))_+ \le f(y),$$

for any $n$ and the function $f(y)$ is integrable. ■

**Exercise 5.23**

1. Let $F_n \to F$ and suppose that $F$ is continuous. Show that $F_n$ converges *uniformly* to $F$, i.e.

   $$\sup_x |F_n(x) - F(x)| \to 0, \quad n \to \infty.$$

   *Hint:* Idea is to consider knot points as in theorem 5.19.

2. Give an example of distribution functions $F_n(x), F(x)$ such that $F_n(x) \xrightarrow{d} F(x)$, but

   $$\sup_x |F_n(x) - F(x)| \not\to 0, \quad n \to \infty.$$

3. Give an example of probability measures $\mathbb{P}, \mathbb{P}_n$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, $n \ge 1$ such that $\mathbb{P}_n \xrightarrow{d} \mathbb{P}$, but convergence $\mathbb{P}_n(B) \to \mathbb{P}(B)$ need not hold for all Borel sets $B \in \mathscr{B}(\mathbb{R})$.

## 5.3 Skorohod Representation Theorem

We now prove that a sequence of weakly convergence probability measures can be represented as random variables defined on a common probability space.

**Theorem 5.24 — Skorohod Representation Theorem.** Suppose $\mu, \mu_n$ $(n = 1, 2, ...)$ are probability measures on a Polish space $(X, \mathscr{B}(X))$ induced by a metric $d$, so that $\mu_n \overset{n\to\infty}{\to} \mu$ weakly. Then there exists a single probability space $(\Omega', \mathscr{F}', \mathbb{P}')$ and random variables $\xi_1, \xi_2, \dots$ on it such that $\xi$ has distribution $\mu$ ($\xi^*\mathbb{P} = \mu$), $\xi_n$ has distribution $\mu_n$ ($\xi_n^*\mathbb{P} = \mu_n$) and that $\xi_n \to \xi$ $\mathbb{P}$-almost surely.

We will not go through the proof for general Polish space, if interested one can read Dudley's paper [5]. However if we limit ourselves to the case when $(X, \mathscr{B}(X)) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ then there is a simple proof.

*Proof.* Let $(\Omega', \mathscr{F}', \mathbb{P}') = ([0,1], \mathscr{B}([0,1], \mathsf{Leb}))$. Note that the probability measures $\mu, \mu_n$ induces distribution functions $F_n, F : \mathbb{R} \to [0,1]$. Let $F^{-1}, F_n^{-1}$ be their right inverses as defined in equation 3.3, then by probability integral transform (proposition 3.2), they have same distribution as $\mu$ and $\mu_n$'s respectively. It remains to prove that $F_n^{-1}(u) \to F^{-1}(u)$ almost surely as $n \to \infty$. In fact we can prove the required limit for any $u$ such that the preimage of $\{u\}$ under $F$ is finite (i.e. a singleton or empty set). Since there are only countably many $u$ such that the above condition doesn't hold (given that $F$ is increasing and right continuous), establishing the above limit will complete the proof.

Let $u$ be a point such that the preimage of $\{u\}$ under $F$ is finite. We make two observations:

- If $x < F^{-1}(u)$, then using the argument in the proof of proposition 3.2 we have $F(x) < u$. If $F$ is continuous at $x$, then by extending the arguments in the above proof of Portmanteau theorem we see that $F_n(x) \to F_\xi(x)$ as $n \to \infty$, and that for sufficiently large $n$ we have $F_n(x) < u \iff x \le F_n^{-1}(u)$. We therefore have $x \le \liminf_{n\to\infty} F_n^{-1}(u)$. Now consider a sequence $x_k \nearrow F^{-1}(u)$ such that $F$ is continuous as $x_k$, which exists by the fact that $F$ is increasing (otherwise $F$ have uncountable discontinuities!) Then for all $k$ one have $x_k \le \liminf_{n\to\infty} F_n^{-1}(u)$, and that $F^{-1}(u) \le \liminf_{n\to\infty} F_n^{-1}(u)$ by sending $k \to \infty$.
- If $x > F^{-1}(u)$, then $F(x) \ge u$. In fact we must have $F(x) > u$, for if $F(x) = u$ then $F(y) = u$ for any $y \in [F^{-1}(u), x]$, contradicting the assumption that the preimage of singleton $\{u\}$ is finite. Repeating the above arguments yields $F^{-1}(u) \ge \limsup_{n\to\infty} F_n^{-1}(u)$.

Combining the observations we have $F_n^{-1}(u) \to F^{-1}(u)$ as desired. ∎

As a result, we may restate any theorems (including Central Limit Theorem) in terms of random variables.

**Example 5.25 — Restating de Moivre's CLT.** We may construct a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, random variables $\eta_n, \eta$ such that $\eta$ is normally distributed and $\sqrt{np(1-p)}\eta_n + np$ has a $\mathsf{B}(n,p)$ distribution.

Of course, more effort is needed to actually construct a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, random variables $\xi_n \overset{\text{iid}}{\sim} \mathsf{B}(1,p)$ and $\eta \sim \mathsf{N}(0,1)$ such that the de Moivre's CLT holds, and this will be omitted in our discussion. Nevertheless, this representation theorem gives us a shortcut for proving results relating to the weak convergence.

**Example 5.26 — Alternative proof of the theorem 5.19.** Let's say we have probability measures $\mathbb{P}_n, \mathbb{P}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that their distribution function converges at all points of continuity. By Skorohod representation theorem, we may construct a sequence of random variables $\bar{\xi}, \bar{\xi}_1, \bar{\xi}_2, ...$ on the common probability space $([0,1], \mathscr{B}([0,1]), \mathsf{Leb})$ with distributions $F_\xi, F_{\xi_1}, F_{\xi_2}, ....$ Let $f \in C_b(\mathbb{R})$, then by change of variable and dominated (or bounded) convergence theorem,

$$\lim \mathbb{E}_{\mathbb{P}_n}[f_n(\xi_n)] = \lim \mathbb{E}_{\mathsf{Leb}}[f_n(\bar{\xi}_n)] \overset{\text{(DCT)}}{=} \mathbb{E}_{\mathsf{Leb}}[f(\bar{\xi})] = \mathbb{E}_{\mathbb{P}}[f(\xi)],$$

noting that $f_n(\bar{\xi}_n) \to f_n(\bar{\xi})$ almost surely (convince yourself!) This completes the proof.

## 5.4 Relative Compactness and Tightness

We will later prove the well-known Central Limit Theorem by looking at characteristic functions. To establish the relationship between characteristic functions and measures, we need some tools relating to the collections of measures, including clarifying the meaning of how a collection is *relatively compact* and *tight*. These concepts are also very useful in proving results relating to stochastic processes and ergodic theory. We begin by the following definition:

**Definition 5.27** A family of probability measures $\mathscr{P} = \{\mathbb{P}_\alpha, \alpha \in A\}$ on a Polish space (and the corresponding set of distribution functions $F_\alpha$ if the underlying Polish space is $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$) is called **relatively compact** if every sequence of measures from $\mathscr{P}$ contains a subsequence that weakly converges to a probability measure.

**Remark 5.28** We emphasise that in this definition the limit measure is to be a probability measure, although it need not belong to the original class $\mathscr{P}$. (This is why the word "relatively" appears in the definition.) In fact, $\mathscr{P}$ is relatively compact if its closure with respect to the Levi-Prokhorov metric is compact.

As an example, the collection containing a weakly convergent sequence of measures is relatively compact. In fact, we have the following.

**Lemma 5.29** If $\{\mathbb{P}_n\}$ and $\mathbb{P}$ are probability measures, then $\mathbb{P}_n \xrightarrow{d} \mathbb{P}$ if and only if every subsequence $\{\mathbb{P}_{n'}\}$ of $\{\mathbb{P}_n\}$ contains a subsequence $\{\mathbb{P}_{n''}\}$ such that $\mathbb{P}_{n''} \xrightarrow{d} \mathbb{P}$.

*Proof.* From the definition of weak convergence we know that $\mathbb{P}_n \xrightarrow{d} \mathbb{P}$ if and only if the following sequence of numbers converges for any continuous bounded function $f$:

$$\int f(x)\,\mathbb{P}_n(\mathrm{d}x) \to \int f(x)\,\mathbb{P}(\mathrm{d}x).$$

So necessity is trivial, just take $\{n''\} = \{n'\}$. For the sufficiency, assume the converse to get a contradiction. Then there exists an $f$, an $\varepsilon > 0$ and a subsequence $\{n'\}$ such that

$$\left| \int f(x)\mathbb{P}_{n'}(\mathrm{d}x) - \int f\,\mathbb{P}(\mathrm{d}x) \right| > \varepsilon \quad \forall n'.$$

This clearly can not be true as by the assumption, $\mathbb{P}_{n''} \xrightarrow{d} \mathbb{P}$, so

$$\left| \int f(x)\mathbb{P}_{n'}(\mathrm{d}x) - \int f\,\mathbb{P}(\mathrm{d}x) \right| < \varepsilon$$

for large enough $n''$. ∎

A given family of probability measures $\mathscr{P}$ is not necessarily relatively compact. This is illustrated by the following examples.

**Example 5.30** Let $\xi_n$ be real-valued random variables and $F_{\xi_n}(x) \to F(x)$ for all $x \in U_F$. Then $F(x)$ is not necessarily a distribution function!

1. Let $\xi_n$ be $U[n, n+1]$. Then $F(x) \equiv 0$ (called runaway to infinity).
2. Let $\xi_n$ be $U[-n, n]$. Then $F(x) \equiv 1/2$ (called spread to infinity).

The following theorem will provide a necessary and sufficient condition for a family of probability measures (or finite measures) to be relatively compact. We first define the notion of tightness:

**Definition 5.31 — Tightness.** A family of probability measures $\mathscr{P} = \{\mathbb{P}_\alpha\}_{\alpha \in A}$ on a Polish space $(X, \mathscr{B}(X))$ induced by metric $d$ (or family of finite measures $\{\mu_\alpha\}_{\alpha \in A}$ with $\mu_\alpha(X) \leq 1$) is **tight** if for all $\varepsilon > 0$ there exists a compact set $K \subset X$ such that

$$\sup_{\alpha \in A} \mathbb{P}_\alpha(X \backslash K) \leq \varepsilon. \tag{5.13}$$

Turns out tightness is a sufficient condition for relative compactness (and is necessary if $(X, \mathscr{B}(X))$ is Polish)

**Theorem 5.32 — Prokhorov's theorem.** A tight family of probability measures $\mathscr{P}$ (or finite measures as specified in definition 5.31) of $(X, \mathscr{B}(X))$ induced by metric is relatively compact. If in addition $(X, \mathscr{B}(X))$ is Polish then the converse also holds.

Here we only discuss the proof for the case when $(X, \mathscr{B}(X)) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. In such case, the Prokhorov's theorem reduces to the Helly's selection theorem. We denote the collection of functions $\mathscr{G} = \{F : \mathbb{R} \to [0,1]\}$ such that $F$ is nondecreasing, continuous from the right. This is called the set of generalised distribution functions.

**Remark 5.33** Distribution functions form a subset of $\mathscr{G}$ for which $F(-\infty) = 0$ and $F(\infty) = 1$.

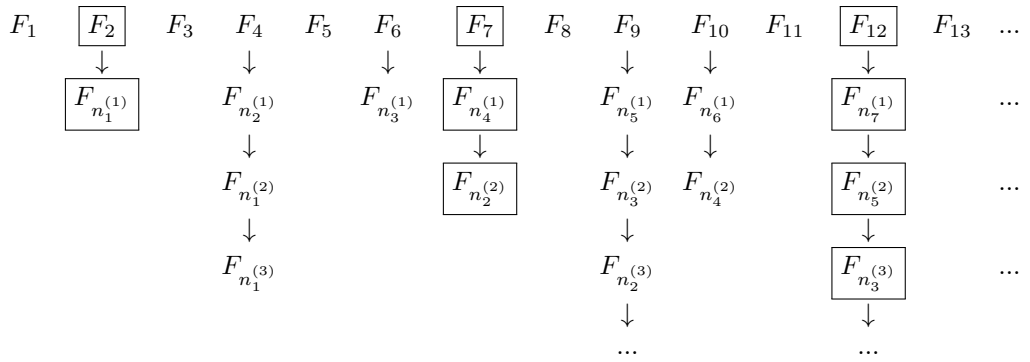
We may then state the Helly's theorem

**Theorem 5.34 — Helly's Selection Theorem.** The set $\mathscr{G}$ of generalised distribution functions is sequentially compact, i.e. for any sequence $F_n \in \mathscr{G}$, $n = 1, 2, \ldots$ there exists a function $F \in \mathscr{G}$ and a subsequence $F_{n_k} \subseteq \{F_n\}$ such that $F_{n_k}(x) \to F(x)$ for every point $x \in \mathbb{R} \setminus U_F$, $U_F$ being set of points of discontinuities of $F$.

*Proof.* The theorem uses a classic diagonal sequence argument. Let $(q_k)_{k\geq 1}$ be an enumeration of $\mathbb{Q}$ (bijection from $\mathbb{N}$ to $\mathbb{Z}$).

- Consider the sequence $(F_n(q_1))_{n\geq 1}$, which is bounded, so by Bolzano-Weierstrass theorem (BW) it has a subsequence $\left(F_{n_k^{(1)}}(q_1)\right)_{k\geq 1}$ which converges to some number $G(q_1) \in [0,1]$.
- Now consider the sequence $\left(F_{n_k^{(1)}}(q_2)\right)_{k\geq 1}$, which is also bounded, so by BW again it has a subsequence $\left(F_{n_k^{(2)}}(q_2)\right)_{k\geq 1}$ which converges to some number $G(q_2) \in [0,1]$.
- We repeat to extract further subsequences...

We can illustrate the above procedure by picture:




We then extract the "diagonal subsequence" is as highlighted above, and write it as $F_{n_k} := F_{n_k^{(k)}}$. An important observation is that the sequence $(F_{n_k})_{k\geq m}$ is a subsequence of $\left(F_{n_k^{(m)}}\right)_{k\geq 1}$ for any $m \geq 1$.

**Claim 1:** For all $q \in \mathbb{Q}$, we have $F_{n_k}(q) \to G(q)$.

*Proof of claim 1:* First find $m$ such that $q_m = q$, then note that $(F_{n_k}(q))_{k\geq m}$ is a subsequence of $\left(F_{n_k^{(m)}}(q)\right)_{k\geq 1}$ which converges to $G(q)$.

**Claim 2:** $G$ is a non-decreasing function, that is if $p, q \in \mathbb{Q}$ and $p \leq q$, then $G(p) \leq G(q)$. This trivially follows from the monotonicity of limits.

As a result, we may "fill in the gaps" and define

$$F(x) = \inf\{f(q) : q \in \mathbb{Q}, q > x\} = \lim_{q \to x+} G(q) \tag{5.14}$$

Then it is quite trivial that $F \in \mathscr{G}$. Moreover, $\forall q \in \mathbb{Q}$ we have $F(q) \geq G(q)$, if $x < q$ then $F(x) \leq G(q)$, and $F(q) = G(q)$. It remains to show that $F_{n_k}(x) \to F(x)$ for all $x \in \mathbb{R} \setminus U_F$, so let us fix $x \in \mathbb{R} \setminus U_F$ and some arbitrary $\epsilon > 0$. By continuity one can choose $y < r < x < q$ with $r, q \in \mathbb{Q}$, such that

$$F(x) - \epsilon < F(y) \leq F(r) = G(r) \leq F(x) \leq F(q) = G(q) < F(x) + \epsilon,$$

and so for sufficiently large $k$ we have

$$F_{n_k}(r), F_{n_k}(q) \in (F(x) - \epsilon, F(x) + \varepsilon) \implies F_{n_k}(x) \in (F(x) - \epsilon, F(x) + \varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, the above arguments complete the proof. ■

We may then prove the Prokhorov's theorem for the case when $(X, \mathscr{B}(X)) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

*Proof. of Prokhorov's theorem for* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. We first prove the partial converse. Let's say that $\mathscr{P} := \{\mu_\alpha\}_{\alpha \in A}$ is a relatively compact but not tight family of measures with $\mu_\alpha(\mathbb{R}) \leq 1$. Then $\exists \epsilon > 0$ such that for any compact $K \subset \mathbb{R}$, $\sup_\alpha \mu_\alpha(\mathbb{R} \setminus K) < \epsilon$. Hence, for any $n$ there is a $\mu_{\alpha_n}$ such that

$$\mu_{\alpha_n}(\mathbb{R} \setminus (-n, n)) > \epsilon. \tag{5.15}$$

By relative compactness, there is a subsequence $(\mu_{\alpha_{n_k}})_{k \geq 1}$ such that $\mu_{\alpha_{n_k}} \overset{k \to \infty}{\to} Q$ weakly, where $Q$ is a finite measure with $Q(\mathbb{R}) \leq 1$. By statement (5) of the Portmanteau theorem, we have

$$\limsup_{k \to \infty} \mu_{\alpha_{n_k}}(\mathbb{R} \setminus (-n, n)) \leq Q(\mathbb{R} \setminus (-n, n)) \tag{5.16}$$

But the RHS $\searrow 0$ as $n \to \infty$, which contradicts with (5.15). Hence $\mathscr{P}$ must be tight.

Now let $\mathscr{P}$ be tight, and $(\mu_n)$ be a sequence of elements in $\mathscr{P}$ associated with the distribution function $(F_n)$, $F_n(x) = \mu_n((-\infty, x])$. By Helly's selection theorem, there exists a subsequence $F_{n_k}(x) \to F(x)$ at $x \in \mathbb{R} \setminus U_F$. We now check that $F(-\infty) = 0$ and $F_{n_k}(+\infty) \to F(+\infty)$. Fix $\epsilon > 0$, then from tightness there is a uniform constant $K \in (0, \infty)$ such that,

$$\mu_{n_k}(\mathbb{R} \setminus (-K, K)) = \underbrace{(F_{n_k}(+\infty) - F_{n_k}(K))}_{\geq 0} + \underbrace{(F_{n_k}(-K) - F_{n_k}(-\infty))}_{\geq 0} < \epsilon \tag{5.17}$$

Of course, this means that for all $x < -K$ and $y > K$

$$\begin{aligned} F_{n_k}(x) = F_{n_k}(x) - F_{n_k}(-\infty) < \epsilon \\ F_{n_k}(+\infty) - F_{n_k}(y) < \epsilon \end{aligned}$$

Choosing $x, y$ such that $F$ is continuous at those points, then

$$\begin{gathered} F(-\infty) \leq F(x) \leq \limsup_{k \to \infty} F_{n_k}(x) < \epsilon. \\ \limsup_{k \to \infty} F_{n_k}(+\infty) \leq \limsup_{k \to \infty} F_{n_k}(y) + \epsilon = F(y) + \epsilon \leq F(+\infty) + \epsilon. \end{gathered}$$

Sending $\epsilon \searrow 0$ yields $F(-\infty) = 0$ and $F(+\infty) \geq \limsup_{k \to \infty} F_{n_k}(+\infty)$. But by remark 5.20 we have $F(+\infty) \leq \liminf_{k \to \infty} F_{n_k}(+\infty)$, so $\lim_{k \to \infty} F_{n_k}(+\infty) = F(+\infty)$ as desired. ■

**Exercise 5.35**

1. Let $\mathbb{P}_\alpha$ be a Gaussian measure on the real line, with parameters $m_\alpha$ and $\sigma_\alpha^2$, $\alpha \in \mathscr{A}$. Show that the family $\{\mathbb{P}_\alpha, \alpha \in \mathscr{A}\}$ is tight if and only if

$$|m_\alpha| \leq \alpha, \quad \sigma_\alpha^2 \leq b, \quad \alpha \in \mathscr{A}.$$

2. Let $\xi_n$ be i.i.d. random variables with a standard exponential distribution. Denote

$$M_n = \max\{\xi_1, \ldots, \xi_n\}.$$

Is the sequence $M_n - \log n$ converging in distribution to a generalised distribution function? Is it tight? Is it weakly converging?

3. Consider Polish space $(X, \mathscr{B}(X))$ with metric $d$. Recall that a class of function $\mathscr{C}$ from $X$ to $\mathbb{R}$ is separating if $\forall f \in \mathscr{C}, \int_X f\,d\mu = \int_X f\,d\nu \implies \mu = \nu$. Show that if $\mu_n, \mu$ are measures on $(X, \mathscr{B}(X))$ such that $\mu_n(X), \mu(X) \leq 1$ then $\mu_n \overset{n\to\infty}{\to} \mu$ weakly $\iff (\mu_n)_{n\geq 1}$ is tight, and there is a separating family $\mathscr{C} \subseteq C_b(X)$ such that

$$\int_X f\,d\mu = \lim_{n\to\infty} \int_X f\,d\mu_n \tag{5.18}$$

The proof for the general case is not very much different, but requires more careful justifications at certain steps. Here we outline the key steps of proving the Prokhorov's theorem.

*Proof.* (Forward direction, sketch).

**Step 0: A simplification:** Let's say $\mathscr{P} := \{\mu_\alpha\}_{\alpha\in A}$ is tight. Then there are compact sets $K_1 \subseteq K_2 \subseteq \ldots$ such that $\mu_\alpha(X \setminus K_n) \leq 1/n$ for all $\alpha$. Define $X' = \lim_{n\to\infty} K_n$, then $\mu_\alpha(X \setminus X') = 0$, and we can treat $\mu_\alpha$ as measures on $X'$. As a result, we may assume that $X$ is $\sigma$-compact (i.e. countable union of compact spaces).

**Step 1: Construction using diagonal arguments:** We note that $X$ is separable, and hence posesses a countable base $\mathscr{U} := \{U_i\}_{i\in\mathbb{N}}$ such that any open sets of $X$ can be written as a countable union of elements in $\mathscr{U}$. Now consider the following countable collections of compact sets:

$$\mathscr{C}' = \left\{\bar{U} \cap K_n \,:\, U \in \mathscr{U}, n \geq 1\right\} \tag{5.19}$$

$$\mathscr{C} = \left\{\bigcup_{i=1}^{n} C_i, n < \infty, C_i \in \mathscr{C}'\right\} \tag{5.20}$$

Note that $\mathscr{C}$ is closed under finite unions, and that $K_n \in \mathscr{C}$.

Now let $(\mu_n)_{n\geq 1}$ be a sequence of elements in $\mathscr{P}$. We may use the diagonal construction in the proof Helly's selection theorem to extract a subsequence $(\mu_{n_k})_{k\geq 1}$, such that for all $C \in \mathscr{C}$, the sequence $(\mu_{n_k}(C))$ possesses a limit:

$$\lim_{k\to\infty} \mu_{n_k}(C) =: \alpha(C).$$

**Step 2: Extending the set function:** This is probably the most technical part, and we will omit the details. If interested, you may refer to page 265-268 of [1]. At the very end, we will be able to construct a measure $\mu$ such that for all open sets $A$,

$$\mu(A) = \sup_{C\in\mathscr{C}, C\subseteq A} \alpha(C). \tag{5.21}$$

From this, we have

$$\mu(X) \geq \sup_{n\in\mathbb{N}} \alpha(K_n) = \sup_{n\in\mathbb{N}} \lim_{n\to\infty} \mu_{n_k}(K_n) \geq \sup_{n\in\mathbb{N}} \limsup_{n\to\infty} \left(\mu_{n_k}(X) - \frac{1}{n}\right) = \limsup_{n\to\infty} \mu_{n_k}(X).$$

Moreover, for any open set $A$ and $C \in \mathscr{C}, C \subset A$,

$$\alpha(C) = \lim_{k\to\infty} \mu_{n_k}(C) \leq \liminf_{k\to\infty} \mu_{n_k}(A)$$

so by taking supremum we have $\mu(A) \leq \liminf_{k\to\infty} \mu_{n_k}(A)$. We therefore have $\lim_{k\to\infty} \mu_{n_k}(X) = \mu(X)$, so by statement (5) of the Portmanteau theorem we have $\mu_{n_k} \overset{n\to\infty}{\to} \mu$ weakly, completing this section of proof.

(Partial converse). This is less useful in practice, but we are including it for completion. Since $X$ is separable, it possesses a countable dense set $\{x_i\}_{i\geq 1}$. For $n \in \mathbb{N}$ we consider the union $A_{n,N} :=$

$\cup_{i=1}^{N} B_{1/n}(x_i)$. Then $\forall n, A_{n,N} \nearrow X$. Now consider the quantity

$$\delta = \sup_{n\in\mathbb{N}} \inf_{N\in\mathbb{N}} \sup_{\mu\in\mathscr{F}} \mu(A_{n,N}^c), \tag{5.22}$$

which exists since the quantity $\mu(A_{n,N}^c)$ is bounded above and below. We want to show that $\delta = 0$, so let's assume that it's not the case. Unfolding the definition, there is an $n \in \mathbb{N}$ such that

$$\begin{aligned}
&\inf_{N\in\mathbb{N}} \sup_{\mu\in\mathscr{F}} \mu(A_{n,N}^c) > \delta/2 > 0, \\
\iff &\forall N \in \mathbb{N}, \sup_{\mu\in\mathscr{F}} \mu(A_{n,N}^c) > \delta/2 \\
\iff &\forall N \in \mathbb{N}, \exists \mu_N \in \mathscr{F} \text{ such that } \mu_N(A_{n,N}^c) > \delta/2.
\end{aligned}$$

From this, we extract a sequence $(\mu_N)$ of elements in $\mathscr{P}$. Since $\mathscr{P}$ is relatively compact, the specific sequence possesses a subsequence $(\mu_{N_k})_{k\geq 1}$ which converges weakly to $\nu$. By statement (6) of the Portmanteau theorem, we know that $\forall n \in \mathbb{N}$,

$$\nu(A_{n,N}^c) \geq \liminf_{k\to\infty} \mu_{n_k}(A_{n,N}^c) \geq \liminf_{k\to\infty} \mu_{n_k}(A_{n,N_k}^c) > \delta/2 > 0$$

But $A_{n,N}^c \searrow \varnothing$ as $N \to \infty$, so $\nu(A_{n,N}) \overset{N\to\infty}{\to} 0$, which is a contradiction. So we must have $\delta = 0$.

