Lecture Notes of STA347H1S

Ziteng Cheng

July 31, 2022

Contents

1 Axioms and Basic Properties of Probabilities

Let $\mathbb{X}$ and Ω be non-empty abstract spaces, with not special structure. We will use $\mathbb{X}$ and Ω interchangeably. In particular, we use Ω to emphasize it as the sample space. $A \subseteq \Omega$ is sometimes called an event. The elements in Ω is denoted by ω.

$2^{\mathbb{X}}$ is called the power set of $\mathbb{X}$, it is the set of all subset of $\mathbb{X}$. $A \subseteq \mathbb{X}$ and $A \in 2^{\mathbb{X}}$ share the same meaning.

def:sigmaAlg **Definition 1.1.** $\mathscr{A} \in 2^{\mathbb{X}}$ is a σ-algebra if

(i) $\emptyset \in \mathscr{A}$, $\Omega \in \mathscr{A}$;

(ii) [closed under complement] $A \in \mathscr{A} \implies A^c \in \mathscr{A}$;

(iii) [closed under countable union] $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{A} \implies \bigcup_{n\in\mathbb{N}} A_n \in \mathscr{A}$.

For $\mathscr{C} \subseteq 2^{\Omega}$, we write $\sigma(\mathscr{C})$ for the smallest σ-algebra containing $\mathscr{C}$.

rmk:SigmaAlg *Remark* 1.2. In view of Theorem A.2 (h), σ-algebra is also closed under intersection, that is, $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{A} \implies \bigcap_{n\in\mathbb{N}} A_n \in \mathscr{A}$. Intersection of σ-algebra is still a σ-algebra, but union is not.

Example 1.3. Here are some examples of σ-algebra:

1. $\{\emptyset, \Omega\}$ is the *trivial* σ-algebra.
2. 2^{Ω} is a σ-algebra.
3. $A \subsetneq \Omega$, then $\sigma(A) = \{\omega, A, A^c, \Omega\}$.
4. Let $\Omega = \{a, b, c, d\}$. $\mathscr{A} = \{\emptyset, \{a, b, c, d\}, \{a\}, \{b\}, \{c, d\}, \{a, b\}, \{a, c, d\}, \{b, c, d\}\}$ is a σ-algebra. Moreover, $\sigma(\{\{a\}, \{b\}, \{c, d\}\}) = \mathscr{A}$.

mk:MonotoneClass *Remark* 1.4. In general, σ-algebra is not quite convenient to describe via definition. An alternative description, which turns out to be more convenient in many cases, is monotone class theorem (cf. [A&B, Secion 4.4]).

def:BorelsigmaAlg **Definition 1.5.**

- We define the Borel σ-algebra on $\mathbb{R}^n$ as $\mathcal{B}(\mathbb{R}^n) := \sigma(\{B \subseteq \mathbb{R}^n : B \text{ is open}\})$. For the rest of the course, we *always* pair $\mathbb{R}^n$ with its Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$.
- The notion of Borel σ-algebra can be extended to, say, a metric space. If $\mathbb{X}$ is a metric space, then $\mathcal{B}(\mathbb{X}) := \sigma(\{B \subseteq \mathbb{X} : B \text{ is open}\})$.

m:SingletonInBorel **Lemma 1.6.** *Let $\mathbb{X}$ be a metric space with metric d. For any $x \in \mathbb{X}$, we have $\{x\} \in \mathcal{B}(\mathbb{X})$.*

Proof. For $r > 0$, we let $B_r(x) := \{x' \in \mathbb{X} : d(x, x') < r\}$, i.e., $B_r(x)$ is the open ball centered at x with radius r. Note that $\mathcal{B}(\mathbb{X})$ is also closed under countable intersection (cf. Remark 1.7), we conclude $\{x\} \in \mathcal{B}(\mathbb{X})$. □

mk:BorelSigmaAlg *Remark* 1.7. It can be shown by combining [A&B, Section 4.9, Theorem 4.44] and monotone class theorem (cf. Remark 1.4) that

$$\begin{aligned}
\mathcal{B}(\mathbb{R}^n) &= \sigma(\{A_1 \times \cdots \times A_n : A_k \in \mathcal{B}(\mathbb{R}), k = 1, ..., n\}) \\
&= \sigma(\{[a_1, b_1] \times \cdots \times [a_n, b_n] : a_k \le b_k \in \mathbb{R}, k = 1, ..., n\}) \\
&= \sigma(\{[a_1, b_1) \times \cdots \times [a_n, b_n) : a_k \le b_k, \in \mathbb{R}, k = 1, ..., n\}) \\
&= \sigma(\{(a_1, b_1] \times \cdots \times (a_n, b_n] : a_k \le b_k, \in \mathbb{R}, k = 1, ..., n\}) \\
&= \sigma(\{(a_1, b_1) \times \cdots \times (a_n, b_n) : a_k \le b_k, \in \mathbb{R}, k = 1, ..., n\}).
\end{aligned}$$

def:Measure **Definition 1.8.** $\mu : \mathscr{A} \to [0,\infty]$ is a measure if

(i) $\mu(\emptyset) = 0$;

(ii) [countable-additivity][1] for any $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{A}$ with $A_i \cap A_j = \emptyset, i \neq j$, we have $\mu(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$.

We say μ is a probability if $\mu(\Omega) = 1$. We usually use $\mathbb{P}$ to denote a probability. If $\mu(\mathbb{X}) = \infty$, we call μ an infinite measure.

rmk:Measure *Remark* 1.9.

1. Regarding the construction of a measure, we refer to the procedure called Carathéodory extension (cf. [D, Theorem 1.1.9], [B, Section 4.5] and [A&B, Section 10.23]).
2. Let μ and μ' be measures on $\sigma(\mathscr{C})$ and $\mu(A) = \mu'(A)$ for $A \in \mathscr{C}$, then $\mu(A) = \mu'(A)$ for $A \in \sigma(\mathscr{C})$. This can be proved by showing $\{A \in 2^\Omega : \mu(A) = \mu(A)\} = \sigma(\mathscr{C})$ using monotone class theorem (cf. Remark 1.4).

exmp:MeasureSp **Example 1.10.**

1. If $\mathbb{X}$ is finite or countable, we can easily construct a measure μ on $2^{\mathbb{X}}$ by assigning a non-negative number α_x to each $x \in \mathbb{X}$ and defining $\mu(A) := \sum_{x\in A} \alpha_x$ for $A \in 2^{\mathbb{X}}$. If $\sum_{x\in\mathbb{X}} \alpha_x = 1$, then μ is a probability.
2. A Dirac measure on x, denoted by δ_x, is a measure that satisfies
$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$
$(\mathbb{X}, \mathscr{X}, \delta_x)$ is a measure space.
3. Let $(\alpha_n)_{n\in\mathbb{N}} \subseteq [0,\infty]$ and $(x_n)_{n\in\mathbb{N}} \subseteq \mathbb{X}$. Then, $\mu(A) := \sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n}(A)$ is a measure on $(\mathbb{X}, \mathscr{X})$. Such μ is called discrete. If $\sum_{n\in\mathbb{N}} \alpha_n = 1$, then μ is a probability on $(\mathbb{X}, \mathscr{X})$.
4. Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Lebesgue measure on $([0,1], \mathcal{B}([0,1]))$ is a probability. See Section 5 for more discussion.

Definition 1.11. Let $\mathscr{A} \subseteq 2^{\mathbb{X}}$ be a σ-algebra and μ be a measure on $\mathscr{X}$.

- We call $(\mathbb{X}, \mathscr{X})$ a measurable space and $(\mathbb{X}, \mathscr{X}, \mu)$ a measure space. If μ is a probability, $(\mathbb{X}, \mathscr{X}, \mu)$ is called a probability space; we usually write $(\Omega, \mathscr{A}, \mathbb{P})$ for probability space.
- On a measure space $(\mathbb{X}, \mathscr{X}, \mu)$, we say N is a null set if there is $A \in \mathscr{X}$ such that $\mu(A) = 0$ and $N \subseteq A$. Note that N may note belong to $\mathscr{X}$.
- We say $(\mathbb{X}, \mathscr{X}, \mu)$ is a complete measure space if $\mathscr{X}$ contains all null sets, i.e., for all $A \in \mathscr{X}$ with $\mu(A) = 0$, we have $N \in \mathscr{A}$ as long as $N \subseteq A$.
- We say $A \in \mathscr{A}$ is true μ-almost surely, if A^c is a null set.

[1] This is also called σ-additivity.

Definition 1.12. Let $A \in 2^{\mathbb{X}}$. The indicator function $\mathbb{1}_A : \mathbb{X} \to \mathbb{R}$ is defined as

$$\mathbb{1}_A(x) := \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

When no confusion arise, we will omit x and simply write $\mathbb{1}_A$.

Lemma 1.13. *The indicator function has the following properties*

(a) $\mathbb{1}_{A\cap B} = \mathbb{1}_A \mathbb{1}_B$, *and in particular, if* $A \subseteq B$, $\mathbb{1}_A = \mathbb{1}_A \mathbb{1}_B$*;*

(b) if $A \cap B = \emptyset$, $\mathbb{1}_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B$, *and in particular,* $\mathbb{1}_A + \mathbb{1}_{A^c} = 1$.

Definition 1.14.

- Let $A, A_1, A_2, \dots \in 2^{\mathbb{X}}$. We say $(A_n)_{n\in\mathbb{N}}$ increases to A if $A_1 \subseteq A_2 \subseteq \dots$ and $\bigcup_{n\in\mathbb{N}} A_n = A$. We $(A_n)_{n\in\mathbb{N}}$ decreases to A if $A_1 \supseteq A_2 \supseteq \dots$ and $\bigcap_{n\in\mathbb{N}} A_n = A$.
- We say $(A_n)_{n\in\mathbb{N}}$ converges to A if $\lim_{n\to\infty} \mathbb{1}_{A_n}(x) = \mathbb{1}_A(x)$ for $x \in \mathbb{X}$. For abbreviation, we write $A_n \uparrow A$, $A_n \downarrow A$ and $\lim_{n\to\infty} A_n = A$, respectively.
- We also define

$$\limsup_{n\to\infty} A_n := \bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k \quad \text{and} \quad \liminf_{n\to\infty} A_n := \bigcup_{n\in\mathbb{N}} \bigcap_{k\geq n} A_k.$$

Note that for any $n, \ell \in \mathbb{N}$ we have $\bigcup_{k\geq n} A_k \supseteq \bigcap_{k\geq \ell} A_k$, and thus

$$\limsup_{n\to\infty} A_n \supseteq \liminf_{n\to\infty} A_n.$$

Remark 1.15. We have

$$\limsup_{n\to\infty} \mathbb{1}_{A_n}(x) = \mathbb{1}_{\limsup_{n\to\infty} A_n}(x) \quad \text{and} \quad \liminf_{n\to\infty} \mathbb{1}_{A_n}(x) = \mathbb{1}_{\liminf_{n\to\infty} A_n}(x), \quad x \in \mathbb{X}.$$

To see this, we first note that $\sup_{k\geq n} \mathbb{1}_{A_k} = \mathbb{1}_{\bigcup_{k\geq n} A_k}$. Note additionally that $\mathbb{1}_{\bigcup_{k\geq n} A_k}(x)$ is decreasing in n and bounded from below by 0, therefore $\lim_{n\to\infty} \mathbb{1}_{\bigcup_{k\geq n} A_k}(x)$ is well-defined. Next, suppose $x \in \mathbb{X}$ satisfies $\lim_{n\to\infty} \mathbb{1}_{\bigcup_{k\geq n} A_k}(x) = 1$, then there must be $N_x \in \mathbb{X}$ such that $\mathbb{1}_{\bigcup_{k\geq n} A_k}(x) = 1$ for $n \geq N_x$, and thus $x \in \bigcup_{k\geq n} A_k$ for $n \geq N_x$. Since $\bigcup_{k\geq n} A_k$ is decreasing in n we have $x \in \bigcup_{k\geq n} A_k$ for $n \in \mathbb{N}$, i.e., $x \in \limsup_{n\to\infty} A_n$. If $x \in \mathbb{X}$ satisfies $\lim_{n\to\infty} \mathbb{1}_{\bigcup_{k\geq 0} A_k}(x) = 0$, there must be $N_x \in \mathbb{X}$ such that $\mathbb{1}_{\bigcup_{k\geq n} A_k}(x) = 0$ for $n \geq N_x$, and thus $x \notin \limsup_{n\to\infty} A_n$.

The theorem below regards the basic properties of probability.

thm:MeasBasic **Theorem 1.16.** *Let* $(\mathbb{X}, \mathscr{X}, \mu)$ *be a measure space. Then, for any* $A, B, A_1, A_2, \dots \in \mathscr{X}$,

(a) $A \subseteq B \implies \mu(A) \leq \mu(B)$*;*

(b) $A \subseteq \bigcup_{n\in\mathbb{N}} A_n \implies \mu(A) \leq \sum_{n\in\mathbb{N}} \mu(A)$*;*

(c) $A_n \uparrow A \implies \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$*;*

(d) if $\mu = \mathbb{P}$ *is a probability, then* $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$*;*

(e) if $\mu = \mathbb{P}$ *is a probability,* $A_n \downarrow A \implies \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

Proof. **(a)** Note $B = A \cup (B \cap A^c)$ and $A \cap (B \cap A^c) = \emptyset$. Then, by Definition 1.1 (ii), $\mu(B) = \mu(A) + \mu(B \cap A^c) \geq \mu(A)$.
(b) Define $B_1 := A_1$ and $B_n := A_n \cap (\bigcup_{k=1}^{n-1} A_k)^c$. Note that $(B_n)_{n\in\mathbb{N}}$ are mutually disjoint. Additionally, $\bigcup_{k=1}^{n} A_k = \bigcup_{k=1}^{n} B_k$, and thus $\bigcup_{k\in\mathbb{N}} A_k = \bigcup_{k\in\mathbb{N}} B_k$. This together with statement (a) and Definition 1.8 (ii) implies that $\mu(A) \leq \mu(\bigcup_{n\in\mathbb{N}} A_n) = \mu(\bigcup_{n\in\mathbb{N}} B_n) = \sum_{n\in\mathbb{N}} \mu(B_n) \leq \sum_{n\in\mathbb{N}} \mu(A_n)$.
(c) Let $(B_n)_{n\in\mathbb{N}}$ be defined as above, and note $B_n = A_n \cap A_{n-1}^c$ for $n \geq 2$. It follows from Definition 1.8 (ii) that $\mu(A) = \sum_{n\in\mathbb{N}} \mu(B_n) = \lim_{m\to\infty} \sum_{n=1}^{m} \mu(B_n) = \lim_{m\to\infty} \mu(\bigcup_{n=1}^{m} B_n) = \lim_{m\to\infty} \mu(A_m)$.
(d)&(e) DIY. ☐

Example 1.17. This is a non-example for Theorem 1.16 (e) when μ is an infinite measure. On the measurable space $(\mathbb{N}, 2^{\mathbb{N}})$, we let μ be a counting measure, that is, $\mu(A)$ be the number of elements in A. It can be verified that μ indeed satisfies Definition 1.8. Let $A_n := \{n, n+1, \dots\}$. Then, $A_n \supset A_{n+1}$ and $\mu(A_n) = \infty$. On the other hand, note that $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$ and thus $\mu(\bigcap_{n\in\mathbb{N}} A_n) = 0$.

The next theorem regards the continuity of probability.

thm:ProbCont **Theorem 1.18** (Continuity of Probability)**.** *Let $\mathscr{A} \subseteq 2^\Omega$ be a σ-algebra. Suppose $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{A}$ and $\lim_{n\to\infty} A_n = A$. Then, $A \in \mathscr{A}$ and $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.*

Proof. In view of Definition 1.1, we have $\limsup_{n\to\infty} A_n \in \mathscr{A}$ and $\liminf_{n\to\infty} A_n \in \mathscr{A}$.

Next, note that by hypothesis, $\omega \in A$ if and only if there is $N_\omega \in \mathbb{N}$ such that $\omega \in A_n$ for any $n \geq N_\omega$ (why?). Therefore, $A \subseteq \limsup_{n\to\infty} A_n$ and $A \subseteq \liminf_{n\to\infty} A_n$. On the other hand, if $\omega \in \limsup_{n\to\infty} A_n$, then for any $n \in \mathbb{N}$, there exists $k \geq n$ such that $\omega \in A_k$. Note additionally that $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$. It follows from hypothesis that $\omega \in A$ (why?), and thus

$$A = \limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n, \tag{1.1}$$ eq:Alimsupliminf

which proves $A \in \mathscr{A}$.

In order to finish the proof, we let $B_n := \bigcap_{k\geq n} A_k$ and $C_n := \bigcup_{k\geq n} A_k$. Note that $B_n \uparrow A$ and $C_n \downarrow A$ due to (1.1). By Theorem 1.16 (c) (e), we yield $\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(C_n) = \lim_{n\to\infty} \mathbb{P}(B_n)$. However, we have $B_n \subseteq A_n \subseteq C_n$. It follows from Theorem 1.16 (a) that $\mathbb{P}(B_n) \leq \mathbb{P}(A_n) \leq \mathbb{P}(C_n)$. Finally, we conclude $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$. ☐

2 Random Variables

def:rv **Definition 2.1.** (i) Let $(\mathbb{X}, \mathscr{X})$ and $(\mathbb{Y}, \mathscr{Y})$ be two measurable spaces. We say a function $f : \mathbb{X} \to \mathbb{Y}$ is $\mathscr{X}$-$\mathscr{Y}$ measurable if $\{x \in \mathbb{X} : f(x) \in B\} \in \mathscr{X}$ for any $B \in \mathscr{Y}$, and we write $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$ for abbreviation. Sometimes it is convenient to write $f^{-1}(B) := \{x \in \mathbb{X} : f(x) \in B\}$. We also define $\sigma(f) := f^{-1}(\mathscr{Y})$, where we note $f^{-1}(\mathscr{Y})$ is a σ-algebra (why?).

(ii) If we set $\mathbb{X} = \Omega$ and consider $Y : (\Omega, \mathscr{A}) \to (\mathbb{Y}, \mathscr{Y})$, to emphasize that Y maps from the event space, we call Y an $\mathscr{A}$-$\mathscr{Y}$ random variable.

(iii) For $Y : (\Omega, \mathscr{A}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, we may call Y an $\mathbb{R}^n$-valued $\mathscr{A}$-random variable. If $n = 1$, we also call Y an real-valued $\mathscr{A}$-random variable. When no confusion arises, we simply call Y a real-valued random variable.

