# Modern Probability Theory

Zhuoran WANG

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)
*Email address*: zwang104@connect.hkust-gz.edu.cn

ABSTRACT. This is a course handout based on **FTEC5031 Advanced Probability Theory** offered by the Hong Kong University of Science and Technology (Guangzhou).

The topics we will discuss including *measure theory*, *probability theory*, *conditional expectations* (the martingale theory will be discussed in stochastic calculus), and Markoc chain (you should read this chapter after the martingales). This handout will reference many textbooks, and the content is mathematically rigorous.

REFERENCES.

1. Çinlar, Erhan. *Probability and stochastics*. Springer, 2011.
2. Durrett, Rick. *Probability: theory and examples*. Cambridge university press, 2019.
3. Le Gall, Jean-François. *Measure theory, probability, and stochastic processes*. Cham: Springer, 2022.
4. Chung, Kai Lai. *A course in probability theory*. Elsevier, 2000.
5. Resnick, Sidney I. *A probability path*. Springer Science & Business Media, 2013.

# Contents

CHAPTER 1

# Measure and Integration

## 1. Measurable Spaces

In this notes, we don't want to discuss whether natural number system contains $0$ or not. We will write $\{0, 1, 2, \cdots\}$ and $\{1, 2, 3, \cdots\}$ both as $\mathbb{N}$, which is not confusing.

Let $E$ be a set, and $A, B$ be some subsets of $E$, we use the usual notations:

$$A \cup B, \quad A \cap B, \quad A \backslash B$$

to denote, respectively, the *union* of $A$ and $B$, the *intersection* of $A$ and $B$, and the *complement* of $B$ in $A$. In particular, we use $A^c$ to denote $E \backslash A$. For an arbitrary collection $\{A_i : i \in I\}$ of subsets of $E$, we write

$$\bigcup_{i \in I} A_i, \quad \bigcap_{i \in I} A_i$$

for the union and intersection, respectively, of all the sets $A_i$, $i \in I$. The *empty set* is denoted by $\varnothing$. Sets $A$ and $B$ are said to be *disjoint* if $A \cap B = \varnothing$. A collection of sets is said to be *disjointed* if its every element is disjoint from every other. A countable disjointed collection of sets whose union is $A$ is called a *partition* of $A$. A collection $\mathcal{C}$ of subsets of $E$ is said to be *closed under intersections* if $A \cap B$ belongs to $\mathcal{C}$ whenever $A$ and $B$ belong to $\mathcal{C}$. If the intersection of every countable collection of sets in $\mathcal{C}$ is in $\mathcal{C}$, then we say that $\mathcal{C}$ is *closed under countable intersections*. The notions of being closed under complements, unions, and countable unions, etc. are defined similarly.

**1.1. $\sigma$-algebra.** A non-empty collection $\mathcal{E}$ of subsets of $E$ is called an **algebra** on $E$ provided that it be closed under *finite unions* and *complements*. It is called a **$\sigma$-algebra** on $E$ if it is closed under *complements* and *countable unions*, that is, if

$$\text{a)} \quad A \in \mathcal{E} \quad \Rightarrow \quad E \backslash A \in \mathcal{E},$$

$$\text{b)} \quad A_1, A_2, \cdots \in \mathcal{E} \quad \Rightarrow \quad \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{E}.$$

A $\sigma$-algebra is also closed under countable intersections. The intersection of an arbitrary (countable or uncountable) family of $\sigma$-algebras on $E$ is again a $\sigma$-algebra on $E$. Given an arbitrary collection $\mathcal{C}$ of subsets of $E$, consider all the $\sigma$-algebras that contain $\mathcal{C}$ (there is at least one such $\sigma$-algebra, namely $2^E$); take the intersection of all those $\sigma$-algebras; the result is the *smallest* $\sigma$-algebra that contains $\mathcal{C}$; it is called the $\sigma$-algebra **generated** by $\mathcal{C}$ and is denoted by $\sigma(\mathcal{C})$.

If $E$ is a topological space, then the $\sigma$-algebra generated by the collection of all open subsets of $E$ is called the **Borel $\sigma$-algebra** on $E$; it is denoted by $\mathcal{B}(E)$; its elements are called Borel sets.

**1.2. $\pi$-systems and $\lambda$-systems.** A collection $\mathcal{C}$ of subsets of $E$ is called a **$\pi$-system** if it is closed under intersections. A collection $\mathcal{D}$ of subsets of $E$ is called a **$\lambda$-system** on $E$ if

$$\text{a)} \quad E \in \mathcal{D},$$

$$\text{b)} \quad A, B \in \mathcal{D} \text{ and } A \supset B \quad \Rightarrow \quad A \backslash B \in \mathcal{D},$$

$$\text{c)} \quad (A_n) \subset \mathcal{D} \text{ and } A_n \uparrow A \quad \Rightarrow \quad A \in \mathcal{D}.$$

It is obvious that a $\sigma$-algebra is both a $\pi$-system and a $\lambda$-system, and the converse will be shown next.

The following lemma shows that $\pi$-systems and $\lambda$-systems are *primitive structures* whose superpositions yield $\sigma$-algebras.

LEMMA 1.1 (Equivalent condition of $\sigma$-algebra). *A collection of subsets of $E$ is a $\sigma$-algebra if and only if it is both a $\pi$-system and a $\lambda$-system on $E$.*

PROOF. *Necessity* is obvious. Now let us consider sufficiency. Now let $\mathcal{E}$ be a collection of subsets of $E$ that is both a $\pi$-system and a $\lambda$-system.

(i). For any $A \in \mathcal{E}$, we have $A^c = E\backslash A \in \mathcal{E}$ since $\mathcal{E}$ is a $\lambda$-system. So $\mathcal{E}$ is closed under complement.

(ii). Since $\mathcal{E}$ is a $\pi$-system, then $A \cap B \in \mathcal{E}$ for all $A, B \in \mathcal{E}$. According to De Morgan's law,

$$A \cup B = (A^c \cap B^c)^c \in \mathcal{E},$$

so $\mathcal{E}$ is closed under *finite union*.

(iii). This closure of finite union can be extended to countable unions. If $(A_n) \subset \mathcal{E}$, let

$$B_1 = A_1,\ B_2 = A_2,\ \cdots,\ B_n = \bigcup_{i=1}^{n} A_i,\ \cdots,$$

then $B_n \in \mathcal{E}$ for all $n \in \mathbb{N}$ according to (ii), and $(B_n)$ is an increasing sequence with limit $\bigcup_n A_n$. Then $\mathcal{E}$ is closed under countable union because $\mathcal{E}$ is a $\lambda$-system. □

LEMMA 1.2. *Let $\mathcal{D}$ be a $\lambda$-system on $E$. Fix $D$ in $\mathcal{D}$ and let $\hat{\mathcal{D}} = \{A \in \mathcal{D} : A \cap D \in \mathcal{D}\}$. Then, $\hat{\mathcal{D}}$ is again a $\lambda$-system.*

PROOF. Just checking the conditions of $\lambda$-system one by one.

(i). $E \in \hat{\mathcal{D}}$ since $E \cap D = D \in \mathcal{D}$.

(ii). For all $A, B \in \hat{\mathcal{D}}$, $A \supset B$, we have $A \cap D, B \cap D \in \mathcal{D}$, then $(A \cap D)\backslash(B \cap D) \in \mathcal{D}$. Besides,

$$\begin{aligned}(A \cap D) \cap (B \cap D)^c &= (A \cap D) \cap (B^c \cup D^c) \\ &= (A \cap B^c \cap D) \cup (A \cap D \cap D^C) = (A \cap B^c) \cap D.\end{aligned}$$

Combine these two equations we know $A \cap B^c = A\backslash B \in \hat{\mathcal{D}}$.

(iii). Let $(A_n)$ be an increasing sequence of $\hat{\mathcal{D}}$ with limit $A$, we know $(A_n \cap D)$ is an increasing sequence of $\mathcal{D}$ with limit $A \cap D$, i.e.,

$$\bigcup_{n=1}^{\infty}(A_n \cap D) = \Big(\bigcup_{n=1}^{\infty} A_n\Big) \cap D \in \mathcal{D},$$

by the definition of $\hat{\mathcal{D}}$, we know that $\bigcup_{n\in\mathbb{N}} A_n \in \hat{\mathcal{D}}$. □

**1.3. Monotone Class Theorem.** This is a very useful tool for showing that certain collections are $\sigma$-algebras. We give it in the form found most useful in probability theory.

THEOREM 1.3 (Monotone class theorem). *If a $\lambda$-system contains a $\pi$-system, then it contains also the $\sigma$-algebra generated by that $\pi$-system.*

PROOF. Let $\mathcal{C}$ be a $\pi$-system and $\mathcal{D}$ be the *smallest* $\lambda$-system on $E$ that contains $\mathcal{C}$. The claim is that $\mathcal{D} \supset \sigma(\mathcal{C})$. Since $\sigma(\mathcal{C})$is the smallest $\sigma$-algebra containing $\mathcal{C}$, it is sufficient to show that $\mathcal{D}$ is a $\sigma$-algebra. In view of Lemma 1.2, it is thus enough to show that the $\lambda$-system $\mathcal{D}$ is also a $\pi$-system.

What we want to prove is:

$$\forall\, A, B \in \mathcal{D}: \quad A \cap B \in \mathcal{D}.$$

What we have now is:

$$\forall\, A, B \in \mathcal{C}: \quad A \cap B \in \mathcal{C}.$$

Remember that $\mathcal{C} \subset \mathcal{D}$, we need to construct a bridge to show the result.

STEP 1. Fix $B \in \mathcal{C}$ and let

$$\mathcal{D}_1 = \{A \in \mathcal{D} : A \cap B \in \mathcal{D}\}.^{1}$$

Since $\mathcal{C} \subset \mathcal{D}$, the set $B \in \mathcal{D}$; and Lemma 1.2 implies that $\mathcal{D}_1$ is a $\lambda$-system. Besides, $\mathcal{C} \subset \mathcal{D}_1$: if $A \in \mathcal{C}$ then $A \cap B \in \mathcal{C}$ since $B \in \mathcal{C}$ and $\mathcal{C}$ is a $\pi$-system. Hence, $\mathcal{D}_1$ must contain the smallest $\lambda$-system containing $\mathcal{C}$, that is, $\mathcal{D} \subset \mathcal{D}_1$.

Now, what we have is

$$\forall\, A \in \mathcal{D},\ B \in \mathcal{C}: \quad A \cap B \in \mathcal{D}.$$

STEP 2. Fix $A \in \mathcal{D}$, the collection

$$\mathcal{D}_2 = \{B \in \mathcal{D} : A \cap B \in \mathcal{D}\}$$

contains $\mathcal{C}$ according to the result showed in step 1. By Lemma 1.2, $\mathcal{D}_2$ is a $\lambda$-system. Thus, $\mathcal{D}_2$ must contain $\mathcal{D}$. In other words, $A \cap B \in \mathcal{D}$ whenever $A$ and $B$ are in $\mathcal{D}$, that is, $\mathcal{D}$ is a $\lambda$-system. □

**1.4. Measurable Spaces.** A **measurable space** is a pair $(E, \mathcal{E})$ where $E$ is a set and $\mathcal{E}$ is a $\sigma$-algebra on $E$. Then, the elements of $\mathcal{E}$ are called **measurable sets**. When $E$ is topological and $\mathcal{E} = \mathcal{B}(E)$, the Borel $\sigma$-algebra on $E$, then measurable sets are also called **Borel sets**.

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. For $A \subset E$ and $B \subset F$, we write $A \times B$ for the set of all pairs $(x, y)$ with $x$ in $A$ and $y$ in $B$; it is called the **product** of $A$ and $B$. If $A \in \mathcal{E}$ and $B \in \mathcal{F}$, then $A \times B$ is said to be a **measurable rectangle**.

We let $\mathcal{E} \otimes \mathcal{F}$ denote the $\sigma$-algebra on $E \times F$ generated by the collection of all measurable rectangles; it is called the **product $\sigma$-algebra**. The measurable space $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is called the **product** of $(E, \mathcal{E})$ and $(F, \mathcal{F})$, and the notation $(E, \mathcal{E}) \times (F, \mathcal{F})$ is used as well.

## 2. Measurable Functions

Let $E$ and $F$ be sets. A **mapping** or **function** $f$ from $E$ into $F$ is a rule that assigns an element $f(x)$ of $F$ to each $x$ in $E$, and then we write $f : E \to F$ to indicate it. If $f(x)$ is an element of $F$ for each $x$ in $E$, we also write $f : x \mapsto f(x)$ to name the mapping involved. Given a mapping $f : E \to F$ and a subset $B$ of $F$, the **inverse image** of $B$ under $f$ is

$$f^{-1}(B) = \{x \in E : f(x) \in B\}.$$

LEMMA 1.4 (Operations of inverse image). *Let $f$ be a mapping form $E$ into $F$. Then,*

$$f^{-1}(\varnothing) = \varnothing, \quad f^{-1}(F) = E, \quad f^{-1}(B \backslash C) = f^{-1}(B) \backslash f^{-1}(C),$$

$$f^{-1}\Big(\bigcup_i B_i\Big) = \bigcup_i f^{-1}(B_i), \quad f^{-1}\Big(\bigcap_i B_i\Big) = \bigcap_i f^{-1}(B_i)$$

*for all subsets $B$ and $C$ of $F$ and arbitrary collections $(B_i)_{i \in I}$ of subsets of $F$.*

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two measurable spaces. A mapping $f : E \to F$ is said to be **measurable** relative to $E$ and $F$ if $f^{-1}(B) \in \mathcal{E}$ for every $B$ in $\mathcal{F}$.

[1] To prove that a set class $\mathcal{C}$ has a certain property, we can usually construct a set class $\mathcal{D}$ which contains all the sets that satisfy this property, and show that this set class $\mathcal{D}$ contains the set class $\mathcal{C}$ to be verified.

THEOREM 1.5 (Equivalent condition of measurable functions). *In order for $f : E \to F$ to be measurable relative to $\mathcal{E}$ and $\mathcal{F}$, it is necessary and sufficient that, for some collection $\mathcal{F}_0$ that generates $\mathcal{F}$, we have $f^{-1}(B) \in \mathcal{E}$ for every $B$ in $\mathcal{F}_0$.*

PROOF. *Necessity* is obvious. Now let us consider *sufficiency*. Let $\mathcal{F}_0$ be a class of subsets of $F$ such that $\mathcal{F} = \sigma(\mathcal{F}_0)$, and suppose that $f^{-1}(B) \in \mathcal{E}$ for all $B \in \mathcal{F}_0$. We need to show that $\mathcal{F}_1 = \{B \in \mathcal{F} : f^{-1}(B) \in \mathcal{E}\} \supset \mathcal{F}_0$ contains $\mathcal{F}$, thus, it is sufficient to show $\mathcal{F}_1$ is a $\sigma$-algebra. Just checking the conditions of $\sigma$-algebra one by one.

(i). Suppose $A \in \mathcal{F}_1 \subset \mathcal{F}$, thus $A^c = F\backslash A \in \mathcal{F}$ and $f^{-1}(A) \in \mathcal{E}$. Use the result showed in Lemma 1.4, we have

$$f^{-1}(F\backslash A) = f^{-1}(F)\backslash f^{-1}(A) = E\backslash f^{-1}(A) \in \mathcal{E}$$

since $\mathcal{E}$ is a $\sigma$-algebra. By definition of $\mathcal{F}_1$, we have showed that $F\backslash A \in \mathcal{F}_1$.

(ii). Suppose $(A_n) \subset \mathcal{F}_1 \subset \mathcal{F}$, thus $\bigcup_n A_n \in \mathcal{F}$ and $f^{-1}(A_n) \in \mathcal{E}$ for all $n \in \mathbb{N}$. Then we have

$$f^{-1}\big(\bigcup_{n=1}^{\infty} A_n\big) = \bigcup_{n=1}^{\infty} f^{-1}(A_n) \in \mathcal{E}$$

using Lemma 1.4 and the condition that $\mathcal{E}$ is a $\sigma$-algebra. Thus $\mathcal{F}_1$ is a $\sigma$-algebra. □

**2.1. Composition of Functions.** Let $(E, \mathcal{E})$, $(F, \mathcal{F})$, and $(G, \mathcal{G})$ be measurable spaces. Let $f$ be a mapping from $E$ into $F$, and $g$ a mapping from $F$ into $G$. The **composition** of $f$ and $g$ is the mapping $g \circ f$ from $E$ into $G$ defined by

$$\forall\, x \in E : \quad (g \circ f)(x) = g(f(x)).$$

LEMMA 1.6 (Measurable functions of measurable functions are measurable). *If $f$ is measurable relative to $\mathcal{E}$ and $\mathcal{F}$, and $g$ relative to $\mathcal{F}$ and $\mathcal{G}$, then $g \circ f$ is measurable relative to $\mathcal{E}$ and $\mathcal{G}$.*

PROOF. Let $f$ and $g$ be measurable. For $C$ in $\mathcal{G}$, obverse that $(g \circ f)^{-1}(C) = f^{-1}\big(g^{-1}(C)\big)$, this is because: for all $C \subset G$, for $x \in E$ we have

$$\begin{aligned} x \in (g \circ f)^{-1}(C) \;&\Leftrightarrow\; (g \circ f)(x) \in C \;\Leftrightarrow\; g(f(x)) \in C \\ &\Leftrightarrow\; f(x) \in g^{-1}(C) \;\Leftrightarrow\; x \in f^{-1}\big(g^{-1}(C)\big). \end{aligned}$$

Now, $g^{-1}(C) \in \mathcal{F}$ by the measurability of $g$ and, hence, $f^{-1}\big(g^{-1}(C)\big) \in \mathcal{E}$ by the measurability of $f$. So, $g \circ f$ is measurable. □

**2.2. Numerical Functions.** Let $(E, \mathcal{E})$ be a measurable space. Recall that $\mathbb{R} = (-\infty, \infty)$, $\bar{\mathbb{R}} = [-\infty, \infty]$, $\mathbb{R}_+ = [0, \infty)$, $\bar{\mathbb{R}}_+ = [0, \infty]$. A **numerical function** on $E$ is a mapping from $E$ into $\bar{\mathbb{R}}$ or some subset of $\bar{\mathbb{R}}$. If all its values are in $\mathbb{R}$, it is said to be *real-valued*. If all its values are in $\bar{\mathbb{R}}_+$, it is said to be *positive*. A numerical function on $E$ is said to be **$\mathcal{E}$-measurable** if it is measurable relative to $\mathcal{E}$ and $\mathcal{B}(\bar{\mathbb{R}})$, the latter denoting the Borel $\sigma$-algebra on $\bar{\mathbb{R}}$ as usual. If $E$ is topological and $\mathcal{E} = \mathcal{B}(E)$, then $\mathcal{E}$-measurable functions are called **Borel functions**.

The following theorem is a corollary of Theorem 1.5 using the fact that $\mathcal{B}(\bar{\mathbb{R}})$ is generated by the collection of intervals $[-\infty, r]$ with $r$ in $\mathbb{R}$. No proof seems needed.

THEOREM 1.7 (Equivalent condition of numerical measurable functions). *A mapping $f : E \to \bar{\mathbb{R}}$ is $\mathcal{E}$-measurable if and only if, for every $r$ in $\mathbb{R}$, $f^{-1}([-\infty, r]) \in \mathcal{E}$.*

REMARK 1.8. The theorem remains true if $[-\infty, r]$ is replaced by $[-\infty, r)$ or by $[r, \infty]$ or by $(r, \infty]$. In the particular case $f : E \to F$, where $F$ is a countable subset of $\bar{\mathbb{R}}$, the mapping $f$ is $\mathcal{E}$-measurable if and only if $f^{-1}(\{a\}) = \{x \in E : f(x) = a\}$ is in $\mathcal{E}$ for every $a$ in $F$. †

**2.3. Positive and Negative Parts.** For $a$ and $b$ in $\mathbb{R}$ we write $a \vee b$ for the *maximum* of $a$ and $b$, and $a \wedge b$ for the *minimum*. Let $f$ be a numerical function on $E$. Then,

$$f^+ = f \vee 0, \quad f^- = -(f \wedge 0)$$

are both positive functions and $f = f^+ - f^-$. The function $f^+$ is called the **positive part** of $f$, and $f^-$ the **negative part**.

LEMMA 1.9 (Another equivalent condition of measurable functions). *The function $f$ is $\mathcal{E}$-measurable if and only if $f^+$ and $f^-$ are.*

PROOF. We will prove a more general result. See Theorem 1.10. First, consider the *necessity*. If $f$ is $\mathcal{E}$-measurable, then $f^+$ and $f^-$ are by Theorem 1.10. Second, if $f^+$ and $f^-$ are $\mathcal{E}$-measurable, then $f = f^+ - f^-$ is measurable by the measurability of sum of measurable functions proved in Theorem 1.10. □

**2.4. Indicators and Simple Functions.** Let $A \subset E$. Its indicator, denoted by $\mathbb{1}_A$, is the function defined by

$$\begin{aligned} \mathbb{1}_A(x) &= 1 \quad \text{if } x \in A, \\ &= 0 \quad \text{if } x \notin A. \end{aligned}$$

We write simply $\mathbb{1}$ for $\mathbb{1}_E$. Obviously, $\mathbb{1}_A$ is $\mathcal{E}$-measurable if and only if $A \in \mathcal{E}$. A function $f$ on $E$ is said to be **simple** if it has the form

$$f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$$

for some $n \in \mathbb{N}$, real numbers $a_1, \cdots, a_n$, and measurable sets $A_1, \cdots, A_n \in \mathcal{E}$. In the case where $a_1, \cdots, a_n$ are distinct real numbers and $\{A_1, \cdots, A_n\}$ is a measurable partition of $E$, this representation is called the **canonical form** of the simple function $f$. It is immediate Remark (b) of Theorem 1.7 applied to the canonical form that *every simple function is $\mathcal{E}$-measurable*. Finally, if $f$ and $g$ are measurable, then so are

$$f + g, \quad f - g, \quad fg, \quad f/g, \quad f \vee g, \quad f \wedge g,$$

except that in the case of $f/g$ one should make sure that $g$ is *nowhere zero*.

**2.5. Operations of Measurable Functions.** Next, we will show that measurable functions remain closed under *finite algebraic operations* and *countable limit operations*. The main difficulty is that $\bar{\mathbb{R}}$ is not a domain, e.g., $(f + g)(x)$ is undefined at points like $f(x) = \infty$, $g(x) = -\infty$. So the next theorem starts with the assumption that the algebraic operations of measurable functions can be defined.

THEOREM 1.10 (Measurable functions remain closed under finite algebraic operations). *If $f$ and $g$ are both $\mathcal{E}$-measurable functions and $k \in \mathbb{R}$, then each of the functions:*

$$f + k, \quad kf, \quad f + g, \quad f^2, \quad fg, \quad 1/f \text{ (where } (1/f)(x) = +\infty \text{ if } f(x) = 0),$$
$$f \vee g, \quad f \wedge g, \quad f_+, \quad f_-, \quad |f|;$$

*which is defined, is measurable.*

Obviously, the above theorem can be generalized to the case of finitely many algebraic operations on finitely many measurable functions.

PROOF. Obviously, the measurability of $f+k$ and $kf$ can be derived from Theorem 1.7 directly.

STEP 1. Prove $f+g$ is measurable. For all $r\in\mathbb{R}$, what we want to prove is

$$(f+g)^{-1}\big((r,\infty]\big)=\bigcup_{s\in\mathbb{Q}} f^{-1}\big((s,\infty]\big)\cap g^{-1}\big((r-s,\infty]\big).$$

*First*, the left hand side is contained in the right hand side. Suppose $x\in(f+g)^{-1}\big((r,\infty]\big)$, i.e., $f(x)>r-g(x)$, which means that $\big(r-g(x),f(x)\big)\neq\varnothing$. Thus, there exist $s\in\mathbb{Q}$ such that $s\in\big(r-g(x),f(x)\big)$ since $\mathbb{Q}$ is *dense* in $\mathbb{R}$. This means that

$$\begin{aligned} s<f(x) \quad &\Rightarrow \quad x\in f^{-1}((s,\infty]),\\ r-g(x)<s \quad &\Rightarrow \quad x\in g^{-1}((r-s,\infty]).\end{aligned}$$

So $x\in f^{-1}\big((s,\infty]\big)\cap g^{-1}\big((r-s,\infty]\big)\subset\bigcup_{s\in\mathbb{Q}} f^{-1}\big((s,\infty]\big)\cap g^{-1}\big((r-s,\infty]\big)$.

*Second*, prove the reverse inclusion relation. For some $s\in\mathbb{Q}$, if $x\in f^{-1}\big((s,\infty]\big)\cap g^{-1}\big((r-s,\infty]\big)$, then $s<f(x)$ and $r-s<g(x)$, so $f(x)+g(x)>r$, i.e. $x\in(f+g)^{-1}\big((r,\infty]\big)$.

*Finally*, since $f^{-1}\big((s,\infty]\big)\in\mathcal{E}$, $g^{-1}\big((r-s,\infty]\in\mathcal{E}\big)$ and $\mathbb{Q}$ is countable, we know that $(f+g)^{-1}\big((r,\infty]\big)\in\mathcal{E}$, i.e., $f+g$ is $\mathcal{E}$-measurable.

STEP 2. $f^2$ and $1/f$ are measurable based on the equations below and the measurability of $f$:

$$\begin{aligned}(f^2)^{-1}([-\infty,r]) &= \varnothing && \text{if } r<0,\\ &= f^{-1}(\{0\}) && \text{if } r=0,\\ &= f^{-1}([-r,r]) && \text{if } r>0.\\ (1/f)^{-1}([-\infty,r]) &= f^{-1}([1/r,0)) && \text{if } r<0,\\ &= f^{-1}([-\infty,0)) && \text{if } r=0,\\ &= f^{-1}([-\infty,0))\cup f^{-1}([1/r,\infty)) && \text{if } r>0.\end{aligned}$$

STEP 3. $f\vee g$ and $f\wedge g$ are $\mathcal{E}$-measurable based on the following equations:

$$\begin{aligned}(f\vee g)^{-1}([-\infty,r]) &= f^{-1}([-\infty,r])\cap g^{-1}([-\infty,r]),\\ (f\wedge g)^{-1}([-\infty,r]) &= f^{-1}([-\infty,r])\cup g^{-1}([-\infty,r]).\end{aligned}$$

STEP 4. The measurability of all the remaining functions follows from the above conclusion, using the following equation:

$$fg=\frac{1}{2}\big[(f+g)^2-f^2-g^2\big],\ f_+=f\vee 0,\ f_-=(-f)\vee 0,\ |f|=f_++f_-.$$

This completes the proof. □

**2.6. Limits of Sequence of Functions.** Let $(f_n)$ be a sequence of numerical functions on $E$. The functions

$$\inf f_n,\quad \sup f_n,\quad \liminf f_n,\quad \limsup f_n \tag{1.1}$$

are defined on $E$ pointwise. If $(f_n)$ is increasing (or decreasing), then $\lim f_n$ exists and is equal to $\sup f_n$ (or $\inf f_n$).

THEOREM 1.11 (the class of measurable functions is closed under countable limits). *Let $(f_n)$ be a sequence of $\mathcal{E}$-measurable functions. Then, each one of the four functions in Equation (1.1) is $\mathcal{E}$-measurable. Moreover, if $(f_n)$ converges, $\lim f_n$ is $\mathcal{E}$-measurable.*

PROOF. STEP 1. Show that $f = \sup f_n$ is $\mathcal{E}$-measurable. $\forall\, x \in E$ and $\forall\, r \in \mathbb{R}$, we have

$$f^{-1}([-\infty, r]) = \{x : f(x) \leqslant r\} = \bigcap_n \{x : f_n(x) \leqslant r\} = \bigcap_n f_n^{-1}([-\infty, r]).$$

Then $f = \sup f_n$ is $\mathcal{E}$-measurable since $f_n$'s are $\mathcal{E}$-measurable and $\mathcal{E}$ is a $\sigma$-algebra.

STEP 2. Obverse that $\inf f_n = -\sup(-f_n)$, and

$$\liminf f_n = \sup_m \inf_{n \geqslant m} f_n, \quad \limsup f_n = \inf_m \sup_{n \geqslant m} f_n$$

since $\inf_{n \geqslant m} f_n$ and $\sup_{n \geqslant m} f_n$ are increasing and decreasing, respectively. Thus we have proved the theorem. □

We now give a more precise characterization of the *upper and lower limits*. We will use this lemma to give an example of Borel-Cantelli lemma in Chapter 3.

LEMMA 1.12 (Characterization of the upper and lower limits). *Let $f_n$ be a sequence of $\mathcal{E}$-measurable functions, then*

1. *the upper limit of $f_n$ equals to $L^+$, i.e., $\limsup f_n = L^+$, if and only if, for all $\varepsilon > 0$,*
   (i) *there exists an integer $N$ such that $f_n < L^+ + \varepsilon$ for all $n \geqslant N$. In other words, for every $\varepsilon > 0$, the elements of the sequence $(f_n)$ are eventually less than $L^+ + \varepsilon$;*
   (ii) *for every integer $N$, there exists an $n \geqslant N$ such that $f_n > L^+ - \varepsilon$. In other words, for every $\varepsilon > 0$, the elements of the sequence $(f_n)$ exceed $L^+ - \varepsilon$ infinitely often.*
2. *the lower limit of $f_n$ equals to $L^-$, i.e., $\liminf f_n = L^-$, if and only if, for all $\varepsilon > 0$,*
   (i) *there exists an integer $N$ such that $f_n > L^- - \varepsilon$ for all $n \geqslant N$.*
   (ii) *for every integer $N$, there exists an $n \geqslant N$ such that $f_n < L^- + \varepsilon$.*

PROOF. We only need to prove (1). First consider the *necessity*. Fix an arbitrary positive $\varepsilon$, then $L^+ + \varepsilon > \limsup f_n$, i.e. $L^+ + \varepsilon > \inf_{m \geqslant 1} f_m^+$, where $f_m^+ := \sup_{n \geqslant m} f_n$. By the definition of infimum, there exists an integer $N$ such that $L^+ + \varepsilon > f_N^+ = \sup_{n \geqslant N} f_n$. By the definition of supremum, we have $L^+ + \varepsilon > a_n$ for all $n \geqslant N$. That is the result we want to show in (i). Next, consider (ii). Since $L^+ - \varepsilon < \limsup f_n$, that is, $L^+ - \varepsilon < \inf_{m \geqslant 1} f_m^+$. Using the definition of infimum again, fix any integer $N \geqslant 1$, we have $L^+ - \varepsilon < f_N^+ = \sup_{n \geqslant N} f_n$. Again, using the definition of supremum, there exists an integer $\ell$ such that $f_\ell > L^+ - \varepsilon$.

Now let us consider the *sufficiency*. Suppose that (i) and (ii) holds for all $\varepsilon > 0$, we want to show that $\limsup f_n = L^+$. It suffices to show that

$$L^+ - \varepsilon \leqslant \inf_{m \geqslant 1} \sup_{n \geqslant m} f_n = \limsup f_n \leqslant L^+ + \varepsilon.$$

From (i), since there is an integer $N$ such that $f_n < L^+ + \varepsilon$ for all $n \geqslant N$, so $L^+ + \varepsilon$ is an upper bounded of the sequence $(f_n)_{n \geqslant N}$, thus $\sup_{n \geqslant N} f_n \leqslant L^+ + \varepsilon$. Besides, since $(\sup_{n \geqslant m} f_n)_{m \geqslant 1}$ is a monotone decreasing sequence, take infimum and we get

$$\inf_{m \geqslant 1} \big( \sup_{n \geqslant m} f_n \big) \leqslant \sup_{n \geqslant N} g_n \leqslant L^+ + \varepsilon.$$

From (ii), for every integer $N$, there exists an integer $n \geqslant N$ such that $f_n > L^+ - \varepsilon$, by the definition of supremum, we get $\sup_{n \geqslant N} f_n \geqslant L^+ - \varepsilon$ holds for all $N$ in $\mathbb{N}$. So $L^+ - \varepsilon$ is a lower bound of $(\sup_{n \geqslant N} f_n)_{N \geqslant 1}$, so

$$\inf_{n \geqslant N} \big( \sup_{n \geqslant N} f_n \big) \geqslant L^+ - \varepsilon.$$

Thus we have the desired result. □

**2.7. Approximation of Measurable Functions.** We start by approximating the identity function on $\bar{\mathbb{R}}_+$ by an increasing sequence of simple functions of a specific form (*dyadic functions*).

LEMMA 1.13 (Dyadic functions). *For each $n \in \mathbb{N}$, let*

$$d_n(r) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{[\frac{k-1}{2^n}, \frac{k}{2^n})}(r) + n\mathbb{1}_{[n,\infty]}(r), \quad r \in \bar{\mathbb{R}}_+.$$

*Then, each $d_n$ is an increasing right-continuous simple function on $\bar{\mathbb{R}}_+$, and $d_n(r)$ increases to $r$ for each $r$ in $\bar{\mathbb{R}}_+$ as $n \to \infty$.*

PROOF. It is obvious that $d_n$ is an increasing right-continuous simple function on $\bar{\mathbb{R}}_+$, and we only need to prove the second claim. In fact, the definition of $d_n$ implies that

$$\forall\, n \in \mathbb{N}: \quad d_n \leqslant d_{n+1}, \quad \text{and} \quad |d_n(r) - r| \leqslant \frac{1}{2^n} \text{ for all } r \in [0, n].$$

Thus $d_n \uparrow r$ as $n \to \infty$. □

The following theorem is important: it reduces many a computation about measurable functions to a computation about simple functions followed by limit taking.

THEOREM 1.14 (Approximation of measurable functions). *A positive function on $E$ is $\mathcal{E}$-measurable if and only if it is the limit of an increasing sequence of positive simple functions.*

PROOF. *Sufficiency* is immediate from Theorem 1.10. Consider the *necessity*. Let $f : E \to \bar{\mathbb{R}}_+$ be $\mathcal{E}$-measurable. We are to show that there is a sequence $(f_n)$ of positive simple functions increasing to $f$. To that end, let $(d_n)$ be as in the preceding lemma and put $f_n = d_n \circ f$. Then, for each $n$, the function $f_n$ is $\mathcal{E}$-measurable, since it is a measurable function of a measurable function. Also, it is positive and takes only finitely many values, because $d_n$ is so. Thus, each $f_n$ is positive and simple. Moreover, since $d_n(r)$ increases to $r$ for each $r$ in $\bar{\mathbb{R}}_+$ as $n \to \infty$, we have that $f_n(x) = d_n(f(x))$ increases to $f(x)$ for each $x$ in $E$ as $n \to \infty$. □

**2.8. Monotone Classes of Functions.** Let $\mathcal{M}$ be a collection of numerical functions on $E$. We write $\mathcal{M}_+$ for the subcollection consisting of *positive functions* in $\mathcal{M}$, and $\mathcal{M}_b$ for the subcollection of *bounded functions* in $\mathcal{M}$. The collection $\mathcal{M}$ is called a **monotone class** provided that it includes the *constant function* 1, and $\mathcal{M}_b$ is a *linear space* over $\mathbb{R}$, and $\mathcal{M}_+$ is closed under increasing limits; more explicitly, $\mathcal{M}$ is a monotone class if

$$\begin{aligned} &\text{a)} \quad 1 \in \mathcal{M}, \\ &\text{b)} \quad f, g \in \mathcal{M}_b \text{ and } a, b \in \mathbb{R} \quad \Rightarrow \quad af + bg \in \mathcal{M}, \\ &\text{c)} \quad (f_n) \subset \mathcal{M}_+,\ f_n \uparrow f \quad \Rightarrow \quad f \in \mathcal{M}. \end{aligned} \tag{1.2}$$

The next theorem is used often to show that a certain property holds for all $\mathcal{E}$-measurable functions. It is a version of Theorem 1.3, it is called the *monotone class theorem for functions*.

THEOREM 1.15 (Monotone class theorem for functions). *Let $\mathcal{M}$ be a monotone class of functions on $E$. Suppose, for some $\pi$-system $\mathcal{C}$ generating $\mathcal{E}$, that $\mathbb{1}_A \in \mathcal{M}$ for every $A$ in $\mathcal{C}$. Then, $\mathcal{M}$ includes all positive $\mathcal{E}$-measurable functions and all bounded $\mathcal{E}$-measurable functions.*

PROOF. We start by showing that $\mathbb{1}_A \in \mathcal{M}$ for every $A$ in $\mathcal{E}$. To this end, let

$$\mathcal{D} = \{A \in \mathcal{E} : \mathbb{1}_A \in \mathcal{M}\}.$$

Using the Eq. (1.2), it is easy to check that $\mathcal{D}$ is a $\lambda$-system. Since $\mathcal{D} \supset \mathcal{C}$ by assumption, and since $\mathcal{C}$ is a $\pi$-system that generates $\mathcal{E}$, we must have $\mathcal{D} \supset \mathcal{E}$ by the monotone class theorem 1.3. So, $\mathbb{1}_A \in \mathcal{M}$ for all $A$ in $\mathcal{E}$.

Therefore, in view of the property (1.2, b), $\mathcal{M}$ includes all simple functions.

Let $f$ be a positive $\mathcal{E}$-measurable function. By Theorem 1.14, there exists a sequence of positive simple functions $f_n$ increasing to $f$. Since each $f_n$ is in $\mathcal{M}_+$ by the preceding step, the property (1.2, c) implies that $f \in \mathcal{M}$.

Finally, let $f$ be a bounded $\mathcal{E}$-measurable function. Then $f^+$ and $f^-$ are in $\mathcal{M}$ by the preceding step and are bounded obviously. Thus, by Eq. (1.2, b), we conclude that $f = f^+ - f^- \in \mathcal{M}$. □

**2.9. Standard Measurable Spaces.** Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let $f$ be a *bijection* from $E$ onto $F$, and let $\hat{f}$ denote its inverse. Then, $f$ is said to be an **isomorphism** of $(E, \mathcal{E})$ and $(F, \mathcal{F})$ if $f$ is measurable relative to $\mathcal{E}$ and $\mathcal{F}$ and $\hat{f}$ is measurable relative to $\mathcal{F}$ and $\mathcal{E}$. $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are said to be **isomorphic** if there exists an *isomorphism* between them.

A measurable space $(E, \mathcal{E})$ is said to be **standard** if it is *isomorphic* to $(F, \mathcal{B}(F))$ for some *Borel subset* $F$ of $\mathbb{R}$. The class of standard spaces is surprisingly *large* and includes almost all the spaces we shall encounter. Here are some examples: The spaces $\mathbb{R}$, $\mathbb{R}^d$, $\mathbb{R}^\infty$ together with their respective Borel $\sigma$-algebras are standard measurable spaces. If $E$ is a *complete separable metric space*, then $(E, \mathcal{B}(E))$ is standard. If $E$ is a *separable Banach space*, or more particularly, a *separable Hilbert space*, then $(E, \mathcal{B}(E))$ is standard.

Clearly, $[0, 1]$ and its Borel $\sigma$-algebra form a *standard* measurable space; so do $\{1, 2, \cdots, n\}$ and its discrete $\sigma$-algebra; so do $\mathbb{N} = \{1, 2, \cdots\}$ and its discrete $\sigma$-algebra. *Every standard measurable space is isomorphic to one of these three* (this is a deep result).

**2.10. Product Spaces and Sections.** We introduced product spaces and the associated product $\sigma$-algebra in Section 1. Now, let us consider *measurable functions* defined on product spaces.

LEMMA 1.16 (Product spaces). *Let $(E, \mathcal{E})$, $(F, \mathcal{F})$ and $(G, \mathcal{G})$ be measurable spaces. Let $f : E \to F$ be measurable relative to $\mathcal{E}$ and $\mathcal{F}$, and let $g : E \to G$ be measurable relative to $\mathcal{E}$ and $\mathcal{G}$. Define $h : E \to F \times G$ by*

$$h(x) = (f(x), g(x)), \quad x \in E.$$

*Show that $h$ is measurable relative to $\mathcal{E}$ and $\mathcal{F} \otimes \mathcal{G}$.*

PROOF. Remember that $\mathcal{F} \otimes \mathcal{G} = \sigma(\mathcal{F} \times \mathcal{G})$, where $\mathcal{F} \times \mathcal{G} := \{A \times B : A \in \mathcal{F},\ B \in \mathcal{G}\}$ is the collection of all measurable rectangles. Now, for all $M = M_1 \times M_2 \in \mathcal{F} \times \mathcal{G}$,

$$\begin{aligned} \{x \in E : h(x) \in M\} &= \{x \in E : (f(x), g(x)) \in M_1 \times M_2\} \\ &= \{x \in E : f(x) \in M_1\} \cap \{x \in E : g(x) \in M_2\} \in \mathcal{E} \end{aligned}$$

since $f$ and $g$ are measurable. Hence, by Theorem 1.5, we get the desired result. □

Lemma 1.16 gives the measurability of a vector function whose elements are all measurable. Now we can discuss the measurability of sections, which are every important in Section 6.

EXAMPLE 1.17 (Sections). Let $f : E \times F \to G$ be measurable relative to $\mathcal{E} \otimes \mathcal{F}$ and $\mathcal{G}$. Then, for fixed $x_0$ in $E$, the mapping $h : y \mapsto f(x_0, y)$ is measurable relative to $\mathcal{F}$ and $\mathcal{G}$. The mapping $h$ is called the **section** of $f$ at $x_0$. †

PROOF. Note that $h = f \circ g$ where $g : F \to E \times F$ is defined by $g(y) = (x_0, y)$. Since the constant and identity functions are both measurable, then $g(y)$ is measurable relative to $\mathcal{F}$ and $\mathcal{E} \otimes \mathcal{F}$ according to Lemma 1.16. Then, by Lemma 1.6, $h$ is measurable since $f$ and $g$ are both measurable. □

## 3. Measures

DEFINITION 1.18 (Measures). Let $(E, \mathcal{E})$ be a measurable space, that is, $E$ is a set and $\mathcal{E}$ is a $\sigma$-algebra on $E$. A **measure** on $(E, \mathcal{E})$ is a mapping $\mu : \mathcal{E} \to \bar{\mathbb{R}}_+$ such that

a) $\mu(\varnothing) = 0$,

b) $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ for every disjoint sequence $(A_n)$ in $\mathcal{E}$.

The latter condition is called **countable additivity**. Note that $\mu(A)$ is always *positive* and can be $+\infty$; the number $\mu(A)$ is called the measure of $A$. A **measure space** is a triplet $(E, \mathcal{E}, \mu)$, where $(E, \mathcal{E})$ is a *measurable space* and $\mu$ is a *measure* on it.

**3.1. Examples.** First, we will give some typical measures that we are frequently used in probability theory.

EXAMPLE 1.19 (Dirac measures). Let $(E, \mathcal{E})$ be a measurable space, and let $x$ be a fixed point of $E$. For each $A$ in $\mathcal{E}$, put

$$\begin{aligned} \delta_x(A) &= 1 \quad \text{if } x \in A, \\ &= 0 \quad \text{if } x \notin A. \end{aligned}$$

Then, $\delta_x$ is a measure on $(E, \mathcal{E})$. It is called the **Dirac measure** sitting at $x$. †

EXAMPLE 1.20 (Counting measures). Let $(E, \mathcal{E})$ be a measurable space. Let $D$ be a fixed subset of $E$. For each $A$ in $\mathcal{E}$, let $\nu(A)$ be the number of points in $A \cap D$. Then, $\nu$ is a measure on $(E, \mathcal{E})$. Such $\nu$ are called *counting measures*. Often, the set $D$ is taken to be countable, in which case

$$\nu(A) = \sum_{x \in D} \delta_x(A), \quad A \in \mathcal{E}.$$

EXAMPLE 1.21 (Discrete measures). Let $(E, \mathcal{E})$ be a measurable space. Let $D$ be a countable subset of $E$. For each $x$ in $D$, let $m(x)$ be a positive number. Define

$$\mu(A) = \sum_{x \in D} m(x)\delta_x(A), \quad A \in \mathcal{E}.$$

Then, $\mu$ is a measure on $(E, \mathcal{E})$. Such measures are said to be *discrete*. We may think of $m(x)$ as the mass attached to the point $x$, and then $\mu(A)$ is the mass on the set $A$. In particular, if $(E, \mathcal{E})$ is a discrete measurable space, then every measure $\mu$ on it has this form. †

EXAMPLE 1.22 (Lebesgue measures). A measure $\mu$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called the **Lebesgue measure** on $\mathbb{R}$ if $\mu(A)$ is the length of $A$ for every interval $A$. As with most measures, it is impossible to display $\mu(A)$ for every Borel set $A$, but one can do integration with it, which is the main thing measures are for. Similarly, the Lebesgue measure on $\mathbb{R}^2$ is the "area" measure, on $\mathbb{R}^3$ the "volume", etc. We shall write $Leb$ for them. Also note the harmless vice of saying, for example, Lebesgue measure on $\mathbb{R}^2$ to mean Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. †

**3.2. Some Properties.** For general measures, we have the following properties:

THEOREM 1.23 (Some properties). *Let $\mu$ be a measure on a measurable space $(E, \mathcal{E})$. Then, the following hold for all measurable sets $A$, $B$ and $A_n$ ($n \in \mathbb{N}$):*

(i) *Finite additivity.* $A \cap B = \varnothing \Rightarrow \mu(A \cup B) + \mu(A) + \mu(B)$.
(ii) *Monotonicity.* $A \subset B \Rightarrow \mu(A) \leqslant \mu(B)$.
(iii) *Sequential continuity.* $A_n \uparrow A \Rightarrow \mu(A_n) \uparrow \mu(A)$.
(iv) *Boole's inequality.* $\mu(\bigcup_n A_n) \leqslant \sum_n \mu(A_n)$.

PROOF. *Monotonicity* follows from finite additivity and the positivity of $\mu$: for $A \subset B$, write $B$ as the union of disjoint sets $A$ and $B \backslash A$, hence $\mu(B) = \mu(A) + \mu(B \backslash A) \geqslant \mu(A)$.

*Sequential continuity* follows from countable additivity: Suppose $A_n \uparrow A$ and define $B_1 = A_1$, $B_n = A_n \backslash A_{n-1}$ for all $n \geqslant 2$. Then $B_n$'s are disjoint and $\bigcup_n B_n = \bigcup_n A_n = A$, so

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

*Boole's inequality*. Obverse that $\mu(A \cup B) = \mu(A) + \mu(B \backslash A) \leqslant \mu(A) + \mu(B)$ for all $A, B \in \mathcal{E}$, this can be extended to finite unions by induction, i.e. $\mu(\bigcup_{k=1}^{n} A_k) \leqslant \sum_{k=1}^{n} \mu(A_k)$. Taking limits on both sides completes the proof since the left side has limit $\mu(\bigcup_{k=1}^{\infty} A_k)$ by *sequential continuity*. □

**3.3. Finite and $\sigma$-finite Measures.** Let $\mu$ be a measure on a measurable space $(E, \mathcal{E})$. It is said to be **finite** if $\mu(E) < \infty$; then $\mu(A) < \infty$ for all $A$ in $\mathcal{E}$ by the monotonicity of $\mu$. It is called a **probability measure** if $\mu(E) = 1$, and we often denote it by $\mathbb{P}$. It is said to be **$\sigma$-finite** if there is a measurable partition $(E_n)$ of $E$ such that $\mu(E_n) < \infty$ for each $n$. It is said to be **$\Sigma$-finite** if there exists a sequence of finite measures $\mu_n$ s.t. $\mu = \sum_n \mu_n$.

PROPOSITION 1.24 (Relation between finite and $\sigma$-finite). *Every finite measure is obviously $\sigma$-finite, and every $\sigma$-finite measure is $\Sigma$-finite.*

PROOF. Assume $\mu$ is a $\sigma$-finite measure. Then there is a countable partition $(E_n)$ of $E$ such that $\mu(E_n) < \infty$ for all $n$. Let $A_1 = E_1$, $A_n = E_n \backslash (\bigcup_{k=1}^{n-1} E_k)$ for $n \geqslant 2$. Then $A_n$'s are disjoint and $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} E_n = E$. For each $n \in \mathbb{N}$, define $\mu_n(A) := \mu(A \cap A_n)$ for $A \in \mathcal{E}$. It is easily to check that $\mu_n$ is a finite measure on $(E, \mathcal{E})$ for all $n$. Moreover, for all $A$ in $\mathcal{E}$, we have

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

Thus, $\mu$ is $\Sigma$-finite. □

**3.4. Specification of Measures.** Given a measure on $(E, \mathcal{E})$, its values over a $\pi$-system generating $\mathcal{E}$ determine its values over all of $\mathcal{E}$, generally. This theorem can be viewed as the uniqueness of extension of measures from a $\pi$-system to its generating $\sigma$-algebra, but we do not want to talk about measure extensions precisely in this notes.

THEOREM 1.25 (Specification of measures). *Let $(E, \mathcal{E})$ be a measurable space. Let $\mu$ and $\nu$ be measures on it with $\mu(E) = \nu(E) < \infty$. If $\mu$ and $\nu$ agree on a $\pi$-system generating $\mathcal{E}$, then $\mu$ and $\nu$ are identical.*

PROOF. Let $\mathcal{C}$ be a $\pi$-system generating $\mathcal{E}$. Suppose that $\mu(A) = \nu(A)$ for every $A$ in $\mathcal{C}$, and $\mu(E) = \nu(E) < \infty$. We need to show that, then, $\mu(A) = \nu(A)$ for every $A$ in $\mathcal{E}$, or equivalently, that

$$\mathcal{D} = \{A \in \mathcal{E} : \mu(A) = \nu(A)\}$$

contains $\mathcal{E}$. Since $\mathcal{D} \supset \mathcal{C}$ by assumption, it is enough to show that $\mathcal{D}$ is a $\lambda$-system, for, then, the monotone class theorem 1.3 yields the desired conclusion that $\mathcal{D} \supset \mathcal{E}$. Now we check the condition for $\mathcal{D}$ to be a $\lambda$-system.

*First*, $E \in \mathcal{D}$ by the assumption that $\mu(E) = \nu(E)$.

*Second*, if $A, B \in \mathcal{D}$, and $A \supset B$, then $A \backslash B \in \mathcal{D}$, because

$$\mu(A \backslash B) = \mu(A) - \mu(B) = \nu(A) - \nu(B) = \nu(A \backslash B),$$

where we used the *finiteness* of $\mu$ to solve $\mu(A) = \mu(B) + \mu(A \backslash B)$ for $\mu(A \backslash B)$ and similarly for $\nu(A \backslash B)$.

*Finally*, suppose that $(A_n) \subset \mathcal{D}$ and $A_n \uparrow A$; then $\mu(A_n) = \nu(A_n)$ for every $n$, the left side increases to $\mu(A)$ by the sequential continuity of $\mu$, and the right side to $\nu(A)$ by the same for $\nu$; hence, $\mu(A) = \nu(A)$ and $A \in \mathcal{D}$. □

COROLLARY 1.26 (Probability measures case). *Let $\mu$ and $\nu$ be probability measures on $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$. Then, $\mu = \nu$ if and only if, for all $r$ in $\mathbb{R}$, we have $\mu([-\infty, r]) = \nu([-\infty, r])$.*

PROOF. It is immediate from the preceding theorem: $\mu(\bar{\mathbb{R}}) = \nu(\bar{\mathbb{R}}) = 1$ since $\mu$ and $\nu$ are probability measures, and the intervals $[-\infty, r]$ with $r$ in $\mathbb{R}$ form a $\pi$-system generating the Borel $\sigma$-algebra on $\bar{\mathbb{R}}$. □

COROLLARY 1.27 (Specification of $\sigma$-finite measures). *Let $\mu$ and $\nu$ be two measures defined on $(E, \mathcal{E})$, and $\mathcal{E}$ be generated by the $\pi$-system $\mathcal{C}$. Suppose either $\mu$ or $\nu$ is $\sigma$-finite on $\mathcal{C}$, and $\mu$ and $\nu$ agree on the $\pi$-system $\mathcal{C}$, then $\mu$ and $\nu$ are identical.*[a]

[a] We use the statement of Theorem 2.2.3 from Chung Kai Lai. *A course in probability theory*. Elsevier, 2001.

PROOF. If $\mu$ is $\sigma$-finite on $\mathcal{C}$, by definition, $\mathcal{C}$ contains a partition $(E_n)$ of $E$ s.t. $\mu(E_n) < \infty$ for all $n \in \mathbb{N}$. Since $\mu$ and $\nu$ agree on $\mathcal{C}$, then we have $\mu(E_n) = \nu(E_n) < \infty$ for all $n \in \mathbb{N}$.