What does this mean? This means that

$$\sup_{n\in\mathbb{N}} \inf_{N\in\mathbb{N}} \sup_{\mu\in\mathscr{P}} \mu(A_{n,N}^c) < \text{ any positive numbers.}$$

In fact for all $n \in \mathbb{N}$ we have

$$\inf_{N\in\mathbb{N}} \sup_{\mu\in\mathscr{P}} \mu(A_{n,N}^c) < \frac{\epsilon}{2^n},$$

so there exists $N_n'$ such that $\sup_{\mu\in\mathscr{P}} \mu(A_{n,N_n'}^c) < \epsilon/2^n$.

Finally, we note that the set $A := \cap_{n=1}^{\infty} A_{n,N_n'}$ is totally bounded and hence pre-compact (with compact closure). Further, for any $\mu \in \mathscr{P}$, we have

$$\mu(\bar{A}^c) \leq \mu(A^c) \leq \sum_{n=1}^{\infty} \mu(A_{n,N_n'}^c) < \epsilon,$$

hence $\mathscr{P}$ is tight. ■

## 5.5 Vague Convergence

We conclude our discussion with another notion of convergence of measures, namely vague convergence. This is even weaker than the notion of weak convergence we have covered, and we need that since this is what we get when the moment generating functions (MGF) / characteristic functions (CF) of a sequence of random variables converges. We will show that in some cases, vague convergence is equivalent to weak convergence, then we might use the convergence of MGF/CF to establish weak convergence of random variables.

The definition of vague convergence is quite similar to the one for weak convergence. However, we have to state an additional assumption on the space $(X, \mathscr{B}(X))$, that it is **locally compact** such that each points $x \in X$ has a compact neighbourhood. The common spaces $(\mathbb{R}, \mathscr{B}(\mathbb{R})), (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ and $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ are all locally compact. To state this definition, recall that the support of a **continuous** real-valued function $f : X \to \mathbb{R}$ is $\overline{X \setminus f^{-1}(\{0\})}$. The definition of vague convergence is as followed:

**Definition 5.36 — Vague convergence of measures and random variables.**

- Let $\mu_n, \mu$ be measures on the above Polish space $(X, \mathscr{B}(X))$. We say that $\mu_n$ converges to $\mu$ **vaguely** as $n \to \infty$ if for all $f \in C_c(X)$, we have

$$\int_X f(x)\,\mu_n(dx) \overset{n\to\infty}{\to} \int_X f(x)\,\mu(dx),$$

  where $C_c(X)$ represents the set of all continuous functions on $X$ with **compact** support.

- Let $\xi_n : (\Omega_n, \mathscr{F}_n, \mathbb{P}_n) \to (X, \mathscr{B}(X))$ be random variables for $n = 1, 2, ..., \infty$. Then $\xi_n \to \xi_\infty$ in distribution as $n \to \infty$ if the push forward measures $\xi_n^* \mathbb{P}_n$ converges to weakly to $\xi_\infty^* \mathbb{P}_\infty$. If $\mathbb{E}_{\mathbb{P}}$ denotes the expectation with respect to $\mathbb{P}$, then by the change of variable formula (theorem 2.11), the above definition is equivalent to saying that for all $f \in C_c(X)$,

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}[f(\xi_n)] = \mathbb{E}_{\mathbb{P}}[f(\xi)]$$

**Remark 5.37** To establish the uniqueness of vague limits, we note that one can establish that if $\int f\,d\mu = \int f\,d\nu$ for all $f \in C_c(X)$, then $\mu = \nu$. The result and its proof are similar to the ones in lemma 5.2.

**Example 5.38 — Continuation of Example 5.30.** We recall the example of $\xi_n \sim U[n, n+1]$ in example 5.30. The distributions of $\xi_n$ then converge vaguely to the zero measure $\mu \equiv 0$.

**Exercise 5.39** Show that if $\xi_n \sim U[-n, n]$, then their distributions do not converge vaguely to any measure.

We note that the total mass cannot immigrate in the following sense:

**Proposition 5.40 — Mass cannot immigrate.** If $\mu, \mu_n$ are measures on $(X, \mathscr{B}(X))$ such that $\mu_n(x) \overset{n\to\infty}{\to} \mu$ vaguely, then

$$\mu(X) \leq \liminf_{n\to\infty} \mu_n(X) \tag{5.23}$$

*Proof.* Let $f_N \in C_c(X)$ such that $f_N \nearrow 1$ as $N \to \infty$, then

$$\mu(X) = \sup_N \int_X f_N\,d\mu = \sup_N \left(\lim_{n\to\infty} \int_X f_N\,d\mu_n\right) \leq \liminf_{n\to\infty} \left(\sup_N \int_X f_N\,d\mu_n\right) = \liminf_{n\to\infty} \mu_n(X).$$

■

Therefore, if $\mathbb{P}_n$ are probability measures on $(X, \mathscr{B}(X))$ such that $\mathbb{P}_n \overset{n\to\infty}{\to} \mu$ vaguely, then $\mu(X) \leq 1$. If in addition we know that $\mu(X) \geq \limsup_{n\to\infty} \mathbb{P}_n(X)$, such that $\mathbb{P}_n(X) \to \mu(X)$ and that $\mu$ is a probability measure, then $\mathbb{P}_n \to \mu$ weakly. More generally,

**Proposition 5.41** Let $\mu_n, \mu$ be measures on $(X, \mathscr{B}(X))$ with $\mu_n(X), \mu(X) \leq 1$. Then $\mu_n \overset{n\to\infty}{\to} \mu$ weakly $\iff$ $\mu_n \overset{n\to\infty}{\to} \mu$ vaguely and $\mu_n(X) \to \mu(X)$ (which can be reduced to $\mu(X) \geq \limsup_{n\to\infty} \mu_n(X)$ by the previous lemma).

*Proof.* The $\Rightarrow$ direction is trivial, so let's prove the $\Leftarrow$ direction. Assume the conditions hold, we would like to prove that statement (6) of the Portmanteau theorem holds. Let $O \subseteq X$ be an open set, and let $\epsilon > 0$. Recall that any finite measures on a Polish space are regular in the sense that

$$\mu(A) = \sup_{K \subseteq A, E \text{ compact}} \mu(K) \tag{5.24}$$

This means that there is a compact set $K \subseteq O$ such that $\mu(K) > \mu(O) - \epsilon$. Since $X$ is locally compact, it is Tynchonoff, and there is compact set $L$ with $K \subseteq L^\circ \subseteq L \subseteq O$. Let $\delta := \rho(K, L^c) = \inf_{x\in K} \rho(x, L^c) > 0$ and let $g_\delta(x) = (1 - \rho(x, E)/\delta) \vee 0$ as defined in equation (5.4). Then $g_\delta \in C_c(X)$ and $\chi_K \leq g_\delta \leq \chi_L$, and thus

$$\liminf_{n\to\infty} \mu_n(O) \geq \liminf_{n\to\infty} \int_X g_\delta \, d\mu_n = \int_X g_\delta \, d\mu \geq \mu(K) \geq \mu(O) - \epsilon.$$

Send $\epsilon \searrow 0$ to conclude. ■

**Exercise 5.42** Let $\mu, \mu_n$ be measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\mu_n(\mathbb{R}), \mu(\mathbb{R}) \leq 1$, and that $F(x) = \mu((-\infty, x])$ and $F_n(x) = \mu((-\infty, x])$. By modifying the proof of theorem 5.19, show that $\mu_n \overset{n\to\infty}{\mu}$ vaguely $\iff F_n(x) \overset{n\to\infty}{\to} F(x)$ for any $x \in \mathbb{R} \setminus U_F$.

**Remark 5.43** **Warning!** Knowing only that the sequence $(F_n(x))$ converges for almost all $x$ does not mean that $\mu_n$ converges vaguely to some measures!

Finally, one could show that the tightness of a vaguely convergent sequence of probability measures prevents the emigration of total mass, and therefore a tight vaguely convergent sequence of probability measures also converges weakly. As an immediate result of question 3 in exercise 5.35, we have

**Theorem 5.44** Let $(X, \mathscr{B}(X))$ be a locally compact Polish space with metric $d$, and let $\mu, \mu_n$ be a finite measure on $(X, \mathscr{B}(X))$ such that $\mu(X), \mu_n(X) \leq 1$. Then $\mu_n \overset{n\to\infty}{\to} \mu$ weakly iff $\mu_n \overset{n\to\infty}{\to} \mu$ vaguely and $(\mu_n)_{n\geq 1}$ being tight.

### 5.5.1 A functional analysis view of vague convergence

We note that $C_c(X)$ is a normed vector space when equipped with the supremum norm. This space is not complete since it is not closed (in fact its closure is $C_0(X)$, space of functions that vanishes at infinity), but we may still talk about bounded linear operators and dual space.

If $\mu$ is a probability measure on $X$, then of course the map $T_\mu := \int_X f(x)\, d\mu$ continues to be a bounded linear functional of $C_c(X)$ with dual norm 1. Moreover, $\mu_n \overset{n\to\infty}{\to} \mu$ vaguely if $T_{\mu_n} \overset{n\to\infty}{\to} T_\mu$ weakly* (when tested with $f \in C_c(X)$, of course). Moreover, $T_\mu$ is positive in the sense that $f \geq 0 \implies T_\mu f \geq 0$. In fact, there is a one-to-one correspondence between a positive functional and a *Radon* measure (which is regular and $\mu(K) < \infty$ whenever $K$ is compact).

**Theorem 5.45 — Riesz-Markov-Katutani.** Let $(X, \mathscr{B}(X))$ with metric $d$ be Polish and locally compact. Let $T$ be a bounded linear functional on $C_c(X)$ which is positive (i.e. $f \geq 0 \implies Tf \geq 0$). Then there is a *unique* Radon measure $\mu$ on $X$ such that for all $f \in C_c(X)$,

$$Tf = \int_X f(x)\, \mu(dx). \tag{5.25}$$

This provides a way to prove theorems about weak/vague convergence. For instance,

**Proposition 5.46** Let $(X, \mathscr{B}(X))$ be a locally compact Polish space, then the set of all measures $\{\mu \,|\, \mu(X) \leq 1\}$ is vaguely relatively compact.

This is a corollary from the Prokhorov theorem [a], but it is really nothing but the Banach-Alaoglu theorem, which is an important theorem in functional analysis:

**Theorem 5.47 — Banach-Alaoglu Theorem.** Let $E$ be a Banach space with norm $\|.\|_E$. If $\|.\|_{E'}$ is the dual norm, then the unit closed ball in $E'$, i.e. $(\{T \in E' \mid \|T\|_{E'} \leq 1\})$, is (sequentially) compact under the weak* topology.

[a] for proof via Prokhorov theorem see [1], corollary 13.31

That means that if $(\mu_n)_{n\geq 1}$ is such that $\mu_n(X) \leq 1$, then $\|T_{\mu_n}\|_{C_c(X)'} \leq 1$, and therefore there is a subsequence $\left(T_{\mu_{n_k}}\right)_{k\geq 1}$ which converges to $T \in C_c(X)'$. It is trivial to see that $T$ is positive, so by the Riesz-Markov-Katutani theorem (theorem 5.45), there is a Radon measure $\mu$ such that $T = T_\mu$. From the definition of vague convergence, we see that $\mu_{n_k} \overset{k\to\infty}{\to} \mu$.

We note that proposition 5.46 implies the Helly's selection theorem when $(X, \mathscr{B}(X)) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. In fact, the proposition guarantees that the subsequential limits are distribution functions of some measures $\mu$ with $\mu(X) \leq 1$.

# 6 Characteristic Functions

In this chapter, we look at characteristic functions of measures of $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, as well as random variables taking value on this measurable space. Let us begin by stating the definition.

**Definition 6.1 — Characteristic Functions.**

- The characteristic function of a measure $\mu$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is

$$\varphi_\xi(t) := \varphi(t) := \int_{-\infty}^{\infty} e^{itx}\, \mu(dx), \quad t \in \mathbb{R}. \tag{6.1}$$

- The characteristic function of a random variable $\xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is

$$\varphi_\xi(t) := \varphi(t) \equiv \mathbb{E}[e^{it\xi}] := \int_\Omega e^{it\xi(\omega)}\, \mathbb{P}(d\omega) = \int_{-\infty}^{\infty} e^{itx}\, [\xi^*\mathbb{P}](dx), \quad t \in \mathbb{R}. \tag{6.2}$$

We may generalise the above definitions to random variables taking value on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. For instance, the characteristic function of a random vector $\xi := (\xi_1, \dots, \xi_n)$ is

$$\varphi_\xi(t_1, \dots, t_n) := \mathbb{E}\left[ \exp\left( i \sum_{k=1}^{n} t_k \xi_k \right) \right].$$

Note that the characteristic function of a random variable only depends on its distribution. If $F(x)$ has density $f(x)$ (with respect to the Lebesgue measure), then

$$\varphi(t) = \int_{\mathbb{R}^n} e^{i(t^\top x)}\, f(x) \mathrm{d}x.$$

In other words, we may view the characteristic function of $\xi$ as the Fourier transform of $f(x)$. Note that in most of the other literatures in Fourier analysis (e.g. [6]) the actual Fourier transform is defined as $\phi(-t)$, but we may translate results from Fourier analysis to this chapter, making sure the the signs are taken consistently. In particular, we may prove a result similar to the $L^1$ inversion formula of Fourier transform, and we will make it precise in section 6.2.

Let us here prove some fundamental properties for characteristic functions.

**Property 6.2**

1. If $\xi$ is a random variable, $a, b$ are constants and $\eta = a\xi + b$, then $\varphi_\eta(t) = e^{itb}\mathbb{E}[e^{iat\xi}]$.
2. $|\varphi(t)| \le \varphi(0) = 1$.
3. Let $\xi$ be a random variable. Then $\varphi_\xi(t)$ is uniformly continuous on $\mathbb{R}$
4. If $\xi_1, \xi_2, \dots, \xi_n$ are independent random variables and $S = \xi_1 + \cdots + \xi_n$, then

$$\varphi_S(t) = \prod_{j=1}^{n} \varphi_{\xi_j}(t).$$

*Proof.* Statements (1) and (2) are trivial. For statement (3), we note that

$$|\varphi(t+h) - \varphi(t)| = \left|\mathbb{E}[e^{it\xi}(e^{ih\xi} - 1)]\right| \le \mathbb{E}[|e^{ih\xi} - 1|],$$

So by dominated convergence theorem, $\mathbb{E}[|e^{ih\xi} - 1|] \to 0$ as $h \to 0$. ∎

Recall the moment generating function (MGF) as defined in section 2.5. We note that MGF also shares properties (1) and (4), but the lack of properties (2) and (3) means that it is more preferrable to use characteristic functions to establish weak convergence.

**Example 6.3 — Examples of characteristic functions.**

1. For a Bernoulli random variable with parameter $p$, we have $\varphi_\xi(t) = pe^{it} + q$, where $q := 1 - p$. Therefore by property (4), if $\xi \sim \mathsf{B}(n, p)$, then $\varphi_\xi(t) = (pe^{it} + q)^n$.
2. Let $\xi \sim \mathsf{N}(m, \sigma^2)$. Then
$$\varphi_\xi(t) = \exp\left(itm - \frac{t^2\sigma^2}{2}\right)$$
3. Let $\xi \sim \mathsf{Po}(\lambda)$. Then
$$\varphi_\xi(t) = e^{-\lambda}\sum_{k=0}^{\infty} e^{itk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda+\lambda e^{it}}.$$

*Proof.* We only show the second example. Let $\eta = (\xi - m)/\sigma$. Then $\eta \sim \mathsf{N}(0, 1)$, i.e. with density

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

It is sufficient to show that $\varphi_\eta(t) = e^{-\frac{t^2}{2}}$. We have

$$\begin{aligned}
\varphi_\eta(t) = \mathbb{E}[e^{it\eta}] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx-\frac{x^2}{2}}\mathrm{d}x \\
&= e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-1/2(x-it)^2}\mathrm{d}x \\
&= e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{\infty-it} e^{-\frac{z^2}{2}}\mathrm{d}z \\
&= e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\mathrm{d}x \\
&= e^{-\frac{t^2}{2}}.
\end{aligned}$$

■

## 6.1 Obtaining moments

It turns out that the existence of moments for a real-valued random variable is determined by the smoothness of its characteristic function at zero. The property is characterised by the following proposition:

**Proposition 6.4 — Moments.** Let $\xi$ be a random variable with characteristic function $\varphi$ and distribution function $F$. Then

- if $\mathbb{E}[|\xi|^n] < \infty$ for some $n \geq 1$ then $\varphi^{(\tau)}(t)$ exists for any $0 \leq \tau \leq n$ and
$$\begin{aligned}
\varphi^{(\tau)}(t) &= \int_{\mathbb{R}} (ix)^\tau e^{itx}\mathrm{d}F(x) \\
\mathbb{E}[\xi^\tau] &= \frac{\varphi^{(\tau)}(0)}{i^\tau} \\
\varphi(t) &= \sum_{\tau=0}^{n}\frac{(it)^\tau}{\tau!}\mathbb{E}[\xi^\tau] + \frac{(it)^n}{n!}\varepsilon_n(t),
\end{aligned}$$
where $|\varepsilon_n(t)| \leq 3\mathbb{E}[|\xi|^n]$ and $\varepsilon_n(t) \to 0, t \to 0$.
- if $\mathbb{E}[|\xi|^n] < \infty$ for all $n \geq 1$ and
$$\limsup_{n\to\infty}\frac{(\mathbb{E}[|\xi|^n])^{1/n}}{n} = \frac{1}{e \cdot R} < \infty, \tag{6.3}$$

then

$$\varphi(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \mathbb{E}[\xi^n]$$

converges for all $|t| < R$.

*Proof.*

- If $\mathbb{E}[|\xi|^n] < \infty$, we have $\mathbb{E}[|\xi|^r] < \infty$ for any $r \leq n$ by Lyapunov's inequality. Consider the difference quotient

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \mathbb{E}\left[e^{it\xi}\left(\frac{e^{ih\xi} - 1}{h}\right)\right].$$

Since

$$\left|\frac{e^{ihx} - 1}{h}\right| \leq |x|,$$

and $\mathbb{E}[|\xi|] < \infty$, it follows from the dominated convergence theorem that the limit

$$\lim_{h \to 0} \mathbb{E}\left[e^{it\xi}\left(\frac{e^{ih\xi} - 1}{h}\right)\right]$$

exists and equals

$$\mathbb{E}\left[e^{it\xi} \lim_{h \to 0}\left(\frac{e^{ih\xi} - 1}{h}\right)\right] = i\mathbb{E}[\xi e^{it\xi}] = i\int_{-\infty}^{\infty} x e^{itx} \mathrm{d}F(x).$$

Hence $\varphi'(t)$ exits

$$\varphi'(t) = i\mathbb{E}[\xi e^{it\xi}] = i\int_{-\infty}^{\infty} x e^{itx} \mathrm{d}F(x).$$

The existence of the derivatives $\varphi^{(r)}(t), 1 < r \leq n$ follows by induction. The second formula follows immediately from the first. To establish the third formula, consider

$$e^{iy} = \cos y + i \sin y = \sum_{k=0}^{n-1} \frac{(iy)^k}{k!} + \frac{(iy)^n}{n!}(\cos\theta_1 y + i \sin\theta_2 y)$$

for real $y$ with $|\theta_1| \leq 1$, $|\theta_2| \leq 1$. We have

$$e^{it\xi} = \sum_{k=0}^{n-1} \frac{(i\xi)^k}{k!} + \frac{(i\xi)^n}{n!}(\cos\theta_1 \xi + i \sin\theta_2 \xi)$$

and

$$\mathbb{E}[e^{it\xi}] = \sum_{k=0}^{n-1} \frac{(it)^k}{k!} \mathbb{E}[\xi^k] + \frac{(it)^n}{n!}(\mathbb{E}[\xi^n] + \varepsilon_n(t)),$$

where

$$\varepsilon_n(t) = \mathbb{E}[\xi^n(\cos\theta_1(\omega)t\xi + i \sin\theta_2(\omega)t\xi - 1)].$$

It is clear that $|\varepsilon_n(t)| \leq 3\mathbb{E}[|\xi^n|]$. The theorem on dominated convergence shows that $\varepsilon_n(t) \to 0, t \to 0$.

- Let $0 \leq t_0 \leq T$. Then, by Stirling's formula we find that

$$\limsup \frac{(\mathbb{E}[|\xi|^n])^{1/n}}{n} \leq \frac{1}{t_0} \implies \limsup \frac{(\mathbb{E}[|\xi|^n t_0^n])^{1/n}}{n} \leq 1 \implies \limsup \left(\frac{\mathbb{E}[|\xi|^n t_0^n]}{n!}\right)^{1/n} < 1.$$

Consequently, the series $\sum \mathbb{E}[|\xi|^n t_0^n / n!]$ converges by Cauchy's test and therefore the series $\sum_{r=0}^{\infty}[(it)^r/r!]\mathbb{E}[\xi^r]$ converges for $|t| \leq t_0$. But by the previous statement for $n \geq 1$

$$\varphi(t) = \sum_{r=0}^{n} \frac{(it)^r}{r!} \mathbb{E}[\xi^r] + R_n(t),$$

where $|R_n(t)| \leq 3(|t|^n/n!)\mathbb{E}[|\xi|^n]$. Therefore

$$\varphi(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r!}\mathbb{E}[\xi^r]$$

for all $|t| < T$.

■

**Remark 6.5** Second part of this theorem gives a sufficient condition for the moments $\mathbb{E}[\xi^n]$ to determine $\varphi(t)$ uniquely. Indeed, under the condition (1), they already determine $\varphi(t)$ for $-R < t < R$. Take $s$ such that $|s| < R/2$. Follow the proof to obtain that

$$\varphi(t) = \sum_{k=0}^{\infty} i^k \frac{(t-s)^k}{k!}\varphi^{(k)}(s),$$

where

$$\varphi^{(k)}(s) = \mathbb{E}[\xi^k e^{is\xi}], \quad -R/2 < s < R/2,$$

is uniquely determined by $\mathbb{E}[\xi^n], n \geq 1$. Therefore, the moments uniquely determine $\varphi(t)$ for $|t| < \frac{3}{2}R$.

**Theorem 6.6 — Carleman's test.** A sufficient condition for unique determination of the characteristic function $\varphi(t)$ is that

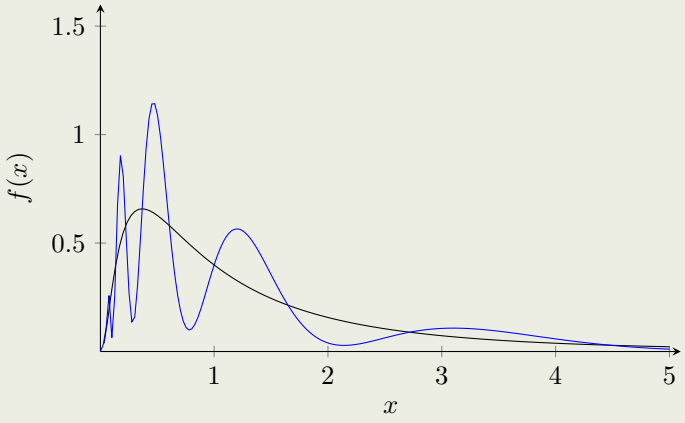
$$\sum_{n=0}^{\infty} \frac{1}{(\mathbb{E}[\xi^{2n}])^{1/2n}} = \infty.$$

**Example 6.7** If $\mathbb{E}[\xi^n]$ grows too fast, there may be multiple characteristic functions $\varphi(t)$ with these moments. As an example, consider a random variable distributed as standard log-normal distribution, which has density.

$$f_0(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\ln x)^2}{2}\right), \quad x \geq 0, \tag{6.4}$$

and another random variable with density

$$f_a(x) = f_0(x) \times (1 + a\sin(2\pi \ln x)), \quad x \geq 0, a \in [-1, 1]. \tag{6.5}$$




We note that these two seemingly different random variables have the same $r$-th moment. To see this, it suffices to evaluate the integral

$$\int_0^{\infty} x^{r-1} \exp\left(-\frac{(\ln x)^2}{2}\right) \sin(2\pi \ln x)\, dx, \quad r = 0, 1, ...$$

With a variable substitution of $s = \ln x$ (such that $ds = dx/x$), the integral equals to

$$\int_{-\infty}^{\infty} \exp\left((r-1)s\right) \exp\left(-\frac{s^2}{2}\right) \sin(2\pi s)\, dx$$

Notice the integrand is an $L^1$ function multiplied by $\sin(2\pi s)$. Therefore, by the Riemann-Lebesgue lemma, the integral is zero, and the two random variables have the same moments!