(iv) Let f and g be $\mathscr{X}$-$\mathcal{B}(\mathbb{R})$ measurable. We say $f = g$, μ-almost surely for $\mu(\{x \in \mathbb{X} : f(x) = g(x)\}^c) = 0$. We write $\mu - a.s.$ for abbreviation. For the rest of this course, unless specified otherwise, the '=' relationship between functions are understood in the almost sure sense. The same is true for '<', '>', '$\leq$' and '$\geq$'. When no confusion arise we will omit μ.

rmk:Measurable *Remark* 2.2. 1. All functions $f : \mathbb{X} \to \mathbb{Y}$ are $2^{\mathbb{X}}$-$\mathscr{Y}$ measurable, regardless of $\mathscr{Y}$. It is tempting to always use $2^{\mathbb{X}}$ when possible. But it turns out that $2^{\mathbb{X}}$ has some pathology when $\mathbb{X}$ is uncountable, say, $\mathbb{X} = \mathbb{R}$. We defer to Remark 5.2 for more discussion.

2. $f : \mathbb{X} \to \mathbb{Y}$ is $\mathscr{X}$-$\sigma(\mathscr{C})$ measurable if and only if $\{x \in \mathbb{X} : f(x) \in C\} \in \mathscr{X}$ for any $C \in \mathscr{C}$. The 'if' direction can be proved by showing $\{B \subseteq \mathbb{X} : f^{-1}(B) \in \mathscr{A}\} = \sigma(\mathscr{C})$ using monotone class theorem (cf. Remark 1.4). The 'only if' direction is clear from definition.

3. Composition preserves measurability. More precisely, consider $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$ and $g : (\mathbb{Y}, \mathscr{Y}) \to (\mathbb{Z}, \mathscr{Z})$, then the composition of g and f, defined as $g \circ f(a) := g(f(a))$, is $\mathscr{X}$-$\mathscr{Z}$ measurable.

4. Suppose $\mathbb{X}$ and $\mathbb{Y}$ are metric spaces. Then, any continuous $f : \mathbb{X} \to \mathbb{Y}$ is $\mathcal{B}(\mathbb{X})$-$\mathcal{B}(\mathbb{Y})$ measurable. This is a consequence of the fact that, $f : \mathbb{X} \to \mathbb{Y}$ is continuous if and only if $f^{-1}(U)$ is open for any open $U \subseteq \mathbb{Y}$.

5. Consider $X_k : (\Omega, \mathscr{A}) \to (\mathbb{X}_k, \mathscr{X}_k)$ for $k = 1, \dots, n$. Then, $(X_1, \dots, X_n)$ as a mapping from Ω to $\mathbb{X}^n$, is $\mathscr{A}$-$\mathscr{X}_1 \otimes \cdots \otimes \mathscr{X}_n$ measurable, where $\mathscr{X} \otimes \mathscr{Y} := \sigma(\{A \times B : A \in \mathscr{X}, B \in \mathscr{Y}\})$.

6. Suppose f and g are real-valued $\mathscr{X}$-measurable function. Then, so are cf (for $c \in \mathbb{R}$), $f + g$, fg, f/g (if $g \neq 0$), $\max\{f, g\}$ and $\min\{f, g\}$.

7. Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of real-valued $\mathscr{X}$-measurable functions. Using point 2, we can show that $\liminf_{n\to\infty} f_n(\omega) := \lim_{n\to\infty} \inf_{k \geq n} f_k(\omega)$ and $\limsup_{n\to\infty} f_n(\omega) := \lim_{n\to\infty} \sup_{k\geq n} f_k(\omega)$ are $\mathscr{X}$-$\mathcal{B}(\mathbb{R})$ measurable. Moreover, if $(f_n(\omega))_{n\in\mathbb{N}}$ converges as $n \to \infty$ for each $\omega \in \Omega$, then $f(\omega) := \lim_{n\to\infty} f_n(\omega)$ is $\mathscr{X}$-$\mathcal{B}(\mathbb{R})$ measurable.

8. For any $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$, $\sigma(f)$ is the smallest σ-algebra on $\mathbb{X}$ such that f is measurable, and we have $f : (\mathbb{X}, \sigma(f)) \to (\mathbb{Y}, \mathscr{Y})$. Moreover, if $\mathscr{Y} = \sigma(\mathscr{C})$, then $f^{-1}(\sigma(\mathscr{C})) = \sigma(f^{-1}(\mathscr{C}))$.

9. Let $\mathcal{I}$ be an uncountable set of indexes and consider $f_i : (\mathbb{X}, \mathscr{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $\sup_{i\in\mathscr{I}} f_i(x)$ may not be measurable (cf. ....).

The next lemma can be proved using element chasing method.

em:PreimageComm **Lemma 2.3.** *Let $f : \mathbb{X} \to \mathbb{Y}$, $B \in 2^{\mathbb{Y}}$ and $(B_i)_{i\in\mathcal{I}} \subseteq 2^{\mathbb{Y}}$, where $\mathcal{I}$ is a set of indexes (possibly uncountable). We have $f^{-1}(B^c) = (f^{-1}(B))^c$, $f^{-1}(\bigcap_{n\in\mathbb{N}} B_n) = \bigcap_{n\in\mathbb{N}} f^{-1}(B_n)$ and $f^{-1}(\bigcup_{n\in\mathbb{N}} B_n) = \bigcup_{n\in\mathbb{N}} f^{-1}(B)$.*

def:SimpleFunc **Definition 2.4.** A function $f : \mathbb{X} \to \mathbb{R}$ is called <u>simple</u> if it takes only finitely many value. In particular, it can be written as

$$f(x) = \sum_{k=1}^{n} r_k \mathbb{1}_{A_k}(x), \quad x \in \mathbb{X}, \tag{2.1}$$ eq:SimpleFuncRepr

for some $n \in \mathbb{N}$, distinct $r_1, \dots, r_n \in \mathbb{R}$ and $A_k = f^{-1}(\{r_k\})$ for $k = 1, \dots, n$. Clearly,

Lemma 2.5. *Let f be the simple function in (2.1). Then, $A_i \cap A_j = \emptyset$ for $i \neq j$. Moreover, $A_1, \dots, A_n \in \mathscr{X}$ if and only if f is $\mathscr{X}$-$\mathcal{B}(\mathbb{R})$ measurable.*

Proof. The first statement follows from Lemma 2.3 and the convention that $f^{-1}(\emptyset) = \emptyset$. Regarding the second statement, if f is measurable, in view of Lemma 1.6, $A_k = f^{-1}(\{r_k\}) \in \mathscr{X}$ due to Definition 2.1. If $A_1, \dots, A_n \in \mathscr{X}$, then for any $B \in \mathcal{B}(\mathbb{R})$, we have

$$f^{-1}(B) = f^{-1}\Big(\bigcup_{k=1,\dots,n;\, r_k \in B} \{r_k\}\Big) = \bigcup_{k=1,\dots,n;\, r_k \in B} f^{-1}(\{r_k\}) \in \mathscr{X},$$

where we have used Lemma 2.3 in the last inequality. $\square$

SimpleFuncApprox **Theorem 2.6** (Simple Function Approximation)**.** *For any $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, there is a sequence of simple functions $(f_n)_{n\in\mathbb{N}}$ such that f_n is $\mathscr{X}$-$\mathcal{B}(\mathbb{R})$ measurable, $|f_n(x)| \le |f(x)|$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for any $x \in \mathbb{X}$. In particular, we can construct f_n as*

$$\begin{aligned} f_n(x) = -n\mathbb{1}_{f^{-1}((-\infty,-n])}(x) &+ \sum_{k=-n2^n}^{-1} \frac{k+1}{2^n} \mathbb{1}_{f^{-1}((\frac{k}{2^n}, \frac{k+1}{2^n}])}(x) \\ &+ \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbb{1}_{f^{-1}([\frac{k}{2^n}, \frac{k+1}{2^n}))}(x) + n\mathbb{1}_{f^{-1}([n,\infty))}(x), \quad x \in \mathbb{X}. \end{aligned}$$

Moreover, if f is non-negative, then $f_n(x) \le f_{n+1}(x)$ for $x \in \mathbb{X}$.

thm:sigmagsigmaf **Theorem 2.7.** *Consider $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$ and $g : (\mathbb{X}, \mathscr{X}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then, $\sigma(g) \subseteq \sigma(f)$ if and only if there is $h : (\mathbb{Y}, \mathscr{Y}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $g = h \circ f$.*

Proof. Regarding the 'if' direction, we have $\sigma(g) = g^{-1}(\mathscr{Z}) = (h \circ f)^{-1}(\mathscr{Z}) \overset{(why?)}{=} f^{-1}(h^{-1}(\mathscr{Z}))$. Since $h^{-1}(\mathscr{Z}) \subseteq \mathscr{Y}$ due to the measurability of h, we conclude $\sigma(g) \subseteq \sigma(f)$.

Now we prove the 'only if' direction. We first assume g is simple and suppose $g(x) = \sum_{k=1}^{m} r_k \mathbb{1}_{A_k}(x)$ for some $r_1, \dots, r_n \in \mathbb{R}$ and $A_1, \dots, A_n \in \mathscr{X}$. Without loss of generality, we assume $r_i \neq r_j$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, where we note $A_k = g^{-1}(\{r_k\}) \in \sigma(g)$. Since $\sigma(g) \subseteq \sigma(f)$, we must have $A_k \in \sigma(f)$ for $k = 1, \dots, n$. If follows that there is B_k such that $f^{-1}(B_k) = A_k$ and $B_i \cap B_j = \emptyset$. Then, $h(x) := \sum_{k=1}^{m} r_k \mathbb{1}_{B_k}(x)$ is the desired.

Now we consider a generic g. There is a sequence of $\mathscr{X}$-measurable simple functions $(g_n)_{n\in\mathbb{N}}$ that approximates g in the sense of Theorem 2.6 and $\sigma(g_n) \subseteq \sigma(f)$. For each $n \in \mathbb{N}$, there is real-valued $\mathscr{Y}$-measurable h_n such that $g_n = h_n \circ f$. Let $L := \{y \in \mathbb{Y} : \liminf_{n\to\infty} h_n(y) = \limsup_{n\to\infty} h_n(y)\}$. Because $\lim_{n\to\infty} h_n(f(x)) = \lim_{n\to\infty} g_n(x) = g(x)$ for $x \in \mathbb{X}$, we have $f(X) \subseteq L$. Define $h(x) := \lim_{n\to\infty} h_n(x)\mathbb{1}_L(x)$, in view of Remark 2.2 (7), the proof is complete. $\square$

3 Distribution as Induced Measure

def:Distn **Definition 3.1.** Let $(\mathbb{X}, \mathscr{X}, \mu)$ be a measure space and consider $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$. For the rest of the course, we will use the following abbreviation/notation

$$\mu(\{x \in \mathbb{X} : f(x) \in B\}) = \mu(f \in B) = \mu(f^{-1}(B)) = \mu \circ f^{-1}(B) =: \mu^f(B), \quad B \in \mathscr{Y},$$

where we note for $B \notin \mathscr{Y}$ the left hand side does not make sense. μ^f is also called the measure induced by f. On a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, for $Y : (\Omega, \mathscr{A}) \to (\mathbb{Y}, \mathscr{Y})$, $\mathbb{P}^Y$ is called the (probabilistic) distribution of Y.

Using Lemma 2.3, we yield the result below.

Theorem 3.2. *μ^f is a measure on $(\mathbb{Y}, \mathscr{Y})$.*

Definition 3.3. Let $\mathbb{P}$ be a probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The cumulative distribution function (CDF) induced by $\mathbb{P}$ is defined as $F(r) := \mathbb{P}((-\infty, r])$. If $\mathbb{P} = \mathbb{P}^Y$ for some real-valued random variable Y, we call F the CDF of Y.

Remark 3.4. Using Remark 1.7 and Remark 1.9 (3), we can show that if two probability measure induces the same distribution function, then the two measures must coincides.

Remark 3.5. One important reason to adopt such framework is that it justifies the existence of continuous time random process, in terms of the result known as see Kolmogorov extension theorem (cf. [A&B, Section 15.6]).

thm:DistFunc **Theorem 3.6.** *Let F be a CDF on $\mathbb{R}$. Then,*

(a) F is non-decreasing;

(b) F is right-continuous on $\mathbb{R}$, that is, $\lim_{z \to r+} F(z) = F(r)$ for $r \in \mathbb{R}$;

(c) $\lim_{r \to -\infty} F(r) = 0$ and $\lim_{r \to \infty} F(r) = 1$;

(d) F has left limit on $\mathbb{R}$, that is, for any $r \in \mathbb{R}$ and $(r_n)_{n \in \mathbb{N}}$ increasing to r we have $(F(r_n))_{n \in \mathbb{N}}$ converges; additionally, $F(r-) := \lim_{z \to r-} F(z) = \mathbb{P}((-\infty, r))$;

(e) F has at most countably many jumps.

Proof. DIY. □

Remark 3.7. In view Remark 1.7, using Carathéodory extension theorem (cf. Remark 1.9(2)), we can show that a function F satisfying conditions (a) (b) (c) above characterizes a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

The result below is an immediate consequence of Theorem 3.6.

Corollary 3.8. *Let $\mathbb{P}$ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and F be the corresponding CDF. Then, for any real numbers $x < y$,*

(a) $\mathbb{P}((x, y]) = F(y) - F(x)$;

(b) $\mathbb{P}([x, y]) = F(y) - F(x-)$;

(c) $\mathbb{P}([x, y)) = F(y-) - F(x-)$;

(d) $\mathbb{P}((x, y)) = F(y-) - F(x-)$;

(e) $\mathbb{P}(\{x\}) = F(x) - F(x-)$.

4 Expectation as Lebesgue Integral

In what follows, we consider the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ with following rules $0 \times \infty = 0$, $0 \times (-\infty) = 0$, $a \pm \infty = \pm\infty$ and $a \times (\pm\infty) = \operatorname{sgn}(a) \cdot \infty$ for $a \in \mathbb{R}$.

Let $(\mathbb{X}, \mathscr{X}, \mu)$ be a measure space. We want to define an integral of a real-valued $\mathscr{X}$-measurable function with respect to μ. We first define the integral for simple random variable.

def:LebIntSimple **Definition 4.1.** Suppose f is a simple function of the form $f(x) = \sum_{k=1}^n r_k \mathbb{1}_{A_k}(x)$ with $r_k \in \mathbb{R}$ and $A_k \in \mathscr{X}$ for $k = 1, \dots, n$. We define

$$\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) := \sum_{k=1}^n r_k \mu(A_k).$$

Note, in particular, $\mu(A) = \int_{\mathbb{X}} \mathbb{1}_A(x)\mu(\mathrm{d}x)$ for $A \in \mathscr{X}$.

The lemma below argues that $\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x)$ is defined uniquely.

em:IntSimpleVerInv **Lemma 4.2.** *Suppose $f(\omega) = \sum_{k=1}^n r_k \mathbb{1}_{A_k}(\omega) = \sum_{k=1}^m \ell_k \mathbb{1}_{B_k}(\omega)$ for some $r_k, \ell_k \in \mathbb{R}$ and $A_k, B_k \in \mathscr{X}$ for $k = 1, \dots, n$. Then, $\sum_{k=1}^n r_k \mu(A_k) = \sum_{k=1}^m s_k \mu(B_k)$.*

Proof. DIY. □

def:LebIntPosX **Definition 4.3.** Suppose $f \geq 0$ (here f may take values in $[0, +\infty]$). We define

$$\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) := \sup\left\{\int_{\mathbb{X}} g(x)\mu(\mathrm{d}x) : g \text{ is simple real-valued } \mathscr{X}\text{-measurable function and } 0 \leq g \leq f\right\}.$$

def:LebInt **Definition 4.4.**

- Let f be a real-valued $\mathscr{X}$-measurable function. We write $f^+ := f\mathbb{1}_{\{f \geq 0\}}$ and $f^- := -f\mathbb{1}_{\{f<0\}}$. If $\int_{\mathbb{X}} f^+(x)\mu(\mathrm{d}x) < \infty$ or $\int_{\mathbb{X}} f^-(x)\mu(\mathrm{d}x) < \infty$, the Lebesgue integral (of f with respect to μ) is defined as

$$\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) := \int_{\mathbb{X}} f^+(x)\mu(\mathrm{d}x) - \int_{\mathbb{X}} f^-(x)\mu(\mathrm{d}x)$$

  We say f is integrable, if both $\int_{\mathbb{X}} f^+(x)\mu(\mathrm{d}x) < \infty$ and $\int_{\mathbb{X}} f^-(x)\mu(\mathrm{d}x) < \infty$, or equivalently, $\int_{\mathbb{X}} |f|(x)\mu(\mathrm{d}x) < \infty$.

- We use $\mathcal{L}^1(\mathbb{X}, \mathscr{X}, \mu)$ for the set of integrable functions. Furthermore, for $p \in (0, \infty)$, we let $\mathcal{L}^p(\mathbb{X}, \mathscr{X}, \mu)$ be the set of real-valued $\mathscr{X}$-measurable functions such that $\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x) < \infty$, and $\mathcal{L}^\infty(\mathbb{X}, \mathscr{X}, \mu)$ the set of real-valued $\mathscr{X}$-measurable functions such that $\mu(\{x : |f(x)| > M\}) = 0$ for some $M > 0$.

Remark 4.5. If $f = u + iv$ is a complex-valued function and $\int_{\mathbb{X}}(|u(x)| + |v(x)|)\mu(\mathrm{d}x)$ is finite, we define $\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) := \int_{\mathbb{X}} u(x)\mu(\mathrm{d}x) + i\int_{\mathbb{X}} v(x)\mu(\mathrm{d}x)$.

Definition 4.6. Let $A \in \mathscr{X}$. We write

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

Definition 4.7. Following Definition 4.4, set $(\mathbb{X}, \mathscr{X}, \mu) = (\Omega, \mathscr{A}, \mathbb{P})$ as a probability space, for a real-valued $\mathscr{A}$-random variable Y, the expectation of Y is defined as the Lebesgue integral,

$$\mathbb{E}^{\mathbb{P}}(Y) := \int_{\Omega} Y(\omega)\mathbb{P}(\mathrm{d}\omega).$$

When no confusion arise, we simply write $\mathbb{E}(Y)$.

The proposition below follows immediately from the definitions above.

prop:ExpnBasic **Proposition 4.8.** *Let $(\mathbb{X}, \mathscr{X}, \mu)$ be a measure space. Let f and g be real-valued $\mathscr{X}$-random variables. Then following is true:*

(a) if f and g are integrable and $f \le g$, then $\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) \le \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x)$;

(b) if f is integrable, then $\int_{\mathbb{X}} cf(x)\mu(\mathrm{d}x) = c\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x)$ for $c \in \mathbb{R}$;

(c) if $A \in \mathscr{X}$ satisfies $\mu(A) = 0$ and $f \ge 0$, then $\int_{\mathbb{X}} f(x)\mathbb{1}_A(x)\mu(\mathrm{d}x) = 0$.