We want to prove: $\mu(A) = \nu(A)$ for *all* $A \in \mathcal{E}$. *First*, if we take $F_n = \bigcup_{i=1}^n E_i$, then $F_n \uparrow E$ and $\mu(F_n) = \nu(F_n) < \infty$ for all $n$ in $\mathbb{N}$ since $\mu(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n \mu(E_i) = \sum_{i=1}^n \nu(E_i) = \nu(\bigcup_{i=1}^n E_i) < \infty$, where the first and third equalities are justified by the finite additivity of measures, the second equality and the last inequality by the assumption of the corollary. *Second*,

$$\lim_n \mu(F_n \cap A) = \mu(A), \quad \text{and} \quad \lim_n \nu(F_n \cap A) = \nu(A) \tag{1.3}$$

since $(F_n)$ is increasing. Besides, $\mu(F_n \cap A) \leqslant \mu(F_n) < \infty$ and $\nu(F_n \cap A) \leqslant \nu(F_n) < \infty$ using the monotonicity of measures. Thus, for each fixed $n \in \mathbb{N}$, $\mu(F_n \cap A)$ and $\nu(F_n \cap A)$ are both finite measures on $\mathcal{E}$, then $\mu(F_n \cap A) = \nu(F_n \cap A)$ by the conclusion of Theorem 1.25. Thus we get the desired result according to Equation (1.3). □

**3.5. Atoms, Purely Atomic and Diffuse Measures.** Let $(E, \mathcal{E})$ be a measurable space, and $\mu$ be a measure on $(E, \mathcal{E})$. A point $x$ is said to be an **atom** of $\mu$ if $\mu(\{x\}) > 0$. The measure $\mu$ is said to be **diffuse** if it has no atoms. It is said to be **purely atomic** if the set $D$ of its atoms is countable and $\mu(E \backslash D) = 0$.

EXAMPLE 1.28. Lebesgue measures are diffuse, a Dirac measure is purely atomic with one atom, and a discrete measures are purely atomic. †

The following theorem applies to $\Sigma$-finite (and therefore, to finite and $\sigma$-finite) measures.

THEOREM 1.29 (Decomposition of a $\Sigma$-finite measure). *Let $\mu$ be a $\Sigma$-finite measure on $(E, \mathcal{E})$. Then $\mu$ can have at most countably many atoms. And,*

$$\mu = \lambda + \nu,$$

*where $\lambda$ is a diffuse measure and $\nu$ is purely atomic.*

PROOF. Any $\Sigma$-finite measure can be represented as a sum of a sequence of finite measures, so it suffices to prove the case for *finite measures*.

(i). Fix $n \in \mathbb{N}$. Let $A_n$ be the set of *atoms* $x$ of the measure $\mu$ such that $\mu(\{x\}) > n^{-1}$. Since $\mu$ is finite, $A_n$ has at most $n\mu(E)$ many elements. Note that $D = \bigcup_{n\in\mathbb{N}} A_n$, since each $A_n$ has finitely many elements, the countable union $D$ has countably many elements, i.e., $\mu$ has countably many atoms.

(ii). Define $\lambda(A) := \mu(A\backslash D)$ and $\nu(A) := \mu(A \cap D)$ for all $A$ in $\mathcal{E}$. One can easily check that $\lambda$ and $\nu$ are measures on $(E, \mathcal{E})$. We claim that $\lambda$ is diffuse and $\nu$ is purely atomic.

Now, for all $x$ in $E$, we have

$$\begin{aligned} \lambda(\{x\}) = \mu(\{x\}\backslash D) = \mu(\varnothing) = 0, &\qquad \text{if } x \in D, \\ = \mu(\{x\}) = 0, &\qquad \text{if } x \notin D. \end{aligned}$$

Therefore $\lambda$ has no atoms, i.e., $\lambda$ is diffuse. Besides, denote by $D'$ the collection of all atoms of $\nu$. Note that $\mu$ and $\nu$ share the same atoms, i.e., $D = D'$. Indeed, let $x \in D$, then

$$\nu(\{x\}) = \mu(\{x\} \cap D) = \mu(\{x\}) > 0,$$

which means that $x \in D'$; besides, let $x \in D'$, then

$$\mu(\{x\}) = \mu(\{x\} \cap D) > 0,$$

which implies that $\{x\} \cap D \neq \varnothing$, i.e., $x \in D$. Hence $D' = D$ is countable and

$$\nu(E\backslash D') = \nu(E\backslash D) = \mu((E\backslash D) \cap D) = \mu(\varnothing) = 0,$$

which means that $\nu$ is purely atomic. □

**3.6. Completeness, Negligible Sets.** Let $(E, \mathcal{E}, \mu)$ be a measure space. A set $B \in \mathcal{E}$ is said to be **negligible** if $\mu(B) = 0$. An arbitrary subset of $E$ is said to be *negligible* if it is contained in a measurable negligible set. The measure space is said to be **complete** if every negligible set is measurable. If it is not complete, the following shows how to enlarge $\mathcal{E}$ to include all negligible sets and to extend $\mu$ onto the enlarged $\mathcal{E}$.

THEOREM 1.30 (Completion of measures). *Let $\mathcal{N}$ be the collection of all negligible subsets of $E$. Let $\bar{\mathcal{E}}$ be the $\sigma$-algebra generated by $\mathcal{E} \cup \mathcal{N}$. Then*

(i) *every $B$ in $\bar{\mathcal{E}}$ has the form $B = A \cup N$, where $A \in \mathcal{E}$ and $N \in \mathcal{N}$,*

(ii) *the formula $\bar{\mu}(A \cup N) = \mu(A)$ defines a unique measure $\bar{\mu}$ on $\bar{\mathcal{E}}$, we have $\bar{\mu}(A) = \mu(A)$ for $A \in \mathcal{E}$, and the measure space $(E, \bar{\mathcal{E}}, \bar{\mu})$ is complete.*

The measure space $(E, \bar{\mathcal{E}}, \bar{\mu})$ described is called the **completion** of $(E, \mathcal{E}, \mu)$. When $E = \mathbb{R}$ and $\mathcal{E} = \mathcal{B}(\mathbb{R})$ and $\mu = Leb$, the elements of $\bar{\mathcal{E}}$ are called the **Lebesgue measurable** sets.

PROOF. (i). Let $\mathcal{F} = \{A \cup N : A \in \mathcal{E}, N \in \mathcal{N}\}$. We will show that $\bar{\mathcal{E}} \subset \mathcal{F}$, it is obvious that, we only need to show that $\mathcal{F}$ is a $\sigma$-algebra on $E$.

(a). $\varnothing = \varnothing \cup \varnothing$ where $\varnothing$ in $\mathcal{E}$ and $\varnothing$ in $\mathcal{N}$.

(b). Let $C = A \cup N$ for some $A$ in $\mathcal{E}$ and $N$ in $\mathcal{N}$. Since $N$ in $\mathcal{N}$, there exists a measurable negligible set $N'$ such that $N \subset N'$. Then

$$C^c = (A \cup N)^c = (A \cup N')^c \cup (N' \backslash (N \cup A)) \in \mathcal{F}.$$

(c). Let $C_n = A_n \cup N_n$ with $A_n \in \mathcal{E}$ and $N_n \in \mathcal{N}$ for all $n$ in $\mathbb{N}$. Then $\bigcup_n C_n = (\bigcup_n A_n) \cup (\bigcup_n N_n)$. Since $N_n \in \mathcal{N}$, then for each $n$, there is a measurable negligible set $N'_n$ such that $N_n \subset N'_n$. By the fact that countable union of null sets is a null set, we deduce $\bigcup_n N'_n \in \mathcal{N}$ and $\bigcup_n N'_n$ is measurable. We also have $\bigcup_n A_n \in \mathcal{E}$, and then $\bigcup_n C_n \in \mathcal{F}$.

(ii). *First, $\bar{\mu}$ is well-defined.* Suppose $A_1 \cup N_1 = A_2 \cup N_2$, where $A_1$ and $A_2$ in $\mathcal{E}$; $N_1$ and $N_2$ in $\mathcal{N}$. We want to show $\mu(A_1) = \mu(A_2)$. Since $N_1 \in \mathcal{N}$, then there is a measurable negligible set $N'_1$ such that $N_1 \subset N'_1$. Then $A_2 \subset A_1 \cup N'_1$, which implies that $\mu(A_2) \leqslant \mu(A_1)$. Same argument gives $\mu(A_1) \leqslant \mu(A_2)$. Thus $\mu(A_1) = \mu(A_2)$.

*Second, $\bar{\mu}$ is a measure on $(E, \bar{\mathcal{E}})$.* (i). It is trivial that $\bar{\mu}(\varnothing) = \mu(\varnothing) = 0$. (ii). Let $C_n = A_n \cup N_n$ in $\bar{\mathcal{E}}$ be disjoint. Therefore, using the fact that $A_n$'s are disjoint we get

$$\bar{\mu}\Big(\bigcup_n C_n\Big) = \bar{\mu}\Big(\bigcup_n A_n \cup \bigcup_n N_n\Big) = \mu\Big(\bigcup_n A_n\Big) = \sum_n \mu(A_n) = \sum_n \bar{\mu}(A_n \cup N_n) = \sum_n \bar{\mu}(C_n).$$

*Third, $\bar{\mu}$ extends $\mu$.* This is obvious since for all $A \in \mathcal{E}$, $A = A \cup \varnothing \in \bar{\mathcal{E}}$, which implies that $\bar{\mu}(A) = \mu(A)$.

*Fourth, $\bar{\mu}$ is the unique extension of $\mu$.* Let $\mu'$ be another measure on $(E, \bar{\mathcal{E}})$ that extends $\mu$, we need to show $\mu'(C) = \bar{\mu}(C)$ for all $C$ in $\bar{\mathcal{E}}$. For all $C \in \bar{\mathcal{E}}$, $C = A \cup N$ for some $A \in \mathcal{E}$ and $N \in \mathcal{N}$. Then,

$$\mu'(C) = \mu'(A \cup N) \geqslant \mu'(A) = \mu(A) = \mu(A \cup N) = \bar{\mu}(C).$$

Besides, since $N \in \mathcal{N}$, there is a negligible measurable set $N' \supset N$, and then

$$\begin{aligned}\mu'(C) = \mu'(A \cup N) &\leqslant \mu'(A \cup N') = \mu(A \cup N') = \mu(A) + \mu(N') - \mu(A \cap N') \\ &= \mu(A) = \bar{\mu}(A \cup N) = \bar{\mu}(C),\end{aligned}$$

where the fourth equality holds since $N'$ is negligible and $A \cap N' \subset N'$.

*Finally, the measure space $(E, \bar{\mathcal{E}}, \bar{\mu})$ is complete.* We need to prove that every negligible set of $(E, \bar{\mathcal{E}}, \bar{\mu})$ is measurable. Let $A \cup N \in \bar{\mathcal{E}}$ be a measurable negligible set, we want to prove: for all $C \subset A \cup N$, $C$ in $\bar{\mathcal{E}}$. Since $N \in \mathcal{N}$, there is a negligible measurable set $N' \supset N$, then

$$\mu(A \cup N') \leqslant \bar{\mu}(A \cup N) + \mu(N') = 0,$$

which shows that $A \cup N'$ is a measurable negligible set with respect to $(E, \mathcal{E}, \mu)$ and $C \subset A \cup N'$ is a negligible set with respect to $(E, \mathcal{E}, \mu)$, i.e., $C \in \mathcal{N}$. Then, $C = \varnothing \cup C \in \bar{\mathcal{E}}$. This completes the proof. □

**3.7. Almost Everywhere.** If a proposition holds for all but a negligible set of $x$ in $E$, then we say that it holds for *almost every* $x$, or **almost everywhere**. If the measure $\mu$ used to define negligibility needs to be indicated, we say *$\mu$-almost everywhere*. If $E$ is replaced by a measurable set $A$, we say almost everywhere on $A$.

EXAMPLE 1.31. For example, given numerical functions $f$ and $g$ on $E$, and a measurable set $A$, saying that $f = g$ almost everywhere on $A$ is equivalent to saying that $\{x \in A : f(x) \neq g(x)\}$ is *negligible*, which is then equivalent to saying that there exists a *measurable* set $M$ with $\mu(M) = 0$ such that $f(x) = g(x)$ for every $x$ in $A \backslash M$. Actually, we need to show that $\{x \in A : f(x) \neq g(x)\}$ is a measurable set.

PROOF. Notice that $\{f \neq g\} = \{f > g\} \cup \{f < g\}$. We only need to prove $\{f > g\} \in \mathcal{E}$. For any $x$ in $\{x \in A : f(x) > g(x)\}$, there exist a rational number $r$ such that $g(x) < r < f(x)$, then

$$\{x \in A : f(x) > g(x)\} = \bigcup_{r\in\mathbb{Q}} \{x \in A : g(x) < r\} \cap \{x \in A : f(x) > r\} \in \mathcal{E}.$$

Thus $\{f \neq g\} \in \mathcal{E}$ and consequently $\{f = g\}$, $\{f \geqslant g\}$ and $\{f > g\}$ are all $\mathcal{E}$-measurable sets. □

## 4. Integration

Let $(E, \mathcal{E}, \mu)$ be a measure space. Assume that $\mathcal{E}$ stands also for the collection of all $\mathcal{E}$-measurable functions on $E$ and that $\mathcal{E}_+$ is the sub-collection consisting of positive $\mathcal{E}$-measurable functions.

Our aim is to define the "integral of $f$ with respect to $\mu$" for all reasonable functions $f$ in $\mathcal{E}$. We shall denote it by any of the following:

$$\mu f = \mu(f) := \int_E f(x)\, \mu(\mathrm{d}x) = \int_E f \,\mathrm{d}\mu.$$

As the notation $\mu(f)$ suggests, integration is a kind of multiplication: for all $a, b \in \mathbb{R}_+$ and $f, g, f_n \in \mathcal{E}_+$:

(1.4)
a) *Positivity* : $\mu(f) \geqslant 0$; $\mu(f) = 0$ if $f = 0$.
b) *Linearity* : $\mu(af + bg) = a\mu(f) + b\mu(g)$.
c) *Monotone convergence theorem* : If $f_n \uparrow f$, then $\mu(f_n) \uparrow \mu(f)$.

We start with the definition of the integral and proceed to proving the properties (1.4) and their extensions. At the end, we shall also show that (1.4) characterizes integration.

DEFINITION 1.32 (Integration with respect to $\mu$).

(i) Let $f$ be *simple and positive*. If its canonical form is $f = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$, then we define

$$\mu(f) = \sum_{i=1}^n a_i\, \mu(A_i).$$

(ii) Let $f \in \mathcal{E}_+$. Put $f_n = d_n \circ f$, where the $d_n$ are as in Lemma 1.13. Then each $f_n$ is simple and positive, and the sequence $(f_n)$ increases to $f$ as shown in the proof of 1.14. The integral $\mu(f_n)$ is defined for each $n$ by the preceding step, and the sequence of numbers $\mu(f_n)$ is increasing (see Remark (d) below). We define

$$\mu(f) = \lim_{n\to\infty} \mu(f_n).$$

(iii) Let $f \in \mathcal{E}$. Then $f^+ = f \wedge 0$ and $f^- = -(f \vee 0)$ belong to $\mathcal{E}_+$, and their integrals $\mu(f^+)$ and $\mu(f^-)$ are defined by the preceding step. Noting that $f = f^+ - f^-$, we define

$$\mu(f) = \mu(f^+) - \mu(f^-)$$

provided that at least one term on the right side be finite. Otherwise, $\mu(f)$ is *undefined*.

REMARK 1.33 (Integration of positive simple functions). Let $f, g$, etc. be simple and positive functions.

(i) The formula for $\mu(f)$ remains the same even when $f = \sum a_i \mathbb{1}_{A_i}$ is NOT the canonical representation of $f$. This is easy to check using the *finite additivity* of $\mu$.
(ii) If $a$ and $b$ are in $\mathbb{R}_+$, then $af + bg$ is simple and positive, and the *linearity property* holds:

$$\mu(af + bg) = a\mu(f) + b\mu(g).$$

(iii) *Monotonicity*. If $f \leqslant g$ then $\mu(f) \leqslant \mu(g)$. This follows from the linearity property above applied to the simple positive functions $f$ and $g - f$:

$$\mu(f) \leqslant \mu(f) + \mu(g - f) = \mu(f + g - f) = \mu(g).$$

(iv) In step (b) of the definition, we have $f_1 \leqslant f_2 \leqslant \cdots$. The remark on monotonicity shows that $\mu(f_1) \leqslant \mu(f_2) \leqslant \cdots$. Thus, $\lim \mu(f_n)$ exists as claimed (it can be $+\infty$). †

**4.1. Examples.** We can give some examples first.

EXAMPLE 1.34 (Discrete measures). Fix $x_0$ in $E$ and consider the Dirac measure $\delta_{x_0}$ sitting at $x_0$. Going through the steps of the definition of the integral, we see that $\delta_{x_0}(f) = f(x_0)$ for every $f$ in $\mathcal{E}$[2]. This extends to discrete measures: if $\mu = \sum_{x \in D} m(x)\delta_x$ for some countable set $D$ and positive masses $m(x)$, then

$$\mu(f) = \sum_{x \in D} m(x) f(x)$$

for every $f$ in $\mathcal{E}_+$. A similar result holds for purely atomic measures as well. †

EXAMPLE 1.35 (Discrete spaces). Suppose that $(E, \mathcal{E})$ is discrete, that is, $E$ is countable and $\mathcal{E} = 2^E$. Then, every numerical function on $E$ is $\mathcal{E}$-measurable, and every measure $\mu$ has the form in the preceding example with $D = E$ and $m(x) = \mu(\{x\})$. Thus, for every positive function $f$ on $\mathcal{E}$,

$$\mu(f) = \sum_{x \in E} \mu(\{x\}) f(x).$$

EXAMPLE 1.36 (Lebesgue integrals). Suppose that $E$ is a Borel subset of $\mathbb{R}^d$ for some $d \geqslant 1$ and suppose that $\mathcal{E} = \mathcal{B}(E)$, the Borel subsets of $E$. Suppose that $\mu$ is the restriction of the Lebesgue measure on $\mathbb{R}^d$ to $(E, \mathcal{E})$. For $f$ in $\mathcal{E}$, we employ the following notations for the integral $\mu(f)$:

$$\mu(f) = Leb_E(f) = \int_E f(x)\, Leb(\mathrm{d}x) = \int_E f(x)\, \mathrm{d}x,$$

the last using $\mathrm{d}x$ for $Leb(\mathrm{d}x)$ in keeping with tradition. This integral is called the **Lebesgue integral** of $f$ on $E$. †

REMARK 1.37 (Riemann integrals and Lebesgue integrals). If the Riemann integral of $f$ exists, then so does the Lebesgue integral, and the two integrals are equal. The *converse* is FALSE; the Lebesgue integral exists for a larger class of functions than does the Riemann integral. †

**4.2. Positivity and Monotonicity.** What we want to do now is to derive some properties of integration of positive measurable functions or general measurable functions.

THEOREM 1.38 (Monotonicity for positive functions). *If $f \in \mathcal{E}_+$, then $\mu(f) \geqslant 0$. If $f$ and $g$ are in $\mathcal{E}_+$ and $f \leqslant g$, then $\mu(f) \leqslant \mu(g)$.*

PROOF. Positivity of $\mu(f)$ for $f$ positive is immediate from Definition 1.32. To show monotonicity, let $f_n = d_n \circ f$ and $g_n = d_n \circ g$ as in Definition 1.32 (ii). Since each $d_n$ is an increasing function (see Lemma 1.13), $f \leqslant g$ implies that $f_n \leqslant g_n$ for each $n$ which in turn implies that $\mu(f_n) \leqslant \mu(g_n)$ for each $n$ by Remark 1.33 (iii). Letting $n \to \infty$, we see from Definition 1.32 that $\mu(f) \leqslant \mu(g)$. □

[2]Actually, we only need to check this result for simple positive functions.

**4.3. Integral Over a Set.** Let $f \in E$ and let $A$ be a measurable set. Then, $f\mathbb{1}_A \in \mathcal{E}$, and the **integral of $f$ over $A$** is defined to be the integral of $f\mathbb{1}_A$. The following notations are used for it:

$$\mu(f\mathbb{1}_A) = \int_A f(x)\,\mu(\mathrm{d}x) = \int_A f\,\mathrm{d}\mu.$$

The following shows that, for each $f$ in $\mathcal{E}_+$, the set function $A \mapsto \mu(f\mathbb{1}_A)$ is *finitely additive*. This property extends to countable additivity as a corollary to the *monotone convergence theorem* 1.40 below (see Theorem 1.41 (iii)).

PROPOSITION 1.39 (Finitely additive of integral over sets). *Let $f \in \mathcal{E}_+$. Let $A$ and $B$ be disjoint sets in $\mathcal{E}$ with union $C$. Then*

$$\mu(f\mathbb{1}_A) + \mu(f\mathbb{1}_B) = \mu(f\mathbb{1}_C).$$

PROOF. If $f$ is simple, this is immediate from the linearity of Remark (b) of Definition 1.32. For arbitrary $f$ in $\mathcal{E}_+$, putting $f_n = d_n \circ f$ as in Definition 1.32 (ii), we get

$$\mu(f_n\mathbb{1}_A) + \mu(f_n\mathbb{1}_B) = \mu(f_n\mathbb{1}_C)$$

since the $f_n$ are simple. Observing that $f_n\mathbb{1}_D = d_n \circ (f\mathbb{1}_D)$ for $D = A, B, C$ and taking limits as $n \to \infty$ we get the desired result through Definition 1.32 (ii). □

**4.4. Monotone Convergence Theorem.** This is the main theorem of integration. It is the KEY tool for *interchanging the order of taking limits and integrals*. It states that the mapping $f \mapsto \mu(f)$ from $\mathcal{E}_+$ into $\bar{\mathbb{R}}_+$ is continuous under increasing limits. As such, it is an extension of the sequential continuity of measures.

THEOREM 1.40 (Monotone convergence theorem). *Let $(f_n)$ be an increasing sequence in $\mathcal{E}_+$. Then,*

$$\mu\big(\lim_{n\to\infty} f_n\big) = \lim_{n\to\infty} \mu(f_n).$$

PROOF. Let $(f_{nm})_{m=1}^{\infty}$ be an *increasing* sequence of *positive simple* functions converging to $f_n$ for all $n$ in $\mathbb{N}$. Define $g_n = \max\{f_{1n}, f_{2n}, \cdots, f_{nn}\} \leqslant f_n$, then $g_n$ is an *increasing* sequence of *positive simple* functions, and $g_n \geqslant f_{kn}$ for all $1 \leqslant k \leqslant n$. Hence $g := \lim_{n\to\infty} g_n \geqslant f_k$ for each $k$ in $\mathbb{N}$, where $g \in \mathcal{E}_+$. By sending $k$ to $\infty$ we get $g \geqslant f$. Otherwise, Since $g_n \leqslant f_n$, then $g \leqslant f$ by letting $n \to \infty$, so $g = f$. Thus $\mu(f) = \mu(g) = \lim_{n\to\infty} \mu(g_n) \leqslant \lim_{n\to\infty} \mu(f_n)$ since $g_n \leqslant f_n$ are both positive simple functions. Moreover, $\lim_{n\to\infty} \mu(f_n) \leqslant \mu(f)$ as $f_n \leqslant f$ for every $n$ and we use the monotonicity showed in Theorem 1.38. Now we have proved the desired result. □

**4.5. Linearity of Integration.** Before talking about more convergence properties of the integration, we turn to some basic properties using monotone convergence theorem.

THEOREM 1.41 (Linearity of integration).

(i) *For $f$ and $g$ in $\mathcal{E}_+$ and $a$ and $b$ in $\mathbb{R}_+$,*

$$\mu(af + bg) = a\mu(f) + b\mu(g).$$

*The same is true for $f, g \in \mathcal{E}$ such that $\mu(f)$ and $\mu(g)$ are fintie, and arbitrary $a, b \in \mathbb{R}$.*

(ii) *Beppo Levi's Theorem. For $(f_n) \subset \mathcal{E}_+$, $\mu(\sum_{n=1}^{\infty} f_n) = \sum_{n=1}^{\infty} \mu(f_n)$.*

(iii) *Let $f \in \mathcal{E}_+$ and $(F_n)_{n=1}^{\infty}$ be a partition of $C \in \mathcal{E}$, then $\mu(f\mathbb{1}_C) = \sum_{n=1}^{\infty} \mu(f\mathbb{1}_{F_n})$.*

PROOF. (i). Suppose that $f, g, a, b$ are all positive. If $f$ and $g$ are simple, the linearity can be checked directly (see Remark 1.33). If not, choose $(f_n)$ and $(g_n)$ to be sequences of simple positive functions

increasing to $f$ and $g$ respectively. Then,

$$\mu(af_n + bg_n) = a\mu(f_n) + b\mu(g_n),$$

and the monotone convergence theorem applied to both sides completes the proof. The remaining statements follow from Definition 1.32 and the linearity for positive functions after putting $f = f^+ - f^-$ and $g = g^+ - g^-$.

(ii). Use the *finite linearity* of integration,

$$\int_E \sum_{k=1}^{n} f_k \,\mathrm{d}\mu = \sum_{k=1}^{n} \int_E f_k \,\mathrm{d}\mu.$$

Then we get the result by letting $n$ to $\infty$ and using the monotone convergence theorem.

(iii). Define $f_n = f\mathbb{1}_{\bigcup_{k=1}^{n} F_k}$, then $(f_n)$ is an increasing sequence of positive measurable functions converging to $f$ for all $x \in C$. Then

$$\int_C f \,\mathrm{d}\mu = \lim_{n\to\infty} \int_C f_n \,\mathrm{d}\mu$$

by using the monotone convergence theorem. Besides, using the definition of $f_n$ and finite linearity of integration,

$$\int_C f_n \,\mathrm{d}\mu = \int_C \sum_{k=1}^{n} f\mathbb{1}_{F_k} \,\mathrm{d}\mu = \sum_{k=1}^{n} \int_{F_k} f \,\mathrm{d}\mu.$$

Then we get the desired result by substituting the RHS of the last equation into the former equation, and letting $n$ to $\infty$. □

**4.6. Integrability.** A function $f$ in $\mathcal{E}$ is said to be **integrable** if $\mu(f)$ exists and is a real number. Thus, $f$ in $\mathcal{E}$ is integrable if and only if $\mu(f^+) < \infty$ and $\mu(f^-) < \infty$, or equivalently, if and only if the integral of $|f| = f^+ - f^-$ is a finite number.

LEMMA 1.42 (Property of integrable functions). *Let $f$ be a $\mathcal{E}$-measurable function.*

(i) *If $f$ is integrable, then $|\mu(f)| \leqslant \mu(|f|)$.*
(ii) *$f$ is integrable if and only if $|f|$ is integrable.*
(iii) *If $f$ is integrable, then $f$ is real-valued (i.e., finite) almost everywhere.*

PROOF. (i). According to Theorem 1.38,

$$\max\left\{\mu(f^+), \mu(f^-)\right\} \leqslant \mu(|f|) \tag{1.5}$$

since $f^+$ and $f^-$ are positive and $|f| = f^+ + f^- \geqslant f^+$ and $f^-$. Thus

$$|\mu(f)| = |\mu(f^+) - \mu(f^-)| \leqslant \max\left\{\mu(f^+), \mu(f^-)\right\} \leqslant \mu(|f|)$$

using the linearity and monotonicity of integration.

(ii). Using Equation (1.5), if $|f|$ is integrable, then $f$ is integrable. Besides, if $f$ is integrable, then $\mu(f^+) < \infty$ and $\mu(f^-) < \infty$. Using the linearity of integration, we have

$$\mu(|f|) = \mu(f^+ + f^-) = \mu(f^+) + \mu(f^-) < \infty,$$

so $|f|$ is integrable.

(iii). According to (b), we only need to prove if $f \in \mathcal{E}_+$ is integrable, then $f < \infty$ almost everywhere. Assume $\mu(\{f = \infty\}) > 0$, then

$$\forall\, n \in \mathbb{N}: \quad \int_E f \,\mathrm{d}\mu \geqslant \int_E f\mathbb{1}_{\{f=\infty\}} \,\mathrm{d}\mu \geqslant n\mu(\{f = \infty\})$$

according to the monotonicity of integration. This means that $\mu(f) = \infty$, which contradicts the condition that $f \in \mathcal{E}_+$ is integrable. So $\mu(\{f = \infty\}) = 0$. □

**4.7. Insensitivity of the Integral.** We show next that the integral of a function remains unchanged if the values of the function are changed over a negligible set.

THEOREM 1.43 (Insensitivity of the Integral).

(i) *If $A$ in $\mathcal{E}$ is negligible, then $\mu(f\mathbb{1}_A) = 0$ for every $f$ in $\mathcal{E}$.*
(ii) *If $f$ and $g$ are in $\mathcal{E}_+$ and $f = g$ almost everywhere, then $\mu(f) = \mu(g)$.*
(iii) *If $f \in \mathcal{E}_+$, then $\mu(f) = 0$ if and only if $f = 0$ almost everywhere.*
(iv) *Let $f \in \mathcal{E}$ be an integrable function such that $\mu(f\mathbb{1}_A) \geqslant 0$ holds for all $A \in \mathcal{E}$, then $f \geqslant 0$ almost everywhere.*

PROOF. (i). Let $A$ be measurable and negligible. If $f \in \mathcal{E}_+$ and simple, then $\mu(f\mathbb{1}_A) = 0$ by Definition 1.32. This extends to non-simple case by the monotone convergence theorem using a sequence of simple $f_n$ increasing to $f$: then $\mu(f_n\mathbb{1}_A) = 0$ for all $n$ and $\mu(f\mathbb{1}_A)$ is the limit of the left side. For $f$ in $\mathcal{E}$ arbitrary, we have $\mu(f^+\mathbb{1}_A) = \mu(f^-\mathbb{1}_A) = 0$ and hence $\mu(f\mathbb{1}_A) = 0$ since $(f\mathbb{1}_A)^+ = f^+\mathbb{1}_A$ and $(f\mathbb{1}_A)^- = f^-\mathbb{1}_A$.

(ii). If $f$ and $g$ are in $\mathcal{E}_+$ and $f = g$ almost everywhere, then $A = \{f \neq g\}$ is measurable and negligible, and the integrals of $f$ and $g$ on $A$ both vanish, i.e., $\mu(f\mathbb{1}_A) = \mu(g\mathbb{1}_A) = 0$. Thus, with $B = A^c$, we have

$$\mu(f) = \mu(f\mathbb{1}_A) + \mu(f\mathbb{1}_B) = \mu(f\mathbb{1}_B),$$

and $\mu(g) = \mu(g\mathbb{1}_B)$, which imply $\mu(f) = \mu(g)$ since $f(x) = g(x)$ for all $x$ in $B$.

(iii). Let $f \in \mathcal{E}_+$ and $\mu(f) = 0$. We need to show that the set $N = \{f > 0\}$ has measure zero. Take a sequence of numbers $\varepsilon_k > 0$ decreasing to 0, let $N_k = \{f > \varepsilon_k\}$, and obverse that $N_k \uparrow N$, which implies that $\mu(N_k) \uparrow \mu(N)$ by the sequential continuity of $\mu$. Thus, it is enough to show that $\mu(N_k) = 0$ for every $k$. This is easy to show: $f \geqslant \varepsilon_k\mathbb{1}_{N_k}$ implies that

$$0 = \mu(f) \geqslant \mu(\varepsilon_k\mathbb{1}_{N_k}) = \varepsilon_k\mu(N_k),$$

and $\varepsilon_k > 0$, we must have $\mu(N_k) = 0$.

(iv). Assume $A = \{f < 0\}$, then we have $f(x) < 0$ for all $x$ in $A$. Besides, the condition shows that $\mu(f\mathbb{1}_A) \geqslant 0$, thus we must have

$$\int_E f\mathbb{1}_A \,\mathrm{d}\mu = \int_A f \,\mathrm{d}\mu = 0,$$

which implies that $f\mathbb{1}_A = 0$ almost everywhere using (c), thus $f \geqslant 0$ almost everywhere. □

**4.8. Absolutely Continuity of the Integral.** The following theorem is very, very important in the discussion of uniform integrability.

THEOREM 1.44 (Absolutely continuity of the integral). *Let $f \in \mathcal{E}$ be an integrable function. Then for all $\varepsilon > 0$, there exists a $\delta > 0$, such that for all $\mu(A) < \delta$, where $A \in \mathcal{E}$, we have*

$$|\mu(f\mathbb{1}_A)| \leqslant \mu(|f|\mathbb{1}_A) < \varepsilon.$$

PROOF. Define $f_n = f\mathbb{1}_{\{f \leqslant n\}} + n\mathbb{1}_{\{f > n\}}$, then $|f_n| \uparrow |f|$. By the monotone convergence theorem, $\mu(|f_n|) \uparrow \mu(|f|)$. Therefore, for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$, such that when $n \geqslant N$, we have

$$\big|\mu(|f|) - \mu(|f_n|)\big| < \frac{\varepsilon}{2} \quad \text{i.e.,} \quad \mu(|f|) < \mu(|f_n|) + \frac{\varepsilon}{2}.$$

Now, choose any $A \in \mathcal{E}$ such that $\mu(A) \leqslant \frac{\varepsilon}{2N}$ $(= \delta)$, then

$$|\mu(f\mathbb{1}_A)| \leqslant \mu(|f|\mathbb{1}_A) = \mu((|f| - |f_N|)\mathbb{1}_A) + \mu(|f_N|\mathbb{1}_A)$$
$$\leqslant \mu(|f| - |f_N|) + N\mu(A) < \frac{\varepsilon}{2} + \frac{N\varepsilon}{2N} = \varepsilon.$$

This completes the proof. □

**4.9. Fatou's Lemma.** We return to the properties of the integral under limits. Next is a useful consequence of the monotone convergence theorem.

THEOREM 1.45 (Fatou's lemma). *Let $(f_n) \subset \mathcal{E}_+$. Then $\mu(\liminf f_n) \leqslant \liminf \mu(f_n)$.*

PROOF. Define $g_m = \inf_{n \geqslant m} f_n$ and recall that $\liminf f_n$ is the limit of the increasing sequence $(g_m)$ in $\mathcal{E}_+$. Hence, by the monotone convergence theorem,

$$\mu(\liminf f_n) = \lim \mu(g_m).$$

On the other hand, $g_m \leqslant f_n$ for all $n \geqslant m$, which implies that $\mu(g_m) \leqslant \mu(f_n)$ for all $n \geqslant m$ by the monotonicity of integration, which in turn means that $\mu(g_m) \leqslant \inf_{n \geqslant m} \mu(f_n)$. Hence, as desired,

$$\lim \mu(g_m) \leqslant \liminf \mu(f_n).$$

This completes the proof. □

COROLLARY 1.46. *Let $(f_n) \subset \mathcal{E}$. If there is an integrable function $g$ such that $f_n \geqslant g$ for every $n$, then*

$$\mu(\liminf f_n) \leqslant \liminf \mu(f_n).$$

*If there is an integrable function $g$ such that $f_n \leqslant g$ for every $n$, then*

$$\mu(\limsup f_n) \geqslant \limsup \mu(f_n).$$

PROOF. Let $g$ be integrable. Then, the measurable set $A = \{|g| = \infty\}$ is negligible according to Lemma 1.42 (iii). Hence, let $B = A^c$, then $f_n\mathbb{1}_B = f_n$, $g_n\mathbb{1}_B = g_n$ almost everywhere, and $g\mathbb{1}_B$ is real-valued.

*The first statement* follows from Fatou's Lemma applied to the well-defined sequence $(f_n\mathbb{1}_B - g\mathbb{1}_B)$ in $\mathcal{E}_+$ together with the linearity and insensitivity of integration.

*The second statement* follows again from Fatou's lemma, now applied to the well-defined sequence $(g\mathbb{1}_B - f_n\mathbb{1}_B)$ in $\mathcal{E}_+$ together with the linearity and insensitivity, and the observation that $\limsup r_n = -\liminf(-r_n)$ for every sequence $(r_n)$ in $\bar{\mathbb{R}}$. □

**4.10. Dominated Convergence Theorem.** A function $f$ is said to be **dominated** by the function $g$ if $|f| \leqslant g$; note that $g \geqslant 0$ necessarily. A sequence $(f_n)$ is said to be *dominated* by $g$ if $|f_n| \leqslant g$ for every $n$. If so, and if $g$ can be taken to be a *finite constant*, then $(f_n)$ is said to be **bounded**.

THEOREM 1.47 (Dominated convergence theorem).

(i) *Dominated convergence. Let $(f_n) \subset \mathcal{E}$. Suppose that $(f_n)$ is dominated by some integrable function $g$. If $\lim f_n$ exists, then it is integrable and*

$$\mu(\lim f_n) = \lim \mu(f_n).$$

(ii) *Bounded convergence. Furthermore, Let $(f_n)$ be bounded and $\mu$ be finite. If $\lim f_n$ exists, then it is a bounded integrable function and the equality holds again.*

PROOF OF THEOREM 1.47 (I). By assumption, $-g \leqslant f_n \leqslant g$ for every $n$, and both $g$ and $-g$ are integrable. Thus, both statements of the last corollary apply:

$$\mu(\liminf f_n) \leqslant \liminf \mu(f_n) \leqslant \limsup \mu(f_n) \leqslant \mu(\limsup f_n). \tag{1.6}$$

If $\lim f_n$ exists, then $\liminf f_n = \limsup f_n = \lim f_n$, and $\lim f_n$ is integrable since it is dominated by $g$. Hence, the extreme members of Equation (1.6) are finite and equal, and all inequality signs are in fact equalities. □

The second part of Theorem 1.47 is called the **bounded convergence theorem**. It is important to note that the assumption of *finite measure* in it is essential!

EXAMPLE 1.48 (A counterexample for bounded convergence theorem). Consider the real line $\mathbb{R}$ equipped with the Borel sets $\mathcal{B}(\mathbb{R})$ and Lebesgue measure $\lambda$. Define $f_n(x) = n^{-1}\mathbb{1}_{[0,n]}$, then it is obvious that $f_n \to 0 := f$ pointwise. However,

$$\int_{\mathbb{R}} f_n \, \mathrm{d}\lambda = \frac{1}{n}\lambda([0,n]) = 1,$$

while $\int_{\mathbb{R}} f \, \mathrm{d}\lambda = 0$. So the conclusion of bounded convergence theorem **fails** since $\lambda(\mathbb{R}) = \infty$. †

PROOF OF THEOREM 1.47 (II). Define $g \equiv M$ where $M$ is a finite constant. Then

$$\mu(g) = \int M \, \mathrm{d}\mu = M\mu(E) < \infty$$

holds since $\mu(E) < \infty$. Then $f$ is dominated by the *integrable* function $g$. Using dominated convergence theorem we get the desired result. □

**4.11. Almost Everywhere Version.** The insensitivity of integration to changes over negligible sets enables us to re-state all the results above by allowing the conditions to fail over negligible sets. We start by extending the definition of integration somewhat.

REMARK 1.49 (Convention). Let $f$ be a numerical function on $E$. Suppose that there exists an $\mathcal{E}$-measurable function $g$ such that $f(x) = g(x)$ for almost every $x$ in $E$. Then, we define the integral $\mu(f)$ of $f$ to be the number $\mu(g)$ provided that $\mu(g)$ is defined. Otherwise, if $\mu(g)$ does not exist, $\mu(f)$ does not exist either.

*The definition here is without ambiguities*: if $h$ is another measurable function such that $f = h$ almost everywhere, then $g = h$ almost everywhere; if $\mu(g)$ exists, then so does $\mu(h)$ and $\mu(g) = \mu(h)$ by the insensitivity property; if $\mu(g)$ does not exist, then neither does $\mu(h)$.

In fact, the convention here is one of notation making, almost. Let $g \in E$ and $f = g$ almost everywhere. Let $(E, \bar{\mathcal{E}}, \bar{\mu})$ be the completion of $(E, \mathcal{E}, \mu)$. Then, $f \in \bar{\mathcal{E}}$ (exercise), and the integral $\bar{\mu}(f)$ makes sense by Definition 1.32 applied on the measurable space $(E, \bar{\mathcal{E}}, \bar{\mu})$. Since $\mathcal{E} \subset \bar{\mathcal{E}}$, the function $g$ is $\bar{\mathcal{E}}$-measurable as well, and $\bar{\mu}(g)$ makes sense and it is clear that $\bar{\mu}(g) = \mu(g)$. Since $f$ and $g$ are $\bar{\mathcal{E}}$-measurable and $f = g$ $\bar{\mu}$-almost everywhere, $\bar{\mu}(f) = \bar{\mu}(g)$ by insensitivity. So, the convention above amounts to writing $\mu(f)$ instead of $\bar{\mu}(f)$.

The preceding convention tells us: when considering properties that hold almost everywhere, we can first discuss the properties in the *completed measure space*, and then, after obtaining the corresponding conclusions, pull back to the *original measure space*. With this convention in place, we now re-state the monotone convergence theorem in full generality.

THEOREM 1.50 (Almost everywhere case of monotone convergence theorem). *Let $(f_n)$ be a sequence of numerical functions on $E$. Suppose that, for each $n$, there is $g_n$ in $\mathcal{E}$ such that $f_n = g_n$ almost*

*everywhere. Further, suppose for each $n$ that $f_n \geqslant 0$ almost everywhere and $f_n \leqslant f_{n+1}$ almost everywhere. Then, $\lim f_n$ exists almost everywhere, is positive almost everywhere, and $\mu(\lim f_n) = \lim \mu(f_n)$.*

We discuss this fully to indicate its meaning and the issues involved first. Let $\mathcal{N}$ denote the collection of all *measurable* negligible sets[3], that is, every $N$ in $\mathcal{N}$ belongs to $\mathcal{E}$ and $\mu(N) = 0$. Now fix $n$. To say that $f_n = g_n$ almost everywhere is to say that there is $N_n$ in $\mathcal{N}$ such that $f_n = g_n$ outside $N_n$. Similarly, $f_n \geqslant 0$ almost everywhere means that there is $M_n$ in $\mathcal{N}$ such that $f_n \geqslant 0$ outside $M_n$. And since $f_n \leqslant f_{n+1}$ almost everywhere, there is $L_n$ in $\mathcal{N}$ such that $f_n \leqslant f_{n+1}$ outside $L_n$. These are the conditions. The claim of the theorem is as follows.

First, there is an $\mathcal{E}$-measurable function $f$, and a set $N$ in $\mathcal{N}$ such that $\lim f_n(x)$ exists and is equal to $f(x)$ for every $x$ outside $N$. Also, there is $M$ in $\mathcal{N}$ such that $f \geqslant 0$ outside $M$. Finally, $\mu(f) = \lim \mu(f_n)$, where the $\mu(f_n)$ are defined by convention to be the numbers $\mu(g_n)$.

PROOF. Let

$$N = \bigcup_{n=1}^{\infty} (L_n \cup M_n \cup N_n).$$

Then, $N \in \mathcal{E}$ and $\mu(N) = 0$ by Boole's inequality, that is, $N \in \mathcal{N}$. For $x \notin N$, we have

$$0 \leqslant f_1(x) = g_1(x) \leqslant f_2(x) = g_2(x) \leqslant \cdots,$$

and hence $\lim f_n(x)$ exists and is equal to $\lim g_n(x)$. Define $f = \lim f_n \mathbb{1}_{E \setminus N}$, then $f$ is the limit of the increasing sequence $(g_n \mathbb{1}_{E \setminus N})$ in $\mathcal{E}_+$. So, $f$ is in $\mathcal{E}_+$ and we may take $M = \varnothing$. There remains to show that $\mu(f) = \lim \mu(g_n)$. Now, in fact

$$\mu(f) = \mu(\lim g_n \mathbb{1}_{E \setminus N}) = \lim \mu(g_n \mathbb{1}_{E \setminus N}) = \lim \mu(g_n),$$

where we used the monotone convergence theorem to justify the second equality, and the insensitivity to justify the third. □

**4.12. Characterization of the Integral.** Definition 1.32 defines the integral $\mu(f)$ for every $f$ in $\mathcal{E}_+$. Thus, in effect, integration extends the domain of $\mu$ from the measurable sets (identified with their indicator functions) to the space $\mathcal{E}_+$ of all positive measurable functions (and beyond), and hence we may regard $\mu$ as the mapping $f \mapsto \mu(f)$ from $\mathcal{E}_+$ into $\bar{\mathbb{R}}_+$.

The mapping $\mu : \mathcal{E}_+ \to \bar{\mathbb{R}}_+$ is necessarily positive, linear, and continuous under increasing limits; these were promised before. We end this section with the following very useful *converse*.

THEOREM 1.51 (Characterization of the integral). *Let $(E, \mathcal{E})$ be a measurable space. Let $L$ be a mapping from $\mathcal{E}_+$ into $\bar{\mathbb{R}}_+$. Then there exists a unique measure $\mu$ on $(E, \mathcal{E})$ such that $L(f) = \mu(f)$ for every $f$ in $\mathcal{E}_+$ if and only if*

*a)* $f = 0 \Rightarrow L(f) = 0.$

*b)* $f, g \in \mathcal{E}_+$ *and* $a, b \in \mathbb{R}_+ \Rightarrow L(af + bg) = aL(f) + bL(g).$

*c)* $(f_n) \subset \mathcal{E}_+$ *and* $f_n \uparrow f \Rightarrow L(f_n) \uparrow L(f).$

PROOF. *Necessity* of the conditions is immediate from the properties of the integral. To show the *sufficiency*, suppose that $L$ has the properties (i) - (iii). Define

$$\mu(A) = L(\mathbb{1}_A), \quad A \in \mathcal{E}. \tag{1.7}$$

[3]From the definition of an negligible set, it follows that any negligible set is contained in a larger measurable negligible set. Thus, we can consider the problem on a larger measurable negligible set.

We show that $\mu$ is a measure. First, $\mu(\varnothing) = L(\mathbb{1}_\varnothing) = L(0) = 0$. Second, if $A_1, A_2, \cdots$ are disjoint sets in $\mathcal{E}$ with union $A$, then the indicator of $\bigcup_{i=1}^n A_i$ is $\sum_{i=1}^n \mathbb{1}_{A_i}$, the latter is increasing to $\mathbb{1}_A$, and hence

$$\mu(A) = L(\mathbb{1}_A) = \lim_n L\Big(\sum_{i=1}^n \mathbb{1}_{A_i}\Big) = \lim_n \sum_{i=1}^n L(\mathbb{1}_{A_i}) = \lim_n \sum_{i=1}^n \mu(A_i) = \sum_{i=1}^\infty \mu(A_i).$$

where we used the conditions (iii) and (ii) to justify the second and third equality signs.

So, $\mu$ is a measure on $(E, \mathcal{E})$. It is unique by the necessity of Equation (1.7). Now, $L(f) = \mu(f)$ for simple $f$ in $\mathcal{E}_+$ by the linearity property (ii) of $L$ and the linearity of integration. This in turn implies that, for every $f$ in $\mathcal{E}_+$, choosing simple $f_n \uparrow f$,

$$L(f) = \lim_n L(f_n) = \lim_n \mu(f_n) = \mu(f)$$

by condition (iii) and the monotone convergence theorem. □

To give an application of Theorem 1.51, let us consider the integral with respect to the measure which is the countable sum of measures.

EXAMPLE 1.52 (Sums of measures). If $\mu_1, \mu_2, \cdots$ are measures on $(E, \mathcal{E})$, then $\mu = \sum_n \mu_n$ is obvious a measure. Show that, for every $f$ in $\mathcal{E}_+$, we have

$$\mu(f) = \sum_n \mu_n(f).$$

PROOF. Define $L : \mathcal{E}_+ \to \bar{\mathbb{R}}_+$ by letting $L(f) = \sum_n \mu_n(f)$. It can be checked that $L$ satisfies the conditions of the integral characterization theorem 1.51. Thus, $L(f) = \nu(f)$ for some *unique* measure $\nu$ on $(E, \mathcal{E})$. That $\nu$ is precisely the measure $\mu = \sum_n \mu_n$ since

$$\nu(B) = L(\mathbb{1}_B) = \sum_n \mu_n(\mathbb{1}_B) = \sum_n \mu_n(B) = \mu(B), \quad \forall\, B \in \mathcal{E}.$$

Thus we have proved the result we desired. □

## 5. Transforms and Indefinite Integrals

**5.1. Image Measures.** This section is about measures defined from other measures via various means and the relationships among integrals with respect to them.

Let $(F, \mathcal{F})$ and $(E, \mathcal{E})$ be measurable spaces. Let $\nu$ be a measure on $(F, \mathcal{F})$ and let $h : F \to E$ be measurable relative to $\mathcal{F}$ and $\mathcal{E}$. We define a mapping $\nu \circ h^{-1}$ from the $\sigma$-algebra $\mathcal{E}$ into $\bar{\mathbb{R}}_+$ by

$$(\nu \circ h^{-1})(B) = \nu(h^{-1}(B)), \quad B \in \mathcal{E},$$

which is well-defined since $h^{-1}(B) \in \mathcal{F}$ by the measurability of $h$. It is easy to check that $\nu \circ h^{-1}$ is a *measure* on $(E, \mathcal{E})$; it is called the **image** of $\nu$ under $h$.

If $\nu$ is *finite*, then so is its image. If $\nu$ is $\Sigma$-*finite*, again, so is its image. BUT, the image of a $\sigma$-finite measure generally *fails* to be $\sigma$-finite (but is $\Sigma$-finite).

The following relates integrals with respect to $\nu \circ h^{-1}$ to integrals with respect to $\nu$.

THEOREM 1.53 (Change of variables formula). *For every $f$ in $\mathcal{E}_+$ we have $(\nu \circ h^{-1})(f) = \nu(f \circ h)$.*

PROOF. Define $L : \mathcal{E}_+ \to \bar{\mathbb{R}}_+$ by letting $L(f) = \nu(f \circ h)$. It can be checked that $L$ satisfies the conditions of the integral characterization theorem 1.51. Thus, $L(f) = \mu(f)$ for some *unique* measure $\mu$ on $(E, \mathcal{E})$. That $\mu$ is precisely the measure $\nu \circ h^{-1}$ since

$$\mu(B) = L(\mathbb{1}_B) = \nu(\mathbb{1}_B \circ h) = \nu(h^{-1}(B)), \quad \forall\, B \in \mathcal{E},$$

where $(\mathbb{1}_B \circ h)(x) = \mathbb{1}_{\{x:\, h(x)\in B\}} = \mathbb{1}_{h^{-1}(B)}$. □

REMARK 1.54. The limitation to positive $\mathcal{E}$-measurable functions can be removed: for arbitrary $f$ in $\mathcal{E}$ the same formula holds provided that the integral on one side be well-defined.

The preceding theorem is a generalization of the *change of variable formula*. In more explicit notation, with $\mu = \nu \circ h^{-1}$, the theorem is that

$$\int_F f(h(x))\,\nu(\mathrm{d}x) = \int_E f(y)\,\mu(\mathrm{d}y), \tag{1.8}$$

that is, if $h(x)$ is replaced with $y$ then $\nu(\mathrm{d}x)$ must be replaced with $\mu(\mathrm{d}y)$. †

Forming image measures is a convenient method of creating new measures from the old, and if the old measure $\nu$ is convenient enough as an integrator, then Theorem 1.53 provides a useful formula for the integrals with respect to the new measure $\mu$.

In fact, the class of measures that can be represented as images of the Lebesgue measure on $\mathbb{R}_+$ is very large. The following is the precise statement; combined with the preceding theorem it reduces integrals over abstract spaces to integrals on $\mathbb{R}_+$ with respect to the Lebesgue measure.

THEOREM 1.55 (Images of the Lebesgue measure). *Let $(E, \mathcal{E})$ be a standard measurable space. Let $\mu$ be a $\Sigma$-finite measure on $(E, \mathcal{E})$ and put $b = \mu(E)$, possibly $+\infty$. Then, there exists a mapping $h$ from $[0, b)$ into $E$, measurable relative to $\mathcal{B}([0, b))$ and $\mathcal{E}$, such that*

$$\mu = \lambda \circ h^{-1},$$

*where $\lambda$ is the Lebesgue measure on $[0, b)$.*

We just omit the proof of this theorem.

**5.2. Indefinite Integrals.** Let $(E, \mathcal{E}, \mu)$ be a measure space. Let $p$ be a positive $\mathcal{E}$-measurable function. Define

$$\nu(A) = \mu(p\mathbb{1}_A) = \int_A p(x)\,\mu(\mathrm{d}x), \quad A \in \mathcal{E}. \tag{1.9}$$

It follows from the monotone convergence theorem that $\nu$ is a measure on $(E, \mathcal{E})$. It is called the **indefinite integral** of $p$ with respect to $\mu$.

LEMMA 1.56 (Indefinite integrals). *For every $f$ in $\mathcal{E}_+$, we have $\nu(f) = \mu(pf)$.*

PROOF. Let $L(f) = \mu(pf)$ and check that $L$ satisfies the conditions of Theorem 1.51. Thus, there exists a unique measure $\hat{\mu}$ on $(E, \mathcal{E})$ such that $L(f) = \hat{\mu}(f)$ for every $f$ in $\mathcal{E}_+$. We have $\hat{\mu} = \nu$ since $\hat{\mu}(A) = L(\mathbb{1}_A) = \mu(p\mathbb{1}_A) = \nu(A)$ for all $A \in \mathcal{E}$. □

REMARK 1.57. The formula (1.9) is another convenient tool for creating new measures from the old. Written in more explicit notation, the preceding proposition becomes

$$\int_E f(x)\,\nu(\mathrm{d}x) = \int_E f(x)p(x)\,\mu(\mathrm{d}x), \quad f \in \mathcal{E}_+,$$

which can be expressed informally by writting

$$\nu(\mathrm{d}x) = p(x)\mu(\mathrm{d}x), \quad x \in E, \tag{1.10}$$

once it is understood that $\mu$ and $\nu$ are measures on $(E, \mathcal{E})$ and that $p$ is positive $\mathcal{E}$-measurable. †

**5.3. Radon-Nikodym Theorem.** What we have done above is to *use a positive $\mathcal{E}$-measurable function to derive a new measure $\nu$ from $\mu$*. Now consider the converse.