**Exercise 6.8 — Computing the moments of log-normal distribution.** What happened? We can compute at the $r$-th moments of the log-normal distribution:

1. Verify that if $\xi$ has a standard normal distribution ($\mathsf{N}(0,1)$), then $\exp(\xi)$ has a standard log-normal distribution.
2. Use LOTUS to compute the $r$-th moment being equal to $\exp(r^2/2)$.

Notice that the moments grows too fast for the characteristic function to be analytical (i.e. possess a Taylor series)!

## 6.2 Inversion Formula

The main objective of this section is to make the inversion formula for characteristic functions precise. Let us state the first part of inversion formula.

**Theorem 6.9 — Inversion formula I.** Let $\mu$ be a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with its corresponding distribution $F(x) = \mu((-\infty, x])$ and characteristic function $\varphi(t) = \int_{\mathbb{R}} e^{itx}\, \mu(dx)$. If $a < b$ then

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt = \mu((a,b)) + \frac{\mu(\{a\}) + \mu(\{b\})}{2} \tag{6.6}$$

In particular when $\mu(\{a\}) = \mu(\{b\}) = 0$, so that $F$ is continuous at $a, b$, then

$$F(b) - F(a) = \mu((a,b]) = \mu((a,b)) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt; \tag{6.7}$$

Before we begin proving this theorem, let us recall some facts about the integral of the sinc function. Define

$$S(T) = \int_0^T \frac{\sin x}{x}\, dx. \tag{6.8}$$

We note that $S(T)$ is a differentiable function with $S(T) > 0$ whenever $T > 0$, and that

**Lemma 6.10**

$$S(+\infty) = \int_0^{\infty} \frac{\sin x}{x} = \frac{\pi}{2} \tag{6.9}$$

This can be proven by standard calculus tricks, either by differentiation under integral or residue calculus. [4] We therefore know that $S(T)$ is a bounded function and $\sup_{T>0} S(T)$ exists.

We also note the following scaling formula

$$\int_0^T \frac{\sin(kx)}{x}\, dx = \int_0^T \frac{\sin(kx)}{kx}\, d(kx) = S(kT), \quad k > 0, \tag{6.10}$$

and that when $k < 0$, we have

$$\int_0^T \frac{\sin(kx)}{x}\, dx = -\int_0^T \frac{\sin(|k|x)}{x}\, dx = -S(|k|T). \tag{6.11}$$

[4] There is an entire article of teaching you how to integrate a since function! See https://www.wikihow.com/Integrate-the-Sinc-Function.

In terms of the sgn function ($\operatorname{sgn} = \chi_{(0,\infty)} - \chi_{(-\infty,0)}$), we have for all $k \in \mathbb{R}$,

$$\int_0^T \frac{\sin(kx)}{x} = \operatorname{sgn}(k) S(|k|T). \tag{6.12}$$

Finally, provided that the sinc function is even, we know that

$$\int_{-T}^T \frac{\sin(kx)}{x} = 2\operatorname{sgn}(k) S(|k|T). \tag{6.13}$$

We are now ready to prove the first inversion formula:

*Proof.* (of inversion formula I, theorem 6.9) So let's first evaluate for fixed $T$,

$$I_T = \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt = \int_{-T}^T \int_{\mathbb{R}} \frac{e^{-ita} - e^{-itb}}{it} e^{itx}\, \mu(dx)\, dt \tag{6.14}$$

We note that the integrand is bounded uniformly in $(t, x)$ (except when $t = 0$, which is a removable singularity):

$$\left| \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \right| \leq \left| \int_a^b e^{-its}\, ds \right| \leq |b - a|, \tag{6.15}$$

so the integrand is integrable. By Fubini's theorem, we may exchange the order of integration:

$$I_T = \int_{\mathbb{R}} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} e^{itx}\, dt\, \mu(dx) = \int_{\mathbb{R}} \int_{-T}^T \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt\, \mu(dx) \tag{6.16}$$

We focus on the inner integral. Since the domain is now symmetric, we may ignore the odd parts of the integral, so we have

$$I_T = \int_{\mathbb{R}} \int_{-T}^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t}\, dt\, \mu(dx) \tag{6.17}$$

We have established a few facts about integrals on since function, and we want to utilise them here. To do so, we have to justify the following exchange the order of limits when evaluating $I_\infty$:

$$I_\infty = \lim_{T\to\infty} I(T) \overset{(?)}{=} \int_{\mathbb{R}} \lim_{T\to\infty} \underbrace{\int_{-T}^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t}}_{J_{T,x}}\, dt\, \mu(dx)$$

and here we will use dominated convergence theorem (DCT). Notice we have

$$|J_{T,x}| \leq \left| \int_{-T}^T \frac{\sin(t(x-a))}{t} \right| + \left| \int_{-T}^T \frac{\sin(t(x-b))}{t} \right| \leq 4 \sup_{T>0} S(T) < \infty,$$

which is integrable with respect to the probability measure $\mu$. Therefore the condition for DCT is satisfied.

Let us now tabulate the values of $J_{\infty,x}$ for different values of $x$:

| $x$ | $\int_{-\infty}^{\infty} \frac{\sin(t(x-a))}{t}\, dt$ | $\int_{-\infty}^{\infty} \frac{\sin(t(x-b))}{t}\, dt$ | $J_{\infty,x}$ |
|---|---|---|---|
| $x > b$ | $\pi$ | $\pi$ | 0 |
| $x = b$ | $\pi$ | 0 | $\pi$ |
| $a < x < b$ | $\pi$ | $-\pi$ | $2\pi$ |
| $x = a$ | 0 | $-\pi$ | $\pi$ |
| $x < a$ | $-\pi$ | $-\pi$ | 0 |

and therefore we have

$$I_\infty = \int_{\{a\}} \pi d\mu + \int_{\{b\}} \pi d\mu + \int_{(a,b)} 2\pi d\mu = 2\pi \left( \frac{\mu(\{a\}) + \mu(\{b\})}{2} + \mu((a,b)) \right),$$

which completes the proof. ∎

Let us also state another inversion formula for obtaining the measures of an atom.

**Theorem 6.11 — Inversion formula II.** Under the setting of the first inversion formula (theorem 6.9), we have

$$\mu(\{a\}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita}\varphi(t)\,dt \tag{6.18}$$

*Proof.* The proof is as similar as above, so we will only give a sketch. We first let

$$\begin{aligned} I_T := \frac{1}{2T}\int_{-T}^{T} e^{-ita}\varphi(t)\,dt &= \frac{1}{2T}\int_{-T}^{T}\int_{\mathbb{R}} e^{it(x-a)}\,d\mu\,dt \\ &\overset{\text{(Fubini)}}{=} \int_{\mathbb{R}} \frac{1}{2T}\int_{-T}^{T} e^{it(x-a)}\,dt\,d\mu \\ &= \int_{\mathbb{R}} \frac{e^{iT(x-a)} - e^{-iT(x-a)}}{2iT(x-a)}\,d\mu \quad \text{see } [5] \\ &= \int_{\mathbb{R}} \frac{\sin(T(x-a))}{T(x-a)}\,d\mu. \end{aligned}$$

We note that the integrand is uniformly bounded by one, which is integrable, so by DCT we have

$$I_\infty = \int_{\mathbb{R}} \left( \lim_{T\to\infty} \frac{\sin(T(x-a))}{T(x-a)}\,d\mu \right). \tag{6.19}$$

But note that the integrand tends to one when $x = a$ as $T \to \infty$ and zero otherwise, so we have

$$I_\infty = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} e^{-ita}\varphi(t)\,dt = \mu(\{a\}). \tag{6.20}$$

■

**Corollary 6.12** Probability distributions on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and characteristic functions are in one-to-one correspondence.

*Proof.* So let's say $\mu, \nu$ are two measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that they have the same characteristic function. By the inversion formulas we see that $\mu((a,b)) = \nu((a,b))$ for any $a < b$. Since the collection of open intervals $\{(a,b) \mid a < b\}$ generates $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, we must have $\mu = \nu$. ■

The final inversion formula concerns the absolute continuity of a probability measure $\mu$ with respect to the Lebesgue measure (i.e. whether a continuous density exists). It turns out that if its characteristic function $\varphi$ is integrable, then we may invert the Fourier transform in the following sense

**Proposition 6.13 — Inversion formula III.** If $\int_{-\infty}^{\infty} |\varphi(t)|\mathrm{d}t < \infty$, then the probability measure $\mu(x)$ has density $f(x)$, in the sense that the distribution function $F$ satisfies

$$F(x) = \mu((-\infty, x]) = \int_{-\infty}^{x} f(y)\,dy, \tag{6.21}$$

and that the density is given by

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt. \tag{6.22}$$

We note that this inversion formula coincides with the $L^1$ inversion formula for Fourier transform: let $f \in L^1(\mathbb{R})$ and its Fourier transform (which is $\varphi(-t)$ with $\varphi$ being the characteristic function of the measure as defined in (6.21)) is in $L^1$, then (6.22) holds with the minus sign replaced by a plus sign.

*Proof.* Utilise inversion formula I (theorem 6.9) and the estimate (6.15), we have

$$\mu((a,b)) + \frac{\mu(\{a\}) + \mu(\{b\})}{2} = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t) \le \frac{b-a}{2}\int_{\mathbb{R}} |\varphi(t)|\,dt \lesssim \mathsf{Leb}((a,b)). \tag{6.23}$$

This proves that $\mu \ll \text{Leb}$, so that $\mu$ possess a density and has no atoms. We now prove that the function given in (6.22) is a density of $\mu$: note that

$$\mu((x, x+h)) = \frac{1}{2\pi} \int_{\mathbb{R}} \left( \int_x^{x+h} e^{-ity}\, dy \right) \phi(t)\, dt \overset{\text{(Fubini)}}{=} \int_x^{x+h} \left( \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ity} \phi(t)\, dt \right) dy \tag{6.24}$$

So $\mu$ has density function $f$ as given in (6.22). It is easy to show that $f$ is continuous (by, e.g. dominated convergence theorem). ■

**Exercise 6.14** Show that the function $f$ as defined above is continuous by dominated convergence theorem.

We may use the third inversion formula to find the characteristic function for even more distributions.

**Exercise 6.15** Let $\xi, \eta$ be a real-valued random variable.

1. Find the characteristic function of $\xi$ if $\xi$ follows a triangular distribution with density

$$f(t) = (1 - |t|)_+ := (1 - |t|) \vee 0, \tag{6.25}$$

then use the third inversion formula to show that the characteristic function of $\eta$ is $\varphi_\eta(t) = (1 - |t|)_+$, where the $\eta$ follows a Polya's distribution [a] with density

$$f(t) = \frac{1 - \cos t}{\pi t^2}. \tag{6.26}$$

We will need this distribution to prove a theorem for the characterisation of characteristic function in section 6.4.

2. Find the characteristic function of $\xi$ if $\xi$ follows a Laplace (bilateral exponential) distribution with density

$$f(t) = \exp(-|t|)/2, \tag{6.27}$$

and hence find the characteristic function of $\eta$ if $\eta$ follows a Cauchy distribution with density

$$f(t) = (\pi(1 + t^2))^{-1}. \tag{6.28}$$

[a] We follow the naming convention in [7].

## 6.3 Central Limit Theorem via Characteristic Functions

We are almost able to prove the Central Limit Theorem. We need to prove the final (and the most important) step, the Levi's continuity theorem, which says if

**Theorem 6.16 — Levi's Continuity theorem.** Let $\varphi_n(t)$ be the characteristic functions of a sequence of probability measures $\mathbb{P}_n$ with on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ distribution functions $F_n$. Then

1. If $\mathbb{P}_n \to \mathbb{P}_\infty$ weakly, where $\mathbb{P}_\infty$ is a measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then $\varphi_n(t) \to \varphi_\infty(t)$ **pointwise** for all $t \in \mathbb{R}$, where $\varphi_\infty$ is the characteristic function of $\mu_\infty$.
2. If $\lim_{n\to\infty} \varphi_n(t)$ exists $\forall t \in \mathbb{R}$, and $\varphi_\infty(t) = \lim_{n\to\infty} \varphi_n(t)$ is **continuous** at $t = 0$, then $\varphi_\infty(t)$ is a characteristic function of some probability measure $\mathbb{P}$ and $\mathbb{P}_n \to \mathbb{P}$ (weakly).
3. If $\varphi_n(t)$ corresponds to $\mathbb{P}_n$ and $\varphi_\infty(t)$ is a characteristic function corresponding to $\mathbb{P}_\infty$, then

$$\varphi_n(t) \to \varphi_\infty(t)\ \forall t \in R \iff \mathbb{P}_n \to \mathbb{P}_\infty \text{ (weakly)}.$$

We note that statement (1) is a direct consequence of the definition of weak convergence when applied to $\Re[e^{it\xi}]$, $\Im[e^{it\xi}]$.

To prove statements 2 and 3, we need the following estimates:

**Lemma 6.17** If $\mathbb{P}$ is a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with characteristic function $\varphi(t)$, then for all $\epsilon > 0$,

$$\mathbb{P}(\{x \,|\, |x| \geq 2/\epsilon\}) \leq \frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} (1 - \varphi(t))\, dt \tag{6.29}$$

This lemma shows that the tail of measure $\mathbb{P}$, hence the existence of moments, is determined by the smoothness of $\varphi$ at zero. The direct connections between the existence of moments and smoothness are established in section 6.1.

*Proof.* First note that for all $x \neq 0$,

$$\int_{-\epsilon}^{\epsilon} (1 - e^{itx})\, dt = 2\epsilon - \frac{e^{it\epsilon} - e^{-it\epsilon}}{ix} = 2u \left(1 - \frac{\sin \epsilon x}{\epsilon x}\right) \tag{6.30}$$

We therefore have

$$\begin{aligned}
\frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} (1 - \varphi(t))\, dt &= \frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} \left(1 - \int_{\mathbb{R}} e^{itx}\, \mu(dx)\right) dt \\
&= \frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} \int_{\mathbb{R}} (1 - e^{itx})\, \mu(dx)\, dt \\
&\overset{\text{(Fubini)}}{=} \int_{\mathbb{R}} \left(\frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} (1 - e^{itx})\, dt\right) \mathbb{P}(dx) \\
&= \int_{\mathbb{R}} \underbrace{2\left(1 - \frac{\sin \epsilon x}{\epsilon x}\right)}_{\geq 0} \mathbb{P}(dx) \\
&\geq 2 \int_{-2/\epsilon}^{2/\epsilon} \left(1 - \frac{\sin \epsilon x}{\epsilon x}\right) \mathbb{P}(dx) \\
&\geq 2 \int_{-2/\epsilon}^{2/\epsilon} \underbrace{\left(1 - \frac{1}{|\epsilon x|}\right)}_{\geq 1/2} \mathbb{P}(dx) \\
&\geq \mathbb{P}(\{x \,|\, |x| \geq 2/\epsilon\})
\end{aligned}$$

■

We are now ready to prove statements 2 and 3 of the continuity theorem.

*Proof.* We really only need to prove the second statement. When this is proven, the third statement follows from this and the inversion theorem.

**We first prove that the sequence $(\mathbb{P}_n)$ is tight.** We are given that $\varphi_\infty$ is continuous at 0 and that $\varphi_\infty(0) = 1$, so for all $\epsilon > 0$, there is $u > 0$ small enough that for all $t \in [-u, u]$, $1 - \varphi_\infty(t) \leq \epsilon/4$, and hence

$$\frac{\epsilon}{2} \geq \frac{1}{u} \int_{-u}^{u} (1 - \varphi_\infty(t))\, dt \overset{\text{(DCT)}}{=} \lim_{n \to \infty} \frac{1}{u} \int_{-u}^{u} (1 - \varphi_n(t))\, dt. \tag{6.31}$$

As a result, there is $n_0$ such that for all $n \geq n_0$ such that

$$\mathbb{P}_n\left(\mathbb{R} \setminus [-2/u, 2/u]\right) \leq \frac{1}{u} \int_{-u}^{u} (1 - \varphi_n(t))\, dt \leq \epsilon \tag{6.32}$$

We may choose smaller $u$ such that the above inequality holds for all $n \geq 1$, and hence we see that $(\mathbb{P}_n)_{n \geq 1}$ is tight. By Prokhorov theorem, for any subsequence of $(\mu_n)_{n \geq 1}$, say $(\mu_{n_k})_{k \geq 1}$, there is a further subsequence that converges weakly to the measure $\nu$. We note that by statement (1) of continuity theorem and uniqueness of limits we know that $\varphi_\infty(t)$ is the characteristic function of $\nu$, which also shows

that the limiting measure must be unique regardless of what subsequence we have chosen.

We finally note that $\mathbb{P}_n \to \nu$ weakly, otherwise there is a point $y \in \mathbb{R} \setminus U_F$, $F$ distribution function of $\nu$, so that there is a subsequence $(\mu_{n'_k})_{k\geq 1}$ such that $|F_{n'_k}(y) - F(y)| \geq \epsilon$ for all $k$, but by above arguments there is a further subsequence $(\mu_{n_{k_j}})_{j\geq 1}$ which converges to $\nu$, which is a contradiction! ■

Once we have the Levi's continuity theorem, we also have the following central limit theorems for free.

**Theorem 6.18 — Central Limit Theorem for Independent Identically Distributed Random Variables.** Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed (nondegenerate) random variables with $\mathbb{E}[\xi_1^2]] < \infty$ and $S_n = \xi_1 + \cdots + \xi_n$. Then as $n \to \infty$

$$\mathbb{P}\left\{\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}} \leq x\right\} \longrightarrow \Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \mathrm{d}u \quad \forall x \in \mathbb{R}.$$

**Remark 6.19** The result can also be written as

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}} \xrightarrow{d} N(0,1).$$

*Proof.* Set $m = \mathbb{E}[\xi_1]$, $\sigma^2 = \mathbb{V}[\xi_1]$, $\varphi(t) = \mathbb{E}[e^{it(\xi_1 - m)}]$. If we put

$$\varphi_n(t) = \mathbb{E}\left[\exp\left(it\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}}\right)\right],$$

by independence

$$\varphi_n(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Since $\mathbb{E}[\xi_1^2] < \infty$, we have by properties of characteristic functions (theorem above) that

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2), \quad t \to 0.$$

So $\varphi_n(t) = [1 - t^2/(2n) + o(1/n)]^n \to e^{-t^2/2}$ for all $t \in \mathbb{R}$. This is the characteristic function of $\mathsf{N}(0,1)$ and so the result follows by continuity theorem. ■

**Theorem 6.20 — Central Limit Theorem for Independent Random Variables.** Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables with finite second moments $\mathbb{E}[\xi_j] < \infty$ and distribution functions $F_j$ for all $j$. Let

$$m_j = \mathbb{E}[\xi_j], \sigma_j^2 = \mathbb{V}[\xi_j] > 0, \quad S_n = \xi_1 + \cdots + \xi_n \text{ and } D_n^2 = \sum_{j=1}^{n} \sigma_j^2.$$

Suppose that the Lindeberg condition is satisfied: for every $\varepsilon > 0$

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x : |x - m_k| \geq \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) \to 0, \quad n \to \infty.$$

Then

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}} \xrightarrow{d} N(0,1).$$

We turn our attention to some special cases in which the Lindeberg condition is satisfied and consequently, the central limit theorem is valid.

- Let the *Lyapunov condition* be satisfied: for some $\delta > 0$

$$\frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}[|\xi_k - m_k|^{2+\delta}] \to 0 \quad n \to \infty.$$

Let $\varepsilon > 0$. Then

$$\begin{aligned}\mathbb{E}[|\xi_k - m|^{2+\delta}] &= \int_{-\infty}^{\infty} |x - m_k|^{2+\delta} \mathrm{d}F_k(x) \\ &\geq \int_{\{x:x|x-m_k|\geq \varepsilon D_n\}} |x - m_k|^{2+\delta} \mathrm{d}F_k(x) \\ &\geq \varepsilon^{\delta} D_n^{\delta} \int_{\{x:|x-m_k|\geq \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x),\end{aligned}$$

and therefore

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x:|x-m_k|\geq \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) \leq \frac{1}{\varepsilon^{\delta}} \frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}[|\xi_k - m_k|^{2+\delta}].$$

Consequently, the Lyapunov condition implies the Lindeberg condition.

- Suppose that there exists $K$ such that for all $n \geq 1$

$$|\xi_k| \leq K < \infty, \quad \forall k,$$

where $D_n \to \infty$ as $n \to \infty$. This condition also implies Lindeberg condition (exercise).

**Theorem 6.21 — Berry-Esseen Inequality.** Let $\xi_1, \xi_2, \ldots$ be a sequence of independent and identically distributed random variables with $\mathbb{E}[|\xi_1|^3] < \infty$. Then

$$\sup_x \left| \mathbb{P}\left( \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}[S_n]}} \leq x \right) - \Phi(x) \right| \leq C \frac{\mathbb{E}[|\xi_1 - \mathbb{E}[\xi_1]|^3]}{\sigma^3 \sqrt{n}},$$

where the absolute constant $C$ satisfies the inequality

$$\frac{1}{\sqrt{2\pi}} \leq C < 0.8.$$

**Remark 6.22** $O(1/\sqrt{n})$ is optimal. Indeed, let $\xi_1, \xi_2, \ldots$ be independent and identically distributed Bernoulli random variables,

$$\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = \frac{1}{2}.$$

Then by symmetry,

$$2\mathbb{P}\left( \sum_{k=1}^{2n} \xi_k < 0 \right) + \mathbb{P}\left( \sum_{k=1}^{2n} \xi_k = 0 \right) = 1$$

and therefore,

$$\left| \mathbb{P}\left( \sum_{k=1}^{2n} \xi_k < 0 \right) - \frac{1}{2} \right| = \frac{1}{2} \mathbb{P}\left( \sum_{k=1}^{2n} \xi_k = 0 \right) = \frac{1}{2} \binom{2n}{n} \frac{1}{2^{2n}} \sim \frac{1}{2\sqrt{\pi n}} = \frac{1}{\sqrt{2\pi}\sqrt{2n}}.$$

Then $\mathbb{E}[|\xi_1|^3] = 1 = \sigma_1$, so the theorem cannot be improved in terms of $O(1/\sqrt{n})$ and $C \geq \frac{1}{\sqrt{2\pi}}$.

**Example 6.23 — Cauchy Distribution.** What happens if $\mathbb{E}[\xi^2] = \infty$? Let $\xi_1, \xi_2, \ldots$ be i.i.d. with Cauchy distribution, i.e. with density

$$f = \frac{\theta}{\pi(x^2 + \theta^2)}, \quad \theta > 0.$$

Then

$$\varphi_{\xi_1}(t) = \frac{\theta}{\pi} \int_{-\infty}^{\infty} \frac{e^{itx}}{x^2 + \theta^2} \mathrm{d}x = e^{-t\theta}, \ t > 0$$

and similarly for $t < 0$. So

$$\varphi_{\xi_1}(t) = e^{-\theta|t|}, t \in \mathbb{R},$$

which implies that

$$\varphi_{S_n/n}(t) = (e^{-\theta|t|/n})^n = e^{-\theta|t|},$$

and thus $S_n/n$ also has the Cauchy distribution!

**Exercise 6.24** Verify the characteristic function for the above (standard) Cauchy distribution.

## 6.4 More about constructing characteristic function

The following theorems determine whether a function $\varphi$ is a characteristic function of some measure on $\mathbb{R}$, and if so, whether we can easily construct the underlying measure. The constructions are usually difficult, and therefore are not usually covered in great detail. Nevertheless, the proofs in this section serve as great examples of using tools developed in the previous chapter. For further discussions please refer to [8].

### 6.4.1 Bochner-Khinchin Theorem

**Theorem 6.25 — Bochner-Khinchin.** Let $\varphi(t)$ be continuous, $t \in \mathbb{R}$, with $\varphi(0) = 1$. A necessary and sufficient condition that $\varphi(t)$ is a characteristic function is that it is positive semi-definite, i.e. that for all real $t_1, \ldots, t_n$ and all complex $\lambda_1, \ldots, \lambda_n$, $n = 1, 2, \ldots$,

$$\sum_{j,k=1}^{n} \varphi(t_j - t_k)\lambda_j \overline{\lambda}_k \geq 0.$$

To show necessity, we note that if $\varphi$ is a characteristic function of a real-valued random variable $\xi$, then

$$\sum_{j,k=1}^{n} \varphi(t_j - t_k)\lambda_j \overline{\lambda}_k \geq 0 = \mathbb{E}\left[\eta\bar{\eta}\right] = \mathbb{E}[|\eta|^2] \geq 0, \quad \eta = \sum_{j=1}^{n} \lambda_j e^{t_j \xi}.$$

The fact that one can construct a probability measure with a positive semi-definite characteristic function is a deep fact in Fourier analysis, and the proof for general cases is beyond our scope. The case for $\varphi$ being integrable is covered in the third inversion formula.