The theorem below is one of the most important result concerning Lebesgue integral.

thm:PreMonoConv **Theorem 4.9** (Monotone Convergence)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of non-negative real-valued $\mathscr{X}$-measurable function such that $f_n' \ge f_n$ for $n' \ge n$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for $x \in \mathbb{X}$. Then,*

$$\lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x).$$

Proof. By Proposition 4.8 (b), $(\int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x))_{n\in\mathbb{N}}$ is an increasing sequence of real numbers and $\int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) \le \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x)$. Let L be the limit. We thus have $\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) \ge L$. What is left to prove is $\int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) \le L$. To this end let $g(x) = \sum_{k=1}^{\ell} r_k \mathbb{1}_{A_k}(x)$ be a simple function such that $g \le f$. Let $c \in (0,1)$ and $B_n = \{x \in \mathbb{X} : f_n(x) \ge cg(x)\}$. Note that $(B_n)_{n\in\mathbb{N}}$ increases to $\mathbb{X}$. It follows from Proposition 4.8 (c) that

$$L \ge \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) \ge \int_{B_n} f_n(x)\mu(\mathrm{d}x) \ge c\int_{B_n} g(x)\mu(\mathrm{d}x) = c\sum_{k=1}^{\ell} r_k\mu(A_k \cap B_n).$$

Note that $(A_k \cap B_n)_{n\in\mathbb{N}}$ increases to A_k. Applying Theorem 1.16 (c) to the right hand side above, we have $L \ge c\int_{\mathbb{X}} g(x)\mu(\mathrm{d}x)$. Since $c \in (0,1)$ is arbitrary, we have $L \ge \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x)$. In view of Definition 4.3, the proof is complete. $\square$

Thanks to monotone convergence theorem, we are now in position to establish the linearity of Lebesgue integral.

lem:IntSum **Lemma 4.10.** *Let f and g be real-valued non-negative $\mathscr{X}$-measurable functions. Then,*

$$\int_{\mathbb{X}} (f(x) + g(x))\mu(\mathrm{d}x) = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) + \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x).$$

Proof. First suppose f and g are simple, say, $f(x) = \sum_{k=1}^{m} r_k \mathbb{1}_{A_k}(x)$ and $g(x) = \sum_{k=1}^{n} s_k \mathbb{1}_{B_k}(x)$. Then, $(f+g)(x) = \sum_{k=1}^{m+n} t_k \mathbb{1}_{C_k}(x)$, where $t_k = r_k, C_k = A_k$ for $k = 1, \ldots, m$, and $t_k = s_{k-m}, C_k = B_{k-m}$ for $k = n+1, \ldots, m+n$. If follows that

$$\int_{\mathbb{X}} (f(x) + g(x))\mu(\mathrm{d}x) = \sum_{k=1}^{m+n} t_k \mu(C_k) = \sum_{k=1}^{m} r_k \mu(A_k) + \sum_{k=1}^{n} s_k \mu(B_k) = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) + \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x).$$

Next, we suppose f and g are non-negative. In view of Theorem 2.6, we let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be sequences of simple function increasing to f and g, respectively. Note that $(f_n + g_n)_{n\in\mathbb{N}}$ also increases to $f + g$. Invoking monotone convergence (Theorem 4.9), the proof is complete. □

The results above imply that the Lebesgue integral is a linear functional on $\mathcal{L}^1(\mathbb{X}, \mathscr{X}, \mu)$. We formulates such linearity into the theorem below.

thm:ExpnLinear **Theorem 4.11.** *Suppose $f, g \in \mathcal{L}^1(\mathbb{X}, \mathscr{X}, \mu)$. Then, for any $a, b \in \mathbb{R}$ we have*

$$\int_{\mathbb{X}} (af(x) + bg(x))\mu(\mathrm{d}x) = a\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) + b\int_{\mathbb{X}} g(x)\mu(\mathrm{d}x).$$

Proof. DIY. □

The next theorem is useful, as it is a vital tools for calculating Lebesgue integral. It shows in particular that the expectations of $g(Y)$ only depends on the distribution of Y.

thm:ExpnRule **Theorem 4.12.** *Consider $f : (\mathbb{X}, \mathscr{X}) \to (\mathbb{Y}, \mathscr{Y})$ and $g : (\mathbb{Y}, \mathscr{Y}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

(a) $g \circ f \in \mathcal{L}^1(\mathbb{X}, \mathscr{X}, \mu)$ if and only if $g \in \mathcal{L}^1(\mathbb{Y}, \mathscr{Y}, \mu^f)$;

(b) If either $g \geq 0$, or the equivalent conditions in (a) is satisfied, then

$$\int_{\mathbb{X}} g(f(x))\mu(\mathrm{d}x) = \int_{\mathbb{Y}} g(y)\mu^f(\mathrm{d}y).$$

Proof. Recall Definition 3.1 and 4.1. Then,

$$\int_{\mathbb{X}} \mathbb{1}_{f^{-1}(B)}(x)\mu(\mathrm{d}x) = \mu(f \in B) = \mu^f(B) = \int_B \mu^f(\mathrm{d}x).$$

This proves (b) for g being simple function. Suppose $g \geq 0$. In view of Theorem 2.6, we let $(g_n)_{n\in\mathbb{N}}$ be a sequence of simple $\mathscr{Y}$-measurable function increasing to g. Note $g_n \circ f$ also increases to $g \circ f$. By monotone convergence (Theorem 4.9),

$$\int_{\mathbb{X}} g \circ f(x)\mu(\mathrm{d}x) = \lim_{n\to\infty} \int_{\mathbb{X}} g_n \circ f(x)\mu(\mathrm{d}x) = \lim_{n\to\infty} \int_{\mathbb{Y}} g_n(y)\mu^f(\mathrm{d}y) = \int_{\mathbb{Y}} g(y)\mu^f(\mathrm{d}y),$$

This proves (b) for $g \geq 0$, and (a) follows immediately by substituting g above with $|g|$. Finally, for $g \in \mathcal{L}^1$, invoking the decomposition that $g = g^+ - g^-$ finishes the proof. □

Proposition 4.13. *We have $\mathcal{L}^p(\Omega, \mathscr{A}, \mathbb{P}) \supseteq \mathcal{L}^q(\Omega, \mathscr{A}, \mathbb{P})$ for $1 \leq p \leq q \leq \infty$.*

Proof. If $q = \infty$, there is $M > 0$ such that $\mathbb{P}(|Y| \le M) = 1$ and thus $\mathbb{P}(|Y|^p \le M) = 1$, which implies that $|Y|^p$ is integrable. Now suppose $q < \infty$. Let $Y \in \mathcal{L}^q$. Since $\mathbb{1}_{\{|Y|\le 1\}} + \mathbb{1}_{\{|Y|>1\}} = 1$, by Lemma 4.10,

$$\mathbb{E}(|Y|^p) = \mathbb{E}\left(|Y|^p \mathbb{1}_{\{|Y|\le 1\}}\right) + \mathbb{E}\left(|Y|^p \mathbb{1}_{\{|Y|>1\}}\right) \le 1 + \mathbb{E}\left(|Y|^q \mathbb{1}_{\{|Y|>1\}}\right) \le 1 + \mathbb{E}(|Y|^q) < \infty,$$

which implies $Y \in \mathcal{L}^p$ and thus completes the proof. □

Remark 4.14. The same is in general not true for infinite measure (why?).

5 Lebesgue Measure and Density Function

sec:LebMeas

def:LebMeasure **Definition 5.1.** A Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, denoted by λ^n, is a measure satisfying

$$\lambda([a_1, b_1] \times \cdots \times [a_n, b_n]) = (b_1 - a_1) \times \cdots \times (b_n - a_n), \quad a_k \le b_k,\ k = 1, \ldots, n.$$

rmk:LebMeas *Remark* 5.2.

1. In view of Remark 1.7 and 1.9 (3), we know Definition 5.1 defines a measure uniquely (if exists). The existence is a consequence of Carathéodory extension theorem (cf. Remark 1.9(2)). In fact, we can define Lebesgue measure for a σ-algebra larger than $\mathcal{B}(\mathbb{R})$, and such σ-algebra is call Lebesgue σ-algebra.
2. We wonder whether we can define Lebesgue measure for $2^{\mathbb{R}^n}$. This turns out to be not possible. A counter example on $\mathbb{R}$ is available at [B, Section 4.4].

Example 5.3. In this example, we show that on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $\lambda(a \times \mathbb{R} \ldots \mathbb{R}) = 0$. To this end let $A_{k,n} := [a - \frac{1}{k}, a + \frac{1}{k}] \times [-n, n] \times \cdots \times [-\ell, \ell]$, and $\lambda(A_{k,\ell}) = 2(2\ell)^n/k$. Invoking Theorem 1.16 (e), we have $\lambda((a \times [-\ell, \ell] \times [-\ell, \ell]) = 0$. It follows from Theorem 1.16 (c) that $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $\lambda(a \times \mathbb{R} \ldots \mathbb{R}) = 0$. A similar (but more tedious) argument shows that a hyperplane has zero Lebesgue measure.

The proposition below argue that Lebesgue integral extends Riemann integral.

Proposition 5.4. *Suppose $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is Riemann integrable on $[a_1, b_1] \times \cdots \times [a_n, b_n]$. Then, f is also Lebesgue integrable on the rectangle, and the Riemann integral coincides with the Lebesgue integral.*

Proof. See Section 7 of 'Lebesgue Integration on Euclidean Space' by Frank Jones. □

Remark 5.5. On the other hand, not every Lebesgue integral make sense as a Riemann integral. An example will be provided in HW.

From now on, we will omit λ from $\lambda(\mathrm{d}x)$ when writing Lebesgue integral with respect to Lebesgue measure. Note that for integral on $\mathbb{R}^n$ with $n > 1$, the dummy variable $x \in \mathbb{R}^n$ is a n-dimensional vector, i.e., $x = (x_1, \ldots, x_n)$. The following notations are equivalent,

$$\int_{\mathbb{R}^n} f(x) dx = \int_{\mathbb{R}^n} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

The notation on the right hand sides deserves more discussion in later sections on product measures and independence.

def:PDF **Definition 5.6.** Let μ be a measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. We say μ is absolutely continuous with respect to Lebesgue measure if there is a non-negative $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\mu(A) = \int_A f(x)\,\mathrm{d}x, \quad A \in \mathcal{B}(\mathbb{R}^n).$$

In this case, we call f the density function of μ. If $\mu = \mathbb{P}$ is a probability, we call f the probabilistic density function (PDF) of $\mathbb{P}$. If $\mu = \mathbb{P}^X$, we call f the PDF of X.

Remark 5.7.

1. The notion of absolute continuity between measures is studied in a boarder setup. Consider a measurable space $(\mathbb{X}, \mathscr{X})$. We say μ is absolutely continuous with respect to ν if for any $A \in \mathscr{X}$ with $\nu(A) = 0$ we have $\mu(A) = 0$. By Radon-Nikodym theorem, if μ is absolutely continuous with respect to ν, there is a real-valued $\mathscr{X}$-measurable f such that

$$\mu(A) = \int_A f(x)\nu(\mathrm{d}x), \quad A \in \mathscr{A},$$

   where the f is unique up to a set with 0 measure under ν and is called the Radon-Nikodym derivative. We refer to [B, Section 13] for the detailed statement and proof.

2. For a measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, it needs not to be absolutely continuous w.r.t. λ, and there may not exists a density function. In general, we have the following decomposition

$$\mu = \mu_D + \mu_C + \mu_S,$$

   where μ_D is a measure with atoms only, μ_C is a measure that is absolutely continuous w.r.t. λ, and μ_S is a measure with no atom but not absolutely continuous w.r.t. λ. We refer to .... for further discussion.

The next result is immediate from Definition 1.8 and 5.6.

Proposition 5.8. *Let f be a PDF of some $\mathbb{R}^n$-valued random variables Y, then*

$$f \geq 0,\ \lambda - a.s. \quad \text{and} \quad \int_{\mathbb{R}^n} f(x)dx = 1. \tag{5.1}$$ eq:PDF

Proof. We claim that for $k \in \mathbb{N}$ and $A_k := \{x \in \mathbb{R}^n : f(x) < -\frac{1}{k}\}$, we must have $\lambda(A_k) = 0$. Indeed, suppose otherwise, we have $\mathbb{P}^Y(A_k) = \int_{A_k} f(x)dx \leq -k^{-1}\lambda(A_k) < 0$, contradicting the hypothesis that $\mathbb{P}^Y$ is a probability. Let $A := \{x \in \mathbb{R}^n : f(x) < -0\}$. Note $(A_k)_{k\in\mathbb{N}}$ increases to A. By Theorem 1.16 (c), we have $\lambda(A) = 0$. Regarding $\int_{\mathbb{R}^n} f(x)dx = 1$, it follows immediately from Definition 5.6 and that $\mathbb{P}^Y(\mathbb{R}^n) = 1$. $\square$

Conversely, we can use f satisfying (5.1) to define a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Proposition 5.9. *Suppose $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies (5.1). Then,*

$$\mu(A) := \int_A f(x)\,\mathrm{d}x, \quad A \in \mathcal{B}(\mathbb{R}^n)$$

is a probability on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Proof. DIY. $\square$

From now on, we call f a PDF as long as f satisfies (5.1).

Below is a continuation of Theorem 4.12.

Theorem 5.10. *Let f be an $\mathbb{R}^n$-valued measurable function and $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose Y admits a PDF f. If $g \geq 0$, or $fg \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, then*

$$\mathbb{E}(g(Y)) = \int_{\mathbb{R}} g(r)f(r)\,\mathrm{d}r.$$

Proof. DIY. □

The proposition below provide an alternative expression for expectation with non-negative real-valued random variable.

prop:NonNegExpn **Proposition 5.11.** *Let F be the CDF of a non-negative real-valued random variable Y. Then,*

$$\mathbb{E}(Y) = \int_{\mathbb{R}_+} (1 - F(r))\,\mathrm{d}r,$$

where the right hand side is understood as a integral with respect to Lebesgue measure (see Definition 5.1).

Proof. DIY. □

Remark 5.12. In fact, a similar formula for random variable that is not necessarily non-negative is also possible. This can be easily proved with the notion of Lebesgue-Stieltjes integral. To heuristically derive an expression, we can assume Y is bounded and has a PDF, then we yield

$$\mathbb{E}(Y) = \int_{\mathbb{R}^+} (1 - F(r))\,\mathrm{d}r - \int_{\mathbb{R}^-} F(r)\,\mathrm{d}r.$$

6 Independence and Product Measures

def:Indep **Definition 6.1.** Consider the probability space $(\Omega, \mathscr{A}, \mathbb{P})$.

- Two event $A, A' \in \mathscr{A}$ are independent if $\mathbb{P}(A \cap A') = \mathbb{P}(A)\mathbb{P}(A')$. A sequence of events $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{A}$ is pairwise independent if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for $i \neq j$, is mutually independent if for any subset $\mathcal{I} \subseteq \mathbb{N}$ we have

$$\mathbb{P}(\bigcap_{n\in\mathcal{I}} A_n) = \prod_{n\in\mathcal{I}} \mathbb{P}(A_n).$$

- Two random variables $Y : (\Omega, \mathscr{A}) \to (\mathbb{Y}, \mathscr{Y})$ and $Z : (\Omega, \mathscr{A}) \to (\mathbb{Z}, \mathscr{Z})$ independent if $\{\omega \in \Omega : Y(\omega) \in B\}$ and $\{\omega \in \Omega : Z(\omega) \in C\}$ are independent for any $B \in \mathscr{Y}$ and $C \in \mathscr{Z}$. Equivalently, we write

$$\mathbb{P}(Y \in B, Z \in C) = \mathbb{P}(Y \in B)\mathbb{P}(Z \in C), \quad B \in \mathscr{Y}, C \in \mathscr{Z}.$$

  A sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ is pairwise independent if Y_i and Y_j are independent for any $i \neq j$, is mutually independent if for any $\mathcal{I} \subseteq \mathbb{N}$

$$\mathbb{P}(\bigcap_{n\in\mathcal{I}} \{Y_n \in B_n\}) = \prod_{n\in\mathcal{I}} \mathbb{P}(Y_n \in B_n), \quad B_n \in \mathscr{Y}_n,\, n \in \mathcal{I}.$$

thm:Indp **Theorem 6.2.** *Let Y and Z be real-valued random variables. Then, Y and Z are independent, if and only if $\mathbb{E}(f(Y)g(Z)) = \mathbb{E}(f(Y))\mathbb{E}(g(Z))$ for any non-negative $f : (\mathbb{Y}, \mathscr{Y}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $g : (\mathbb{Z}, \mathscr{Z}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

Proof. The 'if' direction is immediately when we take f and g as indicators. Regarding the 'only if' direction, an application of simple function approximation and monotone converge finishes the proof. □

Remark 6.3. If we replace the f and g above by bounded measurable functions, the theorem is still true. But be careful when dealing integrable functions in a similar setting, as the product of integrable functions need not be integrable.

Remark 6.4. Suppose $\mathbb{Y}$ and $\mathbb{Z}$ are metric spaces endowed with the corresponding Borel σ-algebra. For Y and Z to be independent, it is sufficient to have $\mathbb{E}(f(Y)g(Z)) = \mathbb{E}(f(Y))\mathbb{E}(g(Z))$ for any bounded continuous f and g. The proof of this statement involves more delicate treatment on the related σ-algebra. We refer to ....

The following technical result will be useful later.

lem:BorelCantelli **Lemma 6.5** (Borel-Cantelli)**.** *On $(\Omega, \mathscr{A}, \mathbb{P})$, let $(A_n)_{n\in\mathbb{N}} \subseteq A_n$. The following is true:*

(a) if $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k) = 0$;

(b) if $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k) = 0$ and $(A_n)_{n\in\mathbb{N}}$ is mutually independent, we have $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$.

Proof. **(a)** By Theorem 1.16 (a) (b),

$$\mathbb{P}(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k) \leq \mathbb{P}(\bigcup_{k\geq m} A_k) \leq \sum_{k\geq m} \mathbb{P}(A_k), \quad m \in \mathbb{N}.$$

By the hypothesis that $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$, the right hand side above tends to 0 as $m \to \infty$. The proof is complete.

