Notes on Probability Theory

Kai-Jyun Wang*

Fall 2025

The notes are based on the lecture of Prof. David Anderson at University of Wisconsin-Madison in 2025-2026. The course structure mainly follows Durrett. The course assumes a certain amount of knowledges in real analysis. For some classic results in real analysis, one can refer to my notes on real analysis.

*National Taiwan University, Department of Economics.

Contents

1. Probability Space

1.1. Probability Space

Definition 1.1
*Let Ω be a set. A collection of subsets $\mathcal{F}$ forms a σ-**algebra** if*

(a) $\varnothing \in \mathcal{F}$.

(b) $A \in \mathcal{F}$ *implies* $A^c \in \mathcal{F}$.

(c) *If* $A_i \in \mathcal{F}$ *are countably many sets,* $\cup_i A_i \in \mathcal{F}$.

The dual $(\Omega, \mathcal{F})$ *is called a **measurable space** and the sets falling in* $\mathcal{F}$ *are said to be **measurable**.*

Definition 1.2
Let $(\Omega, \mathcal{F})$ *be a measurable space. A set function* $\mu : \mathcal{F} \to [0, \infty]$ *is a **measure** if*

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

(b) *For countably many disjoint* $A_i \in \mathcal{F}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.

The triple $(\Omega, \mathcal{F}, \mu)$ *is called a **measure space**.*

Definition 1.3
*A **probability space** is a measure space* $(\Omega, \mathcal{F}, \mathrm{P})$ *such that* $\mathrm{P}(\Omega) = 1$.

Lemma 1.4
Let $\mathcal{S}$ *be a collection of sets. Then there exists the smallest* σ*-algebra containing* $\mathcal{S}$.

Proof. Let $\mathcal{F}$ be the intersection of all σ-algebra containing $\mathcal{S}$. $\mathcal{F}$ is non-empty since the power set is a σ-algebra containing $\mathcal{S}$. Now it is clear that $\varnothing \in \mathcal{F}$ since $\varnothing \in \mathcal{A}$ for every σ-algebra $\mathcal{A}$ contiaining $\mathcal{S}$. If $A \in \mathcal{F}$, $A \in \mathcal{A}$ for all $\mathcal{A}$ containing $\mathcal{S}$ and $A^c \in \mathcal{A}$ for all $\mathcal{A}$. Thus $A^c \in \mathcal{F}$. Finally, if $A_i \in \mathcal{F}$ are countably many sets, then each A_i lies in every $\mathcal{A}$ containing $\mathcal{S}$; so does $\cup_i A_i$ and thus $\cup_i A_i \in \mathcal{F}$. The minimality follows by the construction of $\mathcal{F}$. ■

Definition 1.5
For any collection of sets $\mathcal{S}$, *the smallest* σ*-algebra is denoted as* $\sigma(\mathcal{S})$.

Theorem 1.6
Let $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Then*

(a) *If* $A, B \in \mathcal{F}$ *and* $A \subset B$, *then* $\mathrm{P}(A) \le \mathrm{P}(B)$.

(b) *For countably many* $A_i \in \mathcal{F}$, $\mathrm{P}(\cup_i A_i) \le \sum_i \mathrm{P}(A_i)$.

(c) *If* $A_i \nearrow A$, $\mathrm{P}(A_i) \to \mathrm{P}(A)$.

(d) *If* $A_i \searrow A$, $\mathrm{P}(A_i) \to \mathrm{P}(A)$.

Proof. (a) and (b) are clear. For (c), write $E_i = A_i - A_{i-1}$ and $A_0 = \varnothing$. Then since E_i are disjoint and $A_n = \cup_{i=1}^n E_i$,

$$\mathrm{P}(A_n) = \mathrm{P}\big(\cup_{i=1}^n E_i\big) = \sum_{i=1}^n \mathrm{P}(E_i) \to \sum_i \mathrm{P}(E_i) = \mathrm{P}(\cup_i E_i) = \mathrm{P}(A)$$

as $n \to \infty$.

For (d), note that $A_i^c \nearrow A^c$. Thus $1 - \mathrm{P}(A_i) = \mathrm{P}(A_i^c) \to \mathrm{P}(A^c) = 1 - \mathrm{P}(A)$. Thus $\mathrm{P}(A_i) \to \mathrm{P}(A)$. ∎

Definition 1.7
*The **Borel** σ**-algebra** is the σ-algebra generated by all open sets.*

Definition 1.8
Let P *be a probability measure on* $(\mathbb{R}, \mathcal{B})$*. The **distribution function** F is defined as*

$$F(x) = \mathrm{P}((-\infty, x])$$

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

Proposition 1.9
The distribution function in $(\mathbb{R}, \mathcal{B})$ *satisfies that*

(a) $F(x) \le F(y)$ *for all* $x \le y$.

(b) $F(x) \to F(y)$ *as* $x \to y^+$.

(c) $F(-\infty) = 0$ *and* $F(\infty) = 1$.

Proof. For (a), note that $(-\infty, x] \subset (-\infty, y]$ and

$$F(x) = \mathrm{P}((-\infty, x]) \le \mathrm{P}((-\infty, y]) = F(y).$$

For (b), notice that for $x_n \to y^+$, $(-\infty, x_n] \searrow (-\infty, y]$. Hence

$$F(x_n) = \mathrm{P}((-\infty, x_n]) \to \mathrm{P}((-\infty, y]) = F(y).$$

Similarly, taking $x_n \to \pm\infty$ gives (c). ∎

Definition 1.10
*A collection $\mathcal{S}$ of sets is called an **algebra** if*

(a) $\varnothing \in \mathcal{S}$.

(b) If $A \in \mathcal{S}$*, then* $A^c \in \mathcal{S}$.

(c) If $A, B \in \mathcal{S}$*, then* $A \cup B \in \mathcal{S}$.

Remark
An algebra is closed under finite unions. It is also clear that a σ-algebra is an algebra, while the converse is not true. An example is the collection of all finite unions of intervals in $\mathbb{R}$.

Definition 1.11
A collection $\mathcal{S}$ of sets is called a ***semi-algebra*** *if*

(a) *If $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$.*

(b) *If $A \in \mathcal{S}$, then A^c can be written as a finite disjoint union of sets in $\mathcal{S}$.*

Remark
A semi-algebra must contain $\varnothing$ since for any $A \in \mathcal{S}$, $A^c = \cup_i A_i$, where $A_i \in \mathcal{S}$ are disjoint. Then $A \cap A_1 = \varnothing \in \mathcal{S}$.

Remark
An example of being a semi-algebra but not an algebra is the collection of all intervals of the form $(a_i, b_i]$ for $-\infty \leq a_i < b_i \leq \infty$ with the empty set.

Lemma 1.12
If $\mathcal{S}$ is a semi-algebra, then $\overline{\mathcal{S}} = \{$finite disjoint unions of sets in $\mathcal{S}\}$ forms an algebra.

Proof. It has been shown that $\varnothing \in \mathcal{S}$. For $A, B \in \overline{\mathcal{S}}$, write $A = \cup_{i=1}^n A_i$ and $B = \cup_{j=1}^m B_j$ for disjoint $A_i, B_j \in \mathcal{S}$, respectively. Then $A \cap B = \cup_{i,j}(A_i \cap B_j) \in \overline{\mathcal{S}}$. Thus $\overline{\mathcal{S}}$ is closed under intersection. Now if $A \in \overline{\mathcal{S}}$, $A = \cup_{i=1}^n A_i$ for disjoint $A_i \in \mathcal{S}$. Then $A^c = \cap_{i=1}^n A_i^c$. By the definition of semi-algebra, A_i^c can be written as a finite disjoint union of sets in $\mathcal{S}$ and thus $A_i^c \in \overline{\mathcal{S}}$. Since $\overline{\mathcal{S}}$ is closed under finite intersection, $A^c = \cap_{i=1}^n A_i^c \in \overline{\mathcal{S}}$. Finally, for $A, B \in \overline{\mathcal{S}}$, $A \cup B = (A^c \cap B^c)^c \in \overline{\mathcal{S}}$. We conclude that $\overline{\mathcal{S}}$ is indeed an algebra. ■

Definition 1.13
Suppose $\mathcal{S}$ is a semi-algebra. $\overline{\mathcal{S}} = \{$finite disjoint unions of sets in $\mathcal{S}\}$ is called the ***algebra generated by*** *$\mathcal{S}$.*

Definition 1.14
Let $\mathcal{S}$ be an algebra. A set function $\mu_0 : \mathcal{S} \to [0, \infty]$ is called a ***premeasure*** *if*

(a) *$\mu_0(\varnothing) = 0$.*

(b) *For countable disjoint $A_i \in \mathcal{S}$ such that $\cup_i A_i \in \mathcal{S}$,*

$$\mu_0(\cup_i A_i) = \sum_i \mu_0(A_i).$$

Theorem 1.15
Let ν be a set function on a semi-algebra $\mathcal{S}$ such that $\nu(\varnothing) = 0$. Suppose that

(a) *if $A \in \mathcal{S}$ and $A = \cup_{i=1}^n A_i$ for disjoint $A_i \in \mathcal{S}$, then $\nu(A) = \sum_{i=1}^n \nu(A_i)$;*

(b) *if $A_i \in \mathcal{S}$ are countably many sets and $A = \cup_i A_i \in \mathcal{S}$, then $\nu(A) \leq \sum_i \nu(A_i)$.*

Then ν can be extended to a unique premeasure μ_0 on the algebra generated by $\mathcal{S}$.

Proof. We first show the existence. From lemma 1.12 we know that $\mathcal{S}$ generates an algebra $\mathcal{A} = \{$finite disjoint union of sets in $\mathcal{S}\}$. Define our candidate μ_0 by $\mu_0(A) = \sum_i \nu(A_i)$ for

$A = \cup_i A_i$ where $A_i \in \mathcal{S}$ are disjoint. To see that μ_0 is well-defined, suppose $A = \cup_i B_i$ for disjoint $B_i \in \mathcal{S}$. Observe that

$$A_i = \cup_j (A_i \cap B_j) \quad \text{and} \quad B_j = \cup_i (A_i \cap B_j)$$

are finite disjoint unions. Then

$$\sum_i \nu(A_i) = \sum_i \sum_j \nu(A_i \cap B_j) = \sum_j \sum_i \nu(A_i \cap B_j) = \sum_j \nu(B_j)$$

by (a). Thus μ_0 is well-defined.

Now we check that μ_0 is a premeasure. Clearly $\mu_0(\varnothing) = 0$. For finitely many disjoint $A_i \in \mathcal{A}$ such that $\cup_i A_i \in \mathcal{A}$, we can write $A_i = \cup_j B_{ij}$ for disjoint $B_{ij} \in \mathcal{S}$. Then (a) implies that

$$\mu_0(\cup_i A_i) = \mu_0(\cup_{i,j} B_{ij}) = \sum_{i,j} \nu(B_{ij}) = \sum_i \sum_j \mu_0(B_{ij}) = \sum_i \mu_0(A_i).$$

Next, for countably many disjoint $A_i \in \mathcal{A}$ such that $A = \cup_i A_i \in \mathcal{A}$, write $A_i = \cup_j B_{ij}$, where $B_{ij} \in \mathcal{S}$ are finite disjoint for each i. Then $\mu_0(A_i) = \sum_j \nu(B_{ij})$ and

$$\sum_i \mu_0(A_i) = \sum_i \sum_j \nu(B_{ij}).$$

Without loss of generality, we may choose A_i to be those in $\mathcal{S}$ since otherwise we can replace A_i by B_{ij}. We assume that $A_i \in \mathcal{S}$ from now on. Since $A \in \mathcal{A}$, $A = \cup_i C_i$ for finite disjoint $C_i \in \mathcal{S}$. $C_i = \cup_j (C_i \cap A_j)$. Thus (b) gives that

$$\nu(C_i) \leq \sum_j \nu(C_i \cap A_j).$$

Then

$$\mu_0(A) = \sum_i \nu(C_i) \leq \sum_i \sum_j \nu(C_i \cap A_j) = \sum_j \sum_i \nu(C_i \cap A_j) = \sum_j \nu(A_j) = \sum_j \mu_0(A_j).$$

For the opposite inequality, set $B_n = \cup_{i=1}^n A_i$ and $C_n = A - B_n$. Since $\mathcal{A}$ is an algebra, $C_n \in \mathcal{A}$ and the finite additivity shows that

$$\mu_0(A) = \sum_{i=1}^n \mu_0(A_i) + \mu_0(C_n) \geq \sum_{i=1}^n \mu_0(A_i).$$

Taking $n \to \infty$ gives the desired inequality and thus μ_0 is σ-additive on $\mathcal{A}$.

Finally, if μ_1 is another premeasure on $\mathcal{A}$ extending ν, then for $A = \cup_i A_i$ for disjoint $A_i \in \mathcal{S}$,

$$\mu_1(A) = \sum_i \nu(A_i) = \mu_0(A).$$

$\blacksquare$

Definition 1.16
*A collection of sets $\mathcal{P}$ is called a π-**system** if $A, B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P}$.*

Definition 1.17
*A collection of sets $\mathcal{L}$ is called a λ-**system** if*

(a) $\Omega \in \mathcal{L}$.

(b) If $A, B \in \mathcal{L}$ and $A \subset B$, then $B - A \in \mathcal{L}$.

(c) If $A_i \in \mathcal{L}$ and $A_i \nearrow A$, then $A \in \mathcal{L}$.

Theorem 1.18 (Sierpiński-Dynkin π-λ)
If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system containing $\mathcal{P}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.

Proof. First we show that a collection $\mathcal{S}$ is a σ-algebra if and only if it is both a π-system and a λ-system. Suppose first that $\mathcal{S}$ is a π-system and a λ-system. $\emptyset = \Omega - \Omega \in \mathcal{S}$. If $A \in \mathcal{S}$, then $A^c = \Omega - A \in \mathcal{S}$. For $A, B \in \mathcal{S}$, $A \cup B = (A^c \cap B^c)^c \in \mathcal{S}$ since we have shown that $\mathcal{S}$ is closed under complement and intersection by being a π-system. Thus $\mathcal{S}$ is also closed under finite unions. If $A_i \in \mathcal{S}$ are countably many sets, let $B_n = \cup_{i=1}^n A_i \in \mathcal{S}$. Then $B_n \nearrow \cup_i A_i$ and thus $\cup_i A_i \in \mathcal{S}$.