DEFINITION 1.58 (Absolutely continuous). Let $\mu$ and $\nu$ be measures on a measurable space $(E, \mathcal{E})$. Then, $\nu$ is said to be **absolutely continuous** with respect to $\mu$ if, for every set $A$ in $\mathcal{E}$,

$$\mu(A) = 0 \quad \Rightarrow \quad \nu(A) = 0.$$

THEOREM 1.59 (Radon-Nikodym theorem). *Suppose that $\mu$ is $\sigma$-finite, and $\nu$ is absolutely continuous with respect to $\mu$. Then, there exists a positive $\mathcal{E}$-measurable function $p$ such that*

$$\int_E f(x)\nu(\mathrm{d}x) = \int_E f(x)p(x)\mu(\mathrm{d}x), \quad f \in \mathcal{E}_+.$$

*Moreover, $p$ is unique up to changes on sets with $\mu$-measure zero.*

For more details about this theorem, I highly recommend you to read: Sheldon Axler. *Measure, Integration & Real Analysis*. Springer, 2020. Chapter 9.

## 6. Kernels and Product Spaces

DEFINITION 1.60 (Transition kernel). Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let $K$ be a mapping from $E \times F$ into $\bar{\mathbb{R}}_+$. Then, $K$ is called a **transition kernel** from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ if

(i) the mapping $x \mapsto K(x, B)$ is $\mathcal{E}$-measurable for every set $B$ in $\mathcal{F}$, and
(ii) the mappint $B \mapsto K(x, B)$ is a measure on $(F, \mathcal{F})$ for every $x$ in $E$.

EXAMPLE 1.61. If $\nu$ is a finite measure on $(F, \mathcal{F})$, and $k$ is a positive function on $E \times F$ that is measurable with respect to the product $\sigma$-algebra $\mathcal{E} \otimes \mathcal{F}$, then it will be seen shortly that

$$K(x, B) = \int_B k(x, y)\, \nu(\mathrm{d}y), \quad x \in E,\ B \in \mathcal{F}$$

defines a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$.

In further special case where $E = \{1, \cdots, m\}$ and $F = \{1, \cdots, n\}$ with their discrete $\sigma$-algebras, the transition kernel $K$ is specified by the numbers $K(x, \{y\})$ and can be regarded as an $m$ by $n$ matrix of positive numbers. †

**6.1. Measure-kernel-function.** We begin with the following theorem:

THEOREM 1.62 (Measure-kernel-function). *Let $K$ be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then,*

$$Kf(x) = \int_F f(y)\, K(x, \mathrm{d}y), \quad x \in E$$

*defines a function $Kf$ that is in $\mathcal{E}_+$ for every function $f$ in $\mathcal{F}_+$;*

$$\mu K(B) = \int_E K(x, B)\, \mu(\mathrm{d}x), \quad B \in \mathcal{F}$$

*defines a measure $\mu K$ on $(F, \mathcal{F})$ for each measure $\mu$ on $(E, \mathcal{E})$; and*

$$(\mu K)(f) = \mu(Kf) = \int_E \Big( \int_F f(y)\, K(x, \mathrm{d}y) \Big)\, \mu(\mathrm{d}x)$$

*for every measure $\mu$ on $(E, \mathcal{E})$ and function $f$ in $\mathcal{F}_+$.*

Obviously, we have

$$K(x,B) = K\mathbb{1}_B(x), \quad B \in \mathcal{F}.$$

PROOF. (i). Let $f \in \mathcal{F}_+$. Then $Kf$ is a well-defined positive function on $E$ since $Kf(x)$ is the integral of $f$ with respect to the measure $B \mapsto K(x,B)$.

We show that $Kf$ is $\mathcal{E}$-measurable in two steps: *First*, if $f$ is simple, say $f = \sum_{i=1}^n b_i \mathbb{1}_{B_i}$, then $Kf(x) = \sum_{i=1}^n b_i K(x,B_i)$, which shows that $Kf$ is $\mathcal{E}$-measurable since it is a linear combination of the $\mathcal{E}$-measurable functions $x \mapsto K(x,B_i)$, $1 \leqslant i \leqslant n$. *Second*, if $f$ in $\mathcal{F}_+$ is not simple, we choose simple $f_n$ in $\mathcal{F}_+$ increasing to $f$; then $Kf(x) = \lim_n Kf_n(x)$ for each $x$ by the monotone convergence theorem for the measure $B \mapsto K(x,B)$; and, hence $Kf$ is $\mathcal{E}$-measurable since it is the limit of $\mathcal{E}$-measurable functions $Kf_n$.

(ii). Fix a measure $\mu$ on $(E,\mathcal{E})$. Define $L : \mathcal{F}_+ \to \bar{\mathbb{R}}_+$ by setting

$$L(f) = \mu(Kf).$$

If $f = 0$ then $L(f) = 0$. If $f$ and $g$ are in $\mathcal{F}_+$, and $a$ and $b$ in $\mathbb{R}_+$, then

$$L(af+bg) = \mu(K(af+bg)) = \mu(aKf+bKg) = a\mu(Kf) + b\mu(Kg) = aL(f) + bL(g),$$

where the second equality is justified by the linearity of the integration with respect to the measure $B \mapsto K(x,B)$ for each $x$, and the third equality by the linearity of the integration with respect to $\mu$.

Finally, if $(f_n) \subset \mathcal{F}_+$ and $f_n \uparrow f$, then $Kf_n(x) \uparrow Kf(x)$ by the monotone convergence theorem for $B \mapsto K(x,B)$, and

$$L(f_n) = \mu(Kf_n) \uparrow \mu(Kf) = L(f)$$

by the monotone convergence theorem of $\mu$. Hence, by Theorem 1.51, there exists a measure $\nu$ on $(F,\mathcal{F})$ such that $L(f) = \nu(f)$ for every $f$ in $\mathcal{F}_+$. Taking $f = \mathbb{1}_B$, we see that $\nu(B) = \mu K(B)$ for every set $B$ in $\mathcal{F}$, that is, $\nu = \mu K$. So, $\mu K$ is a measure on $(F,\mathcal{F})$, and $(\mu K)(f) = \nu(f) = L(f) = \mu(Kf)$ as claimed. □

REMARK 1.63. To specify a kernel $K$ from $(E,\mathcal{E})$ into $(F,\mathcal{F})$ it is more than enough to specify $Kf$ for every $f$ in $\mathcal{F}_+$. *Conversely*, as an extension of Theorem 1.51, it is easy to see that a mapping $f \mapsto Kf$ from $\mathcal{F}_+$ into $\mathcal{E}_+$ specifies a transition kernel $K$ if and only if

a) $K0 = 0$;

b) $K(af+bg) = aKf + bKg$ for $f,g \in \mathcal{F}_+$ and $a,b \in \mathbb{R}_+$;

c) $Kf_n \uparrow Kf$ for every sequence $(f_n) \subset \mathcal{F}_+$ increasing to $f$.

**6.2. Product Kernels, Markov Kernels.** Let $K$ be a transition kernel from $(E,\mathcal{E})$ into $(F,\mathcal{F})$ and let $L$ be a transition kernel from $(F,\mathcal{F})$ into $(G,\mathcal{G})$. Then, their **product** is the transition kernel from $(E,\mathcal{E})$ into $(G,\mathcal{G})$ defined by

$$(KL)f = K(Lf), \quad f \in \mathcal{G}_+.$$

Remark 1.63 can be used to show that $KL$ is indeed a kernel. Obviously,

$$KL(x,B) = \int_F L(y,B)\, K(x,\mathrm{d}y), \quad x \in E, B \in \mathcal{G}.$$

A transition kernel from $(E,\mathcal{E})$ into itself is called simply a transition kernel on $(E,\mathcal{E})$. Such a kernel K is called a **Markov kernel** on $(E,\mathcal{E})$ if $K(x,E) = 1$ for every $x$, and a **sub-Markov kernel** if $K(x,E) \leqslant 1$ for every $x$. If $K$ is a transition kernel on $(E,\mathcal{E})$, its **powers** are kernels on $(E,\mathcal{E})$ defined by

$$K^0 = I,\ K^1 = K,\ K^2 = KK,\ K^3 = KK^2,\ \cdots,$$

where $I$ is the *identity kernel* on $(E, \mathcal{E})$:

$$I(x, A) = \delta_x(A) = \mathbb{1}_A(x), \quad x \in E,\ A \in \mathcal{E}.$$

**6.3. Kernels Finite and Bounded.** Let $K$ be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. In analogy with measures, $K$ is said to be **finite** if $K(x, F) < \infty$ for each $x$, and **$\sigma$-finite** if $B \mapsto K(x, B)$ is $\sigma$-finite for each $x$. It is said to be bounded if $x \mapsto K(x, F)$ is **bounded**, and **$\sigma$-bounded** if there exists a measurable partition $(F_n)$ of $F$ such that $x \mapsto K(x, F_n)$ is *bounded* for each $n$. It is said to be **$\Sigma$-finite** if $K = \sum_{n=1}^{\infty} K_n$ for some sequence of finite kernels $K_n$, and **$\Sigma$-bounded** if the $K_n$ can be chosen to be *bounded*.

In the very special case where $K(x, F) = 1$ for all $x$, the kernel is said to be a **transition probability kernel**. Markov kernels are transition probability kernels.

**6.4. Functions on Product Spaces.** We start by re-stating Example 1.17: sections of a measurable function are measurable.

THEOREM 1.64 (Sections). *If $f \in \mathcal{E} \otimes \mathcal{F}$, then $x \mapsto f(x, y)$ is in $\mathcal{E}$ for each $y$ in $F$, and $y \mapsto f(x, y)$ is in $\mathcal{F}$ for each $x$ in $E$.*

Unfortunately, the *converse* is NOT true: it is possible that the conclusions hold, and yet $f$ is not $\mathcal{E} \otimes \mathcal{F}$-measurable. The following is a generalization of the operation $f \mapsto Kf$ of Theorem 1.62 to functions $f$ defined on the product space.

THEOREM 1.65 (Tonelli, Measure-kernel-function). *Let $K$ be a $\Sigma$-finite kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$. Then, for every positive function $f$ in $\mathcal{E} \otimes \mathcal{F}$,*

$$Tf(x) = \int_F f(x, y)\, K(x, \mathrm{d}y), \quad x \in E, \tag{1.11}$$

*defines a function $Tf$ in $\mathcal{E}_+$. Moreover, the transformation $T : (\mathcal{E} \otimes \mathcal{F})_+ \to \mathcal{E}_+$ is linear and continuous under increasing limits.*

We only prove the case where $K$ is *finite* to simplify the proof.

PROOF. Let $f \in (\mathcal{E} \otimes \mathcal{F})_+$. Then, for each $x \in E$, the section $f_x : y \mapsto f(x, y)$ is a function in $\mathcal{F}_+$ by Theorem 1.64, and $Tf(x)$ is the integral of $f_x$ w.r.t. the measure $K_x : B \mapsto K(x, B)$. Thus, $Tf(x)$ is a well-defined *positive* number for each $x$, and the linearity of $T$ is immediate from the *linearity* of integration w.r.t. $K_x$ for all $x$, and $T$ is *continuous under increasing limits* by the monotone convergence theorem for the measure $K_x$.

We still need to show that $Tf \in \mathcal{E}$. To simplify the proof, we assume that $K$ is bounded. Boundedness of $K$ implies that $Tf$ is bounded for each bounded $f \in \mathcal{E} \otimes \mathcal{F}$, and it is checked easily that

$$\mathcal{M} = \{f \in \mathcal{E} \otimes \mathcal{F} :\ f \text{ is positive or bounded},\ Tf \in \mathcal{E}\}$$

is a monotone class. Moreover, $\mathcal{M}$ includes the indicator of every measurable rectangle $A \times B$, since

$$T\mathbb{1}_{A\times B}(x) = \int_F \mathbb{1}_A(x)\mathbb{1}_B(x)\, K(x, \mathrm{d}y) = \mathbb{1}_A(x)K(x, B)$$

and the right side defines an $\mathcal{E}$-measurable function. Since the measurable rectangles generate the $\sigma$-algebra $\mathcal{E} \otimes \mathcal{F}$, it follows from the monotone class theorem **??** that $\mathcal{M}$ includes all positive (or bounded) $f$ in $\mathcal{E} \times \mathcal{F}$ assuming that $K$ is bounded. □

**6.5. Measures on the Product Space.** The following is the general method for constructing measures on the product space.

THEOREM 1.66 (Measures on the product space). *Let $\mu$ be a measure on $(E, \mathcal{E})$. Let $K$ be a $\Sigma$-finite transition kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$. Then,*

$$\pi f = \int_E \mu(\mathrm{d}x) \int_F f(x,y)\, K(x, \mathrm{d}y), \quad f \in (\mathcal{E} \otimes \mathcal{F})_+ \tag{1.12}$$

*defines a measure $\pi$ on the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$. Moreover, if $\mu$ is $\sigma$-finite and $K$ is $\sigma$-bounded, then $\pi$ is $\sigma$-finite and is the unique measure on that product space satisfying*

$$\pi(A \times B) = \int_A K(x, B)\, \mu(\mathrm{d}x), \quad A \in \mathcal{E},\ B \in \mathcal{F}. \tag{1.13}$$

PROOF. (i). Using the notation of Theorem 1.65, the right side of Equation (1.12) is $\mu(Tf)$, the integral of $Tf$ with respect to $\mu$. *In order to show that $\mu(Tf)$ defines a measure, we need to use Theorem 1.51.* Define $L(f) = \mu(Tf)$ for $f \in (\mathcal{E} \otimes \mathcal{F})_+$. Then $L(0) = 0$, and $L$ is linear since $T$ is linear and integration is linear, and $L$ is continuous under increasing limits by the same property of $T$ and the monotone convergence theorem for $\mu$. Hence, there is a *unique* measure $\pi$ such that $L(f) = \mu(f)$ for all $f \in (\mathcal{E} \otimes \mathcal{F})_+$.

(ii). It is obvious that $\pi$ satisfies Equation (1.13): for all $A \in \mathcal{E}$, $B \in \mathcal{F}$, we have

$$\begin{aligned}\pi(\mathbb{1}_{A\times B}) &= \int_E \mu(\mathrm{d}x) \int_F \mathbb{1}_{A\times B}(x,y)\, K(x, \mathrm{d}y) = \int_E \mathbb{1}_A(x)\, \mu(\mathrm{d}x) \int_F \mathbb{1}_B(y)\, K(x, \mathrm{d}y) \\ &= \int_E \mathbb{1}_A(x) K(x, B)\, \mu(\mathrm{d}x) = \int_A K(x, B)\, \mu(\mathrm{d}x).\end{aligned}$$

Now suppose that $\mu$ is $\sigma$-finite and $K$ is $\sigma$-bounded, there remains to show that $\pi$ is $\sigma$-finite and is the only measure satisfying Equation (1.13). To that end, let $\hat{\pi}$ be another measure satisfying (1.13). Since $\mu$ is $\sigma$-finite, there is a measurable partition $(E_m)$ of $E$ such that $\mu(E_m) < \infty$ for all $m$. Since $K$ is $\sigma$-bounded, there is a measurable partition $(F_n)$ of $E$ such that $x \mapsto K(x, F_n)$ is bounded for each $n$. Note that the measurable rectangles $E_m \times F_n$ form a partition of $E \times F$ and that, by the formula (1.13) for $\pi$ and $\hat{\pi}$,

$$\pi(E_m \times F_n) = \hat{\pi}(E_m \times F_n) < \infty$$

for each $m$ and $n$. Thus, the measures $\pi$ and $\hat{\pi}$ are $\sigma$-finite, they agree on the $\pi$-system of measurable rectangles generating $\mathcal{E} \otimes \mathcal{F}$, and that $\pi$-system contains a partition of $E \times F$ over which $\pi$ and $\hat{\pi}$ are finite. It follows from Corollary 1.27 that $\pi = \hat{\pi}$ on $\mathcal{E} \otimes \mathcal{F}$. □

**6.6. Product Measures and Fubini.** In Theorem 1.66, if the kernel $K$ has the form $K(x, B) = \nu(B)$ for some $\Sigma$-*finite* measure $\nu$ on $(F, \mathcal{F})$, then the measure $\pi$ is called the **product** of $\mu$ and $\nu$ and is denoted by $\mu \times \nu$.

THEOREM 1.67 (Fubini's theorem). *Let $\mu$ and $\nu$ be $\Sigma$-finite measures on $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively.*

(i) *There exists a $\Sigma$-finite measure $\pi$ on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ s.t. for every positive $f$ in $\mathcal{E} \otimes \mathcal{F}$,*

$$\pi(f) = \int_E \mu(\mathrm{d}x) \int_F f(x,y)\, \nu(\mathrm{d}y) = \int_F \nu(\mathrm{d}y) \int_E f(x,y)\, \mu(\mathrm{d}x). \tag{1.14}$$

(ii) *If $f$ in $\mathcal{E} \otimes \mathcal{F}$ and is $\pi$-integrable, then $y \mapsto f(x,y)$ is $\nu$-integrable for $\mu$-almost every $x$; and $x \mapsto f(x,y)$ is $\mu$-integrable for $\nu$-almost every $y$ and (1.14) holds again.*

PROOF. (i). Prove that $\pi$ is a $\Sigma$-finite measure satisfying (1.14). Let $\pi f$ be defined by the first integral in (1.14). Taking $K(x, B) = \nu(B)$ in Theorem 1.66 shows that this defines a measure $\pi$ on the product space. Since $\mu = \sum_i \mu_i$ and $\nu = \sum_j \nu_j$ for some finite measures $\mu_i$ and $\nu_j$, we have

$$\pi(f) = \sum_i \sum_j \int_E \mu_i(\mathrm{d}x) \int_F f(x, y)\, \nu_j(\mathrm{d}y) = \sum_i \sum_j (\mu_i \times \nu_j)(f)$$

by Example 1.52 and the monotone convergence theorem. Thus, $\pi = \sum_i \sum_j \mu_i \times \nu_j$ and, arranging the pairs $(i, j)$ into a sequence, we see that $\pi = \sum_n \pi_n$ for some sequence of finite measures $\pi_n$.

(ii). To prove the equality of the integrals in (1.14), we start by observing the second integral is in fact an integral over $F \times E$: defining $\hat{f} : F \times E \to \bar{\mathbb{R}}_+$ by $\hat{f}(y, x) = f(x, y)$, the second integral is

$$\hat{\pi}(\hat{f}) = \int_F \nu(\mathrm{d}y) \int_E \hat{f}(y, x)\, \mu(\mathrm{d}x) = \sum_j \sum_i \int_F \nu_j(\mathrm{d}y) \int_E \hat{f}(y, x)\, \mu_i(\mathrm{d}x) = \sum_i \sum_j (\nu_j \times \mu_i)(\hat{f}).$$

Hence, to prove that $\pi f = \hat{\pi}\hat{f}$, it is sufficient to show that $(\mu_i \times \nu_j)(f) = (\nu_j \times \mu_i)(\hat{f})$ for each pair $i$ and $j$. If this amount holds for each pair $i$ and $j$, then $\pi(f) = \hat{\pi}(\hat{f})$ since $\hat{\pi}(\hat{f})$ is just a *reordering* of the positive series $\pi(f)$.

(iii). Now let us prove $(\mu_i \times \nu_j)(f) = (\nu_j \times \mu_i)(\hat{f})$. Let $h : E \times F \to F \times E$ be the transposition mapping $(x, y) \mapsto (y, x)$. It is obviously measurable relative to $\mathcal{E} \otimes \mathcal{F}$ and $\mathcal{F} \otimes \mathcal{E}$ using Lemma 1.16. For sets $A \in \mathcal{E}$ and $B \in \mathcal{F}$,

$$\big((\mu_i \times \nu_j) \circ h^{-1}\big)(B \times A) = (\mu_i \times \nu_j)(A \times B) = \mu_i(A)\nu_j(B) = \nu_j(B)\mu_i(A) = (\nu_j \times \mu_i)(B \times A),$$

where the second equality is justified by Equation (1.14). And the equality above implies via Theorem 1.25 that $(\mu_i \times \nu_j) \circ h^{-1} = \nu_j \times \mu_i$ since they are both finite measures. Hence,

$$(\nu_j \times \mu_i)(\hat{f}) = \big((\mu_i \times \nu_j) \circ h^{-1}\big)(\hat{f}) = (\mu_i \times \nu_j)(\hat{f} \circ h^{-1}) = (\mu_i \times \nu_j)(f).$$

(iv). Finally, let us prove (ii). Let $f$ be *integrable* with respect to $\pi$. Then (1.14) holds for $f^+$ and $f^-$ separately, and $\pi(f) = \pi(f^+) - \pi(f^-)$ with both terms *finite*[4]. Hence, (1.14) holds for $f$.

As to the integrability of sections, we observe that $x \mapsto \int_F f(x, y)\, \nu(\mathrm{d}y)$ is measurable[5] by letting $\nu(\mathrm{d}y) = K(x, \mathrm{d}y)$ and referring to Theorem 1.65; and is real-valued for $\mu$-almost every $x$ using the integrability of $f$, that is, $y \mapsto f(x, y)$ is $\nu$-integrable for $\mu$-almost every $x$. By symmetry, the finiteness for the second integral implies that $x \mapsto f(x, y)$ is $\mu$-integrable for $\nu$-almost every $y$. □

REMARK 1.68 (Fubini).

(i) Since we have more than one measure, for notions like integrability and negligibility, one needs to point out the measure associated. So, $\pi$-integrable means "integrable with respect to the measure $\pi$".

(ii) It is clear from (1.14) that

$$\pi(A \times B) = \mu(A)\nu(B), \quad A \in \mathcal{E}, B \in \mathcal{F}, \tag{1.15}$$

and for this reason we call $\pi$ the product of $\mu$ and $\nu$ and we use the notation $\pi = \mu \times \nu$.

(iii) If both $\mu$ and $\nu$ are $\sigma$-finite, then Theorem 1.66 applies with $K(x, B) = \nu(B)$ and implies that $\pi$ is the *only* measure satisfying (1.15). †

[4]Since $f$ is integrable with respect to $\pi$. Recall the definition of "integrable".

[5]In some measure theory textbooks, we call this result the *Tonelli-Fubini theorem*.

**6.7. Finite Products.** The concepts and results above extend easily to products of *finitely many* spaces. Let $(E_1, \mathcal{E}_1), \cdots, (E_n, \mathcal{E}_n)$ be measurable spaces. Their **product** is denoted by any of the following three:

$$\bigotimes_{i=1}^{n}(E_i, \mathcal{E}_i) = \Big(\prod_{i=1}^{n} E_i, \bigotimes_{i=1}^{n} \mathcal{E}_i\Big) = (E_1 \times \cdots \times E_n,\ \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n),$$

where $E_1 \times \cdots \times E_n$ is the set of all $n$-tuples $(x_1, \cdots, x_n)$ with $x_i$ in $E_i$ for $1 \leqslant i \leqslant n$, and $\mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n$ is the $\sigma$-algebra generated by the *measurable rectangles* $A_1 \times \cdots \times A_n$ with $A_i$ in $\mathcal{E}_i$, $1 \leqslant i \leqslant n$.

Let $\mu_1, \cdots, \mu_n$ be $\Sigma$-finite measures on $(E_1, \mathcal{E}_1), \cdots, (E_n, \mathcal{E}_n)$ respectively. Then, their **product** $\pi = \mu_1 \times \cdots \times \mu_n$ is the measure defined on the measurable product space by analogy with Fubini's theorem 1.67: for positive functions $f$ in $\otimes_i \mathcal{E}_i$,

$$\pi(f) = \int_{E_1} \mu_1(\mathrm{d}x_1) \int_{E_2} \mu_2(\mathrm{d}x_2) \cdots \int_{E_n} \mu_n(\mathrm{d}x_n)\, f(x_1, \cdots, x_n). \tag{1.16}$$

It is usual to denote the resulting measure space $\bigotimes_{i=1}^{n}(E_i, \mathcal{E}_i, \mu_i)$. Fubini's theorem is generalized to this space and shows that, if $f$ is positive or $\pi$-integrable, the integrals on the right side of (1.16) can be performed in any order desired.

More general measures can be defined on the product space with the help of kernels. We illustrate the skill for $n = 3$: Let $\mu_1$ be a measure on $(E_1, \mathcal{E}_1)$, let $K_2$ be a transition kernel from $(E_1, \mathcal{E}_1)$ into $(E_2, \mathcal{E}_2)$, and let $K_3$ be a transition kernel from $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ into $(E_3, \mathcal{E}_3)$. Consider the formula

$$\pi(f) = \int_{E_1} \mu_1(\mathrm{d}x_1) \int_{E_2} K_2(x_1, \mathrm{d}x_2) \int_{E_3} K_3((x_1, x_2), \mathrm{d}x_3)\, f(x_1, x_2, x_3) \tag{1.17}$$

for positive $f$ in $\mathcal{E}_1 \otimes \mathcal{E}_2 \otimes \mathcal{E}_3$. Assuming that $K_2$ and $K_3$ are $\Sigma$-finite, repeated applications of Theorem 1.66 show that this defines a measure $\pi$ on $(E_1 \times E_2 \times E_3,\ \mathcal{E}_1 \otimes \mathcal{E}_2 \otimes \mathcal{E}_3)$.

We can omit as many parentheses as we can and use a notation analogous to (1.10). For instance, instead of (1.17), we write

$$\pi(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) = \mu_1(\mathrm{d}x_1) K_2(x_1, \mathrm{d}x_2) K_3(x_1, x_2, \mathrm{d}x_3).$$

The notation

$$\pi = \mu_1 \times K_2 \times K_3$$

is also used for the same thing and is in accord with the notation for product measures.

**6.8. Infinite Products.** Let $T$ be an arbitrary set, countable or uncountable. It will play the role of an index set; we think of it as the *time*. For each $t$ in $T$, let $(E_t, \mathcal{E}_t)$ be a measurable space. Let $x_t$ be a point in $E_t$ for each $t$ in $T$. Then we write $(x_t)_{t \in T}$ for the resulting collection and think of it as a function on $T$; this is especially appropriate when $(E_t, \mathcal{E}_t) = (E, \mathcal{E})$ for all $t$, because, then, $x = (x_t)_{t \in T}$ can be regarded as the mapping $t \mapsto x_t$ from $T$ into $E$. The set $F$ of all such functions $x = (x_t)_{t \in T}$ is called the **product space** defined by $(E_t; t \in T)$; and the notation $\prod_{t \in T} E_t$ is used for $F$.

A rectangle in $F$ is a subset of the form

$$\prod_{t \in T} A_t = \{x \in F : x_t \in A_t \text{ for each } t \text{ in } T\}, \tag{1.18}$$

where $A_t$ *differs* from $E_t$ for only a *finite* number of $t$. It is said to be measurable if $A_t \in \mathcal{E}_t$ for every $t$ (for which $A_t$ differs from $E_t$). The $\sigma$-algebra on $F$ generated by the collection of all measurable

rectangles is called the **product $\sigma$-algebra** and is denoted by $\bigotimes_{t\in T}\mathcal{E}_t$. The resulting measurable space is denoted variously by

$$\bigotimes_{t\in T}(E_t,\mathcal{E}_t)=\Big(\prod_{t\in T}E_t,\bigotimes_{t\in T}\mathcal{E}_t\Big).$$

In the special case where $(E_t,\mathcal{E}_t)=(E,\mathcal{E})$ for all $t$, the following notations are also in use for same:

$$(E,\mathcal{E})^T=(E^T,\mathcal{E}^T).$$

Although this is the logical point to describe the construction of measures on the product space, we shall delay it until the end of Chapter 4, at which point the steps involved should look intuitive. For the present, we list the following proposition which allows an arbitrary collection of measurable functions to be thought as one measurable function. It is a many-dimensional generalization of the result in Lemma 1.16.

THEOREM 1.69 (Measurable function on infinite product space). *Let $(\Omega,\mathcal{H})$ be a measurable space. Let $(F,\mathcal{F})=\bigotimes_{t\in T}(E_t,\mathcal{E}_t)$. For each $t$ in $T$, let $f_t$ be a mapping from $\Omega$ into $E_t$. For each $\omega$ in $\Omega$, define $f(\omega)$ to be the point $(f_t(\omega))_{t\in T}$ in $F$. Then, the mapping $f:\Omega\to F$ is measurable relative to $\mathcal{H}$ and $\mathcal{F}$ if and only if $f_t$ is measurable relative to $\mathcal{H}$ and $\mathcal{E}_t$ for all $t$ in $T$.*

PROOF. *First* suppose that $f$ is measurable relative to $\mathcal{H}$ and $\mathcal{F}$. Then, $\{f\in B\}\in\mathcal{H}$ for all $B$ in $\mathcal{F}$. In particular, taking $B$ to be the rectangle in (1.18) with $A_t=E_t$ for all $t$ except $t=s$ for some fixed $s$, we see that $\{f\in B\}=\{f_s\in A_s\}\in\mathcal{H}$ for $A_s$ in $\mathcal{E}_s$. Thus, $f_s$ is measurable relative to $\mathcal{H}$ and $\mathcal{E}_s$ for every $s$ fixed.

*Next*, suppose that each $f_t$ is measurable relative to $\mathcal{H}$ and $\mathcal{E}_t$. If $B$ is a measurable rectangle in $F$, then $\{f\in B\}$ is the intersection of finitely many sets of the form $\{f_t\in A_t\}$ with $A_t$ in $\mathcal{E}_t$, and hence, $\{f\in B\}\in\mathcal{H}$. Since measurable rectangles generate the product $\sigma$-algebra $\mathcal{F}$, this implies via Theorem 1.5 that $f$ is measurable relative to $\mathcal{H}$ and $\mathcal{F}$. □

CHAPTER 2

# Probability Space

## 1. Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The set $\Omega$ is called the **sample space**; its elements are called **outcomes**. The $\sigma$-algebra $\mathcal{F}$ may be called the grand history; its elements are called **events**.

We repeat the properties of the probability measure $\mathbb{P}$, all sets here are events:

$$\begin{aligned}
\text{Norming}: &\quad \mathbb{P}[\varnothing] = 0, \quad \mathbb{P}[\Omega] = 1;\\
\text{Monotonicity}: &\quad H \subset K \quad \Rightarrow \quad \mathbb{P}[H] \leqslant \mathbb{P}[K];\\
\text{Finite additivity}: &\quad H \cap K = \varnothing \quad \Rightarrow \quad \mathbb{P}[H \cup K] = \mathbb{P}[H] + \mathbb{P}[K];\\
\text{Countable additivity}: &\quad (H_n) \text{ disjoint} \quad \Rightarrow \quad \mathbb{P}\Big[\bigcup_n H_n\Big] = \sum_n \mathbb{P}[H_n];\\
\text{Sequential continuity}: &\quad H_n \uparrow H \quad \Rightarrow \quad \mathbb{P}[H_n] \uparrow \mathbb{P}[H];\\
&\quad H_n \downarrow H \quad \Rightarrow \quad \mathbb{P}[H_n] \downarrow \mathbb{P}[H];\\
\text{Boole's inequality}: &\quad \mathbb{P}\Big[\bigcup_n H_n\Big] \leqslant \sum_n \mathbb{P}[H_n].
\end{aligned}$$

All of these are as before for arbitrary measures, except for the *sequential continuity under decreasing limits*, which is made possible by the finiteness of $\mathbb{P}$.

A subset $N$ of $\Omega$ is said to be **negligible** if there exists an *event* $H$ such that $N \subset H$ and $\mathbb{P}[H] = 0$. The probability space is said to be **complete** if every negligible set is an event. It is generally nicer to have $(\Omega, \mathcal{F}, \mathbb{P})$ complete. If it is not, it can be completed. An event is said to be **almost sure** if its probability is one. If a proposition holds for every outcome $\omega$ in an almost sure event, then we say that the proposition holds *almost surely* or *almost everywhere* or for almost every $\omega$ or **with probability one**. Obviously, the concept is *equivalent* to having the proposition fail only over a negligible set.

**1.1. Random Variables.** Let $(E, \mathcal{E})$ be a measurable space. A mapping $X : \Omega \to E$ is called a **random variable** taking values in $(E, \mathcal{E})$ provided that it be measurable relative to $\mathcal{F}$ and $\mathcal{E}$, that is, if

$$X^{-1}(A) = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$$

is an event for every $A$ in $\mathcal{E}$. *Of course, it is sufficient to check the condition for $A$ in a collection that generates $\mathcal{E}$.*

If the $\sigma$-algebra is understood from context, then we merely say that $X$ takes values in $E$ or that $X$ is **$E$-valued**. This is especially the case if $E$ is $\mathbb{R}$ or $\mathbb{R}^d$ or some Borel subset of some such space and $\mathcal{E}$ is the Borel $\sigma$-algebra.

The simplest random variables are indicators of events; we use the usual notation $\mathbb{1}_H$ for the indicator of $H$. A random variable is **simple** if it takes only finitely many values, all in $\mathbb{R}$. It is said to be **discrete** if it is elementary, that is, if it takes only countably many values.

**1.2. Distribution.** Let $X$ be a random variable taking values in some measurable space $(E, \mathcal{E})$. Let $\mu$ be the image of $\mathbb{P}$ under $X$, that is,

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

where the last member is read as "the probability that $X$ is in $A$". Then, $\mu$ is a probability measure on $(E, \mathcal{E})$; it is called the **distribution** of $X$.

From Chapter 1, to specify the distribution $\mu$, it is sufficient to specify $\mu(A)$ for all $X$ belonging to a $\pi$-system that generates $\mathcal{E}$. In particular, if $E = \bar{\mathbb{R}}$ and $\mathcal{E} = \mathcal{B}(E)$, the intervals $[-\infty, x]$ with $x$ in $\mathbb{R}$ form a convenient $\pi$-system; consequently, in this case, it is enough to specify

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

The resulting function $F : \mathbb{R} \to [0, 1]$ is called the **distribution function** of $X$.

It should be noticed that, *each random variable can uniquely determine a distribution measure, but the reverse fails*. For instance, let $X \sim \mathcal{N}(0, 1)$ and $Y = -X$. Then $X$ and $Y$ have the same distribution measure, but $X \neq Y$.

**1.3. Functions of Random Variables.** Let $X$ be a random variable taking values in $(E, \mathcal{E})$. Let $(G, \mathcal{G})$ be another measurable space, and let $f : E \to G$ be measurable relative to $\mathcal{E}$ and $\mathcal{G}$. Then, the composition $Y = f \circ X$ of $X$ and $f$, namely,

$$Y(\omega) = f \circ X(\omega) = f(X(\omega)), \quad \omega \in \Omega,$$

is a *random variable* taking values in $(G, \mathcal{G})$; this follows from Theorem 1.6 of Chapter 1. If $\mu$ is the distribution of $X$, then the distribution $\nu$ of $Y$ is $\nu = \mu \circ f^{-1}$:

$$\nu(B) = \mathbb{P}[Y \in B] = \mathbb{P}[X \in f^{-1}(B)] = \mu(f^{-1}(B)) = (\mu \circ f^{-1})(B), \quad B \in \mathcal{G}.$$

**1.4. Joint Distributions.** Let $X$ and $Y$ be random variables taking values in measurable space $(E, \mathcal{E})$ and $(G, \mathcal{G})$ respectively. Then, the pair

$$Z = (X, Y) : \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$$

is measurable relative to $\mathcal{F}$ and the product $\sigma$-algebra $\mathcal{E} \otimes \mathcal{G}$, that is, $Z$ is a random variable taking values in the product space $(E \times G, \mathcal{E} \otimes \mathcal{G})$.

The distribution of $Z$ is a probability measure $\pi$ on the product space and is also called the **joint distribution** of $X$ and $Y$. Since $\mathcal{E} \otimes \mathcal{F}$ is generated by the $\pi$-system of measurable rectangles, in order to specify $\pi$ it is sufficient to specify

$$\pi(A \times B) = \mathbb{P}[X \in A, Y \in B], \quad A \in \mathcal{E}, B \in \mathcal{G},$$

the right side being the probability that $X$ is in $A$ and $Y$ is in $B$, that is, the probability of $\{X \in A\} \cap \{Y \in B\}$. In the opposite direction, given the joint distribution $\pi$, for $A$ in $\mathcal{E}$ and $B$ in $\mathcal{G}$, we have

$$\mu(A) = \mathbb{P}[X \in A] = \pi(A \times G), \quad \nu(B) = \mathbb{P}[Y \in B] = \pi(E \times B).$$

In this context, the probability measures $\mu$ and $\nu$ are called the **marginal distributions** of $X$ and $Y$ respectively.

**1.5. Independence.** Let $X$ and $Y$ be random variables taking values in measurable space $(E, \mathcal{E})$ and $(G, \mathcal{G})$ respectively, and let $\mu$ and $\nu$ be their respective (marginal) distributions. Then, $X$ and $Y$ are said to be **independent** if their joint distribution is the product measure formed by their marginals, that is, if the distribution of the pair $(X, Y)$ is the product measure $\mu \times \nu$, or in still in other words,

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B], \quad A \in \mathcal{E}, B \in \mathcal{G}.$$

Independence of $X$ and $Y$ is a convenient tool for specifying the joint distribution via its marginals. We shall return to these matters in Chapter 4 for a rigorous treatment.

A *finite collection* $\{X_1, \cdots, X_n\}$ of random variables is said to be an **independency**, or the variables $X_1, \cdots, X_n$ are said to be **independent**, if the distribution of the random vector $(X_1, \cdots, X_n)$ has the product form $\mu_1 \times \cdots \times \mu_n$ where $\mu_1, \cdots, \mu_n$ are probability measures. Then, necessarily, $\mu_i$ is the distribution of $X_i$ for each $i$. An *arbitrary collection* (countable or uncountable) of random variables is said to be an **independency** if every finite sub-collection of it is an **independency**.

**1.6. Stochastic Processes.** Let $(E, \mathcal{E})$ be a measurable space. Let $T$ be an arbitrary set, countable or uncountable. For each $t$ in $T$, let $X_t$ be a random variable taking values in $(E, \mathcal{E})$. Then, the collection $\{X_t : t \in T\}$ is called a **stochastic process** with **state space** $(E, \mathcal{E})$ and **parameter set** $T$.

For each $\omega$ in $\Omega$, let $X(\omega)$ denote the function $t \mapsto X_t(\omega)$ from $T$ into $E$; then, $X(\omega)$ is an element of $E^T$. By Theorem 1.69 theorem of Chapter 1, the mapping $X : \omega \mapsto X(\omega)$ from $\Omega$ into $E^T$ is measurable relative to $\mathcal{F}$ and $\mathcal{E}^T$. In other words, we may regard the stochastic process $(X_t)_{t\in T}$ as a *random variable* $X$ that takes values in $(G, \mathcal{G}) = (E^T, \mathcal{E}^T)$.

The distribution of the random variable $X$, that is, the probability measure $\mathbb{P} \circ X^{-1}$ on $(G, \mathcal{G})$, is called the **probability law** of the stochastic process $(X_t)_{t\in T}$.

Recall that the product $\sigma$-algebra $\mathcal{G}$ is generated by the finite-dimensional rectangles and, therefore, a probability measure on $(G, \mathcal{G})$ is determined by the values it assigns to those rectangles. It follows that the probability law of $X$ is determined by the values

$$\mathbb{P}[X_{t_1} \in A_1, \cdots, X_{t_n} \in A_n] \tag{2.1}$$

with $n$ ranging over $\mathbb{N}$, and $t_1, \cdots, t_n$ over $T$, and $A_1, \cdots, A_n$ over $\mathcal{E}$. Much of the theory of stochastic processes has to do with computing integrals concerning $X$ from the given data regarding Equation (2.1).

## 2. Expectations

Throughout this section $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and all random variables are defined on $\Omega$ and take values in $\bar{\mathbb{R}}$, unless stated otherwise.

Let $X$ be a random variable. Since it is $\mathcal{F}$-measurable, its integral with respect to $\mathbb{P}$ makes sense to talk about. That integral is called the **expectation** of $X$ and is denoted by any of the following

$$\mathbb{E}X = \mathbb{E}[X] = \int_\Omega X(\omega)\, \mathbb{P}[\mathrm{d}\omega] = \int_\Omega X \, \mathrm{d}\mathbb{P}.$$

The expected value $\mathbb{E}X$ exists if and only if the integral does, that is, if and only if we do not have $\mathbb{E}X^+ = \mathbb{E}X^- = +\infty$. Of course, $\mathbb{E}X$ exists whenever $X \geqslant 0$, and $\mathbb{E}X$ exists and is finite if $X$ is bounded.

We shall state most results for positive random variables, because expectations exist always for such, and because the extensions to arbitrary random variables are generally obvious. The basic properties of integrals can all be found in Chapter 1, which we will not repeat here.

The following relates expectations, which are integrals with respect to $\mathbb{P}$, to integrals with respect to *distributions*.

THEOREM 2.1 (Expectation and integrals w.r.t. distribution). *Let $X$ be a random variable taking values in some measurable space $(E, \mathcal{E})$. If $\mu$ is the distribution of $X$, then*

$$\mathbb{E}[f \circ X] = \mu(f) \tag{2.2}$$

*for every $f$ in $\mathcal{E}_+$. Conversely, if Equation (2.2) holds for some measure $\mu$ and all $f$ in $\mathcal{E}_+$, then $\mu$ is the distribution of $X$.*

PROOF. The first statement has been proved in Theorem 1.53, just consider the integration with respect to image measures: if $\mu = \mathbb{P} \circ X^{-1}$, then $\mu(f) = \mathbb{E}[f \circ X]$ at least for $f$ in $\mathcal{E}_+$.

Conversely, if (2.2) holds for all $f$ in $\mathcal{E}_+$, taking $f = \mathbb{1}_A$ in particular, we see that

$$\mu(A) = \mu(\mathbb{1}_A) = \mathbb{E}[\mathbb{1}_A \circ X] = \mathbb{P}[X \in A],$$

that is, $\mu$ is the distribution of $X$. □

REMARK 2.2. The converse statement of the theorem is useful for figuring out the distribution of $X$ in cases where $X$ is a known function of other random variables whose joint distribution is known. Obviously, for a measure $\mu$ to be the distribution of $X$ it is sufficient to have (2.2) hold for all $f$ having the form $f = \mathbb{1}_A$ with $A$ in $\mathcal{E}$, or with $A$ in some $\pi$-system generating $\mathcal{E}$. †

**2.1. Moments & Inequalities.** Let $X$ be a random variable taking values in $\bar{\mathbb{R}}$ and having the distribution $\mu$. The expectation of the $n$th power of $X$, namely $\mathbb{E}[X^n]$, is called the $n$**th moment** of $X$. In particular, $\mathbb{E}[X]$ is called the **mean** of $X$. Assuming that the mean is *finite*, say $\mathbb{E}[X] = \mu$, the $n$th moment of $X - \mu$ is called the $n$**th centered moment** of $X$. In particular, $\mathbb{E}[(X-\mu)^2]$ is called the **variance** of $X$, and we shall denote it by $\mathrm{Var}(X)$; note that

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

LEMMA 2.3 (Moments of positive variables). *Let $X$ be a positive random variable, then for all $p$ in $\mathbb{R}_+$,*

$$\mathbb{E}[X^p] = \int_0^\infty p x^{p-1} \mathbb{P}\{X > x\} \, \mathrm{d}x$$

PROOF. Noting that $X^p(\omega) = \int_0^{X(\omega)} p x^{p-1} \, \mathrm{d}x = \int_0^\infty p x^{p-1} \mathbb{1}_{\{X>x\}}(\omega) \, \mathrm{d}x$, then we can use Fubini's theorem with the product measure $\mathbb{P} \times Leb$ to derive the desired result. □

We may derive some frequently used inequalities in probability theory here.

LEMMA 2.4 (Markov's inequality). *Let $X$ be a positive random variable, then*

$$\mathbb{P}\{X \geqslant b\} \leqslant \frac{1}{b}\mathbb{E}[X]$$

*holds for all $b > 0$.*

PROOF. Noting that $X \geqslant X\mathbb{1}_{\{X \geqslant b\}} \geqslant b\mathbb{1}_{\{X \geqslant b\}} \geqslant 0$, thus

$$\mathbb{E}[X] \geqslant \mathbb{E}[X\mathbb{1}_{\{X \geqslant b\}}] \geqslant b\mathbb{E}[\mathbb{1}_{\{X \geqslant b\}}] > 0$$

holds by the monotonicity of expectation (or integration). Since $\mathbb{E}[\mathbb{1}_{\{X \geqslant b\}}] = \mathbb{P}\{X \geqslant b\}$, then we get

$$\mathbb{P}\{X \geqslant b\} \leqslant \frac{1}{b}\mathbb{E}[X]$$

from the above inequality. □

LEMMA 2.5 (Chebyshev's inequality). *If $\varphi$ is a strictly positive and increasing function on $(0,\infty)$ with $\varphi(u)=\varphi(-u)$, and $X$ is a random variable such that $\mathbb{E}[\varphi(X)]<\infty$, then for each $u>0$:*

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

PROOF. Since $\varphi$ is strictly positive and increasing, then

$$\varphi(X)\geqslant\varphi(X)\mathbb{1}_{\{|X|\geqslant u\}}\geqslant\varphi(u)\mathbb{1}_{\{|X|\geqslant u\}}.$$

Then, by the monotonicity of expectation, we get

$$\mathbb{E}[\varphi(X)]\geqslant\varphi(X)\mathbb{E}[\mathbb{1}_{\{|X|\geqslant u\}}]\geqslant\varphi(u)\mathbb{E}[\mathbb{1}_{\{|X|\geqslant u\}}].$$

Finally, we get the desired result since $\mathbb{E}[\mathbb{1}_{\{|X|\geqslant u\}}]=\mathbb{P}\{|X|\geqslant u\}$. □

LEMMA 2.6 (Jensen's inequality). *Let $X$ have finite mean. Let $\varphi$ be **convex** on $\mathbb{R}$, that is, $\varphi=\sup_n\varphi_n$ for some sequence of functions $\varphi_n$ having the form $\varphi_n(x)=a_n+b_nx$. Then*

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

PROOF. Using again the monotonicity of expectation, we know that

$$\mathbb{E}[\varphi(X)]=\mathbb{E}[\sup_n\varphi_n(X)]\geqslant\sup_n\mathbb{E}[\varphi_n(X)]=\sup_n\left\{a_n+b_n\mathbb{E}X\right\}=\varphi(\mathbb{E}X),$$

where the second inequality is justified since $\sup_n\varphi_n(X)\geqslant\varphi_n(X)$ for all $n$ in $\mathbb{N}$, which implies that $\mathbb{E}[\sup_n\varphi_n(X)]\geqslant\mathbb{E}[\varphi_n(X)]$ for all $n$ in $\mathbb{N}$. □

REMARK 2.7 (Equivalent definition of convex function). Indeed, $\varphi$ is convex on $\mathbb{R}$ if and only if, for all $x,y$ in $\mathbb{R}$, $\lambda$ in $[0,1]$, we have

$$\varphi\big(\lambda x+(1-\lambda)y\big)\leqslant\lambda\varphi(x)+(1-\lambda)\varphi(y). \tag{2.3}$$

PROOF OF REMARK 2.7. The *necessity* is obvious, just using the definition of convex function mentioned in Lemma 2.6. Now consider the *sufficiency* part.

First, for all $z<x<y$, we have

$$\frac{\varphi(x)-\varphi(z)}{x-z}\leqslant\frac{\varphi(y)-\varphi(z)}{y-z}\leqslant\frac{\varphi(y)-\varphi(z)}{y-z}$$

using Equation (2.3). Thus, for all $h_n\downarrow 0$, we have

$$\frac{\varphi(x)-\varphi(x-h_m)}{h_m}\leqslant\frac{\varphi(x)-\varphi(x-h_n)}{h_n}\leqslant\frac{\varphi(x+h_n)-\varphi(x)}{h_n}\leqslant\frac{\varphi(x+h_m)-\varphi(x)}{h_m}$$

holds for all $m<n$. Then we have

$$\varphi'_-(x)=\lim_{h_n\downarrow 0}\frac{\varphi(x)-\varphi(x-h_n)}{h_n}\leqslant\lim_{h_n\downarrow 0}\frac{\varphi(x+h_n)-\varphi(x)}{h_n}=\varphi'_+(x)$$

by the monotonicity showed before.

Now, for all $x\in\mathbb{R}$, let $a$ be any number between the two limits and let $\ell(z)=a(z-x)+\varphi(z)$, then $\ell(x)=\varphi(x)$ and $\varphi(z)\geqslant\ell(z)$ for all $z$ in $\mathbb{R}$. Thus, for any $x_n\in\mathbb{Q}$, we can find a line $\varphi_n$ satisfies these properties. Then, $\varphi(x)=\sup_n\varphi_n(x)$ holds for all $x$ in $\mathbb{Q}$. Finally, we can extend this to the real line, since $\mathbb{Q}$ is dense in $\mathbb{R}$, and $\varphi$ and $\sup_n\varphi_n$ are both continuous on $\mathbb{Q}$.[1] □

[1] For more details, you can find some mathematical analysis book for reference.

**2.2. A Useful Estimation.** Let us prove the following useful theorem as an instructive example.

EXAMPLE 2.8 (A Useful Estimate of Mathematical Expectation). We have

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

so that $\mathbb{E}[|X|]<\infty$ if and only if the series above converges.

PROOF. Using the linearity of integration, if $\Lambda_n=\{n\leqslant|X|<n+1\}$, then

$$\mathbb{E}[|X|]=\sum_{n=0}^{\infty}\int_{\Lambda_n}|X|\,\mathrm{d}\mathbb{P}.$$

Hence by the monotonicity of integration applied to each set $\Lambda_n$:

$$\sum_{n=0}^{\infty}n\mathbb{P}[\Lambda_n]\leqslant\mathbb{E}[|X|]\leqslant\sum_{n=0}^{\infty}(n+1)\mathbb{P}[\Lambda_n]=1+\sum_{n=0}^{\infty}\mathbb{P}[\Lambda_n]. \tag{2.4}$$

It remains to show

$$\sum_{n=0}^{\infty}n\mathbb{P}[\Lambda_n]=\sum_{n=0}^{\infty}\mathbb{P}[|X|\geqslant n]. \tag{2.5}$$

finite or infinite. Now the partial sums of the series on the left may be rearranged to yield, for $N\geqslant 1$,

$$\begin{aligned}\sum_{n=0}^{N}n\mathbb{P}[\Lambda_n]&=\sum_{n=0}^{N}n\big(\mathbb{P}[|X|\geqslant n]-\mathbb{P}[|X|\geqslant n+1]\big)\\&=\sum_{n=0}^{N}n\mathbb{P}[|X|\geqslant n]-\sum_{n=1}^{N+1}(n-1)\mathbb{P}[|X|\geqslant n]\\&=\sum_{n=1}^{N}\big(n-(n-1)\big)\mathbb{P}[|X|\geqslant n]-N\mathbb{P}[|X|\geqslant N+1]\\&=\sum_{n=1}^{N}\mathbb{P}[|X|\geqslant n]-N\mathbb{P}[|X|\geqslant N+1],\end{aligned}$$

where the first equality holds since $\mathbb{P}[\Lambda_n]=\mathbb{P}[|X|\geqslant n]-\mathbb{P}[|X|\geqslant n+1]$. Thus we have

$$\sum_{n=1}^{N}n\mathbb{P}[\Lambda_n]\leqslant\sum_{n=1}^{N}\mathbb{P}[|X|\geqslant n]\leqslant\sum_{n=1}^{N}n\mathbb{P}[\Lambda_n]+N\mathbb{P}[|X|\geqslant N+1]. \tag{2.6}$$

Another application of the monotonicity of integral gives

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

Hence if $\mathbb{E}[|X|]<\infty$, then the last term in (2.6) converges to zero as $N\to\infty$[2] and (2.5) becomes

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

with both sides finite. On the other hand, if $\mathbb{E}[|X|]=\infty$, here, from (2.4) we have

$$\sum_{n=1}^{\infty}n\mathbb{P}[\Lambda_n]=\infty.$$

And then from (2.6) we get $\sum_{n=1}^{\infty}\mathbb{P}[|X|\geqslant n]=\infty$. Thus, (2.5) is true and we complete the proof. □

[2]We can use the absolute continuity of integral to derive this result, since $|X|$ is integrable and $\mathbb{P}[|X|\geqslant N+1]$ converges to zero as $N\to\infty$.

## 3. $L^p$ Space and Uniform Integrability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $X$ be a real-valued random variable. For $p \geqslant 1$, define

$$\|X\|_p = \left(\mathbb{E}[|X|^p]\right)^{1/p},$$

and for $p = \infty$ let

$$\|X\|_\infty = \inf\{b \in \mathbb{R}_+ : |X| \leqslant b \text{ almost surely}\}.$$

It is easy to see that

$$\|X\|_p = 0 \quad \Rightarrow \quad X = 0 \text{ almost surely}.$$

For each $1 \leqslant p \leqslant \infty$, let $L^p$ denote the collection of all real-valued random variables $X$ with $\|X\|_p < \infty$. For $1 \leqslant p < \infty$, $X$ is in $L^p$ if and only if $|X|^p$ is integrable; and $X$ is in $L^\infty$ if and only if $X$ is almost surely bounded. For $X$ in $L^p$, the number $\|X\|_p$ is called the $L^p$**-norm** of $X$; in particular, $\|X\|_\infty$ is called the **essential supremum** of $X$.