(b) Note that, by Theorem 1.16 (c) (e) (d),

$$\begin{aligned}\mathbb{P}(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k) &= \lim_{n\to\infty} \mathbb{P}(\bigcup_{k\geq n} A_k) = \lim_{n\to\infty} \lim_{m\to\infty} \mathbb{P}(\bigcup_{k=1}^{m} A_k) = \lim_{n\to\infty} \lim_{m\to\infty} (1 - \mathbb{P}(\bigcap_{k=1}^{m} A_k^c)) \\ &= \lim_{n\to\infty} \lim_{m\to\infty} (1 - \prod_{k=1}^{m} \mathbb{P}(A_k^c)) = 1 - \lim_{n\to\infty} \lim_{m\to\infty} \prod_{k=1}^{m} (1 - \mathbb{P}(A_k)).\end{aligned}$$

This together with the hypothesis that $\mathbb{P}(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k) = 0$, we have

$$\lim_{n\to\infty} \lim_{m\to\infty} \prod_{k=1}^{m} (1 - \mathbb{P}(A_k)) = 1.$$

By taking log, we have

$$\lim_{n\to\infty} \lim_{m\to\infty} \sum_{k=n}^{m} \log(1 - \mathbb{P}(A_k)) = \lim_{n\to\infty} \sum_{k\geq n} \log(1 - \mathbb{P}(A_k)) = 0.$$

It follows that $\sum_{k\in\mathbb{N}} \log(1 - \mathbb{P}(A_k))$ is a converging sum with non-positive summands. Because $|\log(1 - z)| \geq z$ for $z \in [0, 1]$, we conclude to proof. □

In view of Definition 3.1 and 6.1, for independent Y, Z and non-negative $B \in \mathscr{Y}, C \in \mathscr{Z}$, we have $\mathbb{P}^{(Y,Z)}(B \times C) = \mathbb{P}^{Y}(B)\,\mathbb{P}^{Z}(C)$. This motivates the following notions of product measures.

def:ProdMeas **Definition 6.6.** Consider two measurable spaces $(\mathbb{X}, \mathscr{X})$ and $(\mathbb{S}, \mathscr{S})$. Let μ and ν be measures on $(\mathbb{X}, \mathscr{X})$ and $(\mathbb{S}, \mathscr{S})$, respectively.

- The (Cartesian) product of sets A and B is defined as $A \times B := \{(a, b) : a \in A, b \in B\}$. In particular, $\mathbb{X} \times \mathbb{S} = \{(x, s) : x \in \mathbb{X}, s \in \mathbb{S}\}$.
- The product σ-algebra of $\mathscr{X}$ and $\mathscr{S}$ is defined as $\mathscr{X} \otimes \mathscr{S} := \sigma(\{A \times B : A \in \mathscr{X}, B \in \mathscr{S}\})$.
- The product measure of μ and ν, denoted by $\mu \otimes \nu$, is defined as the measure on $(\mathbb{X}\times\mathbb{S}, \mathscr{X}\otimes\mathscr{S})$ that satisfies $\mu \otimes \nu(A \times B) = \mu(A) \times \nu(B)$ for any $A \in \mathscr{X}$ and $B \in \mathscr{S}$.

Remark 6.7.

1. One way to establish the existence of product measure is to use Carathéodory extension theorem (cf. Remark 1.9). Alternatively, we can also define $\mu\otimes\nu(C)$ as $\int_{\mathbb{Y}} \mu(C_s)\nu(\mathrm{d}s)$ or $\int_{\mathbb{X}} \nu(C_x)\mu(\mathrm{d}x)$ for $C \in \mathscr{X} \otimes \mathscr{S}$, where $C_s := \{x \in \mathbb{X} : (x, s) \in C\}$ and $C_x := \{s \in \mathbb{S} : (x, y) \in C\}$. Note that it is not trivial to justify the above-mentioned definition.
2. We can also separately show the uniqueness using monotone class theorem (cf. Remark 1.4), in case some versions of Carathéodory extension theorem does not cover the uniqueness.

op:IndepProdMeas **Proposition 6.8.** *On $(\Omega, \mathscr{A}, \mathbb{P})$, consider two independent random variables $Y : (\Omega, \mathscr{A}) \to (\mathbb{Y}, \mathscr{Y})$ and $Z : (\Omega, \mathscr{A}) \to (\mathbb{Z}, \mathscr{Z})$. Then, $\mathbb{P}^{(Y,Z)} = \mathbb{P}^{Y} \otimes \mathbb{P}^{Z}$.*

Proof. The proof of this result is out of the scope of this course. It is mainly based on monotone class theorem (cf. Remark 1.4). ☐

prop:LebesgueRn **Proposition 6.9.** *Let λ be the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and λ^n be the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Then, $\mathcal{B}(\mathbb{R}^n) = \mathcal{B}(\mathbb{R})^{\otimes n}$ and $\lambda^n = \lambda^{\otimes n}$.*

Proof. The proof of this result is out of the scope of this course. It is mainly based on Remark 1.7 and monotone class theorem (cf. Remark 1.4). ☐

The following results allow us to interchange the order of integration. The proof, mostly based on monotone class theorem (cf. Remark 1.4), is out of the scope of this course.

:SectionMeasurable **Lemma 6.10.** *Consider $f : (\mathbb{X} \times \mathbb{S}, \mathscr{X} \otimes \mathscr{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let μ and ν be measures on $(\mathbb{X}, \mathscr{X})$ and $(\mathbb{S}, \mathscr{S})$, respectively. The following is true*

(a) for any $x \in \mathbb{X}$, $s \mapsto f(x, s)$ and $s \mapsto \int_{\mathbb{X}} f(x, s)\mu(\mathrm{d}x)$ are $\mathscr{S}$-measurable;

(b) for any $s \in \mathbb{S}$, $x \mapsto f(x, v)$ and $x \mapsto \int_{\mathbb{S}} f(x, s)\nu(\mathrm{d}s)$ are $\mathscr{X}$-measurable.

thm:Fubini **Theorem 6.11** (Fubini-Tonelli)**.** *Suppose $f : (\mathbb{X} \times \mathbb{S}, \mathscr{X} \otimes \mathscr{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies either $f \geq 0$, or, $\int_{\mathbb{X}\times\mathbb{S}} |f(z)|\mu \otimes \nu(\mathrm{d}z) < \infty$ (i.e., $f \in \mathcal{L}^1(\mathbb{X} \times \mathbb{S}, \mathscr{X} \otimes \mathscr{S}, \mu \otimes \nu)$). Then,*

$$\int_{\mathbb{X}\times\mathbb{S}} f(z)\mu \otimes \nu(\mathrm{d}z) = \int_{\mathbb{S}}\int_{\mathbb{X}} f(x, s)\mu(\mathrm{d}x)\nu(\mathrm{d}s) = \int_{\mathbb{X}}\int_{\mathbb{S}} f(x, s)\nu(\mathrm{d}s)\mu(\mathrm{d}x).$$

We note that the analogue for multi-variate f is also true.

Example 6.12. Here is an non-example of Fubini-Tonelli theorem. Let $\mathbb{X} = \mathbb{S} = [0,1]$, both endowed with Lebesgue measure. Let $g_k(z) := (\frac{1}{k} - \frac{1}{k+1})^{-1}\mathbb{1}_{(\frac{1}{k+1},\frac{1}{k})}(z)$ and note $\int_{[0,1]} g_k(z)\,\mathrm{d}z = 1$. We also define

$$f(x,s) = \sum_{k=1}^{\infty} (g_k(x) - g_{k+1}(x))\, g_k(s), \quad (x,s) \in [0,1]^2.$$

Note for each (x,s), at most two terms is non-zero, f is not single signed and $f \notin \mathcal{L}^1$. Note, in addition, that

$$\int_{[0,1]}\int_{[0,1]} f(x,s)\,\mathrm{d}x\,\mathrm{d}s = \int_{[0,1]}\sum_{k=1} 0\cdot g_k(s)\,\mathrm{d}s = 0$$

and

$$\int_{[0,1]}\int_{[0,1]} f(x,s)\,\mathrm{d}s\,\mathrm{d}x = \int_{[0,1]}\sum_{k=1}^{\infty} (g_k(x) - g_{k+1}(x))\,\mathrm{d}x = \int_{[0,1]} g_1(x)\,\mathrm{d}x = \frac{1}{2}.$$

As a consequence of the results above, we yield that the joint PDF of independent random variables is the product of the individual PDFs (if exists).

Corollary 6.13. *Let Y be a $\mathbb{R}^n$-valued random variable and Z be a $\mathbb{R}^d$-valued random variables. Suppose Y and Z are independent, and both have PDFs f_Y and f_Z, respectively. Then, (Y,Z) as an $\mathbb{R}^{n+d}$-valued random variable has density $f_{(Y,Z)}(y,z) = f_Y(y)f_Z(z)$ for $(y,z) \in \mathbb{R}^n \times \mathbb{R}^d$. Moreover, for any $B \in \mathcal{B}(\mathbb{R}^{n+k})$, we have*

$$\mathbb{P}((Y,Z)\in B) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^d} \mathbb{1}_B(y,z)f(y)f(z)\,\mathrm{d}z\,\mathrm{d}y = \int_{\mathbb{R}^d}\int_{\mathbb{R}^n} \mathbb{1}_B(y,z)f(y)f(z)\,\mathrm{d}y\,\mathrm{d}z,$$

Proof. In view of Proposition 6.8, we have

$$\mathbb{P}^{(Y,Z)}(B) = \int_{\mathbb{R}^{n+d}} \mathbb{1}_B(r)\mathbb{P}^{(Y,Z)}(\mathrm{d}r) = \int_{\mathbb{R}^{n+d}} \mathbb{1}_B(r)\mathbb{P}^Y\otimes\mathbb{P}^Z(\mathrm{d}r).$$

Then, by Fubini-Tonelli theorem (Theorem 6.11),

$$\mathbb{P}^{(Y,Z)}(B) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^d} \mathbb{1}_B(y,z)\mathbb{P}^Z(\mathrm{d}z)\,\mathbb{P}^Y(\mathrm{d}y) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^n} \mathbb{1}_B(y,z)\mathbb{P}^Y(\mathrm{d}y)\mathbb{P}^Z(\mathrm{d}z).$$

Invoking the hypothesis that both Y and Z have PDFs and Lemma 6.10, we yield

$$\begin{aligned}\mathbb{P}^{(Y,Z)}(B) &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^d} \mathbb{1}_B(y,z)f_Z(z)\,\mathrm{d}z\mathbb{P}^Y(\mathrm{d}y) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^d} \mathbb{1}_B(y,z)f_Z(z)f_Y(y)\,\mathrm{d}z\ \mathrm{d}y\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^n} \mathbb{1}_B(y,z)f_Y(y)\,\mathrm{d}y\mathbb{P}^Z(\mathrm{d}z) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^n} \mathbb{1}_B(y,z)f_Y(y)f_Z(z)\,\mathrm{d}y\ \mathrm{d}z.\end{aligned}$$

Finally, by Fubini-Tonelli theorem (Theorem 6.11) and Proposition 6.9, we conclude

$$\mathbb{P}^{(Y,Z)}(B) = \int_{\mathbb{R}^{n+d}} \mathbb{1}_B(r)f_{(Y,Z)}(r)\,\mathrm{d}r.$$

$\square$

7 Change of Variables

To illustrate the well-known result called Jacobi's transformation formula, we consider the heuristic argument below on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \lambda^2)$. Let $T \in \mathbb{R}^{2\times 2}$ be a matrix, we treat $v \in \mathbb{R}^2$ as a column vector. Let $B \in \mathcal{B}(\mathbb{R}^2)$, we define $TB := \{Tv : v \in B\}$. We wonder what is $\lambda^2(Tv)$. By the theory of linear algebra, we know that every T arises as a product of elementary matrices: (1) permutation matrix, denoted by T_1^e; (2) summing one row onto the other row, denoted by T_2^e; (3) row stretch, denoted by T_3^e. Note that (1) and (2) do not affect the volume, and $|\det T_i^e| = 1$ for $i = 1, 2$. Only row stretch scale the volume by $|\det T_3^e|$. This together with the fact that $\det TT' = \det T \det T'$, we yield that $\lambda^2(TB) = |\det T|\lambda^2(B)$. Next, let $g : G \subseteq \mathbb{R}^2 \to \mathbb{R}^2$ be one-to-one and continuously differentiable. Additionally, in view of Taylor expansion, note that

$$g(u + \delta u) = g(u) + \begin{pmatrix} \frac{\partial g_1}{\partial r_1}(u) & \frac{\partial g_1}{\partial r_2}(u) \\ \frac{\partial g_2}{\partial r_1}(u) & \frac{\partial g_2}{\partial r_2}(u) \end{pmatrix} \delta u + o(|\delta u|) =: g(u) + J(u)\delta u + o(|\delta u|).$$

The above heuristically leads to the formula that

$$\int_{g(G)} \mathbb{1}_B(g^{-1}(r))\,\mathrm{d}r = \int_{g(G)} \mathbb{1}_{g(B)}(r)\,\mathrm{d}r = \int_G \mathbb{1}_B(r)|\det J(r)|\,\mathrm{d}r, \quad B \subseteq G,\, B \in \mathcal{B}(\mathbb{R}^n).$$

where $g(B) := \{v \in \mathbb{R}^2 : v = g(r) \text{ for some } r \in B\}$. Following the idea of simple function approximation, we yield

$$\int_{g(G)} f(g^{-1}(r))\,\mathrm{d}r = \int_G f(r)|\det J(r)|\,\mathrm{d}r.$$

Since g^{-1} is one-to-one, it is sometimes more convenient to change $f \circ g^{-1}$ above to h,

$$\int_{g(G)} h(r)\,\mathrm{d}r = \int_G h(g(r))|\det J(r)|\,\mathrm{d}r.$$

Below we officially introduce the Jacobi's transformation formula.

Definition 7.1. Let $G \subseteq \mathbb{R}^n$ be open and $g : G \to \mathbb{R}^n$ be continuously differentiable. The <u>Jacobian matrix</u> of g is $J_g : G \to \mathbb{R}^{n\times n}$ with (i,j)-entry being $[J_g(u)]_{i,j} := \frac{\partial g_i}{\partial r_j}(u)$.

thm:JacobiTrans **Theorem 7.2** (Jacobi's transformation formula)**.** *Let $G \subseteq \mathbb{R}^n$ be open. Suppose $g : G \to \mathbb{R}^n$ is one-to-one, continuously differentiable on open $D \in \mathcal{B}(\mathbb{R}^n)$ and $\lambda^n(G \setminus D) = 0$. Then, for any $h : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that is non-negative or integrable, we have*

$$\int_{g(G)} h(r)\,\mathrm{d}r = \int_G h(g(r))|\det J_g(r)|\,\mathrm{d}r.$$

Proof. See Theorem 7.26, W. Rudin (1987) Real and Complex Analysis. □

The next theorem is an application of Jacobi's transformation formula (Theorem 7.2).

thm:CoVPDF **Theorem 7.3.** *Let Y be an $\mathbb{R}^n$-valued random variable with PDF f_Y and $\beta : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Suppose β is one-to-one and $|\det J_\beta(r)| > 0$ (equivalently, $J_\beta(r)$ is invertible) for $r \in \mathbb{R}^n$. Then, $Z = \beta(Y)$ also has a PDF f_Z, and*

$$f_Z(z) = \mathbb{1}_{\beta(\mathbb{R}^n)}(z) f_Y(\beta^{-1}(z))|\det J_{\beta^{-1}}(z)|, \quad z \in \mathbb{R}^n.$$

Proof. Because β has positive Jacobian matrix, $G := \beta(\mathbb{R}^n)$ is open (long story). Moreover, because β is one-to-one, we have $g := \beta^{-1}$ is well-defined on G. We also have $J_g(r)$ is well-defined and equals to the inverse of $J_\beta(g(r))$ for $r \in G$, due to inverse function theorem (cf. [J. Shurman, *Multivariable Calculus*, Theorem 5.2.1]). It follows that $|\det J_g(r)| < \infty$ for $r \in G$. Then, for $B \in \mathcal{B}(\mathbb{R}^n)$, we have

$$\begin{aligned}\mathbb{P}(Z \in B) &= \mathbb{P}(\beta(Y) \in B) = \mathbb{P}(\beta(Y) \in B \cap G) = \mathbb{P}(Y \in \beta^{-1}(B \cap G)) = \mathbb{P}(Y \in g(B \cap G)) \\ &= \int_{g(G)} \mathbb{1}_{g(B\cap G)}(y) f_Y(y)\,\mathrm{d}y.\end{aligned}$$

By Theorem 7.2,

$$\mathbb{P}(Z \in B) = \int_G \mathbb{1}_{g(B\cap G)}(g(r)) f_Y(g(r)) |\det J_g(r)|\,\mathrm{d}r = \int_B \mathbb{1}_G(r) f_Y(g(r)) |\det J_g(r)|\,\mathrm{d}r,$$

which concludes the proof. $\square$

The following is helpful in case β is not one-to-one.