Conversely, if $\mathcal{S}$ is a σ-algebra, then for $A, B \in \mathcal{S}$, $A \cap B = (A^c \cup B^c)^c \in \mathcal{S}$. Thus $\mathcal{S}$ is a π-system. If $A, B \in \mathcal{S}$ and $A \subset B$, then $B - A = B \cap A^c \in \mathcal{S}$. Finally, if $A_i \in \mathcal{S}$ and $A_i \nearrow A$, then $A = \cup_i (A_i - A_{i-1}) \in \mathcal{S}$ with $A_0 = \emptyset$. Thus $\mathcal{S}$ is a λ-system.

Now set $\mathcal{L}$ to be the smallest λ-system containing $\mathcal{P}$. It suffices to show that $\mathcal{L}$ is also a π-system and thus by the above conclusion, $\mathcal{L}$ is a σ-algebra containing $\mathcal{P}$; hence $\sigma(\mathcal{P}) \subset \mathcal{L}$.

To show that $\mathcal{L}$ is a π-system, let $A, B \in \mathcal{L}$. If $A, B \in \mathcal{P}$, $A \cap B \in \mathcal{P} \subset \mathcal{L}$. To extend the result for general $A, B \in \mathcal{L}$, we first fix $B \in \mathcal{P}$ and define

$$\mathcal{L}_B = \{A \mid A \cap B \in \mathcal{L}\}.$$

We claim that $\mathcal{L}_B$ is a λ-system containing $\mathcal{P}$. For $A \in \mathcal{P}$, $A \cap B \in \mathcal{L}$. Thus $\mathcal{P} \subset \mathcal{L}_B$. Clearly $\Omega \in \mathcal{L}_B$. If $E, F \in \mathcal{L}_B$ and $E \subset F$, then

$$(F - E) \cap B = (F \cap B) - (E \cap B) \in \mathcal{L}.$$

Thus $F - E \in \mathcal{L}_B$. Finally, if $E_i \in \mathcal{L}_B$ and $E_i \nearrow E$, then

$$E \cap B = \cup_i (E_i \cap B) \in \mathcal{L}.$$

Hence $E \in \mathcal{L}_B$ and we conclude that $\mathcal{L}_B$ is a λ-system. Since it is a λ-system containing $\mathcal{P}$, it also contains the smallest λ-system $\mathcal{L}$ with the intersection property. Thus $A \cap B \in \mathcal{L}$ whenever $A \in \mathcal{L}$ and $B \in \mathcal{P}$.

Next, fix $A \in \mathcal{L}$ and define $\mathcal{L}_A = \{B \mid A \cap B \in \mathcal{L}\}$. Clearly $\mathcal{L}_A$ contains $\mathcal{L}$ and $\Omega \in \mathcal{L}_A$. If $E, F \in \mathcal{L}_A$ and $E \subset F$, then

$$(F - E) \cap A = (F \cap A) - (E \cap A) \in \mathcal{L}.$$

Thus $F - E \in \mathcal{L}_A$. Finally, if $E_i \in \mathcal{L}_A$ and $E_i \nearrow E$, then

$$E \cap A = \cup_i (E_i \cap A) \in \mathcal{L}.$$

Hence $E \in \mathcal{L}_A$ and we conclude that $\mathcal{L}_A$ is a λ-system. Since it contains $\mathcal{L}$, $A, B \in \mathcal{L}$ implies $A \cap B \in \mathcal{L}$; in other words, $\mathcal{L}$ is a π-system and the proof is complete. ■

Corollary 1.19
Let μ and ν be two probability measures agreeing on a π-system $\mathcal{P}$, i.e., $\mu(A) = \nu(A)$ for all $A \in \mathcal{P}$. Then $\mu(A) = \nu(A)$ for all $A \in \sigma(\mathcal{P})$.

Proof. Put

$$\mathcal{L} = \{A \mid \mu(A) = \nu(A)\}.$$

We claim that $\mathcal{L}$ is a λ-system containing $\mathcal{P}$. It is clear that by our assumption, $\mathcal{P} \subset \mathcal{L}$ and $\Omega \in \mathcal{L}$. If $A, B \in \mathcal{L}$ and $A \subset B$, then

$$\mu(B - A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B - A).$$

Thus $B - A \in \mathcal{L}$. Finally, if $A_i \in \mathcal{L}$ and $A_i \nearrow A$, then

$$\mu(A) = \lim_{i \to \infty} \mu(A_i) = \lim_{i \to \infty} \nu(A_i) = \nu(A).$$

Hence $A \in \mathcal{L}$ and we conclude that $\mathcal{L}$ is a λ-system. By the Sierpiński-Dynkin π-λ theorem, $\sigma(\mathcal{P}) \subset \mathcal{L}$; in other words, μ and ν agree on $\sigma(\mathcal{P})$. ■

Definition 1.20
*A measure μ on a measurable space $(\Omega, \mathcal{F})$ is called **σ-finite** if there exists countable $A_i \in \mathcal{F}$ such that $\cup_i A_i = \Omega$ and $\mu(A_i) < \infty$.*

Definition 1.21
*A set function $\mu^* : 2^\Omega \to [0, \infty]$ is called an **outer measure** if*

(a) $\mu^*(\varnothing) = 0$.

(b) If $A \subset B$, then $\mu^*(A) \leq \mu^*(B)$.

(c) For countably many $A_i \subset \Omega$, $\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$.

Definition 1.22
*Let μ^* be an outer measure. A set $A \subset \Omega$ is said to be **Carathéodory measurable** or **μ^*-***

measurable *if for all* $E \subset \Omega$,

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

Lemma 1.23
Let μ^* *be an outer measure on* Ω. *Then the collection of all* μ^**-measurable sets forms a* σ*-algebra* $\mathcal{F}$ *and* $\mu^*|_{\mathcal{F}}$ *is a measure.*

Proof. Put

$$\mathcal{F} = \{A \subset \Omega \mid \mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) \text{ for all } E \subset \Omega\}.$$

We first show that $\mathcal{F}$ is a σ-algebra. Clearly $\varnothing \in \mathcal{F}$ and if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. For $A, B \in \mathcal{F}$, let $C = A \cup B$. The property of outer measure gives that $\mu^*(E) \le \mu^*(E \cap C) + \mu^*(E \cap C^c)$. To see the opposite inequality, note that $C = A \cup (B \cap A^c)$ and

$$\begin{aligned} \mu^*(E \cap C) + \mu^*(E \cap C^c) &\le \mu^*(E \cap A) + \mu^*(E \cap B \cap A^c) + \mu^*(E \cap A^c \cap B^c) \\ &= \mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E). \end{aligned}$$

Hence $C \in \mathcal{F}$ and $\mathcal{F}$ is closed under finite unions. For countable disjoint $A_i \in \mathcal{F}$ with $A = \cup_i A_i$, let $B_n = \cup_{i=1}^n A_i \in \mathcal{F}$. Then

$$\mu^*(E \cap A) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \ge \mu^*(E \cap B_n) = \sum_{i=1}^n \mu^*(E \cap A_i).$$

Taking $n \to \infty$ gives that

$$\mu^*(E \cap A) \ge \sum_i \mu^*(E \cap A_i) \ge \mu^*(E \cap A)$$

by the σ-subadditivity of outer measure. Hence $\mu^*(E \cap A) = \sum_i \mu^*(E \cap A_i)$. Note also that $E \cap A^c \subset E \cap B_n^c$ so $\mu^*(E \cap A^c) \le \mu^*(E \cap B_n^c)$. Thus

$$\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \ge \sum_{i=1}^n \mu^*(E \cap A_i) + \mu^*(E \cap A^c) \to \mu^*(E \cap A) + \mu^*(E \cap A^c) \ge \mu^*(E)$$

by the σ-subadditivity of outer measure. We conclude that $\mathcal{F}$ is a σ-algebra.

Finally, denote $\mu^*|_{\mathcal{F}}$ by μ. Clearly $\mu(\varnothing) = 0$. For countably many disjoint $A_i \in \mathcal{F}$ such that $A = \cup_i A_i \in \mathcal{F}$, let $B_n = \cup_{i=1}^n A_i \in \mathcal{F}$. Then

$$\mu(A) = \mu(B_n) + \mu(A \cap B_n^c) \ge \mu(B_n) = \sum_{i=1}^n \mu(A_i) \to \sum_i \mu(A_i) \ge \mu(A).$$

Hence $\mu(A) = \sum_i \mu(A_i)$ and μ is a measure on $\mathcal{F}$. ∎

Theorem 1.24 (Carathéodory Extension)
Let ν be a finitely additive, σ-subadditive set function on a semi-algebra $\mathcal{S}$ such that $\nu(\varnothing) = 0$. Then ν can be extended to a measure on $\sigma(\mathcal{S})$.

Proof. By theorem 1.15, ν can be extended to a premeasure μ_0 on the algebra $\mathcal{A}$ generated by $\mathcal{S}$. Define the outer measure by

$$\mu^*(A) = \inf\left\{\sum_i \mu_0(E_i)\ \middle|\ A \subset \cup_i E_i, E_i \in \mathcal{A}\right\}$$

for all $A \subset \Omega$ with the convention that $\inf\varnothing = \infty$. We check that μ^* is indeed an outer measure. Clearly $\mu^*(\varnothing) = 0$. If $A \subset B$, then any cover of B by sets in $\mathcal{A}$ is also a cover of A and hence $\mu^*(A) \le \mu^*(B)$. For countably many $A_i \subset \Omega$, we can find $\left\{E_{ij}\right\}_j$ covering A_i such that

$$\sum_j \mu_0(E_{ij}) \le \mu^*(A_i) + 2^{-i}\epsilon$$

for some $\epsilon > 0$. Then $\cup_{i,j}E_{ij}$ covers $\cup_i A_i$ and

$$\mu^*(\cup_i A_i) \le \sum_i\sum_j \mu_0(E_{ij}) \le \sum_i \mu^*(A_i) + \epsilon.$$

Since ϵ is arbitrary, we conclude that $\mu^*(\cup_i A_i) \le \sum_i \mu^*(A_i)$ and μ^* is indeed an outer measure.

It follows from lemma 1.23 that the collection of all μ^*-measurable sets forms a σ-algebra $\mathcal{F}$ and μ^* restricted on $\mathcal{F}$ is a measure. It is clear that $\mathcal{A} \subset \mathcal{F}$ and $\sigma(\mathcal{S}) \subset \sigma(\mathcal{A}) \subset \mathcal{F}$ and $\mu = \mu^*|_{\sigma(\mathcal{S})}$ is also a measure. Finally, for $A, A_i \in \mathcal{S}$ where A_i covers A,

$$\mu(A) = \mu^*(A) \le \nu(A) \le \sum_i \nu(A\cap A_i) \le \sum_i \nu(A_i).$$

Taking the infimum over all such covers, we get $\nu(A) = \mu^*(A)$ and μ is indeed an extension of ν. ∎

Remark
If the measures are probability measures, then we have that the extension is unique by corollary 1.19.

Theorem 1.25
If F is non-decreasing, right-continuous and satisfies that $F(-\infty) = 0$, $F(\infty) = 1$, then there is a unique probability measure such that

$$\mathrm{P}((-\infty, x]) = F(x).$$

Proof. Define

$$\mathcal{S} = \{(a, b] \mid -\infty \le a < b \le \infty\} \cup \{\varnothing\}.$$

It is clear that $\mathcal{S}$ is a semi-algebra. Define the set function $\mathrm{P} : \mathcal{S} \to [0,1]$ by

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

and $\mathrm{P}(\varnothing) = 0$. For disjoint, at most countable $(a_i, b_i] \in \mathcal{S}$, we define

$$\mathrm{P}(\cup_i (a_i, b_i]) = \sum_i \mathrm{P}((a_i, b_i]) = \sum_i F(b_i) - F(a_i).$$

It is clear that P is finitely additive. If $(a,b] = \cup_i (a_i, b_i]$ for disjoint $(a_i, b_i] \in \mathcal{S}$, we may assume without loss of generality that $a = a_1 < b_1 < b_2 < \cdots < b_n = b$ and

$$\mathrm{P}((a,b]) = F(b) - F(a) = \sum_i F(b_i) - F(a_i) = \sum_i \mathrm{P}((a_i, b_i]).$$

Hence P is σ-additive. It now follows from the Carathéodory extension theorem that P can be extended uniquely to a probability measure on $\sigma(\mathcal{S}) = \mathcal{B}$. ■

Remark
This theorem shows that the distribution function completely characterizes the probability measure. In other words, the term "distribution function" can refer to either the CDF or the probability measure.

1.2. Random Variable

Definition 1.26

Let Ω be a probability space. A ***random variable*** *X is a measurable function $X : \Omega \to (S, \mathcal{S})$, where $(S, \mathcal{S})$ is a measurable space.*

Remark

The codomain is often taken to be $(\mathbb{R}, \mathcal{B})$ or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, but it is also possible to define random functions, i.e., $(S, \mathcal{S})$ is a function space.

Definition 1.27

Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a random variable. The ***distribution*** *of X is the pushforward measure of P under X, i.e.,*

$$\mu_X(A) = \mathrm{P}(X \in A) = \mathrm{P}(X^{-1}(A)), \quad A \in \mathcal{S}.$$

Definition 1.28

Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B})$ be a random variable. The ***cumulative distribution function*** *of X is defined as*

$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X_1 \leq x_1, \ldots, X_d \leq x_d), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Proposition 1.29

Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be a random variable and F be its cumulative distribution function. Then,

(a) *F is non-decreasing, i.e., $x \leq y$ implies $F(x) \leq F(y)$;*

(b) *$F(-\infty) = 0$ and $F(\infty) = 1$;*

(c) *F is right-continuous, i.e., $\lim_{y \to x^+} F(y) = F(x)$;*

(d) *$F(x^-) = \mathrm{P}(X < x)$;*

(e) *$\mathrm{P}(X = x) = F(x) - F(x^-)$.*

Proof. (a) comes from that $\{X \leq x\} \subset \{X \leq y\}$ for $x \leq y$.