These concepts can be deduced into the case of general measure space easily. And we will discuss two important inequalities in the general case below.

### 3.1. Inequalities.

The following theorem summarizes the various connections.

DEFINITION 2.9 (Conjugate exponents). For $1 \leqslant p \leqslant \infty$, the **conjugate exponent** of $p$, denoted by $q$, is the element of $[1, \infty]$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

We agree on that the conjugate exponent of 1 is $\infty$, and the conjugate exponent of $\infty$ is 1.

THEOREM 2.10 (Hölder's inequality). *Let $p$ and $q$ be conjugate exponents. Then for all $f$ in $L^p$, $g$ in $L^q$, we have $fg$ in $L^1$ and*

$$\|fg\|_1 \leqslant \|f\|_p \|g\|_q.$$

PROOF. Suppose $f$ and $g$ are measurable function on $(E, \mathcal{E}, \mu)$. *First* consider the case where $p = 1$ and $q = \infty$, then $|(fg)(x)| \leqslant \|g\|_\infty |f(x)|$ almost surely. By the monotonicity of integral,

$$\mu(|fg|) \leqslant \|g\|_\infty \mu(|f|),$$

that is, $\|fg\|_1 \leqslant \|f\|_1 \|g\|_\infty$. *Second*, consider the case where $1 < p < \infty$. It is obviously to show that

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}$$

holds for all positive $a$ and $b$ (Fix a $b$, and using derivative w.r.t. $a$ to get the desired result).

Assume either $\|f\|_p = 0$ or $\|g\|_q = 0$, then $fg = 0$ almost surely since either $f = 0$ or $g = 0$ holds almost surely.

Then, assume $\|f\|_p = 1$ and $\|g\|_q = 1$, then $|fg| \leqslant p^{-1}|f|^p + q^{-1}|g|^q$ holds for all $x$ in $E$. Then,

$$\|fg\|_1 = \mu(|fg|) \leqslant p^{-1}\mu(|f|^p) + q^{-1}\mu(|g|^q) = p^{-1} + q^{-1} = 1 = \|f\|_p \|g\|_q$$

is justified by the monotonicity of integration.

Finally, assume the general case except the case either $\|f\|_p$ or $\|g\|_q$ equal to zero. Define $\tilde{f} = f/\|f\|_p$ and $\tilde{g} = g/\|g\|_q$, then $\|\tilde{f}\|_p = \|\tilde{g}\|_q = 1$. Then,

$$\int_E \left|\frac{fg}{\|f\|_p \|g\|_q}\right| \mathrm{d}\mu = \frac{1}{\|f\|_p \|g\|_q} \int_E |fg| \, \mathrm{d}\mu \leqslant 1$$

holds by the above case. Then the desired result holds from the last inequality. □

In the study of convergence of measurable function, we will need to discuss inclusion relations in $L^p$ spaces. But we need to assume a ***finite*** measure space in advance. Without loss of generality, we focus on the probability space.

THEOREM 2.11 (Inclusion relation between $L^p$ and $L^q$). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $0 < p < q < \infty$. Then $\|X\|_p \leqslant \|X\|_q$ holds for all $X$ in $L^q$. Furthermore, $L^q \subset L^p$.*

PROOF. Suppose $X \in L^q$, and $r = q/p > 1$. Then the conjugate exponent of $r$ is $\tilde{r} = q/(q-p)$. Then

$$\mathbb{E}[|X|^p] \leqslant \big(\mathbb{E}[|X|^{pr}]\big)^{1/r}\big(\mathbb{E}[1^{\tilde{r}}]\big)^{1/\tilde{r}} = \big(\mathbb{E}[|X|^{pr}]\big)^{1/r}.$$[3]

Taking both ends of the inequality to the $1/p$th power, then

$$\|X\|_p = \big(\mathbb{E}[|X|^p]\big)^{1/p} \leqslant \big(\mathbb{E}[|X|^q]\big)^{1/q} = \|X\|_q$$

holds, which means that $X \in L^p$. □

THEOREM 2.12 (Minkowski's inequality). *Let $1 \leqslant p \leqslant \infty$, then $\|f+g\|_p \leqslant \|f\|_p + \|g\|_p$ holds for all $f$ and $g$ in $L^p$.*

PROOF. The cases $p = 1$ and $p = \infty$ are obvious, just by using the inequality $|f+g| \leqslant |f| + |g|$. So assume that $1 < p < \infty$. Writing

$$|f+g|^p \leqslant (|f| + |g|)^p \leqslant \big(2\max\{|f|, |g|\}\big)^p \leqslant 2^p\big(|f|^p + |g|^p\big),$$

we see that $\mu(|f+g|^p) < \infty$ and thus $f + g \in L^p$. Then, by integrating the inequality

$$|f+g|^p = |f+g|\,|f+g|^{p-1} \leqslant |f|\,|f+g|^{p-1} + |g|\,|f+g|^{p-1}$$

with respect to $\mu$, we get $\mu(|f+g|^p) \leqslant \mu(|f|\,|f+g|^{p-1}) + \mu(|g|\,|f+g|^{p-1})$. By applying the Hölder's inequality to the conjugate exponents $p$ and $q = p/(p-1)$, we get

$$\mu(|f+g|^p) \leqslant \|f\|_p \mu(|f+g|^p)^{(p-1)/p} + \|g\|_p \mu(|f+g|^p)^{(p-1)/p}.$$

If $\mu(|f+g|^p) = 0$, the inequality is trivial. Otherwise, we can divide each side of the preceding inequality by $\mu(|f+g|^p)^{(p-1)/p}$ and we get the desired result. □

**3.2. Uniform Integrability.** This concept plays an important role in *martingale* theory and in the convergence of sequences in the space $L^1$. We start by illustrating the issue involved in the simplest setting.

LEMMA 2.13 (A necessary and sufficient condition of integrable). *Let $X$ be a real-valued random variable. Then $X$ is integrable if and only if*

$$\lim_{b\to\infty} \mathbb{E}[|X|\mathbb{1}_{\{|X|>b\}}] = 0. \tag{2.7}$$

PROOF. Let $Z_b = |X|\mathbb{1}_{\{|X|>b\}}$, then $Z_b$ is dominated by $|X|$ and goes to 0 as $b \to \infty$. Thus, if $X$ is integrable, the dominated convergence yields that $\lim_{b\to\infty} \mathbb{E}[Z_b] = 0$, which is Equation (2.7).

Conversely, if Equation (2.7) holds, then we can choose $b$ large enough to have $\mathbb{E}[Z_b] < 1$, and the inequality $|X| = |X|\mathbb{1}_{\{|X|\leqslant b\}} + |X|\mathbb{1}_{\{|X|>b\}} \leqslant b + Z_b$ shows that $\mathbb{E}[|X|] \leqslant b + 1 < \infty$. □

For a collection of random variables $X$, the uniform integrability of the collection has to do with the possibility of taking the limit in (2.7) uniformly in $X$:

[3] We need the finiteness of measure here.

DEFINITION 2.14 (Uniform integrability). A collection $\mathcal{K}$ of real-valued random variables is said to be **uniformly integrable** if

$$k(b) = \sup_{X\in\mathcal{K}} \mathbb{E}[|X|\mathbb{1}_{\{|X|>b\}}]$$

goes to 0 as $b \to \infty$.

By definition, we have the following basic properties:

REMARK 2.15 (Some basic properties of uniformly integrable random variables).

(i) If $\mathcal{K}$ is *finite* and each $X$ in it is integrable, then $\mathcal{K}$ is *uniformly integrable*.
(ii) If $\mathcal{K}$ is *dominated* by an *integrable* random variable $Z$, then it is uniformly integrable.
(iii) Uniform integrability implies $L^1$-boundedness, that is, if $\mathcal{K}$ is uniformly integrable then $\mathcal{K} \subset L^1$ and $k(0) = \sup_{\mathcal{K}} \mathbb{E}[|X|] < \infty$.
(iv) But $L^1$-boundedness is insufficient for uniform integrability.
(v) However, if $\mathcal{K}$ is $L^p$-bounded for some $p > 1$ then it is uniformly integrable.

PROOF OF REMARK 2.15. (ii). If $|X| \leqslant Z$ for all $X$ in $\mathcal{K}$, then $k(b) \leqslant \mathbb{E}[Z\mathbb{1}_{\{Z>b\}}]$ and the last expectation goes to 0 by Lemma 2.13 applied to $Z$.

(iii). Note that $|X| = |X|\mathbb{1}_{\{|X|\leqslant b\}} + |X|\mathbb{1}_{\{|X|>b\}} \leqslant b + |X|\mathbb{1}_{\{|X|>b\}}$, and take expectation yields

$$\mathbb{E}[|X|] \leqslant b + \mathbb{E}[|X|\mathbb{1}_{\{|X|>b\}}] \leqslant b + k(b)$$

for all $X$ in $\mathcal{K}$. Besides, since $\mathcal{K}$ is uniformly integrable, we can choose a finite number $b$ such that $k(b) \leqslant 1$. Then $X$ is $L^1$ bounded.

(v). This proposition will be showed below: see Theorem 2.19 and take $f(x) = x^p$. □

EXAMPLE 2.16 ($L^1$-boundedness is insufficient for uniform integrability). Suppose that $\Omega = (0, 1)$ with its Borel $\sigma$-algebra for events and the Lebesgue measure as $\mathbb{P}$. Let

$$\begin{aligned} X_n(\omega) &= n \quad \text{if } \omega \leqslant 1/n, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Then, $\mathbb{E}[X_n] = 1$ for all $n$ in $\mathbb{N}$, that is, $\mathcal{K} = (X_n)_{n\geqslant 1}$ is $L^1$-bounded. But $k(b) = 1$ for all $b$ since $\mathbb{E}[X_n\mathbb{1}_{\{X_n\geqslant b\}}] = \mathbb{E}[X_n] = 1$ for $n > b$. †

EXAMPLE 2.17 (Crystal ball condition). For $p > 0$, the collection $\mathcal{K}$ is uniformly integrable, if

$$\sup_{X\in\mathcal{K}} \mathbb{E}[|X|^{p+\delta}] < \infty$$

for some $\delta > 0$. For example, suppose $\mathcal{K}$ is a sequence of random variables satisfying $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] = 1$ for all $X$ in $\mathcal{K}$, then $\mathcal{K}$ is uniformly integrable.

PROOF. To verify sufficiency of the crystal ball condition, write

$$\begin{aligned} \sup_{X\in\mathcal{K}} \mathbb{E}\Big[|X|^p\mathbb{1}_{\{|X|^p>b\}}\Big] &= \sup_{X\in\mathcal{K}} \mathbb{E}\left[|X|^p\mathbb{1}_{\{|X|/b^{1/p}>1\}}\right] = \sup_{X\in\mathcal{K}} \mathbb{E}\left[|X|^p \cdot 1 \cdot \mathbb{1}_{\{|X|^\delta/b^{\delta/p}>1\}}\right] \\ &\leqslant \sup_{X\in\mathcal{K}} \mathbb{E}\left[|X|^p \cdot \frac{|X|^\delta}{b^{\delta/p}}\right] = \frac{1}{b^{\delta/p}} \sup_{X\in\mathcal{K}} \mathbb{E}\Big[|X|^{p+\delta}\Big] \to 0 \end{aligned}$$

as $b \to \infty$, so we get the desired result. □

The following $\varepsilon - \delta$ characterization is the main result on uniform integrability: over all small set, the integrals of $X$ are uniformly small.

THEOREM 2.18 ($\varepsilon-\delta$ characterization). *The collection $\mathcal{K}$ is uniformly integrable if and only if the following are satisfied:*

(i) ***$L^1$-bounded***. $\sup_{X\in\mathcal{K}}\mathbb{E}[|X|]<\infty$.

(ii) ***uniform absolute continuity***. *For every $\varepsilon>0$, there is $\delta>0$ such that for all event $H$,*

$$\mathbb{P}[H]\leqslant\delta \quad\Rightarrow\quad \sup_{X\in\mathcal{K}}\mathbb{E}[|X|\mathbb{1}_H]\leqslant\varepsilon. \tag{2.8}$$

PROOF. Suppose that all $X$ are positive; this amounts to working with $|X|$ throughout. Since

$$X\mathbb{1}_H\leqslant X\mathbb{1}_{\{X\leqslant b\}\cap H}+X\mathbb{1}_{\{X>b\}\cap H}\leqslant b\mathbb{1}_H+X\mathbb{1}_{\{X>b\}}$$

for every event $H$ and every $b$ in $\mathbb{R}_+$, then

$$\sup_{X\in\mathcal{K}}\mathbb{E}[X\mathbb{1}_H]\leqslant b\mathbb{P}[H]+k(b),\quad b\in\mathbb{R}.$$

Consider the *necessity*. Suppose that $\mathcal{K}$ is uniformly integrable. Then, $\mathcal{K}$ is $L^1$-bounded by the Remark above. Also, since $k(b)\to 0$, by definition, for all $\varepsilon>0$, there is $b<\infty$ such that $k(b)\leqslant\varepsilon/2$. By setting $\delta=\varepsilon/2b$ we see that

$$\sup_{X\in\mathcal{K}}\mathbb{E}[X\mathbb{1}_H]\leqslant b\mathbb{P}[H]+k(b)\leqslant b\cdot\frac{\varepsilon}{2b}+\frac{\varepsilon}{2}=\varepsilon$$

holds for all events $H$ such that $\mathbb{P}[H]<\delta$, i.e. $\mathcal{K}$ is uniformly absolutely continuous.

Next, consider the *sufficiency*. Suppose that $\mathcal{K}$ is $L^1$-bounded and that for all $\varepsilon>0$, there is $\delta>0$ such that Equation (2.8) holds for all event $H$. Then, Markov's inequality yields

$$\sup_{X\in\mathcal{K}}\mathbb{P}[X>b]\leqslant\frac{1}{b}\sup_{X\in\mathcal{K}}\mathbb{E}[X]=\frac{1}{b}k(0),$$

where $k(0)=\sup_{X\in\mathcal{K}}\mathbb{E}[X]$ is finite. So there is $b$ such that $\mathbb{P}[X>b]\leqslant\delta$ for all $X$ in $\mathcal{K}$, and, then, for that $b$ we have $k(b)\leqslant\varepsilon$ in view of (2.8) used with $H=\{X>b\}$. In other words, for all $\varepsilon>0$, there is $b<\infty$ such that $k(b)<\varepsilon$, which is the definition of uniform integrability. □

The following theorem is very useful for showing uniform integrability.

THEOREM 2.19 (Superlinear growth implies uniform integrability). *Suppose that there is a positive Borel function $f$ on $\mathbb{R}_+$ such that $\lim_{x\to\infty}f(x)/x=\infty$ and*

$$\sup_{X\in\mathcal{K}}\mathbb{E}[f\circ|X|]<\infty.$$

*Then, $\mathcal{K}$ is uniformly integrable.*

PROOF. We may assume that all $X$ are positive. Also, by replacing $f$ with $f\vee 1$ if necessary[4], we assume that $f\geqslant 1$ in addition to satisfying the stated conditions. Let $g(x)=x/f(x)$ and note that

$$X\mathbb{1}_{\{X>b\}}=(f\circ X)(g\circ X)\mathbb{1}_{\{X>b\}}\leqslant(f\circ X)\sup_{x>b}g(x).$$

This shows that

$$k(b)=\sup_{X\in\mathcal{K}}\mathbb{E}[X\mathbb{1}_{\{X>b\}}]\leqslant\sup_{x>b}g(x)\cdot\sup_{X\in\mathcal{K}}\mathbb{E}[f\circ X],$$

and the right side goes to 0 as $b\to\infty$ since $g(x)\to 0$ as $x\to\infty$. □

We supplement the preceding proposition by a converse and give another characterization.

[4]This assumption ensures that $f\neq 0$, thereby guaranteeing the well-definedness of $g$.

THEOREM 2.20 (De la Vallee-Poussin theorem). *The following are equivalent:*

(i) $\mathcal{K}$ *is uniformly integrable.*
(ii) $h(b) = \sup_{\mathcal{K}} \int_b^\infty \mathbb{P}\{|X| > y\}\,\mathrm{d}y \to 0$ *as* $b \to \infty$.
(iii) $\sup_{\mathcal{K}} \mathbb{E}[f \circ |X|] < \infty$ *for some increasing convex function* $f$ *on* $\mathbb{R}_+$ *with* $\lim_{x\to\infty} f(x)/x = +\infty$.

PROOF. We have showed (iii) $\Rightarrow$ (i) in Theorem 2.19. We now show that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii), again assuming, as we say, that all the $X$ in $\mathcal{K}$ are positive.

(i) $\Rightarrow$ (ii). Suppose $\mathcal{K}$ is uniformly integrable, for all $X$ in $\mathcal{K}$,

$$\mathbb{E}[X\mathbb{1}_{\{X>b\}}] = \int_0^\infty \mathbb{P}[X\mathbb{1}_{\{X>b\}} > y]\,\mathrm{d}y = \int_0^\infty \mathbb{P}[X > b \vee y]\,\mathrm{d}y \geqslant \int_b^\infty \mathbb{P}[X > y]\,\mathrm{d}y,$$

where the first equality is justified by Lemma 2.3. Thus, $k(b) \geqslant h(b)$ for all $b$ in $\mathbb{R}_+$, and the uniform integrability of $\mathcal{K}$ means that $k(b) \to 0$ as $b \to \infty$. Hence, (i) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (iii). Since $h(b) \to 0$ as $b \to \infty$, we can pick $0 = b_0 < b_1 < \cdots$ increasing to $+\infty$ such that

$$h(b_n) \leqslant \frac{h(0)}{2^n} \tag{2.9}$$

holds for all $n$ in $\mathbb{N}$. Note that $h(0)$ is finite since

$$\int_0^\infty \mathbb{P}[X > y]\,\mathrm{d}y = \int_0^b \mathbb{P}[X > y]\,\mathrm{d}y + \int_b^\infty \mathbb{P}[X > y]\,\mathrm{d}y \leqslant b + \int_b^\infty \mathbb{P}[X > y]\,\mathrm{d}y,$$

that is, $h(0) \leqslant b + h(b)$ and $h(b)$ can be made as small as desired. Define

$$g(x) = \sum_{n=0}^\infty \mathbb{1}_{[b_n,\infty)}(x), \quad f(x) = \int_0^x g(y)\,\mathrm{d}y, \quad x \in \mathbb{R}_+;$$

note that $g \geqslant 1$ and is increasing toward $+\infty$, which implies that $f$ is increasing and convex and $\lim_{x\to\infty} f(x)/x = +\infty$. Now for all $X$ in $\mathcal{K}$, we have

$$\mathbb{E}[f \circ X] = \mathbb{E}\Big[\int_0^X g(y)\,\mathrm{d}y\Big] = \sum_{n=0}^\infty \mathbb{E}\Big[\int_{b_n}^\infty \mathbb{1}_{\{X>y\}}\,\mathrm{d}y\Big] \leqslant \sum_{n=0}^\infty h(b_n) \leqslant 2h(0) < \infty,$$

where the second inequality holds by (2.9). Now we see that (ii) $\Rightarrow$ (iii) by Theorem 2.19. □

## 4. Information and Determinability

**4.1. $\sigma$-algebra Generated by Random Variable.** This section is on $\sigma$-algebras generated by random variables and measurability with respect to them. We shall argue that such a $\sigma$-algebra should be thought as a body of *information*, and measurability with respect to it should be equated to being *determined* by that information. We may always assume $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

DEFINITION 2.21 ($\sigma$-algebra generated by random variable). Let $X$ be a random variable taking values in some measurable space $(E, \mathcal{E})$. Then

$$\sigma(X) = X^{-1}(\mathcal{E}) = \{X^{-1}(A) : A \in \mathcal{E}\} \tag{2.10}$$

is a $\sigma$-algebra contained in $\mathcal{F}$. It is called the **$\sigma$-algebra generated by** $X$. Indeed, $\sigma(X)$ is the *smallest* $\sigma$-algebra $\mathcal{G}$ on $\Omega$ such that $X$ is measurable with respect to $\mathcal{G}$ and $\mathcal{E}$.

The formula of $\sigma(X)$ is easy, since, on the one hand the $\sigma$-algebra generated by $X$ must contain all events of the form $X^{-1}(A)$, $A \in \mathcal{E}$, and, on the other hand, the collection of all such events forms a $\sigma$-algebra.

Let $T$ be an arbitrary index set, countable or uncountable. For each $t$ in $T$, let $X_t$ be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. Then

$$\sigma(X_t, t \in T) = \bigvee_{t \in T} \sigma(X_t) \tag{2.11}$$

denotes the $\sigma$-algebra on $\Omega$ generated by the union of $\sigma$-algebras $\sigma(X_t)$, $t \in T$. It is called the **$\sigma$-algebra generated** by the collection $(X_t)_{t \in T}$. Similarly, it is the smallest $\sigma$-algebra $\mathcal{G}$ on $\Omega$ such that, for every $t$ in $T$, the random variable $X_t$ is measurable with respect to $\mathcal{G}$ and $\mathcal{E}_t$.

If $X$ is a random variable taking values in $\bigotimes_{t \in T}(E_t, \mathcal{E}_t)$, we denote by $X_t(\omega)$ the value of the function $X(\omega)$ at the point $t$ in $T$; the resulting mapping $\omega \mapsto X_t(\omega)$ is a random variable with values in $(E_t, \mathcal{E}_t)$ and is called the **$t$-coordinate** of $X$.

LEMMA 2.22 ($\sigma$-algebra generated by random vector). *If $X = (X_t)_{t \in T}$, then $\sigma(X) = \sigma(X_t, t \in T)$.*

PROOF. Review Theorem 1.69. Let $\mathcal{H}$ there be $\sigma(X)$ to conclude that $\sigma(X) \supset \sigma(X_t, t \in T)$, and then let $\mathcal{H}$ be $\sigma(X_t, t \in T)$ to conclude that $\sigma(X_t, t \in T) \supset \sigma(X)$. □

The following theorem shows that a random variable is $\sigma(X)$-measurable if and only if it is a deterministic measurable function of $X$. In other words, the collection $\sigma(X)$ of *random variables* is exactly the set of all measurable functions of $X$.

THEOREM 2.23 (Characterization of $\sigma(X)$). *Let $X$ be a random variable taking values in measurable space $(E, \mathcal{E})$. A mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to $\sigma(X)$ if and only if*

$$V = f \circ X$$

*for some deterministic function $f$ in $\mathcal{E}$.*

PROOF. *Sufficiency* is obvious since measurable functions of measurable functions are measurable. That is, every $V$ having the form $f \circ X$ for some $f$ in $\mathcal{E}$ is $\sigma(X)$-measurable where $X$ is $\sigma(X)$-measurable.

Now consider *necessity*. Define

$$\mathcal{M} = \{V = f \circ X : f \in \mathcal{E}\},$$

we shall use the monotone class theorem for functions to show that $\mathcal{M} \supset \sigma(X)$. First, we need to show that $\mathcal{M}$ is a monotone class of functions on $\Omega$.

(i). $1 \in \mathcal{M}$ since $1 = f \circ X$ with $f(x) = 1$ for all $x$ in $E$.

(ii). Let $U$ and $V$ be bounded and in $\mathcal{M}$, and let $a$ and $b$ be in $\mathbb{R}$. Then, $U = f \circ X$ and $V = g \circ X$ for some $f$ and $g$ in $\mathcal{E}$, and thus, $aU + bV = h \circ X$ with $h = af + bg$. Since $h \in \mathcal{E}$, it follows that $aU + bV \in \mathcal{M}$.

(iii). Let $(V_n) \subset \mathcal{M}_+$ and $V_n \uparrow V$. For each $n$, there is $f_n$ in $\mathcal{E}$ such that $V_n = f_n \circ X$. Then, $f = \sup_n f_n$ belongs to $\mathcal{E}$ and since $V_n \uparrow V$,

$$V(\omega) = \sup_n V_n(\omega) = \sup_n f_n(X(\omega)) = f(X(\omega)), \quad \omega \in \Omega,$$

which shows that $V$ in $\mathcal{M}$. Now $\mathcal{M}$ is a monotone class of functions.

Furthermore, $\mathcal{M}$ includes every indicator variable in $\sigma(X)$: if $H$ in $\sigma(X)$, then $H = X^{-1}(A)$ for some $A$ in $\mathcal{E}$, and $\mathbb{1}_H = \mathbb{1}_A \circ X \in \mathcal{M}$. Thus, by monotone class theorem, $\mathcal{M}$ contains all positive random variables in $\sigma(X)$.

Finally, let $V$ in $\sigma(X)$ be arbitrary. Then, $V^+ \in \sigma(X)$ and is positive, and hence, $V^+ = g \circ X$ for some $g$ in $\mathcal{E}$; similarly, $V^- = h \circ X$ for some $h$ in $\mathcal{E}$. Thus, $V = V^+ - V^- = f \circ X$, where

$$f(x) = (g(x) - h(x))\mathbb{1}_{\{g \wedge h = 0\}}(x).$$

This completes the proof since $f$ in $\mathcal{E}$. □

Put $X = (X_n)_{n \in \mathbb{N}}$ and use Theorem 2.23 we can get the following corollary immediately.

COROLLARY 2.24 (Characterization of $\sigma(X_n, n \in \mathbb{N})$). *For each $n$ in $\mathbb{N}$, let $X_n$ be a random variable taking values in some measurable space $(E_n, \mathcal{E}_n)$. A mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to $\sigma(X_n, n \in \mathbb{N})$ if and only if $V = f(X_1, X_2, \cdots)$ for some $f$ in $\bigotimes_n \mathcal{E}_n$.*

The preceding corollary can be generalized to *uncountable* collections $(X_t, t \in T)$ by using the same device of regarding the collection as one random variable.

In fact, there is a certain amount of *simplification*, reflecting the fact that uncountable products of $\sigma$-algebras $\mathcal{E}_t$, $t \in T$, are in fact generated by the finite-dimensional rectangles.

THEOREM 2.25 (Characterization of $\sigma(X_t, t \in T)$). *Let $T$ be arbitrary. For each $t$ in $T$, let $X_t$ be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. Then, $V : \Omega \to \bar{\mathbb{R}}$ belongs to $\sigma(X_t, t \in T)$ if and only if there exists a sequence $(t_n)$ in $T$ and a function $f$ in $\bigotimes_n \mathcal{E}_{t_n}$ such that*

$$V = f(X_{t_1}, X_{t_2}, \cdots). \tag{2.12}$$

PROOF. *Sufficiency* of the condition is trivial: if $V$ has the form (2.12), then $V \in \sigma(X_{t_n} : n \geqslant 1)$ by Corollary 2.24 above, and $\sigma(X_{t_n} : n \geqslant 1) \subset \sigma(X_t, t \in T)$ obviously.

To show the *necessity*, we use the monotone class theorem for functions together with Lemma 2.23. To that end, let

$$\mathcal{M} = \{V = f(t_1, t_2, \cdots) : \text{ where } (t_n) \subset T,\ f \in \textstyle\bigotimes_n \mathcal{E}_{t_n}\},$$

it is easy to check that $\mathcal{M}$ is a monotone class. We *claim* that $\mathcal{M}$ includes the indicators of a $\pi$-system $\mathcal{G}_0$ that generates $\mathcal{G} := \sigma(X_t, t \in T)$. If this claim is true, then by the monotone class theorem for functions, $\mathcal{M}$ includes all positive $V$ in $\mathcal{G}$, and then, all $V$ in $\mathcal{G}$ since $V = V^+ - V^-$ is obviously in $\mathcal{M}$ if $V^+$ and $V^-$ are in $\mathcal{M}$. Hence, $\mathcal{M} \supset \mathcal{G}$ as desired.

Now let us prove this claim. First we know $\mathcal{G} = \sigma(X)$, where $X = (X_t)_{t \in T}$ takes values in $(E, \mathcal{E}) := \bigotimes_{t \in T}(E_t, \mathcal{E}_t)$, by Lemma 2.23. Recall that $\mathcal{E}$ is generated by the $\pi$-system of all finite-dimensional measurable rectangles. Therefore, the pre-image $X^{-1}(A)$ of those rectangles $A$ form a $\pi$-system $\mathcal{G}_0$ that generates $\mathcal{G}$. Thus, to complete the proof, it is sufficient to show that the indicator of $X^{-1}(A) = \{X \in A\}$ belongs to $\mathcal{M}$ for every such rectangle $A$.

Let $A$ be such a rectangle, that is, $A = \prod_{t \in T} A_t$ with $A_t = E_t$ for all $t$ outside a finite subset $S$ of $T$ and $A_t \in \mathcal{E}_t$ for every $t$ in $S$. Then,

$$\mathbb{1}_{\{X \in A\}} = \mathbb{1}_A \circ X = \Pi_{t \in S} \mathbb{1}_{A_t} \circ X_t,$$

which has the form (2.12), that is, belongs to $\mathcal{M}$. □

**4.2. Filtrations.** Our aim is to use the foregoing to argue that a $\sigma$-algebra on $\Omega$ is the mathematically precise equivalent of the everyday term "*information*". And, random quantities that are determined by that information are precisely the random variables that are measurable with respect to that $\sigma$-algebra.

We are interested in a random experiment taking place over an *infinite* expanse of time. Let $T = \mathbb{R}_+$ or $T = \mathbb{N}$ be the time set. For each time $t$, let $\mathcal{F}_t$ be the "information" gathered during $[0, t]$ by an

observer of the experiment. Then, for $s < t$, we must have $\mathcal{F}_s \subset \mathcal{F}_t$. The family $\mathcal{F} = (\mathcal{F}_t)_{t\in T}$, then, depicts the flow of information as the experiment progresses over time.

DEFINITION 2.26 (Filtrations). Let $T$ be a subset of $\mathbb{R}$. For each time $t$ in $T$, let $\mathcal{F}_t$ be a sub-$\sigma$-algebra of $\mathcal{F}$. The family $(\mathcal{F}_t)_{t\in T}$ is called a **filtration** provided that $\mathcal{F}_s \subset \mathcal{F}_t$ for $s < t$.

EXAMPLE 2.27 (Filtration generated by s stochastic processes). If $X = (X_t)_{t\in T}$ is a stochastic process, then $\mathcal{F}_t = \sigma(X_s : T \ni s \leqslant t)$ yields a filtration $(\mathcal{F}_t)_{t\in T}$. †

The following theorem allows us to approximation eternal random variables by random variables that become known in finite time.

THEOREM 2.28 (Approximation of eternal variables). *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration and put $\mathcal{F}_\infty = \bigvee_{n\in\mathbb{N}} \mathcal{F}_n$. For each bounded random variable $V$ in $\mathcal{F}_\infty$, there are bounded variables $V_n$ in $\mathcal{F}_n$, $n \in \mathbb{N}$, such that*

$$\lim_n \mathbb{E}[|V_n - V|] = 0.^a$$

[a] We will introduce in Chapter 3 that, $V_n$ converges to $V$ in the sence of $L^1$.

REMARK 2.29. In the preceding theorem, the $V_n$ are shown to exist but are unspecified. A very specific version will appear later employing totally new tools; see the *martingale convergence theorem*. †

PROOF. Let $\mathcal{C} = \bigcup_n \mathcal{F}_n$. By definition, $\mathcal{F}_\infty = \sigma(\mathcal{C})$. Obviously $\mathcal{C}$ is a $\pi$-system. We want to use the monotone class theorem for functions to prove this theorem. Let

$$\mathcal{M}_b = \{V \in \mathcal{F}_\infty : V \text{ is } \textit{bounded} \text{ and has the } \textit{approximation property} \text{ described}\}.$$

It is easy to see that $\mathcal{M}_b$ includes *constants*; and the *indicators* of events in $\mathcal{C}$[5]. Besides, $\mathcal{M}_b$ is a vector space: for all $V, W$ in $\mathcal{M}_b$ and $a, b$ in $\mathbb{R}$, we can find $V_n, W_n$ in $\mathcal{F}_n$ for all $n$ in $\mathbb{R}$, such that $\mathbb{E}[|V_n - V|] \to 0$ and $\mathbb{E}[|W_n - W|] \to 0$. Then we have $\mathbb{E}[|aV + bW - (aV_n + bW_n)|] \to 0$ using triangle inequality. Thus, $\mathcal{M}_b$ will include all bounded $V$ in $\mathcal{F}_\infty$ once we check the remaining monotonicity condition.

Let $(U_k) \subset \mathcal{M}_b$ be positive and increasing to a bounded variable $V$ in $\mathcal{F}_\infty$. Then, for each $k \geqslant 1$ there are $U_{k,n}$ in $\mathcal{F}_n$, $n \in \mathbb{N}$, such that $\mathbb{E}[|U_{k,n} - U_k|] \to 0$ as $n \to \infty$. Put $n_0 = 0$, and for each $k \geqslant 1$ we can choose $n_k > n_{k-1}$ such that $\hat{U}_k = U_{k,n_k}$ satisfies $\mathbb{E}[|\hat{U}_k - U_k|] < k^{-1}$. Moreover, since $(U_k)$ is bounded and converges to $V$, the bounded convergence theorem implies that $\mathbb{E}[|U_k - V|] \to 0$. Hence,

$$\mathbb{E}[|\hat{U}_k - V|] \leqslant \mathbb{E}[|\hat{U}_k - U_k|] + \mathbb{E}[|U_k - V|] \to 0$$

as $k \to \infty$. With $n_0 = 0$ choose $V_0 = 0$ and put $V_n = \hat{U}_k$ for all integers $n$ in $(n_k, n_{k+1}]$; then, $V_n \in \mathcal{F}_{n_k} \subset \mathcal{F}_n$, and $\mathbb{E}[|V_n - V|] \to 0$ as $n \to \infty$ in view of the above equation. This is what we need to show that $V$ in $\mathcal{M}_b$. □

## 5. Independence

**5.1. Definitions.** Independence is a truly probabilistic concept. Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. For a fixed integer $n \geqslant 2$, let $\mathcal{F}_1, \cdots, \mathcal{F}_n$ be sub-$\sigma$-algebra of $\mathcal{F}$. Then $(\mathcal{F}_i)_{i=1}^n$ is called an **independency** if

$$\mathbb{E}[V_1 \cdots V_n] = \mathbb{E}[V_1] \cdots \mathbb{E}[V_n] \tag{2.13}$$

[5] You can refer Exercise I.1.18 of Erhan Çinlar's *Probability and Stochastics* for a representation of sets in $\mathcal{C}$, we do not cover this example in this notes.

for all positive random variables $V_1, \cdots, V_n$ in $\mathcal{F}_1, \cdots, \mathcal{F}_n$ respectively.

Let $T$ be an arbitrary index set. Let $\mathcal{F}_t$ be a sub-$\sigma$-algebra of $\mathcal{F}$ for each $t$ in $T$. The collection $(\mathcal{F}_t)_{t\in T}$ is called an **independency** if its every finite subset is an independence. In general, elements of an independency are said to be **independent**, or **mutually independent** if emphasis is needed. In loose language, given some objectives, the objects are said to be *independent* if the $\sigma$-algebras generated by those objectives are independent.

EXAMPLE 2.30 (Independency). For example, a random variable $X$ and a stochastic process $(Y_t)_{t\in T}$ and a collection $(\mathcal{F}_i)_{i\in I}$ of $\sigma$-algebras on $\Omega$ are said to be independent if $\mathcal{G}_1 = \sigma(X)$, $\mathcal{G}_2 = \sigma(Y_t; t \in T)$, $\mathcal{G}_3 = \bigvee_{i\in I} \mathcal{F}_i$ are independent, that is, if $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ is an independency. †

**5.2. Independence of $\sigma$-algebras.** Since a collection of sub-$\sigma$-algebras of $\mathcal{F}$ is an independency if and only if its every finite subset is an independency, we concentrate on the independency of a fintie number of sub-$\sigma$-algebras of $\mathcal{F}$. We start with a test for independence.

THEOREM 2.31 (Equivalent condition of independence). *Let $\mathcal{F}_1, \cdots, \mathcal{F}_n$ be sub-$\sigma$-algebras of $\mathcal{F}$, $n \geqslant 2$. For each $i \leqslant n$, let $\mathcal{C}_i$ be a $\pi$-system that generates $\mathcal{F}_i$. Then, $\mathcal{F}_1, \cdots, \mathcal{F}_n$ are independent if and only if*

$$\mathbb{P}[H_1 \cap \cdots \cap H_n] = \mathbb{P}[H_1] \cdots \mathbb{P}[H_n] \tag{2.14}$$

*for all $H_i$ in $\bar{\mathcal{C}}_i = \mathcal{C}_i \cup \{\Omega\}$, $1 \leqslant i \leqslant n$.*

PROOF. *Necessity* is obvious. For the *sufficiency* part, assume (2.14) for $H_i$ in $\bar{\mathcal{C}}_i$, $1 \leqslant i \leqslant n$. Fix $H_2, \cdots, H_n$ in $\bar{\mathcal{C}}_2, \cdots, \bar{\mathcal{C}}_n$ respectively and let $\mathcal{G} = \{H_1 \in \mathcal{F}_1 : \mathbb{P}[H_1 \cap \cdots \cap H_n] = \mathbb{P}[H_1] \cdots \mathbb{P}[H_n]\}$. By assumption, $\mathcal{G} \supset \mathcal{C}_1$ and $\Omega \in \mathcal{G}$, and the other two conditions for $\mathcal{G}$ to be a $\lambda$-system on $\Omega$ can be checked easily using the properties of probability. It follows from the $\pi$-$\lambda$ theorem that $\mathcal{G} \supset \sigma(\mathcal{C}_1) = \mathcal{F}_1$. Repeating the procedure successively with $H_2, \cdots, H_n$ we see that (2.14) holds for all $H_1, \cdots, H_n$ in $\mathcal{F}_1, \cdots, \mathcal{F}_n$ respectively. In other words, (2.13) holds when $V_i$ are indicators. This is extended to arbitrary positive random variables $V_i$ in $\mathcal{F}_i$ since $V_i$ is the limit of a monotone increasing sequence of simple random variables, and we can get the desired result by applying the monotone convergence theorem repeatedly. □

**5.3. Independence of Collections.** The next proposition shows that independence survives groupings.

THEOREM 2.32 (Independence of collections). *Let $(\mathcal{F}_t)_{t\in T}$ be an independency, $(T_i)_{i\in I}$ be a partition of $T$. Then, the $\sigma$-algebras generated by the subcollections $\mathcal{F}_{T_i} = \bigvee_{t\in T_i} \mathcal{F}_t$, $i \in \mathbb{N}$, form a new independency.*

PROOF. Let $\mathcal{G}_i$ be the $\pi$-system of all sets $A$ having the form

$$A = \bigcap_{t\in J_i} A_t$$

for some *finite* subset $J_i$ of $T_i$ and sets $A_t$ in $\mathcal{F}_t$, $t \in J_i$. We know that $\mathcal{G}_i$ contains all $\mathcal{F}_t$ ($t \in J_i$) and therefore $\bigcup_{t\in J_i} \mathcal{F}_t$; and $\mathcal{G}_i$ generates the $\sigma$-algebra $\mathcal{F}_{T_i}$. It is obvious that $\Omega \in \mathcal{G}_i$.

Now, in order to show that $\mathcal{F}_{T_i}$, $i \in \mathbb{N}$, are independent, we only need to consider every *finite* subset of $\mathcal{F}_{T_i}$, $i \in \mathbb{N}$. And the independence follows by the above characterization and Theorem 2.31 (Notice that we always deal with finite intersections in this procedure). □

**5.4. Pairwise Independence.** A collection of objects ($\sigma$-algebras, random variables) are said to be **pairwise independent** if every pair of them is an independency. It should be noticed that *Pairwise independence* is much weaker than *mutual independence*.

Indeed, independence can be checked by repeated checks for pairwise independence. We state this for a sequence of $\sigma$-algebras; it holds for a finite sequence as well, and therefore can be used to check the independency for arbitrary collections.

THEOREM 2.33 (Pairwise independence). *The sub-$\sigma$-algebras $\mathcal{F}_1, \mathcal{F}_2, \cdots$ of $\mathcal{F}$ are independent if and only if $\bigvee_{i=1}^n \mathcal{F}_i$ and $\mathcal{F}_{n+1}$ are independent for all $n \geqslant 1$.*

PROOF. *Necessity* is obvious from the last theorem. For *sufficiency*, assume that $\mathcal{G}_n = \bigvee_{i=1}^n \mathcal{F}_i$ and $\mathcal{F}_{n+1}$ are independent for all $n$. Then, for $H_1, \cdots, H_m$ in $\mathcal{F}_1, \cdots, \mathcal{F}_m$ respectively, we see (2.14) holds by repeated applications of the independence of $\mathcal{G}_n$ and $\mathcal{F}_{n+1}$ for $n = m-1, m-2, \cdots, 1$ in that order. Thus, $\mathcal{F}_1, \cdots, \mathcal{F}_m$ are independent by Theorem 2.31, and this is true for all $m \geqslant 2$. □

It should be noticed that Theorem 2.32 only shows that mutual independence can be obtained by verifying in turn that $\bigvee_{i=1}^n \mathcal{F}_i$ and $\mathcal{F}_{n+1}$ are pairwise independent. In general, we cannot derive mutual independence from pairwise independence.

EXAMPLE 2.34 (Pairwise independence cannot derive mutual independence). Let $X_1, X_2, X_3$ be independent random variables with $\mathbb{P}(X_k = 0) = \mathbb{P}(X_k = 1) = 1/2$. Let $A_1 = \{X_2 = X_3\}$, $A_2 = \{X_3 = X_1\}$, and $A_3 = \{X_1 = X_2\}$. These events are *pairwise independent* since if $i \neq j$, then

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(X_1 = X_2 = X_3) = 1/4 = \mathbb{P}(A_i)\mathbb{P}(A_j),$$

but they are not *mutually independent* since

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/8 = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3).$$

**5.5. Independence of Random Variables.** For each $t$ in some index set $T$, let $X_t$ be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. The variables $X_t$ are said to be *independent*, and the collection $(X_t)_{t\in T}$ is called an *independency*, if $(\sigma(X_t))_{t\in T}$ is an independency.

Since a collection is an independency if and only if its every finite subset is an independency, we concentrate on the independency of a finite number of them, which amounts to taking $T = \{1, 2, \cdots, n\}$ for some integer $n \geqslant 2$.

THEOREM 2.35 (Independence of random variables). *The random variables $X_1, \cdots, X_n$ are independent if and only if*

$$\mathbb{E}[f_1 \circ X_1 \cdots f_n \circ X_n] = \mathbb{E}[f_1 \circ X_1] \cdots \mathbb{E}[f_n \circ X_n] \tag{2.15}$$

*for all positive functions $f_1, \cdots, f_n$ in $\mathcal{E}_1, \cdots, \mathcal{E}_n$ respectively.*

PROOF. We need to show that (2.13) holds for all positive $V_1, \cdots, V_n$ in $\sigma(X_1), \cdots, \sigma(X_n)$ respectively if and only if (2.15) holds for all positive $f_1, \cdots, f_n$ in $\mathcal{E}_1, \cdots, \mathcal{E}_n$ respectively. But this is immediate form Theorem 2.23. □

Let $\pi$ be the joint distribution of $X_1, \cdots, X_n$, and let $\mu_1, \cdots, \mu_n$ be the corresponding marginals. Then, the left and the right side of (2.15) are equal to, respectively,

$$\int_{E_1 \times \cdots \times E_n} f_1(x_1) \cdots f_n(x_n)\, \pi(\mathrm{d}x_1, \cdots, \mathrm{d}x_n)$$

and

$$\int_{E_1} f_1(x_1)\,\mu_1(\mathrm{d}x_1)\cdots\int_{E_n} f_n(x_n)\,\mu_n(\mathrm{d}x_n).$$

The equality of these two expressions for all positive $f_1,\cdots,f_n$ is equivalent to saying that $\pi = \mu_1\times\cdots\times\mu_n$. That is,

THEOREM 2.36 (Independence and distribution). *The random variables $X_1,\cdots,X_n$ are independent if and only if their joint distribution is the product of their marginal distributions.*

Finally, a comment on functions of independent variables.

If $X$ and $Y$ are random variables taking values in $(E,\mathcal{E})$ and $(D,\mathcal{D})$, then we say that $X$ **determines** $Y$ if $Y = f\circ X$ for some $f: E\to D$ measurable with respect to $\mathcal{E}$ and $\mathcal{D}$. It is obvious that, if $X$ determines $Y$, then $\sigma(Y)\subset\sigma(X)$, this can be verified using Lemma 1.6.

THEOREM 2.37 (Functions of independent variables). *If the random variables $X_1,\cdots,X_n$ are independent, then*

$$f_1(X_1),\cdots,f_n(X_n)$$

*are independent variables provided that $f_1,\cdots,f_n$ are positive and in $\mathcal{E}_1,\cdots,\mathcal{E}_n$, respectively.*

PROOF. It is obvious that $Y_k := f_k(X_k)\in\sigma(Y_k)\subset\sigma(X_k)$ for all $1\leqslant k\leqslant n$ since $X_k$ determines $Y_k$. Then $Y_1,\cdots,Y_n$ are independent since $\sigma(X_1),\cdots,\sigma(X_k)$ are independent. □

The following example is frequently used later.

EXAMPLE 2.38 (Functions of independent variables). Let $X_1,\cdots,X_n$ be independent random variables, and $1\leqslant n_1<n_2<\cdots<n_k=n$ be a partition of $\{1,\cdots,n\}$. Suppose

$$g_j\in\bigotimes_{i=n_{j-1}+1}^{n_j}\sigma(X_i),$$

then

$$g_1(X_1,\cdots,X_{n_1}),\ g_2(X_{n_1+1},\cdots,X_{n_2}),\ \cdots,\ g_k(X_{n_{k-1}+1},\cdots,X_{n_k}) \tag{2.16}$$

are independent.

PROOF. Using Corollary 2.24, we have $g_j(X_{n_{j-1}+1},\cdots,X_{n_j})$ is $\sigma(X_{n_{j-1}+1},\cdots,X_{n_j})$-measurable for all $1\leqslant j\leqslant k$. Then, we know that the variables in (2.16) are independent by Theorem 2.32. □

ANOTHER PROOF. Another PROOF. Indeed, we can use Theorem 2.31 to prove the independence. We only need to show variables in (2.16) are independent on some $\pi$-system that generate the $\sigma$-algebras. Thus $\pi$-systems are very familiar with us, that is, the collection of triangles. □

**5.6. Sums of Independent Random Variables.** We consider finite sum of independent random variables in this chapter first. Let $X$ and $Y$ be $\mathbb{R}^d$-valued independent random variables with distributions $\mu$ and $\nu$ respectively. Then, the distribution of $(X,Y)$ is the product measure $\mu\times\nu$, and the distribution $\mu*\nu$ of $X+Y$ is given by

$$(\mu*\nu)(f) = \mathbb{E}[f(X+Y)] = \int\mu(\mathrm{d}x)\int f(x+y)\,\nu(\mathrm{d}y).$$

This distribution $\mu*\nu$ is called the **convolution** of $\mu$ and $\nu$.

What we will do next is to extend the infinite (countable) sum. Sums of random variables and the limiting behavior of such sums as the number of summands grows to infinity are of constant interest in probability theory.

We shall return to such matters repeatedly in the chapters to follow. For the present, we describe two basic results, zero-one laws due to Kolmogorov and Hewitt-Savage.

**5.7. Kolmogorov's Zero-one Law.** Let $(\mathcal{G}_n)$ be a sequence of sub-$\sigma$-algebras of $\mathcal{F}$. We may think of $\mathcal{G}_n$ as the information revealed by the $n$th trail of an experiment. Then,

$$\mathcal{T}_n = \bigvee_{m>n} \mathcal{G}_m$$

is the information about the future after $n$, and $\mathcal{T} = \bigcap_n \mathcal{T}_n$ is that about the remote future. $\mathcal{T}$ is called the **tail-$\sigma$-algebra**, it consists of events whose occurrences are unaffected by the happening in finite time.

EXAMPLE 2.39 (Tail events). Let $X_n$ be real valued random variables, put $\mathcal{G}_n = \sigma(X_n)$, and $S_n = X_1 + \cdots + X_n$. Then,

(i) The event $\{\lim_n S_n \text{ exists}\}$ belongs to $\mathcal{T}_n$ for all $n$, hence, belongs to the tail-$\sigma$-algebra $\mathcal{T}$.
(ii) Similarly, $\{\limsup_n S_n/n > b\}$ is unaffected by the first $n$ variables, and this is true for all $n$, hence this event belongs to $\mathcal{T}$.
(iii) But, $\{\limsup_n S_n > b\}$ is not in $\mathcal{T}$.
(iv) Let $B$ be a Borel set of $\mathbb{R}$. Let $\{X_n \in B \text{ i.o.}\}$ be the set of $\omega$ for which $\sum_n \mathbb{1}_B \circ X_n(\omega) = \infty$. This event belongs to $\mathcal{T}$.
(v) The event $\{S_n \in B \text{ i.o.}\}$ is not in $\mathcal{T}$.

PROOF. (i). It is obvious since

$$\Big\{\lim_n S_n \text{ exists}\Big\} = \Big\{\sum_n X_n \text{ converges}\Big\} = \bigcap_n \Big\{\sum_{m>n} X_m \text{ converges}\Big\} \in \bigcap_n \Big(\bigvee_{m>n} \mathcal{G}_n\Big).$$

(ii). For all $m \geqslant 1$, we have

$$\limsup_n \frac{S_n}{n} = \lim_n \frac{S_m}{n} + \limsup_n \frac{S_n - S_m}{n} = \limsup_n \frac{S_n - S_m}{n}.$$

Hnece,

$$\Big\{\limsup_n \frac{S_n}{n} > b\Big\} = \bigcap_m \Big\{\limsup_n \frac{X_{m+1} + \cdots + X_n}{n} > b\Big\} \in \bigcap_m \Big(\bigvee_{k>m} \mathcal{G}_k\Big).$$

(iii). $\{\limsup_n S_n > b\} \notin \mathcal{T}$, since its occurrence relies on the values of all the $X_n$'s and cannot be determined by the information encoded in $\{X_{n+1}, X_{n+2}, \cdots\}$ for arbitrary large $n$.

(iv). If $\sum_n \mathbb{1}_B \circ X_n(\omega) = \infty$, that is, for all integer $n$, there is a $m > n$ such that $\omega \in \{X_m \in B\}$. Then,

$$\{X_n \in B \text{ i.o.}\} = \bigcap_{n\geqslant 1} \bigcup_{m>n} \{X_m \in B\} \in \bigcap_n \Big(\bigvee_{m>n} \mathcal{G}_n\Big).$$

(v) is similar to (iii). □

The following theorem, called Kolmogorov's zero-one law, implies in particular that, if the $X_n$ of the preceding example are independent, then each one of the events in $\mathcal{T}$ has probability equal to either 0 or 1.

THEOREM 2.40 (Kolmogorov's zero-one law). *Let $\mathcal{G}_n$, $n$ in $\mathbb{N}$, be independent. Then, $\mathbb{P}[H]$ is either 0 or 1 for every event $H$ in the tail $\mathcal{T}$.*

PROOF. By Theorem 2.32 on partitions of independencies, $\{\mathcal{G}_1, \cdots, \mathcal{G}_n, \mathcal{T}_n\}$ is an independency for every $n$, which implies that so is $\{\mathcal{G}_1, \cdots, \mathcal{G}_n, \mathcal{T}\}$ since $\mathcal{T} \subset \mathcal{T}_n$. Thus, $\{\mathcal{T}, \mathcal{G}_1, \mathcal{G}_2, \cdots\}$ is an independency since every finite subset of it is an independency. And so is $\{\mathcal{T}, \mathcal{T}_0\}$ by Theorem 2.32 again. In other words, for $H$ in $\mathcal{T}$ and $G$ in $\mathcal{T}_0$, we have $\mathbb{P}[H \cap G] = \mathbb{P}[H]\mathbb{P}[G]$, and this holds for $G = H$ as well because $\mathcal{T} \subset \mathcal{T}_0$. Thus, for $H$ in $\mathcal{T}$, we have $\mathbb{P}[H] = \mathbb{P}[H]\mathbb{P}[H]$, which means that $\mathbb{P}[H]$ is either 0 or 1. □

COROLLARY 2.41 (Random variables on tail-$\sigma$-algebra). *Let $\mathcal{G}_n$, $n$ in $\mathbb{N}$, be independent. Then, for every random variable $V$ in the tail-$\sigma$-algebra there is a constant $c$ in $\bar{\mathbb{R}}$ such that $V = c$ almost surely.*

**5.8. * Hewitt-Savage Zero-one Law.** Back to Example 2.39, for instance, $\limsup S_n/n$ is almost surely constant by Corollary 2.41. But can we derive some similar properties for the events $\{\limsup S_n > b\}$ or $\{S_n \in B \text{ i.o.}\}$, which are not in $\mathcal{T}$?

The next theorem will show that, if we add to the independence of $X_n$ the extra condition that they have the same distribution, these two events have probability 0 or 1.