### 6.4.2 Polya's Criterion

**Theorem 6.26 — Polya's criterion.** Let a continuous even real-valued function $\varphi(t)$ satisfy $\varphi(t) \geq 0$, $\varphi(0) = 1$, $\varphi(t) \to 0$ as $t \to \infty$ and let $\varphi(t)$ be convex on $0 \leq t < \infty$ (hence also on $(0, \infty)$). Then $\varphi(t)$ is a characteristic function.

As an observation, we note that the function $\varphi(t)$ must be strictly decreasing over $[0, \infty)$ (see Lemma 2.5 of [9]). To see this, we let $0 < r < s$, then there is $t_0 > s$ such that $0 < f(t_0) < f(r)/2$. By convexity, we have

$$f(s) \leq \frac{f(t_0) - f(r)}{t_0 - r}(s - r) + f(r) < f(r).$$

The Polya's criterion relies on the following simple observation of characteristic function.

**Exercise 6.27 — Convex combination of characteristic function.** Let $\varphi_k(t)$, $k = 1, 2, \ldots$ be characteristic functions and let the nonnegative numbers $\lambda_k$ satisfy $\sum_{k=1}^{n} \lambda_k = 1$. Show that for all $n \in \mathbb{Z}_{\geq 1}$

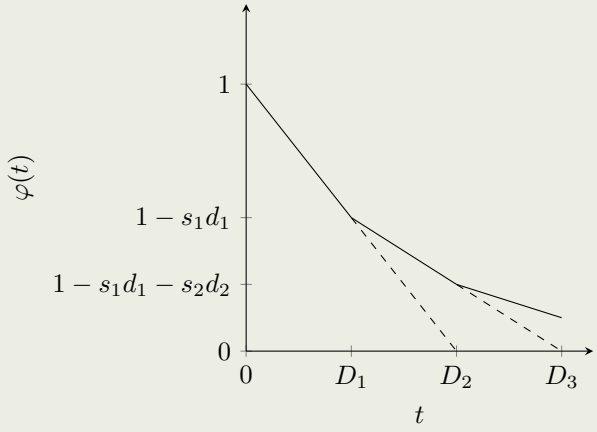
$$\sum_{k=1}^{n} \lambda_k \varphi_k(t)$$

is a characteristic function. Extend the above result for $n = \infty$.

As it turns out, any function $\varphi(t)$ can be approximated by a convex combination of the characteristic function of a Polya's distribution $\varphi(t) = (1 - |t|)_+$. Let us, therefore, follow the arguments in [10] and complete the proof of Polya's criterion.

**Exercise 6.28** Consider a function $\varphi(t)$ as constructed as below: suppose that $d_1, d_2, ... > 0$ with $\sum_{k\geq 1} d_k = \infty$, and $s_1 \geq s_2 \geq ... \geq 0$ such that $s_k \searrow 0$ and $\sum_{k=1}^{\infty} s_k d_k = 1$. We further let $d_0 = 0$, $s_0 = 1$ for convenience, and $D_n = \sum_{k=0}^{n} d_k$. Define $\varphi(t)$ as

$$\varphi(t) = \sum_{n\geq 0} \left(1 - \left(\sum_{k=0}^{n} s_k d_k\right) - s_{n+1}(d_{n+1} - |t|)\right) \chi_{[D_n, D_{n+1})}(|t|) \tag{6.33}$$




The graph of $\varphi(t)$ for $t \geq 0$ is shown above with thick lines. Let further that $\varphi_0(t) := (1 - |t|)_+$ be the characteristic function of the Polya's distribution. Show that $\varphi(t)$ is a convex combination of the characteristic function $\varphi_0(t/t_n)$ for some sequence $t_1, t_2, ...$, and hence is itself a characteristic function.

**Hint.** *We may first assume that $s_k = 0$ for $k > m$ and prove the above statement by induction on $m$, before sending $m \to \infty$ in a suitable way. For the case when $s_k = 0$ for $k > m$ we have $\sum_{k=1}^{m} d_k s_k = 1$, and that $\varphi(t)$ is a **linear** combination of the functions $\varphi(s_1 t), ...\varphi(s_m t)$. We may therefore let*

$$\varphi(t) = \sum_{i=1}^{m} \lambda_i \varphi(s_i t) \tag{6.34}$$

*for some $\lambda_1, ..., \lambda_m$, and solve for $\lambda_i$'s by considering the case when $t = D_1, ..., D_m$. Finally, show that $\lambda_1 + ... + \lambda_m = 1$.*

**Remark 6.29** The construction above indicates that two different distributions might possess characteristic functions which agree with each other over a finite interval containing $0$.

**Exercise 6.30** Complete the proof of Polya's criterion by noting that any function satisfying Polya's criterion itself can be approximated by functions in the form as in exercise 6.28.

**Hint.** *A suggested candidate for the sequence of functions is as followed: define $\varphi_n(t)$ on $t \in [0, n]$ so that it is piecewise linear and passes through the points $(0, 1), (1/n, \varphi(1/n)), (2/n, \varphi(2/n))...$, and extend by $\varphi_n(-t) = \varphi_n(t)$. The slope for each line segment decreases due to the convexity of $\varphi(t)$, so that $\varphi(t)$ is continuous and possesses a right derivative $\varphi'(t)$ with $|\varphi'(|t|)| \to 0$ as $t \to \infty$.*

*Uniform convergence can be shown as follow: fix $\epsilon > 0$, then there exists $M, N$ such that $\varphi_n(t), \varphi(t) < \epsilon$ for all $n \geq N$, uniformly for all $t$ with $|t| > M$. Then in the interval $[-M, M]$, show that $\varphi_n(t) \to \varphi(t)$ uniformly.*

### 6.4.3 Marcinkiewicz Theorem

**Theorem 6.31 — Marcinkiewicz's Theorem.** If a characteristic function $\varphi(t)$ is of the form $e^{p(t)}$, where $p(t)$ is a polynomial, then this polynomial is of degree at most 2.

**Example 6.32** As a quick example, $e^{-t^4}$ is not a characteristic function of any real-valued random variables.

Let us prove a simpler version of Marcinkiewicz's theorem, which assumes $p(t) = t^3 q(t)$ for polynomial $q(t)$. Recall that the first and second derivatives of the characteristic function represent the first and second moments. Therefore,

**Exercise 6.33 — see ((7), exercise 3.3.16).** Show that if $\varphi(t)$ is the characteristic function of the real-valued random variable $\xi$ such that $\lim_{t\to 0}(\varphi(t)-1)/t^2 = c > -\infty$, then $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] = -2c < \infty$. In particular, if $\varphi(t) = 1 + o(t^2)$, then $\mathbb{E}[\xi^2] = 0$ (with variance = 0) and $\varphi(t) = 1$.

So if $\varphi(t)$ is indeed a characteristic function, then by Taylor expansion we have $\varphi(t) = 1 + o(t^2)$, so $\varphi(t) \equiv 1$ and $q(t) \equiv 0$. The actual proof of Marcinkiewicz's theorem is beyond our scope.

### 6.4.4 Cumulants

**Definition 6.34** If there exists an expansion

$$\log \varphi_\xi(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} s_k + o(|t|^n), \quad t \to 0,$$

then the coefficients $s_k$ are called **cumulants** of $\xi$.

**Exercise 6.35** Show that

$$\mathbb{E}[\xi] = s_1, \quad \mathbb{V}[\xi] = s_2.$$

**Remark 6.36**

- If $\xi \sim N(m, \sigma^2)$ then
$$s_1 = m, \quad s_2 = \sigma^2, \quad s_k = 0, \quad k \geq 3.$$
- In general, by Marcinkiewsicz's Theorem if for a random variable $\xi$ there exists $n$ such that $s_k = 0$, for all $k \geq n$, then also $s_k = 0$ for all $k \geq 3$ and $\xi \sim N(s_1, s_2)$.

### 6.4.5 Degenerate distributions

The following theorem shows that a property of the characteristic function of a random variable can lead to a non-trivial conclusion about the nature of the random variable

**Theorem 6.37** Let $\varphi(t)$ be a characteristic function of $\xi$.

1. If $|\varphi(t_0)| = 1$ for some $t_0 \neq 0$, then $\xi$ is concentrated at the points $a + nh$, $h = 2\pi/t_0$, for some $a$, that is,
$$\sum_{-\infty}^{\infty} \mathbb{P}(\xi = a + nh) = 1,$$
where $a$ is a constant.

2. If $|\varphi(t)| = |\varphi(\alpha t)| = 1$ for two different points $t$ and $\alpha t$, where $\alpha$ is irrational, then $\xi$ is degenerate, that is
$$\mathbb{P}(\xi = a) = 1,$$
where $a$ is some constant.
3. If $|\varphi(t)| \equiv 1$, then $\xi$ is degenerate.

*Proof.* 1. If $|\varphi(t_0)| = 1$, $t_0 \neq 0$, there is a number $a$ such that $\varphi(t_0) = e^{it_0 a}$. Then
$$e^{it_0 a} = \int_{-\infty}^{\infty} e^{it_0 x} \mathrm{d}\, F(x) \implies 1 = \int_{-\infty}^{\infty} e^{it_0(x-a)} \mathrm{d}\, F(x) \implies$$
$$1 = \int_{-\infty}^{\infty} \cos t_0(x-a) \mathrm{d}\, F(x) \implies \int_{-\infty}^{\infty} 1 - \cos t_0(x-a) \mathrm{d}\, F(x) = 0.$$
Since $1 - \cos t_0(x-a) \geq 0$, it follows that
$$1 = \cos t_0(x-a) \quad (\mathbb{P}\text{-a.s.}).$$

2. If follows from $|\varphi(t)| = |\varphi(\alpha t)| = 1$ and from the previous statement that
$$\sum_{n=-\infty}^{\infty} \mathbb{P}(\xi = a + \frac{2\pi}{t} n) = \sum_{m=-\infty}^{\infty} \mathbb{P}(\xi = b + \frac{2\pi}{\alpha t} m) = 1.$$
If $\xi$ is not degenerate, then there must be at least two common points:
$$a + \frac{2\pi}{t} n_1 = b + \frac{2\pi}{\alpha t} m_1, \quad a + \frac{2\pi}{t} n_2 = b + \frac{2\pi}{\alpha t} m_2,$$
in the sets
$$\left\{ a + \frac{2\pi}{t} n, n = 0, \pm 1, \dots \right\} \quad \text{and} \quad \left\{ b + \frac{2\pi}{\alpha t} m, m = 0, \pm 1, \dots \right\},$$
whence
$$\frac{2\pi}{t}(n_1 - n_2) = \frac{2\pi}{\alpha t}(m_1 - m_2),$$
and this contradicts the assumption that $\alpha$ is irrational. Conclusion 3. follows from 2.

■

**Exercise 6.38**

1. Let $\varphi(t)$ be a characteristic function. Show that the following are also characteristic functions:
$$|\varphi(t)|^2, \quad e^{\lambda(\varphi(t)-1)}, \quad \lambda \geq 0, \quad \int_0^1 \varphi(ut)\mathrm{d}u, \quad \int_0^{\infty} e^{-u}\varphi(ut)\mathrm{d}u$$
2. Let $X$ and $Y$ be independent identically distributed random variables with zero mean and unit variance. Prove using characteristic functions that if the distribution $F$ of $(X+Y)/\sqrt{2}$ is the same as that of $X$ and $Y$, then $F$ is the normal distribution.
3. Let $\xi$ be an integer-valued random variable and $\varphi_\xi(t)$ be its characteristic function. Show that
$$\mathbb{P}(\xi = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \varphi_\xi(t) \mathrm{d}t, \quad k = 0, \pm 1, \pm 2, \dots$$
4. Show that if $\varphi(t)$ is a characteristic function, then $\Re[\varphi(t)]$ is also a characteristic function, but $\Im[\varphi(t)]$ is not.

# 7 Almost Sure Convergence of Random Series

This chapter is to deal with almost sure convergence of random series. A large part of this chapter is devoted to Kolmogorov's proof of the Strong Law of Large Number. We will also cover some other fundamental results like the Kolmogorov's 0-1 law and the law of iterated logarithm.

## 7.1 Important Zero-One Laws

To begin, let us study a few zero-one laws in probability theory. A zero-one laws states that if an event $A$ in a probability space satisfies certain conditions, then $A$ must be trivial event, i.e. the probability of $A$ must be of probability zero or one. Zero-one laws provide an important shortcuts in establishing "almost sure" statements, including almost-sure convergence.

**Example 7.1** If we prove that an event satisfies the conditions for a particular zero-one law, and that the event has non-zero probability, then it must has probability one.

### 7.1.1 Borel-Cantelli Lemma

Perhaps the example with the easiest proof is the Borel-Cantelli lemma, which you might have seen in elementary probability classes. To proof this important lemma, assume $A_n$ be sequences of events (elements from the $\sigma$-algebra of a probability space $(\Omega, \mathscr{F}, \mathbb{P})$). Recall the following events:

**Definition 7.2 — Infinitely often and eventually events.** Define:

- Infinitely often events:

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k = \{\omega \in \Omega : \omega \in A_m \text{ for infinitely many } m\} \tag{7.1}$$

  If $\omega \in \limsup_{n\to\infty} A_n$, we say that $A_n$ occurs infinitely often (i.o.).

- Eventually (all except finitely often):

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k = \{\omega \in \Omega \,|\, \exists\, m_0(\omega) \text{ such that } \omega \in A_m \text{ for all } m \geq m_0(\omega)\} \tag{7.2}$$

  If $\omega \in \limsup_{n\to\infty} A_n$, we say that $A_n$ occurs eventually (or almost except finitely often, a.e.f.o).

**Exercise 7.3**

1. Try to give a precise interpretation of the descriptions "infinitely often" and "eventually", "limit infimum" and "limit supremum". *Hint*: convince yourself that

$$\chi_{\liminf_{n\to\infty} A_n} \tag{7.3}$$

2. Show that

$$\mathbb{P}\left(\liminf_{n\to\infty} A_n\right) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \tag{7.4}$$

   and prove a similar statement for the probability of event $\liminf_{n\to\infty} A_n$.

3. Describe the complement of $\limsup_{n\to\infty} A_n$.

We are now ready to prove the Borel-Cantelli Lemma.

**Theorem 7.4 — Borel-Cantelli Lemma.**

1. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(A_n \ i.o.) = 0$.
2. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $A_n$ are mutually independent then $\mathbb{P}(A_n \ i.o.) = 1$.

We note that if an event $A$ can be expressed as the lim-sup of an infinite sequence of mutually independent events $(A_n)_{n\geq 1}$, then Borel-Cantelli Lemma states that the probability of $A$ must be either zero and one. As a result, Borel-Cantelli lemma is a zero-one law.

*Proof.*

1. By continuity of measure we have

$$\mathbb{P}(A_n \ i.o.) = \lim_{n\to\infty} \mathbb{P}\bigg(\bigcup_{k\geq n} A_k\bigg) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0.$$

2. Consider $\{A_n \ i.o.\}^c = \{A_n^c \ ev.\} = \bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k^c$. We have

$$1 - \mathbb{P}(A_n \ i.o.) = \lim_{n\to\infty} \mathbb{P}\bigg(\bigcap_{k\geq n} A_k^c\bigg).$$

   Also,

$$\mathbb{P}\bigg(\bigcap_{k\geq n} A_k^c\bigg) = \prod_{k\geq n} \mathbb{P}(A_k^c).$$

   Note that $\log(1-x) \leq -x$ for $x \in [0,1)$, and thus

$$\log \mathbb{P}\bigg(\bigcap_{k\geq n} A_k^c\bigg) = \log \prod_{k\geq n} (1 - \mathbb{P}(A_k)) \leq -\sum_{k\geq n} \mathbb{P}(A_k) = -\infty,$$

   i.e. $\mathbb{P}\bigg(\bigcap_{k\geq n} A_k^c\bigg) = 0$ for all $n$.

■

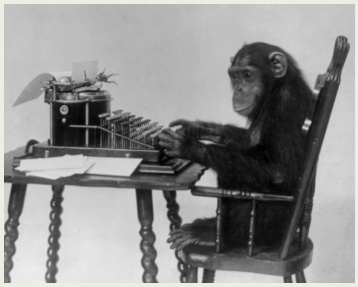
Here we give an example of an application of Borel Cantelli lemma.

**Example 7.5 — Infinite Monkey Theorem.** For Lebesgue almost every real number in $[0,1]$, its binary expansion contains any finite string of $\{0,1\}$ infinitely many times. To see that, assume the desired string to be $(x_1, ..., x_m)$. We consider the sequence of events on $([0,1], \mathscr{B}([0,1]), \mathsf{Leb})$:

$$A_n := \{\omega \,|\, \xi_{nm+1}(\omega) = x_1, \xi_{nm+2}(\omega) = x_2, ..., \xi_{nm+m}(\omega) = x_m\}, \quad n \geq 0$$

where $\xi_k$ are the Radamacher functions as constructed in section 4.1 corresponding to the $k$-th digit of the binary expansion of $\omega \in [0,1)$. The events $A_n$ represents that the desired string appears starting from digit $nm+1$, which are mutually independent given that the Radamacher functions are independent as well. Also note that $\sum_{n\geq 1} \lambda(A_n) = \infty$, so $\mathsf{Leb}(\{A_n \text{ i.o.}\}) = 1$ as desired.

There are many versions of this theorem, with one saying that if we provide an infinite amount of time for a monkey to hit keys on a typewriter keyboard randomly, then it is almost certain that the monkey will type a given finite string of characters infinitely often. The proof of this statement is exactly the same, except here we look at $[0,1)$ with a 26-adic expansion. However, the Borel-Cantelli lemma does not provide information on how much time it would require for a monkey to complete the entire string...



### 7.1.2 Kolmogorov's 0-1 Law

We look at a generalisation of the Borel-Cantelli Lemma. To begin, let us define the notion of tail events and tail $\sigma$-algebra by first introducing some notations:

**Definition 7.6** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space:

1. Let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be sub-$\sigma$-algebra of $\mathscr{F}$. Define the $\sigma$-algebra generated by their union:
$$\bigvee_{i=1}^{n} \mathscr{F}_i = \sigma\left(\bigcup_{i=1}^{n} \mathscr{F}_i\right), \tag{7.5}$$
and extend this definition to the case when $n = \infty$.
2. Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables defined on this probability space. Then
$$\sigma(\xi_1, \xi_2, \ldots, \xi_n) = \bigvee_{i=1}^{n} \sigma(\xi_i), \tag{7.6}$$
with this definition being extended to the infinite case.

**Definition 7.7** Under the above settings, define $\mathscr{F}_n^p = \sigma(\xi_n, \ldots, \xi_p)$ for $p \geq n$ and $\mathscr{F}_n^\infty = \sigma(\xi_n, \ldots)$ (representing the future information), then the $\sigma-$ algebra associated with the sequence of random variables $\xi_1, \xi_2, \ldots$ is
$$\mathscr{T} = \bigcap_{n=1}^{\infty} \mathscr{F}_n^\infty$$
is called the **tail** $\sigma-$**algebra**. Events of $\mathscr{T}$ are called tail events.

We are more interested in studying tail $\sigma$-algebra associated with a mutually independent sequence of random variables $\xi_1, \xi_2, \ldots$. If this assumption holds, then we have the following observation:

**Lemma 7.8** For all $n \geq 2$, $\mathscr{F}_1^n$ and $\mathscr{F}_{n+1}^\infty$ are independent.

*Proof.* Recall the shortcut developed in lemma 3.9! We first prove that $\mathscr{F}_1^n$ is independent with $\mathscr{F}_{n+1}^p$ for any $n+1 \leq p < \infty$. Note that $\mathscr{F}_1^n$ is generated by the $\pi$-system $\cup_{i=1}^n \sigma(\xi_i)$ and $\mathscr{F}_{n+1}^p$ is similarly generated by the $\pi$-system $\cup_{i=n+1}^p \sigma(\xi_i)$, and that these $\pi$-systems are independent, so we know that $\mathscr{F}_1^n$ is independent with $\mathscr{F}_{n+1}^p$ for any finite $p$. Now note that $\mathscr{F}_n^\infty$ is generated by the $\pi$-system $\cup_{p \geq n+1} \mathscr{F}_{n+1}^p$, so $\mathscr{F}_1^n$ is generated by $\mathscr{F}_{n+1}^\infty$. ∎

Note that $\mathscr{T} \subseteq \mathscr{F}_n^\infty$ for any **finite** $n \geq 1$, so we have $\mathscr{T}$ being independent with any $\mathscr{F}_1^n$. This provides an interpretation of the definition of a tail event $A \in \mathscr{T}$, that *its occurrence is independent of a finite number of changes in values of* $\xi_n$. An important example will be to note that the convergence of a sequence of random variable $\xi_1, \xi_2, \ldots$ is not affected when only a finite number of $\xi_n$'s are changed.

**Example 7.9** We see, for example, that $\{\xi_{10} \in B\}$ may not be in $\mathscr{T}$, since its occurrence may be affected by changing a finite number of $\xi_n$, in our case being $\xi_{10}$. We also note that $\{\omega \,|\, \forall n,\, \xi_n \notin \mathscr{B}\}$ is not in $\mathscr{T}$ since its occurrence may change by just affecting a value of $\xi_n$.

In fact, the above observations lead to a more surprising result! Since $\cup_{n \geq 1} \mathscr{F}_1^n$ is a $\pi$-system generating $\mathscr{F}_1^\infty$, we see that $\mathscr{T}$ is independent with $\mathscr{F}_1^\infty$, and hence $\mathscr{T}$ is actually independent with itself! This leads to the well-known Kolmogorov's Zero-One Law.

**Theorem 7.10 — Kolmogorov's Zero-One Law.** Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, and let $A \in \mathscr{T}$. Then $\mathbb{P}(A)$ can only have a value of zero or one.

*Proof.* We have $\mathbb{P}(A) = \mathbb{P}(A \cap A) = (\mathbb{P}(A))^2$, so $\mathbb{P}(A)$ must has value zero or one. ∎

**Example 7.11** Given independent sequence $\xi_1, \xi_2, ...$ and $B_1, B_2, ... \in \mathscr{B}(\mathbb{R})$, then $\{\xi_n \in B_n \text{ i.o.}\} = \cap_{n=1}^{\infty} \cup_{k \geq n} \{\xi_k \in B_k\} \in \mathscr{T}$. Now consider the independent events $A_1, A_2, ...$, then the sequence of random variables $\chi_{A_1}, \chi_{A_2}, ...$ are independent, and that the event $\limsup_n A_n := \{\chi_{A_i} = 1 \text{ i.o.}\}$ is a tail event associated with this independent sequence of random variables. The Kolmogorov's Zero-One Law then says that $\limsup_n A_n$ must have probability zero and one, which coincides with our observation from the Borel-Cantelli lemmas.

Perhaps the most important application of Kolmogorov's Zero-One Law is that it says that a random series of **independent** random variables must converge or diverge almost surely. Let $\xi_1, \xi_2, ...$ be sequence of independent random variables. Then obviously the following function is $\mathscr{F}_k^{\infty}$ measurable:

$$\limsup_{n\to\infty} \sum_{i\geq k} \xi_i, \quad \liminf_{n\to\infty} \sum_{i\geq k} \xi_i, \quad \lim_{n\to\infty} \sum_{i\geq k} \xi_i$$

As a result, the following functions are $\mathscr{T}$ measurable:

$$\limsup_{n\to\infty} \sum_{i\geq 1} \xi_i, \quad \liminf_{n\to\infty} \sum_{i\geq 1} \xi_i, \quad \lim_{n\to\infty} \sum_{i\geq 1} \xi_i$$

In particular, the following set is a tail event associated with the given independent sequence of random variables:

$$A = \left\{ \omega \,\middle|\, \sum_{i\geq 1} \xi_i \text{ exists} \right\},$$

and as a result of the Kolmogorov 0-1 law, we have the following

**Corollary 7.12** If $\xi_1, \xi_2, \ldots$ is a sequence of independent random variables, then $\sum_{n=1}^{\infty} \xi_n$ either converges a.s. or diverges a.s.

We finally note an application of Kolmogorov 0-1 law:

**Example 7.13** Let $\eta$ be a real-valued random variable that is measurable with respect to the $\sigma-$algebra $\mathscr{T}$, i.e. $\{\eta \in B\} \in \mathscr{T}$, $B \in \mathscr{B}(\mathbb{R})$. Then $\eta$ is degenerate, i.e. there is a constant $c$ such that $\mathbb{P}(\eta = c) = 1$.

*Proof.* Note that for all $x \in \mathbb{R}$ one have $\mathbb{P}(\eta \leq x) \in \{0, 1\}$ by the Kolmogorov 0-1 law. Take $c = \inf_{x \in \mathbb{R} \,|\, \mathbb{P}(\eta \leq x) = 1}$. Such an infinmum exists given that the distribution function is increasing, so if the infinmum does not exists then it means $0 = \mathbb{P}(\varnothing) = \mathbb{P}(\eta \leq -\infty) = 1$, which is a contradiction. It is easy to check that $\mathbb{P}(\eta = c) = 1$ by utilising the fact that $\mathbb{P}(\eta \leq x)$ is an increasing cadalag, and we will leave it as an exercise. ■

There are many more zero-one laws in probability theory, but we will defer the discussion of those statements to a later chapter.