Take $a_n \to \infty$. Then $\{X \leq a_n\} \nearrow \Omega$ and $\{X \leq -a_n\} \searrow \emptyset$. By theorem 1.6, we have that

$$F(a_n) = \mathrm{P}(X \leq a_n) \to \mathrm{P}(\Omega) = 1, \quad F(-a_n) = \mathrm{P}(X \leq -a_n) \to \mathrm{P}(\emptyset) = 0.$$

(c) is similar to (b). Take $y_n \to x^+$, then $\{X \leq y_n\} \searrow \{X \leq x\}$. By theorem 1.6, we have that

$$F(y_n) = \mathrm{P}(X \leq y_n) \to \mathrm{P}(X \leq x) = F(x).$$

For (d), take $x_n \to x^-$, then $\{X \leq x_n\} \nearrow \{X < x\}$. By theorem 1.6, we have that

$$F(x_n) = \mathrm{P}(X \leq x_n) \to \mathrm{P}(X < x).$$

For (e), $P(X = x) = P(X \le x) - P(X < x) = F(x) - F(x^-)$. ■

Theorem 1.30
Let F be a non-decreasing, right-continuous function satisfying that $F(-\infty) = 0$ and $F(\infty) = 1$. Then there is a random variable X such that

$$F(x) = \mu_X((-\infty, x]).$$

Proof. Put $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}$, P be the Lebesgue measure and $X(\omega) = \sup\{x \mid F(x) < \omega\}$. Notice that

$$\begin{aligned} \{X \le x\} &= \{\omega \in \Omega \mid \sup\{y \mid F(y) < \omega\} \le x\} \\ &= \{\omega \in \Omega \mid \text{for all } y > x, F(y) \ge \omega\} \\ &= \{\omega \in \Omega \mid F(x) \ge \omega\}. \end{aligned}$$

Hence $P(X \le x) = P(\{\omega \in \Omega \mid \omega \le F(x)\}) = F(x)$. ■

Definition 1.31
If X and Y are random variables mapping to some measurable space $(S, \mathcal{S})$, then X and Y are said to be ***equal in distribution*** *if $\mu_X = \mu_Y$, denoted by $X \overset{d}{=} Y$.*

Definition 1.32
Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution F. $f : \mathbb{R} \to \mathbb{R}$ is said to be the ***density*** *of X if*

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

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

Remark
If f and g are both densities of X, then $f = g$ a.e.

Remark
If $\mu_X \ll \lambda$, where λ is the Lebesgue measure, then by Radon-Nikodym theorem, there is a density f such that

$$\mu_X(A) = \int_A f(x)d\lambda(x)$$

for all $A \in \mathcal{B}$. Or equivalently, F is absolutely continuous.

Example
Not all random variables have densities, even when its CDF is continuous. Consider the

Cantor function

$$F(x)=\begin{cases}\sum_n \frac{a_n}{2^n}, & x=\sum_n \frac{2a_n}{3^n}\in C \textit{ for some } \{a_n\}\in\{0,1\}^{\mathbb{N}}\\ \sup_{y\le x, y\in C} F(y), & x\in[0,1]-C\\ 0, & x<0\\ 1, & x>1\end{cases}$$

where C is the Cantor set. Then F is a valid CDF, but has no density.

Definition 1.33
A probability measure P *is said to be* ***discrete*** *if there is a countable set S such that* $\mathrm{P}(S^c)=0$. *A random variable X is said to be* ***discrete*** *if its distribution is.*

Theorem 1.34
Suppose $X:(\Omega,\mathcal{F})\to(S,\sigma(\mathcal{A}))$ *and* $\mathcal{A}$ *is a collection of subsets in S. If* $X^{-1}(A)\in\mathcal{F}$ *for all* $A\in\mathcal{A}$*, then X is a random variable.*

Proof. Set $\mathcal{G}=\{A\subset S \mid X^{-1}(A)\in\mathcal{F}\}$. Clearly $\varnothing\in\mathcal{G}$ and if $A\in\mathcal{G}$, $X^{-1}(A^c)=(X^{-1}(A))^c\in\mathcal{F}$, so $A^c\in\mathcal{G}$. If $A_n\in\mathcal{G}$, then $X^{-1}(\cup_n A_n)=\cup_n X^{-1}(A_n)\in\mathcal{F}$, so $\cup_n A_n\in\mathcal{G}$. Hence $\mathcal{G}$ is a σ-algebra containing $\mathcal{A}$, so $\sigma(\mathcal{A})\subset\mathcal{G}$. It follows that $X^{-1}(B)\in\mathcal{F}$ for all $B\in\sigma(\mathcal{A})$, so X is a random variable. ■

Corollary 1.35
If X_i *are random variables, then*

$$\inf_i X_i,\quad \sup_i X_i,\quad \liminf_{i\to\infty} X_i,\quad \limsup_{i\to\infty} X_i$$

are all random variables.

Proof. Since the sets of the form $(-\infty,x]$ generate $\mathcal{B}$, it suffices to check that the inverse images of these sets are in $\mathcal{F}$. For $\inf_i X_i$,

$$\left\{\inf_i X_i\le x\right\}=\cup_i\{X_i\le x\}\in\mathcal{F}.$$

For $\sup_i X_i$, since $\sup_i X_i=-\inf_i(-X_i)$, it is also a random variable. Finally, write

$$\liminf_i X_i=\sup_n\inf_{i\ge n} X_i,\quad \limsup_i X_i=\inf_n\sup_{i\ge n} X_i.$$

The results follow from the measurability of $\inf_i X_i$ and $\sup_i X_i$. ■

Definition 1.36
Let X be a random variable. $\sigma(X)$ *is the smallest* σ*-algebra such that X is measurable.*

Remark
If $X:\Omega\to(S,\mathcal{S})$*, then* $\sigma(X)=X^{-1}(\mathcal{S})$.

Definition 1.37
*Let X be a random variable. The **expectation** of X is defined as*

$$\mathrm{E}[X] = \int X d\,\mathrm{P}.$$

Theorem 1.38 (Jensen's Inequality)
Let $X : \Omega \to \mathbb{R}^d$ be a random variable such that $\mathrm{E}[\|X\|_1] < \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then

$$\phi(\mathrm{E}[X]) \leq \mathrm{E}[\phi(X)].$$

Proof. For any given $y \in \mathbb{R}^d$, note that $\left\{x \in \mathbb{R}^d \,\middle|\, \phi(x) > \phi(y)\right\}$ is a open convex set. By the Hahn-Banach seperation theorem, there is a hyperplane $\{f(x) = a + \langle b, x\rangle\}$ separating $\left\{(x, \phi(x)) \in \mathbb{R}^{d+1} \,\middle|\, \phi(x) > \phi(y)\right\}$ and $\{(y, \phi(y))\}$. Note that $\phi(y) = f(y)$ and $\phi(x) \geq f(x)$ for all $x \in \mathbb{R}^d$. Take $y = \mathrm{E}[X]$, then

$$\phi(\mathrm{E}[X]) = f(\mathrm{E}[X]) = \mathrm{E}[f(X)] \leq \mathrm{E}[\phi(X)].$$

∎

Theorem 1.39 (Hölder's Inequality)
Let X, Y be random variables and $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathrm{E}[|XY|] \leq \mathrm{E}[|X|^p]^{1/p}\,\mathrm{E}[|Y|^q]^{1/q}.$$

Proof. If $\mathrm{E}[|X|^p]$ and $\mathrm{E}[|Y^q|]$ are zero or infinite, the result is trivial. We assume that $\mathrm{E}[|X|^p] = \mathrm{E}[|Y|^q] = 1$. For fixed $y \geq 0$, set $\phi(x) = x^p/p + y^p/p - xy$ for $x \geq 0$.

$$\phi'(x) = x^{p-1} - y, \quad \phi''(x) = (p-1)x^{p-2} \geq 0.$$

Thus ϕ is convex and minimized at $x = y^{1/(p-1)}$ with minimum $\phi(y^{1/(p-1)}) = 0$. Hence $x^p/p + y^q/q \geq xy$ for all $x, y \geq 0$.

$$\mathrm{E}[|XY|] \leq \mathrm{E}\left[\frac{|X|^p}{p} + \frac{|Y|^q}{q}\right] = \frac{1}{p} + \frac{1}{q} = 1 = \mathrm{E}[|X|^p]^{1/p}\,\mathrm{E}[|Y|^q]^{1/q}.$$

∎

Theorem 1.40 (Markov's Inequality)
If $X \geq 0$ is a random variable, then for any $c > 0$,

$$\mathrm{P}(X \geq c) \leq \frac{1}{c}\,\mathrm{E}[X].$$

Proof.

$$\mathrm{P}(X \geq c) = \int \mathbf{1}\{X \geq c\}\, d\,\mathrm{P} \leq \int \frac{X}{c} d\,\mathrm{P} = \frac{1}{c}\,\mathrm{E}[X].$$

$\blacksquare$

Example

Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a non-negative function. Put

$$I_A = \inf_{y \in A} \phi(y),$$

where A is some measurable set. Then for any random variable X,

$$I_A \mathbf{1}\{X \in A\} \leq \phi(x)\mathbf{1}\{X \in A\} \leq \phi(x).$$

Thus

$$I_A \mathrm{P}(X \in A) \leq \mathrm{E}[\phi(X)].$$

Corollary 1.41 (Chebyshev's Inequality)
Let X be a random variable. Then for any $c > 0$ and $\alpha \in \mathbb{R}$,

$$\mathrm{P}(|X - \alpha| \geq c) \leq \frac{1}{c^2} \mathrm{E}\left[(X - \alpha)^2\right].$$

Proof. By the Markov's inequality,

$$\mathrm{P}(|X - \alpha| \geq c) = \mathrm{P}((X - \alpha)^2 \geq c^2) \leq \frac{1}{c^2} \mathrm{E}\left[(X - \alpha)^2\right].$$

$\blacksquare$

Theorem 1.42
Suppose X is a random variable of $(S, \mathcal{S})$ with distribution μ and $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$ is measurable. If either

(a) $f \geq 0$, *or*

(b) $\mathrm{E}[|f(X)|] < \infty$,

then

$$\mathrm{E}[f(X)] = \int f(x) d\mu(x).$$

Proof. Suppose first that $f = \mathbf{1}_A$ for some $A \in \mathcal{S}$. Then

$$\mathrm{E}[f(X)] = \mathrm{P}(X \in \mathcal{A}) = \mathrm{P}(X^{-1}(A)) = \mu(A) = \int \mathbf{1}_A d\mu.$$

By linearity we can extend this result to simple functions. Now suppose first that (a) holds. For such f, there is a sequence of simple functions $s_n \nearrow f$ and $s_n \circ X \nearrow f \circ X$. By LMCT,

$$\mathrm{E}[f(X)] = \mathrm{E}\left[\lim_n s_n(X)\right] = \lim_n \mathrm{E}[s_n(X)] = \lim_n \int s_n d\mu = \int f d\mu.$$

Suppose that (b) is the case. Write $f = f^+ - f^-$ and apply the previous result.

$$\mathrm{E}\left[f(X)\right] = \mathrm{E}\left[f^+(X)\right] - \mathrm{E}\left[f^-(X)\right] = \int f^+ d\mu - \int f^- d\mu = \int f d\mu.$$

■

Definition 1.43
*The k-**th moment** of a random variable X is* $\mathrm{E}\left[X^k\right]$.

Definition 1.44
*The **variance** of a random variable X is* $\mathrm{Var}\,\mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right]$.

Definition 1.45
*The **covariance** of two integrable random variables X, Y is*

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])(Y - \mathrm{E}\left[Y\right])\right].$$

Definition 1.46
For $1 \leq p < \infty$, the $\mathcal{L}^p(\Omega, \mathrm{P})$ space is defined as

$$\mathcal{L}^p(\Omega, \mathrm{P}) = \{X : \Omega \to S \mid X \textit{ measurable and } \mathrm{E}\left[|X|^p\right] < \infty\}.$$

For $p = \infty$,

$$\mathcal{L}^\infty(\Omega, \mathrm{P}) = \{X : \Omega \to S \mid X \textit{ measurable and } \operatorname{ess\,sup}_{\omega \in \Omega} X(\omega) < \infty\}.$$

Proposition 1.47
Let $1 \leq p < q \leq \infty$. Then $\mathcal{L}^q(\mathrm{P}) \subset \mathcal{L}^p(\mathrm{P})$.

Proof. Suppose first that $q < \infty$. If $X \in \mathcal{L}^q(\mathrm{P})$, then

$$\mathrm{E}\left[|X|^p\right] \leq \mathrm{E}\left[|X|^q \mathbf{1}\{|X| \geq 1\}\right] + \mathrm{E}\left[|X|^p \mathbf{1}\{|X| < 1\}\right] \leq \mathrm{E}\left[|X|^q\right] + 1 < \infty.$$

Hence $X \in \mathcal{L}^p(\mathrm{P})$. If $q = \infty$, X is essentially bounded, i.e., $X \leq M$ for some $M \in \mathbb{R}$ almost surely. Hence $X \in \mathcal{L}^p$. ■

1.3. Independence

Definition 1.48

*Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Suppose $\mathcal{F}_\beta \subset \mathcal{F}$, $\beta \in B$ are a collection of sub-σ-algebras. Then $\left\{\mathcal{F}_\beta\right\}$ are **independent** if for all finite $\{\mathcal{F}_i\}_{i=1}^n \subset \left\{\mathcal{F}_\beta\right\}$,*

$$\mathrm{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathrm{P}(A_i)$$

where $A_i \in \mathcal{F}_i$.

Definition 1.49

*A collection of random variables $\left\{X_\beta \,\middle|\, \beta \in B\right\}$ on $(\Omega, \mathcal{F}, \mathrm{P})$ is **independent** if the collection of the generating σ-algebras $\left\{\sigma(X_\beta) \,\middle|\, \beta \in B\right\}$ is.*

Remark

In other words,

$$\mathrm{P}(\cap_i \left\{X_{\beta_i} \in A_i\right\}) = \prod_i \mathrm{P}(X_{\beta_i} \in A).$$

Note that these random variables can map into different measurable space.

Definition 1.50

*A collection of events $\mathcal{S}$ is **independent** if $\{\mathbf{1}_A \mid A \in \mathcal{S}\}$ is.*

Proposition 1.51

Let $X_1, \ldots, X_n$ be independent random variables and $g_1, \ldots g_n$ are measurable functions. Then $g_1(X_1), \ldots, g_n(X_n)$ are independent.