Let $X = (X_1, X_2, \cdots)$, where the $X_n$ takes values in some measurable space $(E, \mathcal{E})$. Let $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \cdots)$ be the *filtration* generated by $X$, that is,

$$\mathcal{F}_n = \sigma(X_1, \cdots, X_n), \quad \mathcal{F}_\infty = \bigvee_n \mathcal{F}_n$$

for all $n$ in $\mathbb{N}$. Recall from Theorem 2.23 and its sequel that $\mathcal{F}_\infty$ consists of random variables of the form $V = f \circ X$ with $f$ in $\mathcal{E}^\infty$; and $\mathcal{F}_n$ consists of the random variables of the form

$$V_n = f_n(X_1, \cdots, X_n) = \hat{f}_n \circ X$$

with $f_n$ in $\mathcal{E}^n$ and appropriately defined $\hat{f}_n$.

A **finite permutation** $p$ is meant a bijection $p : \mathbb{N} \to \mathbb{N}$ such that $p(n) = n$ for all but finitely many $n$. For such a permutation $p$, we write

$$X \circ p = (X_{p(1)}, X_{p(2)}, \cdots), \tag{2.17}$$

which is a re-arrangement of the entires of $X$. The notation extends to arbitrary random variables $V$ in $\mathcal{F}_\infty$: if $V = f \circ X$ then

$$V \circ p = f \circ (X \circ p).$$

If the $X_n$'s are *independent and identically distributed*, the probability laws of $X$ and $X \circ p$ are the same, and hence, the distributions of $V$ and $V \circ p$ are the same.

A random variable $V$ in $\mathcal{F}_\infty$ is said to be **permutation invariant** if $V \circ p = V$ for every finite permutation $p$. An event in $\mathcal{F}_\infty$ is said to be *permutation invariant* if its indicator is such. Indeed, variables like $V = \limsup S_n$ or events like $\{S_n \in B \text{ i.o.}\}$ in Example 2.39 are permutation invariant: they are unaffected by the re-arrangements of the entries of $X$ by finite permutation.

The collection of all permutation invariant events is a $\sigma$-algebra which contains the tail-$\sigma$-algebra of $X$[6]. The following, called Hewitt-Savage zero-one law[7], shows that it is almost surely trivial provided that the $X_n$ are identically distribution in addition to being independent.

[6]Since we only re-arrange *finite* entries of $X_n$'s, so we can find the maximum entry of this re-arrangement. But tail-$\sigma$-algebra consists of events whose occurrences are *unaffected* by the happening in *finite* time, so the re-arrangement cannot affect the occurrences of tail events.

[7]We will give another proof after we talked about the convergence of backward martingales.

THEOREM 2.42 (Hewitt-Savage zero-one law). *Suppose that $X_1, X_2, \cdots$ are independent and identically distributed. Then, every permutation invariant event has probability $0$ or $1$. Also, for every permutation invariant random variable $V$ there is a constant $c$ in $\bar{\mathbb{R}}$ such that $V = c$ almost surely.*

PROOF. (i). In order to show that every permutation event $H$ in $\mathcal{F}_\infty$ has probability $0$ or $1$, we need to show that the permutation invariant variable $\mathbb{1}_H$ is a constant almost surely. Since $\mathbb{1}_H$ takes values in $[0,1]$, it is sufficient to show that if $V : \Omega \to [0,1]$ is a permutation invariant variable in $\mathcal{F}_\infty$, then $\mathbb{E}[V^2] = (\mathbb{E}[V])^2$, i.e., $V$ equals to a constant almost surely.

Suppose $V : \Omega \to [0,1]$ is a permutation invariant variable in $\mathcal{F}_\infty$, we will show $\mathbb{E}[V^2] = (\mathbb{E}[V])^2$. By Theorem 2.28, there are $V_n$ in $\mathcal{F}_n$, $n \geqslant 1$, such that each $V_n$ takes values in $[0,1]$ and

$$\lim \mathbb{E}[|V - V_n|] = 0 \quad \Rightarrow \quad \lim \mathbb{E}[V_n] = \mathbb{E}[V]. \tag{2.18}$$

Fix $n$ in $\mathbb{N}$. Let $p$ be a finite permutation. The assumption about $X$ implies that $X$ and $X \circ p$ have the same probability law, which implies that $U$ and $U \circ p$ have the same distribution for every $U$ in $\mathcal{F}_\infty$. Taking $U = V - V_n$, noting that $U \circ p = V \circ p - V_n \circ p = V - V_n \circ p$ by the invariance of $V$, we see that

$$\mathbb{E}[|V - V_n \circ p|] = \mathbb{E}[|V - V_n|]. \tag{2.19}$$

In particular, for the permutation $\hat{p}$ that maps $1, \cdots, n$ to $n+1, \cdots, 2n$ and vice-versa, leaving $\hat{p}(m) = m$ for $m > 2n$. We define $\hat{V}_n = V_n \circ \hat{p}$ and observe that, if $V_n = f_n(X_1, \cdots, X_n)$, then $\hat{V}_n = f_n(X_{n+1}, \cdots, X_{2n})$, which implies that $V_n$ and $\hat{V}_n$ are independent and have the same distribution. Together with (2.19), this yields

$$\mathbb{E}[V_n \hat{V}_n] = (\mathbb{E}[V_n])^2, \quad \mathbb{E}[|V - \hat{V}_n|] = \mathbb{E}[|V - V_n|]. \tag{2.20}$$

Then we have

$$|\mathbb{E}[V^2] - (\mathbb{E}[V])^2| = |\mathbb{E}[V^2 - V_n \hat{V}_n]| \leqslant \mathbb{E}[|V^2 - V_n \hat{V}_n|] \leqslant 2\mathbb{E}[|V - V_n|], \tag{2.21}$$

where the final step used (recalling $|V| \leqslant 1$ and $|V_n| \leqslant 1$)

$$|V^2 - V_n \hat{V}_n| = |(V - V_n)V + (V - \hat{V}_n)V_n| \leqslant |V - V_n| + |V - \hat{V}_n|,$$

and (2.20). Applying (2.18) to (2.21) yields the desired result that $\mathbb{E}[V^2] = (\mathbb{E}[V])^2$.

(ii). For every permutation invariant random variable $V$, we want to prove that $V$ is a constant almost surely, so it is sufficient to prove that $\{V \in B\}$ has probability $0$ or $1$ for all Borel set $B$[8], so we only need to prove $\mathbb{1}_{\{V \in B\}}$ is a permutation invariant variable using (a). Using the fact that $\mathbb{1}_{\{V \in B\}} = \mathbb{1}_B \circ V$, we know that

$$\mathbb{1}_{\{V \in B\}} \circ p = (\mathbb{1}_B \circ V) \circ p = \mathbb{1}_B \circ (V \circ p) = \mathbb{1}_B \circ V = \mathbb{1}_{\{V \in B\}},$$

where the third equality is justified by the condition that $V$ is permutation invariant. So we have gotten the result we desired. □

EXAMPLE 2.43 (Random walks). Returning to Example 2.39, assume further that $X_n$'s have the same distribution. Then, the stochastic process $(S_n)$ is called a **random walk** on $\mathbb{R}$. The avoid the trivial case where $S_n \equiv 0$ almost surely, we assume that $\mathbb{P}[X_1 = 0] < 1$. Then, concerning the limiting behavior of the random walk, there are three possibilities, exactly one of which is almost sure:

[8] If we have proved this result, there must be a $c$ in $\bar{\mathbb{R}}$ such that $\mathbb{P}[V = c] = 1$, which means that $V$ is constant almost surely.

(i) $\lim S_n = +\infty$,
(ii) $\lim S_n = -\infty$,
(iii) $\liminf S_n = -\infty$ and $\limsup S_n = +\infty$.

Indeed, by the preceding theorem, there is a constant $c$ in $\bar{\mathbb{R}}$ such that $\limsup S_n = c$ almost surely. Letting $\hat{S}_n = S_{n+1} - X_1$ yields another random walk $(\hat{S}_n)$ which has the same law as $(S_n)$. Thus, $\limsup \hat{S}_n = c$ almost surely, which means that $c = c - X_1$. Since we excluded the trivial case when $\mathbb{P}[X_1 = 0] = 1$, it follows that $c$ is either $+\infty$ or $-\infty$. Similarly, $\liminf S_n$ is either almost surely $+\infty$ or almost surely $-\infty$. Of the four combinations, discarding the impossible case when $\liminf S_n = +\infty$ and $\limsup S_n = -\infty$, we arrive at the result.

If the common distribution of the $X_n$ is *symmetric*, that is, if $X_1$ and $-X_1$ have the same distribution (like the Gaussian with mean $0$), then $(S_n)$ and $(-S_n)$ have the same law, and it follows that the cases (i) and (ii) are improbable. So then, case (iii) holds almost surely.

CHAPTER 3

# Convergence

This is the core chapter of the probability section of HKUST(GZ) PhD Qualifying Exam in Financial Technology, and we have provided ample examples and exercises.

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

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

and we will study the relation between them.

This will be the most important chapter in probability theory. And I highly recommend a very elegant book: Chung, Kai Lai. *A course in probability theory*. Elsevier, 2000. Besides, we will also add more details from other books.

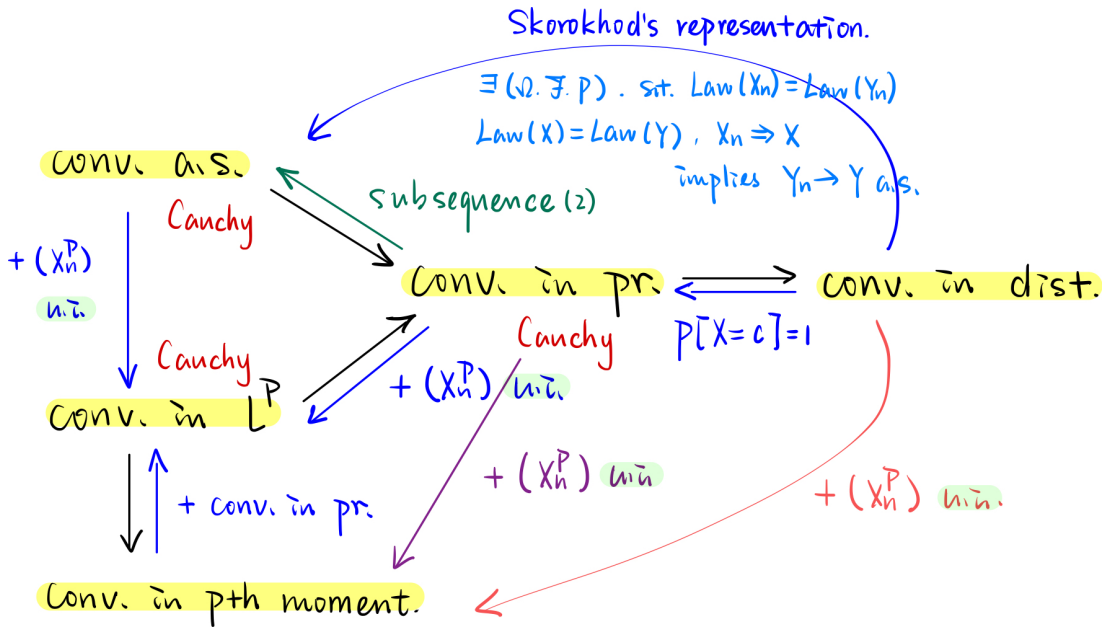


FIGURE 1. Almost all conclusions regarding various types of convergence relationships.

## 1. Various Modes of Convergence

**1.1. Almost Sure Convergence.** The first concept we will discuss is almost sure convergence:

DEFINITION 3.1 (Almost sure convergence). The sequence of random variables $(X_n)$ is said to converge **almost surely** (or **with probability one**) to the random variable $X$ if there exists a *negligible* set $\mathcal{N}$ such that

$$\forall\, \omega \in \Omega\backslash\mathcal{N}: \quad \lim_n X_n(\omega) = X(\omega). \tag{3.1}$$

We allow each random variable a negligible set on which it may be $\pm\infty$. The union of all these sets being still a negligible set, it can be included in $\mathcal{N}$ in (3.1) without modifying the conclusion. When dealing with a *countable* set of random variables that are finite almost surely, to regard them as *finite everywhere*.

THEOREM 3.2 (Equivalent condition of almost sure convergence). *The sequence $(X_n)$ converges almost surely to $X$ if and only if, for all $\varepsilon > 0$, we have*

$$\lim_m \mathbb{P}[|X_n - X| \leqslant \varepsilon \text{ for all } n \geqslant m] = 1. \tag{3.2}$$

REMARK 3.3. The event in (3.2) is same as

$$A_m(\varepsilon) := \bigcap_{n\geqslant m} \{|X_n - X| \leqslant \varepsilon\}.$$

Besides, it is obvious that $A_m(\varepsilon)$ is increasing in $m$, then the (3.2) is actually

$$\mathbb{P}\Big[\bigcup_{m\geqslant 1}\bigcap_{n\geqslant m} \{|X_n - X| \leqslant \varepsilon\}\Big] = 1.$$

Take complement of $\bigcup_{m\geqslant1}\bigcap_{n\geqslant m}\{|X_n - X| \leqslant \varepsilon\}$, we have

$$\mathbb{P}\Big[\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m} \{|X_n - X| > \varepsilon\}\Big] = 0.$$

That is another equivalent criterion of almost sure convergence. †

PROOF OF THEOREM 3.2. First consider the *necessity*. Suppose $(X_n)$ converges to $X$ on $\Omega_0$ with $\mathbb{P}[\Omega_0] = 1$. Now for all $\varepsilon > 0$ and all $\omega_0 \in \Omega_0$, we have $X_n(\omega_0) \to X(\omega_0)$ as $n \to \infty$, that is, for such $\varepsilon > 0$, there is an integer $m(\omega_0, \varepsilon)$ such that $|X_n(\omega_0) - X(\omega_0)| \leqslant \varepsilon$ for all $n \geqslant m(\omega_0, \varepsilon)$. Thus we have

$$\omega \in A_{m(\omega_0,\varepsilon)}(\varepsilon),$$

which implies that $\Omega_0 \subset \bigcup_{m\geqslant1} A_m(\varepsilon)$. Using the monotonicity of probability measure,

$$1 = \mathbb{P}[\Omega_0] \leqslant \mathbb{P}\Big[\bigcup_{m\geqslant1} A_m(\varepsilon)\Big] = \lim_m \mathbb{P}(A_m(\varepsilon)) \leqslant 1.$$

Next, consider the *sufficiency*. Define $A_m(\varepsilon)$ in the same way as before. Set

$$A = \bigcap_{k\geqslant1}\bigcup_{m\geqslant1} A_m(2^{-k}).$$

By the hypothesis, we have $\lim_m \mathbb{P}[A_m(2^{-k})] = 1$. Since the events $A_m(2^{-k})$ is increasing in $m$, we have $\mathbb{P}[\bigcup_{m\geqslant1} A_m(2^{-k})] = 1$, which implies that $\mathbb{P}[A] = 1$. Now, if $\omega \in A$, then for all $k \geqslant 1$, there is $m \geqslant 1$ such that $|X_n - X| \leqslant 2^{-k}$ for all $n \geqslant m$, that is, $X_n \to X$ on $A$. This completes the proof. □

**1.2. Convergence in Probability.** A *weaker* concept of convergence is of basic importance in probability theory.

DEFINITION 3.4 (Convergence in probability). The sequence $(X_n)$ is said to **converge in probability** to $X$ if, for all $\varepsilon > 0$, we have

$$\lim_n \mathbb{P}[|X_n - X| \geqslant \varepsilon] = 0. \tag{3.3}$$

As a consequence of Theorem 3.2, we see that

LEMMA 3.5 (Almost sure convergence implies convergence in probability). *Almost sure convergence implies convergence in probability.*

PROOF. Suppose $(X_n)$ converges to $X$ almost surely, then for all $\varepsilon > 0$, we have

$$\mathbb{P}\Big[\bigcap_{n\geqslant 1}\bigcup_{m\geqslant n}\{|X_m - X| > \varepsilon\}\Big] = \lim_n \mathbb{P}\Big[\bigcup_{m\geqslant n}\{|X_m - X| > \varepsilon\}\Big] = 0.$$

Since $\{|X_n - X| > \varepsilon\} \subset \bigcup_{m\geqslant n}\{|X_m - X| > \varepsilon\}$ holds for all $n$ in $\mathbb{N}$, let $n \to \infty$ then we get the desired result. □

The converse of Lemma 3.5 may not hold:

EXAMPLE 3.6 (Convergence in probability may not imply almost sure convergence). Let $\mathbb{P}[X_n = 0] = 1 - 1/n$, $\mathbb{P}[X_n = 1] = 1/n$, and $X_n$'s are independent. Then $X_n \to 0$ in probability since $\mathbb{P}(|X_n - 0| > \varepsilon) \leqslant 1/n \to 0$. But, $X_n$ does NOT converge to 0 almost surely, since for any $0 < \varepsilon < 1$ we have

$$\begin{aligned}\mathbb{P}\Big[\bigcap_{m\geqslant n}\{|X_m - 0| \leqslant \varepsilon\}\Big] &= \mathbb{P}\Big[\lim_{r\to\infty}\bigcap_{m=n}^{r}\{|X_m| \leqslant \varepsilon\}\Big] = \lim_{r\to\infty}\mathbb{P}\Big[\bigcap_{m=n}^{r}\{|X_m| \leqslant \varepsilon\}\Big] \\ &= \lim_{r\to\infty}\prod_{m=n}^{r}\mathbb{P}[|X_m| \leqslant \varepsilon] = \lim_{r\to\infty}\prod_{m=n}^{r}\Big(1 - \frac{1}{m}\Big) \\ &= \lim_{r\to\infty}\frac{n-1}{n}\frac{n}{n+1}\cdots\frac{r-1}{r} = \lim_{r\to\infty}\frac{n-1}{r} = 0.\end{aligned}$$

By the equivalent definition of almost sure convergence, we see that $X_n$ does NOT converge to 0 almost surely. †

Under what conditions does convergence in probability imply almost sure convergence? We will discuss this question in the next section.

Using the definition of convergence in probability, we can prove the simplest case of weak law of large numbers.

EXAMPLE 3.7 ($L^2$-weak law of large numbers). Suppose $(X_n)$ are independent and identically distributed with $\mathbb{E}[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2 < \infty$. Set $S_n = \sum_{j=1}^n X_j$, then

$$\frac{S_n}{n} \to \mu \quad \text{in probability.}$$

PROOF. Using Chebyshev's inequality, it is obvious that

$$\begin{aligned}\mathbb{P}\Big[\Big|\frac{S_n}{n} - \mu\Big| \geqslant \varepsilon\Big] &\leqslant \frac{\mathbb{E}[|S_n/n - \mu|^2]}{\varepsilon^2} = \frac{\mathrm{Var}[S_n/n]}{\varepsilon^2} = \frac{\mathrm{Var}[S_n]}{n^2\varepsilon^2} \\ &= \frac{1}{n^2\varepsilon^2}\sum_{j=1}^{n}\mathrm{Var}[X_j] = \frac{\sigma^2}{n\varepsilon^2} \to 0\end{aligned}$$

for all $\varepsilon > 0$ since $\sigma^2$ is finite. □

**1.3. Convergence in $L^p$.** Next, we will consider:

DEFINITION 3.8 (Convergence in $L^p$). Assume $0 < p < \infty$. The sequence $(X_n)$ is said to **converge in $L^p$** to $X$ if, $X_n$'s and $X$ in $L^p$, and

$$\lim_n \mathbb{E}[|X_n - X|^p] = 0. \tag{3.4}$$

In all these definitions above, $X_n$ converges to $X$ if and only if $X_n - X$ converges to $0$.

LEMMA 3.9 (Convergence in $L^p$ implies convergence in probability). *Assume $0 < p < \infty$. If $X_n \to X$ in $L^p$, then $X_n \to X$ in probability.*

PROOF. It is obvious that

$$\mathbb{P}[|X_n - X| \geqslant \varepsilon] \leqslant \frac{1}{\varepsilon^p}\mathbb{E}[|X_n - X|^p]$$

by Chebyshev's inequality. □

In general, the converse of Lemma 3.9 is wrong. Besides, there are no general relation between convergence in $L^p$ and almost sure convergence.

EXAMPLE 3.10 (A lot of counterexamples). We have the following relations:

- *Convergence in probability cannot imply convergence in $L^p$.* Let $\mathbb{P}[X_n = 0] = 1 - 1/n$, $\mathbb{P}[X_n = n] = 1/n$, then $X_n \to 0$ in probability since $\mathbb{P}[|X_n - 0| \geqslant \varepsilon] \leqslant 1/n \to 0$. However, $\mathbb{E}[|X_n|] = 1$ does not converge to $0$.
- *Almost sure convergence cannot imply convergence in $L^p$.* Let $\mathbb{P}[X_0 = 0] = 1 - n^{-2}$ and $\mathbb{P}[X_n = n^3] = n^{-2}$, then $X_n \to 0$ almost surely since

  $$\sum_n \mathbb{P}[|X_n - 0| \geqslant \varepsilon] = \sum_n n^{-2} < \infty.$$

  However, $X_n$ not converges to $0$ in $L^1$ as $\mathbb{E}[|X_n - 0|] = n \to \infty$.
- *Convergence in $L^p$ cannot imply almost sure convergence.* Let $\mathbb{P}[X_n = 0] = 1 - n^{-1}$ and $\mathbb{P}[X_n = 1] = n^{-1}$, and they are independent. Then $X_n \to 0$ in $L^1$ since

  $$\mathbb{E}[|X_n - 0|] = 1/n \to 0.$$

  However, it was shown in Example 3.6 that $X_n$ not converges to $0$ almost surely. †

The following theorem provides a *sufficient condition* for obtaining convergence in $L^p$ from convergence in probability.

THEOREM 3.11 (Dominated condition). *If $(X_n)$ converges to $0$ in probability and $(X_n)$ is dominated by $Y$ in $L^p$, then $(X_n)$ converges to $0$ in $L^p$.*

PROOF. Suppose $|X_n| \leqslant Y$ almost surely with $\mathbb{E}[|Y|^p] < \infty$. Then

$$\begin{aligned}\mathbb{E}[|X_n|^p] &= \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n| < \varepsilon\}}] + \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n| \geqslant \varepsilon\}}] \\ &\leqslant \varepsilon^p + \mathbb{E}[|Y|^p \mathbb{1}_{\{|X_n| \geqslant \varepsilon\}}] \to 0\end{aligned}$$

is justified by the arbitrariness of $\varepsilon$; the *absolute continuity* of the integral, and $\mathbb{P}[|X_n| \geqslant \varepsilon] \to 0$ as $n \to \infty$. □

REMARK 3.12. If $X_n$ converges to $X$ in $L^p$, and $(X_n)$ is dominated by $Y$, then $(X_n - X)$ is dominated by $Y + |X|$, which is in $L^p$. Hence there is *no loss of generality* to assume $X \equiv 0$. †

The following example gives a equivalent condition of convergence in probability with the help of Theorem 3.11.

EXAMPLE 3.13 (Equivalent condition of convergence in probability). $X_n$ converges to $0$ in probability if and only if

$$\mathbb{E}\Big[\frac{|X_n|}{1+|X_n|}\Big] \to 0.$$

PROOF. For any random variable $X$, we have $|X|/(1+|X|) \leqslant 1$, where $1$ in $L^p$, then

$$\frac{|X_n|}{1+|X_n|} \to 0 \text{ in probability} \quad \Longleftrightarrow \quad \mathbb{E}\Big[\frac{|X_n|}{1+|X_n|}\Big] \to 0.$$

So it is sufficient for us to show that $X_n \to 0$ in probability is equivalent to $|X_n|/(1+|X_n|) \to 0$ in probability. This is obvious since $|x| \leqslant \varepsilon$ is equivalent to $|x|/(1+|x|) \leqslant \varepsilon$. So

$$\mathbb{P}[|X_n| \leqslant \varepsilon] = \mathbb{P}\Big[\frac{|X_n|}{1+|X_n|} \leqslant \frac{\varepsilon}{1+\varepsilon}\Big],$$

where the left side goes to $1$. □

## 2. Borel-Cantelli Lemma

First let us recall that, let $(E_n)$ be any sequence of subsets of $\Omega$,

$$\limsup_n E_n = \bigcap_{m\geqslant 1}\bigcup_{n\geqslant m} E_n, \quad \liminf_n = \bigcup_{m\geqslant 1}\bigcap_{n\geqslant m} E_n, \quad \liminf_n E_n = \big(\limsup_n E_n^c\big)^c.$$

LEMMA 3.14 (Equivalent conditions of limits of a sequence of sets).

(i) $\omega \in \limsup_n E_n$ *if and only if, for all* $n \in \mathbb{N}$, *there exists a* $k_0 \geqslant n$ *such that* $\omega \in E_{k_0}$, *i.e., there are infinite terms in* $(E_n)$ *that contain* $\omega$.

(ii) $\omega \in \liminf_n E_n$ *if and only if, there exists a* $n_0 \in \mathbb{N}$ *such that for all* $k \geqslant n_0$, *we have* $\omega \in E_k$, *i.e., there are only finite terms in* $(E_n)$ *that do not contain* $\omega$.

In more intuitive language: *the event* $\limsup_n E_n$ *occurs if and only if the events* $E_n$ *occur infinitely often*. Thus we may write

$$\mathbb{P}[\limsup_n E_n] = \mathbb{P}[E_n \text{ i.o.}]$$

where the abbreviation "i.o." stands for "infinitely often".

**2.1. Convergence Part.** The Borel-Cantelli lemma is very simple but still is the basic tool for proving almost sure convergence.

THEOREM 3.15 (Borel-Cantelli lemma - convergence part). *For arbitrary events* $(E_n)$, *we have*

$$\sum_n \mathbb{P}[E_n] < \infty \quad \Rightarrow \quad \mathbb{P}[E_n \text{ i.o.}] = 0. \tag{3.5}$$

PROOF. It is obvious that

$$\mathbb{P}[E_n \text{ i.o.}] = \mathbb{P}\Big[\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m} E_n\Big] = \lim_m \mathbb{P}\Big[\bigcup_{n\geqslant m} E_n\Big] \leqslant \limsup_m \sum_{n\geqslant m} \mathbb{P}[E_n] = 0,$$

where the second equality is justified by continuity of probability measure; the first inequality by Boole's inequality of probability measure; and the last equality since $\sum_n \mathbb{P}[E_n] < \infty$ implies that $\sum_{n \geqslant m} \mathbb{P}[E_n] \to 0$ as $m \to \infty$. □

The following basic theorem characterizes convergence in probability in terms of almost sure convergence, and more.

THEOREM 3.16 (Subsequence principle - part 1). *If $(X_n)$ converges to $X$ in probability, then it has a subsequence converges to $X$ almost surely.*

PROOF. Suppose $X_n \to X$ in probability. Then we have, for any $\varepsilon > 0$, $\lim_n \mathbb{P}[|X_n - X| > \varepsilon] = 0$. That is, for each $k$ in $\mathbb{N}$, there exists an increasing sequence $(n_k)$ such that

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

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

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

By the convergence part of Borel-Cantelli lemma, we have $\mathbb{P}[|X_{n_k} - X| > 2^{-k} \text{ i.o.}] = 0$. That is, there exists a $\Omega_0$ with $\mathbb{P}[\Omega_0] = 1$ such that the follows holds. For each $\omega \in \Omega_0$, there exists a $K(\omega)$ such that $|X_{n_k} - X| \leqslant 2^{-k}$ for all $k \geqslant K(\omega)$. Then it is immediate that $X_{n_k}(\omega) \to X(\omega)$ as $k \to \infty$. Hence, $X_{n_k} \to X$ almost surely. □

As an easy corollary of the Borel-Cantelli lemma, we prove a version of the strong law of large numbers (SLLN) with a finite 4th moment assumption.

EXAMPLE 3.17 ($L^4$-strong law of large numbers). Suppose $(X_n)$ are independent and identically distributed with $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}[X_n^4] < \infty$. Set $S_n = \sum_{j=1}^n X_j$, then

$$\frac{S_n}{n} \to \mu \quad \text{almost surely.}$$

PROOF. We may assume $\mu = 0$ (otherwise, it is sufficient to consider $X_n - \mu$). First,

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

Using Chebyshev's inequality we get $\mathbb{P}[|S_n|/n > \varepsilon] \leqslant C/n^2\varepsilon^4$. Summing over $n$ is finite. Thus the convergence part of Borel-Cantelli lemma implies $\mathbb{P}[|S_n|/n > \varepsilon \text{ i.o.}] = 0$. Thus,

$$\mathbb{P}\Big[\bigcup_{k \geqslant 1} \bigcap_{m \geqslant 1} \bigcup_{n \geqslant m} \{|S_n|/n > 2^{-k}\}\Big] = 0,$$

which implies that $S_n/n \to 0$ almost surely. □

**2.2. Divergence Part.** Under the assumption of independence, Lemma 3.15 has a striking complement.

LEMMA 3.18 (Borel-Cantelli lemma - divergence part). *If the events $(E_n)$ are independent, then*

$$\sum_n \mathbb{P}[E_n] = \infty \quad \Rightarrow \quad \mathbb{P}[E_n \text{ i.o.}] = 1. \tag{3.6}$$

PROOF. It is sufficient to show that $\mathbb{P}[\liminf_n E_n^c] = \lim_m \mathbb{P}[\bigcap_{n\geqslant m} E_n^c] = 0$. Since $E_n^c$'s are independent, we have, for $m \leqslant k$,

$$\mathbb{P}\Big[\bigcap_{n=m}^{k} E_n^c\Big] = \prod_{n=m}^{k} \mathbb{P}[E_n^c] = \prod_{n=m}^{k} (1-\mathbb{P}[E_n]) \leqslant \prod_{n=m}^{k} \exp(-\mathbb{P}[E_n]) = \exp\Big(-\sum_{n=m}^{k} \mathbb{P}[E_n]\Big).$$

Let $k\to\infty$, the right side goes to zero, since $\sum_n \mathbb{P}[E_n]=\infty$. Besides, $\bigcap_{n=m}^k E_n^c$ is decreasing in $k$, then

$$\mathbb{P}\Big[\bigcap_{n\geqslant m} E_n^c\Big] = \lim_k \mathbb{P}\Big[\bigcap_{n=m}^{k} E_n^c\Big] = 0.$$

Thus we get $\lim_m \mathbb{P}[\bigcap_{n\geqslant m} E_n^c]=0$, which is the desired result. □

LEMMA 3.19 (Pairwise independence case of divergence part of Borel-Cantelli lemma). *The implication (3.6) remains true if the events $(E_n)$ are pairwise independent.*

PROOF. It is convenient to introduce indicator random variables. Let $I_n = \mathbb{1}_{E_n}$ for all $n\geqslant 1$. Then $\mathbb{E}[I_n]=\mathbb{P}[E_n]$, $\mathrm{Var}[I_n]=\mathbb{P}[A_n](1-\mathbb{P}[A_n])$, and the pairwise independence translates into $\mathbb{E}[I_iI_j]=\mathbb{E}[I_i]\mathbb{E}[I_j]$ for all $i\neq j$. Define $S_n=\sum_{j=1}^n I_j$, then, the hypothesis is equivalent to $\mathbb{E}[S_n]\to\infty$ as $n\to\infty$ and the conclusion is equivalent to $\mathbb{P}[\lim_n S_n=\infty]=1$. By Chebyshev's inequality, we have

$$\begin{aligned}\mathbb{P}[|S_n-\mathbb{E}[S_n]|>\mathbb{E}[S_n]/2] &\leqslant \frac{\mathbb{E}[|S_n-\mathbb{E}[S_n]|^2]}{(\mathbb{E}[S_n]/2)^2} = \frac{4\mathrm{Var}[S_n]}{(\mathbb{E}[S_n])^2}\\ &= \frac{4\sum_{j=1}^n \mathbb{P}[E_j](1-\mathbb{P}[E_j])}{(\sum_{j=1}^n \mathbb{P}[E_j])^2} \leqslant \frac{4}{\sum_{j=1}^n \mathbb{P}[E_j]} \to 0\end{aligned}$$

since $\mathbb{E}[S_n]=\sum_{j=1}^n \mathbb{P}[E_j]\to\infty$. Thus we have $\mathbb{P}[|S_n-\mathbb{E}[S_n]|\leqslant \mathbb{E}[S_n]/2]\to 1$ as $n\to\infty$. Then,

$$\mathbb{P}[S_n\geqslant \mathbb{E}[S_n]/2] \leqslant \mathbb{P}[\mathbb{E}[S_n]/2\leqslant S_n\leqslant 3\mathbb{E}[S_n]/2]\to 1$$

as $n\to\infty$. Since both sums on the left side increase with $n$ we may let $n$ tend to infinity in $S_n$ and then in $\mathbb{E}[S_n]$, to conclude the desired result, i.e., $\mathbb{P}[\lim_n S_n=\infty]=1$. □

**2.3. Borel-Cantelli Zero-one Law.** Combing Lemma 3.15 and 3.18, we get the following zero-one law:

THEOREM 3.20 (Borel-Cantelli zero-one law). *Suppose the events $(E_n)$ are (pairwise) independent. Then*

(i) $\sum_n \mathbb{P}[E_n]<\infty \iff \mathbb{P}[E_n \text{ i.o.}]=0$.
(ii) $\sum_n \mathbb{P}[E_n]=\infty \iff \mathbb{P}[E_n \text{ i.o.}]=1$.

PROOF. The *necessity* of (a) and (b) are no need to prove again.

Now consider the *sufficiency* of (a). Suppose $\mathbb{P}[E_n \text{ i.o.}]=0$, and we need to show $\sum_n \mathbb{P}[E_n]<\infty$. We just assume it fails and derive a contradiction. If $\sum_n \mathbb{P}[E_n]=\infty$, then use the necessity of (b) we get $\mathbb{P}[E_n \text{ i.o.}]=1$, which contradicts the hypothesis that $\mathbb{P}[E_n \text{ i.o.}]=0$.

The *sufficiency* of (b) is similar and we omit it. □

## 3. More on Convergence

**3.1. Subsequence Principle.** It is well known that a sequence of reals converges to a limit $a$ in $\mathbb{R}$ if and only if *every subsequence* contains a subsequence converging to $a$. For convergence in probability, the following theorem is a further step of Theorem 3.16.

THEOREM 3.21 (Subsequence principle - part 2). *The sequence $(X_n)$ converges to $X$ in probability if and only if, each subsequence $(X_{n_k})$ contains a further subsequence $(X_{n_{k(i)}})$ which converges almost surely to $X$.*

PROOF. Consider the *necessity*. Suppose $X_n \to X$ in probability, that is, for all $\varepsilon > 0$ and $\delta > 0$, there is $N$ in $\mathbb{N}$ such that $\mathbb{P}[|X_n - X| \geqslant \varepsilon] \leqslant \delta$ for all $n \geqslant N$. We need to show, pick any subsequence $(X_{n_k})$, then $X_{n_k} \to X$ in probability, then there is a further subsequence which converges to $X$ almost surely using Theorem 3.16. Indeed, for such $\varepsilon$ and $\delta$, there is $k_0$ in $\mathbb{N}$ such that $n_{k_0} \geqslant N$. Then, $\mathbb{P}[|X_{n_k} - X| \geqslant \varepsilon] \leqslant \delta$ for all $k \geqslant k_0$, that is, $X_{n_k} \to X$ in probability.

Next, consider the *sufficiency*. Suppose any subsequence has a further subsequence converging almost surely to $X$. To show $X_n \to X$ in probability, assume this fails and get a contradiction. If $X_n$ fails to converge in probability, that is, there are $\varepsilon > 0$ and $\delta > 0$, for all $k$ in $\mathbb{N}$, we have $n_k \geqslant k$ such that $\mathbb{P}[|X_{n_k} - X| \geqslant \varepsilon] \geqslant \delta$. But every subsequence, such as $(X_{n_k})$ is assumed to have a further subsequence $(X_{n_{k(i)}})$ converges almost surely, and hence in probability. But

$$\mathbb{P}[|X_{n_{k(i)}} - X| \geqslant \varepsilon] \geqslant \delta,$$

which contradicts the definition of convergence in probability. □

**3.2. Cauchy Criterion.** Sometimes we don't know what the limit of the sequence is, and that's when we need the *Cauchy criterion* to prove convergence.

THEOREM 3.22 (Cauchy criterion - almost sure convergence). *The sequence $(X_n)$ is convergent almost surely if and only if $\lim_{n,m\to\infty} |X_n - X_m| = 0$ almost surely.*

For almost sure convergence, Theorem 3.22 follows from the corresponding result for real numbers, since the assumption is that $\{X_n(\omega), n \geqslant 1\}$ is Cauchy sequence for almost all $\omega$.

To make the preceding practical we provide some details. For this purpose, let

$$Y_n = \sup_{i,j\geqslant n} |X_i - X_j|, \quad Z_n = \sup_k |X_{n+k} - X_n|.$$

The meaning of the Cauchy criterion is that $(X_n(\omega))$ is Cauchy if and only if $Y_n(\omega) \to 0$. And $Y_n(\omega) \to 0$ if and only if $Z_n(\omega) \to 0$, because $Z_n \leqslant Y_n \leqslant 2Z_n$.

LEMMA 3.23 (Practical methods to verify Cauchy sequence). *The following are equivalent: $(X_n)$ is almost surely convergent; $(Y_n)$ converges to $0$ almost surely; and $(Z_n)$ converges to $0$ almost surely.*

EXAMPLE 3.24. Suppose that

$$\liminf_n \lim_m \mathbb{P}\Big[\sup_{k\leqslant m} |X_{n+k} - X_n| \geqslant \varepsilon\Big] = 0$$

for every $\varepsilon > 0$. Then, $(X_n)$ is convergent almost surely.

PROOF. Let $Z_{n,m}$ be the random variable that figures inside the event on the left. Note that $Z_{n,m} \uparrow Z_n$ as $m \to \infty$. Therefore, $\mathbb{1}_{[\varepsilon,\infty)} \circ Z_{n,m} \uparrow \mathbb{1}_{[\varepsilon,\infty)} \circ Z_n$, which together with Fatou's lemma and dominated convergence theorem gives

$$\mathbb{E}[\liminf_n \mathbb{1}_{[\varepsilon,\infty)} \circ Z_n] \leqslant \liminf_n \mathbb{E}[\mathbb{1}_{[\varepsilon,\infty)} \circ Z_n] = \liminf_n \lim_m \mathbb{E}[\mathbb{1}_{[\varepsilon,\infty)} \circ Z_{n,m}] = 0,$$

the last equality being the hypothesis. Since a positive variable with $0$ expectation is almost surely $0$, we have shown that $\liminf_n \mathbb{1}_{[\varepsilon,\infty)} \circ Z_n = 0$ almost surely. This is for every $\varepsilon > 0$. Since $Y_n$ as we defined before is bounded by $2Z_n$, it follows that, for all $\varepsilon > 0$, $\liminf_n \mathbb{1}_{[\varepsilon,\infty)} \circ Y_n = 0$ almost surely.

But $(Y_n)$ is a decreasing sequence and $\mathbb{1}_{[\varepsilon,\infty)}(y)$ is either 0 or 1. So, for all $\varepsilon > 0$, $\sum_n \mathbb{1}_{[\varepsilon,\infty)} \circ Y_n < \infty$ almost surely, which implies that $Y_n \to 0$ almost surely. Finally, we got the desired result through Lemma 3.23. □

THEOREM 3.25 (Cauchy criterion - convergence in probability). *The sequence $(X_n)$ converges in probability if and only if, for all $\varepsilon > 0$,*

$$\lim_{n,m\to\infty} \mathbb{P}[|X_n - X_m| \geqslant \varepsilon] = 0.$$

PROOF. First consider the *necessity*. Assume $X_n \to X$ in probability. Pick $\varepsilon > 0$, let $\delta = \varepsilon/2$. Observe that

$$\{|X_m - X_n| \geqslant \varepsilon\} \subset \{|X_m - X| \geqslant \delta\} \cup \{|X_n - X| \geqslant \delta\},$$ [1]

then

$$\mathbb{P}[|X_m - X_n| \geqslant \varepsilon] \leqslant \mathbb{P}[|X_m - X| \geqslant \delta] + \mathbb{P}[|X_n - X| \geqslant \delta] \to 0$$

by the monotonicity of probability measure, and the assumption that $X_n \to X$ in probability.

Next, consider the *sufficiency*. Suppose $\lim_{n,m\to\infty} \mathbb{P}[|X_n - X_m| \geqslant \varepsilon] = 0$ for all $\varepsilon > 0$. Let $\varepsilon_k = 2^{-k}$ and $n_0 = 0$. Then, for all $k \geqslant 1$, let $n_k > n_{k-1}$ be such that for all $m, n \geqslant n_k$, we have

$$\mathbb{P}[|X_m - X_n| \geqslant \varepsilon_k] \leqslant 2^{-k}.$$

Put $Y_k = X_{n_k}$, then

$$\mathbb{P}[|Y_{k+1} - Y_k| \geqslant \varepsilon_k] \leqslant 2^{-k}.$$

Thus we have $\sum_k \mathbb{P}[|Y_{k+1} - Y_k| \geqslant \varepsilon_k] < \infty$. Using the convergence part of Borel-Cantelli lemma we have $\mathbb{P}[|Y_{k+1} - Y_k| \geqslant \varepsilon_k \text{ i.o.}] = 0$, that is, there is $N$ in $\mathbb{N}$ such that $|Y_{k+1} - Y_k| < \varepsilon_k$ for all $k \geqslant N$. Now, For $n > m \geqslant N$, we have

$$|Y_n - Y_m| = |Y_n - Y_{n-1}| + \cdots + |Y_{m+1} - Y_m| \leqslant \varepsilon_{n-1} + \cdots + \varepsilon_m \leqslant \sum_{j\geqslant m} \varepsilon_k,$$

where the right side goes to $\infty$ as $m \to \infty$ since $\sum_k \varepsilon_k = \sum_k 2^{-k} < \infty$. This means that $(Y_n)$ is a Cauchy sequence, then $Y_n$ converges almost surely; let $X$ be its limit.

Obverse that, for $\varepsilon > 0$ with $\delta = \varepsilon/2$, we have

$$\mathbb{P}[|X_n - X| \geqslant \varepsilon] \leqslant \mathbb{P}[|X_n - X_{n_k}| \geqslant \delta] + \mathbb{P}[|Y_n - X| \geqslant \delta].$$

Now, as $n$ and $k$ tend to $\infty$, the first term on the right side goes to 0 by the hypothesis, and the second term goes to 0 since $Y_k \to X$ almost surely and hence in probability. It follows that $X_n \to X$ in probability. □

**3.3. Convergence and Continuous Functions.** With the help of the subsequence principle, we have the following result:

THEOREM 3.26 (Convergence and continuous functions).

(i) *If $X_n \to X$ almost surely, and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \to g(X)$ almost surely.*
(ii) *If $X_n \to X$ in probability, and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \to g(X)$ in probability.*

PROOF. (i). Suppose $X_n \to X$ almost surely, then there exists a negligible set $\mathcal{N}$ such that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega\backslash\mathcal{N}$. On $\Omega\backslash\mathcal{N}$, we have the following.

Since $g : \mathbb{R} \to \mathbb{R}$ is continuous, that is, for all $\varepsilon > 0$, there is $\delta > 0$ such that $|g(x) - g(y)| < \varepsilon$ for all $|x - y| < \delta$. Besides, for such $\delta > 0$, there is $N$ in $\mathbb{N}$ such that $|X_n(\omega) - X(\omega)| < \delta$ for all

[1] You should show it carefully.

$n \geqslant N$. Then, for all $n \geqslant N$, we have $|g(X_n(\omega)) - g(X(\omega))| < \varepsilon$ holds for all $\omega \in \Omega\backslash\mathcal{N}$. That is, $g(X_n) \to g(X)$ almost surely.

(ii). Let $(g(X_{n_k}))$ be some subsequence of $g(X_n)$. It suffices to find an almost sure convergence subsequence $(g(X_{n_{k(i)}}))$. But we know $(X_{n_k})$ has some almost sure convergence sequence $(X_{n_{k(i)}})$ such that $X_{n_{k(i)}} \to X$ almost surely. Thus $g(X_{n_{k(i)}}) \to g(X)$ almost surely by the result showed in (a). □

**3.4. Convergence and Arithmetic Operations.** Finally, the following theorem shows that convergence (almost sure or in probability) is preserved under arithmetical operations. The method for convergence in probability is similar to Theorem 3.26.

THEOREM 3.27 (Convergence and arithmetic operations). *Suppose that $X_n \to X$ and $Y_n \to Y$ almost surely (or, in probability). Then,*

(i) *$X_n \pm Y_n \to X \pm Y$ almost surely (or, in probability);*
(ii) *$X_nY_n \to XY$ almost surely (or, in probability).*
(iii) *Moreover, $X_n/Y_n \to X/Y$ almost surely (or, in probability) provided that, with probability one, $Y$ and $Y_n$'s are non-zero.*

PROOF. We only need to prove the "almost sure" part. Suppose $X_n \to X$ almost surely, then there exists a negligible set $\mathcal{N}_1$ such that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega\backslash\mathcal{N}_1$. Similarly, there exists a negligible set $\mathcal{N}_2$ such that $Y_n(\omega) \to Y(\omega)$ for all $\omega \in \Omega\backslash\mathcal{N}_2$. Now, consider the convergence on $\Omega\backslash(\mathcal{N}_1 \cup \mathcal{N}_2)$, where $\mathcal{N}_1 \cup \mathcal{N}_2$ is still a negligible set.

(i). It is obvious that $|(X_n \pm Y_n) - (X \pm Y)| \leqslant |X_n - X| + |Y_n - Y|$. Then, for all $\varepsilon > 0$, there exist $N_1$ and $N_2$ in $\mathbb{N}$ such that $|X_n - X| \leqslant \varepsilon/2$ and $|Y_m - Y| \leqslant \varepsilon/2$ for all $n \geqslant N_1$ and $m \geqslant N_2$. Let $N = N_1 \vee N_2$, then

$$|(X_n \pm Y_n) - (X \pm Y)| \leqslant |X_n - X| + |Y_n - Y| \leqslant \varepsilon$$

for all $n \geqslant N$. So $X_n \pm Y_n \to X \pm Y$ almost surely.

(ii). Fix $\omega \in \Omega\backslash(\mathcal{N}_1 \cup \mathcal{N}_2)$. Since every convergent sequence is bounded, there is $K$ in $\mathbb{N}$ such that $|Y_n(\omega)| \leqslant K$ for all $n$ in $\mathbb{N}$. Similarly, there is $L$ such that $|X| \leqslant L$. Then,

$$|X_nY_n - XY| \leqslant |(X_n - X)Y_n| + |X(Y_n - Y)| \leqslant K|X_n - X| + L|Y_n - Y|$$

holds at $\omega \in \Omega\backslash(\mathcal{N}_1 \cup \mathcal{N}_2)$. So the convergence is obvious from the above inequality. We just omit those details because I think you have enough mathematical maturity.

(iii). It is sufficient to prove $1/Y_n \to 1/Y$ almost surely. Similarly, consider $\omega \in \Omega\backslash(\mathcal{N}_1 \cup \mathcal{N}_2)$. For $|X|/2 > 0$, there exists $N_0$ in $\mathbb{N}$ such that $|X_n - X| \leqslant |X|/2$ for all $n \geqslant N_0$. Then,

$$|X_n| = |X - (X - X_n)| \geqslant |X| - |X - X_n| \geqslant |X|/2,$$

and

$$\Big|\frac{1}{X_n} - \frac{1}{X}\Big| = \frac{|X_n - X|}{|X_n|\,|X|} \leqslant \frac{2|X_n - X|}{|X|^2} \to 0$$

because $2/|X|^2$ is finite at $\omega \in \Omega\backslash(\mathcal{N}_1 \cup \mathcal{N}_2)$. Thus we complete the proof. □

## 4. More on Uniform Integrability

**4.1. Convergence, Cauchy and Uniform Integrability.** We will talk about convergence in $L^p$ and in probability in this section. More precisely, we will consider the relation of convergence, Cauchy in $L^p$ and uniform integrability. We will assume $0 < p < \infty$ in this section unless we specify something.

THEOREM 3.28 (Convergence, Cauchy and uniform integrability). *Let $(X_n) \subset L^p$ be a sequence of real-valued variables. For it, the following are equivalent:*

(i) *It converges in $L^p$.*
(ii) *It is Cauchy in $L^p$, that is, $\mathbb{E}[|X_m - X_n|^p] \to 0$ as $m, n \to \infty$.*
(iii) *It converges in probability and $(X_n^p)$ is uniformly integrable.*

REMARK 3.29 ($C_p$ inequality). We will frequently use this inequality:

$$\begin{aligned} |a \pm b|^p &\leqslant 2^{p-1}(|a|^p + |b|^p), && \text{if } p \geqslant 1; \\ &\leqslant |a|^p + |b|^p, && \text{if } 0 < p < 1. \end{aligned}$$

So we can always assume $|a \pm b|^p \leqslant C_p(|a|^p + |b|^p)$ for all $0 < p < \infty$.

PROOF. (i) $\Rightarrow$ (ii). Suppose $X_n \to X$ in $L^p$, that is, $\mathbb{E}[|X_n - X|^p] \to 0$. Then,

$$\mathbb{E}[|X_m - X_n|^p] \leqslant \mathbb{E}[C_p(|X_m - X|^p + |X_n - X|^p)] \leqslant C_p\big(\mathbb{E}[|X_m - X|^p] + \mathbb{E}[|X_n - X|^p]\big) \to 0$$

as $m, n \to \infty$. Hence, $(X_n)$ is Cauchy in $L^p$.

(ii) $\Rightarrow$ (iii). Assume (ii). For every $\varepsilon > 0$, by Chebyshev's inequality,

$$\mathbb{P}[|X_m - X_n| > \varepsilon] \leqslant \frac{1}{\varepsilon^p}\mathbb{E}[|X_m - X_n|^p] \to 0$$

as $m, n \to \infty$. Thus, Theorem 3.25 applies, and the sequence converges in probability. To show that the sequence is uniformly integrable, we use $\varepsilon$-$\delta$ characterization of uniform integrability. That is, we need to verify

1. $\sup_n \mathbb{E}[|X_n^p|] = \sup_n \mathbb{E}[|X_n|^p] < \infty$ and,
2. for all $\varepsilon > 0$, there is $\delta > 0$ such that for all event $H$: $\mathbb{P}[H] < \delta$ implies $\sup_n \mathbb{E}[|X_n|^p \mathbb{1}_H] < \varepsilon$.

Now, fix $\varepsilon > 0$. Since the sequence is Cauchy in $L^p$, there exists $k = k(\varepsilon)$ in $\mathbb{N}$ such that $\mathbb{E}[|X_m - X_n|^p] \leqslant \varepsilon$ for all $m, n \geqslant k$. Thus, for every event $H$,

$$\mathbb{E}[|X_n|^p \mathbb{1}_H] \leqslant C_p\big(\mathbb{E}[|X_n - X_k|^p \mathbb{1}_H] + \mathbb{E}[|X_k|^p \mathbb{1}_H]\big) \leqslant C_p\big(\varepsilon + \mathbb{E}[|X_k|^p \mathbb{1}_H]\big)$$

for all $n \geqslant k$. Consequently,

$$\sup_n \mathbb{E}[|X_n|^p \mathbb{1}_H] \leqslant C_p\big(\varepsilon + \sup_{m \leqslant k} \mathbb{E}[|X_m|^p \mathbb{1}_H]\big). \tag{3.7}$$

On the right side, the *finite* collection $\{X_1^p, \cdots, X_k^p\}$ is uniformly integrable since the $X_n$'s are in $L^p$. Hence, by the $\varepsilon$-$\delta$ characterization, there is $\delta > 0$ such that $p[H] < \delta$ implies that the supremum over $m \leqslant k$ is bounded by $\varepsilon$ and therefore the supremum on the left side is bounded by $C\varepsilon$, where $C$ is a constant. Finally, taking $H = \Omega$, we see that

$$\sup_n \mathbb{E}[|X_n|^p] \leqslant C_p\big(\varepsilon + \sup_{m \leqslant k} \mathbb{E}[|X_m|^p]\big) < \infty.$$

Thus, using the $\varepsilon$-$\delta$ characterization again, we see that $(X_n^p)$ is uniform integrable.

(iii) $\Rightarrow$ (i). Suppose (iii). Since $X_n \to X$ in probability, there is a subsequence $(X_n')$ converges to $X$ almost surely. Then Fatou's lemma yields

$$\mathbb{E}[|X|^p] = \mathbb{E}\big[\liminf_n |X_n'|^p\big] \leqslant \liminf_n \mathbb{E}[|X_n'|^p] \leqslant \sup_n \mathbb{E}[|X_n|^p] < \infty,$$

where the supremum is finite by the assumed uniform integrability. Then, we know that $X$ in $L^p$. To show that $X_n \to X$ in $L^p$. Fix $\varepsilon > 0$ and let $H_n = \{|X_n - X| > \varepsilon\}$. Now, obviously,

$$\begin{aligned} \mathbb{E}[|X_n - X|^p] &= \mathbb{E}[|X_n - X|^p \mathbb{1}_{H_n^c}] + \mathbb{E}[|X_n - X|^p \mathbb{1}_{H_n}] \\ &\leqslant \varepsilon^p + \mathbb{E}[|X_n - X|^p \mathbb{1}_{H_n}]. \end{aligned} \tag{$\star$}$$

Since $X$ is integrable; $(X_n^p)$ is uniform integrable, and

$$|X_n - X|^p \leqslant C_p\big(|X_n|^p + |X|^p\big)$$

holds for all $p > 0$, then $(|X_n - X|^p)$ is uniformly integrable. Thus, for such fixed $\varepsilon$, there is $\delta > 0$ such that all event $H$ satisfying $\mathbb{P}[H] < \delta$ implies

$$\mathbb{E}[|X_n - X|^p \mathbb{1}_H] < \varepsilon.$$

For this $\delta$, notice we can find $n$ so large such that $\mathbb{P}[H_n] < \delta$ since $X_n \to X$ in probability, so

$$\mathbb{E}[|X_n - X|^p \mathbb{1}_{H_n}] < \varepsilon.$$

Finally, we have $\mathbb{E}[|X_n - X|^p] \leqslant M\varepsilon$ using $(\star)$, where $M$ is a constant. This completes the proof that the sequence converges in $L^p$. □

THEOREM 3.30 (Vitali's theorem). *Let $(X_n) \subset L^p$ be a sequence of real-valued variables which converges to $X$ in probability. Then, the following are equivalent:*

(i) *It converges to $X$ in $L^p$.*
(ii) *$(X_n^p)$ is uniformly integrable.*
(iii) $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p] < \infty$.