Corollary 7.4. *Let Y be an $\mathbb{R}^n$-valued random variable with PDF f_Y and $\beta : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Let $(S_k)_{k\in\mathbb{N}_0} \in \mathcal{B}(\mathbb{R}^n)$ be a partition of $\mathbb{R}^n$, that is, $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\mathbb{R}^n = \bigcup_{n\in\mathbb{N}_0}$. Suppose additionally that $\lambda^n(S_0) = 0$; and for $k \in \mathbb{N}$, S_k are open, $\beta_k : S_k \to \mathbb{R}^n$ is one-to-one, continuously differentiable $|\det J_\beta(r)| > 0$ for $r \in S_k$, and $\beta(r) = \mathbb{1}_{S_0}(r)\beta(r) + \sum_{k=1}^\infty \beta_k(r)\mathbb{1}_{S_k}(r)$ for $r \in \mathbb{R}^n$. Then, $Z = \beta(Y)$ has a PDF f_Z, and*

$$f_Z(z) = \sum_{k\in\mathbb{N}} \mathbb{1}_{\beta_k(S_k)}(z) f_Y(\beta_k^{-1}(z)) |\det J_{\beta_k^{-1}}(z)|, \quad z \in \mathbb{R}^n.$$

8 Selections of Inequalities

thm:ChebyshevIneq **Theorem 8.1** (Chebyshev's Inequality)**.** *For $p \in [1,\infty)$ and $a > 0$, we have*

$$\mu(\{x \in \mathbb{X} : |f(x)| \geq a\}) \leq \frac{1}{a^p} \int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x).$$

Proof. This is an immediate consequence of the observation that, for $z \in \{x \in \mathbb{X} : |f(x)| \geq a\}$ we have $|f(z)|^p/a^p \geq 1$. $\square$

cor:ChernoffIneq **Corollary 8.2** (Chernoff's Inequality)**.** *Let $t > 0$. Then,*

$$\mathbb{P}(Y \geq \mathbb{E}(Y) + t) \leq e^{-\lambda t}\mathbb{E}\left(e^{\lambda(Y-\mathbb{E}(Y))}\right) e^{-\lambda t}, \quad \lambda > 0.$$

Proof. Note that for $\lambda > 0$,

$$\mathbb{P}(Y \geq \mathbb{E}(Y) + t) = \mathbb{P}\left(e^{\lambda(Y-\mathbb{E}(Y))} \geq e^{\lambda t}\right) \leq \mathbb{E}\left(e^{\lambda(Y-\mathbb{E}(Y))}\right) e^{-\lambda t}$$

due to Chebyshev's Inequality (Theorem 8.1) with $p = 1$. $\square$

thm:HoeffdingIneq **Theorem 8.3** (Hoeffding's Inequality)**.** *Suppose $Y_1, \ldots, Y_n$ are mutually independent real-valued random variables, and $a_k \leq Y_k \leq b_k$ for $k = 1, \ldots, n$. Then, for $\varepsilon \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{k=1}^n Y_k - \sum_{k=1}^n \mathbb{E}(Y_k)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{k=1}^n (b_k - a_k)^2}\right).$$

Proof. The case of $\varepsilon = 0$ is obvious. Let $\varepsilon > 0$. We first apply Chernoff's inequality (Corollary 8.2) to yield

$$\mathbb{P}\left(\sum_{k=1}^{n} Y_k - \sum_{k=1}^{n} \mathbb{E}(Y_k) \geq \varepsilon\right) \leq e^{-\lambda\varepsilon}\mathbb{E}\left(e^{\lambda\sum_{k=1}^{n}(Y_k-\mathbb{E}(Y_k))}\right) \leq e^{-\lambda\varepsilon}\prod_{k=1}^{n}\mathbb{E}\left(e^{\lambda(Y_k-\mathbb{E}(Y_k))}\right), \quad \lambda > 0.$$

We need to estimate $\mathbb{E}\left(e^{-\lambda(Y_k-\mathbb{E}(Y_k))}\right)$. The estimation is provided in Lemma 8.4, and thus

$$\mathbb{P}\left(\sum_{k=1}^{n} Y_k - \sum_{k=1}^{n} \mathbb{E}(Y_k) \geq \varepsilon\right) \leq e^{-\lambda\varepsilon}\prod_{k=1}^{n} e^{\lambda^2(b_k-a_k)^2/8} = \exp\left(\frac{\lambda^2}{8}\sum_{k=1}^{\infty}(b_k-a_k)^2 - \varepsilon\lambda\right), \quad \lambda > 0.$$

Since $\lambda > 0$ is arbitrary, we pick $\lambda = 4\varepsilon/\sum_{k=1}^{\infty}(b_k-a_k)^2$ and yield

$$\mathbb{P}\left(\sum_{k=1}^{n} Y_k - \sum_{k=1}^{n} \mathbb{E}(Y_k) \geq \varepsilon\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k-a_k)^2}\right).$$

Applying the similar reasoning to $-Y_1, \ldots, -Y_n$, we yield

$$\mathbb{P}\left(\sum_{k=1}^{n} Y_k - \sum_{k=1}^{n} \mathbb{E}(Y_k) \leq -\varepsilon\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k-a_k)^2}\right).$$

Finally, in view of Theorem 1.16 (b), we conclude the proof. $\square$

lem:EstMoment **Lemma 8.4.** *For a real-valued random variable Y with $\mathbb{E}(Y) = 0$ and $a \leq Y \leq b$, we have*

$$\mathbb{E}(e^{\lambda Y}) \leq e^{\lambda^2(b-a)^2/8}, \quad \lambda > 0.$$

Proof. The statement is clearly true for $a = b$. Suppose $a < b$ and define $Z := (b-Y)/(b-a)$. Then, $Y = Za + (1-Z)b$ and

$$e^{\lambda Y} = e^{Z\lambda a+(1-Z)\lambda b} \leq Ze^{\lambda a} + (1-Z)e^{\lambda b} = \frac{b-Y}{b-a}e^{\lambda a} + \frac{Y-a}{b-a}e^{\lambda b}.$$

Taking expectation we yield,

$$\begin{aligned}\mathbb{E}(e^{\lambda Y}) &\leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = \exp\left(\log\left(\frac{be^{\lambda a}-ae^{\lambda b}}{b-a}\right)\right) = \exp\left(\lambda a + \log\left(\frac{b-ae^{\lambda(b-a)}}{b-a}\right)\right)\\ &= \exp\left(u(p-1) + \log(p-(1-p)e^u)\right), \qquad (8.1)\end{aligned}$$

eq:EExplambdaY

where $p := b/(b-a)$ and $u := \lambda(b-a)$. Let $\varphi(u) := u(p-1) + \log(p+(1-p)e^u)$ for $u \geq 0$ (note $b \geq 0$ and thus $p \geq 0$). By Taylor's expansion, we have

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(\xi)u^2$$

for some $\xi \in [0, u]$. Note $\varphi(0) = 0$, and $\varphi'(u) = (p-1) + \frac{(1-p)e^u}{p+(1-p)e^u}$, i.e., $\varphi'(0) = 0$. Regarding φ'',

$$\varphi''(u) = \frac{(1-p)e^u}{p+(1-p)e^u} - \frac{(1-p)^2e^{2u}}{(p+(1-p)e^u)^2} \leq \frac{1}{4}, \quad u \geq 0.$$

It follows that $\varphi(u) \leq u^2/8$. This together with (8.1) concludes the proof. $\square$

Remark 8.5. Extension of Hoeffding's inequality.

thm:HolderIneq **Theorem 8.6** (Hölder's Inequality)**.** *Let $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. On $(\mathbb{X}, \mathscr{X}, \mu)$, then for any real-valued f and g, we have*

$$\int_{\mathbb{X}} |f(x)g(x)|\mu(\mathrm{d}x) \leq \left(\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}} \left(\int_{\mathbb{X}} |g(x)|^q \mu(\mathrm{d}x)\right)^{\frac{1}{q}}.$$

If $p = 1$, then

$$\int_{\mathbb{X}} |f(x)g(x)|\mu(\mathrm{d}x) \leq C_g \int_{\mathbb{X}} |f(x)|\mu(\mathrm{d}x),$$

where $C_g := \inf\{r \geq 0 : \mu(\{x \in \mathbb{X} : |g(x)| > C_g\})\} = 0$.

The proof of Theorem 8.6 relies on the following lemma.

em:YoungIneqProd **Lemma 8.7.** *Let $a, b \geq 0$ and $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

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

The equality holds only when $a^p = b^q$.

Proof. The inequality is true of $a = b = 0$. For the rest of the proof, we assume $a > 0$ and $b > 0$. Because ln is concave, therefore

$$\ln\left(\frac{a^p}{p} + \frac{b^q}{q}\right) \geq \frac{1}{p}\ln(a^p) + \frac{1}{q}\ln(b^q) = \ln(ab).$$

Note that the equality is true only when $a^p = b^q$. Taking exponential on both hand side, we conclude the proof. $\square$

Proof of Theorem 8.6. The case of $p = 1$ is obvious. We suppose $p, q \in (1, \infty)$. If one of $\left(\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}}$ or $\left(\int_{\mathbb{X}} |f(x)|^q \mu(\mathrm{d}x)\right)^{\frac{1}{q}}$ is infinite, the inequality is automatically true. Without loss of generality, we assume both the quantities are 1. By Lemma 8.7, we have

$$|f(x)g(x)| = |f(x)||g(x)| \leq \frac{|f(x)|^p}{p} + \frac{|g(x)|^q}{q}.$$

Integrating both hand side, we yield the statement. $\square$

Below is a special case of Hölder's inequality (Theorem 8.6) with $p = q = 2$.

cor:CSIneq **Corollary 8.8** (Cauchy-Schwartz)**.** *On $(\mathbb{X}, \mathscr{X}, \mu)$, for any real-valued f and g, we have*

$$\int_{\mathbb{X}} |f(x)g(x)|\mu(\mathrm{d}x) \leq \sqrt{\left(\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x)\right)\left(\int_{\mathbb{X}} |g(x)|^q \mu(\mathrm{d}x)\right)}.$$

Theorem 8.9 (Minkowski inequality)**.** *Let $p \in [1, \infty)$. For any real-valued f and g in $\mathcal{L}^p(\mathbb{X}, \mathscr{X}, \mu)$, we have $f + g \in \mathcal{L}^p$ and*

$$\left(\int_{\mathbb{X}} |f(x) + g(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}} \leq \left(\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}} + \left(\int_{\mathbb{X}} |g(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}}.$$

Proof. Note that $x \mapsto |x|^p$ is convex. Therefore,

$$\left|\frac{1}{2}|f(x)| + \frac{1}{2}|g(x)|\right|^p \leq \frac{1}{2}|f(x)|^p + \frac{1}{2}|g(x)|^p, \quad x \in \mathbb{X}.$$

This implies that $f + g \in \mathscr{L}^p$. Next, note that the case of $p = 1$ is immediate due to triangle inequality. We suppose $p > 1$. We also assume $|f + g|$ is no constant 0 as the statement is trivially true otherwise. Note

$$\begin{aligned}\int_{\mathbb{X}} |f(x) + g(x)|^p \mu(\mathrm{d}x) &\leq \int_{\mathbb{X}} |f(x) + g(x)||f(x) + g(x)|^{p-1}\mu(\mathrm{d}x)\\ &\leq \int_{\mathbb{X}} |f(x)||f(x) + g(x)|^{p-1}\mu(\mathrm{d}x) + \int_{\mathbb{X}} |g(x)||f(x) + g(x)|^{p-1}\mu(\mathrm{d}x).\end{aligned}$$

Let $q = \frac{p}{p-1}$ so that $\frac{1}{p} + \frac{1}{q} = 1$. By Hölder's inequality (Theorem 8.6),

$$\int_{\mathbb{X}} |f(x)||f(x) + g(x)|^{p-1}\mu(\mathrm{d}x) \leq \left(\int_{\mathbb{X}} |f(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}} \left(\int_{\mathbb{X}} |f(x) + g(x)|^p \mu(\mathrm{d}x)\right)^{1-\frac{1}{p}}$$

and

$$\int_{\mathbb{X}} |g(x)||f(x) + g(x)|^{p-1}\mu(\mathrm{d}x) \leq \left(\int_{\mathbb{X}} |g(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}} \left(\int_{\mathbb{X}} |f(x) + g(x)|^p \mu(\mathrm{d}x)\right)^{1-\frac{1}{p}}.$$

Combining the above, and simplifying the resulting inequality, we conclude the proof. □

Remark 8.10. One of the major consequence of Minkowski inequality is that we can use

$$d(f, g) := \left(\int_{\mathbb{X}} |f(x) - g(x)|^p \mu(\mathrm{d}x)\right)^{\frac{1}{p}}$$

as a metric for $\mathscr{L}^p$ spaces with $p \in [1, \infty)$.

The following inequality estimates the variance of certain system under the influence multiple independent factor.

Theorem 8.11 (Efron-Stein's Inequality)**.** *Let $Y_1, \ldots, Y_n, Y_1', \ldots, Y_n'$ be mutually independent random variables from $(\Omega, \mathscr{A})$ to $(\mathbb{Y}, \mathscr{Y})$ such that Y_k and Y_k' have the same distribution. Define $Y = (Y_1, \ldots, Y_n)$ and $Y^{(k)} := (Y_1, \ldots, Y_{k-1}, Y_k', Y_{k+1}, \ldots, Y_n)$. Then, for any bounded $f : (\mathbb{Y}^n, \mathscr{Y}^{\otimes n}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we have*

$$\mathrm{Var}(f(Y)) \leq \frac{1}{2}\sum_{k=1}^{n} \mathbb{E}\left(\left(f(Y) - f(Y^{(k)})\right)^2\right).$$

Proof. In what follows, we use the following notations:

$$Y' := (Y_1', \ldots, Y_n'), \quad Y^{[k]} := (Y_1', \ldots, Y_k', Y_{k+1}, \ldots, Y_n), \quad k = 1, \ldots, n.$$

Note $Y^{[n]} = Y'$. We also set $Y^{[0]} := Y$. Observe that

$$\begin{aligned}\mathrm{Var}(f(Y)) &= \mathbb{E}(f(Y)^2) - \mathbb{E}(f(Y))^2 = \mathbb{E}(f(Y)^2) - \mathbb{E}(f(Y)f(Y'))\\ &= \mathbb{E}(f(Y)(f(Y) - f(Y'))) = \sum_{k=1}^{n} \mathbb{E}(f(Y)(f(Y^{[k-1]}) - f(Y^{[k]}))).\end{aligned}$$

Note that the distribution of $(Y_1, \dots, Y_n, Y'_1, \dots, Y'_n)$ remains the same if we switch Y_k and Y'_k. Therefore,

$$f(Y)(f(Y^{[k-1]}) - f(Y^{[k]})) \quad \text{and} \quad f(Y^{(k)})(f(Y^{[k]}) - f(Y^{[k-1]}))$$

have the same distribution. It follows from Theorem 4.12 that

$$\mathbb{E}\left(f(Y)(f(Y^{[k-1]}) - f(Y^{[k]}))\right) = \mathbb{E}\left(f(Y^{(k)})(f(Y^{[k]}) - f(Y^{[k-1]}))\right),$$

and thus the average of both hand sides should remains the same. Consequently, by Cauchy-Schwartz inequality (Corollary 8.8),

$$\mathbb{E}\left(f(Y)(f(Y^{[k-1]}) - f(Y^{[k]}))\right) = \frac{1}{2}\mathbb{E}\left((f(Y) - f(Y^{(k)}))(f(Y^{[k-1]}) - f(Y^{[k]}))\right)$$
$$\leq \frac{1}{2}\sqrt{\mathbb{E}\left((f(Y) - f(Y^{(k)}))^2\right)\mathbb{E}\left((f(Y^{[k-1]}) - f(Y^{[k]}))^2\right)}.$$

Finally, noticing that

$$\mathbb{E}\left((f(Y) - f(Y^{(k)}))^2\right) = \mathbb{E}\left((f(Y^{[k-1]}) - f(Y^{[k]}))^2\right),$$

summing over k, we conclude the proof. $\square$

thm:JensenIneq **Theorem 8.12** (Jensen's Inequality)**.** *On $(\Omega, \mathscr{A}, \mathbb{P})$, consider $Y : (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose $g(Y)$ is integrable. Then, $g(\mathbb{E}(Y)) \leq \mathbb{E}(g(Y))$.*

In order to proof Jensen's inequality, we first prove the following lemma.

lem:Convex **Lemma 8.13.** *Let $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be convex. Then, for any $r_0 \in \mathbb{R}$, there is $c_0 \in \mathbb{R}$ such that $g(r) \geq g(r_0) + c_0(r - r_0)$ for $r \in \mathbb{R}$.*

Proof. We first consider $r > r_0$. Note that for $c \in (0, 1)$, we have

$$g(r_0 + (1-c)(r - r_0)) = g(cr_0 + (1-c)r) \leq cf(r_0) + (1-c)f(r)$$

and thus

$$\frac{g(r_0 + (1-c)(r - r_0)) - g(r_0)}{(1-c)(r - r_0)} \leq \frac{g(r) - g(r_0)}{r - r_0}.$$

It follows that $r \mapsto \frac{f(r) - f(r_0)}{r - r_0}$ is non-decreasing in $r > r_0$. Similar reasoning shows that $r \mapsto \frac{g(r_0) - g(r)}{r_0 - r}$ is non-decreasing in $r < r_0$. Moreover, for $r' < r_0 < r''$, we have

$$g(r_0) \leq g\left(\frac{r'' - r_0}{r'' - r'}r' + \frac{r_0 - r'}{r'' - r'}r''\right) \leq \frac{r'' - r_0}{r'' - r'}g(r') + \frac{r_0 - r'}{r'' - r'}f(r'')$$

and thus

$$\frac{g(r_0) - g(r')}{r_0 - r'} \leq \frac{g(r'') - g(r_0)}{r'' - r_0}.$$

Let $c_0 = \liminf_{r \to r_0+} \frac{g(r'') - g(r_0)}{r'' - r_0}$, we must have $c_0 > -\infty$, and

$$\frac{g(r_0) - g(r')}{r_0 - r'} \leq c_0 \leq \frac{g(r'') - f(r_0)}{r'' - r_0}, \quad r' < r_0 < r'',$$

which completes the proof. $\square$

Proof of Theorem 8.12. In view of Lemma 8.13, we let $r_0 = \mathbb{E}(Y)$ and $c_0 \in \mathbb{R}$ satisfies $g(r) \geq g(r_0) + c_0(r - r_0)$. Then, by Proposition 4.8, we yield

$$\mathbb{E}(g(Y)) \geq g(\mathbb{E}(Y)) + c_0\mathbb{E}(Y - \mathbb{E}(Y)) = g(\mathbb{E}(Y)).$$

The proof is complete. □

9 Convergence of Random Variables

In what follows, we will fix a measure space $(\mathbb{X}, \mathscr{X}, \mu)$. Let f be a real-valued $\mathscr{X}$-measurable and $(f_n)_{n\in\mathbb{N}}$ be a sequence of real-valued $\mathscr{X}$-measurable function. Upon declaration, we will replace $(\mathbb{X}, \mathscr{X}, \mu)$ by a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, and replace f, f_n by real-valued random variables Y, Y_n.

def:FuncConv **Definition 9.1.**

- We say $(f_n)_{n\in\mathbb{N}}$ converges almost surely to f if

$$\mu(\{x : \lim_{n\to\infty} f_n(x) = f(x)\}^c) = 0,$$

and denote $\lim_{n\to\infty} f_n = f$, $\mu - a.s.$. If μ is a probability, we may alternatively say $(f_n)_{n\in\mathbb{N}}$ converges to f with probability 1.

- Let $p \in [1, \infty)$ and suppose $f \in \mathcal{L}^p$, $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{L}^p$. We say $(f_n)_{n\in\mathbb{N}}$ converges to f in $\mathcal{L}^p$ if

$$\lim_{n\to\infty} \int_{\mathbb{X}} |f_n(x) - f(x)|^p \mu(\mathrm{d}x) = 0,$$

and denote $f_n \xrightarrow{\mathcal{L}^p} f$.

- We say $(f_n)_{n\in\mathbb{N}}$ converges to f in measure if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu(\{x \in \mathbb{X} : |f_n(x) - f(x)| > \varepsilon\}) = 0,$$

and denote $\mu\text{-}\lim f_n = f$.

Remark 9.2. Note all the convergence mentioned above depends on the underlying probability space. Sometimes it is necessary to emphasize the dependence on μ.

rmk:ASConv *Remark* 9.3. If $(\mathbb{X}, \mathscr{X}, \mu)$ is complete, following from Remark 2.2 (7), we have almost sure limit (if exists) preserves measurability. This means we need not introduce a measurable f beforehand in Definition 9.1. Instead, we can simply define $f(x) := \lim_{n\to\infty} f_n(x)$ for $x \in \mathbb{X} \setminus N$, when the limit exists except for a null set N.

exmp:FuncConv **Example 9.4.** In this example, we consider the measure space $([0,1], \mathcal{B}([0,1]), \lambda)$.