Proof. Suppose $g_i : (S_i, \mathcal{S}_i) \to (T_i, \mathcal{T}_i)$. For $A_i \in \mathcal{T}_i$, $g^{-1}(A_i) \in \mathcal{S}_i$ and

$$\mathrm{P}(\cap_i \{g_i(X_i) \in A_i\}) = \mathrm{P}(\cap_i \left\{X_i \in g^{-1}(A_i)\right\}) = \prod_i \mathrm{P}(X_i \in g^{-1}(A_i)) = \prod_i \mathrm{P}(g_i(X_i) \in A_i).$$

$g_1(X_1), \ldots, g_n(X_n)$ are independent. ∎

Theorem 1.52

Let $\mathcal{S}_1, \ldots \mathcal{S}_n$ be a collection of π-system. If $\Omega \in \mathcal{S}_i$ for all $i = 1, \ldots, n$ and for all $A_i \in \mathcal{S}_i$,

$$\mathrm{P}(\cap_i A_i) = \prod_i \mathrm{P}(A_i),$$

then $\sigma(\mathcal{S}_1), \ldots, \sigma(\mathcal{S}_n)$ are independent.

Proof. Fix $\mathcal{S}_2, \ldots, \mathcal{S}_n$. Put

$$\mathcal{L} = \left\{A \in \mathcal{F} \,\middle|\, \mathrm{P}(A \cap (\cap_{i=2}^n A_i)) = \mathrm{P}(A) \prod_{i=2}^n \mathrm{P}(A_i), A_i \in \mathcal{S}_i \text{ for } i = 2, \ldots, n\right\}.$$

We claim that $\mathcal{L}$ forms a λ-system. First, by assumption we can pick $A_i = \Omega$ for $i = 2, \ldots, n$ to see that $\Omega \in \mathcal{L}$. Suppose that $A \subset B$, $A, B \in \mathcal{L}$,

$$\begin{aligned} P((B - A) \cap (\cap_{i=2}^n A_i)) &= P((B \cap (\cap_{i=2}^n A_i)) - (A \cap (\cap_{i=2}^n A_i))) \\ &= P(B) \prod_{i=2}^n P(A_i) - P(A) \prod_{i=2}^n P(A_i) = P(B - A) \prod_{i=2}^n P(A_i). \end{aligned}$$

Hence $B - A \in \mathcal{L}$. Let $S_j \nearrow S$, $S_j \in \mathcal{L}$. Then

$$P(S \cap (\cap_{i=2}^n A_i)) = \lim_{j \to \infty} P(S_j \cap (\cap_{i=2}^n A_i)) = \lim_{j \to \infty} P(S_j) \prod_{i=2}^n P(A_i) = P(S) \prod_{i=2}^n P(A_i).$$

Thus $S \in \mathcal{L}$ and $\mathcal{L}$ is a λ-system. By Dynkin's π-λ, $\sigma(\mathcal{S}_1), \mathcal{S}_2, \ldots, \mathcal{S}_n$ satisfies the product property. Repeat the procedure for $\mathcal{S}_2, \ldots, \mathcal{S}_n$. We have that $\sigma(\mathcal{S}_1), \ldots, \sigma(\mathcal{S}_n)$ satisfies the product property. That is, they are independent. ■

Corollary 1.53
Let $X_1, \ldots, X_n$ be $\mathbb{R}$-valued random variables. Then they are independent if and only if

$$P(X_1 \leq s_1, \ldots, X_n \leq s_n) = \prod_{i=1}^n P(X_i \leq s_i)$$

for all $s_i \in \mathbb{R}$, $1 \leq i \leq n$.

Proof. The sufficient part is trivial. For the converse, put $\mathcal{S}_i = \{\{X_i \leq t\} \mid t \in \mathbb{R}\} \cup \{\Omega\}$. Clearly $\mathcal{S}_i$ are π-system and $\Omega \in \mathcal{S}_i$ for all i. $\sigma(\mathcal{S}_i)$ are independent and $\mathcal{S}_i$ generates $\sigma(X_i)$. Applying theorem 1.52 shows that X_i are independent. ■

Corollary 1.54
If $\mathcal{F}_{ij}$, $1 \leq i \leq n, 1 \leq j \leq m(i)$ are independent σ-algebras, then $\mathcal{G}_i = \sigma(\cup_j \mathcal{F}_{ij})$ are independent.

Proof. Put $\mathcal{H}_i = \left\{\cap_j A_j \,\middle|\, A_j \in \mathcal{F}_{ij}\right\}$. We claim that $\sigma(\mathcal{H}_i) = \mathcal{G}_i$. Indeed, by choosing sets of the form

$$(\Omega, \ldots, \Omega, A_j, \Omega, \ldots, \Omega) \in \mathcal{F}_{i1} \times \cdots \times \mathcal{F}_{im(i)},$$

it is clear that $\cup_j \mathcal{F}_{ij} \subset \mathcal{H}_i$. Also, if $A \in \mathcal{H}_i$, then

$$A = \cap_j A_j = (\cup_j (A_j^c))^c \in \sigma(\cup_j \mathcal{F}_{ij}).$$

Thus $\cup_j \mathcal{F}_{ij} \subset \mathcal{H}_i \subset \sigma(\cup_j \mathcal{F}_{ij})$ and $\sigma(\mathcal{H}_i) = \sigma(\cup_j \mathcal{F}_{ij}) = \mathcal{G}_i$. Also notice that $\mathcal{H}_i$ contain Ω and form π-systems. For $A_i \in \mathcal{H}_i$, write $A_i = \cap_j A_{ij}$. Then

$$P(\cap_i A_i) = P(\cap_{ij} A_{ij}) = \prod_{ij} P(A_{ij}) = \prod_i P(\cap_j A_{ij}) = \prod_i P(A_i).$$

From theorem 1.52 we know that $\mathcal{G}_i = \sigma(\mathcal{H}_i)$ are independent. ■

Corollary 1.55
If $X_{ij}, 1 \leq i \leq n, 1 \leq j \leq m(i)$ are independent random variables, then $Y_i = h_i(X_{i1}, \dots, X_{im(i)})$ are independent provided that h_i are measurable.

Proof. Write $\mathcal{F}_{ij} = \sigma(X_{ij})$. We claim that $\sigma(Y_i) \subset \sigma(\cup_j \mathcal{F}_{ij})$. Indeed, if B_i is a measurable set, $h_i^{-1}(B_i)$ is measurable. Write $h_i^{-1}(B_i) = C_{i1} \times \cdots \times C_{im(i)}$ and since each $X_{ij}^{-1}(C_{ij}) \in \mathcal{F}_{ij}$, we see that $\sigma(Y_i) \subset \sigma(\cup_j \mathcal{F}_{ij})$. It then follows from corollary 1.54 that $\sigma(Y_i)$ are independent and Y_i are independent. ■

Theorem 1.56
If $X_1, \dots X_n$ are independent $\mathbb{R}$-valued random variables and the distribution of X_i is μ_i. Then the joint distribution of $(X_1, \dots, X_n)$ is $\mu_1 \times \cdots \times \mu_n$.

Proof. Let μ be the distribution of $(X_1, \dots, X_n)$. By definition,

$$\begin{aligned}\mu((X_1, \dots) \in A_1 \times \cdots \times A_n) &= \mu(X_1 \in A_1, \dots, X_n \in A_n) \\ &= \prod_{i=1}^n \mu_i(X_i \in A_i) = (\mu_1 \times \cdots \times \mu_n)(A_1 \times \cdots \times A_n).\end{aligned}$$

Now the sets of the forms $A = A_1 \times \cdots \times A_n$ is a π-system generating the product σ-algebra. By corollary 1.19, the joint distribution is exactly $\mu_1 \times \cdots \times \mu_n$. ■

Theorem 1.57
Let X, Y be two independent random variables. If $h(x, y)$ satisfies either

(a) $\mathrm{E}\left[|h(X,Y)|\right] < \infty$*, or*

(b) *h is non-negative,*

then

$$\mathrm{E}\left[h(X,Y)\right] = \int \int h d\mu_X d\mu_Y,$$

where μ_X, μ_Y are the distributions of X and Y, respectively.

Proof. The proof follows directly from Fubini-Tonelli theorem. If one of the assumptions is true, then

$$\mathrm{E}\left[h(X,Y)\right] = \int_{\mathbb{R}^2} h d(\mu_X \times \mu_Y) = \int \int h d\mu_X d\mu_Y.$$

■

Remark
If $h(x, y) = h_1(x)h_2(y)$, then

$$\mathrm{E}\left[h_1(X)h_2(Y)\right] = \mathrm{E}\left[h(X,Y)\right] = \int \int h_1 h_2 d\mu_X d\mu_Y = \mathrm{E}\left[h_1(X)\right]\mathrm{E}\left[h_2(Y)\right].$$

Corollary 1.58
If $X_1, \ldots X_n$ are independent random variables and

(a) $\mathrm{E}\left[|X_1 \cdots X_n|\right] < \infty$ *or*

(b) $X_i \geq 0$ *for all i,*

then

$$\mathrm{E}\left[X_1 \cdots X_n\right] = \prod_{i=1}^{n} \mathrm{E}\left[X_i\right].$$

Proof. Let $h(x, y) = xy$. By assumptions, we have either $\mathrm{E}\left[|h(X_1, X_2)|\right] < \infty$ or $h(X_1, X_2) \geq 0$. By theorem 1.57, $\mathrm{E}\left[X_1 X_2\right] = \mathrm{E}\left[X_1\right] \mathrm{E}\left[X_2\right]$. Substitute X_1 by $X_1 X_2$ and X_2 by X_3, we see that $\mathrm{E}\left[X_1 X_2 X_3\right] = \mathrm{E}\left[X_1\right] \mathrm{E}\left[X_2\right] \mathrm{E}\left[X_3\right]$. Repeat the procedure n times and the result follows. ∎

Definition 1.59
Let X, Y be independent random variables with CDF F and G, respectively. The ***convolution*** *of two CDF is defined as*

$$(F * G)(z) = \int F(z - y) dG(y).$$

Remark
If F and G are absolutely continuous with respect to the Lebesgue measure, then they have Radon-Nikodym derivatives f and g. The definition of convolution becomes

$$(F * G)(z) = \int F(z - y) dG(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x) g(y) dx dy.$$

Then

$$(F * G)'(z) = \int f(z - y) g(y) dy = (f * g)(z),$$

which is exactly the definition of convolution of two functions.

Proposition 1.60
Let X and Y be independent random variables. Then

$$\mathrm{P}(X + Y \leq z) = (F * G)(z).$$

Proof. By theorem 1.57,

$$\begin{aligned}\mathrm{P}(X + Y \leq z) = \mathrm{E}\left[\mathbf{1}\left\{X + Y \leq z\right\}\right] &= \int \int \mathbf{1}\left\{x + y \leq z\right\} dF(x) dG(y) \\ &= \int F(z - y) dG(y) = (F * G)(z).\end{aligned}$$

∎

Remark
Note that the convolution is commutative since

$$(F * G)(z) = \mathrm{P}(X + Y \leq z) = \mathrm{P}(Y + X \leq z) = (G * F)(z).$$

Remark
For discrete X and Y, the convolution becomes

$$\mathrm{P}(X+Y=z)=\sum_y \mathrm{P}(X=z-y)\,\mathrm{P}(Y=y).$$

Example
Consider $X\sim\Gamma(\alpha_1,\beta)$ and $Y\sim\Gamma(\alpha_2,\beta)$. Then the density for $X+Y$ is

$$\begin{aligned}
f_{X+Y}(z) &= \int f_X(z-y)f_Y(y)dy \\
&= \int_0^z \frac{1}{\Gamma(\alpha_1)}\beta^{\alpha_1}(z-y)^{\alpha_1-1}e^{-\beta(z-y)}\frac{1}{\Gamma(\alpha_2)}\beta^{\alpha_2}y^{\alpha_2-1}e^{-\beta y}dy \\
&= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\beta^{\alpha_1+\alpha_2}e^{-\beta z}\int_0^z (z-y)^{\alpha_1-1}y^{\alpha_2-1}dy \\
&= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\beta^{\alpha_1+\alpha_2}e^{-\beta z}z^{\alpha_1+\alpha_2-1}\int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}dt \\
&= \frac{B(\alpha_1,\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\beta^{\alpha_1+\alpha_2}e^{-\beta z}z^{\alpha_1+\alpha_2-1} = \frac{1}{\Gamma(\alpha_1+\alpha_2)}\beta^{\alpha_1+\alpha_2}e^{-\beta z}z^{\alpha_1+\alpha_2-1}.
\end{aligned}$$

Hence $X+Y\sim\Gamma(\alpha_1+\alpha_2,\beta)$.

1.4. Convergence of Random Variables

Definition 1.61
A sequence of probability measures P_n *on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *are* ***consistent*** *if*

$$\mathrm{P}_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R} = \mathrm{P}_n((a_1, b_1] \times \cdots \times (a_n, b_n])$$

for every n.

Theorem 1.62 (Kolmogorov Extension)
Suppose that a sequence of probability measures P_n *on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *are consistent. Then there is a unique probability measure* P *on* $(\mathbb{R}^{\mathbb{N}}, \mathcal{B})$ *satisfying that*

$$\mathrm{P}(\left\{\omega \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \omega_i \in (a_i, b_i], 1 \le i \le n\right\}) = \mathrm{P}_n((a_1, b_1] \times \cdots \times (a_n, b_n]),$$

where $\mathcal{B}$ *is generated by the collection*

$$\left\{\omega \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \omega_i \in (a_i, b_i], 1 \le i \le n, n \in \mathbb{N}\right\}.$$

Proof. Let

$$\mathcal{S} = \{(a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R} \times \cdots \mid n \in \mathbb{N}\}.$$

Define P on $\mathcal{S}$ to be

$$\mathrm{P}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R} \times \cdots) = \mathrm{P}_n((a_1, b_1] \times \cdots \times (a_n, b_n])$$

Clearly, $\mathcal{S}$ forms a semi-algebra. From the Carathéodory extension theorem, it suffices to show that P is finitely additive, σ-additive on $\mathcal{S}$ and $\mathrm{P}(\varnothing) = 0$. Note that $\mathrm{P}(\varnothing) = \mathrm{P}(\varnothing \times \mathbb{R} \times \cdots) = \mathrm{P}_1(\varnothing) = 0$. We verify the first two conditions.