## 7.2 More on Almost Sure Convergence

### 7.2.1 Almost sure convergence implies convergence in probability

Recall the definition of almost sure convergence: a sequence $\xi_1, \xi_2, \dots$ of random variables converges almost surely to the random variable $\xi$ (denoted as $\xi_n \xrightarrow{a.s} \xi$) if $\mathbb{P}(\{\omega \,|\, \xi_n(\omega) \not\to \xi(\omega)\}) = 0$. This section is dedicated to explore the relationships between almost sure convergence, $L^p$ convergence and convergence in probability. As a warm up, we begin by the following equivalent definition of almost sure convergence.

**Proposition 7.14** A necessary and sufficient condition that $\xi_n \to \xi$ $\mathbb{P}$-almost surely is that

$$\mathbb{P}\left( \sup_{k \geq n} |\xi_k - \xi| \geq \varepsilon \right) \to 0, \quad n \to \infty,$$

for every $\varepsilon > 0$.

**Hint.** *The key step is to rewrite the set $\{\xi_n \not\to \xi\}$ into a countable union of sets. From the definition of convergence, we have*

$$\xi_n(\omega) \not\to \xi(\omega) \iff \exists\, \epsilon > 0 \text{ s.t. } |\xi_n(\omega) - \xi(\omega)| \geq \epsilon \text{ infinitely often.}$$

*So if we let $A_n^\varepsilon = \{\omega : |\xi_n - \xi| \geq \varepsilon\}$ and $A^\varepsilon = \limsup A_n^\varepsilon = \{A_n^\varepsilon \text{ i.o.}\}$, then we have*

$$\{\omega \,|\, \xi_n(\omega) \not\to \xi(\omega)\} = \bigcup_{\varepsilon \geq 0} A^\varepsilon.$$

*But this is not a countable union. Fortunately, the sets $A^\epsilon$ are nested, so one can restrict $\epsilon$ to the form $\epsilon = 1/m$ for some positive integers $m$, so we have*

$$\{\omega \,|\, \xi_n(\omega) \not\to \xi(\omega)\} = \bigcup_{m=1}^{\infty} A^{1/m}.$$

*The remaining of proof utilises the fact that the sets $A^\epsilon$ are nested.*

*Proof.* By nestedness of $A^\epsilon$ we have the following chain of implications:

$$\begin{aligned}
\mathbb{P}(\{\omega : \xi_n \not\to \xi\}) = 0 &\iff \mathbb{P}\left( \bigcup_{\varepsilon > 0} A^\varepsilon \right) = 0 \\
&\iff \mathbb{P}\left( \bigcup_{m=1}^{\infty} A^{1/m} \right) = 0 \\
&\overset{(*)}{\iff} \forall m \geq 1,\ \mathbb{P}(A^{1/m}) = 0 \\
&\iff \forall \varepsilon > 0,\ \mathbb{P}(A^\varepsilon) = 0
\end{aligned}$$

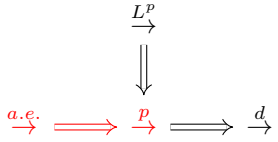
The equivalence $(*)$ is justified below: first note that $\mathbb{P}(A^{1/m}) \leq \mathbb{P}(\cup_{m \geq 1} A^{1/m})$ so $(\Rightarrow)$ clearly follows. The $(\Leftarrow)$ direction follows from union bound: $\mathbb{P}(\cup_{m \geq 1} A^{1/m}) \leq \sum_{m \geq 1} \mathbb{P}(A^{1/m}) = 0$.

To complete the proof, we note that

$$\mathbb{P}(A^\varepsilon) = \mathbb{P}\left( \bigcap_{n \geq 1} \bigcup_{k \geq n} A_k^\varepsilon \right) = \lim_n \mathbb{P}\left( \bigcup_{k \geq n} A_k^\varepsilon \right) = \mathbb{P}\left( \sup_{k \geq n} |\xi_k - \xi| \geq \varepsilon \right),$$

so the result follows. ∎

From this equivalent statement, it is clear that almost sure convergence implies convergence in measure. In our summary diagram, we have proven the following highlighted implication.



**Exercise 7.15 — Cauchy-like condition for almost sure convergence.** Show that the sequence $\{\xi_n\}_{n\geq 1}$ converges almost surely iff, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left( \sup_{k,l\geq n} |\xi_k - \xi_l| \geq \varepsilon \right) = 0,$$

or equivalently,

$$\mathbb{P}\left( \sup_{k\geq 0} |\xi_{n+k} - \xi_n| \geq \varepsilon \right) \to 0, \quad n \to \infty.$$

**Hint.** *Let*

$$B_{k,l}^{\varepsilon} = \{\omega : |\xi_k - \xi_l| \geq \varepsilon\}, \quad B^{\varepsilon} = \bigcap_{n=1}^{\infty} \bigcup_{k,l\geq n} B_{k,l}^{\varepsilon}.$$

*Then*

$$\{\omega : \{\xi_n(\omega)\}_{n\geq 1} \textit{ is not convergent}\} = \bigcup_{\varepsilon>0} B^{\varepsilon},$$

*and it can be shown as in proposition 7.14 that*

$$\mathbb{P}(\{\omega : \{\xi_n(\omega)\}_{n\geq 1} \textit{ is not convergent}\}) = 0$$

*if and only if the statement in the theorem holds.*

### 7.2.2 Convergence in probability does not imply almost sure convergence

The following counter-examples demonstrate that both convergences in probability and convergences in $L^p$ do not imply almost convergence.

**Example 7.16** Let $\xi_1, \ldots, \xi_n$ be independent random variables on a certain probability space $(\Omega, \mathscr{F}, \mathbb{P})$ taking value from $\{0, 1\}$ with

$$\mathbb{P}(\xi_n = 1) = 1/n, \quad \mathbb{P}(\xi_n = 0) = 1 - 1/n.$$

We may choose $(\Omega, \mathscr{F}) = (\{0,1\}, \mathscr{B}(\{0,1\}))^{\mathbb{N}}$ and $\mathbb{P}$ being an appropriate infinite product measure. Then

$$\mathbb{E}[|\xi_n - 0|^p] = 1/n \to 0, \text{ so } \xi_n \xrightarrow{L^p} 0.$$

However,

$$\{\omega : \xi_n \to 0\} = \{\xi_n = 0 \text{ everywhere}\} = \bigcup_{n=1}^{\infty} \underbrace{\bigcap_{k\geq n} \{\xi_k = 0\}}_{\text{increasing sequence of sets}}$$

Therefore,

$$\begin{aligned} \mathbb{P}(\xi_n \to 0) &= \lim_{n\to\infty} \mathbb{P}\left( \bigcap_{k\geq n} \{\xi_k = 0\} \right) \\ &= \lim_{n\to\infty} \prod_{k\geq n} \mathbb{P}(\xi_k = 0) \\ &= \lim_{n\to\infty} \prod_{k\geq n} \left(1 - \frac{1}{k}\right) = 0 \end{aligned}$$

Indeed,

$$\prod_{k\geq n} \left(1 - \frac{1}{k}\right) = \lim_{N\to\infty} \prod_{k=n}^{N} \frac{k-1}{k} = \lim_{N\to\infty} \frac{n-1}{n} \frac{n}{n+1} \cdots \frac{N-1}{N} = 0.$$

Thus $\xi_n \xrightarrow{a.s.} 0$.

**Example 7.17 — Typewriter Sequence.** We provide another example on $([0,1], \mathscr{B}([0,1]), \text{Leb})$ with a similar flavor as the previous example. Consider the sequence

$$f_n := \chi_{A_n}, \quad A_n = \left[\frac{n}{2^k} - 1, \frac{n+1}{2^k} - 1\right] \tag{7.7}$$

whenever $k \geq 0$ and $2^k \leq n < 2^{k+1}$ The first few $A_n$ is as followed:

$$[0,1], [0,1/2], [1/2,1], [0,1/4], [1/4,1/2], [1/2,3/4], [3/4,1], \ldots$$

You can plot the indicator functions yourself, or look at some demonstrations on websites. [a]. You will see that the indicator functions move from left to right over $[0,1]$, then half its width and repeat again, which provides an explanation of the sequence's name. [b]

The convergence in $L^p$ for any $p < \infty$ (hence convergence in $L^p$) of the sequence to zero is proven by noting that the width of indicator function vanishes as $n \to \infty$. However, given that the indicator function moves from left to right infinitely many times, for all $\omega \in [0,1]$, $f_n(\omega) = 1$ (and 0) infinitely often. As a result, $f_n(\omega)$ does not converge almost surely.

[a] e.g. https://math.stackexchange.com/questions/1412091/the-typewriter-sequence

[b] Figure provided by Dr. Shordzi under the Creative Common 2.0 License: https://creativecommons.org/licenses/by-sa/2.0/

However, the implication from almost sure convergence to convergence in probability admits a partial converse: if a sequence converges in probability, then we can extract a subsequence that converges almost surely. To prove this important result, we note that

**Lemma 7.18** A sufficient condition for $\xi_n \xrightarrow{a.s.} \xi$ is that

$$\sum_{n=1}^{\infty} \mathbb{P}(|\xi_n - \xi| \geq \varepsilon) < \infty$$

is satisfied for all $\varepsilon > 0$.

*Proof.* Denote the event $A_n^\varepsilon = \{\omega \,|\, |\xi_n(\omega) - \xi(\omega)| \geq \varepsilon\}$ as required. Since $\sum_{k\geq 1} \mathbb{P}(A^\epsilon) < \infty$, by first Borel-Cantelli lemma we know that $\mathbb{P}(A^\varepsilon) := \mathbb{P}(A_n^\epsilon \text{ i.o.}) = 0$. Following the arguments in proposition 7.14 we know that $\mathbb{P}(\{\omega \,|\, \xi_n \not\to \xi\}) = 0$ as desired. ■

**Corollary 7.19** Let $(\varepsilon_n)_{n\geq 1}$ be a sequence of positive numbers such that $\varepsilon_n \downarrow 0$ as $n \to \infty$. Then if $\xi_n$ converges to $\xi$ in probability sufficiently fast in the sense that

$$\sum_{n=1}^{\infty} \mathbb{P}(|\xi_n - \xi| \geq \varepsilon_n) < \infty,$$

then $\xi_n \xrightarrow{a.s.} \xi$.

*Proof.* Fix an arbitrary $\epsilon > 0$. Choose $N$ such that for all $n \geq N$ we have $\epsilon_n < \epsilon$. Then

$$\sum_{n=1}^{\infty} \mathbb{P}(|\xi_n - \xi| \geq \epsilon) \leq \underbrace{\sum_{n=1}^{N-1} \mathbb{P}(|\xi_n - \xi| \geq \epsilon)}_{\leq N-1 < \infty} + \underbrace{\sum_{n=N}^{\infty} \mathbb{P}(|\xi_n - \xi| \geq \epsilon)}_{< \infty},$$

hence by the above lemma 7.18 we have $\xi_n \xrightarrow{\text{a.s.}} \xi$ as $n \to \infty$. ■

We may now prove our key result for this subsection.

**Theorem 7.20** If $\xi_n \xrightarrow{p} \xi$, then there exists a subsequence such that

$$\xi_{n_k} \xrightarrow{a.s.} \xi.$$

*Proof.* Since $\lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| > 1/k) = 0$ for all $k \geq 1$, we can choose a subsequence such that

$$\mathbb{P}(|\xi_n - \xi| > k^{-1}) \leq 2^{-k} \quad \forall k \geq 1$$

Since $\sum_{k=1}^{\infty} 2^{-k}$ converges, by above corollary 7.19 we have $\xi_{n_k} \xrightarrow{a.s.} \xi$. ■

The trick of extracting a almost surely convergent subsequence is very helpful in justifying limits. Here is an immediate application of the above theorem:

**Corollary 7.21** If $\xi_1 \geq \xi_2 \geq \cdots \geq 0$, are random variables and $\xi_n \xrightarrow{p} 0$, then

$$\xi_n \xrightarrow{a.s.} 0.$$

*Proof.* Note that $\xi_n \xrightarrow{a.s} 0$ if and only if $\limsup \xi_n = 0$ a.s. Let $\varepsilon > 0$. Denote $A_n = \{\xi_n > \varepsilon\}$. Then by continuity,

$$\mathbb{P}(\limsup \xi_n > \varepsilon) = \mathbb{P}(\xi_n > \varepsilon \, i.o.) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k\geq n} A_k\Big).$$

Since $A_n$ is non-increasing, the right hand side equals $\lim_{n\to\infty} \mathbb{P}(A_n)$ which is 0 since $\xi_n \xrightarrow{p} 0$. Thus

$$\mathbb{P}(\limsup \xi_n > \varepsilon) = 0 \quad \forall \varepsilon > 0,$$

and therefore

$$\mathbb{P}(\xi_n \not\to \xi) \leq \sum_{m=1}^{\infty} \mathbb{P}(\limsup \xi_n > 1/m) = 0.$$

■

We finally note an observation regarding convergence in probability:

**Corollary 7.22** Let $\xi, (\xi_n)_{n\geq 1}$ be random variables on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then $\xi_n \to \xi$ in probability as $n \to \infty$ iff for any subsequence, ther is a further subsequence that converges $\mathbb{P}$-a.s. to $\xi$.

*Proof.* The $(\Rightarrow)$ direction is proven in theorem 7.20. For $(\Leftarrow)$, we note that almost sure convergence implies convergence in probability. We also recall a basic fact that in metric space, if $(a_m)$ is a sequence such that for any subsequence there is a further subsequence that converges to some element $a$, then $a_m \to a$. Since convergence in probability is metrisable as shown in exercise 4.16, we may directly use this result to complete the proof. ■

**Exercise 7.23**

1. Let $\xi_1, \xi_2, \ldots, \xi_n$ be i.i.d. random variables with the standard exponential distribution and define

$$M_n = \max\{\xi_1, \ldots, \xi_n\}.$$

Show that as $n \to \infty$

- $\limsup \xi_n / \log n = 1$ a.s.
- $M_n / \log n = 1$ a.s.

### 7.2.3 Almost sure convergence and $L^p$ convergence

It has been proven in elementary measure theory classes that having almost sure convergence does not guarantee $L^p$ convergence. Here we prove a result which states that if we have almost sure convergence and convergence in mean, then we have $L^1$ convergence.

**Theorem 7.24** Let $\xi_n$ be a sequence of nonnegative random variables such that $\xi_n \xrightarrow{a.s.} \xi$ and $\mathbb{E}[\xi_n] \to \mathbb{E}[\xi] < \infty$. Then $\xi_n \xrightarrow{L^1} \xi$, or

$$\mathbb{E}[|\xi_n - \xi|] \to 0, \quad n \to \infty.$$

*Proof.* We have $\mathbb{E}[|\xi_n|] < \infty$ for sufficiently large $n$ and therefore for such $n$ we have

$$\begin{aligned} \mathbb{E}[|\xi_n - \xi|] &= \mathbb{E}[\xi - \xi_n]\chi_{\xi \geq \xi_n} + \mathbb{E}[\xi_n - \xi]\chi_{\xi_n \geq \xi} \\ &= 2\mathbb{E}[\xi - \xi_n]\chi_{\xi \geq \xi_n} + \mathbb{E}[\xi_n - \xi]. \end{aligned}$$

But $0 \leq (\xi - \xi_n)\chi_{\xi \geq \xi_n} \leq \xi$. Therefore, by the dominated convergence theorem, $\mathbb{E}[\xi - \xi_n]\chi_{\xi \geq \xi_n} \to 0$. ■

## 7.3 Strong Law of Large Numbers

We are now ready to prove some strong law of large numbers. A strong law of large number is the statement of the following form: given $\xi_1, \xi_2, \ldots$ is a sequence of integrable random variables, which is assumed to be independent (or "weakly correlated"). Let $S_n = \xi_1 + \xi_2 + ... + \xi_n$, then the sequence $(\xi_n)_{n \geq 1}$ is said to satisfy the strong law of large numbers if

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{\text{a.s.}} 0.$$

It is stronger than the weak law of large number as proven in chapter 4 in the sense that we only have convergence in probability in contrast to almost sure convergence.

Recall that for i.i.d. $\xi_n$ with a second moment assumption (i.e. $\mathbb{V}[\xi_1] < \infty$) we have the $L^2$ weak law of large number. If we impose an excessive moment assumptions, we can prove a strong law of large number.

**Proposition 7.25 — Cantelli's Strong Law of Large Numbers.** Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[\xi_1^4] < \infty$. Then

$$\frac{S_n - \mathbb{E}[S_n]}{n} \xrightarrow{\text{a.s.}} 0, \tag{7.8}$$

i.e. $\{\xi_n\}$ satisfies the strong law of large numbers.

**Hint.** *Without loss of generality we centralise the random variables by substracting its mean, then* $\mathbb{E}[\xi_1] = \mathbb{E}[\xi_1^3] = 0$. *Fix an arbitrary* $\epsilon > 0$, *we hope to use lemma 7.18 to establish almost sure convergence. We will bound the probability* $\mathbb{P}(|S_n/n| \geq \epsilon) = \mathbb{P}(|S_n| \geq n\epsilon)$ *by using Chebyshev inequality.*

*Proof.* Let us therefore bound the probability $\mathbb{P}(|S_n/n| \geq \epsilon)$, so that it decays fast enough and $\sum_{n \geq 1} \mathbb{P}(|S_n/n| \geq \epsilon) < \infty$. We might therefore want $\mathbb{P}(|S_n| \geq n\epsilon) = O(1/n^2)$.

To use the Chebyshev inequality, one has to bound the $k$-th moment of $S_n/n$ (where $k$ is an even number). We want the $k$-th moment to be of order $O(1/n^2)$. We may look at second moments:

$$\mathbb{V}[S_n/n] = n\mathbb{E}[\xi_1]/n^2, \tag{7.9}$$

which is not enough for Chebyshev inequality. We have to look at the forth moments. In the expansion of forth moments, the only non-vanishing terms are terms of the form $\mathbb{E}[\xi_j^4]$ and $\mathbb{E}[\xi_i^2\xi_j^2] = \mathbb{E}[\xi_i^2]\mathbb{E}[\xi_j^2]$ with $i \neq j$, noticing that the odd moments of $\xi_i$ vanishes. We therefore have the expansion

$$\mathbb{E}[S_n^4] = \mathbb{E}\left[\sum_{k=1}^n \xi_k^4 + \binom{4}{2,2} \sum_{j,k=1, j<k} \xi_j^2 \xi_k^2\right] \tag{7.10}$$

where $\binom{4}{2,2} = 4!/(2!)^2 = 6$ comes from the multinomial theorem. Now notice that there are $n(n-1)/2$ unique ways to choose the indices $(j,k)$ such that $j < k$, therefore we have, from the i.i.d. assumption,

$$\mathbb{E}[|S_n/n|^4] = \frac{1}{n^4}\left(n\mathbb{E}[\xi_1^4] + 3n(n-1)(\mathbb{E}[\xi_1^2])^2\right) \tag{7.11}$$

Finally, note from Lyapunov inequality (corollary 2.25) that $\mathbb{E}[\xi_1^2]^{1/2} \leq \mathbb{E}[\xi_1^4]^{1/4}$, so we have

$$\mathbb{E}[|S_n/n|^4] \leq \frac{3n^2 - 2n}{n^4}\mathbb{E}[\xi_1^4] \lesssim \frac{1}{n^2}, \tag{7.12}$$

and therefore $\sum_{n\geq 1} \mathbb{P}(|S_n/n| \geq \epsilon) \lesssim \sum_{n\geq 1} n^{-2} < \infty$ as desired. ■

This strong law of large number is already very useful for practical purposes. It can be used when $\xi_1$ follows common distributions (e.g. Normal, Gamma, Binomial, ...), or have finite supports. However, as seen below, we can impose much weaker assumptions on the moments of $\xi_i$.

We will follow Kolmogorov's approaches in proving his strong law of large numbers. The approach is cumbersome since it involves many technical lemma in analysis, nevertheless it opens up routes of proving more general statements about convergence of random series. Another route is to consider truncation of random variables, which is simpler and actually leads to a stronger version of strong law of large number by Etemadi as proved in [11]. We will not cover this proof in our notes.

Back to Kolmogorov's approach. Let us start with the following Kolmogorov maximal inequality:

**Proposition 7.26 — Kolmogorov's maximal inequality.** Let $\xi_1, \xi_2, \ldots$ be independent random variables with finite variances. Then $\forall n \geq 1$ and $x > 0$, we have

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k - \mathbb{E}[S_k]| \geq x\right) \leq \frac{\mathbb{V}[S_n]}{x^2}.$$

Notice that this is a stronger version of the Chebyshev inequality, since Chebyshev inequality only gives

$$\mathbb{P}\left(|S_k - \mathbb{E}[S_k]| \geq x\right) \leq \frac{\mathbb{V}[S_n]}{x^2}.$$

Combining with union bound we get

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k - \mathbb{E}[S_k]| \geq x\right) = \mathbb{P}\left(\bigcup_{k=1}^n \{|S_k - \mathbb{E}[S_k]| \geq x\}\right) \leq \frac{n\mathbb{V}[S_n]}{x^2}.$$

What this maximal inequality does is to remove the factor of $n$ generated in the union bound. As seen in 7.18, if we are able to bound the probability of maximum deviation using maximal inequalities, then we can use Borel-Cantelli to establish almost sure convergence.

**Hint.** *Again we subtract the random variables with its mean to set $\mathbb{E}[\xi_1] = 0$. The key step is to note that the events $A_k, k = 1, ..., n$ is a partition of $A = \{\omega \mid \max_{1\leq k\leq n} |S_k(\omega)| \geq x\}$, where $|S_j|$ first exceeds $x$ at $j = k$, i.e.*

$$A_k := \{\omega \,|\, |S_j(\omega)| < x \text{ and } |S_k(\omega)| \geq x\}.$$

*Now show that $\mathbb{E}[S_n^2 \chi_{A_k}] \geq x^2 \mathbb{P}(A_k)$ by breaking up $S_n$.*

*Proof.* From the above hint, we have

$$\mathbb{V}[S_n^2] \geq \mathbb{E}[S_n^2 \chi_A] = \sum_{k=1}^{n} \mathbb{E}[S_n^2 \chi_{A_k}].$$

But

$$\begin{aligned}\mathbb{E}[S_n^2 \chi_{A_k}] &= \mathbb{E}[(S_k + \xi_{k+1} + ... + \xi_n)^2 \chi_{A_k}] \\ &= \underbrace{\mathbb{E}[S_k^2 \chi_{A_k}]}_{\geq x^2 \mathbb{P}(A_k)} + \underbrace{2\mathbb{E}[(S_k \chi_{A_k})(\xi_{k+1} + ... + \xi_n)]}_{=2\mathbb{E}[S_k \chi_{A_k}]\mathbb{E}[\xi_{k+1}+...+\xi_n]=0} + \underbrace{\mathbb{E}[(\xi_{k+1} + ... + \xi_n)^2]}_{\geq 0} \\ &\geq x^2 \mathbb{P}(A_k),\end{aligned}$$

so we have

$$\mathbb{V}[S_n^2] \geq x^2 \sum_{k=1}^{n} \mathbb{P}(A_k) = x^2 \mathbb{P}(A). \tag{7.13}$$

■

With this maximal inequality, we may prove the Kolmogorov's three series theorem, which states that the almost sure convergence of a (truncated) random series depends on the convergence of three deterministic series. Here we will prove the statement in three steps:

**Lemma 7.27 — Three series theorem, part I.** Suppose $\xi_1, \xi_2, ...$ are independent real-valued series and have $\mathbb{E}[\xi_i] = 0$ for all $i$. If we know that $\sum_{i\geq 1} \mathbb{V}[\xi_i] < \infty$, then $\sum_{i\geq 1} \xi_i$ converges almost surely.

*Proof.* Since we are not given the limit for which the random series converges, the best way of proving this theorem is to resort on proving the series $S_n := \sum_{i=1}^{n} \xi_i$ being Cauchy as in exercise 7.15. To this we need to show that as $n \to \infty$,

$$\mathbb{P}\left(\sup_{k\geq 1} |S_{n+k} - S_k| \geq \epsilon\right) \to 0$$

We note that for all $m < \infty$ we have, from the above maximal inequality,

$$\mathbb{P}\left(\sup_{1\leq k\leq m} |S_{n+k} - S_k| \geq \epsilon\right) \leq \epsilon^{-2} \sum_{k=1}^{m} \mathbb{V}[\xi_{n+k}] \leq \epsilon^{-2} \sum_{k=1}^{\infty} \mathbb{V}[\xi_{n+k}], \tag{7.14}$$

so one can send $m \to \infty$ (by continuity from below):

$$\mathbb{P}\left(\sup_{k\geq 1} |S_{n+k} - S_k| \geq \epsilon\right) \leq \sum_{k=1}^{\infty} \mathbb{V}[\xi_{n+k}] \overset{n\to\infty}{\to} 0 \tag{7.15}$$

as desired, so the series $S_n$ converges almost surely. ■

**Theorem 7.28 — Three series theorem, part II.** Let $\xi_1, \xi_2, \ldots$ be independent random variables. If $\sum_{n=1}^{\infty} \mathbb{E}[\xi_n]$ and $\sum_{n=1}^{\infty} \mathbb{V}[\xi_n]$ converge, then $\sum_{n=1}^{\infty} \xi_n$ converges a.s.