1. Let $f_n(x) := n^2 \mathbb{1}_{[0,1/n]}$. Then, $(f_n)_{n\in\mathbb{N}}$ converges to 0 almost surely and in measure. But $(f_n)_{n\in\mathbb{N}}$ does not converges to 0 in $\mathcal{L}^p$ for any $p \in [0, \infty)$.

2. Let $s_0 = 0$ and $s_n := s_{n-1} + n^{-1}$ for $n \in \mathbb{N}$. For $n \in \mathbb{N}$ define

$$f_n(x) := \begin{cases} \mathbb{1}_{[s_{n-1}\%1, s_n\%1)}(x), & s_{n-1}\%1 \leq s_n\%1, \\ \mathbb{1}_{[0, s_n\%1) \cup [s_n\%1, 1]}(x), & s_{n-1}\%1 > s_n\%1, \end{cases}$$

where % means modulo. Then, $(f_n)_{n\in\mathbb{N}}$ converges to 0 in $\mathcal{L}^p$ for $p \in [1, \infty)$ and in measure, but not almost surely.

Theorem 9.5. *Let $g : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be continuous.*

(a) if $(f_n)_{n\in\mathbb{N}}$ converges to f almost surely, then $(g \circ f_n)_{n\in\mathbb{N}}$ converges to $g \circ f$ almost surely;

(b) suppose $\mu = \mathbb{P}$ is a probability measure and $(Y_n)_{n\in\mathbb{N}}$ converges to Y in measure, then $(g(Y_n))_{n\in\mathbb{N}}$ converges to $g(Y)$ in measure.

Proof. **(a)** Note $\{x \in \mathbb{X} : \lim_{n\to\infty} f_n(x) = f(x)\} \subseteq \{x \in \mathbb{X} : \lim_{n\to\infty} g(f_n(x)) = g(f(x))\}$. Therefore,

$$\mu(\{x \in \mathbb{X} : \lim_{n\to\infty} g(f_n(x)) = g(f(x))\}^c) \le \mu(\{x \in \mathbb{X} : \lim_{n\to\infty} f_n(x) = f(x)\}^c) = 0.$$

(b) Let $\varepsilon > 0$ and $\ell \in \mathbb{N}$. Note

$$\begin{aligned}\{|g(Y_n) - g(Y)| > \varepsilon\} &= (\{|g(Y_n) - g(Y)| > \varepsilon\} \cap \{Y \in [-\ell, \ell]\}) \cup (\{|g(Y_n) - g(Y)| > \varepsilon\} \cap \{Y \in [-\ell, \ell]^c\}) \\ &\subseteq (\{|g(Y_n) - g(Y)| > \varepsilon\} \cap \{Y \in [-\ell, \ell]\}) \cup \{Y \in [-\ell, \ell]^c\}.\end{aligned}$$

Therefore, by countable additivity and Theorem 1.16 (a),

$$\mu(|g(Y_n) - g(Y)| > \varepsilon) \le \mu(|g(Y_n) - g(Y)| > \varepsilon,\, Y \in [-\ell, \ell]) + \mu(Y \in [-\ell, \ell]^c). \tag{9.1}$$ eq:mugf

Note g is uniformly continuous on $[-\ell-1, \ell+1]$, then there is $\delta \in (0, 1)$ such that $|g(y+r)-g(y)| \le \varepsilon$ for $r \in [-\delta, \delta]$. Thus, under $Y \in [-\ell, \ell]$, for $|g(Y_n)-g(Y)| > \varepsilon$ to be true, we must have $|Y_n-Y| > \delta$. This together with (9.1) implies that

$$\mathbb{P}(|g(Y_n) - g(Y)| > \varepsilon) \le \mathbb{P}(|Y_n - Y| > \delta) + \mathbb{P}(Y \in [-\ell, \ell]^c).$$

Taking lim sup on both hand sides above, we yield

$$\limsup_{n\to\infty} \mu(|g(Y_n) - g(Y)| > \varepsilon) \le \mu(Y \in [-\ell, \ell]^c).$$

Noting that $([-\ell, \ell]^c)_{\ell\in\mathbb{N}}$ decreases to $\emptyset$, invoking Theorem 1.16 (e), we have

$$\limsup_{n\to\infty} \mu(|g(Y_n) - g(Y)| > \varepsilon) = 0,$$

which completes the proof. □

Example 9.6. Here is a non-example for (b) above with $\mu(\mathbb{X}) = \infty$. Consider $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \lambda)$. Let $f(x) := x$ and $f_n(x) := x + n^{-1}$. Clearly, $(f_n)_{n\in\mathbb{N}}$ converges to f in measure. Now let $g(y) := y^2$. Then, for any n and $\varepsilon > 0$, we have $\mu(\{r \in \mathbb{R} : |g(f_n(r)) - r^2| > \varepsilon\}) = \mu([\frac{1}{2}(n\varepsilon - n^{-1}), \infty)) = \infty$.

10 Limit Theorems

In view of the convention that $0 \cdot \infty = 0$ and the notion of almost sure convergence, we can easily extend Theorem 4.9 into the following.

thm:MonoConv **Theorem 10.1** (Monotone Convergence)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of non-negative real-valued measurable function. Suppose $(f_n)_{n\in\mathbb{N}}$ increases to f almost surely. Then,*

$$\lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x).$$

The next theorem is known as *Fatou's lemma.*

thm:Fatou **Theorem 10.2** (Fatou)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of non-negative real-valued measurable function. Then,*

$$\int_{\mathbb{X}} \liminf_{n\to\infty} f_n(x)\mu(\mathrm{d}x) \le \liminf_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x).$$

Proof. Recall that $\liminf_{n\to\infty} f_n(x) = \lim_{n\to\infty} \inf_{k\ge n} f_n(x)$ and we define $g_n(x) := \inf_{k\ge n} f_k(x)$. Note that $(g_n)_{n\in\mathbb{N}}$ increases to $\liminf_{n\to\infty} f_n$ almost surely. In addition, because $g_n \le f_k$ for $k \le n$, we have

$$\int_{\mathbb{X}} g_n(x)\mu(\mathrm{d}x) \le \inf_{k\ge n} \int_{\mathbb{X}} f_k(x)\mu(\mathrm{d}x).$$

Invoking monotone convergence finishes the proof. $\square$

cor:Fatou **Corollary 10.3.** *Suppose $(f_n)_{n\in\mathbb{N}}$ converges almost surely to f and $\int_{\mathbb{X}} |f_n(x)|\mu(\mathrm{d}x) \le K$ for some $K > 0$. Then, $\int_{\mathbb{X}} |f(x)|\mu(\mathrm{d}x) \le K$.*

Example 10.4. Consider $([0,1], \mathcal{B}([0,1]), \lambda)$ and define $f_n(x) := -n^2 \mathbb{1}_{[0,\frac{1}{n}]}(x)$ for $x \in [0,1]$. Note Note $\liminf_{n\to\infty} f_n(x) = 0$ a.s. for $x \in [0,1]$ but $\int_{[0,1]} f_n(x)\mu(\mathrm{d}x) = -n$ and thus

$$\liminf_{n\to\infty} \int_{[0,1]} f_n(x)\mu(\mathrm{d}x) = -\infty.$$

thm:DomConv **Theorem 10.5** (Dominated Convergence)**.** *Suppose $(f_n)_{n\in\mathbb{N}}$ converges almost surely to f and there is $g \in \mathcal{L}^1$ such that $|f_n| \le g$. Then, $f \in \mathcal{L}^1$ and*

$$\lim_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x).$$

Proof. As an immediate consequence of Corollary 10.3, we have $f \in \mathcal{L}^1$. Note $f_n + g \ge 0$. By Theorem 4.11 and Fatou's lemma (Theorem 10.2), we have

$$\begin{aligned}\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) + \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x) &= \int_{\mathbb{X}} \liminf_{n\to\infty} \left(f_n(x) + g(x)\right)\mu(\mathrm{d}x) \\ &\le \liminf_{n\to\infty} \int_{\mathbb{X}} \left(f_n(x) + g(x)\right)\mu(\mathrm{d}x) = \liminf_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x) + \int_{\mathbb{X}} g(x)\mu(\mathrm{d}x). \qquad (10.1)\end{aligned}$$ eq:fliminf

It follows that

$$\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) \le \liminf_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x).$$

On the other hand, we have $g - f_n \ge 0$ and thus with similar reasoning as before,

$$-\int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) \le \liminf_{n\to\infty} \int_{\mathbb{X}} (-f_n)(x)\mu(\mathrm{d}x) = -\limsup_{n\to\infty} \int_{\mathbb{X}} f_n(x)\mu(\mathrm{d}x). \qquad (10.2)$$ eq:flimsupf

Combining (10.1) and (10.2) as well as the fact that $\limsup \ge \liminf$, the proof is complete. $\square$

Remark 10.6. Note that monotone convergence and dominated convergence generalizes Theorem 1.16 (c) (e) and Theorem 1.18.

11 Relations between Convergences

onvLPConvInMeas **Theorem 11.1.** *If $(f_n)_{n\in\mathbb{N}}$ converges to f in converges to in $\mathcal{L}^p$, then $(f_n)_{n\in\mathbb{N}}$ converges to f in measure.*

Proof. Let $\varepsilon > 0$. By Theorem 8.1, we have

$$\mu(\{x \in \mathbb{X} : |f_n(x) - f(x)| \geq \varepsilon\}) \leq \frac{1}{\varepsilon^p} \int_{\mathbb{X}} |f_n(x) - f(x)|^p \mu(\mathrm{d}x).$$

In view of Definition 9.1, we conclude the proof. □

onvInMeasConvAS **Theorem 11.2.** *If $(f_n)_{n\in\mathbb{N}}$ converges to f in measure, then there is a subsequence $(f_{n_j})_{j\in\mathbb{N}}$ converging to f almost surely.*

Proof. Let $n_1 := 1$ and choose $n_k > n_{k-1}$ by induction such that

$$\mu(\{x \in \mathbb{X} : |f_{n_k}(x) - f(x)| > \frac{1}{k}\}) \leq 2^{-k}.$$

We define $A_k := \{x \in \mathbb{X} : |f_{n_k}(x) - f(x)| > \frac{1}{k}\}$ and $A := \bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} A_k$. Note that, by Theorem 1.16 (b), $\mu(\bigcup_{k\in\mathbb{N}} A_k) \leq \sum_{k\in\mathbb{N}} \mu(A_k) < \infty$ and $(\bigcup_{k\geq n} A_k)_{n\in\mathbb{N}}$ decreases to A. Then, by dominated convergence (Theorem 10.5),

$$\mu(A) = \lim_{n\to\infty} \mu(\bigcup_{j\geq n} A_j) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mu(A_k) \leq \lim_{n\to\infty} 2^{-n+1} = 0.$$

In addition, observe that for $x \notin A$, there is $n \in \mathbb{N}$ such that $|f_{n_k}(x) - f(x)| \leq \frac{1}{k}$ for any $k \geq n$, and thus $\lim_{k\to\infty} f_{n_k}(x) = f(x)$. By Theorem 1.16 (a), we have

$$\mu(\{x \in \mathbb{X} : \lim_{k\to\infty} f_{n_k}(x) = f(x)\}^c) \leq \mu(A) = 0.$$

In view of Definition 9.1, we conclude the proof. □

Corollary 11.3. *If $(f_n)_{n\in\mathbb{N}}$ converges to f in converges to in $\mathcal{L}^p$, then there is a subsequence $(f_{n_j})_{j\in\mathbb{N}}$ converging to f almost surely.*

Remark 11.4. measurabilty in complete probability space

Theorem 11.5. *Let $p \in [1, \infty)$. Suppose $(f_n)_{n\in\mathbb{N}}$ converges to f in measure, and there is a non-negative $g \in \mathcal{L}^p$ such that $|f_n| \leq g$ for $n \in \mathbb{N}$. Then, $f \in \mathcal{L}^p$ and $(f_n)_{n\in\mathbb{N}}$ converges to f in $\mathcal{L}^p$.*

Proof. We first prove that $f \in \mathcal{L}^p$. To this end note that

$$\begin{aligned}\{x \in \mathbb{X} : |f(x)| > g(x) + \varepsilon\} &\subseteq \{x \in \mathbb{X} : |f(x)| > |f_n(x)| + \varepsilon\} \\ &\subseteq \{x \in \mathbb{X} : |f(x) - f_n(x)| > +\varepsilon\}.\end{aligned}$$

Therefore, by Theorem 1.16 (a),

$$\mu(\{x \in \mathbb{X} : |f(x)| > g(x) + \varepsilon\}) \leq \mu(\{x \in \mathbb{X} : |f(x) - f_n(x)| > \varepsilon\}) \xrightarrow[n\to\infty]{} 0.$$

Note additionally that $(\{x \in \mathbb{X} : |f(x)| > g(x) + \frac{1}{k}\})_{k\in\mathbb{N}}$ increases to $\{x \in \mathbb{X} : |f(x)| > g(x)\}$. By Theorem 1.16 (c), we conclude $|f| \le g$, and thus $f \in \mathcal{L}^p$.

We proceed to show the $\mathcal{L}^p$ convergence by contradiction. Suppose there is $\varepsilon > 0$ and a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that

$$\int_{\mathbb{X}} |f_{n_k}(x) - f(x)|^p \mu(\mathrm{d}x) \ge \varepsilon_0, \quad k \in \mathbb{N}. \tag{11.1}$$

eq:fnkf

Because $(f_{n_k})_{k\in\mathbb{N}}$ converges to f in measure, by Theorem 11.2, there is a further subsequence $(f_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ converging to f almost surely. Note that $|f_n - f|^p \le 2^p g^p$. Then, by dominated convergence (Theorem 10.5), we have

$$\lim_{\ell\to\infty} \int_{\mathbb{X}} |f_{n_{k_\ell}}(x) - f(x)|^p \mu(\mathrm{d}x) = 0,$$

which contradicts (11.1). This finishes the proof. □

In what follows, we let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space, Y be a real-valued random variable and $(Y_n)_{n\in\mathbb{N}}$ a sequence of real-valued measurable functions.

hm:ASConvInMeas **Theorem 11.6.** *If $(Y_n)_{n\in\mathbb{N}}$ converges to Y almost surely, then $(Y_n)_{n\in\mathbb{N}}$ converges to Y in measure.*

Proof. For $\varepsilon > 0$, let $A_n := \{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| > \varepsilon\}$. Note $\mathbb{1}_{A_n}$ converges to 0 almost surely. Then, by dominated convergence (Theorem 10.5), we have

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \int_{\Omega} \mathbb{1}_{A_n}(\omega)\mu(\mathrm{d}\omega) = \int_{\Omega} \lim_{n\to\infty} \mathbb{1}_{A_n}(\omega)\mu(\mathrm{d}\omega) = 0.$$

In view of Definition 9.1, we conclude the proof. □

Example 11.7. Here we show an example on $(\mathbb{X}, \mathscr{X}, \mu)$ where $\mu(\mathbb{X}) = \infty$ and almost sure convergence does not implies convergence in measure. Let $(\mathbb{X}, \mathscr{X}) = (\mathbb{N}, 2^{\mathbb{N}})$ and μ be the counting measure, i.e., $\mu(A)$ equals to the number of elements in A. Let $f_n(i) = (1 - n^{-1})\mathbb{1}_{[0,n]}(i)$. Note that $(f_n)_{n\in\mathbb{N}}$ converges to 1 almost surely but $\mu(|f_n - 1| > \frac{1}{2}) = \infty$ for all $n \in \mathbb{N}$.

12 Laws of Large Numbers

One of the fundamental result of probability regards the laws of large number. Below we introduce a few different versions. In what follows, we let Y, Z be real-valued random variables, and $(Y_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables. Recall that

$$\mathrm{Var}(Y) := \mathbb{E}((Y - \mathbb{E}(Y))^2).$$

By saying $\mathrm{Var}(Y) < \infty$, we mean $Y \in \mathcal{L}^1$ and $Y - \mathbb{E}(Y) \in \mathcal{L}^2$. This also implies $Y \in \mathcal{L}^2$ (why?). In this case, we have $\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$.

For $Y, Z \in \mathcal{L}^2$, in view of Cauchy-Schwartz inequality (TBA), we define

$$\mathrm{Cov}(Y, Z) := \mathbb{E}((Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z).$$

Lemma 12.1. *For $Y, Z, W \in \mathcal{L}^2$, the following is true:*

(a) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;

(b) $\mathrm{Cov}(aX+bY,W) = a\,\mathrm{Cov}(X,W) + b\,\mathrm{Cov}(Y,W)$ *for any* $a,b\in\mathbb{R}$;

(c) $\mathrm{Cov}(a,W) = 0$ *for any* $a\in\mathbb{R}$;

(d) $\mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$.

The formula below is useful and is a good exercise:

$$\mathrm{Var}(\sum_{k=1}^n Y_k) = \sum_{j=1}^n\sum_{i=1}^n \mathrm{Cov}(Y_i,Y_j) = \sum_{k=1}^n \mathrm{Var}(Y_k) + 2\sum_{\{i,j\}\subseteq\{1,\ldots,n\}} \mathrm{Cov}(Y_i,Y_j).$$

thm:L2LLN **Theorem 12.2.** *[$\mathcal{L}^2$ LLN] Suppose $(Y_n)_{n\in\mathbb{N}}$ satisfies $\mathbb{E}(Y_n)=a$, $\mathrm{Var}(Y_n)=b^2$ and $\mathrm{Cov}(Y_i,Y_j)=0$ for $i\neq j$. Then,*

$$\mathbb{E}\left(\left(\frac{1}{n}\sum_{k=1}^n Y_k - a\right)^2\right) \le \frac{b^2}{n}.$$

Consequently, $(\frac{1}{n}\sum_{k=1}^n Y_k)_{n\in\mathbb{N}}$ converges to a in $\mathcal{L}^2$.