PROOF. We have (i) ⇔ (ii) by Theorem 3.28. Now prove (i) ⇒ (iii). If $X_n \to X$ in $L^p$ for $p \geqslant 1$. Using Minkowski's inequality we get

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

where $X = X_n - (X_n - X)$. Since $X_n \to X$ in $L^p$, we get $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$ as $n \to \infty$.

If $X_n \to X$ in $L^p$ for $p < 1$. Applying $C_p$ inequality: $|x + y|^p \leqslant |x|^p + |y|^p$ we get

$$\mathbb{E}[|X_n|^p] - \mathbb{E}[|X_n - X|^p] \leqslant \mathbb{E}[|X|^p] \leqslant \mathbb{E}[|X_n|^p] + \mathbb{E}[|X_n - X|^p].$$

This completes the proof that $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$.

(iii) ⇒ (ii). For all $A > 0$, define the nonnegative bounded continuous function $f_A$ like:

$$\begin{aligned} f_A(x) &= |x|^p, && \text{if } |x|^p \leqslant A, \\ &\leqslant |x|^p, && \text{if } A < |x|^p \leqslant A + 1, \\ &= 0, && \text{if } |x|^p > A + 1. \end{aligned}$$

Therefore, by Theorem 3.26, we have $f_A \circ X_n \to f_A \circ X$ in probability. Applying the implication (ii) ⇒ (iii) with $p = 1$ to the sequence $(f_A \circ X_n)_{n \geqslant 1}$, it follows that $\mathbb{E}[f_A \circ X_n] \to \mathbb{E}[f_A \circ X]$ (It should be noticed that the sequence $(f_A \circ X_n)$ is uniformly integrable since $f_A$ is bounded).

Since $0 \leqslant |x|^p \mathbb{1}_{\{|x|^p \leqslant A\}} \leqslant f_A(x) \leqslant |x|^p \mathbb{1}_{\{|x|^p \leqslant A+1\}}$, then

$$\liminf_n \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n|^p \leqslant A+1\}}] \geqslant \liminf_n \mathbb{E}[f_A(X_n)] = \mathbb{E}[f_A(X)] \geqslant \mathbb{E}[|X|^p \mathbb{1}_{\{|X|^p \leqslant A\}}].$$

Subtracting this from the assumption (iii) we get

$$\begin{aligned} \limsup_n \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n|^p > A+1\}}] &= \limsup_n \Big[\mathbb{E}[|X_n|^p] - \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n|^p \leqslant A+1\}}]\Big] \\ &= \mathbb{E}[|X|^p] - \liminf_n \mathbb{E}[|X_n|^p \mathbb{1}_{\{|X_n|^p \leqslant A+1\}}] \\ &\leqslant \mathbb{E}[|X|^p] - \mathbb{E}[|X|^p \mathbb{1}_{\{|x|^p \leqslant A\}}] = \mathbb{E}[|X|^p \mathbb{1}_{\{|X|^p > A\}}]. \end{aligned}$$

The last integral does not depend on $n$ and converges to zero as $A \to \infty$. This means: for all $\varepsilon > 0$, there is $A_0 = A_0(\varepsilon)$ and $n_0 = n_0(A_0)$ such that we have

$$\sup_{n \geqslant n_0} \mathbb{E}\big[|X_n|^p \mathbb{1}_{\{|X_n|^p > A+1\}}\big] \leqslant \varepsilon$$

if $A \geqslant A_0$. However, since each $X_n^p$ is integrable, there is $A_1 = A_1(\varepsilon)$ such that

$$\sup_{n \geqslant 1} \mathbb{E}\big[|X_n|^p \mathbb{1}_{\{|X_n|^p > A+1\}}\big] \leqslant \varepsilon$$

if $A \geqslant \max\{A_0, A_1\}$. Thus $(X_n^p)$ is uniformly integrable by definition. □

REMARK 3.31. Theorem 3.30 states that:

$$\text{Convergence in prob. } + \text{ u.i.} \quad \Rightarrow \quad \text{Convergence in } p\text{th mean.}$$

In fact, convergence in probability can be further weakened to convergence in distribution. We will introduce this case in the next section. †

EXERCISE 3.32. Prove (iii) ⇒ (i) directly.

PROOF. Define $Y_n = (|X_n| + |X|)^p - |X_n - X|^p$. Since $X_n \to X$ in probability, then $|X_n| \to |X|$ in probability and $|X_n - X| \to 0$ in probability, which implies that $Y_n \to 2^p|X|^p$ in probability. Using Fatou's lemma, we have

$$\begin{aligned}
\mathbb{E}[2^p|X|^p] &\leqslant \liminf_n \mathbb{E}[(|X_n| + |X|)^p - |X_n - X|^p] \\
&\leqslant \liminf_n \mathbb{E}\big[2^{p-1}\big(|X_n|^p + |X|^p\big) - |X_n - X|^p\big] \quad \text{[using } C_p \text{ inequality]} \\
&\leqslant 2^{p-1} \liminf_n \mathbb{E}[|X_n|^p] + 2^{p-1}\mathbb{E}[|X|^p] - \limsup_n \mathbb{E}[|X_n - X|^p] \\
&\leqslant 2^{p-1}\mathbb{E}[|X|^p] + 2^{p-1}\mathbb{E}[|X|^p] - \limsup_n \mathbb{E}[|X_n - X|^p] \\
&= 2^p\mathbb{E}[|X|^p] - \limsup_n \mathbb{E}[|X_n - X|^p].
\end{aligned}$$

Here, we use the assumption $\|X_n\|_p \to \|X\|_p$ in the fourth inequality. From above inequality, we know that

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

which gives the result that $\|X_n - X\|_p \to 0$ since $\lim_n \mathbb{E}[|X_n - X|^p] \leqslant \limsup_n \mathbb{E}[|X_n - X|^p] = 0$. □

## 5. Weak Convergence

**5.1. Definitions and Examples.** Recall that $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$ denote the set of all *bounded continuous* from $\mathbb{R}$ into $\mathbb{R}$. We equip $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$ with the supremum norm

$$\|\varphi\| = \sup_{x \in \mathbb{R}} |\varphi(x)|.$$

We let $M_1(\mathbb{R})$ denote the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

DEFINITION 3.33 (Weak convergence & convergence in distribution). A sequence $(\mu_n)$ in $M_1(\mathbb{R})$ **converges weakly** to $\mu$ in $M_1(\mathbb{R})$ if for all $\varphi \in \mathcal{C}_b(\mathbb{R}, \mathbb{R})$,

$$\lim_n \mu_n(\varphi) = \mu(\varphi).$$

A sequence $(X_n)$ of random variables with values in $\mathbb{R}$ **converges in distribution** to a random variable $X$ with values in $\mathbb{R}$ if their distribution measures $\mu_n$ converges weakly to $\mu$. This is

equivalent to saying that for all $\varphi \in \mathcal{C}_b(\mathbb{R}, \mathbb{R})$,

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

We will write $\mu_n \Rightarrow \mu$, resp. $X_n \Rightarrow X$, if the sequence $(\mu_n)$ converges weakly to $\mu$, resp. if $(X_n)$ converges in distribution to $X$.

REMARK 3.34. There is some abuse of terminology in saying that the sequence $(X_n)$ converges in distribution to $X$, because the limiting random variable $X$ is NOT determined *uniquely* (only its distribution $\mu$ is determined). For this reason, we will sometimes write that a sequence $(X_n)$ of random variables converges in distribution to a probability measure $\mu$ — one should of course understand that the laws $\mu_n$ converge weakly to $\mu$.

We may also note that the convergence in distribution makes sense even if the random variables $X_n$'s are defined on different probability spaces. This makes the convergence in distribution very different from the other types of convergence studied in this chapter. But we may assume that they are defined on a same probability space. †

EXAMPLE 3.35 (Discrete random variables). Suppose that the random variables $(X_n)$ and $X$ take values in $\mathbb{Z}$. Then $X_n$ converges in distribution to $X$ if and only if for all $x$ in $\mathbb{Z}$, $\mathbb{P}[X_n = x] \to \mathbb{P}[X = x]$. The sufficiency part require a little work, but will be immediate when we have established that $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$ can be replaced by $\mathcal{C}_c(\mathbb{R}, \mathbb{R})$ in the definition of the weak convergence (See Theorem 3.43).†

EXAMPLE 3.36 (Probability density functions). Suppose that, for every $n \in \mathbb{N}$, $X_n$ has a density $p_n(x)$, and that there exists a probability density function $p(x)$ on $\mathbb{R}$ such that

$$p_n(x) \to p(x) \quad \text{almost everywhere},$$

where almost everywhere refers to Lebesgue measure on $\mathbb{R}$. Then $X_n$ converge in distribution to a random variable $X$ with law $\mu(\mathrm{d}x) = p(x)\,\mathrm{d}x$. Indeed, consider a function $\varphi \in \mathcal{C}_b(\mathbb{R})$ such that $0 \leqslant \varphi \leqslant 1$ (clearly, we may restrict our attention to that case). Then, Fatou's lemma shows that

$$\liminf_n \int \varphi(x) p_n(x)\,\mathrm{d}x \geqslant \int \varphi(x) p(x)\,\mathrm{d}x,$$
$$\liminf_n \int (1-\varphi(x)) p_n(x)\,\mathrm{d}x \geqslant \int (1-\varphi(x)) p(x)\,\mathrm{d}x.$$

By combining these two bounds, and using the fact that $p(x)$ is a probability density function, we get

$$\lim_n \int \varphi(x) p_n(x)\mathrm{d}x = \int \varphi(x) p(x)\mathrm{d}x.$$

The following theorem shows that: when we consider the convergence in distribution, we can ignore the "small probabilistic differences" of the sequence.

THEOREM 3.37 (Insensitivity of convergence in distribution). *If $X_n \Rightarrow X$ and $\mathbb{P}[X_n \neq Y_n] \to 0$ as $n \to \infty$, then $Y_n \Rightarrow X$.*

PROOF. By definition, for all $\varphi$ in $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$, we have $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$, and we want to show that $\mathbb{E}[\varphi(Y_n)] \to \mathbb{E}[\varphi(X)]$ as $n \to \infty$. Indeed,

$$\begin{aligned}
\big|\mathbb{E}[\varphi(Y_n) - \varphi(X)]\big| &\leqslant \big|\mathbb{E}[\varphi(Y_n) - \varphi(X_n)]\big| + \big|\mathbb{E}[\varphi(X_n) - \varphi(X)]\big| \\
&\leqslant \mathbb{E}\big[|\varphi(Y_n) - \varphi(X_n)| \mathbb{1}_{\{X_n \neq Y_n\}}\big] + \mathbb{E}\big[|\varphi(Y_n) - \varphi(X_n)| \mathbb{1}_{\{X_n = Y_n\}}\big] + \big|\mathbb{E}[\varphi(X_n) - \varphi(X)]\big| \\
&= \mathbb{E}\big[|\varphi(Y_n) - \varphi(X_n)| \mathbb{1}_{\{X_n \neq Y_n\}}\big] + \big|\mathbb{E}[\varphi(X_n) - \varphi(X)]\big| = (\star).
\end{aligned}$$

Since $\varphi$ in $\mathcal{C}_b(\mathbb{R},\mathbb{R})$, we may assume that $|\varphi(x)| \leqslant M$ for all $x$ in $\mathbb{R}$, so $|\varphi(x)-\varphi(y)| \leqslant 2M$ for all $x$ and $y$ in $\mathbb{R}$. So

$$(\star) \leqslant 2M\,\mathbb{P}[X_n \neq Y_n] + \big|\mathbb{E}[\varphi(X_n)-\varphi(X)]\big| \to 0$$

as $n \to \infty$ since $M$ is finite; $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$ and $\mathbb{P}[X_n \neq Y_n] \to 0$ as $n \to \infty$. This completes the proof. □

Indeed, convergence in probability implies convergence in distribution. Then, we have the following relation between different modes of convergence.

THEOREM 3.38 (Convergence in probability implies convergence in distribution). *If the sequence $X_n$ converges to $X$ in probability, then it also converges to $X$ in distribution.*

PROOF. Suppose first that $X_n \to X$ almost surely. Then, for all $\varphi \in \mathcal{C}_b(\mathbb{R},\mathbb{R})$, Theorem 3.26 gives that $\varphi(X_n) \to \varphi(X)$ almost surely and bounded convergence theorem gives $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$.

If $X_n \to X$ in probability, then we know from Theorem 3.21 that every subsequence of $\mathbb{N}$ has a further subsequence $N$ such that $X_n \to X$ along $N$ almost surely. Thus, $\varphi(X_n) \to \varphi(X)$ along $N$ almost surely by the continuity of $\varphi$, and $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$ along $N$ by bounded convergence theorem. This is only possible if $\mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)]$ as $n \to \infty$. □

ANOTHER PROOF OF THEOREM 3.38. We have known that $\varphi(X_n) \to \varphi(X)$ in probability, that is, $\varphi(X_n)-\varphi(X) \to 0$ in probability. Besides, since $\varphi$ is bounded, we know that $|\varphi(X_n)-\varphi(x)| \leqslant 2\max_{x\in\mathbb{R}}\varphi(x) < \infty$, so $\varphi(X_n)-\varphi(X)$ is dominated by a constant (in $L^1$), then using Theorem 3.11 we know that

$$\varphi(X_n)-\varphi(X) \to 0 \quad \text{in } L^1.$$

Then we get the desired result by Theorem 3.30. □

The converse of Theorem 3.38 may fails in general, because, as we already mentioned, the convergence in distribution of $X_n$ to $X$ does NOT determine the limiting variable $X$. But there is however a very special case where the converse holds.

EXAMPLE 3.39 (A counterexample). Suppose that $(X_n)$ converges to $X$ in distribution and $X = x_0$ for some fixed point $x_0$. Then, in particular, for $\varphi$ defined by $\varphi(x) = |x-x_0| \wedge 1$,

$$\mathbb{E}[|X_n - X| \wedge 1] = \mathbb{E}[\varphi(X_n)] \to \mathbb{E}[\varphi(X)] = \varphi(x_0) = 0.$$

Next, for all $\varepsilon$ in $(0,1)$, we have

$$\varepsilon \mathbb{1}_{\{|X_n-X|>\varepsilon\}} \leqslant |X_n - X| \wedge 1 \leqslant \varepsilon + \mathbb{1}_{\{|X_n-X|>\varepsilon\}},$$

where the second inequality holds since: $|X_n-X|\wedge \leqslant \varepsilon$ if $|X_n-X| \leqslant \varepsilon$; and $|X_n-X|\wedge 1 \leqslant 1$ if $|X_n-X| > \varepsilon$. Taking expectations, and letting $n \to \infty$ we get $X_n \to X$ in probability, which completes the proof. †

**5.2. Characterization Theorem.** If $(X_n)$ is a sequence that converges in distribution to $X$, it is not always true that

$$\mathbb{P}[X_n \in B] \to \mathbb{P}[X \in B]$$

when $B$ is a Borel set of $\mathbb{R}$.

EXAMPLE 3.40 (A counterexample). If, for every $n$ in $\mathbb{N}$, $X_n$ is uniformly distributed on $\{k/n : 1 \leqslant k \leqslant n\}$, then $X_n$ converges in distribution to the uniform distribution on $[0,1]$. This is just a special case of the approximation of the integral of a continuous function by Riemann sums. If we take $B$ the set of all rational numbers, then $\mathbb{P}[X_n \in B] = 1$ while $\mathbb{P}[X \in B] = 0$. †

Still we have the following theorem, where $\partial B$ denotes the boundary of a subset $B$ of $\mathbb{R}$.

THEOREM 3.41 (Portmanteau's theorem). *Let $(\mu_n)$ be a sequence in $M_1(\mathbb{R})$ and let $\mu$ in $M_1(\mathbb{R})$. The following are equivalent:*

(i) *The sequence $\mu_n$ converges weakly to $\mu$.*
(ii) *For every open subset $G$ of $\mathbb{R}$, we have $\liminf_n \mu_n(G) \geqslant \mu(G)$.*
(iii) *For every closed subset $F$ of $\mathbb{R}$, we have $\limsup_n \mu_n(F) \leqslant \mu(F)$.*
(iv) *For every Borel subset $B$ of $\mathbb{R}$ such that $\mu(\partial B) = 0$, $\lim_n \mu_n(B) = \mu(B)$.*

PROOF. (i) $\Rightarrow$ (ii). If $G$ is an open subset of $\mathbb{R}$, we can find a sequence $\varphi_p$ of $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$ such that $0 \leqslant \varphi_p \leqslant \mathbb{1}_G$ and $\varphi_p \uparrow \mathbb{1}_G$ (for instance, $\varphi_p = p \cdot d(x, G^c) \wedge 1$). Then, $\mu_n(\varphi_p) \leqslant \mu_n(\mathbb{1}_G)$, and thus

$$\liminf_n \mu_n(\mathbb{1}_G) \geqslant \liminf_n \mu_n(\varphi_p)$$

holds for all $p$ in $\mathbb{N}$. Taking supremum, we get

$$\liminf_n \mu_n(G) \geqslant \sup_p \big(\liminf_n \mu_n(\varphi_p)\big) = \sup_p \mu(\varphi_p) = \mu(G),$$

where the first equality holds since $\mu_n \Rightarrow \mu$ and the second equality by monotone convergence.

The equivalence (ii) $\Leftrightarrow$ (iii) is obvious since complements of open sets are closed and conversely.

Let us show that (ii) and (iii) imply (iv). If $B$ in $\mathcal{B}(\mathbb{R})$, write $\bar{B}$ for the closure of $B$ and $B^\circ$ for the interior of $B$ so that $\partial B = \bar{B} \backslash B^\circ$. By (b) and (c), we have

$$\limsup_n \mu_n(B) \leqslant \limsup_n \mu_n(\bar{B}) \leqslant \mu(\bar{B}),$$
$$\liminf_n \mu_n(B) \geqslant \liminf_n \mu_n(B^\circ) \geqslant \mu(B^\circ).$$

If $\mu(\partial B) = 0$, then $\mu(\bar{B}) = \mu(B^\circ) = \mu(B)$ and we give (d).

(iv) $\Rightarrow$ (i). Let $\varphi$ in $\mathcal{C}_b(\mathbb{R}, \mathbb{R})$. Choose $a$ and $b$ in $\mathbb{R}$ such that $a < f < b$. Fix $\varepsilon > 0$ arbitrary. Considering the probability measure $\mu \circ \varphi^{-1}$ on $(a, b)$, and pick $a = a_0 < a_1 < \cdots < a_k = b$ such that $a_i - a_{i-1} \leqslant \varepsilon$ for all $i$ and no $a_i$ is an atom for $\mu \circ \varphi^{-1}$ (this is possible since a probability measure has at most countably many atoms). Let $A_i = \varphi^{-1}((a_{i-1}, a_i])$, define

$$g = \sum_{i=1}^k a_{i-1} \mathbb{1}_{A_i}, \quad h = \sum_{i=1}^k a_i \mathbb{1}_{A_i}$$

and observe that

$$\varphi - \varepsilon \leqslant g \leqslant \varphi \leqslant h \leqslant \varphi + \varepsilon. \tag{3.8}$$

If $x \in \partial A_i$ then $\varphi(x)$ is either $a_{i-1}$ or $a_i$, neither of which is an atom for $\mu \circ \varphi^{-1}$. Thus, $\mu(\partial A_i) = 0$ and it follows from assuming (d) that $\mu_n(A_i) \to \mu(A_i)$ as $n \to \infty$ for $1 \leqslant i \leqslant k$. Thus, $\mu_n(g) \to \mu(g)$ and $\mu_n(h) \to \mu(h)$, and (3.8) yields

$$\begin{aligned} \mu(\varphi) - \varepsilon \;\leqslant\; \mu(g) = \lim_n \mu_n(g) \;&\leqslant\; \liminf_n \mu_n(\varphi) \\ &\leqslant\; \limsup_n \mu_n(\varphi) \;\leqslant\; \lim_n \mu_n(h) = \mu(h) \;\leqslant\; \mu(\varphi) + \varepsilon. \end{aligned}$$

In other words, limit inferior and limit superior of the sequence $(\mu_n(\varphi))$ are sandwiched between the numbers $\mu(\varphi) - \varepsilon$ and $\mu(\varphi) + \varepsilon$ for arbitrary $\varepsilon$. So, $\mu_n(\varphi) \to \mu(\varphi)$ as needed to show that (a) holds. □

COROLLARY 3.42 (Distribution functions). *A sequence $(X_n)$ of real random variables converges to $X$ in distribution if and only if the distribution functions $F_n(x)$ converge to $F(x)$ at every point $x$ where $F$ is continuous.*

PROOF. The *necessity* part is obvious from (iv) in Theorem 3.41. In the sufficiency part, suppose that $F_n(x) \to F(x)$ whenever $F$ is continuous at $x$. Since the set of discontinuity points of $F$ is at most countable, it follows that, for all $x$ in $\mathbb{R}$,

$$\liminf_n F_n(x-) \geqslant F(x-), \quad \limsup_n F_n(x) \leqslant F(x).$$

Indeed, take a sequence $x_p \uparrow x$, and such that $F$ is continuous at $x_p$ for each $p$ in $\mathbb{N}$, then we know that $F_n(x_p) \leqslant F_n(x-)$ and $F(x_p) \to F(x-)$ as $p \to \infty$. Now,

$$F(x-) \leftarrow F(x_p) = \lim_n F_n(x_p) \leqslant \liminf_n F_n(x-),$$

that is the first inequality, and the statement about the $\limsup$ is derived in a similar manner. Recalling that $\mu((a,b)) = F(b-) - F(a)$ for any $a < b$, it follows from the preceding display that (ii) of Theorem 3.41 holds for $\mu_n$ and $\mu$, where $\mu_n$ and $\mu$ are the distribution of $X_n$ and $X$, respectively, when $G$ is an open interval. Since any open subset of $\mathbb{R}$ is the disjoint union of at most countably many open intervals, we easily get that (ii) holds for any open subset, proving that $X_n$ converges to $X$ in distribution. □

**5.3. Weaker Conditions for Weak Convergence.** Recall that $\mathcal{C}_K(\mathbb{R},\mathbb{R})$ denote the set of all continuous functions with compact support on $\mathbb{R}$.

THEOREM 3.43 (Weaker equivalent conditions for weak convergence). *Let $\mu_n$ and $\mu$ be probability measures on $\mathbb{R}$. Let $H$ be a subset of $\mathcal{C}_b(\mathbb{R},\mathbb{R})$ whose closure (with respect to the supremum norm) contains $\mathcal{C}_K(\mathbb{R},\mathbb{R})$. The following are equivalent.*

(i) *The sequence $\mu_n$ converges weakly to $\mu$.*
(ii) *For all $\varphi$ in $\mathcal{C}_K(\mathbb{R},\mathbb{R})$, we have $\mu_n(\varphi) \to \mu(\varphi)$ as $n \to \infty$.*
(iii) *For all $\varphi$ in $H$, we have $\mu_n(\varphi) \to \mu(\varphi)$ as $n \to \infty$.*

PROOF. It is obvious that (i) ⇒ (ii) and (i) ⇒ (iii). Now prove (ii) ⇒ (i). Let $\varphi \in \mathcal{C}_b(\mathbb{R},\mathbb{R})$ and let $f_k$ be a sequence in $\mathcal{C}_K(\mathbb{R},\mathbb{R})$ such that $0 \leqslant f_k \leqslant 1$ for every $k$ in $\mathbb{N}$, and $f_k \uparrow 1$ as $k \to \infty$. Then, for every $k$ in $\mathbb{N}$, $\varphi f_k \in \mathcal{C}_K(\mathbb{R},\mathbb{R})$ and thus

$$\mu_n(\varphi f_k) \to \mu(\varphi f_k) \quad \text{as} \quad n \to \infty. \tag{3.9}$$

On the other hand,

$$\big|\mu_n(\varphi) - \mu_n(\varphi f_k)\big| \leqslant \|\varphi\|\mu_n(1-f_k) = \|\varphi\|\big(1-\mu_n(f_k)\big),$$

and similarly

$$\big|\mu(\varphi) - \mu(\varphi f_k)\big| \leqslant \|\varphi\|\big(1-\mu(f_k)\big).$$

Hence, using (3.9), we get for every $k$ in $\mathbb{N}$,

$$\begin{aligned}
\limsup_n \big|\mu_n(\varphi) - \mu(\varphi)\big| &\leqslant \limsup_n \big(\big|\mu_n(\varphi) - \mu_n(\varphi f_k)\big| + \big|\mu_n(\varphi f_k) - \mu(\varphi)\big|\big) \\
&= \limsup_n \big|\mu_n(\varphi) - \mu_n(\varphi f_k)\big| + \big|\mu(\varphi f_k) - \mu(\varphi)\big| \\
&\leqslant \|\varphi\|\big[\limsup_n \big(1-\mu_n(f_k)\big) + \big(1-\mu(f_k)\big)\big] \\
&= 2\|\varphi\|\big(1-\mu(f_k)\big).
\end{aligned}$$

Now we just have to let $k$ tend to $\infty$ (noting that $\mu(f_k) \uparrow 1$ by the monotone convergence theorem) and we find that $\mu_n(\varphi) \to \mu(\varphi)$, so that (a) holds.

It remains to prove (iii) $\Rightarrow$ (ii), so we assume that (iii) holds. If $\varphi \in \mathcal{C}_K(\mathbb{R}, \mathbb{R})$, then, for all $k$ in $\mathbb{N}$, we can find a function $\varphi_k \in H$ such that $\|\varphi - \varphi_k\| \leqslant k^{-1}$, and it follows that

$$\limsup_n \big|\mu_n(\varphi) - \mu(\varphi)\big| \leqslant \limsup_n \big[\big|\mu_n(\varphi - \varphi_k)\big| + \big|\mu_n(\varphi_k) - \mu(\varphi_k)\big| + \big|\mu(\varphi_k - \varphi)\big|\big] \leqslant 2/k,$$

where $\limsup_n \big|\mu_n(\varphi_k) - \mu(\varphi_k)\big| = 0$ by (iii) and $\mu_n(\varphi - \varphi_k) \leqslant \|\varphi - \varphi_k\|\, \mu_n(\mathbb{R}) \leqslant k^{-1}$. Since this inequality holds for all $k$ in $\mathbb{N}$, we get that $\mu_n(\varphi) \to \mu(\varphi)$. □

**5.4. Helly's Selection Principle.** The next result is useful in studying limits of sequences of distributions:

THEOREM 3.44 (Helly's selection principle). *For every sequence $F_n$ of distribution functions, there is a subsequence and a right continuous nondecreasing function $F$ such that $F_n(y) \to F(y)$ along this subsequence at all continuity points $y$ of $F$.*

REMARK 3.45 (Helly's selection principle). *The limit may NOT be a distribution function.* For example, if $a + b + c = 1$ and

$$F_n(x) = a\,\mathbb{1}_{\{x \geqslant n\}} + b\,\mathbb{1}_{\{x \geqslant -n\}} + c\,G(x),$$

where $G$ is a distribution function, then $F_n(x) \to F(x) = b + c\,G(x)$, but

$$\lim_{x \downarrow -\infty} F(x) = b \quad \text{and} \quad \lim_{x \uparrow \infty} F(x) = b + c = 1 - a.$$

In words, an amount of mass $a$ escapes to $+\infty$, and mass $b$ escapes to $-\infty$. The type of convergence that occurs in Theorem 3.44 is called **vague convergence**, and will be denoted here by $\Rightarrow_v$. †

PROOF OF THEOREM 3.44. Denote by $\mathbb{Q}$ the set of all rational numbers and enumerate it as $\{r_k : k \geqslant 1\}$. Consider the sequence $\{F_n(r_1) : n \geqslant 1\}$, which is bounded, and there exists a convergent subsequence, denoted by $F_{1n}(r_1) \to G(r_1)$. Consider the sequence $\{F_{1n}(r_2) : n \geqslant 1\}$. It is bounded, and contains a convergent subsequence, denoted by $F_{2n}(r_2) \to G(r_2)$. Note that $(F_{2n})$ is a subsequence of $(F_{1n})$, thus $F_{2n}(r_1) \to G(r_1)$. Continue in this way, we obtain

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

Choose the diagonal sequence $(F_{nn})$. We assert that it converges along all $r_j$'s. Define

$$G(r_j) = \lim_n F_{nn}(r_j)$$

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

$$F(x) = \inf\{G(r) : x < r \in \mathbb{Q}\}$$

for all $x$ in $\mathbb{R}$. The function $F$ has the following properties.

- First, $F$ is *increasing* because $G$ is increasing on $\mathbb{Q}$.
- Second, $F$ is *right-continuous*. To see this, one needs to check that, for each $x$ in $\mathbb{R}$, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $F(y) \leqslant F(x) + \varepsilon$ as long as $y \leqslant x + \delta$. This is true because, by the definition of infimum, one can find $q$ in $\mathbb{Q}$ such that $x < q$ and $F(x) \leqslant G(q) \leqslant F(x) + \varepsilon$. Then for any $y \leqslant q$, we have $F(y) \leqslant G(q) \leqslant F(x) + \varepsilon$.

• Finally, we will show that

$$\lim_n F_{nn}(x) = F(x)$$

holds for all continuity points $x$ of $F$. For any $p < p' < x < q' < q$ with $p, p', q, q'$ in $\mathbb{Q}$, we have

$$\begin{aligned} F(p) \leqslant G(p') = \lim_n F_{nn}(p') &\leqslant \liminf_n F_{nn}(x) \\ &\leqslant \limsup_n F_{nn}(x) \leqslant \lim_n F_{nn}(q') = G(q') \leqslant F(q). \end{aligned}$$

Thus, for all $p < a < q$ with $p, q$ in $\mathbb{Q}$, we have

$$F(p) \leqslant \liminf_n F_{nn}(x) \leqslant \limsup_n F_{nn}(x) \leqslant F(q).$$

Let $p \uparrow x$ and $q \downarrow x$ with $p, q$ in $\mathbb{Q}$, since $x$ is a continuous point of $F$, we obtain $F_{nn}(x) \to F(x)$. This completes the proof. □

**5.5. Tightness and Prohorov's Theorem.** The subsequential limit of sequence of distribution functions (i.e., probability measures) may nolonger be a distribution function (i.e., probability measure). If we require the subsequential limit to be a probability measure, we need to impose the *tightness* on the sequence of probability measures.

DEFINITION 3.46 (Tightness). A family of probability measures $\{\mu_\alpha : \alpha \in \mathcal{A}\}$ is **tight** if, for any $\varepsilon > 0$, there exists a finite interval $I$ such that

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

THEOREM 3.47 (Prohorov's theorem). *Let $\{\mu_\alpha : \alpha \in \mathcal{A}\}$ be a family of probability measures. In order that any sequence contains a subsequence which converges weakly to a probability measure, it is necessary and sufficient that the family is tight.*

PROOF. First consider the *sufficiency*. Theorem 3.44 asserts that any sequence $(\mu_n)$ contains a convergent subsequence $\mu_{n_k} \Rightarrow \mu$. It remains to show that $\mu(\mathbb{R}) = 1$. For any $\varepsilon > 0$, since the family is tight, there is a finite interval $I$ such that $\mu_{n_k}(I) \geqslant 1 - \varepsilon$. We can find two continuity points $a, b$ of $\mu$ such that $I \subset (a, b)$. Then we have

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

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

Next, consider the *necessity*, we prove it by contradiction. If the family is not tight, then there exists $\varepsilon_0 > 0$ such that for each interval $I_n = (-n, n)$, there exists $\mu_n$ in the family such that

$$\mu_n(I_n) \leqslant 1 - \varepsilon_0$$

for all $n$ in $\mathbb{N}$. Theorem 3.44 asserts that $(\mu_n)$ contains a convergent subsequence $\mu_{n_k} \Rightarrow \mu$. On the other hand, $\mu$ is a probability measure by hypothesis. Thus there exist continuity points $a, b$ of $\mu$ such that $\mu((a, b]) \geqslant 1 - \varepsilon_0/2$. Thus,

$$\lim_k \mu_{n_k}((a, b]) = \mu((a, b]) \geqslant 1 - \varepsilon_0/2.$$

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

$$\lim_k \mu_{n_k}((a, b]) \leqslant \liminf_k \mu_{n_k}(I_{n_k}) \leqslant 1 - \varepsilon_0,$$

which makes a contradiction. □

**5.6. Converging Together Lemma.** Convergence in distribution does not have the usual properties related with convergence. For instance, if $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$, it does not follow by any means that $X_n + Y_n \Rightarrow X + Y$. Nevertheless, the following simple situation still holds.

LEMMA 3.48 (Specific case). *Suppose $X_n$ converges to a constant $c$ in distribution, then $X_n \to c$ in probability.*

PROOF. The limiting distribution $\mu$ is Dirac, and its continuity points is $\mathbb{R}\backslash\{c\}$. In particular, $c-\varepsilon$ and $c+\varepsilon$ are continuity points for any $\varepsilon > 0$. Thus,

$$\mathbb{P}[|X_n - c| > \varepsilon] = \mathbb{P}[X_n < c-\varepsilon] + \mathbb{P}[X_n > c+\varepsilon] \to 0$$

as $n \to \infty$, which gives the convergence in probability. □

LEMMA 3.49 (Slutsky's theorem). *If $X_n \Rightarrow X$ and $Y_n \Rightarrow 0$, then*

(i) *$X_n + Y_n \Rightarrow X$, and*
(ii) *$X_n Y_n \Rightarrow 0$.*

*In general, if $X_n \Rightarrow X$, $\alpha_n \Rightarrow a$, $\beta_n \Rightarrow b$, where $a$ and $b$ are constants, then $\alpha_n X_n + \beta_n \Rightarrow aX + b$.*

PROOF. By Lemma 3.48, we know that $Y_n \to 0$ in probability.

(i). Suppose $\varphi \in \mathcal{C}_K(\mathbb{R},\mathbb{R})$, then $\varphi$ is uniformly continuous and is bounded: that is, $|\varphi| \leqslant M$ and for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|\varphi(x) - \varphi(y)| \leqslant \varepsilon$ as long as $|x-y| \leqslant \delta$. Thus

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

Let $n \to \infty$ and $\varepsilon \to 0$, we obtain $\mathbb{E}[|\varphi(X_n+Y_n) - \varphi(X_n)|] \to 0$, and $\mathbb{E}[\varphi(X_n+Y_n)] \to \mathbb{E}[\varphi(X_n)]$ as desired.

(ii). We choose $M$ large such that $\pm M$ are both continuity points of $X$. We have

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

Since $X_n \Rightarrow X$, then $\mathbb{P}[|X_n| > M] \to \mathbb{P}[|X| > M]$. Since $Y_n \to 0$ in probability, then $\mathbb{P}[|Y_n| > \varepsilon/M] \to 0$. So

$$\limsup_n \mathbb{P}[|X_nY_n| > \varepsilon] \leqslant \mathbb{P}[|X| > M].$$

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

In general, Combining $X_n \Rightarrow X$ and $\alpha_n - a \Rightarrow 0$ we have $(\alpha_n - a)X_n \Rightarrow 0$. Since $aX_n \Rightarrow aX$, we have $\alpha_n X_n \Rightarrow aX$. Now, since $\beta_n - b \Rightarrow 0$, then $\alpha X_n + \beta_n - b \Rightarrow aX$. This implies that $\alpha_n X_n + \beta_n \Rightarrow aX + b$. □

**5.7. Almost Sure Representations.** As almost sure convergence implies convergence in probability which implies convergence in distribution, the direction of implication is clear. In a sense, the following theorem gives the reverse direction.

THEOREM 3.50 (Skorokhod's representation theorem). *Suppose $X_n \Rightarrow X$. Then there is a probability space and random variables $(Y_n)$ and $Y$ defined on this space such that $Y_n \to Y$ almost surely. Moreover, the distributions of $Y_n$ and $Y$ are the same as those of $X_n$ and $X$, respectively.*

PROOF. *First, construct the probability space and $Y_n$ and $Y$.* For the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, take $\Omega = (0,1)$, let $\mathcal{F}$ consist of the Borel subsets of $(0,1)$, and for $\mathbb{P}(A)$ take the Lebesgue measure of $A$.

Consider the distribution functions $F_n$ and $F$ corresponding to $\mu_n$ and $\mu$. For $0 < w < 1$, put

$$Y_n(w) = \inf\{x : w \leqslant F_n(x)\}, \quad Y(w) = \inf\{x : w \leqslant F(x)\}.$$

We claim that $w \leqslant F_n(x)$ if and only if $Y_n(w) \leqslant x$. Indeed, since $F_n$ is increasing, $\{x : w \leqslant F_n(x)\}$ is an interval stretching to $\infty$; since $F_n$ is right continuous, this interval is closed on the left. For $0 < w < 1$, therefore, $\{x : w \leqslant F_n(x)\} = [Y_n(w), \infty)$, and so $Y_n(w) \leqslant x$ if and only if $w \leqslant F_n(x)$. Since $Y_n(w) \leqslant x$ if and only if $w \leqslant F_n(x)$, then

$$\mathbb{P}[\{w : Y_n(w) \leqslant x\}] = \mathbb{P}[\{w : w \leqslant F_n(x)\}] = F_n(x),$$

where the last equality is justified since $\mathbb{P}$ is the *Lebesgue measure* on $(0,1)$. Thus $Y_n$ has the same distribution function $F_n$; similarly, $Y$ has distribution function $F$.

*It remains to show that $Y_n(w) \to Y(w)$.* The idea is that $Y_n$ and $Y$ are *generalized inverse functions*[2] to $F_n$ and $F$; if the direct functions converge, so must the inverses.

Suppose that $0 < w < 1$. Given $\varepsilon$, choose $x$ such that $-\varepsilon < x < Y(w)$ and $\mu(\{x\}) = 0$. Then $F(x) < w$; $F_n(x) \to F(x)$ now implies that, for $n$ large enough, $F_n(x) < w$ and hence $Y(w) - \varepsilon < x < Y_n(w)$. Thus we have $\liminf_n Y_n(w) \geqslant Y(w)$. If $w < w'$ and $\varepsilon$ is positive, choose a $y$ for which $Y(w') < y < Y(w') + \varepsilon$ and $\mu(\{y\}) = 0$. Now

$$w < w' \leqslant F(Y(w')) \leqslant F(y), \tag{3.10}$$

where the second inequality will be verified later. So, for $n$ large enough, $w \leqslant F_n(y)$ and hence we have $Y_n(w) \leqslant y < Y(w') + \varepsilon$. Thus $\limsup_n Y_n(w) \leqslant Y(w')$ if $w < w'$, Hence, $Y_n(w) \to Y(w)$ if $Y$ is continuous at $w$.

Since $Y$ is increasing on $(0,1)$, it has at most countably many discontinuities. At discontinuity points $w$ of $Y$, redefine $Y_n(w) = Y(w) = 0$. With this change, $Y_n(w) \to Y(w)$ for every $w$. Since $Y$ and $Y_n$ have been altered only on a set of Lebesgue measure 0, their distributions are still $\mu_n$ and $\mu$.

*Finally, prove the second inequality of (3.10).* First, $Y(w') < y < \infty$ implies that $A = \{x : F(x) \geqslant w'\} \neq \varnothing$; thus, there exists $(x_n) \subset A$ with $x_n \downarrow \inf A = Y(w')$ for $n \to \infty$. By right-continuity of $F$, we have

$$w' \leqslant F(x_n) \downarrow F(Y(w')),$$

so $w' \leqslant F(Y(w'))$, which completes the proof.[3] □

**5.8. Relation to Uniform Integrability.** Skorokhod's presentation elevates convergence in distribution to the level of almost sure convergence in situations where the desired results concern only the distributions $\mu_n$ and $\mu$.

THEOREM 3.51 (Convergence in distribution and uniform integrability). *Suppose that $X_n \Rightarrow X$. Then the following are equivalent:*

(i) *$(X_n^p)$ is uniformly integrable.*
(ii) *The $X_n$ and $X$ are in $L^p$ and $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$.*

PROOF. Let $(Y_n)$ and $Y$ be as in the last Theorem. Since (i) and (ii) are in fact statements about the *marginal* distributions $\mu_n$ and $\mu$, it is sufficient to show that (i) and (ii) remain equivalent when the

[2] A useful handout about *generalized inverses* can be found here.

[3] The reference of this proof: Theorem 25.6 of Billingsley, Patrick. *Probability and measure*. Wiley, 2017.

$X_n$ and $X$ are replaced with the $Y_n$ and $Y$. But, then, the equivalence is immediate from Theorem 3.30. □

REMARK 3.52. It is worth noting the absence here of the first statement in Theorem 3.30, the one about the convergence of $(X_n)$ to $X$ in $L^p$. This is because convergence in $L^p$ concerns the sequence of joint distributions $\pi_n$ of the pairs $(X_n, X)$, and we have no guarantee that the *joint* distribution of $Y_n$ and $Y$ is $\pi_n$ for each $n$. †

**5.9. Fatou's Lemma.** The following theorem provides the weakest version of Fatou's lemma.

EXAMPLE 3.53 (Fatou's lemma). If $X_n \Rightarrow X$, then show that $\mathbb{E}[|X|] \leqslant \liminf_n \mathbb{E}[|X_n|]$.

PROOF. Using Skorokhod's representation theorem, there is a probability space and random variables $(Y_n)$ and $Y$ defined on this space such that $Y_n \to Y$ almost surely. Besides, we know that the law of $Y_n$ and $Y$ are equal to the law of $X_n$ and $X$, respectively, so $\mathbb{E}[|Y_n|] = \mathbb{E}[|X_n|]$, and $\mathbb{E}[|Y|] = \mathbb{E}[|X|]$. Using Fatou's lemma (almost sure convergence case),

$$\mathbb{E}[|X|] = \mathbb{E}[|Y|] = \mathbb{E}[\liminf_n |Y_n|] \leqslant \liminf_n \mathbb{E}[|Y_n|] = \liminf_n \mathbb{E}[|X_n|],$$

which gives the desired result. □

CHAPTER 4

# Law of Large Numbers and Central Limit Theorems

In order to pass the PhD Qualifying Exam in Financial Technology, you need to know **everything** in this chapter. This chapter we will introduce some classical theory of probability, that is,

- law of large numbers,
- convergence of random series,
- characteristic functions,
- central limit theorem.

We can use convergence of random series to prove the strong law of large numbers, and we also have other techniques to bypass the convergence of random series. We will introduce them both. Characteristic functions is a good tool to prove central limit theorem, but it has little use in more modern probability theory, like stochastic calculus theory. The books by Durrett and Chung Kai Lai are both excellent references, and we will make an effort to integrate these materials, making them more comprehensive and organized.

## 1. Weak Laws of Large Numbers

We will study some more meaningful random variables, such as $(X_1 + \cdots + X_n)/n$ in the following three sections, and the details we talk about mainly follow **Durrett**'s book. We may define $S_n := X_1 + \cdots + X_n$.

DEFINITION 4.1 (Weak law of large numbers (WLLN)). A sequence $(X_n)$ of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to obey the **weak law of large numbers** (WLLN) with normalizing sequences of real numbers $(a_n)$ and $(b_n)$ if

$$\frac{S_n - a_n}{b_n} \to 0 \text{ in probability} \quad \text{as} \quad n \to \infty.$$

DEFINITION 4.2 (Strong law of large numbers (SLLN)). A sequence $(X_n)$ of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to obey the **strong law of large numbers** (SLLN) with normalizing sequences of real numbers $(a_n)$ and $(b_n)$ if

$$\frac{S_n - a_n}{b_n} \to 0 \text{ almost surely} \quad \text{as} \quad n \to \infty.$$

### 1.1. $L^2$-Weak Law.

We recall the simplest weak law of large numbers we introduced in Chapter 3.

THEOREM 4.3 ($L^2$ weak law of large numbers). *Suppose $(X_n)$ are independent and identically distributed with $\mathbb{E}[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2 < \infty$. Set $S_n = \sum_{j=1}^n X_j$, then*

$$\frac{S_n}{n} \to \mu \quad \text{in probability}.$$

PROOF. Using Chebyshev's inequality, it is obvious that

$$\mathbb{P}\Big[\Big|\frac{S_n}{n}-\mu\Big|\geqslant\varepsilon\Big]\leqslant\frac{\mathbb{E}[|S_n/n-\mu|^2]}{\varepsilon^2}=\frac{\mathrm{Var}[S_n/n]}{\varepsilon^2}=\frac{\mathrm{Var}[S_n]}{n^2\varepsilon^2}$$
$$=\frac{1}{n^2\varepsilon^2}\sum_{j=1}^{n}\mathrm{Var}[X_j]=\frac{\sigma^2}{n\varepsilon^2}\to 0$$

for all $\varepsilon>0$ since $\sigma^2$ is finite. □

**1.2. Truncation.** Many limit theorems in probability concern triangular arrays $X_{n,k}$, $1\leqslant k\leqslant n$ of random variables and consider the limiting behavior of their row sums $S_n=X_{n,1}+\cdots+X_{n,n}$. To **truncate** a random variable $X$ at level $M$ means considering

$$\bar{X}=X\mathbb{1}_{\{|X|\leqslant M\}}.$$

To extend the weak law to random variables without a finite second moment, we truncate and then use Chebyshev's inequality.

We begin with a very general but also very useful result. Its proof is easy because we have assumed what we need for the proof.

THEOREM 4.4 (Kolmogorov-Feller: Weak law for triangular arrays). *For all $n$ let $X_{n,k}$, $1\leqslant k\leqslant n$, be independent. Let $b_n>0$ with $b_n\to\infty$, and $\bar{X}_{n,k}=X_{n,k}\mathbb{1}_{\{|X_{n,k}|\leqslant b_n\}}$. Suppose that as $n\to\infty$, we have*

(i) $\sum_{k=1}^{n}\mathbb{P}[|X_{n,k}|>b_n]\to 0$, *and*
(ii) $b_n^{-2}\sum_{k=1}^{n}\mathbb{E}[\bar{X}_{n,k}^2]\to 0$.

*If we let $S_n=X_{n,1}+\cdots+X_{n,n}$ and put $a_n=\sum_{k=1}^{n}\mathbb{E}[\bar{X}_{n,k}]$, then*

$$\frac{S_n-a_n}{b_n}\to 0\quad\textit{in probability.}$$

PROOF. Let $\bar{S}_n=\bar{X}_{n,1}+\cdots+\bar{X}_{n,n}$. Since $|(S_n-a_n)/b_n|\leqslant|(S_n-\bar{S}_n)/b_n|+|(\bar{S}_n-a_n)/b_n|$, then

$$\big\{|(S_n-a_n)/b_n|>\varepsilon\big\}\subset\big\{|(S_n-\bar{S}_n)/b_n|>0\big\}\cup\big\{|(\bar{S}_n-a_n)/b_n|>\varepsilon\big\}$$
$$\subset\big\{S_n\neq\bar{S}_n\big\}\cup\big\{|(\bar{S}_n-a_n)/b_n|>\varepsilon\big\}.$$

Which means that $\mathbb{P}[|(S_n-a_n)/b_n|>\varepsilon]\leqslant\mathbb{P}[S_n\neq\bar{S}_n]+\mathbb{P}[|(\bar{S}_n-a_n)/b_n|>\varepsilon]$. To estimate the first term, we note that if $S_n\neq\bar{S}_n$, there is some $1\leqslant k\leqslant n$ such that $\bar{X}_{n,k}\neq X_{n,k}$, so

$$\mathbb{P}[S_n\neq\bar{S}_n]\leqslant\mathbb{P}\Big[\bigcup_{k=1}^{n}\{\bar{X}_{n,k}\neq X_{n,k}\}\Big]\leqslant\sum_{k=1}^{n}\mathbb{P}[\bar{X}_{n,k}\neq X_{n,k}]=\sum_{k=1}^{n}\mathbb{P}[|X_{n,k}|>b_n]\to 0$$

by the hypothesis (i). For the second term, we note that $a_n=\mathbb{E}[\bar{S}_n]$ and $\mathrm{Var}[X]\leqslant\mathbb{E}[X^2]$ imply

$$\mathbb{P}\Big[\Big|\frac{\bar{S}_n-a_n}{b_n}\Big|>\varepsilon\Big]\leqslant\frac{1}{\varepsilon^2}\mathbb{E}\Big[\Big|\frac{\bar{S}_n-a_n}{b_n}\Big|^2\Big]=\frac{\mathrm{Var}[\bar{S}_n]}{\varepsilon^2b_n^2}$$
$$=\frac{1}{\varepsilon^2b_n^2}\sum_{k=1}^{n}\mathrm{Var}[\bar{X}_{n,k}]\leqslant\frac{1}{\varepsilon^2b_n^2}\sum_{k=1}^{n}\mathbb{E}[\bar{X}_{n,k}^2]\to 0$$

by the assumption (ii), and the proof is complete. □

THEOREM 4.5 (Weak law of large numbers). *Let $X_n$'s be independent and identically distributed with $x\mathbb{P}[|X_n|>x]\to 0$ as $x\to\infty$. Let $S_n=X_1+\cdots+X_n$ and let $\mu_n=\mathbb{E}[X_1\mathbb{1}_{\{|X_1|\leqslant n\}}]$. Then $S_n/n\to\mu_n$ in probability.*

PROOF. We will apply Theorem 4.4 with $X_{n,k} = X_k$ and $b_n = n$. To check (i), we not that

$$\sum_{k=1}^{n} \mathbb{P}[|X_{n,k}| > n] = n\mathbb{P}[|X_k| > n] \to 0$$

by assumption as $n \to \infty$. To check (ii), we will show $n^{-2} \cdot n\mathbb{E}[\bar{X}_{n,1}^2] \to 0$. Since $\bar{X}_{n,1} = X_1 \mathbb{1}_{\{|X_1| \leqslant n\}}$,

$$\mathbb{E}[\bar{X}_{n,1}^2] = \int_0^\infty 2y\mathbb{P}[|\bar{X}_{n,1}| > y]\,\mathrm{d}y = \int_0^n 2y\mathbb{P}[|\bar{X}_{n,1}| > y]\,\mathrm{d}y \leqslant \int_0^n 2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y.$$

The above equation is true since $\mathbb{P}[|\bar{X}_{n,1}| > y] = 0$ for all $y \geqslant n$ and $|\bar{X}_{n,1}| \leqslant |X_1|$ implies that $\mathbb{P}[|\bar{X}_{n,1}| > y] \leqslant \mathbb{P}[|\bar{X}_{n,1}| > y]$. We claim that $y\mathbb{P}[|X_1| > y] \to 0$ implies

$$n^{-1}\mathbb{E}[\bar{X}_{n,1}^2] \leqslant \frac{1}{n}\int_0^n 2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y \to 0 \quad \text{as } n \to \infty.$$

Indeed, we have $0 \leqslant g(y) := 2y\mathbb{P}[|X_1| > y] \leqslant 2y$ and $g(y) \to 0$ as $y \to \infty$, so $M = \sup_{y \geqslant 0} g(y) < \infty$. Let $g_n(y) = g(ny)$, then $g_n$ is bounded and $g_n(y) \to 0$ almost everywhere as $n \to \infty$. Now, using bounded convergence theorem, we have $n^{-1}\int_0^n g(y)\,\mathrm{d}y = \int_0^1 g_n(x)\,\mathrm{d}x \to 0$, which completes the proof. □

**1.3. $L^1$-Weak Law.** Finally, we have the weak law in its most familiar form.

THEOREM 4.6 (Khintchine: $L^1$ weak law of large numbers). *Let $X_n$'s be independent and identically distributed with $\mathbb{E}[|X_n|] < \infty$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu = \mathbb{E}[X_1]$. Then $S_n/n \to \mu$ in probability.*

PROOF. We would like to use Theorem 4.5 to prove this result. First, it should be noticed that $\mathbb{E}[|X_1|\mathbb{1}_{\{|X_1|>n\}}] \to 0$ as $n \to \infty$ since $X_1$ is integrable, so

$$x\mathbb{P}[|X_1| > x] \leqslant \mathbb{E}[|X_1|\mathbb{1}_{\{|X_1|>x\}}] \to 0 \quad \text{as } x \to \infty.$$

Besides, since $X_1\mathbb{1}_{\{|X_1|\leqslant n\}}$ is dominated by $|X_1|$, which is integrable, and $X_1\mathbb{1}_{\{|X_1|\leqslant n\}} \to X_1$, using dominated convergence theorem we get

$$\mu_n = \mathbb{E}[X_1\mathbb{1}_{\{|X_1|\leqslant n\}}] \to \mathbb{E}[X_1] = \mu \quad \text{as } n \to \infty.$$

Using Theorem 4.5, we see that if $\varepsilon > 0$ then $\mathbb{P}[|S_n/n - \mu_n| > \varepsilon/2] \to 0$. Since $\mu_n \to \mu$ (always in probability), it follows that $\mathbb{P}[|S_n/n - \mu| > \varepsilon] \to 0$. □

## 2. Strong Law of Large Numbers

**2.1. $L^4$-Strong Law.** As an easy corollary of the Borel-Cantelli lemma, we prove a version of the strong law of large numbers (SLLN) with a finite 4th moment assumption.