First, if $A, B \in \mathcal{S}$ are disjoint, $m \le n$,

$$A = \left\{\omega \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \omega_i \in (a_i, b_i], 1 \le i \le m\right\} \quad \text{and} \quad B = \left\{\omega \in \mathbb{R}^{\mathbb{N}} \;\middle|\; \omega_i \in (c_i, d_i], 1 \le i \le n\right\},$$

then

$$\mathrm{P}(A \cup B) = \mathrm{P}_n((\pi_n A) \cup (\pi_n B)) = \mathrm{P}_n(\pi_n A) + \mathrm{P}_n(\pi_n B) = \mathrm{P}(A) + \mathrm{P}(B),$$

where $\pi_n : \omega \to (\omega_1, \ldots, \omega_n)$ is the projection onto the first n components. Hence P is finitely additive.

Next, suppose $A_1, \ldots \in \mathcal{S}$ are countably many disjoint measurable sets. Put $A = \cup_i A_i$. We can consider the algebra $\bar{\mathcal{S}} = \{\text{finite disjoint union of sets in } \mathcal{S}\}$ generated by $\mathcal{S}$. $B_n = \cup_{i>n} A_i \in \bar{\mathcal{S}}$. Thus

$$\mathrm{P}(A) = \mathrm{P}(B_n) + \sum_{i=1}^{n} \mathrm{P}(A_n)$$

by the previous result. It now suffices to show that $\mathrm{P}(B_n) \to 0$ for any $B_n \searrow \varnothing$. Suppose not, then there is $\delta > 0$ such that $\mathrm{P}(B_n) \to \delta$ as $B_n \to \varnothing$ by the monotonicity of P.

For such $\{B_n\}$, we claim that there is a sequence of compact set K_n such that $K_n \subset B_n$ and $\mathrm{P}(B_n - K_n) < 2^{-(n+1)}\delta$. Now since $B_1 \in \bar{S}$, there are disjoint $E_1^1, \ldots, E_{m_1}^1$ such that $B_1 = \cup_{i=1}^{m_1} E_i^1$. Now since each E_i^1 is of the product of $(\cdot, \cdot]$. We can find a compact subset K_i^1 of the product of $[\cdot, \cdot]$ such that $\mathrm{P}(E_i^1 - K_i^1) < m_1^{-1} 2^{-2}\delta$. Hence $K_1 = \cup_i K_i^1 \subset B_1$ satisfies that

$$\mathrm{P}(B_1 - K_1) = \sum_{i=1}^{m_1} \mathrm{P}(E_i^1 - K_i^1) < 2^{-2}\delta$$

as desired. Repeat the process and find K_n inductively. The claim follows.

Now, $\cap_{n=1}^m K_n \searrow K$ as $m \to \infty$. Also,

$$\mathrm{P}(B_m - (\cap_{n=1}^m K_n)) \le \sum_{n=1}^{m} \mathrm{P}(B_n - K_n) \le \frac{\delta}{2}.$$

Hence $\delta/2 \le \mathrm{P}(B_m) - \delta/2 \le \mathrm{P}(\cap_{n=1}^m K_n)$. We see that $\cap_{n=1}^m K_n$ is non-empty for each m. But this implies that $K \subset \cap_n B_n$ is non-empty, a contradiction. Thus $\mathrm{P}(B_n) \to 0$.

Finally, the σ-additivity follows from that we can take $n \to \infty$ so that

$$\mathrm{P}(A) = \lim_{n\to\infty} \mathrm{P}(B_n) + \sum_{i=1}^{n} \mathrm{P}(A_n) = \sum_i \mathrm{P}(A_n).$$

Applying Carathéodory extension theorem, such P can be extended on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B})$. ∎

Remark

With Kolmogorov extension theorem, we can consider a sequence of independent variable X_i on the product probability space with $\mathcal{F} = \mathcal{B}$, $\tilde{X}_i : \omega \mapsto \omega_i$ and $\mathrm{P}(B_1 \times \cdots B_n) = \prod_{i=1}^n \mu_i(B_i)$, where μ_i is the distribution of X_i.

Definition 1.63

*Let X_n be a sequence of random variable. X_n **converges almost surely** to X if*

$$\mathrm{P}\left\{\lim_{n\to\infty} X_n = X\right\} = 1.$$

We denote it as $X_n \xrightarrow{a.s.} X$ or $X_n \to X$ a.s.

Definition 1.64

*Let X_n be a sequence of random variable. X_n **converges in probability** to X if for every $\epsilon > 0$,*

$$\mathrm{P}\{|X_n - X| > \epsilon\} \to 0$$

as $n \to \infty$. We denote it as $X_n \xrightarrow{p} X$.

Definition 1.65

A sequence of random variable $X_n \in \mathcal{L}^p$ is said to ***converge in*** *$\mathcal{L}^p$ to X if*

$$\mathrm{E}\left[|X_n - X|^p\right]^{1/p} \to 0$$

as $n \to \infty$. If $p = \infty$, the definition becomes

$$\operatorname{ess\,sup}_{\omega \in \Omega} |X_n(\omega) - X(\omega)| \to 0.$$

We denote it as $X_n \to X$ in $\mathcal{L}^p$.

Proposition 1.66

Let X_n be a sequence of independent and indentically distributed random variables. Then

(a) If $X_n \to X$ almost surely, then $X_n \xrightarrow{p} X$.

(b) If $X_n \to X$ in $\mathcal{L}^p$, then $X_n \xrightarrow{p} X$.

Proof. For (a), given $\epsilon > 0$, put

$$E_k = \cup_{n \geq k} \left\{ |X_n - X| > \epsilon \right\}.$$

Note that $E_k \searrow E = \{|X_n - X| > \epsilon$ for infinitely many $n\} = \{\lim_{n\to\infty} X_n = X\}^c$. Hence

$$\mathrm{P}\left\{|X_k - X| > \epsilon\right\} \leq \mathrm{P}(E_k) \to \mathrm{P}\left\{\lim_{n\to\infty} X_n = X\right\}^c = 0$$

Hence $X_n \to X$ in probability.

For (b), suppose first that $p < \infty$. By Markov inequality,

$$\mathrm{P}\left\{|X_n - X| > \epsilon\right\} = \mathrm{P}\left\{|X_n - X|^p > \epsilon^p\right\} \leq \frac{1}{\epsilon^p} \mathrm{E}\left[|X_n - X|^p\right] \to 0.$$

Let $p = \infty$. Note that $\operatorname{ess\,sup}|X_n - X| = \inf\{c \mid \mathrm{P}\{|X_n - X| > c\} = 0\}$. Convergence in $\mathcal{L}^\infty$ implies that for $\epsilon > 0$, there is N such that if $n \geq N$, $\inf\{c \mid \mathrm{P}\{|X_n - X| > c\} = 0\} < \epsilon$. That is, $\mathrm{P}\{|X_n - X| > \epsilon\} = 0$ for $n \geq N$. Hence $X_n \xrightarrow{p} X$. ∎

2. Law of Large Number and Central Limit Theorem

2.1. Law of Large Number

Definition 2.1
*Let X_i be random variables with $\mathrm{E}\left[X_i^2\right] < \infty$. They are called **uncorrelated** if*

$$\mathrm{E}\left[X_i X_j\right] = \mathrm{E}\left[X_i\right] \mathrm{E}\left[X_j\right].$$

Theorem 2.2 (Weak Law of Large Number I)
Suppose that X_n are uncorrelated random variables with $\mathrm{Var}\left[X_n\right] \leq C\infty$ and $\mathrm{E}\left[X_n\right] = \mu$ for all n. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{n} S_n \to \mu$$

in $\mathcal{L}^2$ and hence in probability.

Proof. Compute that

$$\mathrm{E}\left[\left(\frac{1}{n} S_n - \mu\right)^2\right] = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}(X_i) \leq \frac{C}{n} \to 0.$$

Hence $\frac{1}{n} S_n \to \mu$ in $\mathcal{L}^2$ and thus in probability. ■

Theorem 2.3 (Weak Law of Large Number II, Khinchin)
Suppose that X_i is a sequence of independent and identically distributed random variables with $\mathrm{E}\left[|X_1|\right] < \infty$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = \mathrm{E}\left[X_1\right]$. Then

$$\frac{1}{n} S_n \to \mu$$

in $\mathcal{L}^1$ and hence in probability.

Proof. By replacing X_i with $X_i - \mu$, we may assume without loss of generality that $\mu = 0$. Now, for $C > 0$,

$$0 = \mathrm{E}\left[X_i\right] = \mathrm{E}\left[X_i \mathbf{1}\left\{|X_i| > C\right\}\right] + \mathrm{E}\left[X_i \mathbf{1}\left\{|X_i| \leq C\right\}\right].$$

Also,

$$\begin{aligned}
\frac{1}{n} S_n &= \frac{1}{n} \sum_{i=1}^n X_i \mathbf{1}\left\{|X_i| > C\right\} + \frac{1}{n} \sum_{i=1}^n X_i \mathbf{1}\left\{|X_i| \leq C\right\} \\
&= \frac{1}{n} \sum_{i=1}^n (X_i \mathbf{1}\left\{|X_i| > C\right\} - \mathrm{E}\left[X_i \mathbf{1}\left\{|X_i| > C\right\}\right]) + \frac{1}{n} \sum_{i=1}^n (X_i \mathbf{1}\left\{|X_i| \leq C\right\} - \mathrm{E}\left[X_i \mathbf{1}\left\{|X_i| \leq C\right\}\right]).
\end{aligned}$$

Notice that by LDCT,

$$\mathrm{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i\mathbf{1}\{|X_i|>C\}-\mathrm{E}\left[X_i\mathbf{1}\{|X_i|>C\}\right])\right|\right] \leq 2\,\mathrm{E}\left[|X_1|\,\mathbf{1}\{|X_1|>C\}\right] \to 0$$

as $C \to \infty$ since $|X_1|\,\mathbf{1}\{|X_1|>C\} \leq |X_1|$ and $\mathrm{E}\left[|X_1|\right] < \infty$. Also, by Hölder inequality and the independence,

$$\mathrm{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i\mathbf{1}\{|X_i|\leq C\}-\mathrm{E}\left[X_i\mathbf{1}\{|X_i|\leq C\}\right])\right|\right] \leq \sqrt{\frac{1}{n}\mathrm{Var}(X_i\mathbf{1}\{|X_i|\leq C\})} \leq \frac{C}{\sqrt{n}}$$

For any given $\epsilon > 0$, there is C such that $2\,\mathrm{E}\left[|X_1|\,\mathbf{1}\{|X_1|>C\}\right] < \epsilon$ and

$$\begin{aligned}
\mathrm{E}\left[\left|\frac{1}{n}S_n\right|\right] &\leq \mathrm{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i\mathbf{1}\{|X_i|>C\}-\mathrm{E}\left[X_i\mathbf{1}\{|X_i|>C\}\right])\right|\right] \\
&\quad + \mathrm{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i\mathbf{1}\{|X_i|\leq C\}-\mathrm{E}\left[X_i\mathbf{1}\{|X_i|\leq C\}\right])\right|\right] \\
&\leq \epsilon + \frac{C}{\sqrt{n}} \to \epsilon
\end{aligned}$$

as $n \to \infty$. Since ϵ can be arbitrarily small, we conclude that $\frac{1}{n}S_n \to 0$ in $\mathcal{L}^1$ and hence in probability. ■

Definition 2.4
Let A_n be a sequence of events.

$$\limsup_{n\to\infty} A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$$

and

$$\liminf_{n\to\infty} A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n.$$

Remark
Observe that

$$\limsup_{n\to\infty} A_n = \{\omega \in \Omega \mid \omega \in A_n \textit{ for infinitely many } n\}$$

and

$$\liminf_{n\to\infty} A_n = \{\omega \in \Omega \mid \omega \in A_n \textit{ for all but finitely many } n\}.$$

Theorem 2.5 (Borel-Cantelli I)
Let A_n be a sequence of events. If $\sum_n \mathrm{P}(A_n) < \infty$, then

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

Proof. Let $\epsilon > 0$ be given. By assumption, there is n_0 such that $\sum_{n\geq n_0} \mathrm{P}(A_n) < \epsilon$. Then

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = \mathrm{P}(\cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n) \leq \mathrm{P}(\cup_{n=n_0}^{\infty} A_n) \leq \sum_{n=n_0}^{\infty} \mathrm{P}(A_n) < \epsilon.$$

Since ϵ can be arbitrarily small, $\mathrm{P}(\limsup_{n\to\infty} A_n) = 0$. ∎

Corollary 2.6
Suppose for $\epsilon > 0$, $\sum_n \mathrm{P}(|X_n - X| > \epsilon) < \infty$. Then $X_n \to X$ almost surely.

Proof. Let $E_k = \left\{|X_n - X| > k^{-1} \text{ for finitely many } n\right\}$. Note that $E_{k+1} \subset E_k$ and $E_k \searrow E = \{X_n \to X\}$. Now we claim that $\mathrm{P}(E_k) = 1$. Consider $E_k^n = \left\{|X_n - X| > k^{-1}\right\}$. For fixed k, by assumption we have $\sum_n \mathrm{P}(E_k^n) < \infty$. By Borel-Cantelli, $\mathrm{P}(\limsup_{n\to\infty} E_k^n) = 0$. Hence

$$\mathrm{P}(E_k) = \mathrm{P}(\left\{|X_n - X| > k^{-1} \text{ for infinitely many } n\right\}^c) = 1 - \mathrm{P}(\limsup_{n\to\infty} E_k^n) = 1.$$

It now follows by the monotone convergence of measures that $\mathrm{P}(E) = 1$. ∎

Remark
Intuitively, if the convergence is sufficiently fast, the convergence in probability may recover almost sure convergence.

Theorem 2.7 (Strong Law of Large Number I)
Let X_i be independent and identically distributed with $\mu = \mathrm{E}[X_1]$ and $\mathrm{E}\left[X_1^4\right] < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{n} S_n \to \mu$$

almost surely.