Proof. Because $\mathrm{Cov}(Y_i,Y_j)=0$ for $i\neq j$, we have

$$\mathbb{E}\left(\left(\sum_{k=1}^n Y_k - na\right)^2\right) = \mathbb{E}\left(\left(\sum_{k=1}^n (Y_k-a)\right)^2\right) = \mathbb{E}\left(\sum_{i=1}^n\sum_{j=1}^n (Y_i-a)(Y_j-a)\right)$$
$$= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\left((Y_i-a)(Y_j-a)\right) = \sum_{k=1}^n \mathbb{E}\left((Y_k-a)^2\right) = n\sigma^2.$$

Dividing both hand sides by n^2, we finish the proof. $\square$

thm:StrongLLN **Theorem 12.3.** *[Strong LLN] Suppose $(Y_n)_{n\in\mathbb{N}}$ is be an sequence of pairwise independent identically distributed real-valued random variable such that $\mathbb{E}(Y_1)=a$ and $\mathbb{E}|Y_1|<\infty$. Then, $(\frac{1}{n}\sum_{k=1}^n Y_k)_{n\in\mathbb{N}}$ converges to a almost surely.*

Proof. Without loss of generality, we assume $Y_1\ge 0$. Define $S_n := \sum_{k=1}^n Y_k$, $Z_n := Y_n \mathbb{1}_{\{Y_n<n\}}$, and $T_n := \sum_{k=1}^n Z_k$. For $\alpha>1$, we let $\ell_n := [\alpha^n]$ be the smallest integer larger than α^n. For $\varepsilon>0$, note that by Theorem 8.1,

$$\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{1}{\ell_n}\left(T_{\ell_n} - \mathbb{E}(T_{\ell_n})\right)\right| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\sum_{n=1}^\infty \frac{\mathrm{Var}(T_n)}{\ell_n^2} = \frac{1}{\varepsilon^2}\sum_{n=1}^\infty \frac{1}{\ell_n^2}\sum_{k=1}^{\ell_n}\mathrm{Var}(Z_k).$$

Since the summands are non-negative, we switch the order of sums to yield,

$$\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{1}{\ell_n}\left(T_{\ell_n} - \mathbb{E}(T_{\ell_n})\right)\right| > \varepsilon\right) \le C_\varepsilon \sum_{k=1}^\infty \frac{1}{k^2}\mathrm{Var}(Z_k) \le C_\varepsilon \sum_{k=1}^\infty \frac{1}{k^2}\mathbb{E}\left(Y_1^2 \mathbb{1}_{\{Y_1<k\}}\right)$$
$$= C_\varepsilon \sum_{k=1}^\infty \frac{1}{k^2}\sum_{i=0}^{k-1}\mathbb{E}\left(Y_1^2\mathbb{1}_{\{Y_1\in[i,i+1)\}}\right) = C_\varepsilon \sum_{i=0}^\infty \left(\mathbb{E}\left(Y_1^2\mathbb{1}_{\{Y_1\in[i,i+1)\}}\right)\sum_{k=i}^\infty \frac{1}{k^2}\right)$$
$$\le C_\varepsilon \sum_{i=0}^\infty \frac{1}{i+1}\mathbb{E}\left(Y_1^2\mathbb{1}_{\{Y_1\in[i,i+1)\}}\right) \le C_\varepsilon \mathbb{E}(Y_i) < \infty.$$

for some $C_\varepsilon > 0$. It follows that from Borel-Cantelli lemma (Lemma 6.5) that

$$\mathbb{P}\left(\bigcup_{m\in\mathbb{N}}\bigcap_{n\geq m}\left\{\left|\frac{1}{\ell_n}\left(T_{\ell_n}-\mathbb{E}(T_{\ell_n})\right)\right|\leq\frac{1}{k}\right\}\right)=1.$$

This together with Theorem 1.16 (e) implies that

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}}\bigcup_{m\in\mathbb{N}}\bigcap_{n\geq m}\left\{\left|\frac{1}{\ell_n}\left(T_{\ell_n}-\mathbb{E}(T_{\ell_n})\right)\right|\leq\frac{1}{k}\right\}\right)=1,$$

i.e., $(\ell_n^{-1}(T_{\ell_n}-\mathbb{E}(T_{\ell_n})))_{n\in\mathbb{N}}$ converges to 0 almost surely. Note additionally that

$$\mathbb{E}(Y_1)=\lim_{n\to\infty}\mathbb{E}\left(Y_1\mathbb{1}_{\{Y_1<n\}}\right)=\lim_{n\to\infty}\mathbb{E}(Z_n)=\lim_{n\to\infty}\frac{1}{\ell_n}\mathbb{E}(T_{\ell_n}),$$

where we have used monotone convergence (cf. Theorem 10.1) in the first inequality. We have $(T_{\ell_n}/\ell_n)_{n\in\mathbb{N}}$ converges to $\mathbb{E}(Y_1)$ almost surely. Moreover,

$$\sum_{n=1}^{\infty}\mathbb{P}(Y_n\neq Z_n)=\sum_{n=1}^{\infty}\mathbb{P}(Y_1\geq n)\leq\int_{\mathbb{R}^+}\mathbb{P}(Y_1\geq x)\,\mathrm{d}x=\int_{\mathbb{R}^+}\mathbb{P}(Y_1>x)\,\mathrm{d}x=\mathbb{E}(Y_1)<\infty,$$

where we have used Theorem 3.6 and the fact that $\lambda(Q)=0$ if $Q\subseteq\mathbb{R}$ is countable in the second last inequality, and Proposition 5.11 in the last equality. By Borel-Cantelli lemma (Lemma 6.5) again, we have $\mathbb{P}(Y_n=Z_n,\, n\geq N \text{ for some N})=1$, and thus $(T_{\ell_n}/\ell_n)_{n\in\mathbb{N}}$ also converges to $(S_{\ell_n}/\ell_n)_{n\in\mathbb{N}}$ almost surely. It follows that $(S_{\ell_n}/\ell_n)_{n\in\mathbb{N}}$ converges to $\mathbb{E}(Y_1)$ almost surely. Finally, because $Y_k\geq 0$, for $k\in[\ell_n,\ell_{n+1}]$ we have

$$\frac{1}{\alpha}\frac{S_{\ell_n}}{\ell_n}\leq\frac{\ell_n}{k}\frac{S_{\ell_n}}{\ell_n}\leq\frac{S_k}{k}\leq\frac{\ell_{n+1}}{k}\frac{S_{\ell_{n+1}}}{\ell_{n+1}}\leq\alpha\frac{S_{\ell_{n+1}}}{\ell_{n+1}}.$$

Letting $k\to\infty$, we yield

$$\frac{1}{\alpha}\mathbb{E}(Y_1)\leq\liminf_{k\to\infty}\frac{S_k}{k}\leq\limsup_{k\to\infty}\frac{S_k}{k}\leq\alpha\mathbb{E}(Y_1).$$

Since $\alpha>1$ is arbitrary, we conclude the proof. □

cor:WeakLLN **Corollary 12.4.** *Under the condition of Theorem 12.2 or 12.3, $(\frac{1}{n}\sum_{k=1}^{n}Y_k)_{n\in\mathbb{N}}$ converges to a in measure.*

Proof. This is an immediate consequence of Theorem 12.2 or 11.1, or, Theorem 12.3 and 11.6. □

Remark 12.5. An important consequence of LLN is that, the frequency of event A happening under (pairwise, or mutually) independent trials is asymptotically the same as the probability of A as the number of trials tends to infinity. This justifies the definition of probability.

Proposition 12.6 (Glivenko-Cantelli)**.** *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of pairwise independent and identically distributed real-valued random variables. Let $F_n(r):=\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{(-\infty,r]}(Y_k)$ for $r\in\mathbb{N}$, i.e., F_n is the empirical CDF of n samples. Then,*

$$\lim_{n\to\infty}\sup_{r\in\mathbb{R}}|F_n(r)-F^Y(r)|=0,\ a.s..$$

Proof. To start with, for $j,k\in\mathbb{N}$, $j<k$, let $r_{j,k}:=\inf\{r\in\mathbb{R}:F^Y(r)\ge j/k\}$, we also set $r_{0,k}:=-\infty$ and $r_{k,k}:=\infty$. Thanks to Theorem 3.6 (a), for any j,k and $r\in[r_{j-1,k},r_{j,k}]$

$$\begin{aligned}\left|F_n(r)-F^Y(r)\right| &\le \max\left\{|F_n(r_{j-1,k})-F^Y(r_{j,k})|,|F_n(r_{j,k})-F^Y(r_{j-1,k})|\right\}\\ &\le \max\left\{|F_n(r_{j-1,k})-F^Y(r_{j-1,k})|,|F_n(r_{j,k})-F^Y(r_{j,k})|\right\}+\frac{1}{k}\\ &\le \sup_{r\in(r_{j,k})_{j\in\{1,\dots,k-1\}}}\left|F_n(r)-F^Y(r)\right|+\frac{1}{k}. \end{aligned}\tag{12.1}$$ eq:FnFY

Next, note that for any $r\in\mathbb{R}$, by strong LLN (Theorem 12.3) we have

$$\lim_{n\to\infty}F_n(r)=\lim_{n\to\infty}\sum_{k=1}^n\mathbb{1}_{(-\infty,r]}(Y_k)=\mathbb{P}^Y((-\infty,r])=F^Y(r),\ a.s..$$

and thus

$$\lim_{n\to\infty}\sup_{r\in(r_{j,k})_{j\in\{1,\dots,k-1\}}}\left|F_n(r)-F^Y(r)\right|=0,\ a.s..$$

This together with (12.1) implies that

$$\limsup_{n\in\mathbb{N}}\left|F_n(r)-F^Y(r)\right|\le\frac{1}{k},\ a.s..$$

Since $k\in\mathbb{N}$ is arbitrary, we complete the proof. □

13 Conditional Expectation

Definition 13.1. We consider the probability space $(\Omega,\mathscr{A},\mathbb{P})$.

- Let $Z\in\mathcal{L}^1$. Let $A\in\mathscr{A}$ satisfy $\mathbb{P}(A)>0$. The conditional expectation of Y given A is the quantity

$$\mathbb{E}(Z|A)=\frac{\mathbb{E}(\mathbb{1}_AZ)}{\mathbb{P}(A)}.$$

- Let $Z:(\Omega,\mathscr{A})\to(\mathbb{Z},\mathscr{Z})$. The conditional distribution of Z given A is a probability measure (why?) on $(\mathbb{Z},\mathscr{Z})$ satisfying

$$\mathbb{P}(B|A)=\frac{\mathbb{E}(\mathbb{1}_A\mathbb{1}_{\{Y\in B\}})}{\mathbb{P}(A)}=\frac{\mathbb{E}(\mathbb{1}_A\mathbb{1}_B(Y))}{P(A)}.$$

- Let Y be an $\mathbb{R}^d$-valued random variable and Z be an $\mathbb{R}^n$-valued random variable. Suppose (Y,Z) as an $\mathbb{R}^{d+n}$-valued random variable has PDF $f_{(Y,Z)}$. The marginal PDF of Y in this case is defined as

$$f_Y(y):=\int_{\mathbb{R}^n}f_{(Y,Z)}(y,z)\,\mathrm{d}z,\quad y\in\mathbb{R}^d.$$

The conditional PDF of Z given Y is

$$f_{Z|Y}(z|y):=\frac{f_{(Y,Z)}(y,z)}{f_Y(y)}=\frac{f_{(Y,Z)}(y,z)}{\int_{\mathbb{R}^n}f_{(Y,Z)}(y,z)\,\mathrm{d}z},\quad y\in\mathbb{R}^d,\ z\in\mathbb{R}^n.$$

Proposition 13.2. *Let Z be a non-negative (or integrable) real-valued random variable and $D : (\Omega, \mathscr{A}) \to (\mathbb{N}, 2^{\mathbb{N}})$ be a discrete random variable. Let $H := \{k \in \mathbb{N} : \mathbb{P}(D = k) > 0\}$. Then,*

$$\mathbb{E}(\mathbb{1}_{\{D\in B\}}Z) = \mathbb{E}\left(\mathbb{1}_{\{D\in B\}}\widetilde{Z}\right), \quad B \subseteq \mathbb{N}$$

where

$$\widetilde{Z}(\omega) := \sum_{k\in H} \mathbb{1}_{\{k\}}(D(\omega))\mathbb{E}(Z|\{D = k\}) = \sum_{k\in H} \mathbb{1}_{\{k\}}(D(\omega))\frac{\mathbb{E}(\mathbb{1}_{\{k\}}Z)}{\mathbb{P}(D = k)}.$$

If there is $\widetilde{Z}'(\omega) = \sum_{k\in\mathbb{N}} b_k \mathbb{1}_{\{k\}}(D(\omega))$ that also satisfies

$$\mathbb{E}(\mathbb{1}_{\{D\in B\}}Z) = \mathbb{E}\left(\mathbb{1}_{\{D\in B\}}\widetilde{Z}'\right), \quad B \subseteq \mathbb{N},$$

then $\widetilde{Z} = \widetilde{Z}'$ (we recall that this means $\mathbb{P}(\widetilde{Z} = \widetilde{Z}') = 1$).

Proof. Regarding the first statement, it is sufficient to prove for $B = \{i\}$ with $i \in H$. As it the statement is true for such B, then we can use monotone convergence to extend the statement for any $B \subseteq N$. Note that

$$\mathbb{E}(\mathbb{1}_{\{D=i\}}\widetilde{Z}) = \mathbb{P}(D = i)\frac{\mathbb{E}(\mathbb{1}_{\{D=i\}}Z)}{\mathbb{P}(D = i)}\mathbb{E}(\mathbb{1}_{\{D=i\}}Z).$$

This proves the first statement. Regarding the second statement, we again take $B = \{i\}$, where $i \in H$. Then, by hypothesis, we yield $\mathbb{E}(\mathbb{1}_{\{D=i\}}Z) = \mathbb{E}(\mathbb{1}_{\{D=i\}})b_i$, i.e.,

$$b_k = \frac{\mathbb{E}(\mathbb{1}_{\{k\}}Z)}{\mathbb{P}(D = k)}, \quad k \in H.$$

Finally, because $\mathbb{P}(D \notin H) = 0$, and $\mathbb{P}(\widetilde{Z} - \widetilde{Z}' \neq 0) \le \mathbb{P}(D \notin H)$. The proof is complete. □

Proposition 13.3. *Let Y be an $\mathbb{R}^d$-valued random variable and Z be an $\mathbb{R}^n$-valued random variable. Suppose (Y, Z) as an $\mathbb{R}^{d+n}$-valued random variable has PDF $f_{(Y,Z)}$. Fix a bounded $h : (\mathbb{R}^n, (\mathbb{R}^n)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then,*

$$\mathbb{E}(g(Y)h(Z)) = \mathbb{E}(g(Y)S) \quad \textit{for any bounded} \quad g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

where

$$S(\omega) := \int_{\mathbb{R}^n} h(z) f_{Z|Y}(z|Y(\omega))\, \mathrm{d}z = \frac{\int_{\mathbb{R}^n} h(z) f_{(Y,Z)}(Y(\omega), z)\, \mathrm{d}z}{\int_{\mathbb{R}^n} f_{(Y,Z)}(Y(\omega), z)\, \mathrm{d}z}.$$

Suppose $S' : (\Omega, \mathscr{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies $S'(\omega) = \widetilde{h}(Y(\omega))$ for some $\widetilde{h} : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and

$$\mathbb{E}(g(Y)h(Z)) = \mathbb{E}(g(Y)S') \quad \textit{for any bounded} \quad g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

then $S = S'$.

Proof. DIY. □

The two propositions above motivate a more general definition of conditional expectation. For notional convenience, we define $\mathcal{L}(\Omega, \mathscr{A}, \mathbb{P})$ as the set of random variables Z such that $\mathbb{E}(Z^+)$ or $\mathbb{E}(Z^-)$ is finite. *(The rest of this section depends heavily on σ-algebra. The related context is optional and will NOT appear in the exam.)*

def:CondExpn **Definition 13.4.** Let $\mathscr{G}$ be a sub-σ-algebra fo $\mathscr{A}$ and $Z \in \mathcal{L}(\Omega, \mathscr{A}, \mathbb{P})$. The conditional expectation of Z given $\mathscr{G}$, denoted by $\mathbb{E}(Z|\mathscr{G})$, is a $\mathscr{G}$-measurable random variable satisfying

$$\mathbb{E}(\mathbb{1}_B Z) = \mathbb{E}(\mathbb{1}_B \mathbb{E}(Z|\mathscr{G})) \quad \text{for any} \quad B \in \mathscr{G}.$$

The proof of existence of Definition 13.4 is beyond the scope of this course. Below is a sketch. We first assume $Z \in \mathcal{L}^2(\Omega, \mathscr{A}, \mathbb{P})$. Note $\mathcal{L}^2(\Omega, \mathscr{A}, \mathbb{P})$ is a Hilbert space and $\mathcal{L}^2(\Omega, \mathscr{G}, \mathbb{P})$ is a sub-Hilbert space. In this case, $\mathbb{E}(Z|\mathscr{G})$ is defined as the projection of Z onto $\mathcal{L}^2(\Omega, \mathscr{G}, \mathbb{P})$. Then, we use the approximation $Z \wedge n$ to extend the definition for Z satisfying $\mathbb{E}(Z^+) < \infty$ or $\mathbb{E}(Z^-) < \infty$. During this procedure, we also yield the following technical lemma.

em:CondExpnBasic **Lemma 13.5.** *Let $\mathscr{G}$ be a sub-σ-algebra fo $\mathscr{A}$ and $Z, Z' \in \mathcal{L}(\Omega, \mathscr{A}, \mathbb{P})$. Then,*

(a) If $0 \le Z \le Z'$, then $\mathbb{E}(Z|\mathscr{G}) \le \mathbb{E}(Z'|\mathscr{G})$;

(b) $Z \in \mathcal{L}^2(\Omega, \mathscr{A}, \mathbb{P})$ implies $Z \in \mathcal{L}^2(\Omega, \mathscr{G}, \mathbb{P})$, and $Z \in \mathcal{L}^1(\Omega, \mathscr{A}, \mathbb{P})$ implies $Z \in \mathcal{L}^1((\Omega, \mathscr{G}, \mathbb{P}))$.

In particular, Lemma 13.5 implies that $\mathbb{E}(Z|\mathscr{G}) \ge 0$ if $Z \ge 0$.

The following lemma regard the uniqueness of the definition.

UniquenessCriteria **Lemma 13.6.** *Let $\widetilde{Z}, \widetilde{Z}' \in \mathcal{L}(\Omega, \mathscr{G}, \mathbb{P})$. Suppose*

$$\mathbb{E}(\mathbb{1}_A \widetilde{Z}) = \mathbb{E}(\mathbb{1}_A \widetilde{Z}'), \quad A \in \mathscr{G}.$$

Then, $\widetilde{Z} = \widetilde{Z}'$.