*Proof.* This follows by breaking the random series into two parts:

$$\sum_{i=1}^{\infty} \xi_i = \sum_{i=1}^{\infty} (\xi_i - \mathbb{E}[\xi_i]) + \sum_{i=1}^{\infty} \mathbb{E}[\xi_i], \tag{7.16}$$

provided that the two series in the RHS actually converges, so the theorem follows. ■

The above three series theorem is enough for our application, but for completion let us state the full Kolmogorov's three series theorem:

**Corollary 7.29 — Three series theorem, part III.** Let $\xi_1, \xi_2, \ldots$ be independent random variables. Let further $A > 0$, and $\eta_i = \xi_i \chi_{|\xi_i| \le A}$ be the truncated random variables. Then $\sum_{i\ge1} \xi_i$ converges almost surely if and only if the following three series converges:

$$(1)\ \sum_{i\ge1} \mathbb{P}(|\xi_i| > A),\ (2)\ \sum_{i\ge1} \mathbb{E}[\eta_i] \text{ and } (3)\ \sum_{i\ge1} \mathbb{V}[\eta_i]$$

When $A = \infty$ this theorem reduces to part II.

*Proof.* We only prove the sufficiency part. Notice that part II guarantees that $\sum_{i\ge1} \eta_i$ converges almost surely, and the convergence of $\sum_{i\ge1} \mathbb{P}(|\xi_i| > A)$ means that $\mathbb{P}(|\xi_i| > A \text{ i.o.}) = 0$ by the first Borel-Cantelli lemma. Therefore, almost surely $\xi_i = \eta_i$ eventually, and in such case $\sum_{i\ge1} \xi_i$ converges. ■

We are half-way through proving Kolmogorov's strong law of large number. To complete the proof we have to prove a few technical lemmas concerning the convergence of weighted mean of the first terms of a converging sequence. You should be able to prove the following lemmas yourself using tools from elementary analysis classes.

**Lemma 7.30 — Toeplitz.** Let $\{a_n\}$ be a sequence of nonnegative numbers, $b_n = \sum_{i=1}^n a_i$, $b_1 = a_1 > 0$, and $b_n \uparrow \infty, n \to \infty$. Let $\{x_n\}_{n\ge1}$ be a sequence of numbers converging to $x$. Then

$$\frac{1}{b_n} \sum_{j=1}^n a_j x_j \to x.$$

In particular, if $a_n = 1$, then

$$\frac{x_1 + \cdots + x_n}{n} \to x.$$

**Hint.** *Observe that for all $1 \le N \le n$,*

$$\left| \frac{1}{b_n} \sum_{j=1}^n a_j x_j - x \right| = \left| \frac{1}{b_n} \sum_{j=1}^n a_j (x_j - x) \right| \le \left| \frac{1}{b_n} \sum_{j=1}^N a_j (x_j - x) \right| + \left| \frac{1}{b_n} \sum_{j=N+1}^n a_j (x_j - x). \right|$$

*Now choose $N$ such that $|x_j - x|$ is sufficiently small. Show that the tail (second term of RHS) is very small. Can you complete the proof from here?*

*Proof.* Fix $\epsilon > 0$. Choose $N_0 := N_0(\epsilon)$ such that for all $n \ge N_0$ we have $|x_j - x| < \epsilon/2$. Now choose $N_1 > N_0$ (which depends on $N_0$) such that $b_{N_1}^{-1} \sum_{j=1}^{N_0} |x_j - x| < \epsilon/2$, which exists since $|x_j - x|$ is bounded for $j = 1, \ldots, N_0$. Then for any $n > N_1$, we have

$$\begin{aligned}
\left| \frac{1}{b_n} \sum_{j=1}^n a_j x_j - x \right| &\le \left| \frac{1}{b_n} \sum_{j=1}^{N_0} a_j (x_j - x) \right| + \left| \frac{1}{b_n} \sum_{j=N+1}^n a_j (x_j - x) \right| \\
&\le \frac{1}{b_{N_1}} \left| \sum_{j=1}^N a_j (x_j - x) \right| + \left| \frac{1}{b_n} \sum_{j=N_0+1}^n a_j (x_j - x) \right| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \underbrace{\left( \frac{1}{b_n} \sum_{j=N_0+1}^n a_j \right)}_{\le 1} \le \epsilon,
\end{aligned}$$

so we have the above convergence. ■

Here we give an application of the Toeplitz lemma by the following exercise:

**Exercise 7.31** Suppose $(\xi_i)_{i\ge1}$ is a sequence of independent random variables with common mean $m$ and variance $\mathbb{V}[\xi_k] = k\eta(k)$ with the condition that $\mathbb{V}[\xi_k] \to \infty$, but $\eta(k) > 0$ and $\eta(k) \searrow 0$ as $k \to \infty$. Use Toeplitz lemma to prove that the sequence satisfies weak law of large number, that

$n^{-1}\sum_{i=1}^n \xi_i \to m$ as $n\to\infty$ in $L^2$ and in probability.

We then prove the Kronecker lemma.

**Lemma 7.32 — Kronecker.** Let $b_n$ be a sequence of positive increasing numbers, $b_n \uparrow \infty$ as $n\to\infty$ and let $\{x_n\}$ be a sequence of numbers such that $\sum x_n$ converges. Then

$$\frac{1}{b_n}\sum_{j=1}^n b_j x_j \to 0, \quad n\to\infty.$$

In particular, if $b_n = n$, $x_n = y_n/n$ and $\sum y_n/n$ converges, then

$$\frac{y_1+\cdots+y_n}{n}\to 0, \quad n\to\infty.$$

**Hint.** *We let $b_0 = S_0 = 0, S_n = \sum_{j=1}^n x_j$. In addition (for simplicity notations) we let $a_n$ as in Toeplitz lemma (lemma 7.30), such that $a_1 = b_1$ and $a_n = b_n - b_{n-1}$ for $n\geq 2$. The hint is to note Abel's summation formula:*

$$\begin{aligned}\sum_{j=1}^n b_j x_j = \sum_{j=1}^n b_j(S_j - S_{j-1}) &= b_n S_n - b_0 S_0 - \sum_{j=1}^n S_j(b_j - b_{j-1}) \\ &= b_n S_n - b_0 S_0 - \sum_{j=1}^n a_j S_j.\end{aligned}$$

*Can you draw parallels between the above formula and integration by parts?*

*Proof.* Dividing $b_n$ yields:

$$\frac{1}{b_n}\sum_{j=1}^n b_j x_j = S_n - \underbrace{\frac{b_0 S_0}{b_n}}_{\to\infty} - \frac{1}{b_n}\sum_{j=1}^n a_j S_j. \tag{7.17}$$

So when $n\to\infty$, we see that $b_n^{-1}\sum_{j=1}^n b_j x_j \overset{n\to\infty}{\to} 0$ by Toeplitz lemma. ∎

With this, we may prove Kolmogorov's lemma, which laid a foundation for some of the most important theorems relating to almost sure convergence of random series.

**Theorem 7.33 — Kolmogorov.** Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments, and let there be positive numbers $b_n$ such that $b_n \nearrow \infty$ and

$$\sum_{n\geq 1}\frac{\mathbb{V}[\xi_n]}{b_n^2} < \infty.$$

Then

$$\frac{S_n - \mathbb{E}[S_n]}{b_n} \xrightarrow{a.s.} 0.$$

In the case where $b_n = n$, this is a strong law of large numbers.

**Hint.** *Observe that*

$$\frac{S_n - \mathbb{E}[S_n]}{b_n} = \frac{1}{b_n}\sum_{i=1}^n b_k \frac{\xi_k - \mathbb{E}[\xi_k]}{b_k}$$

*and use Kronecker lemma to conclude.*

*Proof.* Note that

$$\sum_{k\geq 1}\mathbb{V}\left[\frac{\xi_k - \mathbb{E}[\xi_k]}{b_k}\right] = \sum_{k\geq 1}\frac{\mathbb{V}[\xi_k]}{b_k^2} < \infty,$$

so by Kolmogorov's three series theorem the sum $\sum_{k\geq 1}(\xi_k - \mathbb{E}[\xi_k])/b_k$ converges almost surely, and by Toeplitz theorem applied on this series the desired result holds. ∎

In the case where the variables $\xi_1, \xi_2, \dots$ are not only independent but also identically distributed, we can obtain a strong law of large numbers without requiring (as in the theorem above) the existence of the second moment, provided that the first absolute moment exists.

**Theorem 7.34 — Kolmogorov's Strong Law of Large Numbers.** Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables with $\mathbb{E}[|\xi_1|] < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{n\to\infty} m =: \mathbb{E}[\xi_1], n \to \infty.$$

*Proof.* We first recall the relationship between tail probabilities and expectation: if $\xi \geq 0$ is an integrable random variable, then

$$\mathbb{E}[\xi] = \int_{[0,\infty)} \mathbb{P}(\xi \geq x)\, \mathrm{d}x. \tag{7.18}$$

As a result, it is not hard to see that the following inequality holds:

$$\sum_{n\geq 1} \mathbb{P}(\xi \geq n) \leq \mathbb{E}[\xi] \leq 1 + \sum_{n\geq 1} \mathbb{P}(\xi \geq n), \tag{7.19}$$

and hence $\sum_{n\geq 1} \mathbb{P}(|\xi| \geq n) \leq \mathbb{E}[|\xi|] < \infty$. By first Borel-Cantelli lemma, we know that $\mathbb{P}(|\xi_n| \geq n \text{ i.o.}) = 0$, i.e. $|\xi_n| < n$ eventually, $\mathbb{P}$-almost everywhere. So by constructing truncated random variables $\tilde{\xi}_n = \xi_n \chi_{|\xi_n|<n}$, we have

$$\frac{S_n}{n} \xrightarrow{n\to\infty} 0 \text{ a.s.} \iff \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i \xrightarrow{n\to\infty} 0 \text{ a.s.}$$

Note that by dominated convergence theorem, $\mathbb{E}[\tilde{\xi}_n] \xrightarrow{n\to\infty} \mathbb{E}[\xi_1] = 0$; so by Toeplitz lemma (lemma 7.30) we have $n^{-1}\sum_{i=1^n} \mathbb{E}[\tilde{\xi}_i] \overset{n\to\infty}{\to} 0$. Therefore

$$\frac{S_n}{n} \xrightarrow{n\to\infty} 0 \text{ a.s.} \iff \frac{1}{n}\sum_{i=1}^{n} (\tilde{\xi}_i - \mathbb{E}[\tilde{\xi}_i]) \xrightarrow{n\to\infty} 0 \text{ a.s.}$$

But we note that

$$\begin{aligned}
\mathbb{V}\left[\sum_{n\geq 1} \frac{\tilde{\xi}_n - \mathbb{E}[\tilde{\xi}_n]}{n}\right] \leq \sum_{n=1}^{\infty} \frac{\mathbb{E}[\xi_1^2 \chi_{|\xi_1|<n}]}{n^2} &= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \mathbb{E}[\xi_n^2 \chi_{\{|\xi_1|\in[k-1,k)\}}] \\
&= \sum_{k=1}^{\infty} \mathbb{E}[\xi_n^2 \chi_{|\xi_1|\in[k-1,k)}] \underbrace{\left(\sum_{n=k}^{\infty} \frac{1}{n^2}\right)}_{\leq 2/k} \\
&= 2\sum_{k=1}^{\infty} \mathbb{E}\left(|\xi| \underbrace{\frac{|\xi|}{k}}_{\leq 1} \chi_{\{|\xi_1|\in[k-1,k)\}}\right) \leq 2\mathbb{E}[|\xi_1|] < \infty,
\end{aligned}$$

so by three series theorem we know that $n^{-1}\sum_{i=1}^{n}(\tilde{\xi}_i - \mathbb{E}[\tilde{\xi}_i]) \overset{n\to\infty}{0}$ $\mathbb{P}$-a.s, which completes the proof. ∎

**Remark 7.35** The theorem admits a converse in the following sense. Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables such that

$$\frac{\xi_1 + \cdots + \xi_n}{n} \to C < \infty,$$

with probability 1. Then $\mathbb{E}[|\xi_1|] < \infty$ and $C = \mathbb{E}[\xi_1]$. In fact, if $S_n/n \xrightarrow{a.s.} C$, then

$$\frac{\xi_n}{n} = \frac{S_n}{n} - \left(\frac{n-1}{n}\right)\frac{S_{n-1}}{n-1} \xrightarrow{a.s.} 0$$

and therefore $\mathbb{P}(|\xi_n| > n \ i.o.) = 0$. By the Borel-Cantelli lemma

$$\sum \mathbb{P}(|\xi_1| > n) < \infty,$$

and using one of our lemmas above, we have $\mathbb{E}[|\xi_1|] < \infty$. It then follows from the theorem that $C = \mathbb{E}[\xi_1]$. Consequently, for independent identically distributed random variables, the condition $\mathbb{E}[|\xi_1|] < \infty$ is necessary and sufficient for the convergence (with probability 1) of the ratio $S_n/n$ to a finite limit.

Here we state an application in number theory.

**Example 7.36 — Borel's theorem on normal number.** Recall the Radamacher functions on $([0,1], \mathscr{B}[0,1], \text{Leb})$ as defined in section 4.1, denoted as $\xi_1, \xi_2, \ldots$. By the strong law of large numbers, we have the following result of Borel: almost every number in $[0,1)$ is normal, in the sense that with probability 1 the proportion of zeroes and ones in its binary expansion tends to $\frac{1}{2}$, i.e.,

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k(x) \xrightarrow{n\to\infty} \frac{1}{2} \quad \text{Leb-a.s..}$$

## 7.4 Random Walk

Let us define the notion of the rates of convergence. Without loss of generality, we assume $\mu = 0$ by subtracting each random variable $\xi_i$ with its mean.

**Definition 7.37 — Rate of convergence.**

1. A function $\varphi^*(n)$ is called **upper** for $S_n$ if $S_n \le \varphi^*(n)$ for all $n$ with probability 1.

2. A function $\varphi_*(n)$ is called **lower** for $S_n$ if $S_n > \varphi_*(n)$ for infinitely many $n$ with probability 1.

If a function $\psi(n)$ is such that for all $\epsilon > 0$, $(1+\epsilon)\psi(n)$ is upper for $S_n$ and $(1-\epsilon)\psi(n)$ is lower for $S_n$, then the function $\psi(n)$ is an (optimal) rate of convergence for $S_n$.

We make the following observations:

- Consider

$$\begin{aligned}\left\{\limsup \frac{S_n}{\varphi(n)} \le 1\right\} &= \left\{\lim_{n\to\infty} \sup_{m\ge n} \frac{S_m}{\varphi(m)} \le 1\right\} \\ &= \left\{\forall \varepsilon > 0,\ \exists n_1 \ s.t.\ \sup_{m\ge n} \frac{S_m}{\varphi(m)} \le 1+\varepsilon \ \forall n \ge n_1\right\} \\ &= \left\{\forall \varepsilon > 0,\ \exists n_1 \ s.t.\ S_m \le (1+\varepsilon)\varphi(m) \ \forall m \ge n_1\right\}.\end{aligned}$$

  Therefore, if $\mathbb{P}\big(\limsup \frac{S_n}{\varphi(n)} \le 1\big) = 1$, then $(1+\varepsilon)\varphi(n)$ is upper for $S_n$ for all $\varepsilon > 0$.

- In the same way,

$$\begin{aligned}\left\{\limsup \frac{S_n}{\varphi(n)} \ge 1\right\} &= \left\{\forall \varepsilon > 0,\ \exists n_1 \ s.t.\ \sup_{m\ge n} \frac{S_m}{\varphi(m)} \ge 1-\varepsilon \ \forall n > n_1\right\} \\ &= \left\{\forall \varepsilon > 0 : S_m \ge (1-\varepsilon)\varphi(m) \text{ for infinitely many } m\right\},\end{aligned}$$

  so if $\mathbb{P}\big(\limsup \frac{S_n}{\varphi(n)} \ge 1\big) = 1$ then $(1-\varepsilon)\varphi(n)$ is lower for $S_n$ for all $\varepsilon > 0$.

**Example 7.38 — Baby Law of Iterated Logarithms.** Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables with $\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = \frac{1}{2}$. Then, since $\sum[1/(n(\log n)^{2\epsilon+1})] < \infty$ for all

$\epsilon > 0$, we have

$$\forall \epsilon > 0, \quad \frac{S_n}{\sqrt{n(\log n)^{1+2\epsilon}}} \xrightarrow{a.s.} 0$$

Let $\xi_1, \xi_2, \dots$ be a sequence of independent Bernoulli random variables with $\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = \frac{1}{2}$; let $S_n = \xi_1 + \cdots + \xi_n$. We have seen that not only $S_n/n \to 0$, but in fact

$$\frac{S_n}{\sqrt{n}\log n} \xrightarrow{a.s.} 0.$$

However, by the CLT, we have

$$\frac{S_n}{\sqrt{n}} \not\xrightarrow{d} 0.$$

**Theorem 7.39 — Law of Iterated Logarithm.** Let $\xi_1, \xi_2, \dots$ be independent identically distributed random variables with $\mathbb{E}[\xi_1] = 0$ and $\mathbb{E}[\xi_1^2] = \sigma^2 > 0$. Then

$$\mathbb{P}\left( \limsup \frac{S_n}{\psi(n)} = 1 \right) = 1,$$

where

$$\psi(n) = \sqrt{2\sigma^2 n \log\log n},$$

i.e. $\forall \varepsilon > 0$, $(1+\varepsilon)\psi$ is upper and $(1-\varepsilon)\psi$ is lower for $S_n$.

We will defer the proof to later chapters.

# Part III. Foundations of Stochastic Processes

## 8 Conditioning and Disintegration

### 8.1 Conditional Probability

#### 8.1.1 The Discrete Case

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We recall the notions of conditional probability on an event:

**Definition 8.1 — Conditional Probabilities.** Let $B \in \mathscr{F}$ be an event with $\mathbb{P}(B) > 0$. The *conditional probability of $A$ with respect to $B$* is

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad A \in \mathscr{F}. \tag{8.1}$$

By convention, if $\mathbb{P}(B) = 0$ then we set $\mathbb{P}(A \,|\, B) = 0$

**Exercise 8.2** Check that $\mathbb{P}(\cdot|B)$ is a measure on $\mathscr{F}$.

Conditioning should be relative to information one has and $\sigma$-algebra are the natural carriers or descriptions for information content. We would thus like to condition on a $\sigma$-algebra. In the example above, we could replace $B$ by its complement $B^c$ and obtain a new measure $\mathbb{P}(A|B^c)$. Now, for any $\omega \in \Omega$, we have either $\omega \in B$ or $\omega \in B^c$ so it is natural to define

$$\mathbb{P}(A|\sigma(B))(\omega) := \mathbb{P}(A|B)\chi_B(\omega) + \mathbb{P}(A|B^c)\chi_{B^c}(\omega). \tag{8.2}$$

In this way, for a fixed $\omega \in \Omega$, $\mathbb{P}(\cdot|\sigma(B))(\omega)$ is a probability measure but for a fixed $A \in \mathscr{F}$, $\mathbb{P}(A|\sigma(B))(\cdot)$ is a random variable (taking two values).

The sets $B$ and $B^c$ partition $\Omega$, and these ideas carry over to the general partition.

**Definition 8.3 — Conditional Expectation on $\sigma$-algebra generated by partition.** Let $D = \{D_1, D_2, \dots\}$ be a finite or countable partition of $\Omega$ and let $\mathscr{G} = \sigma(D)$. For $A \in \mathscr{F}$ consider the function with values

$$f(\omega) := \mathbb{P}(A|D_i) = \frac{\mathbb{P}(A \cup D_i)}{\mathbb{P}(D_i)}, \quad \text{if } \omega \in D_i. \tag{8.3}$$

This function or random variable $f$ is called the **conditional probability of $A$ given $\mathscr{G}$** and is denoted $\mathbb{P}(A|\mathscr{G})$. This is written as $\mathbb{P}(A|\mathscr{G})_\omega$ or $[\mathbb{P}(A|\mathscr{G})](\omega)$, whenever the argument $\omega$ needs to be explicitly shown.

**Exercise 8.4** Convince yourself that if $\mathscr{G} = \{\emptyset, \Omega\}$ then $f(\omega) := [\mathbb{P}(A \,|\, \mathscr{G})](\omega)$ is the constant function $f(\omega) := \mathbb{P}(A)$.

If the observer learns which element $D_i$ of the partition it is that contains $\omega$, then his new probability for the event $\omega \in A$ is $f(\omega)$. The partition $\{D_i\}$, or equivalently the $\sigma$-algebra, $\mathscr{G}$, can be regarded as an experiment, and to learn which $D_i$ it is that contains $\omega$ is to learn the outcome of the experiment. Thus $\mathbb{P}(A|\mathscr{G})$ is a function whose value on $D_i$ is the ordinary conditional probability $\mathbb{P}(A|D_i)$. This definition needs to be completed, because $\mathbb{P}(A|D_i)$ is not defined if $\mathbb{P}(D_i) = 0$. In this case $\mathbb{P}(A|\mathscr{G})$ will be taken to have any constant value on $D_i$; the value is arbitrary but must be the same over all of the set $D_i$.

**Example 8.5** Suppose that $X_0, X_1, \ldots$ is a Markov chain with state space $S$. The events

$$\{X_0 = i_0, \ldots, X_n = i_n\}, \quad i_0, \ldots, i_n \in S$$

form a finite or countable partition of $\Omega$. If $\mathscr{G}_n$ is the $\sigma$-algebra generated by this partition, then by the defining condition for Markov chains,

$$\mathbb{P}(X_{n+1} = j|\mathscr{G}_n)_\omega = \mathbb{P}(X_{n+1} = j|X_n = i_n),$$

for $\omega \in \{X_0 = i_0, \ldots, X_n = i_n\}$, $i_0, \ldots, i_n \in S$. The sets $\{X_n = i\}$ for $i \in S$ also partition $\Omega$, and they generate a $\sigma$-algebra $\mathscr{G}_n^0$ smaller than $\mathscr{G}_n$. Now

$$\mathbb{P}(X_{n+1} = j|\mathscr{G}_n^0)_\omega = \mathbb{P}(X_{n+1}|X_n = i),$$

for $\omega \in \{X_n = i\}, i \in S$, and the essence of the Markov property is that

$$\mathbb{P}(X_{n+1}|\mathscr{G}_n) = \mathbb{P}(X_{n+1}|\mathscr{G}_n^0).$$

### 8.1.2 The General Case

If $\mathscr{G}$ is the $\sigma$-algebra generated by a partition $D_1, D_2, \ldots$, then the general element of $\mathscr{G}$ is a disjoint union $D_{i_1} \cup \mathscr{D}_{i_2} \cup \cdots$, finite or countable, of certain of the $D_i$. To know which set $D_i$ it is that contains $\omega$ is the same thing as to know which sets in $\mathscr{G}$ contain $\omega$ and which do not. This second way of looking at the matter carries over to the general $\sigma-$algebra $\mathscr{G}$ contained in $\mathscr{F}$ (as always, we have the probability space $(\Omega, \mathscr{F}, \mathbb{P})$). The $\sigma-$algebra $\mathscr{G}$ will not in general come from a partition as above.

One can imagine an observer who knows for each $G$ in $\mathscr{G}$ whether $\omega \in G$ or $\omega \in G^c$. Thus the $\sigma$-algebra $\mathscr{G}$ can in principle be identified with an experiment or observation. Is is natural to try and define conditional probabilities $\mathbb{P}(A|\mathscr{G})$ with respect to the experiment $\mathscr{G}$. To do this, fix an $\mathscr{A}$ in $\mathscr{F}$ and define a finite measure $\nu$ on $\mathscr{G}$ by

$$\nu(G) = \mathbb{P}(A \cap G), \quad G \in \mathscr{G}. \tag{8.4}$$

Then $\mathbb{P}(G) = 0$ implies that $\nu(G) = 0$. The Radon-Nikodym theorem can be applied to the measures $\nu$ and $\mathbb{P}$ on the measurable space $(\Omega, \mathscr{G})$ because the first one is absolutely continuous with respect to the second. It follows that there exists a function or a random variable $f$, which is $\mathscr{G}-$measurable and integrable with respect to $\mathbb{P}$, such that

$$\mathbb{P}(A \cap G) = \nu(G) = \int_G f \mathrm{d}\mathbb{P}, \tag{8.5}$$

for all $G$ in $\mathscr{G}$. Denote this function $f$ by $\mathbb{P}(A|\mathscr{G})$. It is a random variable with two properties:

(i) $\mathbb{P}(A|\mathscr{G})$ is $\mathscr{G}$-measurable and integrable

(ii) $\mathbb{P}(A|\mathscr{G})$ satisfies the functional equation

$$\int_G \mathbb{P}(A|\mathscr{G})\mathrm{d}\mathbb{P} = \mathbb{P}(A \cap G), \quad G \in \mathscr{G}. \tag{8.6}$$

There will in general be many such random variables $\mathbb{P}(A|\mathscr{G})$, but any two of them are equal with probability 1.