Proof. Note that

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{1}{n}S_n - \mu\right)^4\right] &= \frac{1}{n^4}\left(\sum_i \mathrm{E}\left[(X_i - \mu)^4\right] + \sum_{i\neq j} \mathrm{E}\left[(X_i-\mu)^2(X_j-\mu)^2\right]\right) \\
&\leq \frac{1}{n^3}\mathrm{E}\left[(X_1-\mu)^4\right] + \frac{1}{n^4}\binom{n}{2}\binom{4}{2}\mathrm{E}\left[(X_1-\mu)^2\right]^2 \leq \frac{C}{n^2}
\end{aligned}$$

for some constant C. By Checyshev's inequality, for $\epsilon > 0$,

$$\mathrm{P}\left\{\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right\} \leq \frac{1}{\epsilon^4}\mathrm{E}\left[\left(\frac{1}{n}S_n - \mu\right)^4\right] \leq \frac{C}{\epsilon^2 n^2}$$

is absolute summable. Hence by corollary 2.6,

$$\frac{1}{n}S_n \to \mu$$

almost surely. ∎

Theorem 2.8
$X_n \xrightarrow{p} X$ if and only if every subsequence of X_n has a further subsequence converging almost surely.

Proof. Suppose first that $X_n \xrightarrow{p} X$. Given a subsequence $X_{n(k)}$, we can choose $n(k_1) < n(k_2) < \cdots$ such that

$$\mathrm{P}(|X_{n(k_i)} - X| > 2^{-i}) < 2^{-i}.$$

Since 2^{-i} is summable, by Borel-Cantelli we have

$$\mathrm{P}(|X_{n(k_i)} - X| > 2^{-i} \text{ for infinitely many } i) = 0.$$

In other words,

$$\mathrm{P}\left\{X_{n(k_i)} \to X\right\} = \mathrm{P}\left\{|X_{n(k_i)} - X| > 2^{-i} \text{ for infinitely many } i\right\}^c = 1.$$

For the converse, suppose that $X_n \not\to X$ in probability. Then there exist $\epsilon, \delta > 0$ and

$$\mathrm{P}\left\{|X_{n(k)} - X| > \epsilon\right\} \geq \delta.$$

By assumption there is a further subsequence converging almost surely and thus in probability, i.e.,

$$\mathrm{P}\left\{|X_{n(k_j)} - X| > \epsilon\right\} \to 0.$$

This is a contradiction. Hence $X_n \to X$ in probability. ■

Corollary 2.9
Suppose $X_n \xrightarrow{p} X$. Then the followings are true:

(a) If f is continuous, then $f(X_n) \xrightarrow{p} f(X)$.

(b) If $|X_n| \leq Y$ for some $Y \in \mathcal{L}^1$, then $\mathrm{E}[X_n] \to \mathrm{E}[X]$.

Proof. For (a), by theorem 2.8, every subsequence has a further subsequence $X_{n(k_j)} \to X$ almost surely and hence $f(X_{n(k_j)}) \to f(X)$ almost surely. Then by theorem 2.8 again we see that $f(X_n) \xrightarrow{p} f(X)$.

For (b), by theorem 2.8, every subsequence has a furhter subsequence $X_{n(k_j)} \to X$ almost surely and LDCT gives $\mathrm{E}\left[X_{n(k_j)}\right] \to \mathrm{E}[X]$. This implies that $\mathrm{E}[X_n] \to \mathrm{E}[X]$ as well. ■

Definition 2.10
Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space.

$$\mathcal{L}^0(\Omega) = \{X : \Omega \to \mathbb{R} \mid X \textit{ is } \mathcal{F}\textit{-measurable}\}.$$

Remark
In general, the almost convergence notion on $\mathcal{L}^0$ is not metrizable, i.e., there is no metric d on

$\mathcal{L}^0$ such that

$$d(X_n, X) \to 0 \quad \Leftrightarrow \quad X_n \to X \quad a.s.$$

To see this, suppose that the almost sure convergence is metrizable. If $X_n \xrightarrow{p} X$, any subsequence $X_{n(k)}$ converges to X in probability as well. By theorem 2.8, we can find a further subsequence converging almost surely and hence in metric d, but this implies that $d(X_n, X) \to 0$. Then $X_n \to X$ almost surely, which is absurd since convergence in probability does not imply almost sure convergence in general.

However, convergence in probability on $\mathcal{L}^0$ can be metrized. For instance,

$$d(X, Y) = \mathrm{E}\left[\max\left\{|X - Y|, 1\right\}\right].$$

Theorem 2.11 (Borel-Cantelli II)
Let A_n be independent events and $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$. Then

$$\mathrm{P}\left(\limsup_{n\to\infty} A_n\right) = 1.$$

Proof. By assumption we have that $\sum_{n\geq m} \mathrm{P}(A_n) = \infty$ for every $m \in \mathbb{N}$. Notice that $1 + x \leq e^x$. Then

$$\begin{aligned}
\mathrm{P}(\limsup_{n\to\infty} A_n) &= \lim_{m\to\infty} \mathrm{P}(\cup_{n\geq m} A_n) = 1 - \lim_{m\to\infty} \mathrm{P}(\cap_{n\geq m} A_n^c) \\
&= 1 - \lim_{m\to\infty} \lim_{N\to\infty} \mathrm{P}(\cap_{n=m}^N A_n^c) = 1 - \lim_{m\to\infty} \lim_{N\to\infty} \prod_{n=m}^N \mathrm{P}(A_n^c) \\
&= 1 - \lim_{m\to\infty} \prod_{n=m}^\infty (1 - \mathrm{P}(A_n)) \geq 1 - \lim_{m\to\infty} \exp\left(-\sum_{n=m}^\infty \mathrm{P}(A_n)\right) = 1.
\end{aligned}$$

Hence $\mathrm{P}(\limsup_{n\to\infty} A_n) = 1$. ∎

Lemma 2.12
Let X be a non-negative random variable and $h : \mathbb{R} \to \mathbb{R}$ be a differentiable function with $h(0) = 0$ and $h' \geq 0$. Then

$$\mathrm{E}\left[h(X)\right] = \int_0^\infty h'(t)\, \mathrm{P}(X > t) dt.$$

Proof. By Fubini-Tonelli theorem,

$$\begin{aligned}
\mathrm{E}\left[h(X)\right] &= \mathrm{E}\left[\int_0^X h'(t) dt\right] = \mathrm{E}\left[\int_0^\infty \mathbf{1}\left\{t < X\right\} h'(t) dt\right] \\
&= \int_0^\infty h'(t)\, \mathrm{E}\left[\mathbf{1}\left\{t < X\right\}\right] dt = \int_0^\infty h'(t)\, \mathrm{P}(X > t) dt.
\end{aligned}$$

∎

Proposition 2.13
Suppose that X_i are independent and identically distributed random variables with $\mathrm{E}[|X_i|] = \infty$. *Let* $S_n = \sum_{i=1}^n X_i$. *Then*

(a) $\mathrm{P}\{|X_n| > n \textit{ for infinitely many } n\} = 1$.

(b) $\mathrm{P}\left\{\frac{1}{n}S_n \textit{ has finite limit}\right\} = 0$.

Proof. For (a), using lemma 2.12 with h being identity,

$$\begin{aligned}\infty = \mathrm{E}[|X_1|] = \int_0^\infty \mathrm{P}(|X_1| > t)dt &\le \sum_{n=0}^\infty \int_n^{n+1} \mathrm{P}(|X_1| > t)dt \\ \le \sum_{n=0}^\infty \int_n^{n+1} \mathrm{P}(|X_1| > n)dt &= \sum_{n=0}^\infty \mathrm{P}(|X_n| > n).\end{aligned}$$

Now by the second Borel-Cantelli, $\mathrm{P}\{|X_n| > n \text{ for infinitely many } n\} = 1$.

For (b), consider ω with $\frac{S_n(\omega)}{n} \to Y(\omega) \in \mathbb{R}$. Then for such ω,

$$\frac{X_n}{n} = \frac{S_n - S_{n-1}}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1} \to 0$$

as $n \to \infty$. Thus

$$\begin{aligned}\mathrm{P}\left\{\frac{1}{n}S_n \text{ has finite limit}\right\} &\le \mathrm{P}\{|X_n| > n \text{ for finitely many } n\} \\ &= 1 - \mathrm{P}\{|X_n| > n \text{ for infinitely many } n\} = 0.\end{aligned}$$

(b) follows. ∎

Definition 2.14
A collection of σ-algebra $\{\mathcal{H}_k\}$ is ***pairwise independent*** *if for any $\mathcal{H}_1, \mathcal{H}_2 \in \{\mathcal{H}_k\}$,*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

for any $A \in \mathcal{H}_1$ and $B \in \mathcal{H}_2$.

Remark
As before, a sequence of random variables $\{X_k\}$ is pairwise independent if $\{\sigma(X_k)\}$ is.

Theorem 2.15 (Strong Law of Large Number II, Kolmogorov)
Let X_i be pairwise independent, identically distributed random variables with $\mathrm{E}[|X_1|] < \infty$ *and* $S_n = \sum_{i=1}^n X_i$. *Then*

$$\frac{1}{n}S_n \to \mathrm{E}[X_1] = \mu$$

almost surely.

Proof. Since we can always decompose $X_i = X_i^+ - X_i^-$ and X_i^+, X_i^- satisfy the assumption of the theorem, we may assume without loss of generality that $X_i \ge 0$. Let $Y_i = X_i \mathbf{1}\{X_i \le i\}$ and

$T_n = \sum_{i=1}^n Y_i$. Let $\alpha > 1$ and put $k_n = \lfloor \alpha^n \rfloor$. By Chebyshev inequality, for any given $\epsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\left|\frac{T_{k_n} - \mathrm{E}\left[T_{k_n}\right]}{k_n}\right| > \epsilon\right) \le \frac{1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{k_n^2} \mathrm{Var}(T_{k_n})$$
$$= \frac{1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathrm{Var}(Y_i) = \frac{1}{\epsilon^2} \sum_{i=1}^{\infty} \mathrm{Var}(Y_i) \sum_{n:k_n \ge i} \frac{1}{k_n^2}.$$

Since $1/k^2$ is summable and k_n repeat at most m_α times, where m_α is an integer such that $\alpha^{m_\alpha+1} \ge \alpha^{m_\alpha} + 1$, we can find a constant $c_\alpha > 0$ such that

$$\sum_{n:k_n \ge i} \frac{1}{k_n^2} \le \frac{c_\alpha}{i^2}.$$

Let F be the distribution of X. We have

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\left|\frac{T_{k_n} - \mathrm{E}\left[T_{k_n}\right]}{k_n}\right| > \epsilon\right) \le \frac{c_\alpha}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathrm{Var}(Y_i)}{i^2} \le \frac{c_\alpha}{\epsilon^2} \sum_{i=1}^{\infty} \frac{\mathrm{E}\left[Y_i^2\right]}{i^2}$$
$$= \frac{c_\alpha}{\epsilon^2} \sum_{i=1}^{\infty} \frac{1}{i^2} \int_0^i x^2 dF(x) = \frac{c_\alpha}{\epsilon^2} \sum_{i=1}^{\infty} \sum_{k=0}^{i-1} \frac{1}{i^2} \int_k^{k+1} x^2 dF(x)$$
$$= \frac{c_\alpha}{\epsilon^2} \sum_{k=0}^{\infty} \left(\sum_{i=k+1}^{\infty} \frac{1}{i^2}\right) \int_k^{k+1} x^2 dF(x)$$

Also, notice that there is a constant C such that

$$\sum_{i=k+1}^{\infty} \frac{1}{i^2} \le \frac{C}{k+1}.$$

Hence,

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\left|\frac{T_{k_n} - \mathrm{E}\left[T_{k_n}\right]}{k_n}\right| > \epsilon\right) \le \frac{c_\alpha C}{\epsilon^2} \sum_{k=0}^{\infty} \frac{1}{k+1} \int_k^{k+1} x^2 dF(x)$$
$$\le \frac{c_\alpha C}{\epsilon^2} \sum_{k=0}^{\infty} \int_k^{k+1} x dF(x) = \frac{c_\alpha C}{\epsilon^2} \mathrm{E}\left[X_1\right] < \infty.$$

Note that for $\delta > 0$ there is an integer M such that $\mathrm{E}\left[X_1 \mathbf{1}\left\{X_1 > M\right\}\right] \le \delta \le \mathrm{E}\left[X_1\right]$.

$$\frac{\mathrm{E}\left[T_{k_n}\right]}{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathrm{E}\left[Y_i\right] \ge \frac{1}{k_n} \sum_{i=1}^{M} \mathrm{E}\left[Y_i\right] + \frac{1}{k_n} \sum_{i=M+1}^{k_n} \mathrm{E}\left[X_1\right] - \delta.$$

Also,

$$\frac{1}{k_n}\sum_{i=1}^{k_n}\mathrm{E}\left[Y_i\right] \le \frac{1}{k_n}\sum_{i=1}^{k_n}\mathrm{E}\left[X_1\right] = \mathrm{E}\left[X_1\right].$$

Taking $n \to \infty$ and since δ is arbitrary, we conclude that

$$\frac{\mathrm{E}\left[T_{k_n}\right]}{k_n} \to \mathrm{E}\left[X_1\right].$$

Thus, by the Borel-Cantelli lemma,

$$\mathrm{P}\left\{\frac{T_{k_n}}{k_n} \not\to \mathrm{E}\left[X_1\right]\right\} = \mathrm{P}\left\{\left|\frac{T_{k_n} - \mathrm{E}\left[T_{k_n}\right]}{k_n}\right| > \epsilon \text{ for infinitely many } n\right\} = 0.$$

In other words, $T_{k_n}/k_n \to \mathrm{E}\left[X_1\right]$ almost surely. Also,

$$\begin{aligned}\sum_{k=1}^{\infty}\mathrm{P}\left\{X_k \neq Y_k\right\} &= \sum_{k=1}^{\infty}\mathrm{P}\left\{X_k > k\right\} \le \sum_{k=1}^{\infty}\int_{k-1}^{k}\mathrm{P}(X_1 > t)dt \\ &= \int_0^{\infty}\mathrm{P}(X_1 > t)dt = \mathrm{E}\left[X_1\right] < \infty\end{aligned}$$

by lemma 2.12. Hence by Borel-Cantelli lemma, $X_k \neq Y_k$ for finitely many k almost surely. This implies that

$$\lim_{n\to\infty}\frac{1}{k_n}S_{k_n} = \lim_{n\to\infty}\frac{T_{k_n}}{k_n} = \mathrm{E}\left[X_1\right]$$

almost surely. Note that S_m is monotone and for each m, we may find $k(n_m) \le m \le k(n_{m+1})$ so that

$$\frac{S_{k(n_m)}}{k(n_{m+1})} \le \frac{S_m}{m} \le \frac{S_{k(n_{m+1})}}{k(n_m)}.$$

Take $m \to \infty$, we conclude that

$$\frac{1}{\alpha}\mu \le \liminf_{m\to\infty}\frac{S_m}{m} \le \limsup_{m\to\infty}\frac{S_m}{m} \le \alpha\mu$$

almost surely. Taking $\alpha \to 1^+$ gives the desired result. ■

Theorem 2.16
Let X_i be independent and identically distributed with $\mathrm{E}\left[X_1^+\right] = \infty$ and $\mathrm{E}\left[X_1^-\right] < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{n}S_n \to \infty$$

almost surely.