Proof. Let $A_k^+ := \{\omega \in \Omega : \widetilde{Z}(\omega) - \widetilde{Z}'(\omega) > \frac{1}{k}\}$. Note that $A_k \in \mathscr{G}$. Then,

$$\mathbb{P}(A_k) = k\mathbb{E}(\frac{1}{k}\mathbb{1}_{A_k}) \le k\mathbb{E}((\widetilde{Z} - \widetilde{Z}')\mathbb{1}_{A_k}) = 0.$$

Similarly, $\{\widetilde{Z}' - \widetilde{Z} > \frac{1}{k}\}$ also has zero probability. Consequently, let $\mathbb{P}(|\widetilde{Z}' - \widetilde{Z}| > \frac{1}{k}) = 0$. Finally, in view of Theorem 1.16 (e), we have $\mathbb{P}(|\widetilde{Z} - \widetilde{Z}'| > 0) = \mathbb{P}(\bigcup_{k \in \mathbb{N}}\{|\widetilde{Z}' - \widetilde{Z}| > \frac{1}{k}\}) = 0$. □

Caution: We emphasize that $\widetilde{Z} = \widetilde{Z}'$ in Lemma 13.6 means $\mathbb{P}(\widetilde{Z} = \widetilde{Z}') = 1$. This means that Definition 13.4 only specifies the random variable $\mathbb{E}(Z|\mathscr{G})$ almost surely instead of for any $\omega \in \Omega$.

The following result is an immediate consequence of Theorem 2.7.

Proposition 13.7. *Let $Y : (\Omega, \mathscr{A}) \to (\mathbb{Y}, \mathscr{Y})$ and $Z \in \mathcal{L}^1(\Omega, \mathscr{A}, \mathbb{P})$. Then, there is a $h : (\mathbb{Y}, \mathscr{Y}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mathbb{E}(Z|\sigma(Y)) = h(Y)$.*

The next theorem regards the basic properties of conditional expectation.

m:CondExpnBasic **Theorem 13.8.** *Let $Y, Z \in \mathcal{L}(\Omega, \mathscr{A}, \mathbb{P})$ and $\mathscr{G}, \mathscr{H}$ be two σ-algebra such that $\mathscr{H} \subseteq \mathscr{G} \subseteq \mathscr{A}$. The following is true:*

(a) $\mathbb{E}(\mathbb{E}(Z|\mathscr{G})|\mathscr{H}) = \mathbb{E}(Z|\mathscr{H})$*;*

(b) if $\mathscr{H} = \{\emptyset, \Omega\}$*, then* $\mathbb{E}(Z|\mathscr{H}) = \mathbb{E}(Z)$*;*

(c) if $Y, Z \in \mathcal{L}^1$*, then* $\mathbb{E}(aY + bZ|\mathscr{G}) = a\mathbb{E}(Y|\mathscr{G}) + b\mathbb{E}(Z|\mathscr{G})$ *for* $a, b \in \mathbb{R}$*.*

Proof. **(a)** The is an immediate consequence of Lemma 13.6 and the observation that, for any $A \in \mathscr{H}$,

$$\mathbb{E}(\mathbb{1}_A \mathbb{E}(\mathbb{E}(Z|\mathscr{G})|\mathscr{H})) = \mathbb{E}(\mathbb{1}_A \mathbb{E}(Z|\mathscr{G})) = \mathbb{E}(\mathbb{1}_A Z) = \mathbb{E}(\mathbb{1}_A \mathbb{E}(Z|\mathscr{H})).$$

(b) Note that real-valued $\{\emptyset, \Omega\}$-measurable random variable is a constant function of $\omega \in \Omega$. The statement follows by taking $A = \Omega$.
(c) The proof involves the procedure used for proving the existence of Definition 13.4. We refer to ...... for the detailed proof. □

The conditional version of limit theorems also holds true.

hm:CondLimitThm **Theorem 13.9.** *Let* $Z \in \mathcal{L}(\Omega, \mathscr{A}, \mathbb{P})$ *and* $(Z_n)_{n\in\mathbb{N}} \in (\Omega, \mathscr{A}, \mathbb{P})$*. Let* $\mathscr{G}$ *be a sub-*σ*-algebra of* $\mathscr{A}$*. The following is true:*

(a) (Monotone Convergence) if $Z_n \geq 0$ *for* $n \in \mathbb{N}$ *and* $(Z_n)_{n\in\mathbb{N}}$ *increases to* Z *almost surely, then*

$$\lim_{n\to\infty} \mathbb{E}(Z_n|\mathscr{G}) = \mathbb{E}(Z|\mathscr{G}), \text{ a.s.};$$

(b) (Fatou's Lemma) if $Z_n \geq 0$ *for* $n \in \mathbb{N}$*, then*

$$\mathbb{E}(\liminf_{n\to\infty} Z_n|\mathscr{G}) \leq \liminf_{n\to\infty} \mathbb{E}(Z|\mathscr{G}), \text{ a.s.};$$

(c) (Dominated Convergence) if $\lim_{n\in\mathbb{N}} Z_n = Z$ *almost surely, and* $|Z_n| \leq Y$ *for* $n \in \mathbb{N}$ *for some* $Y \in \mathcal{L}^1(\Omega, \mathscr{A}, \mathbb{P})$*, then*

$$\lim_{n\to\infty} \mathbb{E}(Z_n|\mathscr{G}) = \mathbb{E}(Z|\mathscr{G}).$$

Proof. **(a)** In view of Lemma 13.5 (a), we let $\widetilde{Z} := \lim_{n\to\infty} \mathbb{E}(Z_n|\mathscr{G})$ almost surely ($\widetilde{Z}$ may be ∞ with positive probability). Due to Remark 9.3, $\widetilde{Z}$ is $\mathscr{G}$-measurable. Then, by monotone convergence (Theorem 10.1),

$$\mathbb{E}(\widetilde{Z}\mathbb{1}_A) = \lim_{n\to\infty} \mathbb{E}(\mathbb{E}(Z_n|\mathscr{G})\mathbb{1}_A) = \lim_{n\to\infty} \mathbb{E}(Z_n \mathbb{1}_A) = \mathbb{E}(Z\mathbb{1}_A), \quad A \in \mathscr{G},$$

where we have used Definition 13.4 in the second equality. In view of Lemma 13.6, the proof is complete.
(b)&**(c)** The proof are analogous to the proof of Fatou's lemma(Theorem 10.2) and dominated convergence (Theorem 10.5). □

Theorem 13.10. *If* $Y \in \mathcal{L}^\infty(\Omega, \mathscr{G}, \mathbb{P})$ *and* $Z \in \mathcal{L}^1(\Omega, \mathscr{A}, \mathbb{P})$*, then* $\mathbb{E}(YZ|\mathscr{G}) = Y\mathbb{E}(Z|\mathscr{G})$*;*

Proof. We first $Y = \sum_{k=1}^{n} a_k \mathbb{1}_{A_k}$ with $A_k \in \mathscr{G}$ for $k = 1, \ldots, n$. Let $A \in \mathscr{G}$. Then,

$$\mathbb{E}(\mathbb{1}_A \mathbb{E}(YZ|\mathscr{G})) = \mathbb{E}(\mathbb{1}_A \sum_{k=1}^{n} a_k \mathbb{1}_{A_k} Z) = \sum_{k=1}^{k} \mathbb{E}(\mathbb{1}_{A\cap A_k} \mathbb{E}(Z|\mathscr{G})) = \mathbb{E}(\mathbb{1}_A Y \mathbb{E}(Z|\mathscr{G})).$$

This together with Lemma 13.6 proves the statement for simple Y and non-negative Z. Now suppose $Y \in \mathcal{L}^\infty(\Omega, \mathscr{G}, \mathbb{P})$ and $|Y| \le M$ for some $M > 0$. In view of Theorem 2.6, we let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of simple random variable converging to Y almost surely with $|Y_n| \le M$ for $k \in \mathbb{N}$. Thus, $(Y_n Z)_{n\in\mathbb{N}}$ converges to YZ almost surely and $|Y_n Z| \le M|Z| \in \mathcal{L}^1$. By conditional dominated convergence (Theorem 13.9 (c)), we have

$$\lim_{n\to\infty} \mathbb{E}(Y_n Z|\mathscr{G}) = \mathbb{E}(YZ|\mathscr{G}),\ a.s..$$

Consequently, $(Y_n \mathbb{E}(Z|\mathscr{G}))_{n\in\mathbb{N}}$ also converges to $Y\mathbb{E}(Z|\mathscr{G})$ almost surely. Moreover, $|Y_n \mathbb{E}(Z|\mathscr{G})| \le M|\mathbb{E}(Z|\mathscr{G})| \in \mathcal{L}^1$ due to Lemma 13.5 (b). The above together with dominated convergence (Theorem 10.5) and the proved statement for simple random variables implies

$$\mathbb{E}(\mathbb{1}_A \mathbb{E}(YZ|\mathscr{G})) = \lim_{n\to\infty} \mathbb{E}(\mathbb{1}_A \mathbb{E}(Y_n Z|\mathscr{G})) = \lim_{n\to\infty} \mathbb{E}(\mathbb{1}_A Y_n \mathbb{E}(Z|\mathscr{G})) = \mathbb{E}(\mathbb{1}_A Y \mathbb{E}(Z|\mathscr{G})).$$

Since $A \in \mathscr{G}$ is arbitrary, in view of Lemma 13.6, we conclude the proof. □

Now let $Z : (\Omega, \mathscr{A}) \to (\mathbb{Z}, \mathscr{Z})$. We proceed to investigate the function $(\omega, B) \mapsto \mathbb{E}(\mathbb{1}_B(Z)|\mathscr{G})(\omega)$ for $(\omega, B) \in \Omega \times \mathscr{Z}$. It is desirable to view such function as probability measure that depends on the randomness ω. However, because the random variable $\mathbb{E}(\mathbb{1}_B(Z)|\mathscr{G})$ is specified only almost surely, and $\mathscr{Z}$ is uncountable in general, there is no guarantee that $B \mapsto \mathbb{E}(\mathbb{1}_B(Z)|\mathscr{G})(\omega)$ satisfies countable additivity for almost every $\omega \in \Omega$. The following theorem resolve this issue when $\mathbb{Z}$ is a separable metric space.

:RegularCondDistn **Theorem 13.11.** *Let $\mathbb{Z}$ be a separable metric space endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{Z})$. Let $Z : (\Omega, \mathscr{A}) \to (\mathbb{Z}, \mathcal{B}(\mathbb{Z}))$ $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{A}$. Then, there is $P : \Omega \times \mathcal{B}(\mathbb{Z}) \to [0, 1]$ such that*

(i) for each $\omega \in \Omega$, $B \mapsto P(\omega, B)$ is a probability on $(\mathbb{Z}, \mathcal{B}(\mathbb{Z}))$;

(ii) for each $B \in \mathcal{B}(\mathbb{Z})$, $\omega \mapsto P(\omega, B)$ is $\mathscr{A}$-measurable;

(iii) for each $B \in \mathcal{B}(\mathbb{Z})$, $P(\omega, B) = \mathbb{E}(\mathbb{1}_B(Z)|\mathscr{G})(\omega)$ for almost every ω.

Proof. The proof is out-of-scrope. We refer to .... □

The P introduced in Theorem 13.11 is called the regular conditional distribution of Z given $\mathscr{G}$.

14 Weak Convergence of Probability

A Preliminaries

Let $\mathbb{X}$ be a non-empty space, with no specific structure. We will review some algebra of sets. We let $\mathcal{I}$ be a set of indexes and $A, B, A_1, A_2, \cdots \subseteq \mathbb{X}$

Definition A.1.

- $\bigcup_{i\in\mathcal{I}} A_i := \{x \in \mathbb{X} : x \in A_i \text{ for some } i \in \mathcal{I}\}$;
- $\bigcap_{i\in\mathcal{I}} A_i := \{x \in \mathbb{X} : x \in A_i \text{ for all } i \in \mathcal{I}\}$;
- $A^c := \{x \in \mathbb{X} : x \notin A\}$ and $B \setminus A = B \cap A^c$.

thm:SetOp **Theorem A.2.** *The following is true:*

(a) *if* $A \subseteq B$, $A \cap B = A$ *and* $A \cup B = B$*;*

(b) $\bigcap_{i\in\mathcal{I}} A_i \subseteq \bigcup_{i\in\mathcal{I}} A_i$*;*

(c) *if* $A_1, A_2, \dots \subseteq B$ *then* $\bigcup_{i\in\mathcal{I}} A_i \subseteq B$*;*

(d) *if* $A_1, A_2, \dots \supseteq B$ *then* $\bigcap_{i\in\mathcal{I}} A_i \supseteq B$*;*

(e) $A \cup B = B \cup A$, $A \cap B = B \cap A$, *this also extends to infinite union/intersection;*

(f) $\left(\bigcup_{i\in\mathcal{I}} A_i\right) \cap B = \bigcup_{i\in\mathcal{I}}(A_i \cap B)$*;*

(g) $\left(\bigcap_{i\in\mathcal{I}} A_i\right) \cup B = \bigcap_{i\in\mathcal{I}}(A_i \cup B)$*;*

(h) $\left(\bigcup_{i\in\mathcal{I}} A_i\right)^c = \bigcap_{i\in\mathcal{I}} A_i^c$*;*

(i) $\left(\bigcap_{i\in\mathcal{I}} A_i\right)^c = \bigcup_{i\in\mathcal{I}} A_i^c$*.*

In what follows, we let $r \in \mathbb{R}$ and $(r_n)_{n\in\mathbb{R}} \subseteq \mathbb{R}$.

Definition A.3.

- We say $\underline{(r_n)_{n\in\mathbb{R}} \text{ converges to } r}$ if for any $\varepsilon > 0$, there is $N \in \mathbb{N}$ such that $|r_n - r| < \varepsilon$ for any $n \geq N$. We denote this by $\lim_{n\to\infty} r_n = r$.
- We also say $\underline{(r_n)_{n\in\mathbb{R}} \text{ converges to } \infty}$ if for any $M > 0$, there is $N \in \mathbb{N}$ such that $r_n \geq M$ for $n \geq N$. We denote this by $\lim_{n\to\infty} r_n = \infty$.
- We also say $\underline{(r_n)_{n\in\mathbb{R}} \text{ converges to } -\infty}$ if for any $M > 0$, there is $N \in \mathbb{N}$ such that $r_n \leq -M$ for $n \geq N$. We denote this by $\lim_{n\to\infty} r_n = -\infty$.

Theorem A.4. *Let* $(r_n)_{n\in\mathbb{R}} \subseteq \mathbb{R}$*. Suppose* $(r_n)_{n\in\mathbb{R}} \subset \mathbb{R}$ *is non-decreasing (or, non-increasing) and there is* $M > 0$ *such* $|r_n| \leq M$ *for all* $n \in \mathbb{N}$*. Then, there is* $r \in \mathbb{R}$ *such that* $\lim_{n\to\infty} r_n = r$*.*

thm:MonoSeq **Definition A.5.** Let $\mathcal{I}$ be a set of indexes. Let $(r_i)_{i\in\mathcal{I}} \subseteq \mathbb{R}$. We define $\sup_{i\in\mathcal{I}} r_i$ as a number $\bar{r} \in \mathbb{R}$ such that

(i) $\bar{r} \geq r_i$ for all $i \in \mathcal{I}$;

(ii) for any $r' \in \mathbb{R}$ such that $r' \geq r_i$ for all $i \in \mathcal{I}$, we have $r' \geq \bar{r}$.

Similarly, we define $\inf_{i\in\mathcal{I}} r_i$ as a number $\underline{r} \in \mathbb{R}$ such that

(i) $\underline{r} \leq r_i$ for all $i \in \mathcal{I}$;

(ii) for any $r' \in \mathbb{R}$ such that $r' \leq r_i$ for all $i \in \mathcal{I}$, we have $r' \leq \underline{r}$.

We also set $\sup_{i\in\mathcal{I}} r_i := \infty$ if $(r_i)_{i\in\mathcal{I}} \subseteq \mathbb{R}$ is unbounded from above; $\inf_{i\in\mathcal{I}} r_i := -\infty$ if $(r_i)_{i\in\mathcal{I}} \subseteq \mathbb{R}$ is unbounded from below.

Theorem A.6. *If there is $\overline{M} \in \mathbb{R}$ such that $\overline{M} \geq r_i$ for all $i \in \mathcal{I}$, then $\sup_{i\in\mathcal{I}} r_i$ exists uniquely. If there is $\underline{M} \in \mathbb{R}$ such that $\underline{M} \leq r_i$ for all $i \in \mathcal{I}$, then $\inf_{i\in\mathcal{I}} r_i$ exists uniquely.*

Sometimes, it is convenient to consider the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$ with the following rules $0 \times \infty = 0$, $0 \times (-\infty) = 0$, $a \pm \infty = \pm\infty$ and $a \times (\pm\infty) = \operatorname{sgn}(a) \cdot \infty$ for $a \in \mathbb{R}$.

We define

$$\limsup_{n\to\infty} r_n := \lim_{n\to\infty} \sup_{k\geq n} r_k \quad \text{and} \quad \liminf_{n\to\infty} r_n := \lim_{n\to\infty} \inf_{k\geq n} r_k.$$

Note that $(\sup_{k\geq n} r_k)_{n\in\mathbb{N}}$ is non-increasing and $(\inf_{k\geq n} r_k)_{n\in\mathbb{N}}$ is non-decreasing. In view of Theorem A.5, $\limsup_{n\to\infty} r_n$ and $\liminf_{n\to\infty} r_n$ are well-defined and takes finite value if there is $N, M \in \mathbb{N}$ such that $|r_n| \leq M$ for $n \geq N$. It is clear that $\limsup_{n\to\infty} r_n \geq \liminf_{n\to\infty} r_n$.

Theorem A.7. *$(r_n)_{n\in\mathbb{N}}$ converges (to some $r \in \mathbb{R}$) if and only if $\liminf_{n\to\infty} r_n = \limsup_{n\to\infty} r_n$ and they are finite.*

Theorem A.8. *$(r_n)_{n\in\mathbb{N}}$ converges (to some $r \in \mathbb{R}$) if and only if all the subsequences converges (to the same r).*

Below we recall the definitions of pointwise convergence of function. We let $\mathbb{X}$ be an abstract space with no specific structure. We let f be a real-valued function on $\mathbb{X}$ and $(f_n)_{n\in\mathbb{N}}$ a sequence of real-valued functions on $\mathbb{X}$.

Definition A.9. We say $\underline{(f_n)_{n\in\mathbb{N}} \text{ converges (pointwise) to } f}$ if $\lim_{n\to\infty} f_n(x) = f(x)$ for $x \in \mathbb{X}$. We also define two functions $\limsup_{n\to\infty} f_n$ and $\liminf_{n\to\infty} f_n$ as

$$\limsup_{n\to\infty} f_n(x) := \lim_{n\to\infty} \sup_{k\geq n} f_k(x) \quad \text{and} \quad \liminf_{n\to\infty} f_n(x) := \lim_{n\to\infty} \inf_{k\geq n} f_k(x), \quad x \in \mathbb{X},$$

respectively.

Metric space TBA