If $\mathscr{G}$ is generated by a partition $D_1, D_2, \ldots$, the function $f$ defined by (8.1) is $\mathscr{G}$-measurable because $\{\omega : f(\omega) \in H\}$ is the union of those $D_i$ over which the constant value of $f$ lies in $H$. Any $G$ in $\mathscr{G}$ is a disjoint union $G = \bigcup_k D_{i_k}$, and

$$\mathbb{P}(A \cap G) = \sum_k \mathbb{P}(A|D_{i_k})\mathbb{P}(D_{i_k}), \tag{8.7}$$

so that (8.3) satisfies (8.6) as well. Thus the general definition is an extension of the one for the discrete case.

Condition $(i)$ above requires that the values of $\mathbb{P}(A|\mathscr{G})$ depend only on the sets in $\mathscr{G}$. An observer who knows the outcome of $\mathscr{G}$ viewed as an experiment knows for each $G$ in $\mathscr{G}$ whether it contains $\omega$ or not. For each $x$ he knows this in particular for the set $\{\omega' : \mathbb{P}(A|\mathscr{G})_{\omega'} = x\}$, and hence he knows in principle the functional value $\mathbb{P}(A|\mathscr{G})_\omega$ even if he does not know $\omega$ itself.

Condition $(ii)$ has a gambling interpretation. Suppose that the observer, after he has learned the outcome of $\mathscr{G}$ is offered the opportunity to bet on the event $A$ (unless $A$ lies in $\mathscr{G}$, he does not yet know whether or not it occurred). He is required to pay an entry fee of $\mathbb{P}(A|\mathscr{G})$ units and will win 1 unit if $A$ occurs and nothing otherwise. If the observer decides to bet and pays the fee, he gains $1 - \mathbb{P}(A|\mathscr{G})$ if $A$ occurs and $-\mathbb{P}(A|\mathscr{G}$ otherwise, so that his gain is

$$(1 - \mathbb{P}(A|\mathscr{G}))\chi_A + (-\mathbb{P}(A|\mathscr{G}))\chi_{A^c} = \chi_A - \mathbb{P}(A|\mathscr{G}).$$

If he declines to bet, his gain is of course 0. Suppose that he adopts the strategy of betting if $G$ occurs but not otherwise, where $G$ is some set in $\mathscr{G}$. He can actually carry out this strategy, since after learning the outcome of the experiment $\mathscr{G}$ he knows whether or not $G$ occurred. His expected gain with this strategy is his gin integrated over $G$:

$$\int_G (\chi_A - \mathbb{P}(A|\mathscr{G}))\mathrm{d}\mathbb{P}.$$

But (8.6) is exactly the requirement that this vanish for each $G$ in $\mathscr{G}$. Condition $(ii)$ requires then that each strategy be fair in the sense that the observer stands neither to win nor to lose on the average. Thus $\mathbb{P}(A|\mathscr{G})$ is just entry fee, as intuition requires.

## 8.2 Conditional Expectation

We now develop the theory of condition expectation from first principles.

**Definition 8.6 — Conditional expectation.** Suppose that $\xi$ is an integrable random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ and that $\mathscr{G}$ is a $\sigma-$algebra contained in $\mathscr{F}$. There exists a random variable $\mathbb{E}[\xi|\mathscr{G}]$, called **conditional expectation of $\xi$ given $\mathscr{G}$**, having these two properties:

- $\mathbb{E}[\xi|\mathscr{G}]$ is $\mathscr{G}$-measurable and integrable.
- For every $G \in \mathscr{G}$,

$$\mathbb{E}[\chi_G \mathbb{E}[\xi|\mathscr{G}]] = \int_G \mathbb{E}[\xi|\mathscr{G}]\mathrm{d}\mathbb{P} = \int_G \xi \mathrm{d}\mathbb{P} = \mathbb{E}[\chi_G \xi] \tag{8.8}$$

**Theorem 8.7 — Existence of conditional expectation.** Whenever $\xi$ is an integrabl random variable on $(\Omega, \mathscr{F}, \mathbb{P})$, and that $\mathscr{G}$ is a $\sigma$-algebra contained in $\mathscr{F}$, then $\mathbb{E}[\xi\,|\,\mathscr{G}]$ exists, and is unique almost everywhere.

To prove the existence of such a random variable, we recall the following generalised version of the Randon-Nikodym theorem.

**Theorem 8.8 — Radon-Nikodym Theorem.** Let $(\Omega, \mathscr{F})$ be a measure space, $\mu$ - a finite measure on $\mathscr{F}$. Let $\lambda$ be a measure on $\mathscr{F}$ a.c. with respect to $\mu$ (i.e. $\lambda(A) = 0$ whenever $\mu(A) = 0$). Then there exists an $\mathscr{F}$-measurable function $f$ such that

$$\lambda(A) = \int_A f \mathrm{d}\mu \quad \forall A \in \mathscr{F}. \tag{8.9}$$

This function is determined uniquely, up to sets of measure zero. It is called the derivative of $\lambda$ w.r.t. $\mu$ : $f = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu}$.

Consider first the case of nonnegative $\xi$. Define a measure $\nu$ on $\mathscr{G}$ by

$$\nu(G) = \int_G \xi \,\mathrm{d}\mathbb{P}.$$

This measure is finite because $\xi$ is integrable, and it is absolutely continuous with respect to $\mathbb{P}$. By the Radon-Nikodym theorem there is a function $f$, $\mathscr{G}-$measurable, such that $\nu(G) = \int_G f \mathrm{d}\mathbb{P}$. This $f$ has the two properties of our definition. If $\xi$ is not necessarily nonnegative, $\mathbb{E}[\xi^+|\mathscr{G}] - \mathbb{R}[\xi^-|\mathscr{G}]$ clearly has the required properties.

There will in general be many such random variables $\mathbb{E}[\xi|\mathscr{G}]$. Any one of them is called a version of the conditional expected value. Any two versions are equal with probability 1.

**Example 8.9 — Conditioning on $\sigma$-algbera generated by partition.** Suppose that $G_1, G_2, \ldots$ is a finite or countable partition of $\Omega$ generating the $\sigma-$algebra $\mathscr{G}$. Then $\mathbb{E}[\xi|\mathscr{G}]$ must, since it is $\mathscr{G}-$measurable, have some constant value over $G_i$, say $\alpha_i$. Then

$$\int_{G_i} \mathbb{E}[\xi|\mathscr{G}]\mathrm{d}\mathbb{P} = \int_{G_i} \alpha_i \, \mathrm{d}\mathbb{P} = \alpha_i \mathbb{P}(G_i). \tag{8.10}$$

Therefore,

$$\alpha_i \mathbb{P}(G_i) = \int_{G_i} \xi \mathrm{d}\mathbb{P}, \tag{8.11}$$

which implies

$$\mathbb{E}[\xi|\mathscr{G}]_\omega = \frac{1}{\mathbb{P}(G_i)} \int_{G_i} \xi \mathrm{d}\mathbb{P}, \quad \omega \in \mathscr{G}_i. \tag{8.12}$$

If $\mathbb{P}(G_i) = 0$, then the value of $\mathbb{E}[\xi|\mathscr{G}]$ over $G_i$ is constant but arbitrary.

To familiarise yourself with the definition of conditional expectation on a $\sigma$-algebra, we provide the following exercise.

**Exercise 8.10**

1. Convince yourself that if $\mathscr{G} = \{\emptyset, \Omega\}$ then $[\mathbb{E}[\xi \,|\, \mathscr{G}]](\omega)$ is equal to the constant function $f(\omega) := \mathbb{E}(\xi)$ almost everywhere.
2. More challenging: let's say you know that $\xi$ is independent of $\mathscr{G}$, which means that for all $B \in \mathscr{G}$ we have $\xi$ independent of $\chi_B$, then we still have $\mathbb{E}[\xi \,|\, \mathscr{G}] = \mathbb{E}[\xi]$
3. Convince yourself that $\mathbb{E}[\xi \,|\, \mathscr{F}] = \xi$ almost everywhere.
4. Show that $\mathbb{E}[\mathbb{E}[\xi \,|\, \mathscr{G}]] = \mathbb{E}[\xi]$.

**Hint.** *All problems can be solved solely by checking the definition. For example, in question 1 you check that equality (8.6) holds for $G = \phi$ and $G = \Omega$, and in question 4 you directly use the equality (8.6) for appropriate choice of $G \in \mathscr{G}$. You should also try to convince yourself that the above results are intuitive, e.g. in question 3 the conditional expectation is the random variable itself since you know everything about the random variable.*

**Solution.** *(For question 2): This is perhaps the most tricky question, but again we try to resort to the definition (8.6). Here we want to show for all $G \in \mathscr{G}$*

$$E[\xi]\mathbb{E}[\chi_G]\mathbb{E}[\chi_G\mathbb{E}[\xi]] = \mathbb{E}[\chi_G\xi] \tag{8.13}$$

*which is true by the independence assumption. This makes sense, because the $\sigma$ algebra does not give additional information about the random variable $\xi$.*

The example below links conditional expectation and conditional probabilities.

**Example 8.11 — Conditional expectation of indicator function.** For an indicator random variable $\chi_A$ the defining properties of $\mathbb{E}[\chi_A|\mathscr{G}]$ and $\mathbb{P}(A|\mathscr{G})$ coincide, therefore

$$\mathbb{E}[\chi_A|\mathscr{G}] = \mathbb{P}(A|\mathscr{G}) \tag{8.14}$$

almost surely. For a simple random variable $\xi = \sum \alpha_i \chi_{A_i}$,

$$\mathbb{E}[\xi|\mathscr{G}] = \sum \alpha_i \mathbb{P}(A_i|\mathscr{G}) \tag{8.15}$$

almost surely.

## 8.3 Properties of conditional expectation

We list some of the properties for conditional expectation. Most properties are direct applications of the definition (8.6), and we encourage you to prove them by yourselves. The proofs are, however, still included if you want extra hint to familiarise yourself with the notion of conditional expectation.

**Property 8.12 — I. Linearity.** Let $\xi, \eta$ is a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ and let $\mathscr{G}$ be a $\sigma-$algebra contained in $\mathscr{F}$. Then, almost surely,

$$\mathbb{E}[a\xi + b\eta + c\,|\,\mathscr{G}] = a\mathbb{E}[\xi\,|\,\mathscr{G}] + b\mathbb{E}[\eta\,|\,\mathscr{G}] + c \tag{8.16}$$

*Proof.* For all $G \in \mathscr{G}$, we have

$$\begin{aligned}
\mathbb{E}[\chi_G \mathbb{E}[a\xi + b\eta + c\,|\,\mathscr{G}]] &= \mathbb{E}[\chi_G(a\xi + b\eta + c)] \\
&= a\mathbb{E}[\chi_G \xi] + b\mathbb{E}[\chi_G \eta] + c\mathbb{E}[\chi_G] \\
&= a\mathbb{E}[\chi_G \mathbb{E}[\xi\,|\,\mathscr{G}]] + b\mathbb{E}[\chi_G \mathbb{E}[\eta\,|\,\mathscr{G}]] + c\mathbb{E}[\chi_G] \\
&= \mathbb{E}[\chi_G(a\mathbb{E}[\xi\,|\,\mathscr{G}] + b\mathbb{E}[\eta\,|\,\mathscr{G}] + c)]
\end{aligned}$$

■

**Property 8.13 — II. Monotonicity.** Let $\xi, \eta$ is a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\xi \leq \eta$ almost surely, and let $\mathscr{G}$ be a $\sigma-$algebra contained in $\mathscr{F}$. Then, almost surely, $\mathbb{E}[\xi\,|\,\mathscr{G}] \leq \mathbb{E}[\eta\,|\,\mathscr{G}]$. The above result also holds when $\geq$ is replaced by $>$.

*Proof.* Consider the set $G = \{\omega\,|\,\mathbb{E}[\xi\,|\,\mathscr{G}] > \mathbb{E}[\eta\,|\,\mathscr{G}]\}$. If $\mathbb{P}(G) > 0$ then $\mathbb{E}[\chi_G(\mathbb{E}[\xi\,|\,\mathscr{G}] - \mathbb{E}[\eta\,|\,\mathscr{G}])] > 0$ by our definition of $G$. But we also know that $\mathbb{E}[\chi_G(\mathbb{E}[\xi\,|\,\mathscr{G}] - \mathbb{E}[\eta\,|\,\mathscr{G}])] = \mathbb{E}[\chi_G(\xi - \eta)] \leq 0$, which is a contradiction. So $\mathbb{P}(G) = 0$, and $\mathbb{E}[\chi_G(\mathbb{E}[\xi\,|\,\mathscr{G}] - \mathbb{E}[\eta\,|\,\mathscr{G}])] = 0$. So $\mathbb{E}[(\mathbb{E}[\xi\,|\,\mathscr{G}] - \mathbb{E}[\eta\,|\,\mathscr{G}])] = \mathbb{E}[\chi_{G^c}(\mathbb{E}[\xi\,|\,\mathscr{G}] - \mathbb{E}[\eta\,|\,\mathscr{G}])] \leq 0$, and hence $\mathbb{E}[\xi\,|\,\mathscr{G}] \leq \mathbb{E}[\eta\,|\,\mathscr{G}]$ almost surely. ■

Substituting $\xi = 0$, then we know that $\eta \geq 0$ a.s. $\implies \mathbb{E}[\eta\,|\,\mathscr{G}] \geq$ a.s.

**Exercise 8.14** Prove that for all integrable $\xi$, we have $|\mathbb{E}[\xi|\mathscr{G}]| \leq \mathbb{E}[|\xi|\,|\mathscr{G}]$.

**Hint.** *Note that $\xi = \xi^+ - \xi^-$ and utilise triangle inequality. Also note that $|\xi| = \xi^+ + \xi^-$.*

We can hence prove a conditional version of Jensen inequality:

**Theorem 8.15 — Conditional Jensen's Inequality.** Suppose that $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space and $\xi$ is an integrable random variable taking values in an open interval $I \subset \mathbb{R}$. Let $g : I \to \mathbb{R}$ be convex and let $\mathscr{G}$ be a sub $\sigma-$algebra of $\mathscr{F}$. If $\mathbb{E}[|g(\xi)|] < \infty$, then

$$\mathbb{E}[\varphi(\xi)|\mathscr{G}] \geq \varphi(\mathbb{E}[\xi|\mathscr{G}]) \text{ almost surely.} \tag{8.17}$$

**Property 8.16 — III. Tower/Telescopic Properties.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $\xi$ an integrable random variable and $\mathscr{F}_1, \mathscr{F}_2$ be $\sigma-$algebras with $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \mathscr{F}$. Then almost surely

$$\mathbb{E}[\mathbb{E}[\xi|\mathscr{F}_2]|\mathscr{F}_1] = \mathbb{E}[\xi|\mathscr{F}_1] = \mathbb{E}[\mathbb{E}[\xi|\mathscr{F}_1]|\mathscr{F}_2] \tag{8.18}$$

**Hint.** *This is a generalisation of exercise 8.10.*

*Proof.* The first equality can be proven by just using the definition: consider arbitrary $G \in \mathscr{F}_1$, then $G \in \mathscr{F}_2$ and hence

$$\mathbb{E}[\chi_G \mathbb{E}[\xi|\mathscr{F}_2]] = \mathbb{E}[\chi_G \xi] = \mathbb{E}[\chi_G \mathbb{E}[\xi|\mathscr{F}_1]] \tag{8.19}$$

which establish the first equality. The second equality can be proven by using similar method in question 3 from exercise 8.10, noting $\mathbb{E}[\xi \,|\, \mathscr{F}_1]$ is $\mathscr{F}_1$ measurable then it is also $\mathscr{F}_2$ measurable. ■

**Property 8.17 — IV. On Taking Limits Under the Conditional Expectation Sign.** Let $\{\xi_n\}_{n\geq 1}$ be a sequence of extended random variables.

1. (Dominated convergence) If $|\xi_n| \leq \eta$, $\mathbb{E}[\eta] < \infty$, and $\xi_n \to \xi$ $(a.s.)$, then

   $$\mathbb{E}[\xi_n \,|\, \mathscr{G}] \xrightarrow{a.s.} \mathbb{E}[\xi \,|\, \mathscr{G}]$$

   and

   $$\mathbb{E}[|\xi_n - \xi||\mathscr{G}] \xrightarrow{a.s.} 0.$$

2. (Monotone convergence) If $\xi_n \geq \eta$, $\mathbb{E}[\eta] > -\infty$, $\xi_n \uparrow \xi$ $(a.s.)$ and $\mathbb{E}[|\xi|] < \infty$, then

   $$\mathbb{E}[\xi_n|\mathscr{G}] \uparrow \mathbb{E}[\xi|\mathscr{G}] \quad (a.s.)$$

3. If $\xi_n \leq \eta$, $\mathbb{E}[\eta] < \infty$, and $\xi_n \downarrow \xi$ $(a.s.)$, then

   $$\mathbb{E}[\xi_n|\mathscr{G}] \downarrow \mathbb{E}[\xi|\mathscr{G}] \quad (a.s.)$$

4. (Fatou) If $\xi_n \geq \eta$, $\mathbb{E}[\eta] > -\infty$, then

   $$\mathbb{E}[\liminf \xi_n|\mathscr{G}] \leq \liminf \mathbb{E}[\xi_n|\mathscr{G}] \quad (a.s.)$$

5. If $\xi_n \leq \eta$, $\mathbb{E}[\eta] < \infty$, then

   $$\mathbb{E}[\limsup \xi_n|\mathscr{G}] \leq \limsup \mathbb{E}[\xi_n|\mathscr{G}] \quad (a.s.)$$

6. (Summation) If $\xi_n \geq 0$ then

   $$\mathbb{E}\big[\sum \xi_n|\mathscr{G}\big] = \sum \mathbb{E}[\xi_n|\mathscr{G}] \quad (a.s.)$$

*Proof.*

1. Let $\zeta_n = \sup_{m\geq n} |\xi_m - \xi|$. Then $0 \leq |\zeta_n| \leq 2\eta$ and $\zeta_n \to 0$ almost surely, so by DCT we have $\mathbb{E}[\zeta_n] \overset{n\to\infty}{\to} 0$ Now by triangle inequality, one has

   $$0 \leq |\mathbb{E}[X_n|\mathscr{G}] - \mathbb{E}[X|\mathscr{G}]| \leq \mathbb{E}[|X_n - X| \,|\, \mathscr{G}] \leq \mathbb{E}[\zeta_n|\mathscr{G}].$$

   Since the sequence $\mathbb{E}[\zeta_n|\mathscr{G}](\omega)$ is decreasing on $n$ for fixed $\omega$, the sequence $\mathbb{E}[\zeta_n|\mathscr{G}](\omega)$ exhibits limits $\omega$-almost surely. To evaluate the limit, notice that

   $$0 \leq \mathbb{E}\left[\lim_{n\to\infty} \mathbb{E}[\zeta_n|\mathscr{G}]\right] \leq \mathbb{E}\left[\mathbb{E}[\zeta_n|\mathscr{G}]\right] = \mathbb{E}[\zeta_n] \overset{n\to\infty}{\to} 0,$$

   so $\lim_{n\to\infty} \mathbb{E}[\zeta_n|\mathscr{G}] = 0$ almost surely, completing the proof.

2. We note from the proof of MCT that one can set $\eta = 0$. Let $\tilde{\xi}_n = \mathbb{E}[\xi_n|\mathscr{G}]$. Then by property II (monotonicity) we have $\tilde{\xi}_n \geq 0$ almost surely and the events $A_n = \left\{\tilde{\xi}_n < \tilde{\xi}_{n-1}\right\} \in \mathscr{G}$ has $\mathbb{P}(A_n) = 0$. Let $\tilde{\xi} := \limsup_{n\to\infty} \tilde{\xi}_n$ and $A = \cup_{n\geq 2}$. Then $A \in \mathscr{G}$, $\mathbb{P}(A) = 0$ and for all $\omega \in A^c$ we know that $\tilde{\xi}_n(\omega) \nearrow \tilde{\xi}(\omega)$. We evaluate the limits by noting that for all $G \in \mathscr{G}$,

   $$\mathbb{E}\left[\tilde{\xi}\chi_G\right] = \mathbb{E}\left[\tilde{\xi}\chi_{G\cap A^c}\right] \overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbb{E}\left[\tilde{\xi}_n\chi_{G\cap A^c}\right] = \lim_{n\to\infty} \mathbb{E}\left[\xi\chi_{G\cap A^c}\right] \overset{\text{MCT}}{=} \mathbb{E}\left[\xi\chi_{G\cap A^c}\right] = \mathbb{E}\left[\xi\chi_G\right],$$

   so we see that $\tilde{\xi}$ is integrable (by taking $G \in \mathscr{G}$) and that $\xi = \tilde{\xi}$ almost surely as desired.

3. This is trivially equivalent to (2) by considering the sequence $\xi - \xi_n$.
4. This is a direct application to (2) by considering the sequence $\eta_n = \inf_{k \geq n} \xi_n$, which is increasing.
5. This is trivially equivalent to (4).
6. This is a direct application to (2).

■

**Corollary 8.18 — V. Factorisation.** Let $\xi$ and $\eta$ be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ with $\xi$, $\eta$ and $\xi\eta$ integrable. Let $\mathscr{G} \subset \mathscr{F}$ be a $\sigma$-algebra and suppose that $\eta$ is $\mathscr{G}-$measurable. Then

$$\mathbb{E}[\xi\eta|\mathscr{G}] = \eta\mathbb{E}[\xi|\mathscr{G}] \text{ almost everywhere} \tag{8.20}$$

**Hint.** *Sorry this is not an easy proof. We utilise the four step proof, see chapter 1.*

We finally have the following

**Proposition 8.19 — $L^2$ orthogonal projection.** Suppose $\xi$ is a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ such that $\mathbb{E}[\xi^2] < \infty$ (i.e. $\xi \in L^2$). Then $\mathbb{E}[X \,|\, \mathscr{G}]$ is the $L^2$ orthogonal projection onto the subspace

$$\mathscr{P} = \{g \in L^2 \,|\, g \text{ is } \mathscr{G} \text{ measurable.}\} \tag{8.21}$$

In other words, $\mathbb{E}[X \,|\, \mathscr{G}]$ is the unique minimiser of $\mathbb{E}[X - Y]^2$ for $Y \in \mathscr{G}$.

*Proof.* We first note that $\mathbb{E}[\mathbb{E}[X \,|\, \mathscr{G}]^2] \leq \mathbb{E}[X^2]$ by conditional Jensen inequality (8.17), and that $\mathbb{E}[X \,|\, \mathscr{G}]$ is $\mathscr{G}$-measurable, so $\mathbb{E}[X \,|\, \mathscr{G}] \in \mathscr{G}$. Let $Z$ be a $\mathscr{G}$-measurable function, then

$$\begin{aligned}\mathbb{E}[X - Z]^2 &= \mathbb{E}[X - \mathbb{E}[X \,|\, \mathscr{G}] + \mathbb{E}[X \,|\, \mathscr{G}] - Z]^2 \\ &= \mathbb{E}[X - \mathbb{E}[X \,|\, \mathscr{G}]^2 + \mathbb{E}[X \,|\, \mathscr{G}] - Z]^2 + 2\mathbb{E}[(X - \mathbb{E}[X \,|\, \mathscr{G}])(\mathbb{E}[X \,|\, \mathscr{G}] - Z)]\end{aligned}$$

It remains to show that the cross term is zero. Notice that $\mathbb{E}[X \,|\, \mathscr{G}] - Z$ is $\mathscr{G}$-measurable, so by tower property (exercise 8.10)

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X \,|\, \mathscr{G}])(\mathbb{E}[X \,|\, \mathscr{G}] - Z)] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X \,|\, \mathscr{G}])(\mathbb{E}[X \,|\, \mathscr{G}] - Z) \,|\, \mathscr{G}]] \\ &= \mathbb{E}[(\mathbb{E}[X \,|\, \mathscr{G}] - \mathbb{E}[X \,|\, \mathscr{G}])(\mathbb{E}[X \,|\, \mathscr{G}] - Z)] = 0\end{aligned}$$

the last equality is justified by the factorisation property (corollary 8.18), linearity and another application of tower property. So we have

$$\mathbb{E}[X - Z]^2 \geq \mathbb{E}[X - \mathbb{E}[X \,|\, \mathscr{G}]^2 + \mathbb{E}[X \,|\, \mathscr{G}] - Z]^2 \geq \mathbb{E}[X - \mathbb{E}[X \,|\, \mathscr{G}]^2 \tag{8.22}$$

with equality holds if $\mathbb{E}[X \,|\, \mathscr{G}] = Z$ almost everywhere. ■

## 8.4 Conditioning on a random variable

We define the following

**Definition 8.20** The *conditional expectation* of a random variable $\xi$ with respect to a random variable $\eta$ is defined as follows

$$\mathbb{E}[\xi|\eta] \equiv \mathbb{E}[\xi|\sigma(\eta)],$$

where $\sigma(\eta)$ is the $\sigma-$algebra generated by $\eta$.

**Theorem 8.21 — Representation of conditional expectation.** There exists a unique (almost everywhere) Borel function $g : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{E}[\xi|\eta] = g(\eta). \tag{8.23}$$

The proof is a direct application of the following lemma from measure theory

**Lemma 8.22** Let $\mu, \eta$ be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$. Then $\mu$ be $\sigma(\eta)$-measurable $\iff$ there exists a Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\mu = f(\eta)$.