Proof. Write $X_i = X_i^+ - X_i^-$. For X_i^+, consider $Y_i^M = \min\left\{X_i^+, M\right\}$ for some $M > 0$. Note that Y_i^M is independent and identically distributed with finite mean. By the strong law of large number,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^M \to \mathrm{E}\left[Y_1^M\right]$$

almost surely. Hence,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^+ \geq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i^M = \mathrm{E}\left[Y_1^M\right].$$

Notice that $Y_1^M \nearrow X_1^+$ as $M \to \infty$. By LMCT,

$$\lim_{M\to\infty} \mathrm{E}\left[Y_1^M\right] = \mathrm{E}\left[X_1^+\right] = \infty.$$

We conclude that $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^+ = \infty$. On the other hand, by the strong law of large number,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^- \to \mathrm{E}\left[X_1^-\right]$$

almost surely. We end up with

$$\lim_{n\to\infty} \frac{1}{n} S_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^+ - \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^- = \infty$$

almost surely. ■

Example
Let Y_i be independent and identically distributed with density

$$f(y) = \mathbf{1}\{y \geq 1\} \frac{1}{c}\frac{1}{y^2},$$

where c is some normalizing constant. Let $H_i \sim Ber(2^{-i})$. Put $X_i = Y_i H_i$. Then $\mathrm{E}[X_i] = \infty$ for all i, but since

$$\sum_i \mathrm{P}(X_i > 0) = \sum_i 2^{-i} < \infty,$$

by the Borel-Cantelli lemma, $X_i \to 0$ almost surely and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to 0$$

almost surely.

Example
$Y \geq 0$ is a random variable with $\mathrm{E}[Y] = \infty$. Put $X_i = Y$ for all i. Then X_i are identically

distributed with $\mathrm{E}[X_i] = \infty$. *But*

$$\frac{1}{n}\sum_{i=1}^{n} X_i = Y \not\to \infty$$

almost surely.

Example (Event Streaks)
$X_i \overset{iid}{\sim} Ber(2^{-1})$. *Let* L_n *be the longest streaks of* 1 *in the first* n *trials. We have the following:*

$$\lim_{n\to\infty} \frac{L_n}{\log_2(n)} = 1.$$

To see this, let ℓ_n *be the length of the current streaks. For instance, the following sequence*

$$1, 0, 1, 1, 1, 1, 0, \ldots$$

generates $\ell_1 = 1, \ell_2 = 0, \ell_6 = 4$. *Observe that* $L_n = \max_{m\le n} \ell_m$. *Now,*

$$\mathrm{P}(\ell_n \ge k) = \sum_{m=k}^{n} \mathrm{P}(\ell_n = k) = \sum_{m=k}^{n} 2^{-k-1} \le 2^{-k}$$

as $n \to \infty$. *For* $\epsilon > 0$,

$$\mathrm{P}(\ell_n \ge (1+\epsilon)\log_2(n)) = \mathrm{P}(\ell_n \ge \lceil (1+\epsilon)\log_2(n)\rceil) \le 2^{-\lceil (1+\epsilon)\log_2(n)\rceil} \le 2^{-(1+\epsilon)\log_2(n)} = \frac{1}{n^{1+\epsilon}}$$

is summable. By the Borel-Cantelli lemma,

$$\mathrm{P}\left\{\ell_n \ge (1+\epsilon)\log_2(n) \textit{ for infinitely many } n\right\} = 0.$$

Hence

$$\mathrm{P}\left\{\ell_n < (1+\epsilon)\log_2(n) \textit{ for all but finitely many } n\right\} = 1.$$

That is, for almost every ω, *there is* $N(\omega)$ *such that* $\ell_n < (1+\epsilon)\log_2(n)$ *for* $n \ge N(\omega)$. *For such* ω, *we have*

$$L_n(\omega) = \max_{m\le n} \ell_n(\omega) \le \max_{m\le n}(1+\epsilon)\log_2(n) = (1+\epsilon)\log_2(n)$$

as $n > N(\omega)$. *Thus*

$$\limsup_{n\to\infty} \frac{L_n}{\log_2(n)} \le 1+\epsilon$$

almost surely. Note that

$$\left\{\limsup_{n\to\infty} \frac{L_n}{\log_2(n)} \le 1+\epsilon\right\} \searrow \left\{\limsup_{n\to\infty} \frac{L_n}{\log_2(n)} \le 1\right\}$$

as $\epsilon \to 0^+$ and by the monotone convergence of the measures,

$$\limsup_{n\to\infty} \frac{L_n}{\log_2(n)} \leq 1$$

almost surely.

For the other side, note that for large n, we may split the sequence into blocks of size $\lceil (1-\epsilon)\log_2(n) \rceil$ and

$$\frac{n}{\lceil (1-\epsilon)\log_2(n) \rceil} \geq \frac{n}{\log_2(n)}$$

for large n.

$$\begin{aligned}
\mathrm{P}(L_n \leq (1-\epsilon)\log_2(n)) &\leq \mathrm{P}(\textit{each block did not have all 1s}) \\
&\leq (1 - 2^{-\lceil (1-\epsilon)\log_2(n)/2 \rceil})^{n/\lceil (1-\epsilon)\log_2(n)/2 \rceil} \\
&\leq \left(1 - \frac{1}{n^{1-\epsilon}}\right)^{n^{1-\epsilon}\frac{n^{\epsilon}}{\log_2(n)}} \leq \exp\left(-\frac{n^{\epsilon}}{\log_2(n)}\right),
\end{aligned}$$

whcih is summable, so by the Borel Cantelli lemma,

$$\mathrm{P}\left\{L_n \leq (1-\epsilon)\log_2(n) \textit{ for infinitely many } n\right\} = 0.$$

By a similar argument as above,

$$\liminf_{n\to\infty} \frac{L_n}{\log_2(n)} \geq 1 - \epsilon$$

almost surely and by the monotone convergence of the measures

$$\liminf_{n\to\infty} \frac{L_n}{\log_2(n)} \geq 1$$

almost surely. We conclude that

$$1 \leq \liminf_{n\to\infty} \frac{L_n}{\log_2(n)} \leq \limsup_{n\to\infty} \frac{L_n}{\log_2(n)} \leq 1$$

and the claim follows.

Example (Counting Process)
Let $X_i \in (0,\infty)$ be independent and identically distributed random variable. Put $\mu = \mathrm{E}[X_1]$, $T_n = \sum_{i=1}^{n} X_i$ and $N_t = \sup\{n \mid T_n \leq t\}$. Then we have the following claim:

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu}$$

almost surely. To see this, note that since $X_i < \infty$ for all i,

$$\lim_{t\to\infty} N_t = \lim_{t\to\infty} \sup\{n \mid T_n \le t\} = \infty.$$

Now, observe that $T_{N_t} \le t \le T_{N_t+1}$ and hence

$$\frac{T_{N_t}}{N_t} \le \frac{t}{N_t} \le \frac{T_{N_t+1}}{N_t+1}\frac{N_t+1}{N_t}.$$

By the strong law of large number, $T_{N_t}/N_t \to \mu$ almost surely. Thus

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu}.$$

Theorem 2.17 (Glivenko-Cantelli)
Suppose that $X_i \overset{iid}{\sim} F$ with $X_i \in (-\infty, \infty)$ and

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}$$

is the empirical CDF. Then

$$\|F_n - F\|_\infty \to 0$$

almost surely when $n \to \infty$.

Proof. We first claim that for $\epsilon > 0$, we may find a finite partition $\{t_j\}$ such that $-\infty = t_0 < \cdots < t_j = \infty$ and

$$F(t_{j+1}^-) - F(t_j) \le \epsilon$$

for all j. To see the existence of such partition, put $t_0 = -\infty$ and let

$$t_{j+1} = \sup\left\{t \in \mathbb{R} \,\middle|\, F(t) \le F(t_j) + \epsilon\right\}.$$

Observe that $F(t_{j+1}) \ge F(t_j) + \epsilon$. If not, then $F(t_{j+1}) < F(t_j) + \epsilon$. By the right-continuity of F, there is $\delta > 0$ such that $F(t_{j+1} + \delta) \le F(t_j) + \epsilon$, contradicting to the definition of t_{j+1}. It now also follows from the definition that

$$F(t_{j+1}^-) \le F(t_j) + \epsilon.$$

Finally, since F is of finite total variation, the jumps of sizes greater than ϵ can occur only finitely many times and we conclude the existence of such partition.

Next, by the strong law of large number, for almost every ω there is $N(\omega)$ uniform in j such that

$$\left|F_n(t_j) - F(t_j)\right| \le \epsilon$$

for all $n > N(\omega)$. For any $t \in [t_j, t_{j+1})$, we have

$$F(t) - F(t_j) \le F(t_{j+1}^{-1}) - F(t_j) \le \epsilon.$$

Again, by the strong law of the large number,

$$F_n(t_{j+1}^-) - F_n(t_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{t_j < X_i < t_{j+1}\right\} \to \mathrm{E}\left[\mathbf{1}\left\{t_j < X_i < t_{j+1}\right\}\right] = F(t_{j+1}^-) - F(t_j)$$

almost surely. That is, for almost every ω, there is $N'(\omega) > N(\omega)$ such that for all j,

$$F_n(t_{j+1}^-) - F_n(t_j) \le F(t_{j+1}^-) - F(t_j) + \epsilon$$

if $n \ge N'(\omega)$. Combining the above estimates, if $n \ge N'(\omega)$,

$$\begin{aligned}
|F_n(t) - F(t)| &\le \left|F_n(t) - F_n(t_j)\right| + \left|F_n(t_j) - F(t_j)\right| + \left|F(t_j) - F(t)\right| \\
&\le \left|F_n(t_{j+1}^-) - F_n(t_j)\right| + 2\epsilon \\
&\le F(t_{j+1}^-) - F(t_j) + 3\epsilon \le 4\epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we conclude that $F_n \to F$ uniformly for almost every ω and the proof is complete. ■

Theorem 2.18 (Kolmogorov Maximal Inequality)
Suppose that X_i are independent with $\mathrm{E}[X_i] = 0$ and $\mathrm{Var}[X_i] < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\mathrm{P}\left(\max_{1\le k\le n} |S_k| \ge x\right) \le \frac{1}{x^2}\mathrm{Var}[S_n].$$

Proof. Let $A_k = \left\{|S_k| \ge x \text{ and } \left|S_j\right| < x \text{ for } 1 \le j \le k-1\right\}$. Note that

$$\sum_{k=1}^{n} \mathbf{1}_{A_k} = \mathbf{1}\left\{\max_{1\le k\le n} |S_k| \ge x\right\}.$$

$$\begin{aligned}
\mathrm{E}\left[S_n^2\right] \ge \mathrm{E}\left[S_n^2 \sum_{k=1}^{n} \mathbf{1}_{A_k}\right] &= \sum_{k=1}^{n} \mathrm{E}\left[S_n^2 \mathbf{1}_{A_k}\right] = \sum_{k=1}^{n} \mathrm{E}\left[(S_n - S_k + S_k)^2 \mathbf{1}_{A_k}\right] \\
&= \sum_{k=1}^{n} \mathrm{E}\left[(S_n - S_k)^2 \mathbf{1}_{A_k}\right] + 2\,\mathrm{E}\left[(S_n - S_k)S_k \mathbf{1}_{A_k}\right] + \mathrm{E}\left[S_k^2 \mathbf{1}_{A_k}\right] \\
&\ge \sum_{k=1}^{n} \mathrm{E}\left[S_k^2 \mathbf{1}_{A_k}\right] + 2\,\mathrm{E}\left[(S_n - S_k)S_k \mathbf{1}_{A_k}\right].
\end{aligned}$$

Notice that $S_n - S_k \in \sigma(X_{k+1}, \ldots, X_n)$ and $S_k \mathbf{1}_{A_k} \in \sigma(X_1, \ldots, X_k)$ are independent. Thus

$$\begin{aligned} \mathrm{E}\left[S_n^2\right] &\geq \sum_{k=1}^{n} \mathrm{E}\left[S_k^2 \mathbf{1}_{A_k}\right] + 2\,\mathrm{E}\left[(S_n - S_k) S_k \mathbf{1}_{A_k}\right] \\ &= \sum_{k=1}^{n} \mathrm{E}\left[S_k^2 \mathbf{1}_{A_k}\right] \geq x^2 \sum_{k=1}^{n} \mathrm{E}\left[\mathbf{1}_{A_k}\right] = x^2\,\mathrm{P}\Big(\max_{1\leq k\leq n} |S_k| \geq x\Big). \end{aligned}$$

Hence

$$\mathrm{P}\Big(\max_{1\leq k\leq n} |S_k| \geq x\Big) \leq \frac{1}{x^2}\,\mathrm{E}\left[S_n^2\right].$$

■

Definition 2.19

*Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A sub-σ-algebra $\mathcal{G}$ is P-**trivial** if for all $A \in \mathcal{G}$, $\mathrm{P}(A) \in \{0, 1\}$.*

Theorem 2.20 (Kolmogorov Zero-One Law)

Let $\mathcal{F}_i$ be independent σ-algebras, $\mathcal{G}_n = \sigma(\mathcal{F}_n, \ldots)$ and $\mathcal{G}_\infty = \cap_{n=1}^{\infty} \mathcal{G}_n$. Then $\mathcal{G}_\infty$ is P-trivial.