EXAMPLE 4.7 ($L^4$-strong law of large numbers). Suppose $(X_n)$ are independent and identically distributed with $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}[X_n^4] < \infty$. Set $S_n = \sum_{j=1}^n X_j$, then

$$\frac{S_n}{n} \to \mu \quad \text{almost surely.}$$

PROOF. We may assume $\mu = 0$ (otherwise, it is sufficient to consider $X_n - \mu$). First,

$$\mathbb{E}[S_n^4] = \mathbb{E}\Big[\Big(\sum_{1\leqslant i,j,k,l\leqslant n} X_iX_jX_kX_l\Big)^4\Big] = n\mathbb{E}[X_1^4] + 3(n^2-n)\mathbb{E}[X_1^2]^2 \leqslant Cn^2.$$

Using Chebyshev's inequality we get $\mathbb{P}[|S_n|/n > \varepsilon] \leqslant C/n^2\varepsilon^4$. Summing over $n$ is finite. Thus the convergence part of Borel-Cantelli lemma implies $\mathbb{P}[|S_n|/n > \varepsilon \text{ i.o.}] = 0$. Thus,

$$\mathbb{P}\Big[\bigcup_{k\geqslant 1}\bigcap_{m\geqslant 1}\bigcup_{n\geqslant m}\{|S_n|/n > 2^{-k}\}\Big] = 0,$$

which implies that $S_n/n \to 0$ almost surely. □

**2.2. Etemadi's Strong Law.** We are now ready to give Etemadi's (1981) proof of:

THEOREM 4.8 (Etemadi's strong law of large numbers). *Let $X_n$'s be pairwise independent identically distributed random variables with $\mathbb{E}[|X_i|] < \infty$. Let $\mathbb{E}[X_i] = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ almost surely as $n \to \infty$.*

PROOF. As in the proof of the weak law of large numbers, we begin by truncating.

LEMMA 4.9 (Truncating). *Let $Y_k = X_k \mathbb{1}_{\{|X_k|\leqslant k\}}$ and $T_n = Y_1 + \cdots + Y_n$. It is sufficient to prove that $T_n/n \to \mu$ almost surely.*

PROOF OF LEMMA 4.9. Since $\{|X_1| > x\} \supset \{|X_1| > y\}$ for all $y > x$, then

$$\begin{aligned}\infty > \mathbb{E}[|X_i|] &= \int_0^\infty \mathbb{P}[|X_i| > t]\,\mathrm{d}t = \sum_n \int_n^{n+1} \mathbb{P}[|X_1| > t]\,\mathrm{d}t \\ &\leqslant \sum_n \int_n^{n+1} \mathbb{P}[|X_1| > n]\,\mathrm{d}t = \sum_n \mathbb{P}[|X_1| > n] = \sum_n \mathbb{P}[|X_n| > n],\end{aligned}$$

which means that $\mathbb{P}[X_k \neq Y_k \text{ i.o.}] = 0$. This shows that $|S_n - T_n| \leqslant R$ almost surely for all $n$ in $\mathbb{N}$, then $n^{-1}|S_n - T_n| \leqslant n^{-1}R \to 0$ almost surely as $n \to \infty$, from which the desired result follows. □

The second step is not so intuitive, but it is an important part of this proof.

LEMMA 4.10 (Estimation 1). *Continuing with the previous notation, we have $\sum_k \mathrm{Var}[Y_k]/k^2 \leqslant \sum_k \mathbb{E}[Y_k^2]/k^2 \leqslant 4\mathbb{E}[|X_1|] < \infty$.*

PROOF OF LEMMA 4.10. To bound the sum, we observe

$$\mathrm{Var}[Y_k] \leqslant \mathbb{E}[Y_k^2] = \int_0^\infty 2y\mathbb{P}[|Y_k| > y]\,\mathrm{d}y = \int_0^k 2y\mathbb{P}[|Y_k| > y]\,\mathrm{d}y \leqslant \int_0^k 2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y, \tag{4.1}$$

where the second equality holds since $|Y_k| \leqslant k$; so is the last inequality since we have $|Y_k| \leqslant |X_1|$, which implies that $\{|Y_k| > y\} \subset \{|X_1| > y\}$. Using Beppo Levi's Theorem (topics in integration), we have

$$\begin{aligned}\sum_{k=1}^\infty \frac{1}{k^2}\mathbb{E}[Y_k^2] &\leqslant \sum_{k=1}^\infty \frac{1}{k^2}\int_0^k 2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y = \sum_{k=1}^\infty \frac{1}{k^2}\int_0^\infty \mathbb{1}_{[0,k]}(y)\,2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y \\ &= \int_0^\infty \Big(\sum_{k=1}^\infty \frac{1}{k^2}\mathbb{1}_{[0,k]}(y)\Big)\,2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y = \int_0^\infty \Big(\sum_{k>y}^\infty \frac{1}{k^2}\Big)\,2y\mathbb{P}[|X_1| > y]\,\mathrm{d}y,\end{aligned}$$

where we use Equation (4.1) in the first inequality and Beppo Levi's Theorem in the second equality. Notice that $\mathbb{E}[|X_1|] = \int_0^\infty \mathbb{P}[|X_1| > y]\,\mathrm{d}y$, we can complete the proof by showing:

LEMMA 4.11 (Estimation 2). *If $y \geqslant 0$, then $2y\sum_{k>y} k^{-2} \leqslant 4$.*

PROOF OF LEMMA 4.11. We begin with the observation that if $m \geqslant 2$, then

$$\sum_{k\geqslant m} \frac{1}{k^2} = \sum_{k\geqslant m} \int_{k-1}^{k} \frac{1}{k^2}\,\mathrm{d}x \leqslant \sum_{k\geqslant m} \int_{k-1}^{k} \frac{1}{x^2}\,\mathrm{d}x = \int_{m-1}^{\infty} x^{-2}\,\mathrm{d}x = \frac{1}{m-1}. \tag{4.2}$$

When $y \geqslant 1$, then sum starts with $k = \lfloor y\rfloor + 1 \geqslant 2$, where $\lfloor y\rfloor$ denotes the largest integer that is less than or equal to $y$, so

$$2y \sum_{k>y} \frac{1}{k^2} = 2y \sum_{k=\lfloor y\rfloor+1}^{\infty} \frac{1}{k^2} \leqslant 2y \frac{1}{(\lfloor y\rfloor+1)-1} = \frac{2y}{\lfloor y\rfloor} \leqslant 4$$

since $y/\lfloor y\rfloor \leqslant 2$ for $y \geqslant 1$ (Using GeoGebra to draw a picture). To cover $0 \leqslant y < 1$, we note that if $y = 0$, then $2y\sum_{k>y} k^{-2} = 0$, and if $0 < y < 1$, $\sum_{k>y} k^{-2} = 1^{-2} + \sum_{k=2}^{\infty} k^{-2} \leqslant 1 + (2-1)^{-1}$ using Equation (4.2), so

$$2y \sum_{k>y} k^{-2} \leqslant 2\Big(1 + \sum_{k=2}^{\infty} k^{-2}\Big) \leqslant 4.$$

This establishes Lemma 4.11 □

Thus we completes the proof of Lemma 4.10. □

CONTINUE WITH THE PROOF OF THEOREM 4.8. Etemadi's inspiration was that since $X_n^+$ and $X_n^-$, $n \geqslant 1$, satisfy the assumptions of the theorem and $X_n = X_n^+ - X_n^-$, we can without loss of generality suppose $X_n \geqslant 0$.

We will prove the result first for a subsequence and then use monotonicity to control the values in between. We let $\alpha > 1$ and $k(n) = \lfloor \alpha^n\rfloor$. Chebyshev's inequality implies that if $\varepsilon > 0$,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\big[|T_{k(n)} - \mathbb{E}[T_{k(n)}]| > \varepsilon k(n)\big] &\leqslant \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{\operatorname{Var}[T_{k(n)}]}{k(n)^2} \\ &= \frac{1}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{k(n)^2} \sum_{m=1}^{k(n)} \operatorname{Var}[Y_m] = \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \Big(\operatorname{Var}[Y_m] \sum_{n:k(n)\geqslant m} \frac{1}{k(n)^2}\Big).\end{aligned} \tag{4.3}$$

It is obvious that $\sum_{n:k(n)\geqslant m} \frac{1}{k(n)^2} \leqslant \sum_{n=1}^{\infty} n^{-2} = C < \infty$, so the left side of Equation (4.3) is finite. Using Borel-Cantelli lemma and the arbitrary of $\varepsilon$, we know

$$\frac{T_{k(n)} - \mathbb{E}[T_{k(n)}]}{k(n)} \to 0 \quad \text{almost surely.}$$

Besides, $\mathbb{E}[Y_k] = \mathbb{E}[X_1 \mathbb{1}_{\{|X_1|\leqslant k\}}] \to \mathbb{E}[X_1]$ as $k \to \infty$ using the dominated convergence theorem, so we get $\mathbb{E}[T_{k(n)}]/k(n) \to \mathbb{E}[X_1]$ and we have shown that $T_{k(n)}/k(n) \to \mathbb{E}[X_1]$ almost surely.

To handle the intermediate values, we observe that if $k(n) \leqslant m < k(n+1)$, we have

$$\frac{k(n)}{k(n+1)} \frac{T_{k(n)}}{k(n)} = \frac{T_{k(n)}}{k(n+1)} \leqslant \frac{T_{k(n)}}{m} \leqslant \frac{T_m}{m} \leqslant \frac{T_{k(n+1)}}{m} \leqslant \frac{T_{k(n+1)}}{k(n)} = \frac{k(n+1)}{k(n)} \frac{T_{k(n+1)}}{k(n+1)},$$

here we use $Y_i \geqslant 0$ for all $i$ in $\mathbb{N}$. Recalling that $k(n) = \lfloor \alpha^n\rfloor$, we have $k(n+1)/k(n) \to \alpha$ and

$$\frac{1}{\alpha}\mathbb{E}[X_1] \leqslant \liminf_n \frac{T_m}{m} \leqslant \limsup_n \frac{T_m}{m} \leqslant \alpha \mathbb{E}[X_1].$$

Since $\alpha > 1$ is arbitrary, let $\alpha \downarrow 1$ and we completes the proof of Theorem 4.8. □

The next result shows that the strong law holds whenever $\mathbb{E}[X_i]$ exists.

THEOREM 4.12 (Necessary condition of strong law). *Let $X_n$'s be independent identically distributed random variables with $\mathbb{E}[X_i^+] = \infty$, $\mathbb{E}[X_i^-] < \infty$. If $S_n = X_1 + \cdots + X_n$, then $S_n/n \to \infty$ almost surely.*

PROOF. Let $M > 0$ and $X_i^M = X_i \wedge M$. The $X_i^M$'s are independent identically distributed with $\mathbb{E}[|X_i^M|] < \infty$. So if $S_n^M = X_1^M + \cdots + X_n^M$, then Theorem 4.8 implies that $S_n^M/n \to \mathbb{E}[X_i^M]$ almost surely. Since $X_i \geqslant X_i^M$, it follows that

$$\liminf_n \frac{S_n}{n} \geqslant \lim_n \frac{S_n^M}{n} = \mathbb{E}[X_i^M].$$

Besides, the monotone convergence theorem implies that $\mathbb{E}[(X_i^M)^+] \uparrow \mathbb{E}[X_i^+] = \infty$ as $M \to \infty$, so we have $\mathbb{E}[X_i^M] = \mathbb{E}[(X_i^M)^+] - \mathbb{E}[(X_i^M)^-] \uparrow \infty$, and we have

$$\liminf_n \frac{S_n}{n} \geqslant \infty$$

using the above result, which implies the desired result. □

Using the strong law of large numbers, we can prove the following fundamental theorem in mathematical statistics:

DEFINITION 4.13 (Empirical distribution functions). Let $X_n$'s be independent identically distributed with distribution $F$ and let

$$F_n(x) = \frac{1}{n} \sum_{m=1}^{n} \mathbb{1}_{\{X_m \leqslant x\}}.$$

That is, $F_n(x)$ is the observed frequency of values that are less or equal to $x$.

The next result shows that $F_n$ converges uniformly to $F$ as $n \to \infty$.

THEOREM 4.14 (The Glivenko-Cantelli theorem). *As $n \to \infty$, we have*

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \textit{almost surely}.$$

PROOF. ix $x$ and let $Y_n = \mathbb{1}_{\{X_n \leqslant x\}}$. Since the $Y_n$'s are independent identically distributed with

$$\mathbb{E}[Y_n] = \mathbb{P}[X_n \leqslant x] = F(x),$$

the Etemadi's strong law of large numbers implies that $F_n(x) = n^{-1} \sum_{m=1}^n Y_m \to F(x)$ almost surely. Again, fix $x$ and let $Z_n = \mathbb{1}_{\{X_n < x\}}$. Since the $Z_n$'s are independent identically distributed with

$$\mathbb{E}[Z_n] = \mathbb{P}[X_n < x] = F(x-) = \lim_{y \uparrow x} F(y),$$

the Etemadi's strong law of large numbers implies that $F_n(x-) = n^{-1} \sum_{m=1}^n Z_m \to F(x-)$ almost surely. For $1 \leqslant j \leqslant k-1$ let $x_{j,k} = \inf\{y : F(y) \geqslant j/k\}$. The pointwise convergence of $F_n(x)$ and $F_n(x-)$ imply that we can pick $N_k(\omega)$ such that if $n \geqslant N_k(\omega)$, then

$$\big|F_n(x_{j,k}) - F(x_{j,k})\big| < k^{-1} \quad \text{and} \quad \big|F_n(x_{j,k}-) - F(x_{j,k}-)\big| < k^{-1}$$

for $1 \leqslant j \leqslant k-1$. If we let $x_{0,k} = -\infty$ and $x_{k,k} = \infty$, then the last two inequalities hold for $j = 0$ or $k$. If $x$ in $(x_{j-1,k}, x_{j,k})$ with $1 \leqslant j \leqslant k$ and $n \geqslant N_k(\omega)$, then using the monotonicity of $F_n$ and $F$, and the fact that $F(x_{j,k}-) - F(x_{j-1,k}) \leqslant k^{-1}$, we have

$$F_n(x) \leqslant F_n(x_{j,k}-) \leqslant F(x_{j,k}-) + k^{-1} \leqslant F(x_{j-1,k}) + 2k^{-1} \leqslant F(x) + 2k^{-1},$$
$$F_n(x) \geqslant F_n(x_{j-1,k}) \geqslant F(x_{j-1,k}) - k^{-1} \geqslant F(x_{j,k}-) - 2k^{-1} \geqslant F(x) - 2k^{-1}.$$

Thus $\sup_x |F_n(x) - F(x)| \leqslant 2k^{-1}$, and we have proved the result. □

## 3. Convergence of Random Series

**3.1. Inequalities for Maxima.** The maximal inequality and the following series theorems mainly refer **Erhan**'s book, the remain details all refer **Durrett**'s book.

In this section, we will pursue a second approach to the strong law of large numbers based on the convergence of random series. We will consider the almost sure convergence of the sequence $(S_n)$. All results are for the case where the $X_n$'s are independent, in which case Kolmogorov's zero-one law applies, and the convergence of the series has probability 0 or 1, the better case being our aim.

Suppose that the $X_n$ has mean 0, then Chebyshev's inequality yields

$$\varepsilon^2 \mathbb{P}[|S_n| > \varepsilon] \leqslant \mathrm{Var}[S_n] = \mathbb{E}[S_n^2].$$

The following is a considerable improvement when the $X_n$ are independent.

LEMMA 4.15 (Kolmogorov's maximal inequality - Upper bound). *Suppose that the $X_n$'s are independent and have mean 0. Then for every $a > 0$,*

$$a^2 \mathbb{P}[\max_{k \leqslant n} |S_k| > a] \leqslant \mathrm{Var}[S_n].$$

PROOF. Fix $a > 0$ and $n \geqslant 1$. Define $N(\omega) = \inf\{k \geqslant 1 : |S_k(\omega)| \geqslant a\}$ for every $\omega$ in $\Omega$. Note that $N(\omega) = k$ if and only if $|S_k(\omega)| > a$ and $|S_j(\omega)| \leqslant a$ for all $j < k$. Thus $\mathbb{1}_{\{N=k\}}$ is a function of $(X_1, \cdots, X_k)$, which shows that $N$ is a random variable. Moreover, by the same reason, for $k < n$, $U = S_k \mathbb{1}_{\{N=k\}}$ and $V = S_n - S_k$ are functions of independent vectors $(X_1, \cdots, X_k)$ and $(X_{k+1}, \cdots, X_n)$, and thus $\mathbb{E}[UV] = \mathbb{E}[U]\mathbb{E}[V]$; and $\mathbb{E}[V] = 0$ since $E[X_i] = 0$ for all $i$ by hypothesis. Hence, for $k \leqslant n$, we have

$$\mathbb{E}[S_k(S_n - S_k)\mathbb{1}_{\{N=k\}}] = 0. \tag{4.4}$$

Note that $S_n^2 = [S_k + (S_n - S_k)]^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 \geqslant S_k^2 + 2S_k(S_n - S_k)$, and that $|S_k|^2 > a^2$ on the event $\{N = k\}$. Thus

$$\mathbb{E}[S_n^2 \mathbb{1}_{\{N=k\}}] \geqslant a^2 \mathbb{E}[\mathbb{1}_{\{N=k\}}] + 2\mathbb{E}[S_k(S_n - S_k)\mathbb{1}_{\{N=k\}}] = a^2 \mathbb{P}[N = k],$$

in the view of Equation (4.4). Summing both sides over $k \leqslant n$ and reversing the order, we get

$$a^2 \mathbb{P}[N \leqslant n] \leqslant \mathbb{E}[S_n^2 \mathbb{1}_{\{N \leqslant n\}}] \leqslant \mathbb{E}[S_n^2] = \mathrm{Var}[S_n],$$

which completes the proof upon noting that the event $\{N \leqslant n\}$ is the same as the event that $\{\max_{k \leqslant n} |S_k| > a\}$. □

The assumption of independence for the $X_n$ will be relaxed later by martingaling. For the present, the following is an estimate going in the opposite direction.

LEMMA 4.16 (Kolmogorov's maximal inequality - Lower bound). *Suppose that the $X_n$'s are independent and have mean 0, and are dominated by some constant $b$. Then, for every $a > 0$,*

$$\mathbb{P}[\max_{k \leqslant n} |S_k| > a] \geqslant 1 - \frac{(a+b)^2}{\mathrm{Var}[S_n]}.$$

PROOF. Fix $n$ and $a$ and let $N$ be as in the preceding proof. Now we claim that

$$\mathbb{P}[N > n]\mathrm{Var}[S_n] \leqslant (a+b)^2. \tag{4.5}$$

Fix $k \leqslant n$ and note that $|S_k(\omega)| \leqslant a + b$ if $N(\omega) = k$, because $|S_{k-1}(\omega)| \leqslant a$ by the definition of $N(\omega)$ and $|X_k(\omega)| \leqslant b$ by the assumed boundedness. Besides, since

$$S_n^2 = [S_k + (S_n - S_k)]^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2,$$

then

$$S_n^2 \mathbb{1}_{\{N=k\}} \leqslant (a+b)^2 \mathbb{1}_{\{N=k\}} + 2S_k(S_n - S_k)\mathbb{1}_{\{N=k\}} + (S_n - S_k)^2 \mathbb{1}_{\{N=k\}}. \tag{4.6}$$

On the right side, the expectation of the second term is 0 by Equation (4.4), and the reasoning leading to (4.4) shows that the expectation of the third term is

$$\mathbb{E}[(S_n - S_k)^2 \mathbb{1}_{\{N=k\}}] = \mathbb{P}[N = k]\mathbb{E}[(S_n - S_k)^2] \leqslant \mathbb{P}[N = k]\mathrm{Var}[S_n],$$

where the last inequality holds since

$$\begin{aligned}\mathbb{E}[(S_n - S_k)^2] &= \mathbb{E}\big[(S_n - S_k)^2 - (\mathbb{E}[S_n - S_k])^2\big] = \mathrm{Var}[S_n - S_k] = \mathrm{Var}[X_{k+1} + \cdots + X_n] \\ &\leqslant \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[S_n].\end{aligned}$$

Hence, taking expectations on both sides of (4.6) and adding over $k \leqslant n$ we get

$$\mathbb{E}[S_n^2 \mathbb{1}_{\{N \leqslant n\}}] \leqslant \big[(a+b)^2 + \mathrm{Var}[S_n]\big]\,\mathbb{P}[N \leqslant n].$$

On the other hand, for every $\omega$, if $N(\omega) > n$ then $|S_n(\omega)| \leqslant a$. So

$$\mathbb{E}[S_n^2 \mathbb{1}_{\{N>n\}}] \leqslant \mathbb{E}[a^2 \mathbb{1}_{\{N>n\}}] = a^2 \mathbb{P}[N > n].$$

Adding the last two expressions side by side we get an upper bound for $\mathrm{Var}[S_n] = \mathbb{E}[S_n^2]$; and rearranging the terms somewhat we obtain (4.5). □

**3.2. Convergence of Series and Variances.** Indeed, the summability of variances implies the convergence of the associated series:

THEOREM 4.17 (Kolmogorov's one series theorem). *Suppose that the $X_n$'s are independent and have zero mean. If $\sum \mathrm{Var}[X_n]$ converges then $\sum X_n$ converges almost surely.*

PROOF. By Kolmogorov's inequality applied to the sequence $(X_{n+m})_{m \geqslant 1}$, for all $\varepsilon > 0$,

$$\varepsilon^2 \mathbb{P}[\max_{k \leqslant m} |S_{n+k} - S_n| > \varepsilon] \leqslant \sum_{j=n+1}^{n+m} \mathrm{Var}[X_j].$$

Since $\sum \mathrm{Var}[X_n] < \infty$, then, the right side goes to 0 as we let $m \to \infty$ first and $n \to \infty$ next. So the condition of Example 3.24 is satisfied, and $(S_n)$ converges almost surely. □

COROLLARY 4.18 (Kolmogorov's two series theorem). *Let $\mu_n = \mathbb{E}[X_n]$ be the means and $\sigma_n^2 = \mathrm{Var}[X_n]$ the variance of a sequence of independent random variables $(X_n)$. If $\sum \mu_n$ and $\sum \sigma_n^2$ converges then $\sum X_n$ converges almost surely.*

PROOF. The result is immediate if we define $Y_n = X_n - \mu_n$ and apply Theorem 4.17 to $Y_n$. □

The following is nearly a converse to the preceding theorem; within the proof, the most interesting trick in the second step is called *symmetrization*.

THEOREM 4.19 (Converse proposition of one series theorem). *Suppose that $(X_n)$ is a bounded sequence of independent variables. If $\sum(X_n - a_n)$ is almost surely convergent for some sequence $(a_n)$ in $\mathbb{R}$, then $\sum \mathrm{Var}[X_n]$ converges.*

PROOF. (i). First, we prove the assertion under the extra conditions that $a_n = 0$ and $\mathbb{E}[X_n] = 0$ for all $n$. Let $b$ be a bound for $(X_m)$. Note that

$$Z_m = \sup_k |S_{m+k} - S_m| = \lim_n \underbrace{\max_{k\leqslant n} |S_{m+k} - S_m|}_{\text{increasing}},$$

Thus, for all $\varepsilon > 0$, using the continuity of probability measure, we have

$$\mathbb{P}[Z_m > \varepsilon] = \lim_n \mathbb{P}[\max_{k\leqslant n} |S_{m+k} - S_m| > \varepsilon] \geqslant 1 - \frac{(\varepsilon + b)^2}{\sum_{i=m+1}^{\infty} \operatorname{Var}[X_i]}, \tag{4.7}$$

where we used Lemma 4.16 applied to the sequence $(X_{m+n})_{n\geqslant 1}$. If $(S_n)$ converges almost surely, then $Z_m \to 0$ almost surely by Lemma 3.23, and thus the left side of Equation (4.7) tends to 0 as $m \to \infty$. This is impossible if $\sum \operatorname{Var}[X_i] = \infty$.

(ii). Next we remove the extra conditions. Let $(Y_n)$ be independent of $(X_n)$ and have the same law. Suppose that $\sum(X_n - a_n)$ is almost surely convergent. Then, so is $\sum(Y_n - a_n)$ since the sequences $(X_n)$ and $(Y_n)$ have the same law. Thus,

$$\sum(X_n - Y_n) = \sum(X_n - a_n) - \sum(Y_n - a_n)$$

converges almost surely and the sequence $(X_n - Y_n)_{n\geqslant 1}$ is bounded and $\mathbb{E}[X_n - Y_n] = 0$ for all $n$ in $\mathbb{N}$. Hence, part (i) of the proof applies, and we must have $\sum \operatorname{Var}[X_n - Y_n] < \infty$. This finishes the proof since $\operatorname{Var}[X_n - Y_n] = 2\operatorname{Var}[X_n]$. □

REMARK 4.20. In general, random variables may not have finite means or variances. Suppose $(X_n)$ is any sequence of random variables we can take a truncation value $b$ and define $Y_n = X_n \mathbb{1}_{\{|X_n|\leqslant b\}}$. The $(Y_n)$ are independent and dominated by $b$. Corollary 4.18 can be applied to $(Y_n)$ and if we impose the additional condition that $\sum \mathbb{P}[X_n \neq Y_n] = \sum \mathbb{P}[|X_n| > b] < \infty$. By using Borel-Cantelli lemma, with probability 1, $X_n = Y_n$ for all sufficiently large $n$. The convergence of $\sum X_n$ and $\sum Y_n$ are therefore equivalent, we have shown the proof in the proof of Lemma 4.9. Then we have the *three series theorem*. †

Using the remark above, the following theorem gives necessary and sufficient conditions for the almost sure convergence of the series $X_n$.

THEOREM 4.21 (Kolmogorov's three series theorem). *Suppose that the $X_n$'s are independent, and define $Y_n = X_n \mathbb{1}_{\{|X_n|\leqslant b\}}$ for a fixed constant $b$ in $\mathbb{R}_+$. Then $\sum X_n$ is almost sure convergent if and only if the following three series are convergent:*

$$\sum \mathbb{P}[X_n \neq Y_n], \quad \sum \mathbb{E}[Y_n], \quad \sum \operatorname{Var}[Y_n]. \tag{4.8}$$

PROOF. *Sufficiency*. Suppose that all three series in (4.8) are convergent. The independence of the $X_n$ imply the independence of the $Y_n$. Using the two series theorem, the convergence of the third series implies that $\sum(Y_n - \mathbb{E}[Y_n])$ convergence almost surely. Besides, since $\sum \mathbb{E}[X_n]$ is convergent, so $\sum Y_n$ converges almost surely, which implies that $\sum X_n$ converges almost surely.

*Necessity*. Suppose that $\sum X_n$ converges almost surely. Then, for almost every $\omega$, there are at most finitely many $n$ with $|X_n(\omega)| > b$, which means that $X_n(\omega) \neq Y_n(\omega)$ for only finitely many $n$. Thus, $\sum \mathbb{1}_{\{X_n \neq Y_n\}} < \infty$ almost surely, and the independence of $X_n$ implies that the events $\{X_n \neq Y_n\}$ are independent. Using the divergence part of Borel-Cantelli lemma, the first series of (4.8) must converges.

Consequently, $\sum Y_n$ is almost surely convergent since $\sum X_n$ is so. Now, Theorem 4.19 implies that the third series in (4.8) converges. This in turn implies via Corollary 4.18 that $\sum(Y_n - \mathbb{E}[Y_n])$

converges almost surely, which together with the convergence of $\sum Y_n$ implies that the second series in (4.8) is convergent. □

**3.3. Kronecker's Lemma.** In order to introduce the applications of these series theorems, we first introduce the following lemma:

LEMMA 4.22 (Kronecker's lemma). *Suppose that $a_n \uparrow \infty$ and $\sum x_n/a_n$ converges, then $a_n^{-1}\sum_{m=1}^n x_m \to 0$ as $n \to \infty$.*

PROOF. Let $a_0 = 0$ and $b_0 = 0$, and for $m \geqslant 1$, define $b_m = \sum_{k=1}^m x_k/a_k$. So

$$\begin{aligned} a_n^{-1}\sum_{m=1}^n x_m &= a_n^{-1}\Big[\sum_{m=1}^n a_m b_m - \sum_{m=1}^n a_m b_{m-1}\Big] = a_n^{-1}\Big[a_n b_n + \sum_{m=2}^n a_{m-1}b_{m-1} - \sum_{m=1}^n a_m b_{m-1}\Big] \\ &= b_n - \sum_{m=1}^n \frac{a_m - a_{m-1}}{a_n} b_{m-1}. \end{aligned}$$

(Recall that $a_0 = 0$). By hypothesis, $b_n \to b_\infty$ as $n \to \infty$. Since $a_m - a_{m-1} \geqslant 0$, the last sum is an average of $b_0, \cdots, b_n$. Let $B = \sup|b_n|$ and pick $M$ such that $|b_m - b_\infty| < \varepsilon/2$ for $m \geqslant M$, then pick $N$ such that $a_M/a_n < \varepsilon/(4B)$ for $n \geqslant N$. Now. if $n \geqslant N \vee M$, we have

$$\begin{aligned} \Big|\sum_{m=1}^n \frac{a_m - a_{m-1}}{a_n} b_{m-1} - b_\infty\Big| &\leqslant m = 1^n \frac{a_m - a_{m-1}}{a_n}|b_{m-1} - b_\infty| \\ &\leqslant \sum_{m=1}^M \frac{a_m - a_{m-1}}{a_n}\big[|b_{m-1}| + |b_\infty|\big] + \sum_{M+1}^n \frac{a_m - a_{m-1}}{a_n}|b_{m-1} - b_\infty| \\ &\leqslant 2B\sum_{m=1}^M \frac{a_m - a_{m-1}}{a_n} + \frac{\varepsilon}{2}\sum_{M+1}^n \frac{a_m - a_{m-1}}{a_n} = 2B\frac{a_M}{a_n} + \frac{\varepsilon}{2}\frac{a_n - a_M}{a_n} < \varepsilon, \end{aligned}$$

which proving the desired result since $\varepsilon$ is arbitrary. □

A fact that is frequently used is that averages of convergent sequences converge. We will give a detailed proof, whose methods are similar to the proof of Lemma 4.22.

LEMMA 4.23 (Convergence of (weighted) averages). *Suppose that $a_n$ in $\mathbb{R}$ for all $n$ in $\mathbb{N}$. If $a_n \to a$ as $n \to \infty$, then $n^{-1}\sum_{m=1}^n a_m \to a$ as $n \to \infty$. If, in addition, $w_m \geqslant 0$ for all $k$ in $\mathbb{N}$, and $B_n = \sum_{m=1}^n w_m$ for all $n \geqslant 1$, with $B_n \uparrow \infty$ as $n \to \infty$, then*

$$\frac{1}{B_n}\sum_{m=1}^n w_m a_m \to a \quad \text{as } n \to \infty.$$

PROOF. Without loss of generality we may assume that $a = 0$ (consider $a_n - a$, otherwise). Thus, for all $\varepsilon > 0$, we know that $|a_n| < \varepsilon$ as soon as $n > n_0 = n_0(\varepsilon)$. It follows that, for $n > n_0$,

$$\Big|\frac{1}{n}\sum_{m=1}^n a_m\Big| \leqslant \Big|\frac{1}{n}\sum_{m=1}^{n_0} a_m\Big| + \frac{n - n_0}{n}\Big|\frac{1}{n - n_0}\sum_{n_0+1}^n a_m\Big| \leqslant \frac{1}{n}\sum_{m=1}^{n_0}|a_m| + \varepsilon,$$

letting $n \to \infty$ and we have $\limsup_n |n^{-1}\sum_{m=1}^n a_m| \leqslant \varepsilon$ since $\sum_{m=1}^{n_0}|a_m|$ is a constant, and we complete the proof of the first part. The second one follows similarly, we can just default this result. □

**3.4. Application: Strong Laws.** The first application is:

THEOREM 4.24 (Strong law of large numbers). *Let $X_n$'s be independent identically distributed random variables with $\mathbb{E}[|X_i|] < \infty$. Let $\mathbb{E}[X_i] = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ almost surely as $n \to \infty$.*

PROOF. Let $Y_k = X_k \mathbb{1}_{\{|X_k| \leqslant k\}}$ and $T_n = Y_1 + \cdots + Y_n$. Using Lemma 4.9 again it suffices to show that $T_n/n \to \mu$ almost surely. Let $Z_k = Y_k - \mathbb{E}[Y_k]$, so $\mathbb{E}[Z_k] = 0$ and $\operatorname{Var}[Z_k] = \operatorname{Var}[Y_k] \leqslant \mathbb{E}[Y_k^2]$ and Lemma 4.10 implies that

$$\sum_{k=1}^{\infty} \frac{\operatorname{Var}[Z_k]}{k^2} = \sum_{k=1}^{\infty} \frac{\operatorname{Var}[Y_k]}{k^2} \leqslant \sum_{k=1}^{\infty} \frac{\mathbb{E}[Y_k^2]}{k^2} < \infty.$$

Applying the one series theorem 4.17 now, we conclude that $\sum_k Z_k/k$ converges almost surely. So Lemma 4.22 implies

$$\frac{1}{n}\sum_{k=1}^{n} \big(Y_k - \mathbb{E}[Y_k]\big) \to 0 \quad \text{and hence} \quad \frac{T_n}{n} - \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Y_k] \to 0 \text{ almost surely.}$$

The dominated convergence theorem implies $\mathbb{E}[Y_k] \to \mu$ as $k \to \infty$. From this, it follows easily that $n^{-1}\sum_{k=1}^{n} \mathbb{E}[Y_k] \to \mu$, and hence $T_n/n \to \mu$ almost surely. □

**3.5. Application: Rates of Convergence.** As mentioned earlier, one of the advantages of the random series proof is that it provides estimates on the rate of convergence of $S_n/n \to \mu$. By subtracting $\mu$ from each random variable, we can and will suppose without loss of generality that $\mu = 0$.

THEOREM 4.25 (Rates of convergence - Finite variance). *Let $X_n$'s be independent identically distributed random variables with zero mean and finite variance. Let $S_n = X_1 + \cdots + X_n$. If $\varepsilon > 0$ then*

$$\frac{S_n}{n^{1/2}(\log n)^{1/2+\varepsilon}} \to 0 \quad \textit{almost surely.}$$

PROOF. Let $a_n = n^{1/2}(\log n)^{1/2+\varepsilon}$ for $n \geqslant 2$ and $a_1 > 0$.

$$\sum_{n=1}^{\infty} \operatorname{Var}\Big[\frac{X_n}{a_n}\Big] = \sum_{n=1}^{\infty} \frac{1}{a_n^2}\operatorname{Var}[X_n] = \sigma^2\Big(\frac{1}{a_1^2} + \sum_{n=2}^{\infty} \frac{1}{n^{1/2}(\log n)^{1/2+\varepsilon}}\Big) \leqslant \sigma^2\Big(\frac{1}{a_1^2} + \sum_{n=2}^{\infty} \frac{1}{n^{2(1+\varepsilon)}}\Big) < \infty,$$

so using the one series theorem 4.17, we get $\sum X_n/a_n$ converges almost surely and the indicated result follows from Lemma 4.22. □

The next result treats the situation in which $\mathbb{E}[X_i^2] = \infty$ but $\mathbb{E}[|X_i|^p] < \infty$ for some $1 < p < 2$.

THEOREM 4.26 (Marcinkiewicz & Zygmund: Rates of convergence - Infinite variance). *Let $X_n$'s be independent identically distributed with $\mathbb{E}[X_1] = 0$, and $\mathbb{E}[|X_1|^p] < \infty$, $1 < p < 2$. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n}{n^{1/p}} \to 0 \quad \textit{almost surely.}$$

PROOF. Let $Y_k = X_k \mathbb{1}_{\{|X_k| \leqslant k^{1/p}\}}$ and $T_n = Y_1 + \cdots + Y_n$. Then

$$\sum \mathbb{P}[X_k \neq Y_k] = \sum \mathbb{P}[|X_k|^p > k] \leqslant \mathbb{E}[|X_k|^p] < \infty,$$

where the first inequality holds by employing techniques similar to those used in the proof of Lemma 4.9. So the Borel-Cantelli lemma implies that $\mathbb{P}[X_k \neq Y_k \text{ i.o.}] = 0$, and it suffices to show $T_n/n^{1/p} \to 0$

almost surely. In order to prove the desired result, we will show

$$\sum \operatorname{Var}\Big[\frac{Y_n}{n^{1/p}}\Big] < \infty, \tag{4.9}$$

and

$$\frac{\mathbb{E}[T_n]}{n^{1/p}} \to 0. \tag{4.10}$$

Assuming (4.9) holds, by the one series theorem 4.17 and Kronecker's lemma 4.22 we have

$$\frac{T_n - \mathbb{E}[T_n]}{n^{1/p}} \to 0 \quad \text{almost surely.}$$

Combining with (4.10), we have $T_n/n^{1/p} \to 0$ almost surely, which completes the proof. Thus it remains to show (4.9) and (4.10). Consider (4.9) first.

Suppose $Z$ has the same law as $X_1$, so $\mathbb{E}[Y_n^2] = \mathbb{E}[Z^2 \mathbb{1}_{\{|Z| \leqslant n^{1/p}\}}]$. We have

$$\begin{aligned}
\sum_{n=1}^{\infty} \operatorname{Var}\Big[\frac{Y_n}{n^{1/p}}\Big] &\leqslant \sum_{n=1}^{\infty} \frac{1}{n^{2/p}} \mathbb{E}[|Y_n|^2] = \sum_{n=1}^{\infty} \frac{1}{n^{2/p}} \int_0^{\infty} 2y\mathbb{P}[|Y_n| > y]\,\mathrm{d}y \\
&= \sum_{n=1}^{\infty} \frac{1}{n^{2/p}} \sum_{m=1}^{n} \int_{(m-1)^{1/p}}^{m^{1/p}} 2y\mathbb{P}[|Y_n| > y]\,\mathrm{d}y \\
&\leqslant \sum_{n=1}^{\infty} \frac{1}{n^{2/p}} \sum_{m=1}^{n} \int_{(m-1)^{1/p}}^{m^{1/p}} 2y\mathbb{P}[|Z| > y]\,\mathrm{d}y \\
&= \sum_{m=1}^{\infty} \int_{(m-1)^{1/p}}^{m^{1/p}} \Big( \sum_{n=m}^{\infty} \frac{1}{n^{2/p}} \Big) 2y\mathbb{P}[|Z| > y]\,\mathrm{d}y \\
&\overset{\star}{=} \sum_{m=1}^{\infty} \int_{(m-1)^{1/p}}^{m^{1/p}} O(1) y^{p-2}\, 2y\mathbb{P}[|Z| > y]\,\mathrm{d}y \\
&\leqslant O(1) \sum_{m=1}^{\infty} \int_{(m-1)^{1/p}}^{m^{1/p}} p y^{p-1} \mathbb{P}[|Z| > y]\,\mathrm{d}y \\
&= O(1) \int_0^{\infty} p y^{p-1} \mathbb{P}[|Z| > y]\,\mathrm{d}y = O(1)\mathbb{E}[|Z|^p] < \infty,
\end{aligned}$$

where $(\star)$ holds $\sum_{n=m}^{\infty} m^{-2/p} \leqslant \int_m^{\infty} x^{-2/p}\,\mathrm{d}x = C \cdot m^{(p-2)/p} = O(1) \cdot y^{p-2}$ when $y \in [(m-1)^{1/p}, m^{1/p}]$.

Finally, we derive (4.10). For all $1 \leqslant m \leqslant n$, notice that $\mathbb{E}[Y_n] = -\mathbb{E}[Z \mathbb{1}_{\{|Z| > m^{1/p}\}}]$, then

$$\begin{aligned}
\frac{\mathbb{E}[|T_n|]}{n^{1/p}} &\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{n} \mathbb{E}[|Z| \mathbb{1}_{\{|Z| > k^{1/p}\}}] = \frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{1/p} \mathbb{E}\Big[\frac{|Z|}{k^{1/p}} \mathbb{1}_{\{|Z| > k^{1/p}\}}\Big] \\
&\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{1/p} \mathbb{E}\Big[\Big(\frac{|Z|}{k^{1/p}}\Big)^p \mathbb{1}_{\{|Z| > k^{1/p}\}}\Big] = \frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{\frac{1-p}{p}} \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > k^{1/p}\}}] \\
&\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{m} k^{\frac{1-p}{p}} \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > k^{1/p}\}}] + \frac{1}{n^{1/p}} \sum_{k=m}^{n} k^{\frac{1-p}{p}} \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > k^{1/p}\}}] \\
&\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{m} k^{\frac{1-p}{p}} \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > k^{1/p}\}}] + \frac{1}{n^{1/p}} \sum_{k=m}^{n} k^{\frac{1-p}{p}} \underbrace{\mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > m^{1/p}\}}]}_{\text{independent of } k} \\
&\leqslant \frac{1}{n^{1/p}} \underbrace{\sum_{k=1}^{m} k^{\frac{1-p}{p}} \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > k^{1/p}\}}]}_{\text{finite, independent of } n} + O(1)\, \mathbb{E}[|Z|^p \mathbb{1}_{\{|Z| > m^{1/p}\}}].
\end{aligned}$$

We let $n \to \infty$ first, and then let $m \to \infty$, we obtain (4.10).[1] □

**3.6. Application: Infinite Mean.** The next result is a typical application of the second Borel-Cantelli lemma.

EXAMPLE 4.27 (Infinite mean). Let $X_n$'s be independent identically distributed with $\mathbb{E}[|X_n|] = \infty$, then $\mathbb{P}[|X_n| \geqslant n \text{ i.o.}] = 1$. So if $S_n = X_1 + \cdots + X_n$, then

$$\mathbb{P}\Big[\lim_n \frac{S_n}{n} \text{ exists } \in (-\infty, \infty)\Big] = 0.$$

PROOF. Since $\{|X_1| > x\} \supset \{|X_1| > y\}$ for all $y > x$, then

$$\begin{aligned}\infty = \mathbb{E}[|X_1|] &= \int_0^\infty \mathbb{P}\{|X_1| > x\}\,\mathrm{d}x = \sum_{n\geqslant 0}\int_n^{n+1} \mathbb{P}\{|X_1| > x\}\,\mathrm{d}x \\ &\leqslant \sum_{n\geqslant 0}\int_n^{n+1} \mathbb{P}\{|X_1| > n\}\,\mathrm{d}x = \sum_{n\geqslant 0}\mathbb{P}\{|X_1| > n\} = \sum_{n\geqslant 0}\mathbb{P}\{|X_n| > n\},\end{aligned}$$

where the last equality holds since $X_1, \cdots, X_n$ are i.i.d. Now using the Borel-Cantelli lemma, we have $\mathbb{P}\{|X_n| \geqslant n \text{ i.o.}\} = 1$. For the sum $S_n/n = n^{-1}\sum_{i=1}^n X_i$ to have a finite limit, it is necessary for

$$\Big|\frac{S_n}{n} - \frac{S_{n+1}}{n+1}\Big| = \Big|\frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}\Big| \to 0.$$

But if the limit is finite, then $S_n/n(n+1) \to 0$, so $X_{n+1}/(n+1) \to 0$, but we have $\mathbb{P}\{|X_n|/n \geqslant 1 \text{ i.o.}\} = 1$, which shows that $X_{n+1}/(n+1) \to 0$ fails almost surely. □

The next result, due to Feller (1946), shows that when $E|X_1| = \infty$, $S_n/a_n$ cannot converge almost surely to a nonzero limit.

THEOREM 4.28 (Infinite mean). *Let $X_n$'s be independent identically distributed with $\mathbb{E}[|X_n|] = \infty$ and let $S_n = X_1 + \cdots + X_n$. Let $a_n$ be a sequence of positive numbers with $a_n/n$ increasing. Then $\limsup_n |S_n|/a_n = 0$ or $\infty$ according as $\sum \mathbb{P}[|X_1| \geqslant a_n] < \infty$ or $= \infty$.*

PROOF. Since $a_n/n$ is increasing, so $a_{kn} \geqslant ka_n$ for all (fixed) $k$ in $\mathbb{N}$; and $a_n$ is increasing. Thus,

$$\sum_{n=1}^\infty \mathbb{P}[|X_1| \geqslant ka_n] \geqslant \sum_{n=1}^\infty \mathbb{P}[|X_1| \geqslant a_{kn}]. \tag{4.11}$$

Moreover, it is obvious that

$$\sum_{m=k}^\infty \mathbb{P}[|X_1| \geqslant a_m] = \sum_{j=1}^\infty \sum_{n=1}^k \mathbb{P}[|X_1| \geqslant a_{jk+n}] \leqslant \sum_{j=1}^\infty \sum_{m=1}^k \mathbb{P}[|X_1| \geqslant a_{jk}] = k\sum_{j=1}^\infty \mathbb{P}[|X_1| > a_{jk}].$$

Thus we know

$$(4.11) \geqslant \frac{1}{k}\sum_{m=k}^\infty \mathbb{P}[|X_1| \geqslant a_m].$$

The last observation shows that if the sum $\sum \mathbb{P}[|X_1| \geqslant a_n]$ is infinite, then $k^{-1}\sum_{m=k}^\infty \mathbb{P}[|X_1| \geqslant a_m] = \infty$, so the left side of (4.11) is infinite. That is, $\sum_{n=1}^\infty \mathbb{P}[|X_1| \geqslant ka_n] = \sum_{n=1}^\infty \mathbb{P}[|X_n|/a_n \geqslant k] = \infty$ for all $k$ in $\mathbb{N}$. Using Borel-Cantelli lemma, we know that for any $n$ and $k$ in $\mathbb{N}$, there is $m \geqslant n$ such that $|X_m|/a_m \geqslant k$, so $\limsup_n |X_n|/a_n = \infty$. Since $\max\{|S_{n-1}|, |S_n|\} \geqslant |X_n|/2$ and $|S_{n-1}|/a_n \leqslant |S_{n-1}|/a_{n-1}$, so we get $\limsup |X_n|/a_n \leqslant \limsup |S_n| \vee |S_{n-1}|$. If $\limsup |X_n|/a_n \leqslant \limsup |S_n|/a_n$, so we get the desired result that $\limsup |S_n|/a_n = \infty$; otherwise, we must have

[1] This proof references the proof provided in Professor Hao WU's lecture notes on probability theory from Tsinghua University.

$\limsup |X_n|/a_n \leqslant \limsup |S_{n-1}|/a_n \leqslant \limsup |S_{n-1}|/a_{n-1} = \limsup |S_n|/a_n$, so the desired result also holds. To prove the other half, we begin with the identity

$$\sum_{m=1}^{\infty} m\mathbb{P}[a_{m-1} \leqslant |X_i| < a_m] = \sum_{n=1}^{\infty} \mathbb{P}[|X_i| > a_{n-1}]. \tag{4.12}$$

To see this, write $m = \sum_{n=1}^{m} 1$ and then change the order of these sums. We now let $Y_n = X_n \mathbb{1}_{\{|X_n|<a_n\}}$, and $T_n = Y_1 + \cdots + Y_n$. When the sum is $\sum \mathbb{P}[|X_1| \geqslant a_n] = \sum \mathbb{P}[X_n \neq Y_n]$ finite, we have $\mathbb{P}[X_n = Y_n \text{ i.o.}] = 0$, and it suffices to investigate the behavior of the $T_n$. To do this, we let $a_0 = 0$ and compute

$$\begin{aligned}
\sum_{n=1}^{\infty} \operatorname{Var}\Big[\frac{Y_n}{a_n}\Big] &\leqslant \sum_{n=1}^{\infty} \frac{\mathbb{E}[Y_n^2]}{a_n^2} = \sum_{n=1}^{\infty} a_n^{-2} \sum_{m=1}^{n} \mathbb{E}[Y_i^2 \mathbb{1}_{\{a_{m-1}\leqslant |X_i|<a_n\}}] \\
&= \sum_{m=1}^{\infty} \Big( \sum_{n=m}^{\infty} \frac{1}{a_n^2} \Big) \mathbb{E}[Y_i^2 \mathbb{1}_{\{a_{m-1}\leqslant |X_i|<a_m\}}] \\
&\leqslant \sum_{m=1}^{\infty} \Big( \sum_{n=m}^{\infty} \frac{1}{a_n^2} \Big) a_m^2 \mathbb{E}[\mathbb{1}_{\{a_{m-1}\leqslant |X_i|<a_m\}}] \\
&\overset{\star}{\leqslant} \sum_{m=1}^{\infty} \big(Cma_m^{-2}\big) \cdot a_m^2 \mathbb{P}[a_{m-1} \leqslant |X_i| < a_m] \\
&= C \sum_{m=1}^{\infty} m\mathbb{P}[a_{m-1} \leqslant |X_i| < a_m] \\
&= \sum_{n=1}^{\infty} \mathbb{P}[|X_i| > a_{n-1}] < \infty,
\end{aligned}$$

where $(\star)$ holds since $a_n \geqslant na_m/m$, we have $\sum_{n=m}^{\infty} a_n^{-2} \leqslant (m^2/a_m^2) \sum_{n=m}^{\infty} n^{-2} \leqslant Cma_m^{-2}$ because the series $\sum n^{-2}$ converges and $m$ here is a finite integer.

**Can** we derive $(T_n - \mathbb{E}[T_n])/a_n \to 0$ almost surely directly using one series theorem 4.17 and Kronecker's lemma 4.22? Actually, we still need to verify that $a_n \uparrow \infty$ in this case. Indeed, we note that if $\mathbb{E}[|X_i|] = \infty$, $\sum \mathbb{P}[|X_i| > a_n] < \infty$, and $a_n/n$ is increasing, we must have $a_n/n \uparrow \infty$, which implies that $a_n \uparrow \infty$.

We show this by *contradiction*. If $a_n/n \leqslant M$, where $M > 0$ is constant, then $\mathbb{P}[|X_i| > a_n] \geqslant \mathbb{P}[|X_i| > Mn]$, then using the similar technique we used in the proof of Example 4.27, we have

$$\infty > \sum_{n\geqslant 0} \mathbb{P}[|X_i| > a_n] \geqslant \sum_{n\geqslant 0} \mathbb{P}[|X_i| > Mn] \geqslant \mathbb{E}[|X_i|] = \infty,$$

which gives a contradiction! Now, $(T_n - \mathbb{E}[T_n])/a_n \to 0$ almost surely holds!

The last step is to show $\mathbb{E}[T_n]/a_n \to 0$. We observe that for any fixed $N$,

$$\begin{aligned}
\Big|\frac{1}{a_n} \sum_{m=1}^{n} \mathbb{E}[Y_m]\Big| &\leqslant \frac{1}{a_n} \sum_{m=1}^{n} \mathbb{E}[|X_m| \mathbb{1}_{\{|X_m|<a_m\}}] \\
&= \frac{1}{a_n} \sum_{m=1}^{n} \Big( \mathbb{E}[|X_i| \mathbb{1}_{\{|X_i|<a_N\}}] + \mathbb{E}[|X_i| \mathbb{1}_{\{a_N\leqslant |X_i|<a_m\}}] \Big) \\
&= \frac{n}{a_n} \cdot a_N \mathbb{P}[|X_i| < a_N] + \frac{n}{a_n} \mathbb{E}[|X_i| \mathbb{1}_{\{a_N\leqslant |X_i|<a_n\}}] \\
&\leqslant \frac{na_N}{a_n} + \frac{n}{a_n} \mathbb{E}[|X_i| \mathbb{1}_{\{a_N\leqslant |X_i|<a_n\}}].
\end{aligned}$$

Since $a_n/n \to \infty$, the first term on the right side converges to 0. Besides, since $m/a_m$ is decreasing, the second term

$$\frac{n}{a_n}\mathbb{E}[|X_i|\mathbb{1}_{\{a_N\leqslant|X_i|<a_n\}}] \leqslant \sum_{m=N+1}^{n}\frac{m}{a_m}\mathbb{E}[|X_i|\mathbb{1}_{\{a_{m-1}\leqslant|X_i|<a_m\}}] \leqslant \sum_{m=N+1}^{\infty} m\mathbb{P}[a_{m-1}\leqslant|X_i|<a_m].$$

Equation (4.12) shows that the sum above is finite, so it is small arbitrarily if $N$ is large and the desired result follows. □

## 4. Characteristic Functions

**4.1. Definition and Basic Properties.** The topics in this chapter mainly follow **Chung Kai Lai**'s book, but we do not want to cover all details in his book. So, if you need or want more, check it by yourself.

DEFINITION 4.29 (Characteristic functions). If X is a random variable, we define its **characteristic function** by

$$\varphi(t) = \mathbb{E}[e^{\mathrm{i}tX}] = \mathbb{E}[\cos tX] + \mathrm{i}\mathbb{E}[\sin tX].$$

The last formula requires taking expectation of a complex-valued random variable, but, as indicated by the second equality, no new theory is needed.