**Hint.** *We apply four-step proof on $\mu$.*

*Proof.*

1. Let $\mu = \chi_A$ for some $A \in \mathscr{F}$. For $\mu$ to be $\sigma(\eta)$ measurable we must have $A \in \sigma(\eta)$. (Why?) This means there exists $B \in \mathscr{B}(\mathbb{R})$ such that $\eta^{-1}(B) = A$. We immediate see that $\mu = \chi_B(\eta)$.
2. Let $\mu$ be a simple random variable such that $\mu = \sum_{i=1}^n c_j \chi_{A_j}$, with $\{A_j\}_{j=1}^n$ partitions $\Omega$. Again we must have $A_j \in \sigma(\eta)$, so for all $j$ there exists $B \in \mathscr{B}(\mathbb{R})$ such that $\eta^{-1}(B_j) = A_j$. Then $\{B_j\}_{j=1}^n$ partitions $\eta(\Omega)$, and that $\mu = f(\eta)$ with $f = \sum_{i=1}^n c_j \chi_{B_j}(x)$.
3. We assume $\mu$ being non-negative, then $\mu$ can be approximated by a pointwise non-decreasing sequence of simple random variable which converges pointwise to $\mu$: $\mu_n \nearrow \mu$. Each $\mu_n$ can be represented in the form of $f_n(\eta)$. Choose $f = \sup_{n\geq 1} f_n$, which is Borel measurable, and for all $\omega$,
$$f(\eta(\omega)) = \sup_{n\geq 1} \mu_n(\omega) = \mu(\omega) \tag{8.24}$$
4. For general $\mu$, we decompose it into $\mu^+ - \mu^-$, and write $\mu^+ = f^+(\eta)$ and $\mu^- = f^-(\eta)$ by step 3. We let $f = f^+ - f^-$ to complete the proof.

■

**Example 8.23 — Conditional expectation of random variables with joint density.** Consider real valued random variables $X, Y$ on same probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Assume the random vector $(X, Y)$ has continuous joint density $f_{X,Y}(x, y) > 0$. Recall that $X$ has density $f_X(x) = \int_\Omega f_{X,Y}(x, y)\, dy$ and $Y$ has density $f_Y(y) = \int_\Omega f_{X,Y}(x, y)\, dx$. Assume $f_X(x), f_Y(y) > 0$ almost everywhere in $\mathbb{R}$. We want to compute $\mathbb{E}[h(X) \,|\, Y]$ for all Borel-measurable function $h$ with $\mathbb{E}[|h(X)|] < \infty$. By theorem 8.21, we know that $\mathbb{E}[h(X) \,|\, Y] = \phi(Y)$ for a unique (almost everywhere) Borel-measurable $\phi$. We claim that
$$\phi : y \mapsto \int_{\mathbb{R}} h(x) \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx \tag{8.25}$$
We can show this by utilising the definition of conditional expectation: for all $A \in \sigma(Y)$ we have
$$\mathbb{E}[\chi_A \phi(Y)] = \mathbb{E}[\chi_A h(X)] \tag{8.26}$$
Since $A \in \sigma(Y) \subseteq \mathscr{F}$, we know that $A = Y^{-1}(B)$ for some $B \in \mathscr{B}(\mathbb{R})$.

$$\begin{aligned}
\mathbb{E}[\chi_A h(X)] &= \mathbb{E}[\chi_B(Y) h(X)] \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x) \chi_B(y) f_{X,Y}(x, y)\, dy\, dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x) \chi_B(y) \frac{f_{X,Y}(x, y)}{f_Y(y)} f_Y(y)\, dy\, dx \\
&\overset{\text{(Tonelli)}}{=} \int_{\mathbb{R}} \int_{\mathbb{R}} h(x) \chi_B(y) \frac{f_{X,Y}(x, y)}{f_Y(y)} f_Y(y)\, dx\, dy \\
&= \int_{\mathbb{R}} \chi_B(y) \left( \int_{\mathbb{R}} h(x) \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx \right) f_Y(y) dy \\
&= \mathbb{E}[\chi_B(Y) \phi(Y)] \\
&= \mathbb{E}[\chi_A \phi(Y)]
\end{aligned}$$

which completes the proof. In particular, this shows us that if $C = X^{-1}(D)$
$$\mathbb{P}(X \in C \,|\, Y) = Q(Y; C), \text{ with } Q(y; C) = \int_{\mathbb{R}} \chi_C(x) \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx \tag{8.27}$$

This example can be generalised to the case when $X, Y$ are random vectors.

**Exercise 8.24 — Conditional expectation of discrete random variables.**

- Consider random variables $X, Y$ taking value in $\mathbb{N}$, and assume they have joint mass $p_{X,Y}(x, y)$ for $x, y \in \mathbb{N}$. Assume $h : \mathbb{N} \to \mathbb{R}$ such that $\mathbb{E}[|h(X)|] < \infty$. Using (8.3), verify that $\mathbb{E}[h(X) \,|\, Y] = \phi(Y)$, where for $y$ such that $p_Y(y) \neq 0$,

$$\phi(y) = \sum_{x \in \mathbb{N}} h(x) \frac{p_X(x, y)}{p_Y(y)} \tag{8.28}$$

  Notice how this formula is similar to (8.25). What value should we assign to $\phi(y)$ for $p_Y(y) = 0$ if we were to follow our convention about conditional probability on zero-probability events?

- Here is an application to the above formula. Consider random variables $Z_1, Z_2$ on $(\Omega, \mathscr{F}, \mathbb{P})$ with $Z_1 \sim \mathsf{Po}(\lambda_1)$ and $Z_2 \sim \mathsf{Po}(\lambda_2)$. Assume $p = \lambda_1/(\lambda_1 + \lambda_2)$, show that

$$\mathbb{P}[Z_1 = k \,|\, Z_1 + Z_2 = n] = \binom{n}{k} p^k (1 - p)^{n-k} \tag{8.29}$$

We will study the above examples further in the next section. Before that, let us highlight the projection property of conditional expectation.

**Proposition 8.25 — $L^2$ orthogonal projection II.** If $\mathbb{E}[\xi^2] < \infty$, then

$$\min_f \mathbb{E}[(\xi - f(\eta)^2] = \mathbb{E}[(\xi - \mathbb{E}[\xi|\eta])^2],$$

where min is over all $\sigma(\eta)$-measurable functions such that $\mathbb{E}[f^2(\eta)] < \infty$.

*Proof.* This is a direct application of proposition 8.19. ■

We can therefore obtain conditional expectation by obtaining the minimiser of $L^2$ error.

**Exercise 8.26 — Conditional expectation of normal distribution.** Consider random variables $X, Y$ such that they are jointly normally distributed:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathbf{N}_2 \left( \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right), \tag{8.30}$$

then $\mathbb{E}[Y|X] = f(X)$. Assume we know that $f(x) = ax + b$, then we know that $a, b$ are minimiser of $\mathbb{E}[Y - aX - b]^2$. Verify that

$$\mathbb{E}(Y \,|\, X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) = \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \rho\frac{\sigma_Y}{\sigma_X}X. \tag{8.31}$$

## 8.5 Regular conditional distribution

Let us revisit the above examples. We know from example 8.23 that if $(X, Y)$ has joint density $f_{X,Y}(x, y) > 0$ then we have

$$\mathbb{P}(X \in C \,|\, Y) = Q(Y; C), \text{ with } Q(y; C) = \int_{\mathbb{R}} \chi_C(x) \frac{f_{X,Y}(x, y)}{f_Y(y)}\, dx. \tag{8.32}$$

There are two ways to interpret $Q(y; C)$. If we fix a set $C \in \mathscr{F}$, then the function $Q(.; C)$ is a $\sigma(Y)$-measurable. On the other hand, if we fix $y \in Y(\Omega)$, then the set function $Q(y; .)$ is a measure on

$(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. To see this it is trivial to see that $Q(y;\varnothing) = 0$ and

$$Q(y, \mathbb{R}) = \frac{1}{f_Y(y)} \int_{\mathbb{R}} f_{X,Y}(x,y)\, dx = 1 \tag{8.33}$$

We only need to prove $\sigma$-additivity. Let $A_1, A_2, ...$ be disjoint sets in $\mathscr{B}(\mathbb{R})$, and let $A = \sqcup_{i\geq 1} A_i$, then by monotone convergence theorem

$$\begin{aligned} Q(y,A) = \int_{\mathbb{R}} \chi_A(x) \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx &= \int_{\mathbb{R}} \sum_{i\geq 1} \chi_{A_i}(x) \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx \\ &\overset{\text{(MCT)}}{=} \sum_{i\geq 1} \int_{\mathbb{R}} \chi_{A_i}(x) \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx = \sum_{i\geq 1} Q(y, A_i) \end{aligned}$$

$Q(y, C)$ is hence considered as a *(Markov) stochastic kernel*. In this chapter, we would like to prove that we can construct such transitional kernel $Q_{Y,\mathscr{G}}(\omega, C) = [\mathbb{P}(Y \in C \,|\, \mathscr{G})](\omega)$ for any random variables $Y$ and $\sigma$-algebra $\mathscr{G}$. We begin by formally defining the notion of transitional kernel

**Definition 8.27 — Transitional, Stochastic kernel.** Let $(\Omega_1, \mathscr{F}_1)$, $(\Omega_2, \mathscr{F}_2)$ be measurable spaces. A map $Q : \Omega_1 \times \mathscr{F}_2 \to [0, \infty]$ is a *transitional kernel* if

- when $A_2 \in \mathscr{F}_2$ is fixed, $Q(., A_2)$ is a $\mathscr{F}_1$-measurable function.
- when $\omega_1 \in \Omega_1$ is fixed, the set function $Q(\omega_1, .)$ is a measure on $(\Omega_2, \mathscr{F}_2)$

In addition if $\forall \omega_1 \in \Omega_1$, $Q(\omega_1, \Omega_2) = 1$ (i.e. $Q(\omega_1, .)$ is a probability measure), then $Q$ is a *(Markov) stochastic kernel*. If we have $\forall \omega_1 \in \Omega_1$, $Q(\omega_1, \Omega_2) \leq 1$, then $Q$ is *substochastic*.

**Remark 8.28** A time-homogeneous Markov chain is characterise by a family of stochastic kernels, which explains why stochastic kernels are named after Markov.

From this, we can define the notion of regular conditional distribution. Here we assume $\xi$ takes value on a general measurable space $(E, \mathscr{E})$.

**Definition 8.29 — Regular conditional distribution.** Let $\xi : (\Omega, \mathscr{F}) \to (E, \mathscr{E})$ be a random variable, and let $\mathscr{G}$ be a $\sigma$-algebra of $\Omega$ contained in $\mathscr{F}$. A regular conditional distribution of $\xi$ **given the $\sigma$-algebra** $\mathscr{G}$ is a stochastic kernel $Q : \Omega \times \mathscr{E} \to [0, \infty]$ such that for almost all $\omega \in \Omega$, we have

$$\forall B \in \mathscr{E}, \quad Q(\omega, B) = [\mathbb{P}(\xi \in B \,|\, \mathscr{G})](\omega) \tag{8.34}$$

In addition, assume $\eta : (\Omega, \sigma(\eta)) \to (E', \mathscr{E}')$ is a random variable (not necessary equal to $(E, \mathscr{E})$), then the conditional distribution of $\xi$ given the **random variable** $\eta$ is the conditional distribution of $\xi$ given on the $\sigma$-algebra $\sigma(\eta)$.

**Remark 8.30** Let $Q : \Omega \times \mathbb{E}$ be a regular conditional distribution of $\xi$ given $\eta$, then for almost all $\omega \in \Omega$, and for all $B \in \mathscr{E}$, we have

$$Q(\omega, B) = [\mathbb{P}(\xi \in B \,|\, \eta)](\omega) \tag{8.35}$$

When $Q$ is restricted so that the above equality is satisfied, then $Q(., B)$ is $\eta$ measurable, and hence $Q(\omega, B) = \tilde{Q}(\eta(\omega), B)$ for a unique function $\tilde{Q} : (E', \mathscr{E}') \times \mathscr{E} \to [0, \infty]$, such that $\tilde{Q}(., B)$ is $\mathscr{E}$ measurable for all $B \in \mathscr{E}$. We will also refer to this function $\tilde{Q}$ when talking about regular conditional distribution of $\xi$ on $\eta$.

**Example 8.31 — Continuation of Example 8.23.** Under same settings, we see that $Q(y, C)$ is a regular conditional distribution of $X$ given $Y$ in the sense as described in remark 8.30. Note that when $y$ is fixed, then the measure $Q(y, .)$ is absolutely continuous with density $f_{X|Y}(x|y) := f_{X,Y}(x, y)/f_Y(y)$. The function $f_{X|Y}(x|y)$ is known as the *conditional density* of $X$ given $Y$.

### 8.5.1 Existence of regular conditional distribution

We first assume $(E, \mathscr{E})$ to be $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ for simplicity. The main difficulty of constructing appropriate regular conditional distribution is that the conditional expectation $[\mathbb{P}(Y \in B \,|\, \mathscr{G})](\omega)$ is unique up to measure-zero set. Moreover, many properties of probability measures only hold almost surely, e.g. monotonicity and continuity from above/below. Fortunately, we only need the equality (8.34) to hold for almost all $\omega$. Our plan is therefore to construct the stochastic kernel $Q(\omega, B)$ for almost all "good" $\omega$ when all desired properties hold, and extend this kernel to other "bad" $\omega$. This will work if the set of "bad" $\omega$ is measure zero.

**Theorem 8.32 — Existence of regular conditional distribution for real-valued random variables.** Let $\xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a random variable, then for all $\sigma$-algebra $\mathscr{G} \subseteq \mathscr{F}$ we can construct a regular conditional distribution $\xi$ given $\mathscr{G}$.

**Hint.** *We only need to look at $Q(\omega, B)$ for all $B = (-\infty, r]$ for some $r$, since the half intervals $(-\infty, r]$ generates $\mathscr{B}(\mathbb{R})$. From chapter 1, we list out the desired property for a version of $F(\omega, r) = [\mathbb{P}[\xi \in (-\infty, r] \,|\, \mathscr{G}]](\omega)$ to be a valid distribution function*

1. *$F(\omega, r) \overset{r \to -\infty}{\to} 0$ and $F(\omega, r) \overset{r \to \infty}{\to} 1$*
2. *(Monotonicity) $F(\omega, r) \leq F(\omega, s)$ whenever $r \leq s$, and*
3. *(Right continuity) $F(\omega, (-\infty, r_n)) \searrow F(\omega, r)$ whenever $r_n \searrow r$.*

*So there are four cases for $F(\omega, r)$ not to be a distribution function:*

- *$\omega \in A_{r,s} := \{\omega \,|\, F(\omega, r) > F(\omega, s)\}$ for $r \leq s$.*
- *$\omega \in B'_r := \{\omega \,|\, \text{there exists a sequence } (r_n) \searrow r \text{ such that } F(\omega, r_n) \not\to F(\omega, r)\}$*
- *$\omega \in C' := \left\{\omega \,|\, F(\omega, r) \overset{r \to \infty}{\not\to} 1\right\}$*
- *$\omega \in D' := \left\{\omega \,|\, F(\omega, r) \overset{r \to -\infty}{\not\to} 0\right\}$*

*If we can show that the "bad" set of $\omega$, $G' = (\cup_{r \leq s} A_{r,s}) \cup (\cup_r B_r) \cup C \cup D$, has measure zero, then we can define $F(\omega, .)$ as followed:*

$$F(\omega, r) = \begin{cases} [\mathbb{P}(\xi \in (-\infty, r] \,|\, \mathscr{G})](\omega) & \omega \in \Omega \setminus G' \\ F_0(r) & \omega \in F \end{cases} \tag{8.36}$$

*where $F_0$ can be any appropriate distribution function. Then for all $\omega$, $F(\omega, r)$ is a distribution function, and for almost all $\omega$, $F(\omega, r) = \mathbb{P}(\xi \in (-\infty, r] \,|\, \mathscr{G})$.*

*We note from property 8.13 that $A_{r,s}$ has measure zero for any fixed $r, s$, but this does not imply that $\cup_{r \leq s} A_{r,s}$ has measure zero if we only assume $r, s$ are real, since this is an uncountable union. Moreover, even though we know that for a fixed sequence $(r_n) \searrow r$, we have $F(\omega, r_n) \searrow F(\omega, r)$ almost surely in $\omega$, we don't know whether $B_r$ itself has measure zero, since*

$$B'_r = \bigcup_{(r_n) \text{ sequence} (r_n) \searrow r} \{\omega \,|\, F(\omega, r_n) \not\to F(\omega, r)\} \tag{8.37}$$

*which is again an uncountable union of measure zero sets. Fortunately, we can refine our analysis by noting the following elementary facts in analysis: (exercise!)*

- *If $f : \mathbb{R} \to \mathbb{R}$ is a monotonic increasing function, then $f$ is right continuous iff for all $x$, there exists a sequence $(x_n) \searrow x$ such that $f(x_n) \to x$ as $n \to \infty$.*
- *$\mathbb{Q}$ is dense in $\mathbb{R}$. As a result, the half-intervals $(-\infty, a], a \in \mathbb{Q}$ generate $\mathscr{B}(\mathbb{R})$.*

*As a result, we may refine our analysis by considering $F(\omega, r)$ for all $r \in \mathbb{Q}$, then extend our analysis to $r \in \mathbb{R}$.*

*Proof.* We redefine the sets of "bad" $\omega$.

- $\omega \in A_{r,s} := \{\omega \,|\, F(\omega, r) > F(\omega, s)\}$ for $r \leq s$
- $\omega \in B_r := \{\omega \,|\, F(\omega, r + 1/n) \not\to F(\omega, r)\}$
- $\omega \in C := \left\{\omega \,|\, F(\omega, n) \overset{n\to\infty}{\not\to} 1\right\}, n \in \mathbb{Z}_{\geq 1}$
- $\omega \in D := \left\{\omega \,|\, F(\omega, -n) \overset{n\to\infty}{\not\to} 0\right\}, n \in \mathbb{Z}_{\geq 1}$.

Define

$$G = \left(\bigcup_{r\leq s, r, s\in\mathbb{Q}} A_{r,s}\right) \cup \left(\bigcup_{r\in\mathbb{Q}} B_r\right) \cup C \cup D \tag{8.38}$$

Then $G$ is a countable union of measure-zero sets, and so $G$ is also measure-zero. Hence if we define

$$F(\omega, z) = \begin{cases} \inf\left([\mathbb{P}(\xi \in (-\infty, r] \,|\, \mathscr{G})](\omega) \,:\, r \in \mathbb{Q}, r > z\right) & \omega \in \Omega \setminus G \\ F_0(z) & \omega \in F \end{cases} \tag{8.39}$$

then it is a valid distribution function for all $\omega$, and by chapter 1 we can assign a measure $Q(\omega, .)$ to each of the distribution $F(\omega, .)$. Be reminded that $Q(\omega, B)$ is itself a measurable function for all $B \in \mathscr{B}(\mathbb{R})$, we have therefore constructed a stochastic kernel.

We finally check that $Q$ is actually a regular conditional distribution, i.e. for all $A \in \mathscr{G}$,

$$\mathbb{E}[\chi_A(\omega) Q(\omega, B)] = \mathbb{E}[\chi_{A\cap\{\xi\in B\}}] = \mathbb{P}(A \cap \{\xi \in B\}) \tag{8.40}$$

Fix an arbitrary $A \in \mathscr{G}$, then we see that the above equality holds for all $B = (-\infty, r]$ where $r \in \mathbb{Q}$, and hence we can extend the equality to for all $B \in \mathscr{B}(\mathbb{R})$. ■

**Remark 8.33** We can extend the above result to any Polish spaces $(E, \mathscr{E})$, for instance $\mathbb{R}^n$ and $C^0([0, 1])$ equipped with supremum norm. To do so (theoretically) we note that there is a bijective map $\varphi : (E, \mathscr{E}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\varphi$ is $\mathscr{E}$ measurable and $\varphi^{-1}$ is $\mathscr{B}(\mathbb{R})$ measurable. We can hence construct regular conditional distribution on $\xi : (\Omega, \mathscr{F}) \to (E, \mathscr{E})$ by constructing a regular conditional distribution on $\xi' = \phi \circ \xi : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, denote as $\overline{Q} : \Omega \times \mathscr{B}(\mathbb{R}) \to [0, \infty]$. Then the stochastic kernel $Q(\omega, A) = \overline{Q}(\omega, \phi(A)) : \Omega \times \mathscr{E} \to [0, \infty]$ is indeed a regular conditional distribution of $\xi$.

### 8.5.2 Further Examples

**Example 8.34 — Continuation of exercise 8.24.** Consider random variables $Z_1, Z_2$ on $(\Omega, \mathscr{F}, \mathbb{P})$ with $Z_1 \sim \mathsf{Po}(\lambda_1)$ and $Z_2 \sim \mathsf{Po}(\lambda_2)$. Assume $p = \lambda_1/(\lambda_1 + \lambda_2)$. We have shown that the following regular conditional distribution of $Z_1$ given $Z_1 + Z_2$

$$\mathbb{P}[Z_1 = k \,|\, Z_1 + Z_2 = n] = \binom{n}{k} p^k (1-p)^{n-k} \tag{8.41}$$

is actually the probability mass of a binomial distribution $\mathsf{B}(n, p)$. As a result, we may abuse definitions to say that of $Z_1$ given $Z_1 + Z_2$ is "$\mathsf{B}(n, p)$" distributed, written as $Z_1 \,|\, Z_1 + Z_2 \sim \mathsf{B}(n, p)$.

**Exercise 8.35 — Evaluating conditional expectation from conditional distribution.** Let $Q(\omega, B)$ be a regular conditional distribution of random variable $\xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (E, \mathscr{E})$ given $\sigma$-algebra $\mathscr{G} \subseteq \mathscr{F}$. Assume $h : (E, \mathscr{E}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is a $\mathscr{E}$-measurable function such that $\mathbb{E}|h(\xi)| < \infty$.

1. Show that
$$[\mathbb{E}[h(\xi) \,|\, \mathscr{G}]](\omega) = \int h(x) Q(\omega, dx) \tag{8.42}$$
where the integral at the RHS is defined so that
$$\int \chi_B(x) Q(\omega, dx) = \int \chi_B(x)\, dQ(\omega, .) = Q(\omega, B) \tag{8.43}$$

2. Define the map $\overline{Q}$ mapping $f$ as a bounded function on $(E, \mathscr{E})$ to a function $\overline{Q}f$ on $(\Omega, \mathscr{F})$ such that:
$$\overline{Q} : f \mapsto \left( \omega \mapsto \int f(x)\, Q(\omega, dx) \right) \tag{8.44}$$
Show that $\overline{Q}f$ is bounded, and that $\overline{Q}$ is a contraction int the following sense:
$$\sup_{\omega \in \Omega} |\overline{Q}f(\omega)| \leq \sup_{x \in E} |f(x)| \tag{8.45}$$

**Hint.** *For question 1, one can prove the desired result to $h$ being simple function by using linearity of integrals. Then follow the remaining steps of the four-step proof by utilising appropriate convergence theorems.*

**Exercise 8.36 — Bayesian Analysis.**

1. Construct a random vector $(\lambda, N) : (\Omega, \mathscr{F}) \to ([0, \infty), \mathscr{B}([0, \infty))) \otimes (\mathbb{N}, 2^{\mathbb{N}})$ such that $\lambda \sim \Gamma(\alpha, \beta)$ and $N \,|\, \lambda \sim \mathsf{Po}(\lambda)$.
2. Show that $\lambda \,|\, N$ follows a Poisson distribution with a suitable parameter as a function of $N$.

**- End of Probability Theory -**

# References

[1] Klenke A. Probability theory : a comprehensive course. Universitext. London: Springer; 2008.

[2] Parthasarathy KR. Probability measures on metric spaces. Providence, R.I: AMS Chelsea Pub.; 2005 - 1967.

[3] Brezis H. Functional Analysis, Sobolev Spaces and Partial Differential Equations. 1st ed. Universitext. New York, NY: Springer New York; 2011.

[4] Kreyszig E. Introductory functional analysis with applications. Wiley classics library ed. ed. Wiley classics library. New York: Wiley; 1989 - 1978.

[5] Dudley RM. Distances of Probability Measures and Random Variables. The Annals of Mathematical Statistics. 1968;39(5):1563 1572. Available from: https://doi.org/10.1214/aoms/1177698137.

[6] Stein EM. Fourier analysis : an introduction. Princeton lectures in analysis ; 1. Princeton ;: Princeton University Press; 2003.

[7] Durrett R. Probability : theory and examples. Fifth edition. ed. Cambridge series in statistical and probabilistic mathematics ; 49. Cambridge: Cambridge University Press; 2019.

[8] Lukacs E. A Survey of the Theory of Characteristic Functions. Advances in Applied Probability. 1972;4(1):1-38. Available from: http://www.jstor.org/stable/1425805.

[9] Johnsen J. Characterization of Log-convex decay in non-selfadjoint dynamics. Electronic Research Announcements. 2018;25(0):72-86.

[10] Billingsley P. Probability and Measure. 3rd ed. John Wiley and Sons; 1995.

[11] Etemadi N. An elementary proof of the strong law of large numbers. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 1981;55(1):119-22. ID: Etemadi1981. Available from: https://doi.org/10.1007/BF01013465.