Proof. Observe that a σ-algebra $\mathcal{G}$ satisfies that for $A \in \mathcal{G}$, $\mathrm{P}(A) \in \{0, 1\}$ if $\mathcal{G}$ is independent of itself. Indeed, if $\mathcal{G}$ is independent of itself, then for any $A \in \mathcal{G}$,

$$\mathrm{P}(A) = \mathrm{P}(A \cap A) = \mathrm{P}(A)^2$$

implies that $\mathrm{P}(A) = 0$ or 1. Now, for any given n, $\sigma(\mathcal{F}_1, \ldots, \mathcal{F}_{n-1})$ is independent with $\mathcal{G}_n$ and $\mathcal{G}_\infty \subset \mathcal{G}_n$. Hence $\mathcal{G}_\infty$ is independent of $\sigma(\mathcal{F}_1, \ldots, \mathcal{F}_{n-1})$ for all n.

In particular, $\mathcal{G}_\infty$ is independent of $\sigma(\cup_n \mathcal{F}_n)$. To see this, note that $\cup_n \sigma(\cup_{k=1}^{n} \mathcal{F}_k)$ is a π system that generates $\sigma(\cup_n \mathcal{F}_n)$ and for $A \in \cup_n \sigma(\cup_{k=1}^{n} \mathcal{F}_k)$, $A \in \sigma(\cup_{k=1}^{n} \mathcal{F}_k)$ for some n. For $B \in \mathcal{G}_\infty$,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

since $\sigma(\cup_{k=1}^{n} \mathcal{F}_k)$ and $\mathcal{G}_\infty$ is independent. Now it follows from theorem 1.52 that $\mathcal{G}_\infty$ and $\sigma(\cup_n \mathcal{F}_n)$ are independent. Notice that $\mathcal{G}_\infty \subset \sigma(\cup_n \mathcal{F}_n)$ and hence $\mathcal{G}_\infty$ is independent of itself. The proof is complete. ■

Corollary 2.21

Let X_i be independent and identically distributed and put $S_n = \sum_{i=1}^{n} X_i$. Then

(a) S_n is either almost surely convergent or almost surely divergent.

(b) If $\frac{1}{n} S_n$ converges almost surely, its limit is almost surely a constant.

Proof. Define $\mathcal{F}_i = \sigma(X_i)$ and note that $\{S_n \text{ converges}\} \in \mathcal{G}_\infty = \cap_n \sigma(\cup_{i\geq n} \mathcal{F}_i)$ since

$$\{S_n \text{ converges}\} = \Big\{\lim_{n\to\infty} \sum_{i\geq n} X_i = 0\Big\}$$

is $\mathcal{G}_\infty$-measurable. By the Kolmogorov zero-one law,

$$\mathrm{P}\left\{S_n \text{ converges}\right\} \in \{0,1\}.$$

This proves (a).

For (b), note that by a similar argument, we have

$$\left\{\frac{1}{n}S_n \text{ converges}\right\} \in \mathcal{G}_\infty,$$

where $\mathcal{G}_\infty$ is P-trivial. Also, $\lim_{n\to\infty} \frac{1}{n}S_n$ is $\mathcal{G}_\infty$-measurable. Since $\mathcal{G}_\infty$ is P-trivial,

$$F(t) := \mathrm{P}\left\{\lim_{n\to\infty} \frac{1}{n}S_n \le t\right\} \in \{0,1\}.$$

Thus

$$\lim_{n\to\infty} \frac{1}{n}S_n = \sup\{t \mid F(t) = 0\}$$

almost surely, proving that the limit is almost surely a constant. ∎

2.2. Convergence in Distribution

Definition 2.22
Let F_n and F be CDFs. We say that $F_n \to F$ ***in distrbution*** *or* ***weakly*** *if $F_n(x) \to F(x)$ for every x such that F is continuous at x, denoted as $F_n \xrightarrow{d} F$.*

Definition 2.23
Let X_n and X be random variables. $X_n \xrightarrow{d} X$ if the corresponding distributions $F_n \xrightarrow{d} F$.

Remark
If X_n, X are integer-valued, then $X_n \xrightarrow{d} X$ if and onliy if $\mathrm{P}(X_n = a) \to \mathrm{P}(X = a)$ for all $a \in \mathbb{Z}$.

Theorem 2.24 (Scheffé)
If f_n are density functions such that $f_n \to f$ almost everywhere, where f is a density function, then

$$\sup_{B \in \mathcal{B}} \left| \int_B f_n dx - \int_B f dx \right| \to 0$$

as $n \to 0$. In particular, taking $B = [-\infty, x]$ gives the uniform convergence of the CDFs.

Proof. Since

$$\sup_{B \in \mathcal{B}} \left| \int_B f_n dx - \int_B f dx \right| \leq \sup_{B \in \mathcal{B}} \int_B |f_n - f| \, dx \leq \int |f_n - f| \, dx,$$

the theorem follows once we prove that $f_n \to f$ in $\mathcal{L}^1$. Now, since $|f_n - f| \to 0$ almost everywhere and

$$|f_n - f| \leq |f_n| + |f| \quad \Rightarrow \quad 0 \leq |f_n| + |f| - |f_n - f| \, .$$

By the assumptions that f_n and f are density functions,

$$\int f_n dx = 1 = \int f dx.$$

By the Fatou's lemma,

$$\begin{aligned} 2 \int |f| \, dx &= \int \liminf_{n \to \infty} |f_n| + |f| - |f_n - f| \\ &\leq \liminf_{n \to \infty} \int f_n dx + \int f dx - \int |f_n - f| \, dx \\ &= 2 \int f dx - \limsup_{n \to \infty} \int |f_n - f| \, dx. \end{aligned}$$

Hence

$$\limsup_{n \to \infty} \int |f_n - f| \, dx \leq 0 \quad \Rightarrow \quad \int |f_n - f| \, dx \to 0.$$

Hence $f_n \to f$ in $\mathcal{L}^1$ and the proof is complete. ■

Proposition 2.25
If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

Proof. Let F_n and F be the corresponding CDFs for X_n and X. Suppose that x is a continuity point of F. For $\epsilon > 0$,

$$\begin{aligned} F_n(x) = \mathrm{P}(X_n \le x) &\ge \mathrm{P}(X_n \le X + \epsilon, X \le x) \ge \mathrm{P}(|X_n - X| \le \epsilon, X \le x) \\ &\ge \mathrm{P}(X \le x) - \mathrm{P}(|X_n - X| > \epsilon) = F(x) - \mathrm{P}(|X_n - X| > \epsilon) \end{aligned}$$

due to $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) \le \mathrm{P}(A \cap B) + \mathrm{P}(B^c)$ for measurable sets A, B. Taking $n \to \infty$ gives

$$\liminf_{n\to\infty} F_n(x) \ge F(x).$$

Similarly,

$$\begin{aligned} F(x + \epsilon) = \mathrm{P}(X \le x + \epsilon) &\ge \mathrm{P}(X \le X_n + \epsilon, X_n \le x) \\ &\ge \mathrm{P}(X_n \le x) - \mathrm{P}(|X_n - X| > \epsilon) = F_n(x) - \mathrm{P}(|X_n - X| > \epsilon). \end{aligned}$$

Taking $n \to \infty$ gives

$$\limsup_{n\to\infty} F_n(x) \le F(x + \epsilon).$$

Since ϵ is arbitrary, by the continuity of F at x we have

$$\limsup_{n\to\infty} F_n(x) \le F(x).$$

Hence

$$F(x) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x)$$

and we conclude that $F_n \xrightarrow{d} F$, i.e., $X_n \xrightarrow{d} X$. ■

Theorem 2.26 (Skorokhod Representation)
Suppose $F_n \xrightarrow{d} F$. Then there are corresponding random variables X_n, X for F_n and F such that $X_n \sim X$, $X \sim F$ and $X_n \to X$ almost surely.

Proof. Take $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}$ and P be the Lebesgue measure on $[0, 1]$. Put

$$X_n(\omega) = \sup\{x \in \mathbb{R} \mid F_n(x) < \omega\} \quad \text{and} \quad X(\omega) = \sup\{x \in \mathbb{R} \mid F(x) < \omega\}$$

with the convention that $\sup \varnothing = -\infty$. Then

$$\mathrm{P}\{X_n \le x\} = \mathrm{P}\{\omega \mid \omega \le F_n(x)\} = F_n(x) \quad \text{and} \quad \mathrm{P}\{X \le x\} = \mathrm{P}\{\omega \mid \omega \le F(x)\} = F(x).$$

It now suffices to show that $X_n \to X$ almost surely. Indeed, since F_n, F are CDFs, there are only at most countable discontinuities. Let ω be a point of continuity of X. We may find another continuity point such that $F(y) < \omega$. The convergence in distribution implies that

$F_n(y) \to F(y)$. Hence for n large enough, we have $F_n(y) < \omega$ and hence $X_n(\omega) > y$. Thus

$$\liminf_{n\to\infty} X_n(\omega) \geq y$$

for all $y \leq X(\omega)$. Thus

$$\liminf_{n\to\infty} X_n(\omega) \geq X(\omega).$$

Similarly, pick a continuity point y such that $F(y) \geq \omega$ would give

$$\limsup_{n\to\infty} X_n(\omega) \leq y$$

for all $y \geq X(\omega)$ and thus

$$\limsup_{n\to\infty} X_n(\omega) \leq X(\omega).$$

Combining the above results gives that $X_n \to X$ almost surely, since X is continuous almost surely. ■

Corollary 2.27
Let $g \geq 0$ be a continuous measurable function and $X_n \xrightarrow{d} X$. Then

$$\mathrm{E}\left[g(X)\right] \leq \liminf_{n\to\infty} \mathrm{E}\left[g(X_n)\right].$$

Proof. Let Y_n and Y be the Skorokhod representations for X_n and X, respectively. Since now $g(Y_n) \to g(Y)$ almost surely, the Fatou's lemma shows that

$$\mathrm{E}\left[g(X)\right] = \mathrm{E}\left[g(Y)\right] \leq \liminf_{n\to\infty} \mathrm{E}\left[g(Y_n)\right] = \liminf_{n\to\infty} \mathrm{E}\left[g(X_n)\right].$$

■

Theorem 2.28 (Helly-Bray)
Suppose that X_n and X are $\mathbb{R}$-valued random variables. Then $X_n \xrightarrow{d} X$ if and only if

$$\mathrm{E}\left[g(X_n)\right] \to \mathrm{E}\left[g(X)\right]$$

for all $g \in C_b(\mathbb{R})$.

Proof. Assume first that $X_n \xrightarrow{d} X$. By the Skorokhod representation theorem, we may assume that X_n and X are defined on the same probability space and $X_n \to X$ almost surely. Now, since for all $g \in C_b(\mathbb{R})$, $g(X_n) \to g(X)$ almost surely and are uniformly bounded, the bounded convergence theorem implies that

$$\mathrm{E}\left[g(X_n)\right] \to \mathrm{E}\left[g(X)\right].$$

Conversely, suppose that $\mathrm{E}\left[g(X_n)\right] \to \mathrm{E}\left[g(X)\right]$ for all $g \in C_b(\mathbb{R})$. Let F_n and F be the distribution functions for X_n and X respectively and x be a continuity point of F. For $\epsilon > 0$, consider

$$g_\epsilon(y) = \begin{cases} 1 & y \leq x \\ 1 - \frac{y-x}{\epsilon} & x < y \leq x + \epsilon \\ 0 & y \geq x + \epsilon. \end{cases}$$

Clearly $g_\epsilon \in C_b(\mathbb{R})$. Let $g(y) = \mathbf{1}\{y \leq x\}$.

$$\limsup_{n\to\infty} F_n(x) = \limsup_{n\to\infty} \mathrm{E}[g(X_n)] \leq \limsup_{n\to\infty} \mathrm{E}[g_\epsilon(X_n)] = \mathrm{E}[g_\epsilon(X)] \leq F(x+\epsilon).$$

Since ϵ is arbitrary and F is continuous at x, we have

$$\limsup_{n\to\infty} F_n(x) \leq F(x).$$

On the other hand,

$$\liminf_{n\to\infty} F_n(x) = \liminf_{n\to\infty} \mathrm{E}[g(X_n)] \geq \liminf_{n\to\infty} \mathrm{E}[g_\epsilon(X_n+\epsilon)] = \mathrm{E}[g_\epsilon(X+\epsilon)] \geq F(x+\epsilon) \geq F(x).$$

Hence $F_n(x) \to F(x)$ and the proof is complete. ■

Remark
The theorem gives an alternative characterization for the convergence in distribution. In particular, we can define the notion of convergence in distribution of general random element $X_n : \Omega \to (S, d)$ as $\mathrm{E}[g(X_n)] \to \mathrm{E}[g(X)]$ for all bounded $g : S \to \mathbb{R}$.

Theorem 2.29 (Continuous Mapping Theorem)
Let $X_n \xrightarrow{d} X$ and g be a measurable function continuous μ_X-almost surely. Then $g(X_n) \xrightarrow{d} g(X)$.

Proof. By the Skorokhod representation theorem, we may assume that X_n and X are on the same space and $X_n \to X$ almost surely. By the continuity of g, we have $g(X_n) \to g(X)$ almost surely. For all $f \in C_b(\mathbb{R})$, $f(g(X_n)) \to f(g(X))$ almost surely as well. Since $f \circ g$ is bounded, the bounded convergence theorem gives $\mathrm{E}[f(g(X_n))] \to \mathrm{E}[f(g(X))]$. By the Helly-Bray theorem, this implies that $g(X_n) \xrightarrow{d} g(X)$. ■

Remark
If g is bounded, then $\mathrm{E}[g(X_n)] \to \mathrm{E}[g(X)]$ directly by applying bounded convergence theorem on $g(X_n)$ and $g(X)$.

Example
$X_n \sim U[-\frac{1}{n}, \frac{1}{n}] \xrightarrow{d} \delta_0$. Let $g(x) = \mathbf{1}\{x \geq 0\}$. Then $g(X_n) \sim Ber(\frac{1}{2}) \overset{d}{\not\to} g(X) \sim \delta_1$.