THEOREM 4.30 (Basic properties). *All characteristic functions have the following properties:*

(i) $\varphi(0) = 1$, $|\varphi(t)| = |\mathbb{E}[e^{\mathrm{i}tX}]| \leqslant \mathbb{E}[|e^{\mathrm{i}tX}|] = 1$,
(ii) $\varphi(-t) = \overline{\varphi(t)}$,
(iii) $|\varphi(t+h) - \varphi(t)| \leqslant \mathbb{E}[|e^{\mathrm{i}hX} - 1|]$, *so* $\varphi(t)$ *is uniformly continuous on* $\mathbb{R}$,
(iv) $\mathbb{E}[e^{\mathrm{i}t(aX+b)}] = e^{\mathrm{i}tb}\varphi(at)$.
(v) *It* $\varphi$ *is a characteristic function, so is* $|\varphi|^2$.

PROOF. (i) is obvious. For (ii) we note that

$$\varphi(-t) = \mathbb{E}\big[\cos(-tX) + \mathrm{i}\sin(-tX)\big] = \mathbb{E}\big[\cos(tX) - \mathrm{i}\sin(tX)\big] = \overline{\varphi(t)}.$$

For (iii), it is obvious that

$$\begin{aligned}\big|\varphi(t+h) - \varphi(t)\big| &= \big|\mathbb{E}[e^{\mathrm{i}(t+h)X} - e^{\mathrm{i}tX}]\big| = \big|\mathbb{E}[e^{\mathrm{i}tX}(e^{\mathrm{i}hX} - 1)]\big| \\ &\leqslant \mathbb{E}\big[|e^{\mathrm{i}tX}|\,|e^{\mathrm{i}hX} - 1|\big] = \mathbb{E}\big[|e^{\mathrm{i}hX} - 1|\big],\end{aligned}$$

so uniform convergence follows from the bounded convergence theorem. For (iv), it is obvious that $\mathbb{E}[e^{\mathrm{i}t(aX+b)}] = e^{\mathrm{i}tb}\mathbb{E}[e^{\mathrm{i}(ta)X}] = e^{\mathrm{i}tb}\varphi(at)$. Finally, for (v), suppose the characteristic function of $X$ is $\varphi$, and let $Y$ be independent identically distributed with $X$, then

$$\mathbb{E}[e^{\mathrm{i}t(X-Y)}] = \mathbb{E}[e^{\mathrm{i}tX}]\mathbb{E}[e^{-\mathrm{i}tY}] = \varphi(t)\varphi(-t) = |\varphi(t)|^2,$$

which completes the proof. □

The main reason for introducing characteristic functions is the following:

THEOREM 4.31 (Sum of two independent random variables). *If* $X_1$ *and* $X_2$ *are independent and have characteristic functions* $\varphi_1$ *and* $\varphi_2$*, then* $X_1 + X_2$ *has characteristic function* $\varphi_1(t)\varphi_2(t)$.

PROOF. $X_1$ and $X_2$ are independent implies $\mathbb{E}[e^{\mathrm{i}t(X_1+X_2)}] = \mathbb{E}[e^{\mathrm{i}tX_1}e^{\mathrm{i}tX_2}] = \mathbb{E}[e^{\mathrm{i}tX_1}]\mathbb{E}[e^{\mathrm{i}tX_2}]$. □

EXAMPLE 4.32 (The characteristic function of Gaussian variable). Every $X \sim \mathcal{N}(0,1)$ has probability density function

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

Then the characteristic function of $X$ is $\varphi_X(t) = \exp\{-t^2/2\}$. In general, the characteristic function of $Y = \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$ is $\exp\{\mathrm{i}\mu t - \sigma^2 t^2/2\}$ by using the property (v). †

**4.2. Uniqueness and Inversion Formula.** We defined the characteristic function for each probability distribution (or measure). The *question* is: how can we find the probability measure from a given characteristic function? This is done using the inversion formula.

THEOREM 4.33 (Inversion formula). *Suppose $\varphi$ is the characteristic function for a probability measure $\mu$. For $x < y$, we have*

$$\mu\big((x,y)\big) + \frac{1}{2}\mu(\{x\}) + \frac{1}{2}\mu(\{y\}) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}tx} - e^{-\mathrm{i}ty}}{\mathrm{i}t} \varphi(t)\,\mathrm{d}t.$$

Observe that the integrand in the right hand side is bounded by $O(|t|^{-1})$ as $|t| \to \infty$; yet we cannot assert the "infinite integral" exists (in Lebesgue sense). Indeed, it does not in general. The fact that the limit in the right hand side does exist is part of the assertion.

PROOF. We claim that

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}tx} - e^{-\mathrm{i}ty}}{\mathrm{i}t} \varphi(t)\,\mathrm{d}t &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}tx} - e^{-\mathrm{i}ty}}{\mathrm{i}t} \int_{\mathbb{R}} e^{\mathrm{i}tz}\,\mu(\mathrm{d}z)\,\mathrm{d}t \\ &= \int_{\mathbb{R}} \mu(\mathrm{d}z) \int_{-T}^{T} \frac{e^{-\mathrm{i}t(z-x)} - e^{-\mathrm{i}t(z,y)}}{2\pi\mathrm{i}t}\,\mathrm{d}t. \end{aligned} \tag{4.13}$$

This is true due to Fubini's theorem, because the integrand in the right side is bounded by a integrable function with respect to $\mu(\mathrm{d}z)\mathrm{d}t$ on $\mathbb{R} \times [-T, T]$:

$$\left| \frac{e^{-\mathrm{i}t(z-x)} - e^{-\mathrm{i}t(z,y)}}{2\pi\mathrm{i}t} \right| \leqslant |x - y|.$$

This proves (4.13). Define

$$I(T, z; x, y) = \int_{-T}^{T} \frac{e^{-\mathrm{i}t(z-x)} - e^{-\mathrm{i}t(z,y)}}{2\pi\mathrm{i}t}\,\mathrm{d}t.$$

It is clear that

$$I(T, z; x, y) = \int_0^T \frac{\sin\big(t(z-x)\big)}{\pi t}\,\mathrm{d}t - \int_0^T \frac{\sin\big(t(z-y)\big)}{\pi t}\,\mathrm{d}t.$$

The quantity $I$ is bounded in $T$, because for any $w \geqslant 0$,

$$0 \leqslant \operatorname{sgn}(\alpha) \int_0^w \frac{\sin(\alpha t)}{t}\,\mathrm{d}t \leqslant \int_0^\pi \frac{\sin t}{t}\,\mathrm{d}t \leqslant \pi. \quad \left[\frac{\sin t}{t} \leqslant 1\right]$$

Therefore, we can interchange the limit and the integral. It remains to derive the limit of $I$ as $T \to \infty$. Note that

$$\lim_{T\to\infty} \int_0^T \frac{\sin(\alpha t)}{t}\,\mathrm{d}t = \lim_{T\to\infty} \int_0^T \sin(\alpha t) \int_0^\infty e^{-tu}\,\mathrm{d}u\,\mathrm{d}t = \frac{\pi}{2}\operatorname{sgn}(\alpha).$$

As a consequence, we have

$$\begin{aligned}\lim_{T\to\infty} I(T,z;x,y) &= 0 && \text{if } z<x<y \text{ or } x<y<z,\\ &= \frac{1}{2} && \text{if } z=x \text{ or } z=y,\\ &= 1 && \text{if } x<Z<y.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-\mathrm{i}tx}-e^{-\mathrm{i}ty}}{\mathrm{i}t}\varphi(t)\,\mathrm{d}t &= \int_{\mathbb{R}}\mu(\mathrm{d}z)\lim_{T\to\infty} I(T,z;x,y)\\ &= \mu\big((x,y)\big)+\frac{1}{2}\mu(\{x\})+\frac{1}{2}\mu(\{y\}),\end{aligned}$$

this completes the proof.[2] □

The following corollary shows that the characteristic function can uniquely determine the distribution measure.

COROLLARY 4.34 (Uniqueness). *If two probability measures $\mu$ and $\nu$ have the same characteristic function, then $\mu=\nu$.*

PROOF. Suppose $A_\mu$ is the set of atoms of $\mu$ and $A_\nu$ the set of atoms of $\nu$. From the inversion formula, we have

$$\mu\big((a,b)\big)=\nu\big((a,b)\big) \quad \text{for all } a,b\in\mathbb{R}\backslash(A_\mu\cup A_\nu).$$

Note that $A_\mu$ and $A_\nu$ are countable using Theorem 1.29, thus $\mathbb{R}\backslash(A_\mu\cup A_\nu)$ is dense. Using the fact that $\{(a,b): a,b\in\mathbb{R}\backslash(A_\mu\cup A_\nu)\}$ generates $\mathcal{B}(\mathbb{R})$, we have $\mu=\nu$. □

COROLLARY 4.35 (Symmetry distribution theorem). *If $\varphi$ is real, then $X$ and $-X$ have the same distribution.*

COROLLARY 4.36 (Sum of independent Gaussian random variables). *If $X_1$ and $X_2$ are independent and have normal distributions with mean $0$ and variance $\sigma_1^2$ and $\sigma_2^2$, respectively, then $X_1+X_2$ has a normal distribution with mean $0$ and variance $\sigma_1^2+\sigma_2^2$.*

The inversion formula is simpler when $\varphi$ is integrable, but as the next result shows this only happens when the *underlying measure* (or *underlying distribution*) is nice.

THEOREM 4.37 (Integrable characteristic function). *Suppose $\varphi$ is the characteristic function for the distribution function $F$. If $\varphi$ in $L^1(\mathbb{R})$, then $F$ is differentiable and the density function*

$$p(x)=F'(x)=\frac{1}{2\pi}\int_{\mathbb{R}}e^{-\mathrm{i}tx}\varphi(t)\,\mathrm{d}t, \quad \textit{and} \quad \varphi(t)=\int_{\mathbb{R}}e^{-\mathrm{i}tx}p(x)\,\mathrm{d}t.$$

PROOF. Write the inversion formula in terms of $F$:

$$\frac{F(y)+F(y-)}{2}-\frac{F(x)+F(x-)}{2}=\frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-\mathrm{i}tx}-e^{-\mathrm{i}ty}}{\mathrm{i}t}\varphi(t)\,\mathrm{d}t.$$

Let $x\to y-$, the right side goes to zero (we are allowed to interchange the limit and the integral due to the hypothesis on $\varphi$). Thus $F$ is continuous. The above formula then writes:

$$F(y)-F(x)=\frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-\mathrm{i}tx}-e^{-\mathrm{i}ty}}{\mathrm{i}t}\varphi(t)\,\mathrm{d}t.$$

[2] We have skipped many calculations. You can check Section 6.2 of Chung Kai Lai's textbook for more details.

Divide both sides by $y-x$ and then let $y \to x+$, we obtain the conclusion after interchanging the limit and the integral. □

The following theorem yield information on the atoms of $\mu$ by means of $\varphi$.

THEOREM 4.38 (Inversion formula). *For each $x$, we have*

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-\mathrm{i}tx} \varphi(t)\,\mathrm{d}t = \mu(\{x\}). \tag{4.14}$$

PROOF. Indeed, it is obvious that

$$\begin{aligned} \frac{1}{2T}\int_{-T}^{T} e^{-\mathrm{i}tx}\varphi(t)\,\mathrm{d}t &= \int_{\mathbb{R}} \mu(\mathrm{d}z) \frac{1}{2T}\int_{-T}^{T} e^{\mathrm{i}t(z-x)}\,\mathrm{d}t = \int_{\mathbb{R}} \mu(\mathrm{d}z)\frac{1}{T}\int_{0}^{T} \cos\big(t(z-x)\big)\,\mathrm{d}t \\ &= \int_{\mathbb{R}\setminus\{x\}} \mu(\mathrm{d}z)\frac{\sin\big(T(z-x)\big)}{T(z-x)} + \mu(\{x\}). \end{aligned} \tag{4.15}$$

Then, Equation (4.14) holds by the following observation:

$$\lim_{T\to\infty} \int_{\mathbb{R}\setminus\{x\}} \mu(\mathrm{d}z)\frac{\sin\big(T(z-x)\big)}{T(z-x)} = 0.$$

This completes the proof. □

**4.3. Lévy-Crámer Continuity Theorem.** The goal of this section is the following convergence theorem.

THEOREM 4.39 (Lévy-Crámer continuity theorem I). *Let $(\mu_n)$ be a sequence of probability distribution measures with characteristic functions $(\varphi_n)$. Suppose that*

1. *$\varphi_n$ converges everywhere in $\mathbb{R}$ and defines the limiting function $\varphi$;*
2. *$\varphi$ is continuous at $t=0$.*

*Then we have*

(i) *$\mu_n \Rightarrow \mu$, where $\mu$ is a probability measure;*
(ii) *the characteristic function of $\mu$ is $\varphi$.*

We first discuss the converse direction of this theorem.

THEOREM 4.40 (Lévy-Crámer continuity theorem II). *Let $(\mu_n)$ and $\mu$ be a sequence of probability measure with characteristic functions $(\varphi_n)$ and $\varphi$, respectively. If $\mu_n \Rightarrow \mu$, then $\varphi_n$ converges to $\varphi$ uniformly in every finite interval. Furthermore, the family $(\varphi_n)$ is equicontinuous on $\mathbb{R}$.*

PROOF OF THEOREM 4.40. Since the real and imaginary part of $e^{\mathrm{i}tx}$ are bounded continuous, the *weak convergence* implies $\varphi_n = \int_{\mathbb{R}} e^{\mathrm{i}tx}\,\mu_n(\mathrm{d}x) = \int_{\mathbb{R}} e^{\mathrm{i}tx}\,\mu(\mathrm{d}x) \to \varphi$ pointwise.

We first show the *equicontinuity*. That is, for any $\varepsilon > 0$, there is $\delta > 0$ such that for all $n$ in $\mathbb{N}$, $|\varphi_n(t+h) - \varphi_n(t)| \leqslant \varepsilon$ as long as $|h| \leqslant \delta$. Indeed, for any $t$ in $\mathbb{R}$ and $h$ in $\mathbb{R}$, we have

$$\begin{aligned} |\varphi_n(t+h)-\varphi_n(t)| &\leqslant \int_{\mathbb{R}} |e^{\mathrm{i}hx}-1|\,\mu_n(\mathrm{d}x) \leqslant \int_{\{|x|\leqslant A\}} |hx|\,\mu_n(\mathrm{d}x) + \int_{\{|x|>A\}} 2\,\mu_n(\mathrm{d}x) \\ &\leqslant |h|A + 2\int_{\{|x|>A\}} \mu_n(\mathrm{d}x). \end{aligned}$$

For any $\varepsilon > 0$, there exists $n_0 = n_0(A, \varepsilon)$ such that

$$|\varphi_n(t+h) - \varphi_n(t)| \leqslant |h|A + 2\int_{\{|x|>A\}} \mu(\mathrm{d}x) + \varepsilon/4.$$

This gives the equicontinuity for $(f_n)$: for any $\varepsilon > 0$, we choose $A$ large enough such that $\mu(\{|x| > A\}) \leqslant \varepsilon/4$, then for any $h$ such that $|h| \leqslant \delta := \varepsilon/(4A)$, we have

$$|\varphi_n(t+h) - \varphi_n(t)| \leqslant \varepsilon, \quad \text{for all } t \in \mathbb{R}, \text{ and } n \geqslant n_0$$

as desired.

Next, we show the *uniform convergence* on compact interval $I$, i.e., we show that, for any $\varepsilon > 0$, there is $n_0 = n_0(I, \varepsilon)$ such that $|\varphi_n(t) - \varphi(t)| \leqslant \varepsilon$ for $n \geqslant n_0$ any all $t$ in $I$.

From the equicontinuity, there is $\delta$ such that $|\varphi_n(t) - \varphi(s)| \leqslant \varepsilon$ as long as $|t - s| \leqslant \delta$. Choose finite sequence of points $\{a_1, \cdots, a_{m_0}\} \subset I$ such that $\bigcup_{k=1}^{m_0}(a_k - \delta, a_k + \delta)$ is a finite cover of $I$. By the pointwise convergence, $\varphi_n(a_k) \to \varphi(a_k)$ for $1 \leqslant k \leqslant m_0$. Thus there is $n_0 = n_0(I, \varepsilon)$ such that

$$|\varphi_n(a_k) - \varphi(a_k)| \leqslant \varepsilon, \quad \text{for all } 1 \leqslant k \leqslant m_0, \text{ and } n \geqslant n_0.$$

For any $t$ in $I$, there exists $k$ such that $t$ in $(a_k - \delta, a_k + \delta)$, thus

$$|\varphi_n(t) - \varphi(t)| \leqslant |\varphi(t) - \varphi_n(a_k)| + |\varphi_n(t_k) - \varphi(a_k)| + |\varphi(a_k) - \varphi(t)| \leqslant 3\varepsilon, \quad \text{for all } n \geqslant n_0.$$

This gives the uniform convergence. □

Using the Helly's selection principle, we know that $(\mu_n)$ contains a convergent subsequence: $\mu_{n_k} \Rightarrow \mu$ where $\mu$ is a finite measure satisfying $\mu(\mathbb{R}) \leqslant 1$ (subprobability measure). To show Theorem 4.39, we need to argue that all subsequential limit are the same with the probability measure which is given by $\varphi$. To this end, we first argue that, under the assumption of the theorem, any subsequential limit is indeed a *probability* measure, which needs the following lemma.

LEMMA 4.41 (Lower bound given by characteristic function). *Suppose that $\varphi$ is the characteristic function of a probability distribution measure $\mu$. For each $A > 0$, we have*

$$\mu\big([-2A, 2A]\big) \geqslant A\left|\int_{-A^{-1}}^{A^{-1}} \varphi(t)\,\mathrm{d}t\right| - 1.$$

PROOF. First, using the proof of Equation (4.15), we have

$$\frac{1}{2T}\int_{-T}^{T} \varphi(t)\,\mathrm{d}t = \int_{\mathbb{R}} \frac{\sin Tx}{Tx}\,\mu(\mathrm{d}x).$$

Thus,

$$\begin{aligned}\frac{1}{2T}\left|\int_{-T}^{T} \varphi(t)\,\mathrm{d}t\right| &\leqslant \mu\big([-2A, 2A]\big) + \frac{1}{2TA}\Big[1 - \mu\big([-2A, 2A]\big)\Big] \\ &= \Big(1 - \frac{1}{2TA}\Big)\mu\big([-2A, 2A]\big) + \frac{1}{2TA}.\end{aligned}$$

Set $T = A^{-1}$, we obtain the desired result. □

PROOF OF THEOREM 4.39. Suppose $(\mu_{n_k})$ is a convergent subsequence of $(\mu_n)$ and denote the limit by $\mu$. *First*, we argue that $\mu$ is a probability measure. By the above lemma, we have, when $\pm 2\delta^{-1}$ are the continuity points of the distribution function derived from $\mu$,

$$\mu(\mathbb{R}) \geqslant \mu\big([-2\delta^{-1}, 2\delta^{-1}]\big) = \lim_k \mu_{n_k}\big([-2\delta^{-1}, 2\delta^{-1}]\big) \geqslant \limsup_k \frac{1}{\delta}\left|\int_{-\delta}^{\delta} \varphi_{n_k}(t)\,\mathrm{d}t\right| - 1. \tag{4.16}$$

Since $\varphi_{n_k} \to \varphi$ pointwise everywhere by hypothesis, and by bounded convergence theorem, we have

$$\limsup_k \frac{1}{\delta}\left|\int_{-\delta}^{\delta} \varphi_{n_k}(t)\,\mathrm{d}t\right| = \frac{1}{\delta}\left|\int_{-\delta}^{\delta} \varphi(t)\,\mathrm{d}t\right|.$$

Since $\varphi$ is continuous at zero and $\varphi(0) = 1$, we have

$$\lim_{\delta\to 0} \frac{1}{2\delta}\left|\int_{-\delta}^{\delta} \varphi(t)\,\mathrm{d}t\right| = \lim_{\delta\to 0}\left|\frac{1}{2\delta}\int_{-\delta}^{\delta} \varphi(t)\,\mathrm{d}t\right| = 1,$$

where the last equality holds using the derivative of *indefinite integral*. Thus, for any $\varepsilon > 0$, there is $\delta_0 > 0$ such that, for any $0 < \delta \leqslant \delta_0$,

$$\frac{1}{\delta}\left|\int_{-\delta}^{\delta} \varphi(t)\,\mathrm{d}t\right| \geqslant 2 - \varepsilon.$$

Therefore,

$$\mu(\mathbb{R}) \geqslant 1 - \varepsilon$$

using Equation (4.16). This holds for any $\varepsilon > 0$. Thus $\mu(\mathbb{R}) = 1$ and $\mu$ is a probability measure.

Let $\phi$ be the characteristic function of $\mu$, by Theorem 4.40, we know that $\varphi_{n_k} \to \phi$ everywhere. By the hypothesis, $\phi \equiv \varphi$. We see that any subsequential limit has the characteristic function $\varphi$, and hence, by Theorem 4.33, any subsequential limit has the same probability measure $\mu$ whose characteristic function is given by $\varphi$, since $\mu$ can be uniquely determined by the class of open intervals whose endpoints are not atomic points of $\mu$. □

EXAMPLE 4.42. Let $\mu_n$ be the probability measure such that $\mu(\{0\}) = \mu(\{n\}) = 1/2$. Then $\mu_n \Rightarrow \mu$ where $\mu$ has mass $1/2$ at $\{0\}$ and zero elsewhere. For the characteristic functions, we have

$$\varphi_n = 2^{-1} + 2^{-1}e^{\mathrm{i}tn},$$

they do not converge. †

EXAMPLE 4.43. Let $\mu_n$ be the uniform distribution measure on $[-n, n]$. Then $\mu_n \Rightarrow \mu$, where $\mu$ is identically zero. For the characteristic functions, we have

$$\begin{aligned}\varphi_n(t) &= \frac{\sin nt}{nt}, && \text{if } t \neq 0,\\ &= 1, && \text{if } t = 0.\end{aligned}$$

They converge and the limiting function is

$$\begin{aligned}\varphi(t) &= 0, && \text{if } t \neq 0,\\ &= 1, && \text{if } t = 0.\end{aligned}$$

The limiting function is not continuous at zero. So $\mu$ is NOT a probability measure obviously and theoretically. †

**4.4. Moments and Derivatives.** The last topic we discussed in the section on characteristic functions is the relationship between the *higher-order derivatives* of the characteristic function and the *higher-order moments* of the random variable.

THEOREM 4.44 (Derivatives and moments). *Suppose $\mu$ is a probability measure and $\varphi$ is its characteristic function.*

(i) *If $\mu$ has a finite moment of order $k \geqslant 1$, then $\varphi$ has a bounded continuous derivative of order $k$ given by*

$$\varphi^{(k)}(t) = \int_{\mathbb{R}} (\mathrm{i}x)^k e^{\mathrm{i}tx}\, \mu(\mathrm{d}x), \quad \Rightarrow \quad \varphi^{(k)}(0) = \mathrm{i}^k \mathbb{E}[X^k].$$

*Using Taylor's expansion of $\varphi(t)$, we have*

$$\varphi(t) = 1 + \sum_{k=1}^{n} \frac{\mathrm{i}^k \mathbb{E}[X^k]}{k!} t^k + o(|t|^n).$$

(ii) *If $\varphi$ has a finite derivative of even order $k$ at $t = 0$, then $\mu$ has a finite moment of order $k$.*

PROOF. (i). For $k = 1$, the first assertion follows from the formula:

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \int_{\mathbb{R}} \frac{e^{\mathrm{i}(t+h)x} - e^{\mathrm{i}tx}}{h}\, \mu(\mathrm{d}x) = \int_{\mathbb{R}} \frac{e^{\mathrm{i}hx} - 1}{h} \cdot e^{\mathrm{i}tx}\, \mu(\mathrm{d}x).$$

It is obvious that the integrand above it dominated by $|x|$, since

$$\left| \frac{e^{\mathrm{i}hx} - 1}{h} \cdot e^{\mathrm{i}tx} \right| \leqslant \left| \frac{hx}{h} \right| \cdot |e^{\mathrm{i}tx}| = |x|.$$

Hence if $\mu$ has a finite moment of order $k \geqslant 1$, then $\mathbb{E}[|X|] = \int_{\mathbb{R}} |x|\, \mu(\mathrm{d}x) < \infty$. So we may let $h \to 0$ under the integral sign and obtain (a) when $k = 1$. The higher order can be derived using induction, which is omit here.

(ii). We start with $k = 2$ and suppose $\varphi''(0)$ exists and is finite. We have

$$\varphi''(0) = \lim_{\delta \to 0} \frac{\varphi(\delta) - 2\varphi(0) + \varphi(\delta)}{\delta^2}.$$

For the right side, we have

$$\frac{\varphi(\delta) - 2\varphi(0) + \varphi(\delta)}{\delta^2} = \int_{\mathbb{R}} \frac{e^{\mathrm{i}\delta x} - 2 + e^{-\mathrm{i}\delta x}}{\delta^2}\, \mu(\mathrm{d}x) = -2 \int_{\mathbb{R}} \frac{1 - \cos \delta x}{\delta^2}\, \mu(\mathrm{d}x).$$

By Fatou's lemma, we have

$$\int_{\mathbb{R}} x^2\, \mu(\mathrm{d}x) = 2 \int_{\mathbb{R}} \lim_{\delta \to 0} \frac{1 - \cos(\delta x)}{\delta^2}\, \mu(\mathrm{d}x) \leqslant 2 \liminf_{\delta \to 0} \int_{\mathbb{R}} \frac{1 - \cos(\delta x)}{\delta^2}\, \mu(\mathrm{d}x) = -\varphi''(0).$$

Therefore, $\mu$ has finite second moment. For *general* $2k$, suppose the conclusion holds for $2k - 2$ and $\varphi^{(2k)}(0)$ exists and is finite. By the induction hypothesis, we have

$$\varphi^{(2k-2)}(0) = \int_{\mathbb{R}} (\mathrm{i}x)^{2k-2} e^{\mathrm{i}tx}\, \mu(\mathrm{d}x).$$

Define

$$G(x) = \int_{-\infty}^{x} y^{2k-2}\, \mu(\mathrm{d}y) \quad \text{for all } x \in \mathbb{R}.$$

If $G(\infty) > 0$, then $G(\cdot)/G(\infty)$ is a distribution. For its corresponding probability measure, its characteristic function is given by

$$\phi(t) = \frac{1}{G(\infty)} \int_{\mathbb{R}} e^{\mathrm{i}tx} x^{2k-2}\, \mu(\mathrm{d}x) = \frac{(-1)^{k-1} \varphi^{(2k-2)}(t)}{G(\infty)}.$$

By the induction hypothesis, $\phi''(0)$ exists and is finite. By the proof of the case with $k = 2$, we know that $G$ has finite moment of order 2, and thus $\mu$ has finite moment of order $2k$ as desired.

If $G(\infty) = 0$, then $\mu = \delta_0$. Then $\varphi \equiv 1$ and $\mu$ has finite moment of order $2k$. □

**4.5. Applications.** We give two basic applications:

EXAMPLE 4.45 (Khintchine: $L^1$ weak law of large numbers). Let $X_n$'s be independent and identically distributed with $\mathbb{E}[|X_n|] < \infty$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu = \mathbb{E}[X_1]$. Then $S_n/n \to \mu$ in probability.

PROOF. Suppose $\varphi$ is the characteristic function of $X_1$. Then the characteristic function of $S_n/n$ is given by

$$\varphi_n(t) = \mathbb{E}\Big[\exp\Big(\frac{\mathrm{i}t(X_1+\cdots+X_n)}{n}\Big)\Big] = \prod_{i=1}^{n}\mathbb{E}\Big[\exp\Big(\mathrm{i}\cdot\frac{t}{n}\cdot X_i\Big)\Big] = \varphi\Big(\frac{t}{n}\Big)^n.$$

Taking Taylor's expansion of $\varphi(t/n)$ at $t=0$ and using Theorem 4.44 we get

$$\varphi\Big(\frac{t}{n}\Big) = \varphi(0) + \varphi'(0)\cdot\frac{t}{n} + o\Big(\frac{t}{n}\Big) = 1 + \mathrm{i}\mu\frac{t}{n} + o\Big(\frac{t}{n}\Big).$$

Therefore,

$$\varphi_n(t) = \Big(1 + \mathrm{i}\mu\frac{t}{n} + o(t/n)\Big)^n \to e^{\mathrm{i}\mu t}$$

as $n\to\infty$. In other words, the characteristic functions $\varphi_n$ converge to the characteristic function of the Dirac measure $\delta_\mu$, i.e., $X_n \to \mu$ in distribution. This gives the convergence in probability using converging together lemma introduced in the last chapter. □

Using the characteristic function, we can present the simplest version of the Central Limit Theorem, which is already introduced in elementary courses.

EXAMPLE 4.46 (Lindeberg-Lévy: Central limit theorem (i.i.d. case)). Let $X_n$'s be independent and identically distributed with $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}[X_1] < \infty$. Let $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow \Phi,$$

where $\Phi$ is the *standard normal distribution*.

PROOF. We may assume that $\mu = 0$. Consider the characteristic function of $S_n/(\sigma\sqrt{n})$:

$$\varphi_n(t) := \mathbb{E}\Big[\exp\Big(\frac{\mathrm{i}t(X_1+\cdots+X_n)}{\sigma\sqrt{n}}\Big)\Big] = \varphi\Big(\frac{t}{\sigma\sqrt{n}}\Big)^n,$$

where $\varphi$ and $\varphi_n$ are the characteristic functions of $X_1$ and $S_n/(\sigma\sqrt{n})$, respectively. Using Theorem 4.44,

$$\varphi\Big(\frac{t}{\sigma\sqrt{n}}\Big) = 1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{\sigma^2 n}\Big).$$

Therefore we have $\varphi_n(t) \to \exp(-t^2/2)$ as $n\to\infty$, which is the characteristic function of standard normal distribution. This gives the weak convergence. □

## 5. Central Limit Theorem

**5.1. Lyapunov's Central Limit Theorem.** In the last section of this chapter, we only want to talk about two types of central limit theorem, that is, Lyapunov's theorem and Lindeberg-Feller theorem.

The version of these theorem we provide is *consistent* with Durrett's version, but our proof mainly refers to **Erhan**'s textbook. We will deal with an triangular array $X_{n,k}$, $1 \leqslant k \leqslant n$ of random variables again, and discuss the limiting behavior of their row sums $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,n}$. Lyapunov's theorem is formulated with a condition on the *third moments* of the $X_{n,k}$.

THEOREM 4.47 (Lyapunov's central limit theorem). *For all $n$ let $X_{n,k}$, $1 \leqslant k \leqslant n$, be independent. Suppose that*

(i) $\mathbb{E}[X_{n,k}] = 0$ *for all $n$ in $\mathbb{N}$ and $1 \leqslant k \leqslant n$,*
(ii) $\mathrm{Var}[S_n] = 1$ *for all $n$ in $\mathbb{N}$, where $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,n}$,*
(iii) $\lim_n \sum_{k=1}^n \mathbb{E}[|X_{n,k}|^3] = 0$.

*Then, we have $S_n \Rightarrow \Phi$, where $\Phi$ is the standard normal distribution.*

The most essential part of the proof of Lyapunov's CLT is given as a lemma below. The lemma needed, due to *Lindeberg*, considers a sum of $k$ independent variables and approximates the distribution of the sum by the *Gaussian distribution* with the same mean and variance as the sum.

LEMMA 4.48 (Approximation lemma). *Let $Y_1, \cdots, Y_k$ be independent and have mean 0. Let $S$ be their sum and assume that $\mathrm{Var}[S] = 1$. Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable thrice and assume that the derivatives $f', f'', f'''$ are bounded and continuous, with $C$ a bound for $|f'''|$. Then,*

$$\big|\mathbb{E}[f \circ S] - \mathbb{E}[f \circ \Phi]\big| \leqslant C \sum_{j=1}^{k} \mathbb{E}[|Y_j|^3],$$

*where $\Phi$ is the standard normal distribution.*

PROOF. Let $Z_1, \cdots, Z_k$ be independent Gaussian variables with means $\mathbb{E}[Z_j] = \mathbb{E}[Y_j] = 0$ and variances $\mathrm{Var}[Z_j] = \mathrm{Var}[Y_j]$. Then $T = Z_1 + \cdots + Z_k$ has the same distribution as $\Phi$ and the claim is that

$$\big|\mathbb{E}[f \circ S] - \mathbb{E}[f \circ T]\big| \leqslant C \sum_{j=1}^{k} \mathbb{E}[|Y_j|^3]. \tag{4.17}$$

The idea is to replace, one at a time, each $Y_j$ with $Z_j$. So we define $V_1, \cdots, V_k$ recursively by

$$S = V_1 + Y_1; \quad V_j + Z_j = V_{j+1} + Y_{j+1}, \quad 1 \leqslant j < k,$$

and note that $V_k + Z_k = T$. Then, using the above equation and by *canceling terms* one by one, we have

$$f \circ S - f \circ T = \sum_{j=1}^{k} \big[f(V_j + Y_j) - f(V_j + Z_j)\big],$$

and to prove (4.17), it is enough to show that, for each $j$,

$$\big|\mathbb{E}[f(V_j + Y_j)] - \mathbb{E}[f(V_j + Z_j)]\big| \leqslant C\mathbb{E}[|Y_j|^3]. \tag{4.18}$$

To this end, *fix $j$ and drop it from notation*. Using Taylor's expansion for $f$:

$$f(v + x) = f(v) + f'(v)x + \frac{1}{2}f''(v)x^2 + \frac{1}{6}R(v)x^3, \tag{4.19}$$

where $|R(v)| \leqslant C$, the bound for $|f'''|$. Note that $V, Y, Z$ are independent, and $f, f', f''$ are bounded, and $\mathbb{E}[Y] = \mathbb{E}[Z] = 0$, and $\mathbb{E}[Y^2] = \mathbb{E}[Z^2] = b^2$ ($b$ does depend on $j$). Now, replace $v$ with $V$ and $x$ with $Y$ and take expectations on both sides of (4.19) we get

$$\mathbb{E}[f(V + Y)] = \mathbb{E}[f(V)] + \mathbb{E}[f'(V)]\mathbb{E}[Y] + \frac{1}{2}\mathbb{E}[f''(V)]\mathbb{E}[Y^2] + \frac{1}{6}R(V)\mathbb{E}[Y^3],$$

next, replace $v$ with $V$ and $x$ with $Z$ and take expectations on both sides of (4.19) we get

$$\mathbb{E}[f(V + Z)] = \mathbb{E}[f(V)] + \mathbb{E}[f'(V)]\mathbb{E}[Z] + \frac{1}{2}\mathbb{E}[f''(V)]\mathbb{E}[Z^2] + \frac{1}{6}R(V)\mathbb{E}[Z^3],$$

taking differences and noting that $\mathbb{E}[Y] = \mathbb{E}[Z] = 0$, and $\mathbb{E}[Y^2] = \mathbb{E}[Z^2]$, we get

$$\big|\mathbb{E}[f(V+Y)] - \mathbb{E}[f(V+Z)]\big| \leqslant C/6\big(\mathbb{E}[|Y|^3] + \mathbb{E}[|Z|^3]\big). \tag{4.20}$$

Since $Z$ has the Gaussian distribution with mean 0 and variance $\mathbb{E}[Z^2] = \mathbb{E}[Y^2] = b^2$, a direct computation shows that $\mathbb{E}[|Z|^3] = b^3\sqrt{8/\pi} \leqslant 2b^3$. Since the $L^2$-norm is less or equal than the $L^3$-norm, so $b = (\mathbb{E}[Y^2])^{1/2} \leqslant (\mathbb{E}[|Y|^3])^{1/3}$. Hence

$$\mathbb{E}[|Y|^3] + \mathbb{E}[|Z|^3] \leqslant \mathbb{E}[|Y|^3] + 2\mathbb{E}[|Y|^3] = 3\mathbb{E}[|Y|^3].$$

Putting this into (4.20) shows (4.18) and completes the proof. □

PROOF OF THEOREM 4.47. We need to apply Lemma 4.48. Let $S = S_n = X_{n,1} + \cdots + X_{n,n}$, and define $f = \sin tx$ for a fixed $t$ in $\mathbb{R}$, separately. Since $f''' \leqslant |t|^3$, then we get

$$\big|\mathbb{E}[\sin tS_n] - \mathbb{E}[\sin t\Phi]\big| \leqslant |t^3| \sum_{k=1}^{n} \mathbb{E}[|X_{n,k}|^3],$$

and the same result if we replace $\sin tx$ by $\cos tx$. Thus we have

$$\begin{aligned}\big|\mathbb{E}[e^{\mathrm{i}tS_n}] - \mathbb{E}[e^{\mathrm{i}t\Phi}]\big| &= \big|(\mathbb{E}[\cos tS_n] - \mathbb{E}[\cos t\Phi]) + \mathrm{i}(\mathbb{E}[\sin tS_n] - \mathbb{E}[\sin t\Phi])\big| \\ &\leqslant \big|\mathbb{E}[\cos tS_n] - \mathbb{E}[\cos t\Phi]\big| + \big|\mathbb{E}[\sin tS_n] - \mathbb{E}[\sin t\Phi])\big| \leqslant 2|t^3| \sum_{k=1}^{n} \mathbb{E}[|X_{n,k}|^3],\end{aligned}$$

where the first inequality holds since $(a^2+b^2)^{1/2} \leqslant (a^2)^{1/2} + (b^2)^{1/2}$. Letting $n \to \infty$ and we get

$$\big|\mathbb{E}[e^{\mathrm{i}tS_n}] - \mathbb{E}[e^{\mathrm{i}t\Phi}]\big| \leqslant 2|t^3| \sum_{k=1}^{n} \mathbb{E}[|X_{n,k}|^3] \to 0$$

for all $t$ in $\mathbb{R}$. Using Theorem 4.39 we get $S_n \Rightarrow \Phi$, which completes the proof. □

REMARK 4.49. There is another method to prove Lyapunov's CLT, which can be founded in **Chung Kai Lai**'s book, and we omit this method here.

In general, we may suppose that $S_n$ has mean $a_n$ and variance $b_n^2$. Then, the Lyapunov's theorem applies to the triangular array $(Y_{n,k})$ with $Y_{n,k} = (X_{n,k} - \mathbb{E}[X_{n,k}])/b_n$ to show that, if $\lim b_n^{-3} \sum_{j=1}^{n} \mathbb{E}[|X_{n,k} - \mathbb{E}[X_{n,k}]|^3] = 0$, then $(Z_n - a_n)/b_n \Rightarrow \Phi$. †

COROLLARY 4.50 (An extension of Lyapunov's CLT). *Let $a_n = \mathbb{E}[S_n]$ and $b_n^2 = \mathrm{Var}[S_n]$. Suppose that $a_n \to a$ and $b_n \to b$, where $b \neq 0$. Assume that for each $n$ and $k$ there is a constant $C_{n,k}$ such that $|X_{n,k}| \leqslant C_{n,k}$ and that $\lim_n \sup_{k \geqslant n} C_{n,k} = 0$, then $(Z_n - a)/b \Rightarrow \Phi$.*

PROOF. Put $Y_{n,k} = (X_{n,k} - \mathbb{E}[X_{n,k}])/b_n$. Since $|Y_{n,k}| \leqslant 2C_{n,k}/b_n \leqslant \varepsilon_n$, where

$$\varepsilon_n = 2(\max_{1 \leqslant k \leqslant n} C_{n,k})/b_n$$

and therefore $|Y_{n,k}|^3 \leqslant \varepsilon_n |Y_{n,k}|^2$. Thus,

$$\sum_{k=1}^{n} \mathbb{E}[|Y_{n,k}|^3] \leqslant \varepsilon_n \sum_{k=1}^{n} \mathrm{Var}[X_{n,k}]/b_n = \varepsilon_n,$$

and $\varepsilon_n \to 0$ in view of the assumptions on $b_n$ and $C_{n,k}$. Hence, Theorem 4.47 applies to the array $(Y_{n,k})$ to show that $(Z_n - a_n)/b_n = \sum_{k=1}^{n} Y_{n,k} \Rightarrow \Phi$. Besides, since $a_n \to a$ and $b_n \to b$, Theorem 3.49 shows that $(Z_n - a)/b \Rightarrow \Phi$. □

**5.2. Lindeberg's Central Limit Theorem.** The idea here is to replace the condition on third moments by something weaker, called **Lindeberg's condition**: For every $\varepsilon > 0$,

$$L_n(\varepsilon) = \sum_{k=1}^{n} \mathbb{E}\Big[|X_{n,k}|^2 \mathbb{1}_{\{|X_{n,k}|>\varepsilon\}}\Big] \to 0 \tag{4.21}$$

as $n \to \infty$. This being after the reduction assumptions that $\mathbb{E}[X_{n,k}] = 0$ and $\mathrm{Var}[S_n] = 1$. The general case can be derived using the Remark of Lyapunov's CLT.

THEOREM 4.51 (Lindeberg's CLT). *For all $n$ let $X_{n,k}$, $1 \leqslant k \leqslant n$, be independent. Suppose that $\mathbb{E}[X_{n,k}] = 0$ and $\mathrm{Var}[S_n] = 1$ for all $n$ and $1 \leqslant k \leqslant n$. If (4.21) holds then $S_n \Rightarrow \Phi$, where $\Phi$ is the standard normal distribution.*

PROOF. Assume (4.21). Then, for every $\varepsilon > 0$, there is an integer $m(\varepsilon)$ such that $L_n(\varepsilon) \leqslant \varepsilon^3$ for all $n \geqslant m(\varepsilon)$, and we may choose $m(\varepsilon)$ to be increasing to $+\infty$ as $\varepsilon$ decreases to 0. Choose $\varepsilon_n$ for each $n$ so that $m(\varepsilon_n) \leqslant n$ for all $n$ large enough. Then

$$\lim_n \Big(\frac{1}{\varepsilon_n}\Big)^2 L_n(\varepsilon_n) = 0. \tag{4.22}$$

Let $Y_{n,k} = X_{n,k}\mathbb{1}_{\{|X_{n,k}|\leqslant\varepsilon_n\}}$ and put $T_n = Y_{n,1} + \cdots + Y_{n,n}$. Then

$$\mathbb{P}[S_n \neq T_n] \leqslant \sum_{k=1}^{n} \mathbb{P}[X_{n,k} \neq Y_{n,k}] = \sum_{k=1}^{n} \mathbb{P}[|X_{n,k}| > \varepsilon] \leqslant \Big(\frac{1}{\varepsilon_n}\Big)^2 L_n(\varepsilon_n),$$

where the last inequality holds since $X^2\mathbb{1}_{\{|X|>\varepsilon\}} \geqslant \varepsilon^2\mathbb{1}_{\{|X|>\varepsilon\}}$ for arbitrary $\varepsilon$ and $X$. As $n \to \infty$, $\mathbb{P}[S_n \neq T_n] \to 0$ using Equation (4.22). To completes the proof, it is enough to show that $T_n \Rightarrow \Phi$ using the *insensitivity* property of convergence in distribution. But, since $|Y_{n,k}| \leqslant \varepsilon_n$ and $\varepsilon_n \to 0$, Corollary 4.50 implies that $T_n \Rightarrow \Phi$ once we show that

$$\mathbb{E}[T_n] \to 0, \quad \text{and} \quad \mathrm{Var}[T_n] \to 1. \tag{4.23}$$

To show (4.23), we estimate the mean and variance of $Y_{n,k}$ with all subscripts dropped.

Since $\mathbb{E}[X] = 0$, we have $\mathbb{E}[Y] = \mathbb{E}[Y] - \mathbb{E}[X] = -\mathbb{E}[X\mathbb{1}_{\{|X|>\varepsilon\}}]$, which yields

$$\big|\mathbb{E}[Y]\big| \leqslant \mathbb{E}[|X|\mathbb{1}_{\{|X|>\varepsilon\}}] \leqslant \frac{1}{\varepsilon}\mathbb{E}[X^2\mathbb{1}_{\{|X|>\varepsilon\}}]. \tag{4.24}$$

Second, using again the fact that $\mathbb{E}[Y] = -\mathbb{E}[X\mathbb{1}_{\{|X|>\varepsilon\}}]$, we have

$$\mathrm{Var}[Y] = \mathrm{Var}[X\mathbb{1}_{\{|X|\leqslant\varepsilon\}}] = \mathbb{E}[X^2\mathbb{1}_{\{|X|\leqslant\varepsilon\}}] - \big(\underbrace{\mathbb{E}[X\mathbb{1}_{\{|X|>\varepsilon\}}]}_{=-\mathbb{E}[Y]}\big)^2 = (\star).$$

Besides, using Cauchy-Schwarz inequality,

$$(\star) \geqslant \mathbb{E}[X^2\mathbb{1}_{\{|X|\leqslant\varepsilon\}}] - \mathbb{E}[X^2\mathbb{1}_{\{|X|>\varepsilon\}}] = \mathbb{E}[X^2] - 2\mathbb{E}[X^2\mathbb{1}_{\{|X|>\varepsilon\}}]. \tag{4.25}$$

Third, in the other direction,

$$\mathrm{Var}[X] = \mathbb{E}[X^2] \geqslant \mathbb{E}[Y^2] \geqslant \mathrm{Var}[Y]. \tag{4.26}$$

Finally, we put back the subscripts $n$ and $k$ in (4.24), (4.25) and (4.26), and sum them over $k$ to get (recall that $\mathrm{Var}[S_n] = 1$)

$$\big|\mathbb{E}[T_n]\big| \leqslant \frac{1}{\varepsilon_n} L_n(\varepsilon_n), \quad \text{and} \quad 1 - 2L_n(\varepsilon_n) \leqslant \mathrm{Var}[T_n] \leqslant 1, \tag{4.27}$$

where the first inequality is verified by (4.24); the second by (4.25) and the third by (4.26).

Now (4.22) and (4.27) together imply (4.23), and the proof is complete. □

We give several **remark** here:

- In order to ensuring that $S_n \Rightarrow \Phi$, Lindeberg's condition implies that the $X_{n,k}$ are uniformly small compared with $S_n$ for large $n$, that is

$$\lim_n \mathbb{P}\Big[\max_{1\leqslant k\leqslant n} |X_{n,k}| > \varepsilon\Big] = 0 \quad \text{for every } \varepsilon > 0. \tag{4.28}$$

To see this, first we claim that Lindeberg's condition implies that $\max_{1\leqslant k\leqslant n} \mathrm{Var}[X_{n,k}] \to 0$. This is obvious since

$$\begin{aligned}\mathrm{Var}[X_{n,k}] = \mathbb{E}[X_{n,k}^2] &= \mathbb{E}[X_{n,k}^2 \mathbb{1}_{\{X_{n,k}^2 \leqslant \varepsilon\}}] + \mathbb{E}[X_{n,k}^2 \mathbb{1}_{\{|X_{n,k}|>\varepsilon\}}] \\ &\leqslant \varepsilon^2 + \sum_{k=1}^n \mathbb{E}[X_{n,k}^2 \mathbb{1}_{\{|X_{n,k}|>\varepsilon\}}] \to \varepsilon^2 \quad \text{as } n \to \infty.\end{aligned}$$

This holds for any $\varepsilon > 0$, so $\lim_n \max_{1\leqslant k\leqslant n} \mathrm{Var}[X_{n,k}] \leqslant \varepsilon^2 \to 0$ as $\varepsilon \to 0$. Next, using Chebyshev's inequality, we get

$$\mathbb{P}[|X_{n,k}| > \varepsilon] \leqslant \frac{\mathrm{Var}[X_{n,k}]}{\varepsilon^2} \leqslant \frac{\max_{1\leqslant k\leqslant n} \mathrm{Var}[X_{n,k}]}{\varepsilon^2} \to 0$$

holds for all $1 \leqslant k \leqslant n$ as $n \to \infty$. So we get (4.28).

- It follows that Lindeberg's condition is not *necessary* for $S_n \Rightarrow \Phi$. For example, assuming $\mathbb{E}[S_n] = 0$ and $\mathrm{Var}[S_n] = 1$ as before, suppose that all the $X_{n,k}$ have Gaussian distributions and, in particular, $\mathrm{Var}[X_{n,1}] = 1/2$ for all $n$. In this case, $S_n$ has the same standard Gaussian distribution as $\Phi$, and $S_n \Rightarrow \Phi$ trivially, but (4.28) fails because of the chunk $X_{n,1}$ that does not get small with $n$.
- However, if we assume that $\mathbb{E}[X_{n,k}] = 0$ and $\mathrm{Var}[S_n] = 1$ as before, Lindeberg's condition is necessary and sufficient in order that $S_n \Rightarrow \Phi$ **and**

$$\lim_n \max_{1\leqslant k\leqslant n} \mathbb{P}[|X_{n,k}| > \varepsilon] = 0 \quad \text{for all } \varepsilon > 0. \tag{4.29}$$

Its sufficiency is immediate from the preceding theorem and the Remark above, since (4.28) implies (4.29). The proof of its necessity requires hard work, and we just omit it here.

- In general, we may suppose that $S_n$ has variance $\sigma_n^2$, and $\sigma_n^2 \to \sigma^2$. Then, Lindeberg's condition and CLT applies to the triangular array $(Z_{n,k})$ with $Z_{n,k} = X_{n,k}/\sigma_n$ shows that $M_n/\sigma_n \Rightarrow \Phi$, where $M_n = Z_{n,1} + \cdots + Z_{n,n}$. Since $\sigma_n \to \sigma$, so $M_n \Rightarrow \sigma\Phi \sim \mathcal{N}(0, \sigma^2)$ using Slutsky's theorem.

There are some other method to prove this theorem, you may refer **Durrett**'s book.

**5.3. Applications.** Finally, we still want to introduce some applications:

EXAMPLE 4.52 (Cycles in a random permutation and record values). Suppose $(Y_n)$ are independent Bernoulli variables with $\mathbb{P}[Y_n = 1] = 1/n$ and $\mathbb{P}[Y_n = 0] = 1 - 1/n$. Define $S_n = Y_1 + \cdots + Y_n$. Then we have

$$\frac{S_n - \log n}{\sqrt{\log n}} \Rightarrow \Phi.$$

PROOF. Since $\mathbb{E}[Y_n] = 1/n$ and $\mathrm{Var}[Y_n] = 1/n - 1/n^2$, we find $\mathbb{E}[S_n] \sim \log n$ and $\mathrm{Var}[S_n] \sim \log n$. For $1 \leqslant k \leqslant n$, define

$$X_{n,k} = \frac{Y_k - 1/k}{\sqrt{\log n}}.$$

Then we have $\mathbb{E}[X_{n,k}] = 0$, and $\sum_{k=1}^n \mathbb{E}[X_{n,k}^2] \to 1$, and for any $\varepsilon > 0$,

$$\lim_n \sum_{k=1}^n \mathbb{E}[X_{n,k}^2 \mathbb{1}_{\{|X_{n,k}|>\varepsilon\}}] = 0$$

since $|X_{n,k}| \leqslant 1/\sqrt{\log n}$. Then by Theorem 4.51, we obtain the conclusion. □

EXAMPLE 4.53 (Lyapunov's theorem). Suppose $(Y_n)$ are independent and define $S_n = Y_1 + \cdots + Y_n$. Define $\alpha_n = \sqrt{\mathrm{Var}[S_n]}$. If there is $\delta > 0$ such that

$$\lim_n \alpha_n^{-2-\delta} \sum_{k=1}^n \mathbb{E}\Big[|Y_k - \mathbb{E}[Y_k]|^{2+\delta}\Big] = 0,$$

then we have

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} \Rightarrow \Phi.$$

PROOF. For $1 \leqslant k \leqslant n$, define $X_{n,k} = (Y_k - \mathbb{E}[Y_k])/\alpha_n$. Then $\mathbb{E}[X_{n,k}] = 0$ and $\sum_{k=1}^n \mathbb{E}[X_{n,k}^2] = 1$. For any $\varepsilon > 0$, we have

$$\begin{aligned}
\frac{1}{\alpha_n^2} \sum_{k=1}^n \mathbb{E}\Big[X_{n,k}^2 \mathbb{1}_{\{|X_{n,k}|>\varepsilon\}}\Big] &= \sum_{k=1}^n \mathbb{E}\left[\frac{(Y_k - \mathbb{E}[Y_k])^2}{\alpha_n^2} \cdot 1 \cdot \mathbb{1}_{\{|Y_k - \mathbb{E}[Y_k]|>\varepsilon\alpha_n\}}\right] \\
&\leqslant \sum_{k=1}^n \mathbb{E}\left[\frac{(Y_k - \mathbb{E}[Y_k])^2}{\alpha_n^2} \cdot \left|\frac{Y_k - \mathbb{E}[Y_k]}{\varepsilon\alpha_n}\right|^{\delta} \cdot \mathbb{1}_{\{|Y_k - \mathbb{E}[Y_k]|>\varepsilon\alpha_n\}}\right] \\
&\leqslant \frac{1}{\varepsilon^\delta \alpha_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}\Big[|Y_k - \mathbb{E}[Y_k]|^{2+\delta}\Big] \to 0.
\end{aligned}$$

Thus we can apply Theorem 4.51 to $(X_{n,k})$ and we obtain the conclusion. □